Xiangxiang Huang; Yonghui Huang

Mean-variance optimality for semi-Markov decision processes under first passage criteria

Kybernetika, Vol. 53 (2017), No. 1, 59–81

Persistent URL: `http://dml.cz/dmlcz/146708`

KYBERNETIKA — VOLUME *53* (2017), NUMBER 1, PAGES 59–81

MEAN-VARIANCE OPTIMALITY FOR SEMI-MARKOV DECISION PROCESSES UNDER FIRST PASSAGE CRITERIA

Xiangxiang Huang and Yonghui Huang

This paper deals with a first passage mean-variance problem for semi-Markov decision processes in Borel spaces. The goal is to minimize the variance of a total discounted reward up to the system's first entry to some target set, where the optimization is over a class of policies with a prescribed expected first passage reward. The reward rates are assumed to be possibly unbounded, while the discount factor may vary with states of the system and controls. We first develop some suitable conditions for the existence of first passage mean-variance optimal policies and provide a policy improvement algorithm for computing an optimal policy. Then, two examples are included to illustrate our results. At last, we show how the results here are reduced to the cases of discrete-time Markov decision processes and continuous-time Markov decision processes.

Keywords: semi-Markov decision processes, first passage time, unbounded reward rate, minimal variance, mean-variance optimal policy

Classification: 90C40, 60J27

1. INTRODUCTION

Since Markowitz's work in 1950s [26], mean-variance problems have been an important class of stochastic optimization problems in economics and finance, where one seeks to minimize the variance of a total reward among policies with (at least) a certain expected reward. Compared with the classical optimality criteria concentrated on maximizing expected returns [3, 8, 16], mean-variance optimality criteria consider not only the mean of a random return but also the variability of the random return. The background of mean-variance problems arises from the tradeoff between the mean and variance, and the fact that a risk-aversion investor usually prefers to a return lower than the maximal one to keep a smaller variance risk. Due to this, mean-variance problems have been widely studied for various dynamic systems described by stochastic differential equations [5, 7, 22, 31], Markov decision processes (MDPs) [2, 3, 8, 10, 13, 21, 27, 28], and so on.

For the issue of mean-variance in MDPs, there have been a lot of references; see, [4, 19, 25, 28] for the finite horizon reward variance; [6, 10, 12, 20, 28, 30] for the infinite horizon discounted reward variance; [11, 24, 30] for the first passage variance;

DOI: 10.14736/kyb-2017-1-0059

and [2, 6, 8, 9, 13, 14, 21, 27, 29, 32] for the limiting average variance. To the best of our knowledge, most of the aforementioned works in MDPs focus on solving mean-variance problems in discrete-time MDPs (DTMDPs) [3, 4, 6, 13, 14, 21, 24, 25, 28, 30, 32] as well as in continuous-time MDPs (CTMDPs) [8, 9, 10, 11, 12, 20, 27], nevertheless, only a few works address mean-variance problems in semi-Markov decision processes (SMDPs); see [2, 28] for finite SMDPs and [19] with a finite time horizon. Moreover, it should be noted that most of the existing works on mean-variance problems for MDPs deal with *fixed* finite or infinite time horizons. As far as we know, a few works [11, 24, 30] deal with the first passage variance, where the early results on first passage variance in Section 2 of [24] is limited to the case of finite states and actions DTMDPs, where the existence of homogeneous Markovian controls with the maximal expected reward and the minimal variance is established. In many real-world situations, however, the control horizon may be a *random* duration. For example, in a maintenance system, people may be often concerned with the expected total repair costs before the system is restored; and in the medical research, one's interest may usually center on the expected total cost before he/she is cured. These situations motivate the first passage problems [16, 18], where the aim is usually to maximize the expected total reward before the systems fall in (or reach) a target set which represents the set of all good or bad states according to practical considerations. This paper attempts to consider a first passage mean-variance (FPMV) problem for SMDPs, i. e., the problem of minimizing the variance of a first passage total reward over a class of policies with a common expected first passage reward, which is a *new* issue in SMDPs. The purpose of selecting this issue is twofold. First, as explained above, both first passage problems and mean-variance problems are meaningful and significant topics in reality. Second, SMDPs are a sort of more general stochastic dynamic systems than DTMDPs and CTMDPs. As is known, the sojourn times in SMDPs are allowed to follow an arbitrary probability distribution, while the ones in DTMDPs are a fixed constant and the ones in CTMDPs are exponentially distributed.

The goal of the paper is to find an optimal policy with the minimal variance and a prescribed mean of a first passage total discounted reward. We assume that the state and control sets are Borel spaces, while the reward rates are possibly unbounded from both above and below. The discount factor may depend on states and controls, which is an extension of the usual constant ones in previous studies [6, 10, 12, 20, 28] and just state-dependent ones [30]. The consideration of a *varying* discount factor rather than a *fixed* constant one derives from the practical cases such as the interest rate in economic and financial systems [1, 15, 23], which can be adjusted according to the real circumstances. To investigate the FPMV optimality problem, we first characterize the policies with a common expected first passage reward (see Theorem 3.4), and then transform the variance of the first passage reward to the expected first passage reward with a new reward rate and another discount factor, which plays a crucial role in solving FPMV-optimal policies; see Theorem 3.5. Further, we establish the dynamic programming equation for our FPMV problem and the existence of FPMV-optimal policies under suitable conditions, and, in addition, we derive a value iteration algorithm and a policy improvement algorithm for calculating the value function and an FPMV-optimal policy, respectively; see Theorem 3.9. Then, two examples are shown to illustrate the application of our main result; see Examples 4.1 and 4.3. Finally, we exhibit the reduction of the results here.

The rest of this paper is organized as follows. Section 2 formulates the control model and the optimization problem. Our main results on the existence and computation of FPMV-optimal policies are stated in Section 3. Two examples are presented to illustrate our results in Section 4 before giving the reductions to the cases for DTMDPs and CTMDP in Section 5. We conclude with a summary in Section 6.

2. THE CONTROL MODEL

An FPMV model of SMDPs consists of the following objects

$$\{E, B, (A(x) \subset A, x \in E), Q(\cdot, \cdot|x, a), r(x, a), \alpha(x, a), g(x)\}, \tag{1}$$

where E is a Borel state space endowed with the Borel σ-field $\mathcal{B}(E)$, and A is a Borel action space endowed with the Borel σ-field $\mathcal{B}(A)$; $B \in \mathcal{B}(E)$ is a given target set such as a set of failure or working states; $A(x) \in \mathcal{B}(A)$ is the collection of all actions available to a controller at state $x \in E$. The semi-Markov kernel $Q(\cdot, \cdot|x, a)$ on $R_+ \times E$ given K describes the transition mechanism, where $R_+ := [0, +\infty)$ and $K := \{(x, a)|x \in E, a \in A(x)\}$ denotes the set of all feasible state-action pairs and is assumed to be a (Borel) measurable subset of $E\times A$. If an action $a \in A(x)$ is selected in state x, then $Q(t, D \mid x, a)$ is the joint probability that the sojourn time in state x is not greater than $t \in R_+$, and the next state is in $D \in \mathcal{B}(E)$. Furthermore, $r(x, a)$ and $\alpha(x, a)$ are measurable functions on K, denoting the reward rate and the discount factor, respectively. Finally, $g(x)$ is a measurable function on E, representing the mean reward one expects to earn.

Remark 2.1. (a) Our FPMV model (1) differs from the usual ones in previous studies in the following two aspects: first, a target set B (and thus a first passage problem) is introduced; second, the discount factor varies with states and controls rather than is a fixed constant.

(b) The reward rate can be negative, in which case it is interpreted as a cost rate. The target set B may represent the set of good or bad states, the interpretation of which depends on the practical consideration.

(c) Note that if the semi-Markov kernel $Q(\cdot, \cdot|x, a)$ is taken some particular forms, our model can be reduced to the corresponding one of CTMDPs [10, 11, 12, 20] or of DTMDPs [6, 24, 28, 30]; see Section 5 for further details.

To formulate the FPMV optimality problem, the concept of policies is needed.

Definition 2.2. A (deterministic stationary) policy is a measurable function from E to A such that $f(x) \in A(x)$ for every $x \in E$.

The collection of all such policies is denoted by F. Given $(s, x) \in R_+ \times E$ and $f \in F$, by the well-known Tulcea's theorem , there exist a unique probability space $(\Omega, \mathcal{F}, P^f_{(s,x)})$ and a stochastic process $\{T_n, J_n, A_n\}_{n\geq 0}$ such that for each $t \in R_+$, $C \in \mathcal{B}(E)$ and $n \geq 0$,

$$P^f_{(s,x)}(T_0 = s, J_0 = x) = 1, \tag{2}$$

$$P^f_{(s,x)}(A_n = f(J_n)|T_0, J_0, A_0, \ldots, T_{n-1}, J_{n-1}, A_{n-1}, T_n, J_n) = 1, \tag{3}$$

$$P^f_{(s,x)}(T_{n+1} - T_n \leq t, J_{n+1} \in C|T_0, J_0, A_0, \ldots, T_n, J_n, A_n) = Q(t, C|J_n, A_n), \tag{4}$$

where T_n, J_n, A_n denote the nth jump epoch, the state and the action taken at the nth jump epoch, respectively. For convenience, we denote by $\Theta_0 := 0, \Theta_n := T_n - T_{n-1} (n \geq 1)$ the sojourn times between two successive jump epochs. The expectation operator associated with $P^f_{(s,x)}$ is denoted by $E^f_{(s,x)}$. Particularly, we will write $P^f_{(0,x)}$ and $E^f_{(0,x)}$ as P^f_x and E^f_x, respectively. In what follows, we always set the initial jump epoch $T_0 := 0$ without loss of generality.

To avoid the possibility of an infinite number of jumps within finite time, we make the following assumption that the sequence $\{T_n\}_{n\geq 0}$ does't have finite accumulation points.

Assumption 2.3. For all $x \in E$ and $f \in F$, $P^f_x(\{\lim_{n\to\infty} T_n = \infty\}) = 1$.

Assumption 2.3 obviously holds for DTMDPs. To verify it, we give a sufficient condition below, which is the standard regular condition widely used in SMDPs; see, for instance, [16, 17, 18].

Condition 2.4. There exist constants $\delta > 0$ and $\varepsilon > 0$ such that

$$Q(\delta, E|x, a) \leq 1 - \varepsilon \quad \forall (x, a) \in K. \tag{5}$$

By Proposition 2.1 in [17], Condition 2.4 indeed implies Assumption 2.3. Moreover, it is more easily verified since the condition (5) is imposed on the *primitive* data of the model (1). Under Assumption 2.3, we define an underlying continuous-time state-action process $\{x(t), a(t), t \in R_+\}$ related to the stochastic process $\{T_n, J_n, A_n\}_{n\geq 0}$ by

$$x(t) = J_n, \;\; a(t) = A_n \quad \text{for } T_n \leq t < T_{n+1} \text{ and } n \geq 0.$$

Now, for the given target set $B \in \mathcal{B}(E)$, let

$$\tau_B := \begin{cases} \inf\{t \geq 0 : x(t) \in B\} & \text{if } \{t \geq 0 : x(t) \in B\} \neq \emptyset, \\ +\infty & \text{otherwise} \end{cases} \tag{6}$$

be the first passage time into the set B of the state process $\{x(t), t \in R_+\}$. Then, for each $x \in E$, *the first passage variance* under a policy $f \in F$ is defined as

$$\sigma^2(x, f) := E^f_x \left[\Big(\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s),a(s))\mathrm{d}s} r(x(t), a(t))\,\mathrm{d}t - V(x, f) \Big)^2 \right],$$

where $V(x, f)$, denoting *the first passage mean* of f, is given by

$$V(x, f) := E^f_x \Big[\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s),a(s))\mathrm{d}s} r(x(t), a(t))\,\mathrm{d}t \Big],$$

provided that these expectations are well-defined. Obviously, it follows from (6) that

$$V(x, f) = \sigma^2(x, f) = 0 \;\; \forall x \in B \text{ and } f \in F.$$

Thus, we will restrict our discussion to the case of the initial state $x \in B^c := E - B$ in the upcoming argument.

For the function g given in (1), let $F_g \subseteq F$ denote the collection of all policies having the same expected first passage reward g starting from every state in B^c, i. e.,

$$F_g := \{f \in F | \ V(x, f) = g(x) \ \ \forall x \in B^c\}.$$

F_g is assumed to be nonempty throughout the following. In most cases, there are usually more than one elements in F_g (see Examples 4.1 and 4.3). In this situation, it is natural to seek a policy with the minimal first passage variance in F_g, which leads to the FPMV (optimality) problem we are concerned with:

$$MV_g : \text{minimize } \sigma^2(x, f) \text{ over } f \in F_g \text{ for all } x \in B^c. \tag{7}$$

Our aim is to find a so-called FPMV-optimal policy $f^* \in F_g$ satisfying

$$\sigma^2(x, f^*) = \sigma^2_*(x) \quad \forall x \in B^c, \tag{8}$$

where $\sigma^2_*(x) := \inf_{f \in F_g} \sigma^2(x, f)$ is the FPMV value function, or simply, the value function.

Remark 2.5. (a) If the target set $B = \emptyset$(then $\tau_B = +\infty$) and $\alpha(x, a) \equiv \alpha$, in which case the optimality here is reduced to the standard infinite horizon discounted mean-variance problem. Hence, our FPMV problem is an improvement of the ones in [6, 28] for DTMDPs and the ones in [10, 12, 20] for CTMDPs.

(b) As is known, SMDPs are generalizations of DTMDPs and CTMDPs. That is, in SMDPs the sojourn times are allowed to follow an arbitrary probability distribution, while the ones in DTMDPs are a fixed constant and the ones in CTMDPs are exponentially distributed. In this setting, our PFMV problem extends the previous works on mean-variance problems for DTMDPs [6, 24, 28, 30] and CTMDPs [10, 11, 12, 20]; see Section 5 for more details.

(c) If g has the special form

$$g(x) = \sup_{f \in F} V(x, f) \ \ \forall x \in B^c,$$

then the corresponding FPMV problem is similar to the so-called variance minimization problem in [12, 20] for CTMDPs and [24, 30] for DTMDPs .

3. MAIN RESULTS

3.1. Characterization of policies in F_g

To solve the problem MV_g in (7), it is necessary to properly characterize the policies in F_g, which will be done in this subsection. First, we give some basic assumptions and preliminary facts.

Assumption 3.1. There exist constants $0 < \rho < 1$, $M > 0$, $\alpha_0 > 0$, and a measurable function $w \geq 1$ on E such that

(i) $|r(x,a)| \le Mw(x)$ and $\alpha_0 \le \alpha(x,a)$ for all $x \in B^c$,$a \in A(x)$.

(ii) $\int_{B^c} w^2(y) m_1(\mathrm{d}y|x,a) \le \rho^2 w^2(x)$ for all $x \in B^c$,$a \in A(x)$, where

$$m_1(\mathrm{d}y|x,a) := \int_0^\infty e^{-\alpha(x,a)t} Q(\mathrm{d}t, \mathrm{d}y|x,a).$$

Remark 3.2. Letting $\overline{r}(x,a) := r(x,a) \int_0^\infty e^{-\alpha(x,a)t}(1 - Q(t,E|x,a))\,\mathrm{d}t$, we have

$$|\overline{r}(x,a)| = |r(x,a)| \int_0^\infty e^{-\alpha(x,a)t}(1 - Q(t,E|x,a))\,\mathrm{d}t \le \frac{|r(x,a)|}{\alpha_0} \le \frac{Mw(x)}{\alpha_0}. \quad (9)$$

Moreover, Jensen's inequality gives

$$\int_{B^c} w(y) m_1(\mathrm{d}y|x,a) \le \rho w(x), \ x \in B^c \text{ and } a \in A(x),$$

which implies Assumption 11.2.3 in [18] for constrained first passage SMDPs with M and β replaced by $\frac{M}{\alpha_0}$ and ρ, respectively. The role of Assumption 3.1 is to ensure the uniqueness of the solution to the related dynamic programming equation (see Theorems 3.4 and 3.9) as well as the finiteness of $V(f)$ and $\sigma^2(f)$ (see Lemma 3.3).

To show the finiteness of $\sigma^2(f)$, it would be more convenient to consider the second moment instead of the variance. Thus, for each $x \in E$ and $f \in F$, we denote by

$$V^{(2)}(x,f) := E_x^f \left[\Big(\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s),a(s))ds} r(x(t),a(t))\,\mathrm{d}t \Big)^2 \right]$$

the second moment of the first passage discounted reward. For every $x \in B^c$, it is obvious that

$$\begin{aligned} V^{(2)}(x,f) &= \sigma^2(x,f) + V^2(x,f) \quad \forall f \in F & (10)\\ &= \sigma^2(x,f) + g^2(x) \quad \forall f \in F_g. & (11)\end{aligned}$$

Hence, the FPMV problem MV_g is equivalent to the following problem

$$\text{minimize } V^{(2)}(x,f) \text{ over all } f \in F_g \text{ for all } x \in B^c. \quad (12)$$

In relation to (10), for any given policy $f \in F$, the finiteness of $V^{(2)}(f)$ and $V(f)$ indicates the finiteness of $\sigma^2(f)$, for which we have the following fact.

Lemma 3.3. Suppose that Assumptions 2.3 and 3.1 are satisfied. Then, for each $x \in B^c$ and $f \in F$, we have

$$|V(x,f)| \le Mw(x)/\alpha_0(1-\rho), \text{ and } 0 \le V^{(2)}(x,f) < M^2 w^2(x)/\alpha_0^2(1-\rho)^2.$$

Proof. By Remark 3.2, Assumption 3.1 indicates Assumption 11.2.3 in [18]. Then, it follows from Lemma 11.3.1(a) in [18] that

$$|V(x,f)| \le Mw(x)/\alpha_0(1-\rho) \ \forall x \in B^c, f \in F.$$

We now turn to proving the second statement. Since the positive of the discount factor implies $\int_0^\infty e^{-2\alpha(x,a)t}Q(\mathrm{d}t, D|x,a) < m_1(D|x,a)$ for all $D \in \mathcal{B}(E)$ and $(x,a) \in K$, we obtain

$$\int_{B^c} w^2(y)\int_0^\infty e^{-2\alpha(x,a)t}Q(\mathrm{d}t,\mathrm{d}y|x,a) < \int_{B^c} w^2(y)m_1(\mathrm{d}y|x,a) \ \ \forall (x,a)\in K,$$

which, together with Assumption 3.1(ii), yields

$$\int_{B^c} w^2(y)m_2(\mathrm{d}y|x,a) < \rho^2 w^2(x) \ \ \forall x \in B^c, a\in A(x) \tag{13}$$

with m_2 defined by

$$m_2(\mathrm{d}y|x,a) := \int_0^\infty e^{-2\alpha(x,a)t}Q(\mathrm{d}t,\mathrm{d}y|x,a). \tag{14}$$

Under Assumptions 2.3 and 3.1(ii), by (13) and the same argument to (11.11) in Lemma 11.3.1(a) of [18] with ρ^2, w^2 and 2α in lieu of β, w and α, respectively, we have

$$E_x^f\Big[\prod_{k=0}^{n-1} e^{-2\alpha(J_k,A_k)\Theta_{k+1}}I_{\{J_0\in B^c,\dots,J_n\in B^c\}}w^2(J_n)\Big] < \rho^{2n}w^2(x) \ \ \forall x\in B^c, f\in F. \tag{15}$$

Here and below, I_D is the indicator function on a set D, and $\prod_{k=n}^{m} y_k = 1$ when $m < n$ for any sequence $\{y_k\}$.

By Lemma 11.3.1 in [18], the second moment $V^{(2)}(f)$ can be expressed as

$$V^{(2)}(x,f) = E_x^f\Big[\Big(\sum_{n=0}^\infty\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}}I_{\{J_0\in B^c,\dots,J_n\in B^c\}}r(J_n,A_n)\Delta_{n+1}\Big)^2\Big] \qquad \forall x\in B^c, f\in F, \tag{16}$$

with $\Delta_{n+1} := \int_0^{\Theta_{n+1}} e^{-\alpha(J_n,A_n)t}\,\mathrm{d}t$, which, along with (15) and a straightforward calculation, gives that

$$\begin{aligned} V^{(2)}(x,f) &\le E_x^f\Big[\Big(\sum_{n=0}^\infty\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}}I_{\{J_0\in B^c,\dots,J_n\in B^c\}}|r(J_n,A_n)|\Delta_{n+1}\Big)^2\Big] \\ &= E_x^f\Big[\sum_{n=0}^\infty\sum_{k+l=n}\Big(\prod_{m=0}^{k-1} e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\dots,J_k\in B^c\}}|r(J_k,A_k)|\Delta_{k+1}\Big) \\ &\qquad\times\Big(\prod_{m=0}^{l-1} e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\dots,J_l\in B^c\}}|r(J_l,A_l)|\Delta_{l+1}\Big)\Big] \\ &= \sum_{n=0}^\infty\sum_{k+l=n} E_x^f\Big[\Big(\prod_{m=0}^{k-1} e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\dots,J_k\in B^c\}}|r(J_k,A_k)|\Delta_{k+1}\Big) \\ &\qquad\times\Big(\prod_{m=0}^{l-1} e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\dots,J_l\in B^c\}}|r(J_l,A_l)|\Delta_{l+1}\Big)\Big] \end{aligned}$$

$$
\begin{aligned}
&\le \sum_{n=0}^{\infty}\sum_{k+l=n}\Big\{E_x^f\Big[\Big(\prod_{m=0}^{k-1}e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_k\in B^c\}}|r(J_k,A_k)|\Delta_{k+1}\Big)^2\Big]\Big\}^{\frac{1}{2}}\\
&\quad\times\Big\{E_x^f\Big[\Big(\prod_{m=0}^{l-1}e^{-\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_l\in B^c\}}|r(J_l,A_l)|\Delta_{l+1}\Big)^2\Big]\Big\}^{\frac{1}{2}}\\
&= \sum_{n=0}^{\infty}\sum_{k+l=n}\Big[E_x^f\Big(\prod_{m=0}^{k-1}e^{-2\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_k\in B^c\}}r^2(J_k,A_k)\big(\Delta_{k+1}\big)^2\Big)\Big]^{\frac{1}{2}}\\
&\quad\times\Big[E_x^f\Big(\prod_{m=0}^{l-1}e^{-2\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_l\in B^c\}}r^2(J_l,A_l)\big(\Delta_{l+1}\big)^2\Big)\Big]^{\frac{1}{2}}\\
&\le \frac{M^2}{\alpha_0^2}\sum_{n=0}^{\infty}\sum_{k+l=n}\Big[E_x^f\Big(\prod_{m=0}^{k-1}e^{-2\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_k\in B^c\}}w^2(J_k)\Big)\Big]^{\frac{1}{2}}\\
&\quad\times\Big[E_x^f\Big(\prod_{m=0}^{l-1}e^{-2\alpha(J_m,A_m)\Theta_{m+1}}I_{\{J_0\in B^c,\ldots,J_l\in B^c\}}w^2(J_l)\Big)\Big]^{\frac{1}{2}}\\
&< \frac{M^2}{\alpha_0^2}\sum_{n=0}^{\infty}\sum_{k+l=n}[\rho^{2k}w^2(x)]^{\frac{1}{2}}[\rho^{2l}w^2(x)]^{\frac{1}{2}}\\
&= \frac{M^2}{\alpha_0^2}w^2(x)\sum_{n=0}^{\infty}(n+1)\rho^n = M^2w^2(x)/\alpha_0^2(1-\rho)^2.
\end{aligned}
$$

On the other hand, by the definition of $V^{(2)}(f)$, it directly follows that $V^{(2)}(x,f)\ge 0$ for every $x\in B^c, f\in F$, which completes the proof. □

To facilitate the upcoming argument, we introduce the concept of the w-weighted norm similar to that in [18] with w as in Assumption 3.1. A (real-valued) function u on B^c is called w-bounded if the w-weighted norm of u, i. e., $\|u\|_w := \sup_{x\in B^c}\frac{|u(x)|}{w(x)}$, is finite. The Banach space of all w-bounded measurable functions on B^c is denoted by $M_w(B^c)$. Obviously, from Lemma 3.3 we have

$$V(\cdot,f)\in M_w(B^c) \text{ and } V^{(2)}(\cdot,f)\in M_{w^2}(B^c)$$

for each $f\in F$.

Since our optimization is over all policies in F_g, it is helpful to characterize elements of F_g in terms of primitive data in (1). The following theorem gives such a characterization.

Theorem 3.4. Under Assumptions 2.3 and 3.1, the following assertions hold.

(a) For each $f\in F$, $V(\cdot,f)$ is a unique solution in $M_w(B^c)$ to the equation

$$u(x)=\overline{r}(x,f(x))+\int_{B^c}u(y)m_1(\mathrm{d}y|x,f(x))\quad \forall x\in B^c,$$

with $\overline{r}$ as in (9).

(b) A policy $f \in F_g$ if and only if $f(x) \in A_g(x)$ for every $x \in B^c$, where $A_g(x)$ is given by

$$A_g(x) := \big\{a \in A(x) | g(x) = \bar{r}(x,a) + \int_{B^c} g(y) m_1(\mathrm{d}y|x,a)\big\}, \quad x \in B^c. \tag{17}$$

Proof. (a) As indicated in Lemma 3.3, $V(\cdot, f)$ is in $M_w(B^c)$ under Assumptions 2.3 and 3.1. In addition, by Remark 3.2, $\int_{B^c} w(y) m_1(\mathrm{d}y|x,a) \le \rho w(x)$ for every $x \in B^c$ and $a \in A(x)$, and thus, the statement can be obtained using the similar way as in Lemma 11.3.2(a) in [18].

(b) Since the nonempty of F_g implies $g \in M_w(B^c)$, part (a), together with the definition of F_g, completes the proof. □

3.2. Transformation of the second moment into a mean

As mentioned in (12), our original problem MV_g is equivalent to minimizing $V^{(2)}(f)$ over F_g. In this subsection, we will place our concern on the second moment $V^{(2)}(f)$ and show how to transform it into a mean.

Theorem 3.5. Under Assumptions 2.3 and 3.1, for each $f \in F_g$, $V^{(2)}(\cdot, f)$ is the unique solution in $M_{w^2}(B^c)$ to the following equation

$$u(x) = R(x, f(x)) + \int_{B^c} u(y) m_2(\mathrm{d}y|x, f(x)), \ x \in B^c, \tag{18}$$

where m_2 is as in (14) and

$$\begin{aligned} R(x,a) : \ &= \ \frac{r^2(x,a)}{\alpha^2(x,a)}[1 + m_2(E|x,a) - 2m_1(E|x,a)] \\ &\quad + \frac{2r(x,a)}{\alpha(x,a)} \int_{B^c} g(y)[m_1(\mathrm{d}y|x,a) - m_2(\mathrm{d}y|x,a)]. \end{aligned}$$

Moreover, the second moment $V^{(2)}(f)$ is expressed by means of the first moment, i. e.,

$$V^{(2)}(x,f) = E_x^f \left[\int_0^{\tau_B} e^{-\int_0^t 2\alpha(x(s),a(s))\,\mathrm{d}s} c_g(x(t), a(t))\,\mathrm{d}t \right] =: J(x,f),$$

where

$$c_g(x,a) := \frac{R(x,a)}{\int_0^\infty e^{-2\alpha(x,a)t}(1 - Q(t,E|x,a))\,\mathrm{d}t} \quad \forall (x,a) \in K_g \tag{19}$$

with $K_g := \{(x,a) | x \in B^c, a \in A_g(x)\}$.

Proof. Let $f \in F_g$ and $x \in B^c$ be arbitrarily fixed. By (16), a direct calculation gives

$$\begin{aligned} V^{(2)}(x,f) &= E_x^f\Big[\Big(\sum_{n=0}^{\infty}\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\Big] \\ &= E_x^f\Big[\Big(r(J_0,A_0)I_{\{J_0\in B^c\}}\Delta_1 \\ &\qquad +\sum_{n=1}^{\infty}\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\Big] \\ &=: L_1+L_2+L_3, \end{aligned}$$

where

$$\begin{aligned} L_1 &:= E_x^f\Big[\Big(r(J_0,A_0)I_{\{J_0\in B^c\}}\Delta_1\Big)^2\Big], \\ L_2 &:= 2E_x^f\Big[r(J_0,A_0)I_{\{J_0\in B^c\}}\Delta_1 \\ &\qquad \times\Big(\sum_{n=1}^{\infty}\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)\Big], \\ L_3 &:= E_x^f\Big[\Big(\sum_{n=1}^{\infty}\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\Big]. \end{aligned}$$

Noting that $J_0 = x \in B^c$, we have

$$\begin{aligned} L_1 &= E_x^f\Big[r^2(J_0,A_0)I_{\{J_0\in B^c\}}\Big(\frac{1-e^{-\alpha(J_0,A_0)\Theta_1}}{\alpha(J_0,A_0)}\Big)^2\Big] \\ &= \frac{r^2(x,f(x))}{\alpha^2(x,f(x))}\Big[1+\int_0^{+\infty}\big(e^{-2\alpha(x,f(x))t}-2e^{-\alpha(x,f(x))t}\big)Q(\mathrm{d}t,E|x,f(x))\Big] \\ &= \frac{r^2(x,f(x))}{\alpha^2(x,f(x))}[1+m_2(E|x,f(x))-2m_1(E|x,f(x))]. \end{aligned}$$

Furthermore, it follows from the property of conditional expectation and the Markov property that

$$\begin{aligned} L_2 &= 2E_x^f\Big\{E_x^f\Big[\Big(\sum_{n=1}^{\infty}\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big) \\ &\qquad \times r(J_0,A_0)I_{\{J_0\in B^c\}}\Delta_1|T_0,J_0,A_0,T_1,J_1\Big]\Big\} \\ &= 2E_x^f\Big[r(J_0,A_0)e^{-\alpha(J_0,A_0)\Theta_1}I_{\{J_0\in B^c,J_1\in B^c\}}\Delta_1 \\ &\qquad \times E_x^f\Big(\sum_{n=1}^{\infty}\prod_{k=1}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\dots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}|T_0,J_0,A_0,T_1,J_1\Big)\Big] \end{aligned}$$

$$\begin{aligned}
&= \frac{2r(x,f(x))}{\alpha(x,f(x))}\int_{B^c}\int_0^\infty [e^{-\alpha(x,f(x))t}-e^{-2\alpha(x,f(x))t}]Q(\mathrm{d}t,\mathrm{d}y|x,f(x))\\
&\quad \times E_x^f\Big(\sum_{n=1}^\infty\prod_{k=1}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\\
&\qquad |T_0=0,J_0=x,A_0=f(x),T_1=t,J_1=y\Big)\\
&= \frac{2r(x,f(x))}{\alpha(x,f(x))}\int_{B^c}[m_1(\mathrm{d}y|x,f(x))-m_2(\mathrm{d}y|x,f(x))]\\
&\quad E_y^f\Big(\sum_{n=0}^\infty\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)\\
&= \frac{2r(x,f(x))}{\alpha(x,f(x))}\int_{B^c}[m_1(\mathrm{d}y|x,f(x))-m_2(\mathrm{d}y|x,f(x))]V(y,f)\\
&= \frac{2r(x,f(x))}{\alpha(x,f(x))}\int_{B^c} g(y)[m_1(\mathrm{d}y|x,f(x))-m_2(\mathrm{d}y|x,f(x))]
\end{aligned}$$

where the third equality is due to the properties (2)−(4).

Similarly, we have

$$\begin{aligned}
L_3 &= E_x^f\Big\{E_x^f\Big[\Big(\sum_{n=1}^\infty\prod_{k=0}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\\
&\qquad \Big|T_0,J_0,A_0,T_1,J_1\Big]\Big\}\\
&= E_x^f\Big\{e^{-2\alpha(J_0,A_0)\Theta_1} I_{\{J_0\in B^c,J_1\in B^c\}} E_x^f\Big[\Big(\sum_{n=1}^\infty\prod_{k=1}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}}\\
&\qquad I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\Big|T_0,J_0,A_0,T_1,J_1\Big]\Big\}\\
&= \int_{B^c}\int_0^\infty e^{-2\alpha(x,f(x))t}Q(\mathrm{d}t,\mathrm{d}y|x,f(x))\\
&\quad \times E_x^f\Big[\Big(\sum_{n=1}^\infty\prod_{k=1}^{n-1} e^{-\alpha(J_k,A_k)\Theta_{k+1}} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\Delta_{n+1}\Big)^2\\
&\qquad \Big|T_0=0,J_0=x,A_0=f(x),T_1=t,J_1=y\Big]\\
&= \int_{B^c} V^{(2)}(y,f)m_2(\mathrm{d}y|x,f(x)).
\end{aligned}$$

Therefore, taking all the above results of L_1, L_2, L_3 into consideration, we obtain

$$V^{(2)}(x,f) = R(x,f(x)) + \int_{B^c} V^{(2)}(y,f)m_2(\mathrm{d}y|x,f(x)),$$

which, together with Lemma 3.3, shows that $V^{(2)}(f)$ is a solution in $M_{w^2}(B^c)$ to the equation (18).

On the other hand, consider the following expected first passage cost

$$J(x,f)=E_x^f\Big[\int_0^{\tau_B} e^{-\int_0^t 2\alpha(x(s),a(s))\,\mathrm{d}s}c_g(x(t),a(t))\,\mathrm{d}t\Big]$$

with $c_g(x,a)$ as in (19). Under Assumption 3.1, for all $x\in B^c$ and $a\in A(x)$, we have

$$\begin{aligned}\Big|c_g(x,a)\int_0^\infty e^{-2\alpha(x,a)t}(1-Q(t,E|x,a))\,\mathrm{d}t\Big| &= \big|R(x,a)\big|\\ &< \frac{3M^2+2M\rho\alpha_0\|g\|_w}{\alpha_0^2}w^2(x), \end{aligned}\tag{20}$$

and $\int_{B^c}w^2(y)m_2(\mathrm{d}y|x,a)<\rho^2w^2(x)$ (see (13)). Thus, a similar argument to the finiteness of $V(f)$ in Lemma 3.3 yields

$$|J(x,f)|<\frac{3M^2+2M\rho\alpha_0\|g\|_w}{\alpha_0^2(1-\rho^2)}w^2(x),\tag{21}$$

and hence $J(\cdot,f)\in M_{w^2}(B^c)$ for all $f\in F_g$. Moreover, letting w in Theorem 3.4(a) be replaced with w^2, we conclude that $J(f)$ is the unique solution in $M_{w^2}(B^c)$ to the equation (18). Recalling that $V^{(2)}(f)$ is a solution in $M_{w^2}(B^c)$ to (18), we immediately obtain that $V^{(2)}(x,f)=J(x,f)$ for each $x\in B^c$ and $f\in F_g$. □

Remark 3.6. In fact, Theorem 3.5 holds for any policy $f\in F$ with $g(x)$ replaced by $V(x,f)$ for each $x\in B^c$. More precisely, as an argument of Theorem 3.5, we get that $V^{(2)}(\cdot,f)$ is the unique solution within $M_{w^2}(B^c)$ to the following equation

$$u(x)=\bar{R}(x,f)+\int_{B^c}u(y)m_2(\mathrm{d}y|x,f(x))\ \ \forall x\in B^c, f\in F,$$

with

$$\begin{aligned}\bar{R}(x,f):\ &=\ \frac{r^2(x,f(x))}{\alpha^2(x,f(x))}[1+m_2(E|x,f(x))-2m_1(E|x,f(x))]\\ &\quad +\frac{2r(x,f(x))}{\alpha(x,f(x))}\int_{B^c}V(y,f)[m_1(\mathrm{d}y|x,f(x))-m_2(\mathrm{d}y|x,f(x))].\end{aligned}$$

Furthermore, for each $f\in F$, the second moment $V^{(2)}(x,f)$ has the following expression

$$V^{(2)}(x,f)=E_x^f\left[\int_0^{\tau_B} e^{-\int_0^t 2\alpha(x(s),f(x(s)))\,\mathrm{d}s}\overline{c}(x(t),f)\,\mathrm{d}t\right],$$

where

$$\overline{c}(x,f):=\frac{\bar{R}(x,f)}{\int_0^\infty e^{-2\alpha(x,f(x))t}(1-Q(t,E|x,f(x)))\,\mathrm{d}t}\quad \forall x\in B^c.$$

3.3. On the existence of FPMV-optimal policies

In this subsection, we will establish the existence and computation of an FPMV-optimal policy via the so-called FPMV optimality equation (23) below.

From Theorem 3.5 and the relation (11), the original problem MV_g (7) can be further reduced to

$$\text{minimize } J(x,f) \text{ over all } f \in F_g \text{ for all } x \in B^c,$$

which is a classical expectation (rather than variance) minimization problem and can be solved via the following first passage SMDPs model

$$\{E, B, (A_g(x) \subset A, x \in E), Q(\cdot,\cdot|x,a), c_g(x,a), 2\alpha(x,a)\}, \tag{22}$$

with c_g as in (19), A_g as in (17) and the other data as in (1). In this setup, the existing results on first passage SMDPs [18] can be applied.

For simplicity of notation, we introduce dynamic programming operators T^f and T on $M_{w^2}(B^c)$ as follows: for each $f \in F_g$, $x \in B^c$ and $u \in M_{w^2}(B^c)$, with $R(x,a)$ as in Theorem 3.5,

$$\begin{aligned} T^f u(x) &:= R(x, f(x)) + \int_{B^c} u(y) m_2(\mathrm{d}y|x, f(x)), \\ Tu(x) &:= \inf_{a \in A_g(x)} \Big\{ R(x,a) + \int_{B^c} u(y) m_2(\mathrm{d}y|x,a) \Big\}. \end{aligned}$$

To obtain the existence of the FPMV-optimal policy, we also require the continuous-compactness condition as below.

Assumption 3.7. Let w and m_1 be as in Assumption 3.1, and m_2 be as in (14).

(i) $A(x)$ is compact for each $x \in B^c$.

(ii) For each fixed $x \in B^c$, $t \in R_+$, Borel set $D \subset B^c$ and $D = E$, the functions $r(x,a)$, $\alpha(x,a)$, and $Q(t, D|x,a)$ are continuous in $a \in A(x)$.

(iii) The functions $\int_{B^c} w(y) m_1(\mathrm{d}y|x,a)$ and $\int_{B^c} w^2(y) m_2(\mathrm{d}y|x,a)$ are continuous in $a \in A(x)$ for each $x \in B^c$.

Lemma 3.8. Under Assumptions 3.1 and 3.7, for each fixed $x \in B^c$, we have the following statements.

(a) $m_1(D|x,a)$ and $m_2(D|x,a)$ are continuous in $a \in A(x)$ for every Borel set $D \subset B^c$ and $D = E$.

(b) $u'(x,a) := \int_{B^c} u(y) m_1(\mathrm{d}y \mid x,a)$ and $v'(x,a) := \int_{B^c} v(y) m_2(\mathrm{d}y \mid x,a)$ are continuous in $a \in A(x)$ for every function $u \in M_w(B^c)$ and $v \in M_{w^2}(B^c)$, respectively.

(c) $A_g(x)$ is compact.

Proof. (a) For each fixed $x \in B^c$ and Borel set $D \subset B^c$, let $\{a_n\} \subset A(x)$ be an arbitrary sequence such that $a_n \to a \in A(x)$, and m_1 be as in Assumption 3.1. For every n, we have

$$m_1(D|x, a_n) \quad = \quad \int_0^\infty e^{-\alpha(x,a_n)t} Q(\mathrm{d}t, D|x, a_n).$$

Letting $n \to \infty$ in the above equality, by Assumption 3.7(ii) and the generalized dominated convergence theorem in Proposition A.4 [8], we conclude that $m_1(D|x, a)$ is continuous in $a \in A(x)$ for each $x \in B^c$ and Borel set $D \subset B^c$ and $D = E$. Similarly, the conclusion for $m_2(D|x, a)$ can be achieved.

(b) Using a similar argument to the proof of Lemma 8.3.7 in [13], part (b) follows from Assumption 3.7(iii) and part (a).

(c) Let $x \in B^c$ be arbitrarily fixed. To show that $A_g(x)$ is compact, it suffices to prove that $A_g(x)$ is closed because $A_g(x) \subset A(x)$ and $A(x)$ is compact. Indeed, let $\{a_n\} \subset A_g(x)$ be an arbitrary sequence such that $a_n \to a \in A(x)$. Then, for each n, we have

$$g(x) = \overline{r}(x, a_n) + \int_{B^c} g(y) m_1(\mathrm{d}y \mid x, a_n)$$

Since $g \in M_w(B^c)$, it follows from part (b) that $\int_{B^c} g(y) m_1(\mathrm{d}y \mid x, a)$ is continuous in $a \in A(x)$. Moreover, $\overline{r}(x, a_n)$ is continuous in $a \in A(x)$ by the dominated convergence theorem. Thus, taking $n \to \infty$ in the above equality, under Assumption 3.7, we obtain

$$g(x) = \overline{r}(x, a) + \int_{B^c} g(y) m_1(\mathrm{d}y \mid x, a),$$

which shows that $a \in A_g(x)$. □

Now, we are ready to state the main result concerning the existence and computation of FPMV-optimal policies.

Theorem 3.9. Under Assumptions 2.3, 3.1 and 3.7, the following assertions hold.

(a) $(\sigma_*^2 + g^2)$ is the unique solution within $M_{w^2}(B^c)$ to the so-called FPMV optimality equation $u(x) = Tu(x)$, i. e.,

$$(\sigma_*^2 + g^2)(x) = T(\sigma_*^2 + g^2)(x) \quad \forall x \in B^c, \tag{23}$$

where σ_*^2 is as in (8).

(b) A policy $f \in F_g$ is FPMV-optimal if and only if $f(x)$ attains the minimum in (23) for each $x \in B^c$, i.e, $(\sigma_*^2 + g^2)(x) = T^f(\sigma_*^2 + g^2)(x), \forall x \in B^c$.

(c) There exists a policy f^* such that $(\sigma_*^2 + g^2) = T^{f^*}(\sigma_*^2 + g^2)$, and such a policy f^* is FPMV-optimal.

(d) The value function σ_*^2 can be approximated by the following iteration sequence:

$$\sigma_*^2 = \lim_{n \to \infty} u_n - g^2 \text{ with } u_{n+1} := Tu_n, n \geq 0, \text{ and } u_0 = 0.$$

(e) An FPMV-optimal policy can be obtained by the following algorithm:

Policy improvement algorithm:

(1) For a given expected reward g, compute $A_g(x)$ for each $x \in B^c$, and then get F_g by Theorem 3.4(b).

(2) Pick an arbitrary policy $f \in F_g$. Let $k = 0$ and set $h_k := f$.

(3) Policy evaluation: Compute $J(h_k) = \sigma^2(h_k) + g^2$ as the unique solution in $M_{w^2}(B^c)$ to the equation $u(x) = T^{h_k}u(x)$ for all $x \in B^c$.

(4) Policy improvement: For any $k \geq 0$, take $h_{k+1} \in F$ as follows:

$$h_{k+1}(x) \in D(h_k, x) \text{ if } D(h_k, x) \neq \emptyset, \text{ and}$$
$$h_{k+1}(x) := h_k(x) \text{ if } D(h_k, x) = \emptyset, x \in B^c,$$

where

$$D(h_k, x) := \{a \in A_g(x) | R(x,a) + \int_{B^c} J(y, h_k) m_2(\mathrm{d}y|x,a) < J(x, h_k)\}.$$

(5) If $h_{k+1}(x) = h_k(x)$ for all $x \in B^c$, then, h_k is FPMV-optimal. Otherwise, increase k by 1 and return to step (3).

Proof. (a) First, we prove that T is a contraction operator from $M_{w^2}(B^c)$ to itself. Indeed, by Lemma 3.8, $R(x,a) + \int_{B^c} u(y) m_2(\mathrm{d}y|x,a)$ is continuous in $a \in A(x)$ for each $x \in B^c$ and $u \in M_{w^2}(B^c)$. Then, it follows from the measurable selection theorem (e. g., Lemma 8.3.8 in [13]) that there exists $f \in F$ such that $Tu(x) = T^f u(x)$. Therefore, $Tu(\cdot)$ is measurable on B^c for all $u \in M_{w^2}(B^c)$, and for any function $u \in M_{w^2}(B^c)$, by (20) and (13), we have

$$\begin{aligned}
|Tu(x)| &\leq \big|R(x, f(x))\big| + \Big|\int_{B^c} u(y) m_2(\mathrm{d}y|x, f(x))\Big| \\
&< \frac{3M^2 + 2M\rho\alpha_0\|g\|_w}{\alpha_0^2} w^2(x) + \|u\|_{w^2}\rho^2 w^2(x) \\
&= \big((3M^2 + 2M\rho\alpha_0\|g\|_w)/\alpha_0^2 + \|u\|_{w^2}\rho^2\big) w^2(x) \quad \forall x \in B^c,
\end{aligned}$$

which shows that $Tu(\cdot)$ is w^2-bounded.

Furthermore, for each $u, v \in M_{w^2}(B^c)$, we have

$$\begin{aligned}
& |Tu(x) - Tv(x)| \\
= & \Big| - \sup_{a \in A_g(x)} \Big[- R(x,a) - \int_{B^c} u(y) m_2(\mathrm{d}y|x,a)\Big] \\
& + \sup_{a \in A_g(x)} \Big[- R(x,a) - \int_{B^c} v(y) m_2(\mathrm{d}y|x,a)\Big]\Big| \\
\leq & \sup_{a \in A_g(x)} \Big| \int_{B^c} [u(y) - v(y)] m_2(\mathrm{d}y|x,a)\Big| \\
\leq & \|u - v\|_{w^2} \sup_{a \in A_g(x)} \int_{B^c} w^2(y) m_2(\mathrm{d}y|x,a) \\
< & \|u - v\|_{w^2} \rho^2 w^2(x) \quad \forall x \in B^c.
\end{aligned}$$

Hence, $\|Tu - Tv\|_{w^2} < \|u - v\|_{w^2}\rho^2$, and thus T is a contraction operator from the Banach space $M_{w^2}(B^c)$ to itself. By Banach's Fixed Point Theorem, T has a unique fixed point u_* in $M_{w^2}(B^c)$. Note that $J_*(\cdot) := \inf_{f\in F_g} J(\cdot, f)$ is in $M_{w^2}(B^c)$ (see (21)). Thus, it remains to prove $u_* = J_*$, which, for the model (22) with w^2, 2α and c_g in lieu of w, α and r in [18], respectively, can be verified by the similar manner as in Lemma 11.3.2(b) of [18]. Finally, using Theorem 3.5 again, we have $J_*(\cdot) = \sigma_*^2(\cdot) + g^2(\cdot)$.

(b) By Theorem 3.5, we see that $J(f) = \sigma^2(f) + g^2$ is the unique solution in $M_{w^2}(B^c)$ to the equation (18). This, together with the definition of FPMV-optimal policies, gives the statement.

(c) It is an immediate result of parts (a)–(b) and the measurable selection theorem (e. g., Lemma 8.3.8 in [13]).

(d) Note that from the proof of part (a), T is a contraction operator from $M_{w^2}(B^c)$ to itself. Hence, by Banach's Fixed Point Theorem and part (a), we have $\sigma_*^2 + g^2 = \lim\limits_{n\to\infty} T^n u$ for some $u \in M_{w^2}(B^c)$, which proposes the iteration algorithm of the value function.

(e) It follows from Theorem 7.5.1 in [3]. □

Remark 3.10. It is worth to mention that the way we used here to ensure the existence of the optimality equation is Banach's theory rather than the method of dynamic programming in continuous-time Markov decision processes such as [9, 10, 11].

4. EXAMPLES

In this section, we use two examples to illustrate the application of our main result. One shows that elements in F_g are not unique and an FPMV-optimal policy is derived when the set F_g is finite, another justifies the existence of FPMV-optimal policies when the set F_g is infinite.

Example 4.1. The control model under consideration is given by: $E = \{y_1, y_2, y_3\}$; $B = \{y_3\}$; $A(y_1) = \{a_{11}, a_{12}\}$, $A(y_2) = \{a_{21}, a_{22}\}$, $A(y_3) = \{a_{31}\}$; $Q(t, y|x, a) = (1 - e^{-t})p(y|x, a)$ for every $t \in R_+$, $a \in A(x)$ and $x \in B^c$, where $p(y|x, a)$ are the transition probabilities defined by

$$\begin{aligned}
p(y_1|y_1, a_{11}) = 0,\ p(y_2|y_1, a_{11}) &= \frac{3}{10},\ p(y_3|y_1, a_{11}) = \frac{7}{10};\\
p(y_1|y_1, a_{12}) = 0,\ p(y_2|y_1, a_{12}) &= \frac{1}{2},\ p(y_3|y_1, a_{12}) = \frac{1}{2};\\
p(y_1|y_2, a_{21}) = \frac{3}{4},\ p(y_2|y_2, a_{21}) &= 0,\ p(y_3|y_2, a_{21}) = \frac{1}{4};\\
p(y_1|y_2, a_{22}) = \frac{1}{5},\ p(y_2|y_2, a_{22}) &= \frac{1}{5},\ p(y_3|y_2, a_{22}) = \frac{3}{5}.
\end{aligned}$$

Moreover, let

$$\begin{aligned}
&r(y_1, a_{11}) = 3,\ r(y_1, a_{12}) = \frac{5}{2},\ r(y_2, a_{21}) = 3,\ r(y_2, a_{22}) = 5;\\
&\alpha(y_1, a_{11}) = \alpha(y_2, a_{21}) = \frac{1}{2},\ \alpha(y_1, a_{12}) = \alpha(y_2, a_{22}) = 1;\ g(y_1) = 2,\ g(y_2) = 3.
\end{aligned}$$

The policy set $F = \{f_1, f_2, f_3, f_4\}$, where $f_1(y_1) = a_{11}, f_1(y_2) = a_{21}$; $f_2(y_1) = a_{11}, f_2(y_2) = a_{22}$; $f_3(y_1) = a_{12}, f_3(y_2) = a_{21}$; $f_4(y_1) = a_{12}, f_4(y_2) = a_{22}$.

Now we have the following result.

Proposition 4.2. For the control model in Example 4.1, the set F_g is equal to $\{f_3, f_4\}$, and the policy f_3 is FPMV-optimal.

Proof. By a direct calculation, we get

$$\begin{aligned}
&m_1(y_1|y_1,a_{11}) = 0,\ m_1(y_2|y_1,a_{11}) = \frac{1}{5},\ m_1(y_1|y_1,a_{12}) = 0,\ m_1(y_2|y_1,a_{12}) = \frac{1}{4},\\
&m_1(y_1|y_2,a_{21}) = \frac{1}{2},\ m_1(y_2|y_2,a_{21}) = 0,\ m_1(y_1|y_2,a_{22}) = \frac{1}{10},\ m_1(y_2|y_2,a_{22}) = \frac{1}{10};\\
&m_2(y_1|y_1,a_{11}) = 0,\ m_2(y_2|y_1,a_{11}) = \frac{3}{20},\ m_2(y_1|y_1,a_{12}) = 0,\ m_2(y_2|y_1,a_{12}) = \frac{1}{6},\\
&m_2(y_1|y_2,a_{21}) = \frac{3}{8},\ m_2(y_2|y_2,a_{21}) = 0,\ m_2(y_1|y_2,a_{22}) = \frac{1}{15},\ m_2(y_2|y_2,a_{22}) = \frac{1}{15};
\end{aligned}$$

and

$$\overline{r}(y_1,a_{11}) = 2, \overline{r}(y_1,a_{12}) = \frac{5}{4},\ \overline{r}(y_2,a_{21}) = 2,\ \overline{r}(y_2,a_{22}) = \frac{5}{2}.$$

Obviously, Example 4.1 satisfies Assumptions 2.3, 3.1 and 3.7. Thus, by Theorem 3.9, there exists an FPMV-optimal policy. Furthermore, it follows from Theorem 3.4(a) that

$$\begin{pmatrix} V(y_1,f_1) \\ V(y_2,f_1) \end{pmatrix} = \begin{pmatrix} \frac{8}{3} \\ \frac{10}{3} \end{pmatrix},\ \begin{pmatrix} V(y_1,f_2) \\ V(y_2,f_2) \end{pmatrix} = \begin{pmatrix} \frac{115}{44} \\ \frac{135}{44} \end{pmatrix},$$

and

$$\begin{pmatrix} V(y_1,f_3) \\ V(y_2,f_3) \end{pmatrix} = \begin{pmatrix} V(y_1,f_4) \\ V(y_2,f_4) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} g(y_1) \\ g(y_2) \end{pmatrix}.$$

Therefore, the set F_g is given by $F_g = \{f_3, f_4\}$.

On the other hand, using Theorem 3.5 and Remark 3.6, we have

$$\begin{pmatrix} V^{(2)}(y_1,f_1) \\ V^{(2)}(y_2,f_1) \end{pmatrix} = \begin{pmatrix} \frac{1520}{151} \\ \frac{2080}{151} \end{pmatrix}, \begin{pmatrix} V^{(2)}(y_1,f_2) \\ V^{(2)}(y_2,f_2) \end{pmatrix} = \begin{pmatrix} \frac{58425}{6094} \\ \frac{3225}{277} \end{pmatrix},$$

and

$$\begin{pmatrix} V^{(2)}(y_1,f_3) \\ V^{(2)}(y_2,f_3) \end{pmatrix} = \begin{pmatrix} \frac{232}{45} \\ \frac{164}{15} \end{pmatrix}, \begin{pmatrix} V^{(2)}(y_1,f_4) \\ V^{(2)}(y_2,f_4) \end{pmatrix} = \begin{pmatrix} \frac{430}{83} \\ \frac{920}{83} \end{pmatrix},$$

which implies that $\sigma^2(x,f_3) < \sigma^2(x,f_4) < \sigma^2(x,f_2) < \sigma^2(x,f_1)$ for every $x \in B^c$. Therefore, the policy f_3 is FPMV-optimal although the objective expected reward g here is not the maximal one. □

Finally, we illustrate the application of the policy improvement algorithm.

- For each $x \in \{y_1, y_2\}$, solving (17) gives that $A_g(y_1) = \{a_{12}\}$ and $A_g(y_2) = \{a_{21}, a_{22}\}$. Therefore (by Theorem 3.4(b)), $F_g = \{f_3, f_4\}$.

- Pick a policy $f_4 \in F_g$ and take $h_0 := f_4$.

- Obtain $J(h_0) = \left(\begin{array}{c} \frac{430}{83} \\ \frac{920}{83} \end{array} \right)$.

- Policy improvement: $h_1(y_1) = a_{12}$, $h_1(y_2) = a_{21}$.

- Since $h_0 \neq h_1$, a further iteration yields $h_1 = h_2 = \left(\begin{array}{c} a_{12} \\ a_{21} \end{array} \right)$. Thus, f_3 is FPMV-optimal.

Example 4.3. Consider a model $\{E, B, (A(x) \subset A, x \in E), Q(\cdot, \cdot|x, a), r(x, a), \alpha(x, a), g(x)\}$, where $E = A := (-\infty, +\infty), A(x) := [-|x|, |x|]$ for every $x \in E$, $B := (1, +\infty)$, and $Q(\cdot, \cdot|x, a), r(x, a), \alpha(x, a), g(x)$ are defined by

$$\begin{aligned} Q(t, D|x, a) &:= (1 - e^{-t}) \int_D \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-|x|+|a|)^2}{2}} \, \mathrm{d}y, \\ r(x, a) &:= (x^2 - 1) \frac{10|x| + 9 + a}{|x| + 1} + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}, \\ \alpha(x, a) &:= \frac{a - 1}{|x| + 1} + 9, \quad g(x) := x^2 - 1 \end{aligned}$$

for each $(x, a) \in K$, $D \in \mathcal{B}(E)$ and $t \in R_+$.

Now, we have the following result.

Proposition 4.4. For the control model in Example 4.3, $A_g(x) = \{-|x|, |x|\}$ for all $x \in B^c = (-\infty, 1]$. Moreover, there exists an FPMV-optimal policy.

Proof. Let $w(x) = x^2 + 1$ for all $x \in E$. A direct calculation yields that

$$\begin{aligned} Q(t, E|x, a) &= 1 - e^{-t}, \ \overline{r}(x, a) = x^2 - 1 + \frac{|x| + 1}{10|x| + 9 + a} \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}}, \\ m_1(D|x, a) &= \frac{|x| + 1}{10|x| + 9 + a} \int_D \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-|x|+|a|)^2}{2}} \, \mathrm{d}y, \\ m_2(D|x, a) &= \frac{|x| + 1}{19|x| + 17 + 2a} \int_D \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-|x|+|a|)^2}{2}} \, \mathrm{d}y, \\ \int_{B^c} w(y) m_1(\mathrm{d}y|x, a) &= \frac{|x| + 1}{10|x| + 9 + a} \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} (y^2 + 1) e^{-\frac{(y-|x|+|a|)^2}{2}} \, \mathrm{d}y, \\ \int_{B^c} w^2(y) m_2(\mathrm{d}y|x, a) &= \frac{|x| + 1}{19|x| + 17 + 2a} \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} (y^2 + 1)^2 e^{-\frac{(y-|x|+|a|)^2}{2}} \, \mathrm{d}y \end{aligned}$$

for every $(x,a) \in K$, $D \in \mathcal{B}(E)$ and $t \in R_+$. Using the dominated convergence theorem, from the above expressions we see that Assumptions 2.3 and 3.7 are obviously holds. Next, we will verify Assumption 3.1. Indeed, for each $(x,a) \in K$ $\alpha(x,a) \geq 8$,

$$|r(x,a)| = \big|(x^2-1)\frac{10|x|+9+a}{|x|+1} + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\big| \leq (11 + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}})(x^2+1),$$

and

$$\begin{aligned}\int_{-\infty}^{1}(y^2+1)^2 m_1(\mathrm{d}y|x,a) &\leq \frac{|x|+1}{10|x|+9+a}\int_{-\infty}^{\infty}(y^4+2y^2+1)\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-|x|+|a|)^2}{2}}\,\mathrm{d}y \\ &= \frac{|x|+1}{10|x|+9+a}[6+8(|x|-|a|)^2+(|x|-|a|)^4] \\ &\leq \frac{1}{9}[6+12x^2+6x^4] = \frac{2}{3}(x^2+1)^2,\end{aligned}$$

which shows that Assumption 3.1 holds with $\alpha_0 = 8$, $M = 11 + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}$ and $\rho = \frac{\sqrt{6}}{3}$. As analyzed above, all conditions required for Theorem 3.9 are fulfilled. Thus, by Theorem 3.9, there exists an FPMV-optimal policy.

For each $x \in (-\infty, 1]$, letting the date in the equation of (17) be replaced with ones in Example 4.3, we immediately get

$$\int_{-\infty}^{1}(y^2-1)\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-|x|+|a|)^2}{2}}\,\mathrm{d}y = -\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}.$$

Solving the above equation, we obtain two solutions $a_1 := -|x|$ and $a_2 := |x|$. Thus, $A_g(x) = \{-|x|, |x|\}$ for all $x \in B^c = (-\infty, 1]$, which implies that there are an infinite number of policies in F_g. □

Remark 4.5. Example 4.3 only justifies the existence of an FPMV-optimal policy but not gives an explicit FPMV-optimal policy. To obtain an explicit FPMV-optimal policy, numerical experiments by executing the value iteration or policy improvement algorithms proposed in Theorem 3.9 are needed.

5. REDUCTION TO THE CASES FOR DTMDPS AND CTMDPS

In this section, we will show how a first passage variance model for SMDPs becomes that for DTMDPs and CTMDPs if the semi-Markov kernel Q is of the special forms as below.

Case 1. the first passage variance model for DTMDPs

Suppose that the semi-Markov kernel Q has the expression

$$Q(t,D|x,a) = \begin{cases} p(D|x,a), & t \geq 1, \\ 0, & \text{otherwise} \end{cases}$$

for all $D \in \mathcal{B}(E)$ and $(x,a) \in K$, with $p(\cdot|x,a)$ as a stochastic kernel on E given K. In this case, $T_n = n$ almost everywhere for each $n \geq 0$, and thus $\{T_n, J_n, A_n\}$ is the standard DTMDPs.

By some transformations of structure, for each $x \in E, f \in \mathbb{F}$, $V(x,f)$ can be rewritten as

$$\begin{aligned} V(x,f) &= E_x^f\Big[\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s),a(s))\,\mathrm{d}s} r(x(t),a(t))\,\mathrm{d}t\Big] \\ &= E_x^f\Big[\sum_{n=0}^{\infty}\int_n^{n+1} e^{-\sum_{k=0}^{n-1}\alpha(J_k,A_k)-\alpha(J_n,A_n)(t-n)} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\,\mathrm{d}t\Big] \\ &= E_x^f\Big[\sum_{n=0}^{\infty} e^{-\sum_{k=0}^{n-1}\alpha(J_k,A_k)} I_{\{J_0\in B^c,\ldots,J_n\in B^c\}} r(J_n,A_n)\int_0^1 e^{-\alpha(J_n,A_n)t}\,\mathrm{d}t\Big]. \end{aligned}$$

Let $\widetilde{\alpha}(x,a) := e^{-\alpha(x,a)}$ and $\widetilde{r}(x,a) := r(x,a)\int_0^1 e^{-\alpha(x,a)t}\,\mathrm{d}t$. Then

$$\begin{aligned} \sigma^2(x,f) &= E_x^f\left[\Big(\sum_{n=0}^{\infty}\prod_{k=0}^{n-1}\widetilde{\alpha}(J_k,A_k) I_{\{J_0\in B^c,\ldots,J_n\in B^c\}}\widetilde{r}(J_n,A_n) - V(x,f)\Big)^2\right] \\ &= E_x^f\left[\Big(\sum_{n=0}^{\tau_B-1}\prod_{k=0}^{n-1}\widetilde{\alpha}(J_k,A_k)\widetilde{r}(J_n,A_n) - V(x,f)\Big)^2\right], \end{aligned}$$

which coincides with the first passage discounted variance for DTMDPs. Thus, the optimality equation becomes

$$\sigma_*^2(x) + g^2(x) = \inf_{a\in A_g(x)}\Big\{\widetilde{r}(x,a)[2g(x)-\widetilde{r}(x,a)] + \widetilde{\alpha}^2(x,a)\int_{B^c}[\sigma_*^2(y)+g^2(y)]p(\mathrm{d}y|x,a)\Big\}.$$

This extends the results of [24, 30] to the case of state-action dependent discount factors and an arbitrary function g. In fact, the above equation is the optimality equation in the first passage variance model for DTMDPs with a target set B, the discount factor $\widetilde{\alpha}(x,a)$, the transition probability $p(\cdot|x,a)$ and reward function $\widetilde{r}(x,a)$.

Case 2. the first passage variance model for CTMDPs

Suppose that the semi-Markov kernel Q has the form

$$Q(t,D|x,a) = \begin{cases} (1-e^{-q(x,a)t})\frac{q(D|x,a)}{q(x,a)}, & x\notin D, t\geq 0, \\ 0, & \text{otherwise} \end{cases}$$

for all $D \in \mathcal{B}(E)$ and $(x,a)\in K$, where $q(\cdot|x,a)$ is a conservative and stable transition rates on E given K, $q(x,a) := -q(\{x\}|x,a)$. In this case, we have

$$\begin{aligned} m_1(D|x,a) &= \int_0^{\infty} e^{-\alpha(x,a)t}Q(\mathrm{d}t,D|x,a) = \frac{1}{\alpha(x,a)+q(x,a)}q(D|x,a), \\ m_2(D|x,a) &= \int_0^{\infty} e^{-2\alpha(x,a)t}Q(\mathrm{d}t,D|x,a) = \frac{1}{2\alpha(x,a)+q(x,a)}q(D|x,a),\ \forall\, x\notin D. \end{aligned}$$

Although the policies for CTMDPs are different from those for SMDPs, they are identical for the stationary policies. Then, the model (1) becomes

$$\{E, B, (A(x)\subset A, x\in E), q(\cdot|x,a), r(x,a), \alpha(x,a), g(x)\}.$$

Under the first passage variance criterion, the optimality equation (23) can be written by

$$\inf_{a\in A_g(x)}\Big\{2r(x,a)g(x)+\int_{B^c}[\sigma_*^2(y)+g^2(y)]q(\mathrm{d}y|x,a)-2[\sigma_*^2(x)+g^2(x)]\alpha(x,a)\Big\}=0,$$

which is exactly the optimality equation in the first passage model with the variance criterion for CTMDPs in [11].

6. CONCLUDING REMARKS

In this paper, we have studied a new and interesting issue – the FPMV problem for SMDPs. Our work extends the previous works from DTMDPs and CTMDPs to SMDPs, from fixed finite or infinite time horizons to a random time horizon, and from a constant discount factor to a varying one. To solve the FPMV problem of SMDPs, we characterize the policies with a common expected first passage reward, transform the variance of the first passage reward to a mean, and develop suitable conditions under which an FPMV-optimal policy is ensured. Also, we have derived a policy improvement algorithm as well as a value iteration algorithm to compute an FPMV-optimal policy. To show the application of our main results, we give two illustrative examples. At last, a reduction of the results to the cases of DTMDPs and CTMDPs is exhibited.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China(Grant No.: 11626058) and the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University(Grant No.: 2016008). The authors are also grateful to the Associate Editor and the referee for many valuable suggestions that have improved this paper.

(Received November 24, 2015)

REFERENCES

[1] H. Berument, Z. Kilinc, and U. Ozlale: The effects of different inflation risk premiums on interest rate spreads. Phys. A *333* (2004), 317–324. DOI:10.1016/j.physa.2003.10.039

[2] M. Baykal-Gürsoy and K. Gürsoy: Semi-Markov decision processes: nonstandard criteria. Probab. Engrg. Inform. Sci. *21* (2007), 635–657.

[3] N. Bäuerle and U. Rieder: Markov decision processes with applications to finance. In: Universitext, Springer, Heidelberg 2011. DOI:10.1007/978-3-642-18324-9

[4] E. Collins: Finite-horizon variance penalised Markov decision processes. OR Spektrum *19* (1997), 35–39. DOI:10.1007/s002910050017

[5] O. L. V. Costa, A. C. Maiali, and A. de C. Pinto: Sampled control for mean-variance hedging in a jump diffusion financial market. IEEE Trans. Automat. Control *55* (2010), 1704–1709. DOI:10.1109/tac.2010.2046923

[6] J. A. Filar, L. C. M. Kallenberg, and H. M. Lee: Variance-penalized Markov decision processes. Math. Oper. Res. *14* (1989), 147–161. DOI:10.1287/moor.14.1.147

[7] C. P. Fu, A. Lari-Lavassani, and X. Li: Dynamic mean-variance portfolio selection with borrowing constraint. European J. Oper. Res. *200* (2010), 312–319. DOI:10.1016/j.ejor.2009.01.005

[8] X. P. Guo and O. Hernández-Lerma: Continuous-Time Markov Decision Processes: Theory and Applications. Springer-Verlag, Berlin 2009. DOI:10.1007/978-3-642-02547-1

[9] X. P. Guo and X. Y. Song: Mean-variance criteria for finite continuous-time Markov decision processes. IEEE Trans. Automat. Control *54* (2009), 2151–2157. DOI:10.1109/tac.2009.2023833

[10] X. P. Guo, L. E. Ye, and G. Yin: A mean-variance optimization problem for discounted Markov decision processes. European J. Oper. Res. *220* (2012), 423–429.

[11] X. P. Guo, X. X. Huang and Y. Zhang: On the first passage g-mean variance optimality for discounted continuous-time Markov decision processes. SIAM J. Control Optim. *53* (2015), 1406–1424. DOI:10.1137/140968872

[12] Q. Y. Hu: Continuous time Markov decision processes with discounted moment criterion. J. Math. Anal. Appl. *203* (1996), 1–12. DOI:10.1006/jmaa.1996.9999

[13] O. Hernández-Lerma and J. B. Lasserre: Further Topics on Discrete-Time Markov Control Processes. Springer-Verlag, New York 1999. DOI:10.1007/978-1-4612-0561-6

[14] O. Hernández-Lerma, O. Vega-Amaya and G. Carrasco: Sample-path optimality and variance-minimization of average cost Markov control processes. SIAM J. Control Optim. *38* (1999), 79–93.

[15] S. Haberman and J. H. Sung: Optimal pension funding dynamics over infinite control horizon when stochastic rates of return are stationary. Insurance Math. Econom. *36* (2005), 103–116. DOI:10.1016/j.insmatheco.2004.10.006

[16] Y. H. Huang and X. P. Guo: First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs. Acta Math. Appl. Sin. Engl. Ser. *27* (2011), 177–190. DOI:10.1007/s10255-011-0061-2

[17] Y. H. Huang, X. P. Guo, and X. Y. Song: Performance analysis for controlled semi-Markov systems with application to maintenance. J. Optim. Theory Appl. *150* (2011), 395–415. DOI:10.1007/s10957-011-9813-7

[18] Y. H. Huang and X. P. Guo: Constrained optimality for first passage criteria in semi-Markov decision processes. Optimization, Control, and Applications of Stochastic Systems, pp. 181–202, Systems Control Found. Appl., Birkhäuser/Springer, New York 2012.

[19] Y. H. Huang and X. P. Guo: Mean-variance problems for finite horizon semi-Markov decision processes. Appl. Math. Optim. *72* (2015), 233–259. DOI:10.1007/s00245-014-9278-9

[20] S. C. Jaquette: Markov decision processes with a new optimality criterion: continuous time. Ann. Statist. *3* (1975), 547–553. DOI:10.1214/aos/1176343087

[21] M. Kurano: Markov decision processes with a minimum-variance criterion. J. Math. Anal. Appl. *123* (1987), 572–583. DOI:10.1016/0022-247x(87)90332-5

[22] I. Kharroubi and T. Lim: A. Ngoupeyou, Mean-variance hedging on uncertain time horizon in a market with a jump. Appl. Math. Optim. *68* (2013), 413–444. DOI:10.1007/s00245-013-9213-5

[23] M. J. Lee and W. J. Li: Drift and diffusion function specification for short-term interest rates. Econom. Lett. *86* (2005), 339–346. DOI:10.1016/j.econlet.2004.09.002

[24] P. Mandl: On the variance in controlled Markov chains. Kybernetika *7* (1971), 1–12.

[25] S. Mannor and J. N. Tsitsiklis: Algorithmic aspects of mean-variance optimization in Markov decision processes. European J. Oper. Res. *231* (2013), 645–653. DOI:10.1016/j.ejor.2013.06.019

[26] H. M. Markowitz: Portfolio Selection: Efficient Diversification of Investments. John Wiley and Sons, Inc., New York 1959.

[27] T. Prieto-Rumeau and O. Hernández-Lerma: Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. Math. Methods Oper. Res. *70* (2009), 527–540. DOI:10.1007/s00186-008-0276-z

[28] M. J. Sobel: The variance of discounted Markov decision processes. J. Appl. Probab. *19* (1982), 794–802. DOI:10.1017/s0021900200023123

[29] D. J. White: Computational approaches to variance-penalised Markov decision processes. OR Spektrum *14* (1992), 79–83. DOI:10.1007/bf01720350

[30] X. Wu and X. P. Guo: First passage optimality and variance minimisation of Markov decision processes with varying discount factors. J. Appl. Probab. *52* (2015), 441–456. DOI:10.1017/s0021900200012560

[31] X. Y. Zhou, G. Yin: Markowitz's mean-variance portfolio selection with regime switching: a continuous-time model. SIAM J. Control Optim. *42* (2003), 1466–1482. DOI:10.1137/s0363012902405583

[32] Q. X. Zhu and X. P. Guo: Markov decision processes with variance minimization: a new condition and approach. Stoch. Anal. Appl. *25* (2007), 577–592. DOI:10.1080/07362990701282807

Xiangxiang Huang, School of Computer Science and Network Security, Dongguan University of Technology, Dongguan, 523000. P. R. China.
e-mail: hxiangx3@163.com

Yonghui Huang, Corresponding author. School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, 510275. P. R. China.
e-mail: hyongh5@mail.sysu.edu.cn