ON THE FIRST PASSAGE G-MEAN-VARIANCE OPTIMALITY FOR DISCOUNTED CONTINUOUS-TIME MARKOV DECISION PROCESSES*

XIANPING GUO†, XIANGXIANG HUANG†, AND YI ZHANG‡

Abstract. This paper considers the discounted continuous-time Markov decision processes (MDPs) in Borel spaces and with unbounded transition rates. The discount factors are allowed to depend on states and actions. The main attention is concentrated on the set F_g of stationary policies attaining a given mean performance g up to the first passage of the continuous-time MDP to an arbitrarily fixed target set. Under suitable conditions, we prove the existence of a g-mean-variance optimal policy that minimizes the first passage variance over the set F_g using a transformation technique, and also give the value iteration and policy iteration algorithms for computing the g-variance value function and a g-mean-variance optimal policy respectively. Two examples are analytically solved to demonstrate the application of our results.

Key words. continuous-time Markov decision processes, state-action-dependent discount factors, first passage mean-optimality, first passage g-mean-based variance minimization

AMS subject classifications. 90C40, 93E20

1. Introduction.

The present paper studies a so-called first passage g-mean-based variance minimization problem for a discounted continuous time Markov decision process (MDP) in Borel spaces.

As an important class of stochastic optimal control problems, continuous-time MDPs have been widely studied; see [6, 23, 24, 25] for instance. One of the most commonly used performance measures for continuous-time MDPs, which is further considered in the present paper, is the expected discounted reward criterion; see [6, 21] for finite state and action spaces, [16, 17, 24, 25] for denumerable state spaces and bounded transition rates, [6, 23] for denumerable states but unbounded transition rates, [2, 11] for (general) possibly uncountable state spaces and bounded transition rates, and [3, 5, 7, 22] for possibly uncountable state spaces and unbounded transition rates. In all the mentioned works, the focus is on the existence and computation of an optimal policy.

In many real situations, there exist more than one such optimal policies, and thus it is meaningful to identify the policies which are with the smallest variance in this class of optimal policies. The corresponding variance minimization problem is considered in [14, 15], where the authors focus on discounted continuous-time MDPs in a finite or countable state space and with a bounded reward rate. On the other hand, a risk-averse decision maker might prefer a policy with a reasonable mean performance g (not necessarily the value function) but also a very attractive variance performance. In fact, the Nobel laureate Markowitz suggests in [19] that one should select a policy with a mean-variance performance in the efficient frontier in the set of all attainable mean-variance vectors to incorporate the trade-off between the mean and variance performance; see also [4, 20, 27]. In light of this, in the present paper we consider the so called g-mean-based variance minimization problem, and aim at the so called g-mean-variance optimal policies, i.e., those with the minimal variance out of

*Research supported by NSFC.

†The School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, 510275, P. R. China (mcsgxp@mail.sysu.edu.cn, hxiangx3@163.com).

‡Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 7ZL, UK (zy1985@liv.ac.uk).

the concerned class of policies, whose mean performance is given by the function g. In this connection, the g-mean-based variance minimization problem is a generalization of the variance minimization problem considered in [14, 15].

To the best of our knowledge, the g-mean-based variance minimization problem for discounted continuous-time MDPs was firstly considered in [10], where, however, the controlled process is assumed to be in finite state and action spaces. The variance minimization problems considered in [14, 15] are also restricted to the denumerable state spaces, and bounded transition and reward rates. In the present paper, we consider the g-mean-based variance minimization problem for discounted continuous-time MDPs in Borel state and action spaces with possibly unbounded transition and reward rates. Furthermore, we allow the following more general features as compared to [10, 14, 15] in our model (see also Remark 2.4 for greater details):

(i) The mean and variance of the discounted total reward for each policy are valuated up to the (random) first passage time of the controlled process to a target set, instead of over the infinite time horizon. Such the first passage optimality, as considered also in [1, 18] for discrete-time MDPs and in [8] for continuous-time MDPs in denumerable state spaces with the mean performance measure, has rich applications to, e.g., reliability, where one is interested in the mean performance of a system before it fails, and the target set can be taken as the collection of failure states of the system.

(ii) The discount factors are state-action-dependent in response to, e.g., the fact that the interest rate offered by a bank may differ with the investor's decision of depositing in a fixed long-term saver account or in a flexible short-term basic account, and the interest may also change with the amount of the depositing money of the investor. (Due to this and the above features, the resulting continuous-time MDP model is the extension of those in [1, 5, 6, 7, 9, 10, 16, 22, 23, 24, 25, 26].)

To solve the first passage g-mean-based variance minimization problem, we first need to deal with the first passage mean optimality problem in Borel spaces, which gives preliminary facts for analyzing the first passage g-mean-based variance minimization. Thus, the main contributions of the present paper are as follows.

(1) (*On the first passage mean-optimality.*) We show that the mean-value function is the unique solution to the first passage mean-optimality equation by a value iteration technique, and also establish the existence and an approximation algorithm of a first passage mean-optimal policy; see Theorem 3.4.

(2) (*On the first passage g-mean-based variance minimization.*) Based on the characterization of the class of policies with the given mean performance g, by reducing the first passage g-mean-based variance minimization problem to a first passage mean minimization problem, we show the existence of a first passage g-mean-variance optimal policy, and provide its characterization by a so-called first passage g-mean-based variance optimality equation. A value iteration algorithm is justified for computing the g-variance value function too; see Theorem 4.4 below. Moreover, a policy iteration algorithm for computing a first passage g-mean-variance optimal policy is given in Theorem 4.5.

(3) (*On applications.*) To demonstrate the application of our results, we present two examples, which are solved in closed-forms, and which can be used to show the difference among the three kinds of discount factors; see Proposition 5.2 for the details.

The rest of this paper is organized as follows. In Section 2 we formulate the

mathematical model and state the optimality problems under consideration. In Section 3, some technical preliminaries about the existence of mean-optimal policies and the calculation of the mean-value function are given. The existence and computation (approximation) of g-mean-variance optimal policies together with the g-mean-based variance optimality equation is established in Section 4. In Section 5 we explicitly solve two examples to illustrate our main results. Finally, we finish this paper with some remarks in Section 6.

2. Optimal control problem. *Notation.* If X is a Borel space, we denote by $\mathcal{B}(\mathrm{X})$ the Borel σ-algebra, and by D^c the complement of a set D in $\mathcal{B}(\mathrm{X})$ with respect to X. For any real-valued measurable function $V \geq 1$ on X, a real-valued measurable function u on D^c is called V-*bounded* if $\|u\|_V := \sup_{x\in D^c} \frac{|u(x)|}{V(x)} < \infty$. Denote by $M_V(D^c)$ the Banach space of all V-bounded measurable functions on D^c.

The concerned continuous-time MDP model is specified by the eight-tuple

$$\mathcal{M} := \{S, A, (A(x) \subseteq A, x \in S), q(\cdot|x,a), r(x,a), \alpha(x,a), B, g\} \tag{2.1}$$

with the following components.

- S is the nonempty Borel state space.
- A is the nonempty Borel action space.
- $A(x)$, a Borel subset of A, denotes the set of all admissible actions at the state $x \in S$. The set $K := \{(x,a)|\ x \in S, a \in A(x)\}$ of admissible state-action pairs is assumed to be a Borel subset of $S \times A$, and to contain the graph of a measurable mapping from S to A.
- $q(\cdot|x,a)$ specifies the transition rates, that is, the following conditions are satisfied:

  $\mathbf{T_1}$: for each fixed $(x,a) \in K$, $q(\cdot|x,a)$ is a signed measure on $\mathcal{B}(S)$, while for each fixed $D \in \mathcal{B}(S)$, $q(D|\cdot,\cdot)$ is a measurable function on K;

  $\mathbf{T_2}$: for all $(x,a) \in K$ and $x \notin D \in \mathcal{B}(S)$, $0 \leq q(D|x,a) < \infty$;

  $\mathbf{T_3}$: $q(S|x,a) = 0$, and for each $x \in S$,

  $$q^*(x) := \sup_{a\in A(x)} q(x,a) < \infty, \tag{2.2}$$

  where $q(x,a) := -q(\{x\}|x,a) \geq 0$.
- The real-valued function $r(x,a)$ denotes the reward rate and is assumed to be Borel-measurable on K. (Since $r(x,a)$ is allowed to take positive and negative values, it can be also interpreted as a cost rate rather than a reward rate.)
- The measurable function $\alpha(x,a) > 0$ is the state-action-dependent discount factor.
- The measurable set $B \in \mathcal{B}(S)$ is any given target set.
- The measurable function g on S is any given expected mean performance.

Definition 2.1. *A (stationary) policy f is a measurable mapping from S to A such that $f(x) \in A(x)$ for each $x \in S$. The set of all such policies is denoted by F.*

Suppose the decision maker adopts a policy f. Then the continuous-time MDP evolves like the following. If the current state is $x(t) \in S$ at time $t \geq 0$, the process stays there for a sojourn time, whose tail function is given by

$$e^{-q(x(t),f(x(t)))t},$$

and if $q(x(t), f(x(t))) > 0$, then the new state obeys the distribution given by

$$\frac{q(dy \setminus \{x(t)\}|x(t), f(x(t)))}{q(x(t), f(x(t)))}.$$

For any initial distribution γ on S and $f \in F$, as in [5, 7, 9, 22] one can construct a probability space $(\Omega, \mathcal{F}, P_\gamma^f)$ and a Markov jump process $\{x(t), t \geq 0\}$ with values in S, which evolves as described in the above. In particular, for each $t > 0, x \in S, x \notin D \in \mathcal{B}(S)$,

$$P_\gamma^f((x(0) \in D) = \gamma(D),$$

$$P_\gamma^f(\tau_1 \leq t | x(0) = x) = 1 - e^{-tq(x,f(x))}, \tag{2.3}$$

$$P_\gamma^f(\tau_1 \in dt, x(\tau_1) \in D | x(0) = x) = e^{-tq(x,f(x))} q(D|x, f(x))dt, \tag{2.4}$$

where

$$\tau_1 := \inf\{t > 0 : x(t) \neq x(0)\} \tag{2.5}$$

denotes the first jumping time of $\{x(t), t \geq 0\}$. The expectation operator associated with P_γ^f is denoted by E_γ^f. We will write P_γ^f and E_γ^f as P_x^f and E_x^f, respectively, when γ is a Dirac measure concentrated at $x \in S$.

Assumption A. There exist a measurable function $w \geq 1$ on S, constants $c > 0, b \geq 0$, and $M > 0$ such that

(1) $\int_S w(y)q(dy|x, a) \leq cw(x) + b$ for all $a \in A(x)$ and $x \in S$; and

(2) $q^*(x) \leq Mw(x)$ for all $x \in S$, with $q^*(x)$ as in (2.2).

Assumption A ensures that the process $\{x(t), t \geq 0\}$ is nonexplosive under each policy f [9, 22] (i.e., $P_x^f(x(t) \in S) \equiv 1$), and it is also required for the finiteness of the expected first passage reward $V_B(x, f)$ defined in (2.7) below.

For the given target set $B \in \mathcal{B}(S)$, we denote by

$$\tau_B := \begin{cases} \inf\{t \geq 0 : x(t) \in B\} & \text{if } \{t \geq 0 : x(t) \in B\} \neq \emptyset, \\ +\infty & \text{otherwise} \end{cases} \tag{2.6}$$

the first passage time to B of the process $\{x(t), t \geq 0\}$. In particular, $\tau_B = \infty$ if $B = \emptyset$.

Definition 2.2. *(The first passage discounted mean and variance criteria.) For each $x \in S$ and $f \in F$, the mean of the first passage discounted total reward for f is defined as*

$$V_B(x, f) := E_x^f\Big[\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt\Big], \tag{2.7}$$

and the variance of the first passage discounted total reward for f is given by

$$\sigma_B^2(x, f) := E_x^f\Big[\Big(\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt - V_B(x, f)\Big)^2\Big]. \tag{2.8}$$

To state the optimality problem we are concerned with, let us introduce some notation as below. Let

$$V_B^*(x) := \sup_{f \in F} V_B(x, f) \quad \forall\, x \in S \tag{2.9}$$

denote the *mean-value function (of the first passage mean criterion)*.

By (2.6) we see that $\tau_B = 0$ when the initial state $x(0)$ is in B, and thus it follows from (2.7) and (2.8) that $V_B(x, f) = \sigma_B^2(x, f) = 0$ for all $x \in B$ and $f \in F$. Hence, in the coming arguments we will restrict our attention to the initial states in B^c.

For the given g, let F_g be the set of all policies with the performance g on B^c, i.e.,

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

The condition of $F_g \neq \emptyset$ will be given in Lemma 4.1 below.

Just as Markowitz's mean-variance portfolio problem [4, 20, 27], we assume that the set F_g is nonempty throughout this paper, and then consider the following so-called *first passage g-mean-based variance minimization problem*:

$$P_g : \text{minimize } \sigma_B^2(x,f) \text{ over all } f \in F_g \text{ for all } x \in B^c. \tag{2.10}$$

In particular, when g is taken as the mean-value function V_B^* for the special case that $B = \emptyset$ and $\alpha(x,a)$ is a constant, the corresponding V_B^*-mean-based variance minimization problem is the variance minimization problem in the previous literature [10, 14, 15] for the case of infinite horizon, denumerable states, and a constant discount factor, whereas the variance minimization problem for the first passage continuous-time MDPs with varying discount factors has not been studied yet.

DEFINITION 2.3. *(a) A policy $f^* \in F$ is said to be (first passage) mean-optimal if*

$$V_B(x,f^*) = V_B^*(x) \quad \forall\ x \in B^c. \tag{2.11}$$

(b) A policy $f^* \in F_g$ is called (first passage) g-mean-variance optimal if

$$\sigma_g^2(x) := \inf_{f \in F_g} \sigma_B^2(x,f) = \sigma_B^2(x,f^*) \quad \forall\ x \in B^c,$$

where the function σ_g^2 on B^c is called the g-variance value function.

REMARK 2.4.

(a) When $\alpha(x,a)$ is a positive constant (denoted by α) and $B = \emptyset$, it follows from (2.6) and (2.7) that $V_B(x,f) = E_x^f\big[\int_0^\infty e^{-\alpha t} r(x(t), f(x(t)))dt\big] =: V_\alpha(x,f)$, which is the infinite horizon discounted reward criterion and widely studied; see [2, 5, 6, 7, 9, 10, 16, 22, 23] for instance. Moreover, When $\alpha(x,a)$ depends on states only (denoted by $\alpha(x)$) and $B = \emptyset$, then the $V_B(x,f) = E_x^f\big[\int_0^\infty e^{-\int_0^t \alpha(x(s))ds} r(x(t), f(x(t)))dt\big] =: V_{\alpha(x)}(x,f)$, which is the same as in [26].

(b) In Corollary 5.6 below we have $V_B(x,f) \neq V_\alpha(x,f) \neq V_{\alpha(x)}(x,f)$ for some f, which shows the difference between the first passage discounted criterion and the infinite horizon discounted criterion [5, 6, 7, 9, 10, 16, 22, 23].

(c) The function g and the set B are arbitrarily given but fixed. Hence, when g is taken as the mean-value function and B is the empty set, our g-mean-based variance minimization problem is degenerated to the variance minimization problem in [14, 15].

3. Preliminaries. In this section, we will establish a so-called first passage discounted mean-optimality equation with state-action-dependent discounting, show the existence of a mean-optimal policy, and also provide a value iteration algorithm for computing the mean-value function.

For a policy $f \in F$ and $x \in B^c$, since the reward $r(x,a)$ can be unbounded, to ensure the finiteness of $V_B(x,f)$, we give the following condition and a fact.

Assumption B. There exist constants $\alpha_0, c_1, M_1 > 0$ and $b_1 \geq 0$, such that

(1) $|r(x,a)| \le M_1 w(x)$ for all $x \in B^c$ and $a \in A(x)$;
(2) $c < \alpha_0 \le \alpha(x,a)$ for all $x \in B^c$ and $a \in A(x)$, with c as in Assumption A(1), and $\alpha(x,a)$ as in (2.1);
(3) $\int_S w^2(y)q(dy|x,a) \le c_1 w^2(x) + b_1$ for all $x \in B^c$ and $a \in A(x)$.

REMARK 3.1. In fact, the role of Assumption B(3) is for the finiteness of $E_x^f[w(x(t))q(x(t), f(x(t)))]$, which is required for the usage of Theorem 3.2 in [26] in proving Theorem 3.2 below.

THEOREM 3.2. *Under Assumptions A and B, for each fixed $f \in F$, the following statements hold.*

(a) $V_B(x,f)$ is the unique solution within $M_w(B^c)$ to the following equation:

$$\alpha(x,f(x))u(x) = r(x,f(x)) + \int_{B^c} u(y)q(dy|x,f(x)) \quad \forall\ x \in B^c. \tag{3.1}$$

(b) If there is a function $u \in M_w(B^c)$ such that

$$\alpha(x,f(x))u(x) \ge r(x,f(x)) + \int_{B^c} u(y)q(dy|x,f(x)) \quad \forall\ x \in B^c. \tag{3.2}$$

Then, $u(x) \ge V_B(x,f)$ for all $x \in B^c$.

Proof. From the Definition 2.2, we see that the process $\{x(t), t \ge 0\}$ can be ignored when it leaves the set B^c. Thus, we view the $\{x(t), t \ge 0\}$ to be absorbed in some cemetery state (say, $\Delta \notin S$), and consider a new model

$$\mathcal{M}_\Delta := \{S_\Delta, A_\Delta, (A_\Delta(x) \subseteq A_\Delta, x \in S_\Delta), q_\Delta(\cdot|x,a), r_\Delta(x,a), \alpha_\Delta(x,a)\}$$

of standard continuous-time MDPs, where $S_\Delta := S \cup \{\Delta\}$, $A_\Delta := A \cup \{a_\Delta\}$, $A_\Delta(\Delta) := \{a_\Delta\}$,$A_\Delta(x) := A(x)$ for all $x \in S$, $q_\Delta(\Delta|\Delta, a_\Delta) := 0, r(\Delta, a_\Delta) := 0$, and

$$q_\Delta(dy|x,a) = q(dy|x,a)I_{B^c}(x) \quad \text{for } dy \in \mathcal{B}(\mathrm{S});$$
$$r_\Delta(x,a) = r(x,a)I_{B^c}(x), \quad \alpha_\Delta(x,a) = \alpha(x,a)I_{B^c}(x) + \alpha_0 I_{B\cup\Delta}(x)$$

for all $x \in S_\Delta$ and $a \in A_\Delta(x)$. Moreover, for any $f \in F$, we define the corresponding policy f_Δ for the model $\mathcal{M}_\Delta$ by $f_\Delta(x) := f(x)$ for all $x \in S$ and $f_\Delta(\Delta) := a_\Delta$. For any given $f \in F$, since every state in $B \cup \{\Delta\}$ is absorbing and has null reward for the model $\mathcal{M}_\Delta$ under f_Δ, the first passage discounted mean criterion $V_B(x,f)(x \in S)$ in (2.7) is obviously equivalent to the classical infinite discounted expected criterion $U(x, f_\Delta)$ in [26, (2.4)] for the model $\mathcal{M}_\Delta$, and thus the statements (a) and (b) follow from Theorem 3.2 and Lemma 6.3 in [26], respectively. □

Inspired by (3.1), we call the following equation (3.3) *the first passage discounted mean-optimality equation (with state-action-dependent discounting)*:

$$\sup_{a\in A(x)} \left\{ r(x,a) + \int_{B^c} u(y)q(dy|x,a) - u(x)\alpha(x,a) \right\} = 0 \quad \forall\ x \in B^c. \tag{3.3}$$

A function u in $M_w(B^c)$ satisfying (3.3) is called a solution to the optimality equation.

To show the existence of a solution to the optimality equation (3.3), we introduce the operator T_B as follows: for $u \in M_w(B^c)$, let

$$T_B u(x) := \sup_{a\in A(x)} \left\{ \frac{r(x,a)}{\alpha(x,a) + q(x,a)} + \frac{\int_{B^c - \{x\}} u(y)q(dy|x,a)}{\alpha(x,a) + q(x,a)} \right\} \quad \forall x \in B^c. \tag{3.4}$$

Also, we need an additional condition below.

Assumption C. Let w be as in Assumption A, and $x \in B^c$.

(1) $A(x)$ is a compact set.
(2) For each $x \in B^c$ and Borel set $D \subseteq B^c$, the functions $r(x,a)$, $\alpha(x,a)$, $q(x,a)$, and $q(D \setminus \{x\}|x,a)$ are continuous in $a \in A(x)$.
(3) The function $\int_{B^c} w(y)q(dy|x,a)$ is continuous in $a \in A(x)$.

REMARK 3.3.

(a) Assumption C is similar to the standard continuity-compactness hypotheses for discrete-time and continuous-time MDPs; see, for instance, [5, 6, 12, 24] and their references.
(b) Under Assumption C(2,3), as in the proof of Lemma 8.3.7(a) in [12], we can show that $\int_{B^c-\{x\}} u(y)q(dy|x,a)$ is continuous in $a \in A(x)$ for each $u \in M_w(B^c)$, and so is $r(x,a) + \int_{B^c-\{x\}} u(y)q(dy|x,a)$ in $a \in A(x)$.

Under Assumptions A, B and C, it follows from Remark 3.3(b) that for each $n \geq 0$ and $x \in B^c$ we can legally define

$$u^*_{n+1} := T_B u^*_n, \text{ with } \mathrm{T_B} \text{ as in (3.4)}, \ u^*_0(x) := \frac{M_1 b}{\alpha_0(\alpha_0 - c)} + \frac{M_1}{\alpha_0 - c} w(x). \tag{3.5}$$

THEOREM 3.4. *Under Assumptions A, B and C, the following assertions hold.*
(a) $u^*_n \geq u^*_{n+1}$ for all $n \geq 0$, and $\sup_{n\geq 0} \|u^*_n\|_w \leq \frac{M_1(b+\alpha_0)}{\alpha_0(\alpha_0-c)}$.
(b) $V^*_B(x) = \lim_{n\to\infty} u^*_n(x)$, and V^*_B is the unique solution within $M_w(B^c)$ of the first passage discounted mean-optimality equation (3.3).
(c) For each $n \geq 1$, there exists $f_n \in F$ such that

$$u^*_{n+1}(x) = \frac{r(x, f_n(x)) + \int_{B^c-\{x\}} u^*_n(y)q(dy|x, f_n(x))}{\alpha(x, f_n(x)) + q(x, f_n(x))}, \qquad x \in B^c. \tag{3.6}$$

Moreover, there exists an $f^* \in F$ such that
(c_1) $f^*(x)$ is an accumulation point of $\{f_n(x)\}$ for each $x \in B^c$; and
(c_2) f^* is conserving, i.e.,

$$\begin{aligned} 0 &= \sup_{a\in A(x)} \left\{ r(x,a) + \int_{B^c} V^*_B(y)q(dy|x,a) - V^*_B(x)\alpha(x,a) \right\} \\ &= r(x, f^*(x)) + \int_{B^c} V^*_B(y)q(dy|x, f^*(x)) - V^*_B(x)\alpha(x, f^*(x)), \quad x \in B^c. \end{aligned}$$

(d) A policy in F is mean-optimal if and only if it attains the maximum in (3.3) with $u(x)$ being replaced by $V^*_B(x)$ for all $x \in B^c$, and so f^* in (c) is mean-optimal.

Proof. (a) By Theorem 3.3 in [26] we see that (a) is true.

(b) and (c) will be proved together. Let $u^*(x) := \lim_{n\to\infty} u^*_n(x)$ for each $x \in B^c$. For each $n \geq 0$ and $x \in B^c$, since $u^*_{n+1}(x) = T_B u^*_n(x)$, by (3.4) and the dominated convergence theorem we have

$$u^*(x) \geq \frac{r(x,a)}{\alpha(x,a) + q(x,a)} + \frac{\int_{B^c-\{x\}} u^*(y)q(dy|x,a)}{\alpha(x,a) + q(x,a)} \qquad \forall\, a \in A(x), \tag{3.7}$$

which, together with $\mathbf{T_3}$, implies

$$\sup_{a\in A(x)} \left\{ r(x,a) + \int_{B^c} u^*(y)q(dy|x,a) - u^*(x)\alpha(x,a) \right\} \leq 0. \tag{3.8}$$

On the other hand, for each $n \geq 1$, under Assumption C, Remark 3.3(b) ensures the existence of $f_n \in F$ satisfying (3.6)

Since the multifunction $x \longmapsto A(x)$ $(x \in B^c)$ is compact-valued, and the set $\{(x,a)\ x \in B^c, a \in A(x)\}$ is also measurable, Propositions D.4 and D.7 in [13] ensure the existence of $f^* \in F$ such that $f^*(x)$ is an accumulation point of $\{f_n(x)\}$ for each $x \in B^c$.

Thus, for any fixed $x \in B^c$, there exists a subsequence $\{f_{n_m}(x)\}$ of $\{f_n(x)\}$ such that the limit $\lim_{m\to\infty} f_{n_m}(x) = f^*(x)$ exists and belongs to $A(x)$. Hence, by (3.6) and the extension of Fatou's lemma (i.e., Lemma 8.3.7 in [12]) we have

$$u^*(x) \leq \frac{r(x,f^*(x)) + \int_{B^c-\{x\}} u^*(y)q(dy|x,f^*(x))}{\alpha(x,f^*(x)) + q(x,f^*(x))}$$

which, together with $\mathbf{T_3}$, implies

$$r(x,f^*(x)) + \int_{B^c} u^*(y)q(dy|x,f^*(x)) - u^*(x)\alpha(x,f^*(x)) \geq 0. \tag{3.9}$$

Thus, by (3.8) and (3.9) we have

$$\begin{aligned} 0 &\geq \sup_{a\in A(x)} \left\{ r(x,a) + \int_{B^c} u^*(y)q(dy|x,a) - u^*(x)\alpha(x,a) \right\} \\ &\geq r(x,f^*(x)) + \int_{B^c} u^*(y)q(dy|x,f^*(x)) - u^*(x)\alpha(x,f^*(x)) \geq 0. \end{aligned} \tag{3.10}$$

This means

$$\sup_{a\in A(x)} \left\{ r(x,a) + \int_{B^c} u^*(y)q(dy|x,a) - u^*(x)\alpha(x,a) \right\} = 0. \tag{3.11}$$

Moreover, from (3.10) we have

$$\begin{aligned} 0 &= r(x,f^*(x)) + \int_{B^c} u^*(y)q(dy|x,f^*(x)) - \alpha(x,f^*(x))u^*(x) \\ &\geq r(x,f(x)) + \int_{B^c} u^*(y)q(dy|x,f(x)) - \alpha(x,f(x))u^*(x) \quad \forall\ x \in B^c \text{ and } f \in F. \end{aligned}$$

This fact, along with Theorem 3.2, gives $V_B(x,f^*) = u^*(x) \geq V_B(x,f)$ for all $x \in B^c$ and $f \in F$. Therefore, $u^*(x) = V_B(x,f^*) = V_B^*(x)$. From (3.11) we see that $V_B^*(x)$ solves (3.3). To show the uniqueness of the solution to (3.3), let $v \in M_w(B^c)$ be an arbitrary solution to the equation (3.3). The measurable selection theorem together with Remark 3.3(b) ensures the existence of $f' \in F$ satisfying

$$\begin{aligned} &r(x,f'(x)) + \int_{B^c} v(y)q(dy|x,f'(x)) = v(x)\alpha(x,f'(x)), \\ &r(x,f(x)) + \int_{B^c} v(y)q(dy|x,f(x)) \leq v(x)\alpha(x,f(x)) \quad \forall\ f \in F. \end{aligned}$$

Hence, Theorem 3.2 yields that $v(x) = V_B^*(x)$, which shows the uniqueness of the solution to (3.3). This completes both (b) and (c).

(d) Obviously, it follows from Theorem 3.2(a) and part (b) of this theorem. □

REMARK 3.5.

(a) Theorem 3.4(a,b) gives an iteration approach for the calculation of $V_B^*(x)$. In particular, since the sequence of iterations $\{u_n^*\}$ in (3.5) is constructed from the *primitive* data in the model (2.1), the corresponding approximation method to calculate $V_B^*(x)$ can be implemented.
(b) Theorem 3.4(c,d) show that a mean-optimal policy can be approximated from the policy sequence $\{f_n\}$ obtained in Theorem 3.4(c).
(c) As the arguments in [6, 23, 24, 25], we can give a policy iteration for computing a mean-optimal policy, but the details are omitted here.

4. On g-mean-variance optimal policies. The main goal of this section is to show the existence and computation of a g-mean-variance optimal policy, based on the results established in the previous section.

Since the objective of the g-mean-based variance minimization problem is to minimize $\sigma_B^2(f)$ over f in the set F_g, it is helpful to characterize the policies in F_g. To this end, we need to introduce the following notation

$$A_g(x) := \begin{cases} \{a \in A(x) | r(x,a) + \int_{B^c} g(y) q(dy|x,a) - g(x)\alpha(x,a) = 0\} & x \in B^c \\ A(x) & \text{otherwise} \end{cases}. \tag{4.1}$$

The following fact gives a characterization of F_g in terms of $A_g(x)$.

LEMMA 4.1. *Under Assumptions A and B, a policy $f \in F_g$ if and only if $f(x) \in A_g(x)$ for each $x \in B^c$. (Hence, $F_g \neq \emptyset$ if and only if $A_g(x) \neq \emptyset$ for each $x \in B^c$).*

Proof. Since F_g is assumed to be nonempty, the function g is in $M_w(B^c)$. It follows from the uniqueness of the solution to (3.1) and the definition of F_g. □

Lemma 4.1 implies that $A_g(x) \neq \emptyset$ for all $x \in B^c$ is the condition of $F_g \neq \emptyset$.

Next, we will show that the variance $\sigma_B^2(x,f)$ can be transformed into a mean of a first passage discounted total utility and another discount factor.

To make arguments more convenient, for each $f \in F$, we denote by

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

the second moment of the first passage total reward

$$\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t), f(x(t)))dt.$$

Obviously, it follows from the definitions of $\sigma_B^2(f)$ and F_g that

$$V^{(2)}(x,f) = \sigma_B^2(x,f) + g^2(x) \quad \forall\, x \in B^c, f \in F_g.$$

Thus, the g-mean-based variance minimization problem P_g in (2.10) is equivalent to the following problem minimizing the second moment $V^{(2)}$ over F_g:

$$Q_g : \text{minimize } V^{(2)}(x,f) \text{ over all } f \in F_g \text{ for all } x \in B^c. \tag{4.2}$$

For the finiteness of both $V^{(2)}(x,f)$ and $E_x^f[w^2(x(t))q(x(t), f(x(t)))]$ $(f \in F, x \in B^c)$, as the introduction of Assumption B for the mean-optimality above, we need the following condition.

Assumption D. Suppose that the following conditions hold.

(1) $0 < c_1 < 2\alpha_0$, with α_0 and c_1 as in Assumption B.

(2) There exist constants $c_2 > 0$ and $b_2 \geq 0$ such that

$$\int_S w^3(y)q(dy|x,a) \leq c_2 w^3(x) + b_2 \qquad \forall\ x \in B^c, a \in A(x),$$

with w as in Assumption A.

THEOREM 4.2. *Under Assumptions A, B and D, the following assertions hold.*
(a) $V^{(2)}(\cdot, f)$ *is in* $M_{w^2}(B^c)$ *for each* $f \in F$.
(b) $V^{(2)}(\cdot, f)$ *is the unique solution in* $M_{w^2}(B^c)$ *to the equation*

$$2\alpha(x, f(x))u(x) = 2r(x, f(x))V_B(x, f) + \int_{B^c} u(y)q(dy|x, f(x)), \qquad x \in B^c.$$

Consequently,

$$\sigma_B^2(x, f) = 2E_x^f \left[\int_0^{\tau_B} e^{-\int_0^t 2\alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))V_B(x(t), f)dt\right] - V_B^2(x, f)$$

for each $x \in B^c$ *and* $f \in F$.

Proof. (a) Since $0 < c_1 < 2\alpha_0$ (by Assumption D), there exists $0 < \varepsilon_0 < 1$ such that $0 < c_1 < 2\alpha_0(1 - \varepsilon_0)$. Therefore, for each $x \in B^c$ and $f \in F$, by Assumption B we have

$$\begin{aligned}
V^{(2)}(x, f) &= E_x^f \left[\left(\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt\right)^2\right] \\
&\leq E_x^f \left[\left(\int_0^{\tau_B} e^{-t\varepsilon_0\alpha_0} e^{-t(1-\varepsilon_0)\alpha_0} |r(x(t), f(x(t)))|dt\right)^2\right] \\
&\leq E_x^f \left[\Big(\int_0^{\tau_B} e^{-2\varepsilon_0\alpha_0 t}dt\Big)\Big(\int_0^{\tau_B} e^{-2\alpha_0(1-\varepsilon_0)t} r^2(x(t), f(x(t)))dt\Big)\right] \\
&= E_x^f \left[\frac{1 - e^{-2\varepsilon_0\alpha_0\tau_B}}{2\varepsilon_0\alpha_0}\Big(\int_0^{\tau_B} e^{-2\alpha_0(1-\varepsilon_0)t} r^2(x(t), f(x(t)))dt\Big)\right] \\
&\leq \frac{1}{2\varepsilon_0\alpha_0} E_x^f \left[\int_0^{\tau_B} e^{-2\alpha_0(1-\varepsilon_0)t} r^2(x(t), f(x(t)))dt\right] \\
&\leq \frac{M_1^2}{2\varepsilon_0\alpha_0} E_x^f \left[\int_0^{\infty} e^{-2\alpha_0(1-\varepsilon_0)t} w^2(x(t))dt\right] \\
&\leq \frac{M_1^2}{2\varepsilon_0\alpha_0}\Big[\frac{b_1}{2\alpha_0(1-\varepsilon_0)[2\alpha_0(1-\varepsilon_0) - c_1]} + \frac{1}{2\alpha_0(1-\varepsilon_0) - c_1} w^2(x)\Big],
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality, and the last inequality follows from Theorem 3.3(a) in [5] with w replaced by w^2 here.

(b) For any fixed $f \in F$, and $x(0) := x \notin B$, by a straightforward calculation we have

$$\begin{aligned}
V^{(2)}(x, f) &= E_x^f \left[\left(\int_0^{\tau_B} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt\right)^2\right] \\
&= E_x^f \left[\left(\int_0^{\tau_1} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt \right.\right. \\
&\qquad \left.\left. + \int_{\tau_1}^{\tau_B} I_{\{\tau_1 < \tau_B\}} e^{-\int_0^t \alpha(x(s), f(x(s)))ds} r(x(t), f(x(t)))dt\right)^2\right] \\
&=: I_1 + I_2 + I_3,
\end{aligned}$$

where

$$I_1 := E_x^f\left[\left(\int_0^{\tau_1} e^{-\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)^2\right],$$
$$I_2 := E_x^f\left[\left(\int_{\tau_1}^{\tau_B} I_{\{\tau_1<\tau_B\}} e^{-\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)^2\right],$$
$$I_3 := 2E_x^f\left[\left(\int_0^{\tau_1} e^{-\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)\right.$$
$$\left.\times\left(\int_{\tau_1}^{\tau_B} I_{\{\tau_1<\tau_B\}} e^{-\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)\right].$$

Note that $x(s)=x$ for all $s\le\tau_1$, by (2.3) and a direct calculation, we have

$$\begin{aligned}
I_1 &= \frac{r^2(x,f(x))}{\alpha^2(x,f(x))}E_x^f\left[1-2e^{-\alpha(x,f(x))\tau_1}+e^{-2\alpha(x,f(x))\tau_1}\right]\\
&= \frac{r^2(x,f(x))}{\alpha^2(x,f(x))}\left[1+\int_0^\infty\left(e^{-(2\alpha(x,f(x))+q(x,f(x)))t}-2e^{-(\alpha(x,f(x))+q(x,f(x)))t}\right)q(x,f(x))dt\right]\\
&= \frac{1}{2\alpha(x,f(x))+q(x,f(x))}\times\frac{2r^2(x,f(x))}{\alpha(x,f(x))+q(x,f(x))}.
\end{aligned}$$

Moreover, a straightforward calculation, along with the property of conditional expectation and the strong Markov property, yields

$$\begin{aligned}
I_2 &= E_x^f\left[\left(e^{-\alpha(x,f(x))\tau_1}I_{\{x(\tau_1)\notin B\}}\int_{\tau_1}^{\tau_B} e^{-\int_{\tau_1}^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)^2\right]\\
&= E_x^f\left[e^{-2\alpha(x,f(x))\tau_1}I_{\{x(\tau_1)\notin B\}}E_x^f\left[\left(\int_{\tau_1}^{\tau_B} e^{-\int_{\tau_1}^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt\right)^2|\tau_1,x(\tau_1)\right]\right]\\
&= E_x^f\left[e^{-2\alpha(x,f(x))\tau_1}I_{\{x(\tau_1)\notin B\}}V^{(2)}(x(\tau_1),f)\right]\\
&= \int_0^\infty e^{-(2\alpha(x,f(x))+q(x,f(x)))t}\int_{B^c-\{x\}}V^{(2)}(y,f)q(dy|x,f(x))dt\\
&= \frac{1}{2\alpha(x,f(x))+q(x,f(x))}\int_{B^c-\{x\}}V^{(2)}(y,f)q(dy|x,f(x)).
\end{aligned}$$

Similarly, by (2.3)–(2.4) and (3.1) we have

$$\begin{aligned}
I_3 &= \frac{2r(x,f(x))}{\alpha(x,f(x))}E_x^f\left[e^{-\alpha(x,f(x))\tau_1}(1-e^{-\alpha(x,f(x))\tau_1})I_{\{x(\tau_1)\notin B\}}\right.\\
&\qquad\left.\times E_x^f\left[\int_{\tau_1}^{\tau_B} e^{-\int_{\tau_1}^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))dt|\tau_1,x(\tau_1)\right]\right]\\
&= \frac{2r(x,f(x))}{\alpha(x,f(x))}E_x^f\left[\left(e^{-\alpha(x,f(x))\tau_1}-e^{-2\alpha(x,f(x))\tau_1}\right)I_{\{x(\tau_1)\notin B\}}V_B(x(\tau_1),f)\right]\\
&= \frac{2r(x,f(x))}{2\alpha(x,f(x))+q(x,f(x))}\left[V_B(x,f)-\frac{r(x,f(x))}{\alpha(x,f(x))+q(x,f(x))}\right].
\end{aligned}$$

Thus, taking all the above results of I_1, I_2, I_3 into consideration, we obtain

$$V^{(2)}(x,f)=\frac{2r(x,f(x))V_B(x,f)+\int_{B^c-\{x\}}V^{(2)}(y,f)q(dy|x,f(x))}{2\alpha(x,f(x))+q(x,f(x))},$$

which is equivalent to

$$2\alpha(x,f(x))V^{(2)}(x,f) = 2r(x,f(x))V_B(x,f) + \int_{B^c} V^{(2)}(y,f)q(dy|x,f(x)). \tag{4.3}$$

On the other hand, Assumptions A and B imply that $r(x,f(x))V_B(x,f) \in M_{w^2}(B^c)$. Thus, by Theorem 3.3(a) in [5] and Assumptions B and D we have

$$\begin{aligned} &E_x^f\Big[\int_0^{\tau_B} e^{-2\int_0^t \alpha(x(s),f(x(s)))ds}|r(x(t),f(x(t)))V_B(x(t),f)|dt\Big] \\ &\le M_1||V_B||_w E_x^f\Big[\int_0^\infty e^{-2\alpha_0 t} I_{\{x(t)\notin B\}} w^2(x(t))dt\Big] \\ &\le M_1||V_B||_w\Big[\frac{b_1}{2\alpha_0(2\alpha_0-c_1)} + \frac{1}{2\alpha_0-c_1}w^2(x)\Big] \quad \forall\ x\in B^c, \end{aligned}$$

which, together with Theorem 3.2 and Assumption D(2), implies that

$$V^{(2)}(x,f) = 2E_x^f\Big[\int_0^{\tau_B} e^{-2\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))V_B(x(t),f)dt\Big]$$

is the unique solution within $M_{w^2}(B^c)$ to the equation (4.3). Moreover, since $\sigma_B^2(x,f) = V^{(2)}(x,f) - V_B^2(x,f)$, the proof is completed. □

COROLLARY 4.3. If S is finite, then

$$\sigma^2(f) = [2diag(\alpha(f)) - Q_B(f)]^{-1}c_g(f) - g^2, \quad \text{for all } f\in F_g,$$

where

$$\begin{aligned} diag(\alpha(f)) &:= diag(\alpha(x,f(x)), x\in B^c), \\ Q_B(f) &:= (q(y|x,f(x)), x,y\in B^c), \\ c_g(f) &:= (2r(x,f(x))g(x), x\in B^c)^T. \end{aligned}$$

Let

$$J_B(x,f) := 2E_x^f\Big[\int_0^{\tau_B} e^{-2\int_0^t \alpha(x(s),f(x(s)))ds} r(x(t),f(x(t)))g(x(t))dt\Big].$$

Since $V_B(x,f) = g(x)$ for all $f\in F_g$ and $x\in B^c$, Theorem 4.2 implies

$$\sigma_B^2(x,f) = J_B(x,f) - g^2(x), \quad \text{for each } f\in F_g \text{ and } x\in B^c.$$

Therefore, we can conclude that, under suitable Assumptions A, B and D, the problem Q_g in (4.2) (and so the original problem P_g) is equivalent to the following one:

$$Q_g^*:\ \text{minimize}\ \ J_B(x,f) \text{ over } F_g \ \text{ for all } x\in B^c, \tag{4.4}$$

which is a first passage mean-optimality problem, and can be solved by combining Lemma 4.1 and the results developed in Section 3 above.

Note that $J_B(\cdot,f)$ is w^2-bounded. In order to solve the problem Q_g in (4.2), in spirit of Assumption C above we introduce the hypothesis below.

Assumption C*. Let w be as in Assumption A, and $x\in B^c$.

(1) Assumption C(1,2) are satisfied.

(2) The function $\int_{B^c} w^2(y)q(dy|x,a)$ is continuous in $a \in A(x)$.

It follows from the proof of Remark 3.3(b) that, under Assumption C*, the function $\int_{B^c} w(y)q(dy|x,a)$ is continuous in $a \in A(x)$ for any fixed $x \in B^c$. Hence, Assumption C* implies Assumption C, that is, all results in Section 3 still hold when Assumption C is replaced with Assumption C*.

We next state our main results about the g-mean-based variance minimization problem.

THEOREM 4.4. *Under Assumptions A, B, C* and D, the following statements hold.*

(a)$\sigma_g^2 + g^2$ is a unique solution within $M_{w^2}(B^c)$ to the so-called discounted g-mean-variance optimality equation

$$\inf_{a\in A_g(x)} \left\{ 2r(x,a)g(x) + \int_{B^c} u(y)q(dy|x,a) - 2u(x)\alpha(x,a) \right\} = 0, \tag{4.5}$$

where σ_g^2 and $A_g(x)$ are as in Definition 2.3 and (4.1), respectively.

(b) For any $n \geq 0, x \in B^c$, let $u'_{n+1}(x)$ and $h_n \in F_g$ be such that

$$\begin{aligned} u'_{n+1}(x) &:= \inf_{a\in A_g(x)} \left\{ \frac{2r(x,a)g(x)}{2\alpha(x,a)+q(x,a)} + \frac{1}{2\alpha(x,a)+q(x,a)} \int_{B^c-\{x\}} u'_n(y)q(dy|x,a) \right\} \\ &=: \frac{1}{2\alpha(x,h_n(x))+q(x,h_n(x))} \Big[2r(x,h_n(x))g(x) + \int_{B^c-\{x\}} u'_n(y)q(dy|x,h_n(x)) \Big] \end{aligned}$$

with $u'_0(x) := -2M_1||g||_w \Big[\frac{b_1}{2\alpha_0(2\alpha_0-c_1)} + \frac{1}{2\alpha_0-c_1} w^2(x) \Big]$.

Then,

(b_1) there exists a g-mean-variance optimal policy (denoted as h^*), such that $h^*(x)$ is an accumulation point of $\{h_n(x)\}$ for each $x \in B^c$;

(b_2) (A value iteration algorithm): $(u'_n(x) - g^2(x)) \uparrow \sigma_g^2(x)$ $(x \in B^c)$ as $n \uparrow \infty$.

(c) A policy $f \in F$ is g-mean-variance optimal if and only if $f(x)$ attains the minimum in (4.5) for every $x \in B^c$ with u being replaced by $\sigma_g^2 + g^2$.

Proof. In line with the discussions before the statement, for problem (4.4), we consider a new continuous-time MDP model

$$\bar{\mathcal{M}} := \big\{ S, A, (A_g(x), x \in S), q(\cdot|x,a), \bar{c}(x,a), \bar{\alpha}(x,a), B \big\}, \tag{4.6}$$

where $A_g(x)$ as in (4.1),

$$\bar{c}(x,a) := 2r(x,a)g(x),\ \bar{\alpha}(x,a) := 2\alpha(x,a)$$

for all $x \in S$ and $a \in A_g(x)$, and the other elements are the same as in (2.1). Then, by Theorem 4.2 we have

$$\inf_{f\in F_g} J_B(x,f) - g^2(x) = \sigma_g^2(x)$$

for each $x \in B^c$. For this new model $\bar{\mathcal{M}}$, the corresponding versions of Assumptions A, B and C are satisfied. The statements follow from Theorem 3.4 applied to model (4.6). □

Using Theorem 4.4(b), we can give a value iteration algorithm for the g-variance value function, and the details are omitted. Moreover, as the arguments in [6, 23, 24, 25], we can give a policy iteration for computing a g-mean-variance optimal policy.

For the simplicity of statements of the policy iteration, we only consider the case when S and $A(x)$ are all finite.

For any given $f \in F_g, x \in B^c$, and $a \in A_g(x)$, let

$$u_f(x,a) := 2r(x,a)g(x) + \sum_{y \notin B} J_B(y,f)q(y|x,a),$$

with $J_B(f) = [2diag(\alpha(f)) - Q_B(f)]^{-1}c_g(f)$, and

$$B_f(x) := \{a \in A_g(x) : u_f(x,a) < 2\alpha(x,a)J_B(x,f)\}.$$

Define an *improvement policy* h of f as follows:

$$h(x) \in B_f(x) \text{ if } B_f(x) \neq \emptyset;\ h(x) := f(x) \text{ if } B_f(x) = \emptyset. \tag{4.7}$$

The policy iteration algorithm:

1. Compute $A_g(x)$ in (4.1), and then get $F_g = \sqcap_{x \in S} A_g(x)$.
2. Pick an arbitrary $f \in F_g$. Let $k = 0$, and take $f_k := f$.
3. Policy evaluation: Obtain $J_B(f_k) = [2diag(\alpha(f_k)) - Q_B(f_k)]^{-1}c_g(f_k)$.
4. Policy improvement: Obtain a policy f_{k+1} from (4.7) (with f_k and f_{k+1} in lieu of f and h, respectively).
5. If $f_{k+1} = f_k$, then stop because f_{k+1} is mean-variance optimal (by Theorem 4.5 below). Otherwise, increase k by 1 and return to step 3.

THEOREM 4.5. Suppose that S and $A(x)(x \in S)$ are all finite. Let $\{f_k\}$ be sequence obtained by the policy iteration algorithm. Then, the following assertions hold.

(a) $J_B(f_{k+1}) \leq J_B(f_k)$ and $J_B(f_{k+1}) \neq J_B(f_k)$ when $f_{k+1} \neq f_k$.

(b) There exists a finite number k^* such that f_{k^*} is optimal.

Proof. Since S and $A(x)$ are all finite, F (and hence F_g) is also finite. As in the proof of Theorem 3.2(b), for the continuous-time MDP model (4.6), we get that any function $v \in M_{w^2}(B^c)$ satisfying $2r(x,f(x))g(x) + \sum_{y \notin B} v(y)q(y|x,f(x)) \leq 2\alpha(x,f(x))v(x)$ implies $v(x) \geq J_B(x,f)$ for all $x \in B^c$. By the definition of f_k and f_{k+1} in (4.7) and $f_{k+1} \neq f_k$, we have

$$2r(x,f_{k+1}(x))g(x) + \sum_{y \notin B} J_B(y,f_k)q(y|x,f_{k+1}(x)) \leq 2\alpha(x,f_{k+1}(x))J_B(x,f_k),$$

which, together with the uniqueness in Theorem 4.2, implies the statement (a). Part (b) directly follows from the finiteness of F_g and part (a). □

Theorem 4.5 shows that a (g-mean-variance) optimal policy can be obtained by the policy iteration approach in a finite number of iterations.

5. Examples. In this section, we give two examples to illustrate the application of our main results. The first one with finite states and actions is used to show the difference between our results and those in the previous literature, and the second one about a cash flow model shows the potential applications of Theorem 4.4.

EXAMPLE 5.1. The continuous-time control model we are concerned with is given as follows: $S = \{x_1, x_2, x_3\}; B = \{x_3\}$; $A(x_1) = \{a_{11}, a_{12}\}$, $A(x_2) = \{a_{21}, a_{22}\}$ and $A(x_3) = \{a_{31}\}$; the transition rates, $q(\cdot|x,a)$, are defined by

$$q(x_1|x_1,a_{11}) = -1,\ q(x_2|x_1,a_{11}) = \frac{1}{8};\ q(x_1|x_1,a_{12}) = -4,\ q(x_2|x_1,a_{12}) = 3;$$

$$q(x_1|x_2,a_{21}) = \frac{3}{2},\ q(x_2|x_2,a_{21}) = -2;\ q(x_1|x_2,a_{22}) = 1,\ q(x_2|x_2,a_{22}) = -1,$$

the reward rates $r(x,a)$ and the expected mean performance $g(x)$, are given as

$$r(x_1,a_{11})=1,\ r(x_1,a_{12})=2,\ r(x_2,a_{21})=3,\ r(x_2,a_{22})=\frac{17}{13};\ g(x_1)=1,\ g(x_2)=2.$$

The policy set $F=\{f_1,f_2,f_3,f_4\}$, where $f_1(x_1)=a_{11}, f_1(x_2)=a_{21}$; $f_2(x_1)=a_{11}, f_2(x_2)=a_{22}$; $f_3(x_1)=a_{12}, f_3(x_2)=a_{21}$ and $f_4(x_1)=a_{12}, f_4(x_2)=a_{22}$.

Now we have the following result.

PROPOSITION 5.2. *For the control model in Example 5.1, we have the following assertions.*

(a)(On the case of state-action-dependent discount factors.) Suppose that the discount factors are given by

$$\alpha(x_1,a_{11})=\frac{1}{4},\ \alpha(x_1,a_{12})=1,\ \alpha(x_2,a_{21})=\frac{1}{4},\ \alpha(x_2,a_{22})=\frac{2}{13}.$$

Then, the set F_g is equal to $\{f_1,f_2\}$, and the policy f_2 is g-mean-variance optimal.

(b)(On the case of state-dependent discount factors.) Suppose that the discount factors are given by

$$\alpha(x_1,a_{11})=\alpha(x_1,a_{12})=4,\ \alpha(x_2,a_{21})=\alpha(x_2,a_{22})=\frac{1}{4}.$$

Then, the set F_g is equal to $\{f_3\}$, and the policy f_3 is g-mean-variance optimal.

(c)(On the case of a constant discount factor.) Suppose that a constant discount factor is given by

$$\alpha(x,a)\equiv\frac{1}{4}.$$

Then, the set F_g is equal to $\{f_1\}$, and the policy f_1 is g-mean-variance optimal.

Proof. Obviously, Example 5.1 satisfies all Assumptions needed in Theorem 4.4. (a) We next solve a g-mean-variance optimal policy by using the policy iteration algorithm with the following steps:

1) For each $x\in\{x_1,x_2\}$, solving the equation in (4.1) gives that $A_g(x_1)=\{a_{11}\}$ and $A_g(x_2)=\{a_{21},a_{22}\}$. Therefore (by Lemma 4.1), $F_g=\{f_1,f_2\}$.
2) Pick a policy $f_1\in F_g$. Take $h_0:=f_1$.
3) Policy evaluation: Obtain $J_B(h_0)=[2diag(\alpha(f_1))-Q_B(f_1)]^{-1}c_g(f_1)=\begin{pmatrix}\frac{104}{57}\\ \frac{112}{19}\end{pmatrix}$.
4) Policy improvement: From (4.7) with h_0 and h_1 in lieu of f and h respectively, we obtain a policy h_1 given as follows: $h_1(x_1)=a_{11}$, and $h_1(x_2)=a_{22}$.
5) Since $h_1\neq h_0$, a further iteration yields $h_2(x_1)=h_1(x_1)=a_{11}$ and $h_2(x_2)=h_1(x_2)=a_{22}$. Thus, f_2 is a g-mean-variance optimal policy.

(b) It follows from the equation in (4.1) that $A_g(x_1)=\{a_{12}\}$ and $A_g(x_2)=\{a_{21}\}$, and by the policy iteration algorithm we see that f_3 is g-mean-variance optimal.

(c) Similarly, we have $A_g(x_1)=\{a_{11}\}$ and $A_g(x_2)=\{a_{21}\}$, and also see that f_1 is g-mean-variance optimal. □

REMARK 5.3.

(a) Note that the discount factor in Example 5.1 may not be constant. Thus, to the best of our knowledge, the example here is not covered by the previous literature on mean-variance optimality problems with any constant discount factor.

(b) After a brief look at the optimal policies in Proposition 5.2, we see that the g-mean-variance optimal policies in the three cases of discount factors are all different. This verifies the existence of differences between our results and those in the previous literature.

To further illustrate the application of the obtained results, we introduce a cash flow model in the following.

EXAMPLE 5.4. (A cash flow model.) Consider a continuous-time controlled problem of cash flow in an economic market, in which the amount of the cash is referred to as the state of cash flow. Thus, the state space of the cash flow is $S = (-\infty, +\infty)$. Given the current state of cash flow $x \in S$, a control action $a \in A(x)$ is performed by withdrawing money with the amount $-a$ if $a < 0$ or taking a supply of money with the amount a for $a \geq 0$. When the current state is $x \in S$ and an action $a \in A(x)$ is chosen, a reward $r(x, a)$ is earned. In addition, the amount of cash x is assumed to keep invariable for an exponential-distributed random time with parameter $k(x, a) \geq 0$, and then the cash flow is supposed to jump to other states with the normal distribution $N(x, 1)$. Therefore, the transition rates of cash flow is represented by

$$q(D|x,a) := k(x,a)\Big[\int_D \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} dy - \delta_x(D)\Big] \qquad \text{for each } D \in \mathcal{B}(S). \tag{5.1}$$

Moreover, the discount factor is defined by

$$\alpha(x,a) := \beta(x,a) + \alpha \quad \forall\ (x,a) \in K$$

with some nonnegative function $\beta(\cdot,\cdot)$ on K and some constant $\alpha > 0$.

For this cash flow model, the decision maker wishes to minimize the variance over all policies having some given expected reward g before the state of cash flow falls in some target set $B \subset S$.

To ensure the existence of g-mean-variance optimal policies for the cash flow model, we consider the following hypotheses:

(C_1) $|k(x,a)| \leq M(x^2+1)$ and $|r(x,a)| \leq M_1(x^2+1)$ for all $x \in B^c, a \in A(x)$ with some positive constants M and M_1.

(C_2) $\alpha > 3M$ and $A(x)$ is assumed to be compact for each $x \in B^c$.

(C_3) $k(x,a)$, $r(x,a)$ and $\beta(x,a)$ are measurable on K and continuous in $a \in A(x)$ for each fixed $x \in B^c$.

Under the above conditions, we have the following fact.

PROPOSITION 5.5. *Under the hypotheses C_1–C_3, Example 5.4 satisfies Assumptions A, B, C^* and D, and hence (by Theorem 4.4) there exists a g-mean-variance optimal policy.*

Proof. To verify conditions required in Theorem 4.4, let $w(x) := x^2 + 1$ for all $x \in S$. Then, it follows from (5.1) that

$$\begin{aligned}
\int_S w(y)q(dy|x,a) &= k(x,a)\Big[\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}(y^2+1)e^{-\frac{(y-x)^2}{2}} dy - (x^2+1)\Big] \\
&= k(x,a) \leq M(x^2+1). \\
\int_S w^2(y)q(dy|x,a) &= k(x,a)\Big[\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}(y^2+1)^2 e^{-\frac{(y-x)^2}{2}} dy - (x^2+1)^2\Big] \\
&= k(x,a)(6x^2+5) \leq M(x^2+1)(6x^2+5) \\
&\leq 6M(x^2+1)^2,
\end{aligned}$$

which, together with the hypotheses C_1–C_3, yields Assumptions A, B, C^* and $D(1)$ with $c := M$, $b := 0$, $c_1 := 6M$, $b_1 := 0$ and $\alpha_0 := \alpha > 3M$. Using a similar argument, we can also verify Assumption $D(2)$. Thus, by Theorem 4.4, there exists a g-mean-variance optimal policy. □

COROLLARY 5.6. *For the following special data in Example 5.4 with the target set $B := \emptyset$,*

$$r(x,a) := x^2 - a^2 + 6x + 6,\ \beta(x,a) := 0,\ \alpha := 6,$$
$$k(x,a) := \frac{1}{2}(a - \frac{1}{2}x)^2,\ A(x) := [-|x|, |x|],\ g(x) := x + 1$$

for all $x \in S$ and $a \in A(x)$, then we have

(a) $A_g(x) = \{-x, x\}$ for all $x \in S$ (hence the set F_g is infinite);

(b) the policy $f^*(x) = x$ $(x \in S)$ is g-mean-variance optimal, and $\sigma_g^2(x) = \frac{1}{95}x^2$ for all $x \in S$.

Proof. Under the data given in Corollary 5.6, the hypothesis C_3 is obviously true. Also, we have

$$|r(x,a)| \le x^2 + 6 + 3(x^2+1) \le 9(x^2+1),\ k(x,a) \le \frac{9}{8}(x^2+1),$$

which implies the hypotheses C_1–C_2 with $M = \frac{9}{8}$, $M_1 = 9$ and $\alpha > \frac{27}{8}$. Thus, Proposition 5.5 ensures the existence of a g-mean-variance optimal policy. To solve $A_g(x)$ in (4.1), let the given data here in lieu of the ones in (4.1). A direct calculation yields that $A_g(x) = \{-x, x\}$ for all $x \in (-\infty, \infty)$. Then, Lemma 4.1 implies that there exist infinite policies in F_g. To further find a g-mean-variance optimal policy over the set F_g, (by Theorem 4.4) it only need to seek the policy which attains the minimum in (4.5) with $\sigma_g^2 + g^2$ in lieu of u. Indeed, for each $a \in A_g(x)$, we have $2r(x,a)g(x) = 12(x+1)^2$. Furthermore, we suppose for a moment that $\sigma_g^2(x) + g(x) =: k_0 + k_1 x + k_2 x^2$, with some constants k_0, k_1 and k_2 to be specified below. Then, by Theorem 4.4, the discounted g-mean-based variance optimality equation (4.5) becomes

$$\inf_{a \in A_g(x)} \{12(x+1)^2 + \frac{1}{2}k_2(a - \frac{1}{2}x)^2 - 12(k_0 + k_1 x + k_2 x^2)\} = 0. \tag{5.2}$$

Suppose that $k_2 > 0$ and let $\nu(x,a) := \frac{1}{2}k_2(a - \frac{1}{2}x)^2$. Then, the minimum of $\nu(x,a)$ over $a \in A_g(x)$ is $\nu(x,x) = \frac{1}{8}k_2 x^2$. Now, the function in the parenthesis in (5.2) is further reduced to

$$(12 + \frac{1}{8}k_2 - 12k_2)x^2 + (24 - 12k_1)x + (12 - 12k_0) = 0,$$

which implies that $k_2 = \frac{96}{95}, k_1 = 2$ and $k_0 = 1$. Indeed, $k_2 = \frac{96}{95}$ shows that our result based on the hypothesis $k_2 > 0$ is true. Therefore, the g-mean-variance optimal policy and the g-variance function are given by $f^*(x) := x$ and $\sigma_g^2(x) = \frac{1}{95}x^2$ for all $x \in S$. □

6. Concluding remarks. To sum up, this paper considered the g-mean-based variance minimization problem for continuous-time MDPs, in which the state and action spaces are general, transition and reward rates are unbounded, and in which the discount factors are state-action-dependent. One focuses on the variance minimization over the set F_g of all policies, whose total discounted reward (over the first passage of the controlled continuous-time MDP to some target set) attains a given reward

g. The g-mean-variance optimality equation was established, and the existence and characterization of a g-mean-variance optimal policy were given. The value and policy iteration algorithms were justified, too. In particular, when the states and actions are all finite, the policy iteration algorithm can be used to obtain a g-mean-variance optimal policy in finite iterations. The applications of the obtained results were demonstrated by two analytically solved examples, which can be used to show the difference among the three kinds of discount factors, and which seem not be covered by the previous literature on continuous-time MDPs.

Our study on the g-mean-based variance minimization problem is based on the setup $F_g := \{f \in F \mid V_B(x,f) = g(x) \text{ for all } x \in B^c\}$ (i.e. the all admissible policies obtaining the given reward g). The advantage of such setup of F_g is that the existence and computation of a g-mean-variance optimal policy can be established.

Unsolved problems: In fact, it is more adequate for applications that a controller is interested in policies such that the expected discounted reward is at least the quantity g. Hence, it is more natural and desirable to study the g-mean-based variance minimization problem (2.10) with F_g defined in one of the following two cases:

1) $F_g := \{f \in F \mid V_B(x,f) \geq g(x) \text{ for all } x \in B^c\}$;

2) $F_g := \{f \in F \mid \int_S V_B(x,f)\gamma(dx) \geq \int_S g(x)\gamma(dx)\}$, with a initial distribution γ on S.

However, it is a challenge and unsolved problem to consider the g-mean-based variance minimization problems (2.10) with F_g replaced with one in the two cases above.

REFERENCES

[1] N. Baüerle and U. Rieder, *Markov decision processes with applications to finance*, Springer, Heidelberg, 2011.

[2] E. A. Feinberg, *Continuous time discounted jump Markov decision processes: a discrete-event approach*, Math. Oper. Res., 29(2004), pp. 492–524.

[3] E. A. Feinberg, *Reduction of discounted continuous-time MDPs with unbounded jump and reward rates to discrete-time total-reward MDPs*, in Optimization, control, and applications of stochastic systems, Systems Control Found. Appl., D. Hernández-Hernández and A. Minjarez-Sosa, Eds., Birkhäuser/Springer, New York, 2012, pp. 77–97.

[4] Ch. P. Fu, A. Lari-Lavassani and X. Li, *Dynamic mean-variance portfolio selection with borrowing constraint*, European J. Oper. Res., 200(2010), pp. 312–319.

[5] X. P. Guo, *Continuous-time Markov decision processes with discounted rewards: the case of Polish spaces*, Math. Oper. Res., 32(2007), pp. 73–87.

[6] X. P. Guo and O. Hernández-Lerma, *Continuous-time Markov decision processes*, Springer, Heidelberg Dordrecht, New York, 2009.

[7] X. P. Guo and X. Y. Song, *Discounted continuous-time constrained Markov decision processes in Polish spaces*, Ann. Appl. Probab., 21(2011), pp. 2016–2049.

[8] X. P. Guo, X. Y. Song and Y. Zhang, *First passage optimality for continuous-time Markov decision processes with varying discount factors and history-dependent policies*, IEEE Trans. Automat, Control, 59(2014), pp. 163–174.

[9] X. P. Guo and A. Piunovskiy, *Discounted continuous-time Markov decision processes with constraints: unbounded transition and loss rates*, Math. Oper. Res., 36(2011), pp. 105–132.

[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), pp. 423–429.

[11] O. Hernández-Lerma and T. E. Govindan, *Nonstationary continuous-time Markov control processes with discounted costs on infinite horizon*, Acta Appl. Math., 67(2001), pp. 277–293.

[12] O. Hernández-Lerma and J. B. Lasserre, *Further topics on discrete-time Markov control processes*, Springer-Verlag, New York, 1999.

[13] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, Springer-Verlag, New York, 1996.

[14] Q. Y. Hu, *Continuous-time Markov decision processes with discounted moment criterion*, J. Math. Anal. Appl., 203(1996), pp. 1–12.
[15] S. C. Jaquette, *Markov decision processes with a new optimality criterion: continuous time*, Ann. Statist., 3(1975), 547–553.
[16] P. Kakumanu, *Continuously discounted Markov decision models with countable state and action space*, Ann. Math. Statist, 42(1971), 919–926.
[17] S. A. Lippman, *Applying a new device in the optimization of exponential queuing systems*, Oper. Res., 23(1975), pp. 687–710.
[18] J. Y. Liu and S. M. Huang, *Markov decision processes with distribution function criterion of first-passage time*, Appl. Math. Optim., 43(2001), pp. 187–201.
[19] H. M. Markowitz, *Portfolio selection*, J. Finance, 7(1952), pp. 77–91.
[20] H. M. Markowitz, *Mean-variance analysis in portfolio choice and capital markets*, Basil Blackwell, Oxford, UK, 1987.
[21] B. L. Miller, *Finite state continuous time Markov decision processes with an infinite planning horizon*, J. Math. Anal. Appl., 22(1968), pp. 552–569.
[22] A. Piunovskiy and Y. Zhang, *Discounted continuous-time Markov decision processes with unbounded rates: the convex analytic approach*, SIAM J. Control. Optim., 49(2011), pp. 2032–2061.
[23] T. Prieto-Rumeau and O. Hernández-Lerma, *Selected topics on continuous-time controlled Markov chains and Markov games*, Imperial College Press, London, 2012.
[24] M. L. Puterman, *Markov decision processes*, Wiley, New York, 1994.
[25] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*, Wiley, New York, 1999.
[26] L. E.Ye and X. P. Guo, *Continuous-time Markov decision processes with state-dependent discount factors*, Acta Appl. Math., 121(2012), pp. 5–27.
[27] G. Yin and X. Y. Zhou, *Markowitz's mean-variance portfolio selection with regime switching: from discrete-time models to their continuous-time limits*, IEEE Trans. Automat. Control, 49(2004), pp. 349–360.