J. Appl. Prob. **52,** 441–456 (2015)
Printed in England

FIRST PASSAGE OPTIMALITY AND VARIANCE MINIMISATION OF MARKOV DECISION PROCESSES WITH VARYING DISCOUNT FACTORS

XIAO WU * ** AND
XIANPING GUO,* *** *Sun Yat-Sen University*

Abstract

This paper deals with the first passage optimality and variance minimisation problems of discrete-time Markov decision processes (MDPs) with varying discount factors and unbounded rewards/costs. First, under suitable conditions slightly weaker than those in the previous literature on the standard (infinite horizon) discounted MDPs, we establish the existence and characterisation of the first passage expected-optimal stationary policies. Second, to further distinguish the expected-optimal stationary policies, we introduce the variance minimisation problem, prove that it is equivalent to a *new* first passage optimality problem of MDPs, and, thus, show the existence of a variance-optimal policy that minimises the variance over the set of all first passage expected-optimal stationary policies. Finally, we use a *computable* example to illustrate our main results and also to show the difference between the first passage optimality here and the standard discount optimality of MDPs in the previous literature.

Keywords: Discrete-time Markov decision process; varying discount factor; unbounded reward; first passage optimality; variance minimisation

2010 Mathematics Subject Classification: Primary 90C40
Secondary 93E20; 60J27

1. Introduction

This paper is a *first* attempt to study the first passage optimality and the variance minimisation for discrete-time Markov decision processes (MDPs) with varying discount factors and unbounded rewards.

As is well known, infinite horizon discounted (discrete-time) MDPs have been widely discussed according to the various forms of discount factors, which can be classified into the following four groups: (i) MDPs with a fixed constant discount factor α, (ii) MDPs with multiple constant discount factors $\alpha_1, \alpha_2, \ldots, \alpha_n$, (iii) MDPs with randomised discount factors, in which the process of randomised discount factors is assumed to be independent of the state process of MDPs, (iv) MDPs with varying (state-dependent) discount factors $\alpha(x)$. In group (i), the focus of study is on the conditions for the existence and computation of optimal stationary policies (see, for instance, [9], [10], [20] and the references therein). Concerning group (ii), an optimal Markov policy that is stationary from some finite time onwards was obtained in [3] for the case of finite states and finite actions, and an algorithm for calculating such optimal policies was given. In group (iii), the existence of asymptotically discounted optimal policies

Received 5 July 2013; revision received 1 July 2014.
* Postal address: School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, P. R. China.
** Email address: jxwuxiao@126.com
*** Email address: mcsgxp@mail.sysu.edu.cn

was proved in [1] and [4]. In group (iv), theory and applications on varying discount factors are given in [18], [21], and [23] for the stationary MDPs and [7] for the nonstationary MDPs, where the optimality equation and existence of optimal policies are established under suitable conditions. In most of the existing works on the four groups of discounted MDPs, the time horizon of MDPs is either finite or infinite. However, in real situations, we usually need to find an optimal policy of the MDPs, in which the time horizon is *random*, such as the first passage time to some given set. For instance, in portfolio or insurance optimisation it is usually desirable to maximise the wealth before falling below a certain baseline capital or bankruptcy. This motivates the work on the so-called first passage optimality of MDPs. This paper will further study the first passage optimality problem and focus on the group (iv).

The *first passage optimality problem* of MDPs usually refers to maximising an expected total reward until the first time that the state process enters a given *target set*, also known as a *stopping set*. Evidently, the first passage criterion in the first optimality problem of MDPs is a generalization of the standard discount criterion in group (i) above, see Remark 2.2 for details. The first passage problem has been studied in [2], [7], [13], [14], [16], [17], and [24] for the existence of the optimal policies and [21] for the applications to the so-called ruin problem, as well as [1], [14], and [18] for the applications to reliability, maintenance, and quality control problems. More precisely, Derman [2] studied a first passage problem for discrete-time MDPs with finite state and action spaces, where a target state is absorbing, proving the existence of an optimal stationary policy, and also giving successive approximations, policy improvement, and linear programming to obtain optimal solutions. Furthermore, Liu and Liu [17] discussed the first passage model for discrete-time MDPs with denumerable states and bounded rewards, and proved that the uniqueness of the bounded solution to the optimality equation and the existence of an optimal stationary policy under some assumptions. In addition, Liu and Huang [16] and Yu *et al.* [24] discussed the first passage distribution function optimality criterion of MDPs with countable state and action spaces, and then gave some properties of several kinds of optimal policies and the existence and algorithms for these optimal policies. Afterwards, Huang and Guo [13], [14] considered a first passage model for discounted semi-MDPs with denumerable states and nonnegative costs, and proved that the optimal value function satisfies the optimality equation and there exists an optimal stationary policy under suitable conditions by using a minimum nonnegative solution approach. Recently, Guo *et al.* [7] have studied the first passage problems for nonstationary nonlinear discrete-time stochastic control systems with the time- and state-dependent discount factors and nonnegative unbounded or uniformly bounded rewards, and proved that the optimal reward functions satisfy the optimality equations and there exists an optimal Markov policy. It is worth noting that most of the above works about the first passage problems focus on the MDPs with bounded or nonnegative unbounded rewards. However, the first passage optimality problem of MDPs with state-dependent discount factors and unbounded costs/rewards from above and below has yet to be studied.

As is well known, for a given standard discount or the first passage criterion of MDPs, one can calculate the optimal policies under appropriate conditions. However, it is desirable to distinguish the *best* policy when the optimal policies are not unique. Thus, to further distinguish optimal policies, a decision-maker aims to take a policy that has minimal variance over the set of the optimal policies. This is *the variance minimisation problem* of MDPs. There are many papers on the variance minimisation problem of MDPs, see, for instance, [5], [6], [10], [11], [15], and [19] for the expected average criterion in MDPs, and [8], [10], and [22] for the infinite horizon discounted criterion of MDPs. To the best of the authors' knowledge, the variance minimisation problem for the first passage criterion of discounted MDPs with unbounded

rewards has not been studied. Hence, this paper concerns the first passage optimality and the corresponding variance minimisation problems of the MDPs with varying discount factors. For the first passage optimality problem, we present some suitable conditions weaker than those in [10] for the infinite horizon MDPs with a constant discount factor. Under these conditions, we prove the existence of a first passage expected-optimal stationary policy and the uniqueness of a solution to the first passage optimality equation (Theorem 3.2), and, thus, extend the corresponding results in [10] to the case of MDPs with varying discount factors. To solve the variance minimisation problem, we prove that this problem can be transformed into an *equivalent* first passage optimality problem of another discounted MDP with a *new* reward function and *new* action sets. Afterwards, we establish the optimality equation for the variance minimisation problem, and show the existence of a so-called variance-optimal policy (Theorem 5.1). Finally, to further illustrate the main results in this paper, we present a *computable* cash-balance model (Example 6.1) with varying discount factors, for which we obtain the different first passage expected-optimal policies and optimal value functions for different target sets, and also obtain a variance-optimal policy. This also shows the difference between the first passage optimality and the infinite horizon discounted optimality of MDPs with varying discount factors; see Proposition 6.1 and Remark 6.1 for details.

The organization of this paper is as follows. In Section 2 we formulate the control model for discrete-time MDPs and state the concerned first passage optimality and variance minimisation problems. In Section 3 we establish the first passage optimality equation and give conditions for the existence of the first passage expected-optimal stationary policies. After the establishment of a relationship between the first passage variance and a new discounted expected reward of a policy in Section 4, we obtain the variance optimality results in Section 5. Finally, we illustrate our main results with an application example in Section 6.

2. The control model

The model of discount MDPs is of the form

$$\{X, A, (A(x),\ x \in X),\ Q(\cdot \mid x, a),\ B, \alpha(x), r(x, a)\}, \tag{2.1}$$

where X and A are state and action spaces, which are assumed to be Borel spaces with Borel σ-fields $\mathcal{B}(X)$ and $\mathcal{B}(A)$, respectively, and $A(x)$ denotes the set of admissible actions at state $x \in X$. Let $K := \{(x, a) \mid x \in X, a \in A(x)\}$ be the set of all feasible state-action pairs and suppose that K is a measurable Borel subset of $X \times A$. The transition law $Q(\cdot \mid x, a)$ is a stochastic kernel on X given K, which denotes the one-step (homogeneous) transition probability. Moreover, $B \in \mathcal{B}(X)$ is a given set of target states, and $\alpha(x)$, the discount factors, are measurable functions from X to $[0, 1)$. Finally, $r(x, a)$ is a real-valued reward function and is assumed to be measurable on K.

Remark 2.1. In some cases, $r(x, a)$ is allowed to take positive and negative values. So, it can also be interpreted as a cost.

To introduce a policy, we require some notation. For each $n \geq 0$, let H_n be the family of admissible histories up to time n, that is, $H_0 := X$ and $H_n := K^n \times X = K \times H_{n-1}$.

Definition 2.1. A *randomised history-dependent policy* is a sequence $\pi := \{\pi_n,\ n \geq 0\}$ of stochastic kernels π_n on A given H_n satisfying

$$\pi_n(A(x_n) \mid h_n) = 1 \quad \text{for all } h_n := (x_0, a_0, x_1, a_1, \ldots, x_n) \in H_n,\ n \geq 0.$$

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

The set of all randomised history-dependent policies is denoted by Π. Furthermore, we denote by Φ the set of all stochastic kernels φ on A given X such that $\varphi(A(x) \mid x) = 1$ for all $x \in X$, and F the set of all measurable functions $f\colon X \to A$ satisfying $f(x) \in A(x)$ for all $x \in X$. It is clear that $F \subset \Phi$.

Definition 2.2. A policy $\pi = \{\pi_n\} \in \Pi$ is said to be a *randomised stationary policy* if there is a stochastic kernel $\varphi \in \Phi$ such that $\pi_n(\cdot \mid h_n) = \varphi(\cdot \mid x_n)$ for each $h_n \in H_n$ and $n \geq 0$. Such a randomised stationary policy will be denoted by φ. A randomised stationary policy $\varphi \in \Phi$ is said to be *(deterministic) stationary* if there is a function $f \in F$ such that $\varphi(\cdot \mid x)$ is the Dirac measure at $f(x)$ for all $x \in X$. We will denote such a stationary policy by f, and regard F as the set of all stationary policies for simplicity.

For any initial distribution μ on X and $\pi = \{\pi_n\} \in \Pi$, by the well-known Tulcea's theorem [9, p. 178], there exist a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_\mu^\pi)$ and a stochastic process $\{x_n, a_n, n \geq 0\}$ such that for each $C \in \mathcal{B}(X)$, $D \in \mathcal{B}(A)$ and $n \geq 0$,

$$\mathbb{P}_\mu^\pi\{x_0 \in C\} = \mu(C), \tag{2.2}$$

$$\mathbb{P}_\mu^\pi\{a_n \in D \mid h_n\} = \pi_n(D \mid h_n), \tag{2.3}$$

$$\mathbb{P}_\mu^\pi\{x_{n+1} \in C \mid h_n, a_n\} = Q(C \mid x_n, a_n). \tag{2.4}$$

We will denote by $\mathbb{E}_\mu^\pi$ the expectation with respect to $\mathbb{P}_\mu^\pi$ in (2.2)–(2.4). In particular, if μ is the Dirac measure concentrated at some state x, we write $\mathbb{P}_\mu^\pi$ and $\mathbb{E}_\mu^\pi$ as $\mathbb{P}_x^\pi$ and $\mathbb{E}_x^\pi$, respectively.

In addition, for each $(x, a) \in K$ and $f \in F$, let

$$r(x, f) := r(x, f(x)) \quad \text{and} \quad Q(\cdot \mid x, f) := Q(\cdot \mid x, f(x)).$$

For the target set B, we denote the first passage time of the state process $\{x_n, n \geq 0\}$ into B by $\tau_B := \inf\{n \geq 0 \mid x_n \in B\}$ (with $\inf \varnothing := \infty$).

Definition 2.3. For each $\pi \in \Pi$ and $x \in X$, the first passage expected criterion in this paper is defined as follows:

$$V(x, \pi) := \mathbb{E}_x^\pi \left\{ r(x_0, a_0) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha(x_i) r(x_n, a_n) \right\}, \tag{2.5}$$

which is called the *first passage discounted expected reward* of π.

Remark 2.2. (a) In (2.5) and what follows for any sequence $\{y_k\}$, we use the convention

$$\sum_{k=n}^{m} y_k := 0 \quad \text{and} \quad \prod_{k=n}^{m} y_k := 1 \quad \text{if } m < n.$$

(b) If $B = \varnothing$ then $\tau_B = \infty$, and if $\alpha(\cdot) \equiv \alpha$ is a constant in (0,1) for all $n \geq 0$, then $V(x, \pi)$ in (2.5) becomes the standard (infinite-horizon) α-discounted reward in [10, p. 43].

Definition 2.4. Denote by $V^*(x)$ the *optimal value function*, where $V^*(x) := \sup_{\pi \in \Pi} V(x, \pi)$ for all $x \in B^c$, and a policy $\pi^* \in \Pi$ is called the *first passage expected-optimal* if $V^*(x) = V(x, \pi^*)$ for all $x \in B^c$.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

The first passage optimality problem in this paper is to provide conditions for the existence of first passage expected-optimal stationary policies and also characterise the optimal value function V^*. Under suitable conditions, the class of first passage expected-optimal stationary policies, denoted by

$$F^* := \{f \in F \mid V^*(x) = V(x, f) \text{ for all } x \in B^c\}, \tag{2.6}$$

is nonempty. Moreover, in many situations we find that F^* has more than one element (see Example 6.1 below). For this case, how do we further choose the *best* policy in F^*? To deal with such a question, as in [8], [10], and [11] for the infinite-horizon MDPs, we now introduce a so-called first passage variance of a policy, which is used to distinguish the best policy in F^*.

Definition 2.5. For a policy $f \in F$ and $x \in B^c$, the first passage variance of f is given by

$$\sigma^2(x, f) := \mathbb{E}_x^f \left\{ r(x_0, a_0) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha(x_i) r(x_n, a_n) - V(x, f) \right\}^2, \tag{2.7}$$

and the *variance minimisation problem* is as follows:

$$\text{P: minimise } \sigma^2(x, f) \text{ over } f \in F^* \text{ for all } x \in B^c.$$

Definition 2.6. A policy $f^* \in F^*$ is called *variance-optimal* for the first passage criterion if

$$\sigma^2(x, f^*) \le \sigma^2(x, f) \quad \text{for all } f \in F^*,\ x \in B^c.$$

The minimal variance value is $\sigma_*^2(x) := \min_{f \in F^*} \sigma^2(x, f)$ for all $x \in B^c$.

Obviously, a variance-optimal policy f^* is a first passage expected-optimal policy, and it has the minimal variance among all the policies in F^*.

The main goal of this paper is to find conditions for the existence of a policy that minimises the variance over the set of all stationary policies with the maximum first passage discounted expected reward.

3. First passage discounted-reward optimality results

We assume throughout that $\omega\colon B^c \to [1, \infty)$ denotes a Borel-measurable function that will be referred to as a weight function, and denote by $B_\omega(B^c)$ the Banach space of real-valued measurable functions u on B^c with the finite norm $\|u\|_\omega := \sup_{x \in B^c}(|u(x)|/\omega(x))$.

Since $r(x, a)$ may be unbounded, in order to guarantee the finiteness of $V(x, \pi)$, we need the conditions below.

Assumption 3.1.

(a) *There exists a constant* $\alpha \in (0, 1)$ *such that* $\sup_{x \in B^c} \alpha(x) \le \alpha$.

(b) *There exist nonnegative constants* β *and* γ, *with* $0 < \gamma < 1/\alpha$, *and a weight function* $\omega \ge 1$ *on* B^c *such that for each* $x \in B^c$,

$$\sup_{a \in A(x)} |r(x, a)| \le \beta\omega(x) \quad \textit{and} \quad \sup_{a \in A(x)} \int_{B^c} \omega^2(y) Q(\mathrm{d}y \mid x, a) \le \gamma^2 \omega^2(x).$$

Remark 3.1. (a) Assumption 3.1(a) is satisfied when the set B^c is finite, and it is the generalization of the constant discount factor case $\alpha(x) \equiv \alpha \in (0, 1)$. Assumption 3.1(b) implies that the reward function is allowed to have neither upper nor lower bounds.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

(b) By the Jensen inequality, Assumption 3.1(b) implies that $\sup_{a\in A(x)} \int_{B^c} \omega(y) Q(\mathrm{d}y \mid x, a) \le \gamma\omega(x)$, which is the same as Assumption 8.3.2 of [10, p. 44] for the case of $B = \varnothing$.

Lemma 3.1. *Suppose that Assumption 3.1 holds. Then for any $\pi \in \Pi$ and $x \in B^c$,*

$$|V(x,\pi)| \le \frac{\beta}{1-\alpha\gamma}\omega(x). \tag{3.1}$$

Proof. For each $x \in B^c$, the definition of $V(x,\pi)$ means

$$V(x,\pi) = \mathbb{E}_x^\pi \left\{ r(x_0,a_0)\,\mathbf{1}_{\{x_0\in B^c\}} + \sum_{k=1}^{\infty} \mathbf{1}_{\{x_0\in B^c, x_1\in B^c,\dots,x_k\in B^c\}} \prod_{i=0}^{k-1} \alpha(x_i) r(x_k,a_k) \right\}, \tag{3.2}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. By Assumption 3.1(b) and the Markov properties, we obtain

$$\mathbb{E}_x^\pi \{\mathbf{1}_{\{x_0\in B^c, x_1\in B^c,\dots,x_k\in B^c\}}\, \omega(x_k)\} \le \gamma^k \omega(x), \qquad k = 0, 1, 2, \dots. \tag{3.3}$$

From (3.2) and Assumption 3.1, we obtain

$$|V(x,\pi)| \le \beta\omega(x) + \beta \sum_{k=1}^{\infty} (\alpha\gamma)^k \omega(x) = \frac{\beta}{1-\alpha\gamma}\omega(x) \quad \text{for all } x \in B^c,\ \pi \in \Pi.$$

Theorem 3.1. *Suppose that Assumption 3.1 holds. Then for each fixed $f \in F$, $V(x,f)$ is the unique solution in $B_\omega(B^c)$ to the following:*

$$V(x,f) = r(x,f) + \alpha(x) \int_{B^c} V(y,f) Q(\mathrm{d}y \mid x, f) \quad \textit{for all } x \in B^c. \tag{3.4}$$

Proof. For each $x \in B^c$ and $f \in F$, by (11), we have

$$\begin{aligned} V(x,f) &= \mathbb{E}_x^f \left\{ \mathbb{E}_x^f \left[r(x_0,a_0) + \sum_{k=1}^{\infty} \mathbf{1}_{\{x_0\in B^c, x_1\in B^c,\dots,x_k\in B^c\}} \prod_{i=0}^{k-1} \alpha(x_i) r(x_k,a_k) \right] \,\middle|\, h_1 \right\} \\ &= r(x,f) + \int_{A(x)\times B^c} \alpha(x) \mathbb{E}_{x_1}^f \left\{ \mathbf{1}_{\{x_1\in B^c\}}\, r(x_1,a_1) \right. \\ &\quad \left. + \sum_{k=1}^{\infty} \mathbf{1}_{\{x_1\in B^c,\dots,x_k\in B^c\}} \prod_{i=1}^{k-1} \alpha(x_i) r(x_k,a_k) \right\} Q(\mathrm{d}x_1 \mid x, f) \\ &= r(x,f) + \alpha(x) \int_{B^c} V(y,f) Q(\mathrm{d}y \mid x, f). \end{aligned}$$

Note that, by (3.1), it is clear that $V(x,f) \in B_\omega(B^c)$. Now, for any $u(x) \in B_\omega(B^c)$, we define an operator T_f on $B_\omega(B^c)$ as follows:

$$(T_f u)(x) := r(x,f) + \alpha(x) \int_{B^c} u(y) Q(\mathrm{d}y \mid x, f) \quad \text{for all } x \in B^c.$$

Then, by Assumption 3.1, we have $|(T_f u)(x)| \le \beta\omega(x) + \alpha\gamma \|u\|_\omega \omega(x)$. Furthermore, we can prove that T_f is a contraction operator on $B_\omega(B^c)$, and so T_f has a unique fixed point u^* in $B_\omega(B^c)$, that is,

$$u^*(x) = (T_f u^*)(x) = r(x,f) + \alpha(x) \int_{B^c} u^*(y) Q(\mathrm{d}y \mid x, f) = V(x,f) \quad \text{for all } x \in B^c.$$

Hence, $V(x,f)$ is a unique solution to (3.4) in $B_\omega(B^c)$.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

Theorem 3.1 provides a characterisation of $V(x, f)$. To further establish the existence of a first passage expected-optimal policy, we consider the usual conditions below.

Assumption 3.2. *For each $x \in B^c$,*

(a) *the set $A(x)$ is compact;*

(b) *the reward function $r(x, a)$ is upper semi-continuous (u.s.c.) in $a \in A(x)$;*

(c) *for each $x \in X$ and $C \in \mathcal{B}(X)$, the functions $Q(C \mid x, a)$ and $\int_{B^c} \omega(y) Q(\mathrm{d}y \mid x, a)$ are u.s.c. in $a \in A(x)$.*

Remark 3.2. Assumption 3.2 is referred to as the *continuity-compactness conditions*. If we consider the reward function $r(x, a)$ as $-c(x, a)$, then Assumption 3.2 is essentially the same as Assumptions 8.3.1 and 8.3.3 of [10, pp. 44–45].

In the following, we will show that under Assumptions 3.1 and 3.2, $V^*(x) = \sup_{\pi \in \Pi} V(x, \pi)$ is the unique solution in $B_\omega(B^c)$ to the *first passage optimality equation* below:

$$u(x) = \sup_{a \in A(x)} \left\{ r(x, a) + \alpha(x) \int_{B^c} u(y) Q(\mathrm{d}y \mid x, a) \right\} \quad \text{for all } x \in B^c, \tag{3.5}$$

where u is in $B_\omega(B^c)$.

Theorem 3.2. *Under Assumptions 3.1 and 3.2, the following assertions hold.*

(a) *The optimal reward function V^* satisfies the first passage optimality equation (3.5), and it is the unique solution in $B_\omega(B^c)$ to (3.5).*

(b) *There exists a stationary policy $f^* \in F$ such that*

$$V^*(x) = r(x, f^*) + \alpha(x) \int_{B^c} V^*(y) Q(\mathrm{d}y \mid x, f^*) \quad \textit{for all } x \in B^c, \tag{3.6}$$

and f^* is a first passage expected-optimal policy.

(c) *A policy $f \in F$ is in F^* if and only if $f(x)$ attains the maximum in (3.5) (with the function u replaced by V^*) for every $x \in B^c$.*

Proof. (a) Define the operator T on $B_\omega(B^c)$ as follows:

$$Tu(x) := \sup_{a \in A(x)} \left\{ r(x, a) + \alpha(x) \int_{B^c} u(y) Q(\mathrm{d}y \mid x, a) \right\} \quad \text{for all } x \in B^c \text{ and } u \in B_\omega(B^c).$$

Then, as in the proof of Theorem 3.1, we can conclude that T is a contraction map from $B_\omega(B^c)$ into itself. By Banach's fixed point theorem, T has a unique fixed point u^* in $B_\omega(B^c)$, i.e. $Tu^* = u^*$. Hence, to prove part (a) we need to verify that $u^* = V^*$ for each $x \in B^c$.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

First, we show that $u^*(x) \le V^*(x)$ for each $x \in B^c$. In fact, by the equality $Tu^* = u^*$ and Lemma 8.3.8 of [10], there exists an $f^* \in F$ such that

$$\begin{aligned} u^*(x) &= r(x, f^*(x)) + \alpha(x) \int_{B^c} u^*(y) Q(\mathrm{d}y \mid x, f^*(x)) \\ &= \mathbb{E}_x^{f^*} \{ r(x_0, f^*(x_0)) \mathbf{1}_{\{x_0 \in B^c\}} + \alpha(x_0) \mathbf{1}_{\{x_0 \in B^c, x_1 \in B^c\}} u^*(x_1) \} \\ &= \mathbb{E}_x^{f^*} \Bigg\{ r(x_0, f^*(x_0)) + \sum_{k=1}^{N-1} \mathbf{1}_{\{x_0 \in B^c, x_1 \in B^c, \ldots, x_k \in B^c\}} \prod_{i=0}^{k-1} \alpha(x_i) r(x_k, f^*(x_k)) \\ &\quad + \mathbf{1}_{\{x_0 \in B^c, x_1 \in B^c, \ldots, x_N \in B^c\}} \prod_{i=0}^{N-1} \alpha(x_i) u^*(x_N) \Bigg\}, \end{aligned} \tag{3.7}$$

where the last equality can be deduced by iteration. Since, by (3.3),

$$\lim_{N \to \infty} \left| \mathbb{E}_x^{f^*} \left\{ \mathbf{1}_{\{x_0 \in B^c, \ldots, x_N \in B^c\}} \prod_{i=0}^{N-1} \alpha(x_i) u^*(x_N) \right\} \right| \le \lim_{N \to \infty} (\alpha\gamma)^N \|u^*\|_\omega \omega(x) = 0, \tag{3.8}$$

it follows from (3.2) and (3.7) that $u^*(x) = V(x, f^*)$, and so $u^* \le V^*$ for each $x \in B^c$.

Second, we show that $u^*(x) \ge V^*(x)$ for each $x \in B^c$. Since

$$u^*(x) \ge r(x, a) + \alpha(x) \int_{B^c} u^*(y) Q(\mathrm{d}y \mid x, a) \quad \text{for all } x \in B^c \text{ and } a \in A(x).$$

Then for each $n \ge 0$, $\pi \in \Pi$, and $x \in B^c$, we have

$$\mathbb{E}_x^\pi \{ r(x_n, a_n) \mathbf{1}_{\{x_n \in B^c\}} + \mathbf{1}_{\{x_n \in B^c, x_{n+1} \in B^c\}} \alpha(x_n) u^*(x_{n+1}) - \mathbf{1}_{\{x_n \in B^c\}} u^*(x_n) \mid h_n, a_n \} \le 0.$$

Hence,

$$\begin{aligned} &\mathbb{E}_x^\pi \Bigg\{ \mathbf{1}_{\{x_0 \in B^c, \ldots, x_n \in B^c\}} \prod_{i=0}^{n-1} \alpha(x_i) r(x_n, a_n) + \mathbf{1}_{\{x_0 \in B^c, \ldots, x_{n+1} \in B^c\}} \prod_{i=0}^{n} \alpha(x_i) u^*(x_{n+1}) \\ &\qquad - \mathbf{1}_{\{x_0 \in B^c, \ldots, x_n \in B^c\}} \prod_{i=0}^{n-1} \alpha(x_i) u^*(x_n) \mid h_n, a_n \Bigg\} \\ &\quad \le 0. \end{aligned}$$

Taking expectation $\mathbb{E}_x^\pi$ and then summing from 0 to $N-1$, we obtain

$$\begin{aligned} &\mathbb{E}_x^\pi \Bigg\{ r(x_0, f(x_0)) \mathbf{1}_{\{x_0 \in B^c\}} + \sum_{n=1}^{N-1} \mathbf{1}_{\{x_0 \in B^c, \ldots, x_n \in B^c\}} \prod_{i=0}^{n-1} \alpha(x_i) r(x_n, a_n) \\ &\qquad + \mathbf{1}_{\{x_0 \in B^c, \ldots, x_N \in B^c\}} \prod_{i=0}^{N-1} \alpha(x_i) u^*(x_N) \Bigg\} \\ &\quad \le u^*(x). \end{aligned}$$

Letting $N \to \infty$ and using (3.8), we obtain $V(x, \pi) \le u^*(x)$ for each $\pi \in \Pi$ and $x \in B^c$. Then, it is clear that $V^*(x) \le u^*(x)$ for each $x \in B^c$. Therefore, part (a) holds.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

(b) Since $V^*(x) = V(x, f^*)$ for every $x \in B^c$ (with f^* as in (3.7)), by part (a) and Theorem 3.1, we see that (3.6) holds.

(c) If $f \in F$ is a first passage expected-optimal policy, then $V^*(x) = V(x, f)$ for each $x \in B^c$, together with (3.4) implies that $f(x)$ attains the maximum in (3.5) for all $x \in B^c$. Conversely, it obviously holds by the definition of the first passage expected-optimal policy.

Remark 3.3. (a) Theorem 3.2 extends Theorem 8.3.6 of [10, p. 47] for a constant discount factor to the case of state-dependent discount factors, and also generalizes Theorem 3.2 of [23] to the case of the first passage criterion.

(b) The uniqueness of the solution to the first passage optimality equation (3.5) plays a crucial role in solving expected-optimal stationary policies (see the proof of Theorem 4.1(b) and Example 6.1 below).

Since we will find the minimum of $\sigma^2(x, f)$ over F^* in this paper, it is helpful to characterise the set F^*. Let

$$A^*(x) := \left\{ a \in A(x) \mid V^*(x) = r(x, a) + \alpha(x) \int_{B^c} V^*(y) Q(\mathrm{d}y \mid x, a) \right\} \quad \text{for all } x \in B^c, \tag{3.9}$$

where $V^*(x)$ is the optimal reward function.

Then, by Theorem 3.2(c), we have the following fact:

$$F^* = \{ f \in F \mid f(x) \in A^*(x) \text{ for all } x \in B^c \},$$

which gives an exact characterisation of F^* by means of $A^*(x)$ defined in (3.9).

4. Transformation of the variance problem

In this section we will show that the variance of $f (\in F^*)$ can be transformed into a first passage discounted expected reward of f for a new reward function and other discount factors.

Theorem 4.1. *For any $x \in B^c$ and $f \in F$, let*

$$R(x, f) := r(x, f)[2V(x, f) - r(x, f)] + \alpha^2(x) \int_{B^c} V^2(y, f) Q(\mathrm{d}y \mid x, f) - V^2(x, f). \tag{4.1}$$

Under Assumptions 3.1 and 3.2, we have the following:

(a) *for each $f \in F$, $\sigma^2(x, f)$ is the unique solution in $B_{\omega^2}(B^c)$ to the following equation*

$$u(x) = R(x, f) + \alpha^2(x) \int_{B^c} u(y) Q(\mathrm{d}y \mid x, f) \quad \textit{for all } x \in B^c, \tag{4.2}$$

(b) *for each $f \in F$ and $x \in B^c$,*

$$\sigma^2(x, f) = \mathbb{E}_x^f \left\{ R(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) R(x_n, f) \right\} \tag{4.3}$$

$$= \mathbb{E}_x^f \left\{ \tilde{r}(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) \tilde{r}(x_n, f) \right\} - V^2(x, f), \tag{4.4}$$

where $\tilde{r}(x, f) := r(x, f)[2V(x, f) - r(x, f)]$.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

Proof. (a) For each $f \in F$ and $x \in B^c$ from (2.7), we obtain

$$\sigma^2(x, f) = r^2(x, f) + 2\mathbb{E}_x^f\left\{r(x_0, f)\sum_{k=1}^{\infty}\mathbf{1}_{\{x_0\in B^c, x_1\in B^c, \dots, x_k\in B^c\}}\prod_{i=0}^{k-1}\alpha(x_i)r(x_k, f)\right\}$$
$$+ \mathbb{E}_x^f\left\{\sum_{k=1}^{\infty}\mathbf{1}_{\{x_0\in B^c, x_1\in B^c, \dots, x_k\in B^c\}}\prod_{i=0}^{k-1}\alpha(x_i)r(x_k, f)\right\}^2 - V^2(x, f). \tag{4.5}$$

Let

$$\begin{aligned} I_1 &:= 2\mathbb{E}_x^f\left\{r(x_0, f)\,\mathbf{1}_{\{x_0\in B^c\}}\sum_{k=1}^{\infty}\mathbf{1}_{\{x_0\in B^c, x_1\in B^c, \dots, x_k\in B^c\}}\prod_{i=0}^{k-1}\alpha(x_i)r(x_k, f)\right\} \\ &= 2\alpha(x)r(x, f)\mathbb{E}_x^f\{V(x_1, f)\,\mathbf{1}_{\{x_1\in B^c\}}\} \\ &= 2\alpha(x)r(x, f)\int_{B^c} V(y, f)Q(\mathrm{d}y \mid x, f), \end{aligned}$$

which together with (3.4) gives

$$I_1 = 2r(x, f)[V(x, f) - r(x, f)] = 2r(x, f)V(x, f) - 2r^2(x, f). \tag{4.6}$$

Let

$$I_2 := \mathbb{E}_x^f\left\{\sum_{k=1}^{\infty}\mathbf{1}_{\{x_0\in B^c, x_1\in B^c, \dots, x_k\in B^c\}}\prod_{i=0}^{k-1}\alpha(x_i)r(x_k, f)\right\}^2 \tag{4.7}$$
$$= \alpha^2(x)\mathbb{E}_x^f\left\{\mathbb{E}_x^f\left[\sum_{k=1}^{\infty}\mathbf{1}_{\{x_1\in B^c, x_2\in B^c, \dots, x_k\in B^c\}}\prod_{i=1}^{k-1}\alpha(x_i)r(x_k, f)\right]^2 \,\middle|\, h_1\right\}$$
$$= \alpha^2(x)\int_{B^c}\sigma^2(y, f)Q(\mathrm{d}y \mid x, f) + \alpha^2(x)\int_{B^c} V^2(y, f)Q(\mathrm{d}y \mid x, f). \tag{4.8}$$

Hence, it follows from (4.1) and (4.5)–(4.8) that

$$\sigma^2(x, f) = r^2(x, f) + I_1 + I_2 - V^2(x, f) = R(x, f) + \alpha^2(x)\int_{B^c}\sigma^2(y, f)Q(\mathrm{d}y \mid x, f).$$

This shows that $\sigma^2(x, f)$ satisfies (4.2). Furthermore, as the proof of the uniqueness of a solution to (3.4), we can prove that the solution to (4.2) is also unique.

(b) Using the conditional expectation and Markov properties, we obtain

$$\mathbb{E}_x^f\left\{\alpha^2(x_0)\int_{B^c} V^2(y, f)Q(\mathrm{d}y \mid x_0, f) + \sum_{n=1}^{\tau_B-1}\prod_{i=0}^{n}\alpha^2(x_i)\int_{B^c} V^2(y, f)Q(\mathrm{d}y \mid x_n, f)\right\}$$
$$= \mathbb{E}_x^f\left\{\alpha^2(x_0)\,\mathbf{1}_{\{x_0\in B^c\}}\int_{B^c} V^2(y, f)Q(\mathrm{d}y \mid x_0, f)\right.$$
$$\left. + \sum_{k=1}^{\infty}\mathbf{1}_{\{x_0\in B^c, x_1\in B^c, \dots, x_k\in B^c\}}\prod_{i=0}^{k}\alpha^2(x_i)\int_{B^c} V^2(y, f)Q(\mathrm{d}y \mid x_k, f)\right\}$$

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

$$= \mathbb{E}_x^f \Bigg\{ \alpha^2(x_0)\, \mathbf{1}_{\{x_0 \in B^c\}}\, \mathbb{E}_x^f [\mathbf{1}_{\{x_1 \in B^c\}}\, V^2(x_1, f)] + \sum_{k=1}^{\infty} \mathbf{1}_{\{x_0 \in B^c, x_1 \in B^c, \ldots, x_k \in B^c\}} \prod_{i=0}^{k} \alpha^2(x_i) \mathbb{E}_x^f [\mathbf{1}_{\{x_{k+1} \in B^c\}}\, V^2(x_{k+1}, f) \mid h_k] \Bigg\}$$

$$= \mathbb{E}_x^f \Bigg\{ V^2(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) V^2(x_n, f) \Bigg\} - V^2(x, f),$$

which together with (4.1) yields (4.4).

To complete the proof of part (b), it remains to prove (4.3). As in the proof of (3.3), we can obtain for any $\pi \in \Pi$ and $x \in B^c$,

$$\mathbb{E}_x^f \{ \mathbf{1}_{\{x_0 \in B^c, x_1 \in B^c, \ldots, x_k \in B^c\}}\, \omega^2(x_k) \} \le \gamma^{2k} \omega^2(x), \qquad k = 0, 1, 2, \ldots.$$

Then, by Lemma 3.1, we have

$$\left| \mathbb{E}_x^f \Bigg\{ \tilde{r}(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) \tilde{r}(x_n, f) \Bigg\} - V^2(x, f) \right|$$

$$\le \beta^2 \left(1 + \frac{2}{1 - \alpha\gamma} \right) \left(1 + \sum_{n=1}^{\infty} (\alpha\gamma)^{2n} \right) \omega^2(x) + \frac{\beta^2}{(1 - \alpha\gamma)^2} \omega^2(x) \le M \omega^2(x),$$

where M is a finite constant. This implies that

$$\mathbb{E}_x^f \Bigg\{ R(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) R(x_n, f) \Bigg\} \in B_{\omega^2}(B^c).$$

Moreover, from Theorem 3.1 it follows that $\mathbb{E}_x^f \{ R(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) R(x_n, f) \}$ is the unique solution in $B_{\omega^2}(B^c)$ to (4.2), which together with part (a) yields (4.3).

Remark 4.1. Theorem 4.1 provides another expression of the variance of a policy $f \in F$, and gives an exact relationship between the discounted expected reward and variance of the policy. If we restrict ourselves to the class of stationary policies and consider $R(x, f)$ defined in (4.1) as a new reward function, then the variance minimisation problem can be transformed into an equivalent first passage optimality problem with $r(x, a)$ and $\alpha(x)$ replaced by $\tilde{r}(x, a) := r(x, a)[2V^*(x) - r(x, a)]$ and $\alpha^2(x)$, respectively.

By Theorem 4.1, the problem P is equivalent to the following problem:

$$\text{P}^*\text{: minimise } V_{\alpha^2}(x, f) \text{ over } f \in F^* \quad \text{for all } x \in B^c,$$

where F^* is from (2.6) and $V_{\alpha^2}(x, f) := \mathbb{E}_x^f \{ \tilde{r}(x_0, f) + \sum_{n=1}^{\tau_B - 1} \prod_{i=0}^{n-1} \alpha^2(x_i) \tilde{r}(x_n, f) \}$.

To solve problem P*, we construct a new model

$$\{X, A, (A^*(x), x \in X), Q(\cdot \mid x, a), B, \alpha^2(x), \tilde{r}(x, a)\}, \tag{4.9}$$

where $A^*(x)$ comes from (3.9). Then, all the results of the first passage optimisation can be used to obtain the similar consequences for the variance minimisation problem.

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

5. Variance-optimal policies

In this section we establish the existence of a variance-optimal stationary policy for the first passage discounted reward criterion by means of the new model (4.9).

Theorem 5.1. *Let V^* be the optimal value function. Under Assumptions 3.1 and 3.2, the following assertions hold:*

(a) *$\sigma_*^2 + (V^*)^2$ is the unique solution in $B_{\omega^2}(X)$ to*

$$u(x) = \inf_{a \in A^*(x)} \Big\{ r(x,a)[2V^*(x) - r(x,a)] + \alpha^2(x) \int_{B^c} u(y) Q(\mathrm{d}y \mid x, a) \Big\} \quad \text{for all } x \in B^c, \tag{5.1}$$

(b) *there exists a stationary policy $f_v^* \in F^*$ such that $f_v^*(x)$ attains the minimum of the right-hand side in (5.1) for every $x \in B^c$, and f_v^* is a variance-optimal policy,*

(c) *a stationary policy $f \in F$ is variance-optimal if and only if $f(x)$ attains the minimum in (5.1) for every $x \in B^c$.*

Proof. For each $x \in B^c$, let $V_{\alpha^2}^*(x) := \inf_{f \in F^*} V_{\alpha^2}(x, f)$. Then, by Theorems 4.1(b) and (2.6), we have $V_{\alpha^2}^*(x) = \sigma_*^2(x) + (V^*(x))^2$. Moreover, the model (4.9) satisfies the hypothesis in Theorem 3.2 with ω and $r(x,a)$ replaced by ω^2 and $-\tilde{r}(x,a)$, respectively. Thus, the results in (a)–(c) follow from Theorem 3.2.

Remark 5.1. Although Theorem 5.1 ensures the existence of a variance-optimal stationary policy (and the main goal of this paper is achieved), a complete calculation procedure of the variance-optimal policy is not given yet. However, when B^c and $A(x)$ are finite, we can use the *policy iteration algorithm* [8], [20] to compute the variance-optimal policies.

6. Example of an application

It is easy to verify that Examples (8.6.2)–(8.6.4) of [10, pp. 68–71] satisfy Assumptions 3.1 and 3.2, and then the existence of the variance-optimal policies can be established. Besides the existence of the variance-optimal policies, in this section we aim to obtain the exact forms of the first passage expected-optimal stationary policies and variance-optimal policies in a cash-balance model (6.1), which is considered in [12] and [23] for the infinite horizon discount MDPs.

Example 6.1. (*A cash-balance model*) Consider the controlled cash-balance system

$$x_{n+1} = x_n + a_n + \varepsilon_n, \qquad n = 0, 1, \ldots, \tag{6.1}$$

where the state x_n and action a_n denote the amount of cash balance, and a withdrawal of size $-a_n$ (if $a_n < 0$) of the money in cash, or a supply in the amount a_n (if $a_n > 0$) at time n, respectively. Suppose that the state of the system is $x \in X := (-\infty, +\infty)$. A decision-maker takes an action a in a given set $A(x)$, which is assumed to be $[-|x|, |x|]$. The disturbances $\{\varepsilon_n, n \geq 0\}$ are assumed to be independent standard normal random variables, that is, the transition law $Q(\cdot \mid x, a)$ is given by

$$Q(C \mid x, a) = \int_C \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - x - a)^2}{2} \right) \mathrm{d}y \quad \text{for all } (x,a) \in K,\ C \in \mathcal{B}(X).$$

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

Finally, the function $r(x, a)$ is assumed to be u.s.c. in $a \in A(x)$ for each state x in X.

Proposition 6.1. *In Example 6.1, the following assertions hold.*

(a) *If the discount factors $\alpha(x)$ satisfy $\sup_{x\in B^c} \alpha(x) < \frac{1}{4}$, and there exists a constant $M > 0$ such that $|r(x, a)| \le M(x^2 + 1)$ for all $x \in X$ and $a \in [-|x|, |x|]$, then a variance-optimal policy exists (by Theorem 5.1), and it is also first passage expected-optimal.*

(b) *(The special cases of varying discount factors.) Suppose that the discount factors $\alpha(x)$ and reward function $r(x, a)$ are given as follows:*

$$\alpha(x) := \delta e^{-x^2/4} \quad \text{and} \quad r(x, a) := x^2 - 1 - \delta e^{-x^2/4}[(a^2 - x^2)^2 + (x + a)^2]$$

for each $x \in B^c$ and $a \in A(x)$, where δ is a constant in $(0, \frac{1}{4})$. Then, we have

(bi) *(On the infinite horizon discounted criterion.) Take the target set $B := \varnothing$, then the first passage expected-optimal policies are given by $f_1^*(x) := -x$ and $f_2^*(x) := x$ for every $x \in X$, and the optimal value function is given by $V^*(x) = x^2 - 1$ for all $x \in X$. Moreover, the policy $f_1^*(x) = -x (x \in X)$ is a variance-optimal policy.*

(bii) *(On the first passage criterion.) Let $B^c := [-m, m]$, where m is some positive constant (such as m represents the boundary which makes the system cease to be effective). Then, the first passage expected-optimal policy is only given by $f^*(x) := -x$, and the optimal reward function is given by*

$$V^*(x) = x^2 - 1 - \frac{2\delta m e^{-m^2/2}}{\sqrt{2\pi} - (2\sqrt{3\pi}\delta/3)[2\phi((\sqrt{6}/2)m) - 1]} e^{-x^2/4}, \quad x \in X,$$

where $\phi(x) := \int_{-\infty}^{x} (1/\sqrt{2\pi}) e^{-t^2/2}\, dt$ (the standard normal distribution function). Of course, the policy $f^*(x) = -x$ is also the variance-optimal policy.

(c) *(The special cases of a fixed discount factor.) Suppose that the discount factors $\alpha(x) :\equiv \alpha \in (0, \frac{1}{4})$ and reward function $r(x, a)$ is given by*

$$r(x, a) := x^2 - 1 - \alpha[(a^2 - x^2)^2 + (x + a)^2] \quad \text{for all } x \in B^c,\ a \in A(x).$$

Then, we have the following results.

(ci) *(On the infinite horizon discounted criterion.) Take the target set $B := \varnothing$, then the first passage expected-optimal policies are given by $f_1^*(x) := -x$ and $f_2^*(x) := x$ for every $x \in X$, and the optimal value function is given by $V^*(x) = x^2 - 1$.*

(cii) *(On the first passage criterion.) Let $B^c := [-m, m]$ with some constant $m > 0$, then the first passage expected-optimal policy is given only by $f^*(x) = -x$, and the optimal value function is as follows*

$$V^*(x) = x^2 - 1 - \frac{2\alpha m e^{-m^2/2}}{\sqrt{2\pi}[1 - \alpha(2\phi(m) - 1)]}.$$

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

Proof. (a) Let $\omega(x) := x^2 + 1$ for each $x \in B^c$, then it follows that

$$\int_{B^c} \omega(y) Q(\mathrm{d}y \mid x, a) \le \int_{-\infty}^{+\infty} \omega(y) Q(\mathrm{d}y \mid x, a) = (x+a)^2 + 2 < 4\omega(x),$$

$$\begin{aligned}\int_{B^c} \omega^2(y) Q(\mathrm{d}y \mid x, a) &\le \int_{-\infty}^{+\infty} \omega^2(y) Q(\mathrm{d}y \mid x, a) \\ &= (x+a)^4 + 8(x+a)^2 + 6 \\ &< 16\omega^2(x),\end{aligned}$$

$$|r(x,a)| \le M\omega(x)$$

for all $x \in X, a \in A(x)$. Since Assumption 3.2(c) has been verified from Proposition 4.1 of [23], it follows from the three inequalities (just established) that Assumptions 3.1 and 3.2 are satisfied. Thus, Theorem 5.1 ensures the existence of a variance-optimal stationary policy.

(bi) Since

$$|r(x,a)| \le x^2 + 1 + \frac{1}{4}\frac{(a^2 - x^2)^2 + (x+a)^2}{1 + (x^2/4)} < 10\omega(x)$$

for all $x \in X$ and $a \in A(x)$, by Theorem 3.2, the optimality equation (3.5) becomes

$$\begin{aligned}V^*(x) = \sup_{a \in A(x)} \Big\{ & x^2 - 1 - \delta \mathrm{e}^{-x^2/4}[(a^2 - x^2)^2 + (x+a)^2] \\ & + \delta \mathrm{e}^{-x^2/4} \int_{-\infty}^{+\infty} V^*(y) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(y-x-a)^2/2} \, \mathrm{d}y \Big\}. \end{aligned} \tag{6.2}$$

Now, suppose that $V^*(x) = k_1 x^2 + k_2 x + k_3$, then, we have

$$\begin{aligned}\int_{-\infty}^{+\infty} V^*(y) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(y-x-a)^2/2} \, \mathrm{d}y &= \int_{-\infty}^{+\infty} (k_1 y^2 + k_2 y + k_3) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(y-x-a)^2/2} \, \mathrm{d}y \\ &= k_1[(x+a)^2 + 1] + k_2(x+a) + k_3,\end{aligned}$$

which together with (6.2) yields $k_1 = 1$, $k_2 = 0$, $k_3 = -1$ (i.e. $V^*(x) = x^2 - 1$), and the maximum in (6.2) is attained at $f_1^*(x) := -x$ and $f_2^*(x) := x$ for each $x \in X$. Because of the uniqueness of the solution in $B_\omega(X)$ to (6.2), by Theorem 3.2 the policies f_1^* and f_2^* are first passage expected-optimal and the optimal reward function is given by $V^*(x) = x^2 - 1$.

In order to distinguish the variance-optimal policy between f_1^* and f_2^*, we need only to find the policy which makes $V_{\alpha^2}(x, f)$ attain the optimal value $V_{\alpha^2}^*(x)$ in problem P^*. For each $x \in X$, let $A^*(x) := \{f_1^*(x), f_2^*(x)\}$. Then, we have

$$r(x,a)[2V^*(x) - r(x,a)] = (x^2 - 1)^2 - \delta^2 \mathrm{e}^{-x^2/2}(x+a)^4 \quad \text{for all } a \in A^*(x).$$

Thus, by Theorem 5.1 for each $x \in X$,

$$\begin{aligned}V_{\alpha^2}^*(x) = \inf_{a \in A^*(x)} \Big\{ & (x^2 - 1)^2 - \delta^2 \mathrm{e}^{-x^2/2}(x+a)^4 \\ & + \delta^2 \mathrm{e}^{-x^2/2} \int_{-\infty}^{+\infty} V_{\alpha^2}^*(y) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(y-x-a)^2/2} \, \mathrm{d}y \Big\}. \end{aligned} \tag{6.3}$$

https://doi.org/10.1239/jap/1437658608 Published online by Cambridge University Press

Furthermore, we suppose that $V^*_{\alpha^2}(x) := (x^2-1)^2 + k_4 e^{-x^2/2}$, then

$$\begin{aligned}\int_{-\infty}^{+\infty} V^*_{\alpha^2}(y)\frac{1}{\sqrt{2\pi}}e^{-(y-x-a)^2/2}\,dy &= \int_{-\infty}^{+\infty}\frac{y^4-2y^2}{\sqrt{2\pi}}e^{-(y-x-a)^2/2}\,dy + 1 + k_4 e^{-(x+a)^2/4}\int_{-\infty}^{+\infty} e^{-(y-(x+a)/2)^2}\,dy \\ &= (x+a)^4 + 6(x+a)^2 + 3 - 2[(x+a)^2+1] + 1 + \frac{k_4}{\sqrt{2}}e^{-(x+a)^2/4}. \qquad (6.4)\end{aligned}$$

Note that the function $h(y) := 4y^2 + (k_4/\sqrt{2})e^{-y^2/4}$ attains the minimum at y=0 as $k_4 \le 16\sqrt{2}$, which together with (6.4) implies that the minimum in (6.3) is attained at $f_1^*(x) = -x$, and yields $k_4 = 2\sqrt{2}\delta^2/(\sqrt{2}-\delta^2)$ $(\le 16\sqrt{2})$, that is,

$$V^*_{\alpha^2}(x) = (x^2-1)^2 + \frac{2\sqrt{2}\delta^2}{\sqrt{2}-\delta^2}e^{-x^2/2} \quad \text{for all } x \in X.$$

Because of the uniqueness of the solution in $B_{\omega^2}(X)$ to (6.3), it follows from Theorem 5.1 that the policy $f_1^*(x) = -x (x \in X)$ is a variance-optimal policy.

(bii) For the case of $B^c = [-m, m]$, the optimality equation (6.2) becomes

$$\begin{aligned}V^*(x) = \sup_{a\in A(x)}\Big\{&x^2 - 1 - \delta e^{-x^2/4}[(a^2-x^2)^2 + (x+a)^2] \\ &+ \delta e^{-x^2/4}\int_{-m}^{m} V^*(y)\frac{1}{\sqrt{2\pi}}e^{-(y-x-a)^2/2}\,dy\Big\}. \qquad (6.5)\end{aligned}$$

Using the arguments in (bi), by a straightforward calculation we can verify that the maximum in (6.5) is only attained at $f^*(x) := -x$, and

$$V^*(x) := x^2 - 1 - \frac{2\delta m e^{-m^2/2}}{\sqrt{2\pi} - (2\sqrt{3\pi}\delta/3)[2\phi((\sqrt{6}/2)m) - 1]}e^{-x^2/4},$$

where $\phi(x)$ is the standard normal distribution function. Because of the uniqueness of the solution in $B_\omega(B^c)$ to the equation (6.5), by Theorem 3.2 the policy $f^*(x) = -x$ is first passage expected-optimal. Evidently, the unique first passage expected-optimal policy $f^*(x) = -x$ is also the variance-optimal policy.

(ci) As the same calculations of Proposition 6.1 (bi), we see that (ci) is true.

(cii) For the case of $B^c = [-m, m]$, it follows that the first passage expected-optimal policy is given only by $f^*(x) = -x$, and the optimal reward function is given by

$$V^*(x) = x^2 - 1 - \frac{2\alpha m e^{-m^2/2}}{\sqrt{2\pi}[1 - \alpha(2\phi(m) - 1)]},$$

and so (cii) follows.

Remark 6.1. Proposition 6.1(bi) shows that there are indeed more than one first passage expected-optimal policies for the case of varying discount factors, and it presents an *exact* variance-optimal policy. Moreover, the difference of the optimal value functions in Proposition 6.1(b) further shows the difference between the first passage criterion here and the infinite

horizon criterion for the discrete-time MDPs with varying discount factors [23]. Furthermore, the difference of the optimal value functions in Proposition 6.1(c) implies that in general the first passage criterion is also *different* from the (well known) infinite horizon expected discounted criterion for discrete-time MDPs with a constant discount factor given in [9], [10], and [20].

Acknowledgements

The authors thank the Editor and the anonymous referees for valuable comments and suggestions that have helped to improve this paper. The research of the authors was supported by National Natural Science Foundation of China (NSFC) and Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2011).

References

[1] Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications in Finance.* Springer, Heidelberg.
[2] Derman, C. (1970). *Finite State Markovian Decision Processes* (Math. Sci. Eng. **67**). Academic Press, New York.
[3] Feinberg, E. A. and Shwartz, A. (1994). Markov decision models with weighted discounted criteria. *Math. Operat. Res.* **19,** 152–168.
[4] González-Hernández, J., López-Martínez, R. R. and Minjárez-Sosa, J. A. (2008). Adaptive policies for stochastic systems under a randomized discounted cost criterion. *Bol. Soc. Mat. Mexicana (3)* **14,** 149–163.
[5] Guo, X. and Hernández-Lerma, O. (2009). *Continuous-Time Markov Decision Processes.* Springer, Berlin.
[6] Guo, X. and Song, X. (2009). Mean-variance criteria for finite continuous-time Markov decision processes. *IEEE Trans. Automatic Control* **54,** 2151–2157.
[7] Guo, X., Hernández-del-Valle, A. and Hernández-Lerma, O. (2012). First passage problems for nonstationary discrete-time stochastic control systems. *Europ. J. Control* **18,** 528–538.
[8] Guo, X., Ye, L. and Yin, G. (2012). A mean-variance optimization problem for discounted Markov decision processes. *Europ. J. Operat. Res.* **220,** 423–429.
[9] Hernández-Lerma, O. and Lasserre, J. B. (1996). *Discrete-Time Markov Control Processes.* Springer, New York.
[10] Hernández-Lerma, O. and Lasserre, J. B. (1999). *Further Topics on Discrete-Time Markov Control Processes.* Springer, New York.
[11] Hernández-Lerma, O., Vega-Amaya, O. and Carrasco, G. (1999). Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM J. Control Optimization* **38,** 79–93.
[12] Hordijk, A. and Yushkevich, A. A. (1999). Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards. *Math. Meth. Operat. Res.* **50,** 421–448.
[13] Huang, Y. and Guo, X. (2009). Optimal risk probability for first passage models in semi- Markov decision processes. *J. Math. Anal. Appl.* **359,** 404–420.
[14] Huang, Y.-H. and Guo, X.-P. (2011). First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs. *Acta. Math. Appl. Sinica (English Ser.)* **27,** 177–190.
[15] Kurano, M. (1987). Markov decision processes with a minimum-variance criterion. *J. Math. Anal. Appl.* **123,** 572–583.
[16] Liu, J. and Huang, S. (2001). Markov decision processes with distribution function criterion of first-passage time. *Appl. Math. Optimization* **43,** 187–201.
[17] Liu, J. Y. and Liu, K. (1992). Markov decision programming—the first passage model with denumerable state space. *Systems Sci. Math. Sci.* **5,** 340–351.
[18] Mamabolo, R. M. and Beichelt, F. E. (2004). Maintenance policies with minimal repair. *Econ. Qual. Control* **19,** 143–166.
[19] Prieto-Rumeau, T. and Hernández-Lerma, O. (2009). Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. *Math. Meth. Operat. Res.* **70,** 527–540.
[20] Puterman, M. L. (1994). *Markov Decision Processes.* John Wiley, New York.
[21] Schäl, M. (2005). Control of ruin probabilities by discrete-time investments. *Math. Meth. Operat. Res.* **62,** 141–158.
[22] Sobel, M. J. (1982). The variance of discounted Markov decision processes. *J. Appl. Prob.* **19,** 794–802.
[23] Wei, Q. and Guo, X. (2011). Markov decision processes with state-dependent discount factors and unbounded rewards/costs. *Operat. Res. Lett.* **39,** 369–374.
[24] Yu, S. X., Lin, Y. and Yan, P. (1998). Optimization models for the first arrival target distribution function in discrete time. *J. Math. Analysis Appl.* **225,** 193–223.