Stony Brook University

On Variance Minimization for Constrained Discounted Continuous-Time MDPs

A Dissertation Presented

by

Jun Fei

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2009

Stony Brook University

The Graduate School

Jun Fei

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Eugene Feinberg – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Joseph Mitchel – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Jiaqiao Hu
Assistant Professor, Department of Applied Mathematics and Statistics

Petar M. Djurić
Professor
Department of Electric and Computer Engineering

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

On Variance Minimization for Constrained Discounted Continuous-Time MDPs

by

Jun Fei

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2009

We discuss the minimization of variances of the total discounted rewards for constrained continuous-time Markov Decision Processes (MDPs) with countable state spaces and its application in dynamic power management for portable electronic devices. The rewards consist of cumulative rewards earned between jumps and instant rewards earned at jump epochs. According to the existing theory, for the expected total discounted rewards optimal policies exist in the forms of randomized stationary and switching stationary strategies. While the former is typically unique, the latter forms a finite set whose number of elements grows exponentially with the number of constraints.

There are two natural definitions of total discounted rewards: (i)by interpreting discounting as a coefficient in front of the future reward rates (multiplicative discount), and (ii) by interpreting discounting as stopping times (probabilistic discounting). We show through conditional variance that the variance under the multi-

plicative discounting is less than or equal to the variance under the probabilistic discounting. For the second interpretation of discounting and for rewards up to the first jump we provide an index for selection of actions by switching stationary strategies and show that an index policy achieves the smaller variance than the randomized stationary policy. In particular, for problems with zero instant rewards, the index policy achieves the minimum variance of rewards up to the first jump among all the equivalent switching strategies. For rewards beyond the first jump, we provide an example for which the index strategy is not the best among switching stationary strategies. We also give an example that under the multiplicative discounting the best switching strategy may not outperform the randomized policy even for problems with rewards up to the first jump.

We also discuss an application of the results to dynamic power management for portable electronic devices. We propose an optimal switching strategy that has two advantages over the "best" nonrandomized stationary policy suggested in the power management literature. First, our approach yields better performance in saving energy consuption. Second, while computing the "best" nonrandomized policy is NP-hard finding the optimal switching strategy is a P-hard problem.

Contents

Acknowledgements

I would like to thank Eugene Feinberg, my supervisor, for his many suggestions and constant support during this research. I would also like to thank my wife and my parents for their constant support during my study at Stony Brook.

Stony Brook, NY 11794 Jun Fei
October, 2009

Chapter 1

Introduction

Since the introduction by Markowitz in his Nobel-Prize winning paper [39], variance has played an important role in stochastic optimization. With the risk of a portfolio measured by the variance of its return Markowitz showed how to formulate the problem of minimizing a portfolio's variance subject to the constraint that its expected return equals a prescribed level as a quadratic programming problem. Such an optimal portfolio is said to be variance minimizing and if it also achieves the maximum expected return among all portfolios having the same variance of return then it is said to be efficient. The set of all points in the two dimensional plane of variance or standard deviation and expected return that are produced by efficient portfolios is called the efficient frontier.

Owing to its practical importance, especially in portfolio management in finance, the mean-variance problem has drawn continuing attention; see for example, [2, 8, 9, 19, 21, 22, 35, 42, 44, 51, 58, 61, 62] among others.

In the area of Markov Decision Processes the usual optimization criteria have been the expected total rewards over a finite horizon, the expected discount total rewards over a finite or infinite horizon, or the limiting average reward per unit time over an infinite horizon as long as the limit exists. These standard criteria may be quite insufficient to fully capture the various aspects considered by the decision maker. The biggest weakness lies in the fact that they fail to take into account the risk - an important characteristic decision makers may weigh more than the expectation. The risk can be characterized by probabilities of upper or lower tails, e.g., minimizing the ruin probability in insurance problem, maximizing the one percentile of profits or minimizing the 99 percentile of costs. There have been some papers devoted to the probability criteria for various rewards. Filar [16] and Filar et al. [17] studied the percentile performance criteria for the limiting average return. Zhang et al. [60] and White [56] considered the threshold probability criteria for

discounted MDPs and focused on the properties of the optimality equations without discussing the existence and properties of the optimal policies.

More often the risk is characterized by the variance or standard deviation. As far as we know variance of MDPs has been studied primarily for discrete time models; see the review paper by White [55] and references therein.

The study of variances of MDPs was initiated by Mandl [37, 38] where the asymptotic behavior of the variance of the sum of costs under the policy which minimizes the mean cost per unit time was first investigated. Jaquette [26] studied the variance of finite state and action discrete time MDPs with small discount rate and the asymptotic behavior as the discount rate approaches zero. Jaquette [27] further studied the moment optimality of finite state and action discrete time MDPs by examining the negative of the Laplace transform of the total return random variable. The results were extended to the continuous time case in Jaquette [28] where the author showed that a moment optimal policy can always be found among the class of piecewise constant policies. In Jaquette [29] the author formulated an explicit utility function to combine both mean and variance into one function and developed the so-called utility optimality with constant aversion to risk. It was shown that optimal policies exist but are not necessarily stationary for an infinite horizon stationary MDP with finite state and action spaces. Benito [5] presented formulas to calculate the expected value and the variance of the reward in discrete time MDPs. Sobel [48] derived formulas for the variance and higher moments of the present value of single-stage rewards for a finite MDP and for a semi-Markov Decision Process. Sobel [49] and Chung [6] studied the problem by maximizing the mean/standard deviation ratio in an undiscounted setting. Kawai [31] tries to find an optimal randomized policy that minimizes the variance of the reward among the policies that give the mean not less than a specified value by introducing a parametric Markov Decision Process with an average cost criterion. It was shown that there exists an optimal policy which is a mixture of at most two pure policies.

Filar et al. [18] consider variance-penalized models. Kadota [30] studied the average variance of Markov Decision Processes with countable states and finite actions and developed sufficient conditions to assure that there is a stationary deterministic policy that minimizes the average-variance in a class of mean-optimal policies. Altman and Shwartz [1] and Baykal-Gursoy and Ross [3] provided further results on variance-penalized models. In particular Baykal-Gursoy and Ross [3] introduced two definitions of variability, namely, the expected time-average variability and time-average expected variability. Sobel [50] and Puterman [45] considered mean-variance tradeoffs and considered the problem of finding Pareto-optimal policies in the sense of having high

means and low variances. White [57] investigated some algorithms to solve the mean-variance problem. Chung [7] discussed the computation of a Pareto-optimal policy in the sense of having high means and low variances of the stationary distribution under the unichain condition. Liu and Zhao [36] investigated the average reward semi-Markov Decision Processes with a general multichain structure using a data-transformation method.

For continuous-time Markov Decision Processes, the research is relative less. In addition to Jaquette [28], Van Dijk and Sladky [54] investigated the variance of the undiscounted total rewards for continuous-time Markov chains with finite state spaces and showed that the variance growth rate is asymptotically linear in time. Baykal-Gürsoy and Gürsoy [4] defined the expected time-average variability criterion for communicating Semi-Markov Decision Processes and showed that under certain assumptions an $\epsilon-$optimal (randomized) stationary policy exists for this criterion when the state space is finite and the action sets are compact.

As an important application of Markov Decision Processes in engineering, Dynamic Power Management (DPM) of portable electronic systems, such as laptops, PDAs, and cellular phones, has drawn increasing attention. The goal of dynamic power management is to extend the battery life while meeting the performance requirements. The term "dynamic" is contrasted to "static" where a constant power scheme is used regardless of the variation in requests for services. High power consumption not only reduces the battery life, but also results in increased packaging and cooling costs as well as potential reliability problems.

The problem of finding a power management scheme (or policy) that minimizes power dissipation under performance constraints is of great interest to system designers. A simple power management system includes four components: Service Provider (SP), Service Requestor (SR), Service Queue (SQ), and Power Manager (PM). The SR generates service requests. The SQ buffers the service requests. The SP provides service in a top-down manner.

The working modes of the SP can be modelled into three states: "busy", "idle" and "sleep". In busy states, the SP is fully powered and fully operational. For convenience of modelling it is assumed that the SP cannot switch to any other state when it has working on some request. The transition from busy state to other states only occurs when the SP finishes one service request. For each busy state, there exists a corresponding idle state. In the idle states, the SP is fully powered, but it is not working on any request. An idle state is the only state that connects to its corresponding busy state. When the SP finishes a service, it will automatically switch from the busy state to its corresponding idle state. Conversely, when the SP wants to switch from some

other state to a busy state, it first switches to the corresponding idle state and then goes to the busy state. When the SP is in an idle state, it is in the same power mode as when it is in the corresponding busy state. We remark that the idle states are not physical states of the SP but merely for the convenience of modelling. In power-down ("sleep") states the SP is partially or completely shut down and is not operational.

Throughout the time the PM monitors the states of the SR, SQ, and SP and issues state-transition commands to the SP. A simple and well-known heuristic policy is the "time-out" policy, which shuts down the SP after it has been idle for a certain amount of time. This "time-out" policy is widely used in today's portable computers. A predictive system shutdown approach was proposed in [53] and [24]. The approach tries to predict the "on" and "off" time of each component using a regression model based on the component's previous "on" and "off" times such that the Service Provider (SP) can be turned on just before the request comes. This method works best for the special cases where the requests are highly correlated and highly predictable.

A power management approach based on a Markov decision process has been proposed in [24] where the system is modeled as a discrete-time Markov decision process by combining the stochastic models of each component. Once the model and its parameters are determined, an optimal power management policy for achieving the best power-delay trade-off in the system is generated. This approach offers significant improvements over previous power management techniques in terms of its theoretical framework for modeling and optimizing the system. There are, however, some shortcomings; see Qiu et al. [46]. In [46] power management problem was formulated as a constrained continuous time Markov Decision Process. A linear programming approach that minimizes the average cost was proposed, and optimal randomized policies are found by solving the LP. Since in practice it is hard to implement the randomized policy they tried to search for the "best" nonrandomized stationary policy using either nonlinear programming approach or a heuristic policy iteration. Finding such a policy is an NP-hard problem; see Feinberg [11]. In addition, it typically has worse performance than the optimal randomized policy or may not exist even for some feasible problems [12].

In this dissertation we study constrained Continuous-Time Markov Decision Processes (MDPs) with countable state spaces. According to the existing theory, see Feinberg [13], for a feasible problem with compact action sets, optimal policies for the expected total discounted rewards can be found in several forms. The most natural forms are randomized stationary policies and switching stationary strategies. Randomized stationary policies select actions at jump epochs and the selected action is used until the next jump. Switching

stationary strategies do not utilize randomization procedure and may change actions between jumps. For each randomized stationary policy, the finite set of equivalent switching stationary strategies can be defined. These policies are equivalent in the sense that the total discounted expected rewards for them are equal for any reward function. Once an order of actions used between jumps at each state is fixed, a particular equivalent switching stationary strategy can be computed.

If the number of constraints is K, the optimal policies exist in the forms of K-randomized stationary policies and K-switching stationary strategies. The former means that the total number of randomization procedures at all states is bounded by K, while the latter means that the total number of switching epoches between jumps at all states is limited by K. If a randomized policy uses m actions at one state, different permutations of these actions define $m!$ switching stationary strategies. According to Feinberg [13, Theorem 4.5], their expected total discounted rewards are equal. Thus, in terms of the expected total rewards these two classes of policies are equivalent. However, their variances may be different. In this paper we are interested in finding a policy that has the smallest variance of total discounted rewards among the equivalent randomized stationary and switching stationary policies.

There are two natural definitions of total discounted rewards: (i) by interpreting discounting as a coefficient in front of the future reward rates (multiplicative discount), and (ii) by interpreting discounting as stopping times (probabilistic discounting). Each of these interpretations defines the total discounted reward as a random variable. The two random variables are different, but they have the same expectations. However, their second moments are different and hence their variances are different too.

Our first research topic is naturally to see whether there is any relation (equality or inequality) between the two definitions of variances. We show through conditional variance that the variance under the multiplicative discounting is less than or equal to the variance under the probabilistic discounting. The results are accepted for publication by Journal of Applied Probability and will appear in December 2009 [14].

Both K-randomized stationary policies and K-switching stationary strategies use no more than $K+1$ actions at any state. Therefore, it is sufficient to limit action sets to the finite subsets. In this paper we define an index such that an index K-switching stationary strategy achieves a smaller variance of total discounted rewards than the equivalent K-randomized stationary policy. In particular, if there are no instant rewards, we show that the index K-switching stationary strategy has the smallest variance among all the K-switching stationary strategies and the K-randomized stationary policy.

In the end we examined the application of switching strategies to the Dynamic Power Management problem studied in [46]. Instead of finding the "best" nonrandomized stationary policies as suggested by [46] we compute the best switching strategies by fixing the order of actions available at each state. We will further compare the performance of our proposed approach and the suggested approach in [46].

Chapter 2

Model Definition and Known Results

2.1 Model definition

A continuous-time MDP is defined by a set $\{S, A, A(\cdot), q, K, r_k, R_k,\ k = 0, \ldots, K\}$, where

$S = \{1, 2, 3, \ldots, \}$ is a countable state space;

A is an action space, which is assumed to be a complete separable metric space, and $A(i)$, $\in S$, are action sets available at states $i \in S$, which are assumed to be compact subsets of A;

$q(i, j, a)$ is the transition rate from state i to state j, $i \neq j$, when action a is selected. Let $q(i, a) := -q(i, i, a) = \sum_{j=1, j\neq i}^{n} q(i, j, a) \leq C$ for some $0 < C < \infty$;

$K = \{0, 1, \ldots\}$ is the number of constraints;

$r_k(i, a)$ is the reward rate for criterion $k, k = 0, 1, \ldots, K$, at state i when action a is selected, $i \in S, a \in A(i)$;

$R_k(i, j, a)$ is the instant reward for criterion $k, k = 0, 1, \ldots, K$, earned when transiting from state i to state j and action a is selected. The functions $r_k(i, a)$ and $R_k(i, j, a)$ are assumed to be upper semi-continuous in a and uniformly bounded above, $k = 0, 1, \ldots, K$.

The definitions of strategies and explanations on how strategies define the corresponding multivariate point processes $\{T_n, X_n : n \geq 0\}$, where T_n is the nth jump epoch, and X_n is the state after nth jump, $0 = T_0 \leq T_1 \leq \ldots \leq T_n \leq \ldots$, $X_n \in S$, is given in Appendix A. Here we repeat the main definitions, give definitions of some important subclasses of strategies including K-switching stationary strategies and K-randomized stationary policies, and

introduce the objective criteria. Since T_0 is always 0, the history up to time t is $i_0, t_1, x_1, \ldots, t_n, i_n, t$, where n is the number of jumps up to time t, i_0 is the observed initial state, and t_i and i_i are ith jump epochs and states immediately after them, $i = 1, \ldots, n$. We write t_i and i_i instead of T_i and X_i to indicate that they are given observed values rather than random values. Let $\Omega^* = \cup_{n\geq 0}(S \times [0,\infty))^n$ and let $\mathcal{F}^*$ be the Borel σ-field on Ω^* induced by the Borel σ-field on $[0,\infty)$. Consider the set of all finite histories $\Omega = \{(i_0, t_1, i_1, t_2 \ldots, i_n, t) : 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq t < \infty,\ n = 0, 1, \ldots\}$. Then Ω is a measurable subset of Ω^*. We denote by $\mathcal{F} = \{B \in \mathcal{F}^* | \, B \subset \Omega\}$ the restriction of $\mathcal{F}^*$ to Ω.

A (nonrandomized) strategy ψ is a Borel mapping from Ω to A such that $\psi(i_0, t_1, \ldots,\ i_{n-1}, t_n, i_n, t) \in A(i_n)$ for each $(i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) \in \Omega$. For a strategy ψ and given the history $\omega_n = i_0, t_1 \ldots t_n, i_n, t$, the joint probability distribution that the jump happens during the interval $[t, t + dt)$ and $i_{n+1} = j$ is $q(i, j, \psi(i_0, t_1, \ldots, t_n, i_n, t))dt$. Each probability distribution μ of the initial state X_0 and each randomized strategy π define the unique multivariate point process on the probability space $(\Omega_\infty, \mathcal{F}_\infty)$, where $\Omega_\infty = (S \times [0,\infty])^\infty$ and $\mathcal{F}_\infty$ is the Borel σ-field on Ω_∞ induced by the Borel σ-fields on $[0,\infty]$. We denote by P_μ^ψ and E_μ^ψ the probabilities and expectations for this process. We also write P_i^ψ and E_i^ψ instead of P_μ^ψ and P_μ^ψ when $\mu(i) = 1$ for some $i \in S$.; see Remark in the appendix for details.

The condition $q(i, a) \leq C < \infty$ implies that $P_\mu^\psi(T_\infty = \infty) = 1$; see Ross [47]. For a multivariate point process $\{T_n, X_n : n \geq 0\}$ defined by a strategy ψ and by some initial distribution, we define $X(t) = X_n$ and $a(t) = \psi(X_0, T_1, \ldots, X_n, t)$ for $T_n \leq t < T_{n+1}$. Let $a_n = \psi(X_0, T_1, \ldots, X_n, T_{n+1})$, if $T_{n+1} < \infty$, and a_n be an arbitrary element from $A(X_n)$, if $T_{n+1} = \infty$, where $n = 0, 1, \ldots$. The values of a_n define transition probabilities from X_n to X_{n+1} at jump epoches.

Let $V_k(i, \psi)$ be the expected total discounted rewards for criterion $k, k = 0, 1, \ldots, K$, when $i \in S$ is the initial state and the strategy ψ is selected,

$$V_k(i,\psi) = E_i^\psi\left[\sum_{n=0}^{\infty} e^{-\alpha T_n} R_k(X_n, X_{n+1}, a_n) + \int_0^\infty e^{-\alpha t} r_k(X(t), a(t))\, dt\right], \tag{2.1.1}$$

where $\alpha > 0$ is the discount rate.

It is also possible to consider randomized strategies; see Appendix A. These strategies can be defined as (nonrandomized) strategies when actions are replaced with probability distributions on the action set A. Then the action sets A and $A(i)$ are replaced with the sets $\tilde{A}$ and $\tilde{A}(i)$ of probability measures on A and $A(i)$ respectively. In addition, the transition intensities q and reward

functions r_k and R_k, $k = 1, \ldots, K$, are replaced respectively with $\tilde{q}$, $\tilde{r}_k$, and $\tilde{R}_k$ defined in (A.0.1), (A.0.2) and (A.0.3) in the appedix. Then the expected total discounted rewards $V_k(i, \pi)$ for the new model are the expected total discounted rewards for randomized policies in the original model.

Let Π be the set of strategies and $R\Pi$ be the set of randomized strategies. Then $\Pi \subseteq R\Pi$.

Now consider a situation when a decision maker can make decisions only at jump epochs and the selected actions remain unchanged until the next jump. We also consider the simplest situation when the decisions depend only on current states. We shall call such decision rules *randomized stationary policies*.

A randomized stationary policy is defined by probabilities $\sigma(da|i)$ on A such that $\sigma(A(i)|i) = 1$. In each state i_n the decision a_n is selected according to the probability distribution $\sigma(da|i_n)$ and this action controls the process until the next jump. Ionescu Tulchea theorem in Jacod [25] implies that any initial distribution μ and any randomized stationary policy σ define a unique multivariate point process on the probability space $(\Omega_\infty, \mathcal{F}_\infty)$.

At first glance, the definition of a randomized stationary policy is not relevant to the definition of a randomized strategy. In fact, any randomized stationary policy can be represented as a randomized strategy. Indeed, for a randomized stationary policy σ, the distribution of the sojourn time until the next jump is the mixture of exponential distributions with intensities $q(i, a)$ and taken with probabilities $\pi(da|i)$. Let $\theta_n = T_{n+1} - T_n$ be the time spent until the next jump. Then

$$P\{t < \theta_n \leq t + dt, i_{n+1} = j | S_n > t, i_n = i\} = \frac{\int_A e^{-q(i,a)t} q(i, j, a)\sigma(da|i)}{\int_{A(i)} e^{-q(i,a)t}\sigma(da|i)} dt.$$

It is easy to see that any randomized stationary policy σ can be presented as a randomized strategy π for which the selection of actions depend only on the current state i and time t passed since the last jump. This representation is defined by the formula

$$\pi(B|i, t) = \frac{\int_B e^{-q(i,a)t}\sigma(da|i)}{\int_{A(i)} e^{-q(i,a)t}\sigma(da|i)}$$

for each measurable sunset B of $A(i)$. Thus, any randomized stationary policy can be defined as a randomized strategy.

For a given initial state i and for given constants $C_k, k = 1, \ldots, K$, consider

the following problem:

$$\begin{array}{l} \text{Maximize } V_0(i,\pi), \\ \text{s.t. } \pi \in R\Pi \\ V_k(i,\pi) \geq C_k, k = 1,\ldots,K. \end{array} \tag{2.1.2}$$

A randomized stationary policy is called K-randomized stationary, $K = 0, 1, \ldots,$ if the number of additional actions used by randomization procedures is limited by K. This means that for each $i \in S$ there exists a finite subset $A^\pi(i)$ of $A(i)$ such that $\pi(A^\pi(i)|i) = 1$ for all $i \in S$ and $\sum_{i\in S}\sum_{a\in A^\pi(i)}[I\{\pi(a|i) > 0\} - 1] \leq K$.

A (nonrandomized) strategy is called switching stationary if the action selection depends only on the current state, and time passed since the last jump. Thus, a switching stationary strategy ϕ is defined as a function $\phi(i,t)$ such that $i \in S, t \in [0,\infty)$, and function $\phi(i,t)$ is measurable in t.

Similar to K-randomized stationary policies we consider K-switching stationary strategies in which the number of switching points is not greater than K. This means that function $\phi(i,t)$ is discontinuous in t at most at K points (i,t).

2.2 Existing results

Theorem 6.1 and A.9 in Feinberg [13] provide various structures of optimal policies for problem (2.1.2). In particular, Theorem 6.1(ii) and Theorem A.9(ii) imply the following statement.

Theorem 2.2.1 *If problem (2.1.2) is feasible, then (i) there exists an optimal K-randomized stationary policy, and (ii) there exists an optimal K-switching stationary strategy.*

In addition the results of sections 4 and 5 in Feinberg (2004) explain how to link optimal K-randomized policies and K-switching stationary strategies. Let σ be an arbitrary K-randomized stationary policy. For example, σ can be an optimal K-randomized policy.

Let $A^\sigma(i) = \{a \in A(i) : \sigma(a|i) > 0\}$ be the set of actions used by the policy σ at state $i \in S$. We denote by S^* the set of states $i \in S$, for which $A^\sigma(i)$ consists of more than one point. For each $i \in S$, fix an arbitrary order

in which these elements are ordered, $A^\sigma(i) = \{a_1(i), a_2(i), \dots, a_{m(i)}(i)\}$, where $\sum_{i\in S}[m(i)-1] = \sum_{i\in S^*}[m(i)-1] \le K$. For each state $i \in S$ there are $m(i)!$ possible orders. Therefore, there are $\prod_{i\in S^*} m(i)!$ possibilities to fix such orders at all states. When all these orders are fixed, a policy σ define a K-switching stationary strategy ϕ defined as follows

$$\phi(i,t) = a_k(i) \quad \texttt{when } S_{k-1}(i) \le t < S_k(i),\ k = 1, \dots, m(i),\ i \in S, \qquad (2.2.1)$$

where

$$S_0(i) = 0, S_k(i) = S_{k-1}(i) + s_k(i), \qquad k = 1, 2, \dots, m(i),$$

and

$$s_k(i) = -\frac{1}{\alpha + q(i, a_k(i))} \ln\big(1 - \frac{\sigma(a_k(i)|i)}{\sum_{l=k}^{m} \sigma(a_l(i)|i)}\big).$$

In the above formulas $s_k(i)$ are the lengths of time intervals between switching epoches $S_{k-1}(i)$ and $S_k(i)$. Note that $S_{m(i)}(i) = s_{m(i)}(i) = +\infty$.

We use an ordered sequence $< a_1(i), a_2(i) \dots, a_{m(i)}(i) >$ to represent a K-switching randomized strategy ϕ at state i. Then the set of orders $< a_1(i), a_2(i) \dots, a_{m(i)}(i) >_{i\in S^*}$ and formula (2.2.1) define a K-switching strategy ϕ.

Thus formula (2.2.1) defines $N = \prod_{i\in S^*} m(i)!$ different K-switching policies ϕ for the K-randomized stationary policy σ. We introduce the notation $M = \sum_{i\in S}[m(i)-1]$. Since σ is a K-randomized stationary policy, $M \le K$. We observe that $2^M \le N \le (M+1)!$. In particular, $N = 2^M$ when S consists of M points and each set $A^\sigma(i)$, $i \in S^*$, consists of two points, and $N = (M+1)!$ when S^* is a singleton and $A(i^*)$ consists of $(M+1)$ points, where $A(i^*) = \{i^*\}$.

According to [13, Corollary 5.3], the policies ϕ satisfy the following property.

Theorem 2.2.2 *Given a K-randomized stationary policy σ, for any K-switching strategy ϕ defined by (2.2.1), the equalities $V_k(i,\phi) = V_k(i,\sigma)$, $k = 0, 1, \dots, K$, hold for all $i \in S$.*

Theorem 2.2.3 implies that if σ is an optimal policy for problem (2.1.2) then ϕ is also optimal for this problem. In fact, in Feinberg [13, Corollary 5.3] implies that $V_k(i,\sigma) = V_k(i,\phi)$ for any reward functions R and r. We remark that [13, Corollary 5.3] is formulated for a fixed initial state distribution. However the definition of ϕ does not depend on the initial state distribution;

see (2.2.1). Thus, the equality of the expected total discounted rewards in [13, Corollary 5.3] and Theorem 2.2.3 in holds for any initial measure.

For each K-randomized stationary strategy ϕ defined by a described order and (2.2.1) for $i \in S$ and $k = 1, \dots, m(i)$

$$E_i^{\phi} e^{-\alpha T_1} I\{a_0 = a_k(i)\} = E_i^{\sigma} e^{-\alpha T_1} I\{a_0 = a_k(i)\} = \frac{\sigma(a_k(i)|i) q(i, a_k(i))}{\alpha + q(i, a_k(i))}, \tag{2.2.2}$$

The second equality follows from the fact that $Ee^{-\alpha\xi} = q/(\alpha+q)$ for an exponential random variable with an intensity q and T_1 is a mixture of exponential random variables with the intensities $q(i, a_k(i))$ and each $a_k(i)$ is selected with the probability $\sigma(a_k(i)|i)$. The first equality follows from [13, (5.10)]. Since at jump epoch T_1 the transition probabilities are equal to $q(i, a_0)/(\alpha + q(i, a_0))$ for the both policies σ and ϕ, equalities (2.2.2) imply

$$E_i^{\phi} e^{-\alpha T_1} I\{a_0 = a_k(i), X_1 = j\} = E_i^{\sigma} e^{-\alpha T_1} I\{a_0 = a_k(i), X_1 = j\} = \frac{\sigma(a_k(i)|j) q(i, j, a_k(i))}{\alpha + q(i, a_k(i))}. \tag{2.2.3}$$

In addition,

$$E_i^{\sigma} \int_0^{T_1} e^{-\alpha t} I\{a(t) = a_k(i)\} dt = E_i^{\phi} \int_0^{T_1} e^{-\alpha t} I\{a(t) = a_k(i)\} dt = \frac{\sigma(a_k(i)|j)}{\alpha + q(i, a_k(i))}, \tag{2.2.4}$$

where the second equation follows from the fact that $E \int_0^{\xi} e^{-\alpha t} dt = 1/(\alpha + q)$ for an exponential random variable with an intensity q and T_1 is a mixture of exponential random variables with the intensities $q(i, a_k(i))$ and each $a_k(i)$ is selected with the probability $\sigma(a_k(i)|i)$. The first equation follows from (2.2.2) and [13, Lemma 4.3] or from [10, Theorem 1].

According to [13, Corollary 5.3], the policies ϕ satisfy the following property. This property can also be proved by using (2.2.2-2.2.4).

Theorem 2.2.3 *Given a K-randomized stationary policy σ, for any K-switching strategy ϕ defined by (2.2.1), the equalities $V_k(i, \phi) = V_k(i, \sigma)$, $k = 0, 1, \dots, K$, hold for all $i \in S$.*

Theorem 2.2.3 implies that if σ is an optimal policy for problem (2.1.2) then ϕ is also optimal for this problem. In fact, in Feinberg [13, Corollary 5.3] implies that $V_k(i, \sigma) = V_k(i, \phi)$ for any reward functions R and r. We remark that [13, Corollary 5.3] is formulated for a fixed initial state distribution.

However the definition of ϕ does not depend on the initial state distribution; see (2.2.1). Thus, the equality of the expected total discounted rewards in [13, Corollary 5.3] and Theorem 2.2.3 in holds for any initial measure.

2.3 Problem formulation

We are interested in comparing the variances of the total discounted rewards corresponding to the objective criterion V_0 for a K-randomized stationary policy and the corresponding K-switching strategies. The aim is to select an optimal strategy with the smallest possible variance of the objective function. In the rest of this paper we deal only with objective criterion and omit everywhere index $k = 0$. For example, we shall write V, r, and R instead of V_0, r_0, and R_0.

Consider a K-randomized stationary policy σ. Since at each state it uses a finite number of actions $A^\sigma(i)$ and the corresponding switching stationary strategies use the same actions we can limit the action sets to finite sets $A^\sigma(i)$. Therefore, without loss of generality, in the rest of this paper we consider only finite action set $A(i), i \in S$. In fact, since $|A^\sigma(i)| \le K$ for all $i \in S$, we can assume that A is finite. So, consider an MDP with a finite action set A.

Unlike the expectation V, the second moments of the total discounted reward depends on the particular definition of discounting. There are two natural ways to interpret discounting. One way is to interpret discounting as the coefficient in front of the future reward rates. One unit of reward earned at a future time t is worth only $e^{-\alpha t}$ at time 0 if the compounding is continuous. The discounted reward earned under this interpretation is

$$J^{[1]}(\omega_\infty) = \sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R(X_n, a_n, X_{n+1}) + \int_0^\infty e^{-\alpha t} r(X(t), a(t))\, dt,$$

where $\omega_\infty = X_1, T_1, X_2, T_2, \ldots$ and $a_n = a(T_n-)$.

Another way is to interpret discounting as a stopping intensity. Let the time horizon of the process, T be an exponentially distributed random variable independent of the stochastic sequence $X_0, T_1, X_1, \ldots$ for any initial state i and for any strategy $\sigma \in R\Pi$. The random variable T is defined on its probability space and with a slight abuse notations we shall use P_i^σ for the probability on the product of this space and $(\Omega_\infty, \mathcal{F}_\infty)$. In particular, $P_i^\sigma(T \le t, A) = e^{-\alpha t} P_i^\sigma(A)$ for any $t \in [0, \infty]$ and for any $A \in \mathcal{F}_\infty$. We shall also keep the notation E_i^σ for this extended probability space.

The discounted reward earned under this interpretation is

$$J^{[2]}(\omega_\infty) = \sum_{n=0}^{n(T)-1} R(X_n, a_n, X_{n+1}) + \int_0^T r(X(t), a(t))\, dt,$$

where $n(T) = \sup\{n : T_n \leq T\}$.

In this paper we primarily follow the second interpretation of discounting and, for a given K-randomized stationary policy σ we consider a finite collection of strategies Ψ consisting of this policy σ and all equivalent K-switching stationary strategies ϕ defined by (2.2.1). The aim is to select a natural optimal policy with minimal variance. In other words we are to solve such an optimization problem.

$$min\{E_i^\pi (J^{[\ell]}(\omega_\infty))^2 : \pi \in \Psi\}, \ell = 1, 2$$

We comment that discounting as a stopping intensity is equivalent to adding an absorbing state to the model. Each other state jumps to the absorbing state with an intensity α, where α is the discount rate.

Chapter 3

Inequality for Variances of the Discounted Rewards

3.1 Introduction

As we mentioned in the previous chapter there are two natural definitions of total discounted rewards: (i) by interpreting discounting as a coefficient in front of the future reward rates (multiplicative discount), and (ii) by interpreting discounting as the probability that the process has not been stopped if the stopping time has an exponential distribution independent of the process (probabilistic discounting). It is well-known that the expected total discounted rewards corresponding to these definitions are the same. In this chapter we show that for the first definition the variance of the total discounted rewards is smaller than for the second one. Instead of restriction within the settings of Markov Decision Processes we consider a general probability space. We remark that the inequality relation discussed in this chapter is true for general probability space endowed with reward processes.

3.2 Results in general stochastic process

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\mathcal{F}_t$, $t \in [0, \infty)$, where $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for all $0 \leq s < t < \infty$. Consider a nondecreasing sequence of stopping times T_n, $n = 1, 2, \dots$. Let $\mathcal{F}_\infty = \bigcup_{t \in [0,\infty)} \mathcal{F}_t$.

We consider an $\mathcal{F}_t$-adapted stochastic process r_t, $t \in [0, \infty)$, and an $\mathcal{F}_{T_n}$-adapted stochastic sequence R_n, $n = 1, 2, \dots$. The process r_t can be inter-

preted as the reward rate at time t. In addition, a lump sum R_n is collected at time T_n.

There are two natural ways to define the total discounted rewards. One way is to interpret discounting as the coefficient in front of the reward rate. In this case, the total discounted rewards are defined as

$$J_1 = \int_0^\infty e^{-\alpha t} r_t \, dt + \sum_{n=1}^{\infty} e^{-\alpha T_n} R_n, \tag{3.2.1}$$

where $\alpha > 0$ is the discount rate.

Another way is to define the total discounted rewards as the total rewards until a stopping time T that has an exponential distribution with rate α. Let T be independent of $\mathcal{F}_\infty$ and $P\{T > t\} = e^{-\alpha t}$. Then the total discounted reward can be defined as

$$J_2 = \int_0^T r_t \, dt + \sum_{n=1}^{N(T)} R_n, \tag{3.2.2}$$

where $N(t) = \sup\{n : T_n \le t\}$, $t \ge 0$.

It is well known that

$$E\big[J_1\big] = E\big[J_2\big], \tag{3.2.3}$$

if at least one side of this equation is well-defined (a random variable has a well-defined expectation if either the expectation of its positive part is finite or the expectation of its negative part is finite).

Indeed,

$$\begin{aligned} & E\sum_{n=1}^{N(T)} R_n = \sum_{n=1}^{\infty} E\, R_n I\{T \ge T_n\} = \sum_{n=1}^{\infty} E\, E[R_n I\{T \ge T_n\} | \mathcal{F}_{T_n}] \\ = \;& E\sum_{n=1}^{\infty} R_n E[I\{T \ge T_n\} | \mathcal{F}_{T_n}] = E\sum_{n=1}^{\infty} R_n P\{T \ge T_n | \mathcal{F}_{T_n}\} = E\sum_{n=1}^{\infty} R_n e^{-\alpha T_n} \end{aligned}$$

and

$$E\int_0^T r_t\,dt = E\,[\int_0^\infty r_t I\{T\ge t\}\,dt] = \int_0^\infty E\,[r_t I\{T\ge t\}]\,dt$$
$$= \int_0^\infty E\,[r_t]\cdot E[I\{T\ge t\}]\,dt = \int_0^\infty E\,r_t P\{T\ge t\}\,dt = E\int_0^\infty e^{-\alpha t} r_t\,dt.$$

In particular, (3.2.3) holds for deterministic functions r and R and therefore

$$E\big[J_1|\mathcal{F}_\infty\big] = E\big[J_2|\mathcal{F}_\infty\big] \qquad P-\text{a.s.}, \tag{3.2.4}$$

if either $E\big[|J_1|\,|\mathcal{F}_\infty\big] < \infty$ or $E\big[|J_2|\,|\mathcal{F}_\infty\big] < \infty$ P-a.s.

However, the second moments can be different. Indeed, we have the following statement.

Theorem 3.2.1 *If either* $E\big[|J_1|\,|\mathcal{F}_\infty\big] < \infty$ *or* $E\big[|J_2|\,|\mathcal{F}_\infty\big] < \infty$ P*-a.s. then*

$$Var(J_1) \le Var(J_2),$$

and the equality holds if and only if $Var(J_2|\mathcal{F}_\infty) = 0$ P*-a.s.*

Proof. By the total variance formula [52, p. 83] or [20, p. 454] for $i = 1, 2$

$$\mathrm{Var}(J_i) = E[\mathrm{Var}(J_i|\mathcal{F}_\infty)] + \mathrm{Var}(E[J_i|\mathcal{F}_\infty]).$$

Therefore, because of (3.2.4),

$$\mathrm{Var}(E[J_1|\mathcal{F}_\infty]) = \mathrm{Var}(E[J_2|\mathcal{F}_\infty]).$$

In addition, $E[\mathrm{Var}(J_1|\mathcal{F}_\infty)] = 0$ and $E[\mathrm{Var}(J_2|\mathcal{F}_\infty)] \ge 0$. Hence, $\mathrm{Var}(J_2) - \mathrm{Var}(J_1) = E[\mathrm{Var}(J_2|\mathcal{F}_\infty)] \ge 0$, i.e., $\mathrm{Var}(J_1) \le \mathrm{Var}(J_2)$.

3.3 Results in a Semi-Markov Process

A semi-Markov process is a stochastic process $\{X(t), t \geq 0\}$ with a finite or countable set of states $N = \{1, 2, \ldots\}$, having stepwise trajectories with jumps at time $0 < T_1 < T_2 < \ldots$ and such that values $X(T_n)$ at its jumps form a Markov chain with transition probabilities

$$p_{ij} = P\{X(T_n) = j)|X(T_{n-1}) = i\}.$$

The distribution of the jump times T_n are described in terms of the distribution functions $F_{ij}(t)$ as follows

$$P\{T_n - T_{n-1} \leq t, X(T_n) = j|X(T_{n-1}) = i\} = p_{ij}F_{ij}(t)$$

and moreover, they are independent of the states of the process at earlier moments of time.

Consider a reward structure specified by $\{r, R\}$ where $r(x)$ is the reward rate at state x and $R(i, j)$ is the instant reward earned when transiting from state i to state j.

Under the multiplicative discounting the total discounted rewards are defined in (3.2.1). Under the probabilistic discounting the total discounted rewards are defined in (3.2.2).

In the sequel we will present and prove results specific to semi-Markov process though they are proved in section 3.2.

Corollary 3.3.1 *Consider a semi-Markov process $\{X(t)\}$ endowed with a reward structure $\{r, R\}$. In terms of the first moment, the two definitions of discounting are equivalent, i.e. $E[J^{[1]}] = E[J^{[2]}]$.*

Proof. We will show

$$E\sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R(X_n, X_{n+1}) = E\sum_{n=0}^{n(T)-1} R(X_n, X_{n+1}),$$

and

$$E\int_0^{\infty} e^{-\alpha t} r(X(t))\, dt = E\int_0^{T} r(X(t))\, dt.$$

Indeed,

$E\sum_{n=0}^{n(T)-1} R(X_n, X_{n+1}) = \sum_{n=0}^{\infty} E\, R(X_n, X_{n+1}) I\{T > T_{n+1}\}$
$= \sum_{n=0}^{\infty} E\, E[R(X_n, X_{n+1}) I\{T > T_{n+1}\} | X_n, a_n, X_{n+1}, T_{n+1}]$
$= \sum_{n=0}^{\infty} E\, R(X_n, X_{n+1}) E[I\{T > T_{n+1}\} | X_n, a_n, X_{n+1}, T_{n+1}]$
$= \sum_{n=0}^{\infty} E\, R(X_n, X_{n+1}) P\{T > T_{n+1} | T_{n+1}\}$
$= \sum_{n=0}^{\infty} E\, R(X_n, X_{n+1}) e^{-\alpha T_{n+1}} = E\sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R(X_n, X_{n+1})$,

where we use that, given T_{n+1}, the event $\{T > T_{n+1}\}$ and the random vector (X_n, X_{n+1}) are independent, and

$E\left[\int_0^{\infty} r(X(t)) I\{T > t\}\, dt\right] = \int_0^{\infty} E\left[r(X(t)) I\{T > t\}\right] dt =$
$\int_0^{\infty} E\left[r(X(t))\right] \cdot E[I\{T > t\}]\, dt = \int_0^{\infty} E\, r(X(t)) P\{T > t\}\, dt$
$= \int_0^{\infty} E\, r(X(t)) e^{-\alpha t}\, dt = E\int_0^{\infty} e^{-\alpha t} r(X(t))$,

where the third equality follows from the independence of $r(X(t))$ and T.

■

As to the inequality of variances we will show the proof for two special cases: one with cumulative rewards r only and the other with instant rewards R only.

Corollary 3.3.2 *Consider a semi-Markov process $\{X(t)\}$ endowed with a reward structure $\{r, R\}$. When $r = 0$ or $R = 0$, variance of total discounted rewards under the multiplicative discounting is less than or equal to that under the probabilistic discounting, i.e., $Var_i^{\pi}[J^{[1]}] \le Var_i^{\pi}[J^{[2]}]$.*

Proof. For simplicity we simplify the notation as follows: $R(X_n, X_{n+1}) \triangleq R_n$ and $r(X(t)) \triangleq r_t$.

We repeat the definition of $J^{[1]}$ and $J^{[2]}$ here:

$$J^{[1]} = \sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R_n + \int_0^{\infty} e^{-\alpha t} r_t\, dt,$$

$$J^{[2]} = \sum_{n=0}^{n(T)-1} R_n + \int_0^{T} r_t\, dt,$$

When $R = 0$ we need to show

$$E\left(\int_0^{\infty} e^{-\alpha t} r_t\, dt\right)^2 \le E\left(\int_0^{T} r_t\, dt\right)^2 \tag{3.3.1}$$

When $r = 0$ we need to show

$$E\Big(\sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R_n\Big)^2 \le E\Big(\sum_{n=0}^{n(T)-1} R_n\Big)^2 \tag{3.3.2}$$

Proof of inequality (3.3.1) is as follows:

$$\begin{aligned} E\Big(\int_0^T r_t\,dt\Big)^2 &= EE\Big[\big(\int_0^t r_t\,dt\big)^2 \,|T = t\Big] \\ &= \int_0^\infty \alpha e^{-\alpha t} E\big(\int_0^t r_s\,ds\big)^2\,dt \end{aligned}$$

$$\begin{aligned} \int_0^\infty e^{-\alpha t} r_t\,dt &= \int_0^\infty e^{-\alpha t} d\int_0^t r_s ds \\ &= e^{-\alpha t}\int_0^t r_s ds\,|_0^\infty - \int_0^\infty \big(\int_0^t r_s ds\big)\,de^{-\alpha t} \\ &= 0 + \int_0^\infty \alpha e^{-\alpha t}\big(\int_0^t r_s ds\big)\,dt \\ &= \int_0^\infty \alpha e^{-\alpha t}\,\big(\int_0^t r_s ds\big)\,dt \end{aligned}$$

Now we want to use the inequity $\big(\int fg\big)^2 \le \big(\int f^2\big)\big(\int g^2\big)$. Set $f = \sqrt{\alpha e^{-\alpha t}}$ and $g = \sqrt{\alpha e^{-\alpha t}}\int_0^t r_s ds$. Then we have

$$\begin{aligned} \big(\int_0^\infty \alpha e^{-\alpha t}\int_0^t r_s ds\,dt\big)^2 &\le \int_0^\infty \alpha e^{-\alpha t} dt \int_0^\infty \alpha e^{-\alpha t}\big(\int_0^t r_s ds\big)^2\,dt \\ &= \int_0^\infty \alpha e^{-\alpha t}\big(\int_0^t r_s ds\big)^2\,dt \end{aligned}$$

This proves (3.3.1).

For inequality (3.3.2), we proceed as follows:

$$E_i^\pi\Big(\sum_{n=0}^{n(T)-1} R_n\Big)^2 = E_i^\pi\Big(\sum_{n=0}^{\infty} R_n I\{n(T)\geq n+1\}\Big)^2 = E_i^\pi\Big(\sum_{n=0}^{\infty} R_n I\{T\geq T_{n+1}\}\Big)^2$$

$$= E_i^\pi \sum_{n=0}^{\infty} R_n^2\, I\{T\geq T_{n+1}\} + 2E_i^\pi \sum_{n=1}^{\infty}\sum_{m=0}^{n-1} R_n R_m I\{T\geq T_{n+1}\}$$

$$= E_i^\pi \sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R_n^2 + 2E_i^\pi \sum_{n=1}^{\infty}\sum_{m=0}^{n-1} e^{-\alpha T_{n+1}} R_n R_m$$

On the other hand,

$$E_i^\pi\Big(\sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R_n\Big)^2 = E_i^\pi \sum_{n=0}^{\infty} e^{-2\alpha T_{n+1}} R_n^2 + 2E_i^\pi \sum_{n=1}^{\infty}\sum_{m=0}^{n-1} e^{-\alpha(T_{n+1}+T_{m+1})} R_n R_m$$

$$\leq E_i^\pi \sum_{n=0}^{\infty} e^{-\alpha T_{n+1}} R_n^2 + 2E_i^\pi \sum_{n=1}^{\infty}\sum_{m=0}^{n-1} e^{-\alpha T_{n+1}} R_n R_m = E_i^\pi\Big(\sum_{n=0}^{n(T)-1} R_n\Big)^2$$

This completes the proof of (3.3.2). ∎

We remark that when both r and R are present in the reward structure the above traditional method of proof has difficulty due to the "cross-product" terms between the cumulative rewards and instant rewards.

3.4 Results in a Continuous-time Markov Chain

In the semi-Markov process if

$$F_i' j(t) = e^{-q_{ij}t}, t\geq 0$$

for all $i, j \in N$, then the semi-Markov process $\{X(t)\}$ is a continuous-time Markov chain. In particular if all the distributions degenerate to a constant (unit inter-arrival times) the process is further reduced to a discrete-time Markov chain.

In the sequel we give two examples that illustrate the inequality presented in theorem (3.2.1).

Example 3.4.1 *Consider a Markov process with two states:* 1 *and* 0*, where* 0 *is an absorbing state. Let state* 1 *be the initial state. The process spends an exponential time* $T_1 \sim exp(\lambda)$ *at state 1 and then jumps to state 0. At state 1 the reward rate is 1 and at the jump epoch there is no lump-sum reward. At state 0 the process collects no rewards. Let the discount factor be* α *and* $T \sim exp(\alpha)$*.*

The total discounted rewards under the two definitions are

$$\begin{aligned} J_1 &= \int_0^{T_1} e^{-\alpha t}\, dt = \frac{1}{\alpha}(1 - e^{-\alpha T_1}), \\ J_2 &= \int_0^{T \wedge T_1} dt = T \wedge T_1. \end{aligned}$$

For the first definition,

$$Var(J_1) = \frac{1}{\alpha^2} Var(e^{-\alpha T_1}) = \frac{1}{\alpha^2}(M_{T_1}(-2\alpha) - (M_{T_1}(-\alpha))^2) = \frac{\lambda}{(\lambda+\alpha)^2(\lambda+2\alpha)},$$

where $M_X(s)$ *is the moment generating function of a random variable* X*. In particular,* $M_{T_1}(s) = \lambda/(\lambda - s)$*.*

Since $T \wedge T_1$ *is an exponential random variable with intensity* $\lambda + \alpha$*,*

$$Var(J_2) = \frac{1}{(\lambda+\alpha)^2}.$$

Thus, $Var(J_1) < Var(J_2)$*.*

Example 3.4.2 *Consider a discrete time Markov chain where at each jump the process receives a lump sum reward of 1. Let the time interval between jumps be 1 unit of time. The discount factor is* α *and* $T \sim exp(\alpha)$*.*

The total discounted rewards under the two definitions are respectively

$$\begin{aligned} J_1 &= \sum_{n=1}^{\infty} e^{-\alpha n} = \frac{e^{-\alpha}}{1-e^{-\alpha}}, \\ J_2 &= \sum_{n=1}^{N(T)} 1 = N(T). \end{aligned}$$

Note that J_1 is a deterministic number and J_2 is a random variable depending on T. Thus, $Var(J_1) = 0 < Var(J_2)$. In fact, since the inter-arrival time is 1 $N(T) = [T]$ where $[x]$ is the integer part of $x \in R_+$. We have

$$\begin{aligned} E[J_2] &= \sum_{n=0}^{\infty} \int_n^{n+1} n\alpha e^{-\alpha t} dt = \sum_{n=0}^{\infty} (1-e^{-\alpha}) n e^{-\alpha n} \\ &= (1-e^{-\alpha}) \sum_{n=0}^{\infty} n e^{-\alpha n} = (1-e^{-\alpha}) \frac{e^{-\alpha}}{(1-e^{-\alpha})^2} = \frac{e^{-\alpha}}{1-e^{-\alpha}} \\ E[J_2^2] &= \sum_{n=0}^{\infty} \int_n^{n+1} n^2 \alpha e^{-\alpha t} dt = \sum_{n=0}^{\infty} (1-e^{-\alpha}) n^2 e^{-\alpha n} \\ &= (1-e^{-\alpha}) \sum_{n=0}^{\infty} n^2 e^{-\alpha n} = (1-e^{-\alpha}) \frac{e^{-\alpha}(1+e^{-\alpha})}{(1-e^{-\alpha})^3} = \frac{e^{-\alpha}(1+e^{-\alpha})}{(1-e^{-\alpha})^2} \\ Var(J_2) &= E[J_2^2] - E^2[J_2] = \frac{e^{-\alpha}(1+e^{-\alpha})}{(1-e^{-\alpha})^2} - \frac{e^{-2\alpha}}{(1-e^{-\alpha})^2} = \frac{e^{-\alpha}}{(1-e^{-\alpha})^2}. \end{aligned}$$

Chapter 4

Variance under probabilistic discounting

4.1 Formulas for the moments for the model up to the first jump

In this section we derive formulas for the first and second moments of the model up to the first jump. Consider a $K-$randomized stationary policy σ. Assume there are m actions available at state i, namely $\{a_1, a_2, \ldots, a_m\}$. For convenience let $p_k = \sigma(a_k), q_k = q(i, a_k), r_k = r(i, a_k), k = 1, 2, \ldots, m$.

When policy π, which could be a randomized stationary policy or a switching stationary strategy, is adopted we define the first and second moment of the instant reward earned at the first jump as follows

$$m_a = E_i^{\pi}[R(X_0, a_0, X_1)|a_0 = a],$$

$$w_a = E_i^{\pi}[R^2(X_0, a_0, X_1)|a_0 = a].$$

We first consider a switching strategy ϕ defined by (2.2.1). Assume the order is fixed as $< a_1, a_2, \ldots, a_m >$. During the time interval $[S_{k-1}, S_k]$, action a_k is taken and the jump intensity is q_k. In this interval the process may jump to the next state before reaching the changing epoch S_k, or go to the next time interval $[S_k, S_{k+1})$ and take action a_{k+1}. Let U_k be the total rewards earned starting S_{k-1} up to the first jump given that the process does not jump before

S_{k-1}. That is,

$$U_k = \int_{S_{k-1}}^{T_1} r(X_0, a(t))dt + R(X_0, a_0, X_1) I\{T_1 > S_{k-1}\}$$

Note that U_1 is the total rewards earned starting $t = 0$ up to the first jump. Let M_k and W_k be the first and second moment of U_k.

Let ζ_k be the exponential random variable with intensity q_k. There are two cases for each interval:

Case 1: If $\zeta_k < s_k$: the process jumps before S_k. The rewards earned starting S_{k-1} consists only of the reward earned between S_{k-1} and $S_{k-1} + \zeta_k$.

Case 2: If $\zeta_k > s_k$: the process continues and selects the next action a_{k+1}. The rewards earned starting S_{k-1} consists of two parts: reward earned between S_{k-1} and S_k, and the rewards earned starting S_k.

Backward from the last interval, we derive formulas to compute the moments of total reward up to the first jump.

Theorem 4.1.1 *For a switching strategy $\phi =< a_1, \ldots, a_m >$, the first and second moment of the total reward earned up to the first jump can be computed recursively as follows,*

$$\begin{aligned}
M_m &= \frac{r_m}{q_m} + m_{a_m}, \\
M_k &= (1 - e^{-q_k s_k})\big(\frac{r_k}{q_k} + m_{a_k}\big) + e^{-q_k s_k} M_{k+1}, \\
W_m &= \frac{2{r_m}^2}{{q_m}^2} + 2\frac{r_m}{q_m} m_{a_m} + w_{a_m}, \\
W_k &= \frac{2r_k}{q_k}(1 - e^{-q_k s_k} - q_k s_k e^{-q_k s_k})\big(\frac{r_k}{q_k} + m_{a_k}\big) + \\
&\quad e^{-q_k s_k}(W_{k+1} + 2r_k s_k M_{k+1}) + w_{a_k}, k = 1, 2, \ldots, m-1.
\end{aligned}$$

Proof. First, if $T_1 > S_{m-1}$, action a_m will be taken. The process will continue until the first jump. Before its jump the sojourn time is an exponential random variable with rate q_m. Thus,

$$M_m = \frac{r_m}{q_m} + E[R(X_0, a_m, X_1)] = \frac{r_m}{q_m} + m_{a_m},$$

and

$$\begin{aligned} W_m &= E\big[\Big(r_m T_m + R(X_0, a_m, X_1)\Big)^2\big] \\ &= E[r_m^2 T_m^2] + 2r_m E[T_m] m_{a_m} + w_{a_m} \\ &= \frac{2{r_m}^2}{{q_m}^2} + 2\frac{r_m}{q_m} m_{a_m} + w_{a_m}. \end{aligned}$$

To find M_k we condition it on ζ_k. From formula of iterated expectation, we have

$$\begin{aligned} M_k &= EE[U_k | T_1 > S_{k-1}, \zeta_k] \\ &= \int_0^{s_k} \Big(r_k\zeta_k + m_{a_k}\Big) f(\zeta_k) d\zeta_k + \int_{s_k}^{+\infty} f(\zeta_k)\, d\zeta_k\, (r_k s_k + E[U_{k+1} | T_1 > S_k]) \\ &= \int_0^{s_k} r_k \zeta_k f(\zeta_k)\, d\zeta_k + m_{a_k} \int_0^{s_k} f(\zeta_k)\, d\zeta_k + \int_{s_k}^{+\infty} f(\zeta_k)\, d\zeta_k\, (r_k s_k + M_{k+1}) \\ &= (1 - e^{-q_k s_k} - e^{-q_k s_k} q_k s_k)\frac{r_k}{q_k} + m_{a_k}(1 - e^{-q_k s_k}) + e^{-q_k s_k}(r_k s_k + M_{k+1}) \\ &= (1 - e^{-q_k s_k})\big(\frac{r_k}{q_k} + m_{a_k}\big) + e^{-q_k s_k}\, M_{k+1} \end{aligned}$$

To find W_k, we also condition it on ζ_k:

$$\begin{aligned} W_k &= EE[{U_k}^2 | T_1 > S_{k-1}, \zeta_k] \\ &= \int_0^{s_k} {r_k}^2 {\zeta_k}^2 f(\zeta_k) d\zeta_k + 2m_{a_k} \int_0^{s_k} r_k \zeta_k f(\zeta_k) d\zeta_k + \int_0^{s_k} f(\zeta_k) d\zeta_k\, W_{a_k} \\ &\quad + \int_{s_k}^{+\infty} f(\zeta_k) d\zeta_k\, ({r_k}^2 {s_k}^2 + 2r_k s_k E[U_{k+1} | T_1 > S_k] + E[U_{k+1}^2 | T_1 > S_k]) \\ &= [2 - e^{-q_k s_k}(2 + 2q_k s_k + q_k^2 s_k^2)]\frac{r_k^2}{q_k^2} + 2\alpha(1 - e^{-q_k s_k} - e^{-q_k s_k} q_k s_k)\frac{r_k}{q_k} m_{a_k} \\ &\quad + (1 - e^{-q_k s_k}) w_{a_k} + e^{-q_k s_k}({r_k}^2 {s_k}^2 + 2r_k s_k M_{k+1} + W_{k+1}) \\ &= \frac{2r_k^2}{q_k^2}(1 - e^{-q_k s_k} - q_k s_k e^{-q_k s_k}) + e^{-q_k s_k}(W_{k+1} + 2r_k s_k M_{k+1}) \\ &\quad + 2(1 - e^{-q_k s_k} - e^{-q_k s_k} q_k s_k)\frac{r_k}{q_k} m_{a_k} + (1 - e^{-q_k s_k}) w_{a_k} \\ &= \frac{2r_k}{q_k}(1 - e^{-q_k s_k} - q_k s_k e^{-q_k s_k})\big(\frac{r_k}{q_k} + m_{a_k}\big) + e^{-q_k s_k}(W_{k+1} + 2r_k s_k M_{k+1} + w_{a_k} \end{aligned}$$

The formulas for the randomized policy are much simpler. They are simply weighted average of each pure Markov process when a single action is played.

Theorem 4.1.2 *For a randomized policy $\sigma = \{a_1, \ldots, a_m\}$, the first and second moment of the total reward earned up to the first jump can be computed using the following formula,*

$M = \sum_{k=1}^{m} p_k \left(\frac{r_k}{q_k} + m_{a_k} \right)$,

$W = \sum_{k=1}^{m} p_k \left(\frac{2{r_k}^2}{{q_k}^2} + 2\frac{r_k}{q_k} m_{a_k} + w_{a_k} \right)$.

4.2 "Action index" for switching strategies

We first consider the simplest case where the action set contains only two actions a and b. We have the following result:

Lemma 4.2.1 *For a two-action set, $A(i) = \{a, b\}$, if $r_a/q_a + m_a > r_b/q_b + m_b$, then $W_{<a,b>} < W_{\{a,b\}} < W_{<b,a>}$.*

Proof. Using formula in Theorem (4.1.1), for the switching strategy $< a, b >$:

$$\begin{aligned} M_b &= \frac{r_b}{q_b} + m_b, \\ W_b &= 2\frac{r_b}{q_b}\left(\frac{r_b}{q_b} + m_b\right) + w_b. \end{aligned}$$

The first and second moment of the total reward up to the first jump is then

$$\begin{aligned} M_{<a,b>} &= \left(\frac{r_a}{q_a} + m_a\right)p_a + \left(\frac{r_b}{q_b} + m_b\right)p_b, \\ W_{<a,b>} &= \left(2\frac{r_b}{q_b}\left(\frac{r_b}{q_b} + m_b\right) - 2\frac{r_a}{q_a}\left(\frac{r_b}{q_b} + m_b\right)\ln(p_b)p_b + \right. \\ &\quad \left. 2\frac{r_a}{q_a}\left(\frac{r_a}{q_a} + 2m_a\right)\left(p_a + p_b\ln(p_b)\right) + w_b p_b + w_a p_a. \end{aligned}$$

To get $W_{<b,a>}$ we just need to swap a and b in $W_{<a,b>}$.
$W_{\{a,b\}} = \Big[2\frac{r_a}{q_a}(\frac{r_a}{q_a} + m_a) + w_a\Big]p_a + \Big[2\frac{r_b}{q_b}(\frac{r_b}{q_b} + m_b) + w_b\Big]p_b$

To compare we take the differences:

$$\begin{aligned}
W_{<a,b>} - W_{\{a,b\}} &= \frac{2r_a}{q_a}p_b \ln(p_b)\Big[(\frac{r_a}{q_a} + m_a) - (\frac{r_b}{q_b} + m_b)\Big], \\
W_{<b,a>} - W_{\{a,b\}} &= \frac{2r_b}{q_b}p_a \ln(p_a)\Big[(\frac{r_b}{q_b} + m_b) - (\frac{r_a}{q_a} + m_a)\Big], \\
W_{<a,b>} - W_{<b,a>} &= 2\Big[\frac{r_a}{q_a}p_b \ln(p_b) + \frac{r_b}{q_b}p_a \ln(p_a)\Big]\Big[(\frac{r_a}{q_a} + m_a) - (\frac{r_b}{q_b} + m_b)\Big].
\end{aligned}$$

Apparently, if $r_a/q_a + m_a > r_b/q_b + m_b$, then $W_{<a,b>} - W_{\{a,b\}} < 0$, i.e., $W_{<a,b>} < W_{\{a,b\}}$. Similarly, $W_{<b,a>} > W_{\{a,b\}}$ and $W_{<a,b>} < W_{<b,a>}$.

It seems that the quantity $r_a/q_a + m_a$ works as an index. If we sort the actions in the descending order of this index we will obtain an indexed switching strategy. For the case of two actions the indexed switching strategy is better than the randomized policy and is better than the other switching strategy too. Thus, the indexed switching strategy is the best for the case of two actions.

4.3 Interchanging of two neighboring actions

Next we compare two switching strategies obtained by interchanging two neighboring actions where action set has more than two actions. We use $< A, a, b, B >$ and $< A, b, a, B >$ to denote two switching strategies that differ only in the order of action a and b, where A and B are action sequences and there are $m - k$ actions in B. The following theorem shows that it suffices to compare the variances of subsequences of actions $< a, b, B >$ and $< b, a, B >$ in order to compare the variances of $< A, a, b, B >$ and $< A, b, a, B >$.

Lemma 4.3.1 *If $W_{<a,b,B>} > W_{<b,a,B>}$, then $W_{<A,a,b,B>} > W_{<A,b,a,B>}$.*

Proof. Let $A = \{a_1, \ldots, a_{k-2}\}$. Using the formula in Theorem (4.1.1), we obtained the second moment of the subsequence $< a_{k-2}, \ldots >$ as
$W_{k-2} = \frac{2r_{k-2}}{q_{k-2}}(1 - e^{-q_{k-2}s_{k-2}} - q_{k-2}s_{k-2}e^{-q_{k-2}s_{k-2}})\big(\frac{r_{k-2}}{q_{k-2}} + m_{a_{k-2}}\big) +$
$e^{-q_{k-2}s_{k-2}}(W_{k-1} + 2r_{k-2}s_{k-2}M_{k-1}) + w_{a_{k-2}}$.

From the definition of switching strategy in (2.2.1) we know that

$$s_{k-2} = -\frac{1}{\alpha + q_{k-2}} \ln\big(1 - \frac{p_{k-2}}{\sum_{l=k-2}^{m} p_l}\big).$$

Notice that for $< a_{k-2}, a, b, B >$ and $< a_{k-2}, b, a, B >$, s_{k-2} are the same. Also, $r_{k-2}, q_{k-2}, m_{a_{k-2}}$ and $w_{a_{k-2}}$ are the same. Hence, the larger W_{k-1} is, the larger W_{k-2} is. By principle of induction, we can deduce that if $W_{<a,b,B>} > W_{<b,a,B>}$, $W_{<A,a,b,B>} > W_{<A,b,a,B>}$.

Theorem 4.3.1 *The difference between $W_{<a,b,B>}$ and $W_{<b,a,B>}$ is as follows,*

$$W_{<a,b,B>} - W_{<b,a,B>} = \frac{2r_b^2/q_b^2\big[\frac{H_1(0,\rho)-H_1(p_a,\rho)}{1-P} + \frac{1-K}{(1-P)^2}(H_2(0,\rho) - H_2(p_b,\rho))\big]}{p_B + p_a + p_b}.$$

where

$$\begin{aligned}
H_1(u,\rho) &= ((1-Q)p_B - u(\rho-1))\ln(1 - \frac{p_b}{p_B + p_b + u}),\\
H_2(u,\rho) &= ((Qp_B + u)\rho - (p_B + u)\rho^2)\ln(1 - \frac{p_a}{p_B + p_a + u}),\\
\rho &= \frac{r_a/q_a + m_a}{r_b/q_b + m_b}, Q = \frac{M_{k+1}}{r_b/q_b + m_b}, K = \frac{m_a}{r_a/q_a + m_a}, P = \frac{m_b}{r_b/q_b + m_b}, p_B = \sum_{j\in B} p_j
\end{aligned}$$

Proof. For the subsequence $< b, B >$

$$\begin{aligned}
M_k &= \big(\frac{r_b}{q_b} + m_b\big)\frac{p_b}{p_b + p_B} + M_{k+1}\big(1 - \frac{p_b}{p_b + p_B}\big),\\
W_k &= \big(2\ln\big(1 - \frac{p_b}{p_b + p_B}\big)\frac{r_b}{q_b}\big(\frac{r_b}{q_b} + m_b - M_{k+1}\big) + W_{k+1}\big)\big(1 - \frac{p_b}{p_b + p_B}\big)\\
&\quad + \big(2\frac{r_b}{q_b}\big(m_b + \frac{r_b}{q_b}\big) + w_b\big)\frac{p_b}{p_b + p_B}.
\end{aligned}$$

First and second moment of the subsequence $< a, b, B >$ are

$$
\begin{aligned}
M_{<a,b,B>} &= \frac{p_B}{p_a+p_b+p_B}M_{k+1} + \frac{p_a}{p_a+p_b+p_B}\big(\frac{r_a}{q_a}+m_a\big) + \frac{p_b}{p_a+p_b+p_B}\big(\frac{r_b}{q_b}+m_b\big),\\
W_{<a,b,B>} &= \frac{1}{p_B+p_a+p_b}\big\{2\ln\big(1-\frac{p_a}{p_B+p_a+p_b}\big)\\
&\quad \frac{r_a}{q_a}\big(\big(\big(\frac{r_a}{q_a}+m_a\big)-\big(\frac{r_b}{q_b}+m_b\big)\big)p_b + \big(\frac{r_a}{q_a}+m_a-M_{k+1}\big)p_B\big) +\\
&\quad 2\ln\big(\frac{p_B}{p_B+p_b}\big)\frac{r_b}{q_b}\big(\frac{r_b}{q_b}+m_b-M_{k+1}\big)p_B + \big(w_a+2\frac{r_a}{q_a}\big(\frac{r_a}{q_a}+m_a\big)\big)\\
&\quad p_a + \big(w_b+2\frac{r_b}{q_b}\big(\frac{r_b}{q_b}+m_b\big)\big)p_b + W_{k+1}p_B\big\}.
\end{aligned}
$$

To get $W_{<b,a,B>}$ we only need to swap a and b in $W_{<a,b,B>}$.

The difference between the two is: $W_{<a,b,B>} - W_{<b,a,B>} = \frac{2}{q_a^2 q_b^2 p_B}\Bigg($ $\ln\big(1-\frac{p_a}{p_B+p_a+p_b}\big)$
$\frac{r_a}{q_a}\big[\big(\big(\frac{r_a}{q_a}+m_a\big)-\big(\frac{r_b}{q_b}+m_b\big)\big)p_b+\big(\frac{r_a}{q_a}+m_a-M_{k+1}\big)p_B\big] + \ln$
$\big(\frac{p_B}{p_B+p_b}\big)\frac{r_b}{q_b}\big(\frac{r_b}{q_b}+m_b-M_{k+1}\big)p_B - \ln\big(1-\frac{p_b}{p_B+p_a+p_b}\big)$
$\frac{r_b}{q_b}\big[\big(\big(\frac{r_b}{q_b}+m_b\big)-\big(\frac{r_a}{q_a}+m_a\big)\big)p_a+\big(\frac{r_b}{q_b}+m_b-M_{k+1}\big)p_B\big] -$
$\ln\big(\frac{p_B}{p_B+p_a}\big)\frac{r_a}{q_a}\big(\frac{r_a}{q_a}+m_a-M_{k+1}\big)p_B\Bigg)$

Outside the big parenthesis is positive. So we only need to consider the expression inside the parenthesis, i.e.,

$$
\begin{aligned}
&\ln\big(1-\frac{p_a}{p_B+p_a+p_b}\big)\frac{r_a}{q_a}\big[\big(\big(\frac{r_a}{q_a}+m_a\big)-\big(\frac{r_b}{q_b}+m_b\big)\big)p_b+\big(\frac{r_a}{q_a}+m_a-M_{k+1}\big)p_B\big]\\
&+\ln\big(\frac{p_B}{p_B+p_b}\big)\frac{r_b}{q_b}\big(\frac{r_b}{q_b}+m_b-M_{k+1}\big)p_B - \ln\big(1-\frac{p_b}{p_B+p_a+p_b}\big)\frac{r_b}{q_b}\big[\big(\big(\frac{r_b}{q_b}+m_b\big)\\
&-\big(\frac{r_a}{q_a}+m_a\big)\big)p_a+\big(\frac{r_b}{q_b}+m_b-M_{k+1}\big)p_B\big] - \ln\big(\frac{p_B}{p_B+p_a}\big)\frac{r_a}{q_a}\big(\frac{r_a}{q_a}+m_a-M_{k+1}\big)p_B.
\end{aligned}
\tag{3.1}
$$

Dividing the above equation by $q_a^2 r_b^2$ and substituting the variables, we get
$\frac{(1-Q)p_B}{1-P}\ln(1-\frac{p_b}{p_B+p_b}) - \frac{((1-Q)p_B-p_a(\rho-1))}{1-P}\ln(1-\frac{p_b}{p_B+p_b+p_a}) +$
$\frac{1-K}{(1-P)^2}(Qp_B\rho - p_B\rho^2)\ln(\frac{p_B}{p_B+p_a}) - \frac{1-K}{(1-P)^2}((Qp_B+p_b)\rho-(p_B+p_b)\rho^2)\ln(1-$
$\frac{p_a}{p_B+p_a+p_b})$

Further let

$$H_1(u,\rho) = ((1-Q)p_B - u(\rho-1))\ln(1-\frac{p_b}{p_B+p_b+u}), \tag{4.3.2}$$

$$H_2(u,\rho) = ((Qp_B+u)\rho - (p_B+u)\rho^2)\ln(1-\frac{p_a}{p_B+p_a+u}). \tag{4.3.3}$$

We get

$$W_{<a,b,B>} - W_{<b,a,B>} = \frac{2r_b^2/q_b^2\left[\frac{H_1(0,\rho)-H_1(p_a,\rho)}{1-P} + \frac{1-K}{(1-P)^2}(H_2(0,\rho)-H_2(p_b,\rho))\right]}{p_B+p_a+p_b}.$$

Among the policies we are interested in the following three special policies: *randomized policy*, *indexed switching strategy* and the *minimum-variance policy*.

We comment that unlike the two-action case where the indexed switching strategy has the smallest variance, for general cases the indexed switching strategy may not have the smallest variance. A counterexample is given below:

Example 4.3.1 *Consider a continuous-time MDP. At the initial state there are three actions a_1, a_2 and a_3. The corresponding reward rates are 1, 5 and 9. Jump intensities are all 1. The expected total instant rewards are 10, 3, and 8. The second moment of the instant rewards are 150, 10 and 100. The indices are:*

$a_1 : 1/1 + 10 = 11$

$a_2 : 5/1 + 3 = 8$

$a_3 : 9/1 + 8 = 17$

So the indexed switching strategy is $< a_3, a_1, a_2 >$. However, the calculated variance for this switching strategy is 173.90, while for the policy $< a_2, a_1, a_3 >$, the variance is 147.45. This example shows that the indexed switching strategy may not be the best among all the policies.

However, the the indexed switching strategy does outperform the randomized policy. In other words, the indexed switching strategy has a smaller variance than the randomized policy.

4.4 Indexed switching strategy outperform randomized policy

Theorem 4.4.1 *The indexed switching strategy has a smaller variance than the randomized policy.*

Proof. We will show by induction over the number of actions in the action set. Consider adding actions one by one in the descending order of $\frac{r_k}{q_k} + m_k$, to the front of the existing action sequence. In other words, the first action (a_1) to be added has the smallest value of $\frac{r_k}{q_k} + m_k$.

Step 1: When $m = 2$, from Theorem (4.2.1), the descending-ordered switching strategy has smaller variance than the randomized policy.

Step 2: Suppose when $m = k$, the descending-ordered switching strategy has a smaller variance than the randomized policy. Consider $m = k + 1$. We add a new action a to the front of the action sequence. Note that action a has the largest value of $r_k/q_k + m_k$ among the $k + 1$ actions.

Let $p_\Sigma = \sum_{j=1}^{k} p_j$. Let $Ms^{(k)}$ and $Ws^{(k)}$ be the first and second moment of the ordered switching strategy with k actions, then from Theorem (4.1.1), we have

$$\begin{aligned} Ms^{(k+1)} &= \big(m_a + \frac{r_a}{q_a}\big)\frac{p_a}{p_\Sigma + p_a} + Ms^{(k)}\big(1 - \frac{p_a}{p_\Sigma + p_a}\big), \\ Ws^{(k+1)} &= 2\ln\big(1 - \frac{p_a}{p_\Sigma + p_a}\big)\frac{r_a}{q_a}\big(\frac{r_a}{q_a} + m_a - Ms^{(k)}\big)\big(1 - \frac{p_a}{p_\Sigma + p_a}\big) + \\ &\quad 2\frac{r_a}{q_a}\big(\frac{r_a}{q_a} + m_a\big)\frac{p_a}{p_\Sigma + p_a} + Ws^{(k)}\big(1 - \frac{p_a}{p_\Sigma + p_a}\big) + w_a\frac{p_a}{p_\Sigma + p_a}. \end{aligned}$$

Let $Mr^{(k)}$ and $Wr^{(k)}$ be the first and second moment of the randomized

policy with k actions, then:

$$\begin{aligned} Mr^{(k+1)} &= Mr^{(k)}\frac{p_\Sigma}{p_\Sigma+p_a} + \big(m_a + \frac{r_a}{q_a}\big)\frac{p_a}{p_\Sigma+p_a}, \\ Wr^{(k+1)} &= Wr^{k}\frac{p_\Sigma}{p_\Sigma+p_a} + \big(\frac{2m_a r_a}{q_a} + \frac{2r_a^2}{q_a^2} + w_a\big)\frac{p_a}{p_\Sigma+p_a}. \end{aligned}$$

From Feinberg [13] we know that a switching strategy has the same first moment of discounted rewards as the randomized policy. Now we consider the difference between $Ws^{(k+1)}$ and $Wr^{(k+1)}$:

$$\begin{aligned} Ws^{(k+1)} - Wr^{(k+1)} &= \frac{p_\Sigma}{(p_\Sigma+p_a)^2}\Bigg(2\ln\big(\frac{p_\Sigma}{p_\Sigma+p_a}\big)\frac{r_a}{q_a}\big(\frac{r_a}{q_a} + m_a - Ms^{(k)}\big)(p_\Sigma+p_a) + \\ &\quad (Ws^{(k)} - Wr^{(k)})(p_\Sigma+p_a) + \\ &\quad 2(Mr^{(k)} - Ms^{(k)})q_a x_a\big(\frac{r_a}{q_a} + M_a - Mr^{(k)} - Ms^{(k)}\big)\Bigg) \\ &= \frac{p_\Sigma}{p_\Sigma+p_a}\Bigg(2\frac{r_a}{q_a}\big(\frac{r_a}{q_a} + m_a - Ms^{(k)}\big)\ln\big(\frac{p_\Sigma}{p_\Sigma+p_a}\big) + (Ws^{(k)} - Wr^{(k)})\Bigg). \end{aligned}$$

It is easy to see that $\frac{r_a}{q_a} + m_a > Ms^{(k)}$. On the other hand, $\ln\big(\frac{p_\Sigma}{p_\Sigma+p_a}\big) < 0$, so the first term is negative. By assumption of induction, $Ws^{(k)} < Wr^{(k)}$, so, the second term is also negative. Therefore, the whole equation evaluates negative, and we have $Ws^{(k)} < Wr^{(k)}$ for all k.

4.5 Results when instant rewards are zero

In this section we consider a special case where the instant rewards are zero.

Theorem 4.5.1 *When the instant rewards are zero, the indexed switching strategy has the smallest variance among all the switching strategies and thus is the best switching strategy.*

Proof. In equation (4.3.1), let $m_a = m_b = w_a = w_b = 0$. ρ is reduced to $\frac{r_a/q_a}{r_b/q_b} = \frac{r_a q_b}{r_b q_a}$. Divide the equation by $(q_a r_b)^2$, we get:

$G(\rho) = (\rho-1)\big(\rho p_b \ln\big[1-\frac{p_a}{p_a+p_b+p_B}\big] + q_a x_a \ln\big[1-\frac{p_b}{p_a+p_b+p_B}\big]\big) + p_B\big(\rho^2\big(\frac{M_{k+1}}{r_a/q_a}-1\big)\ln\big[1-\frac{p_a}{p_a+p_B}\big] - \big(\rho\frac{M_{k+1}}{r_a/q_a}-1\big)\ln\big[1-\frac{p_b}{p_b+p_B}\big] - \rho^2\big(\frac{M_{k+1}}{r_a/q_a}-1\big)\ln\big[1-\frac{p_a}{p_a+p_b+p_B}\big] + \big(\rho\frac{M_{k+1}}{r_a/q_a}-1\big)\ln\big[1-\frac{p_b}{p_a+p_b+p_B}\big]\big)$.

For convenience, we make the following substitutions:

$A_0 = p_b \ln\big[1-\frac{p_a}{p_a+p_b+p_B}\big]$, $B_0 = p_a \ln\big[1-\frac{p_b}{p_a+p_b+p_B}\big]$,

$C_0 = \ln\big[1-\frac{p_a}{p_a+p_B}\big]$, $D_0 = \ln\big[1-\frac{p_b}{p_b+p_B}\big]$,

$E_0 = \ln\big[1-\frac{p_a}{p_a+p_b+p_B}\big]$, $F_0 = \ln\big[1-\frac{p_b}{p_a+p_b+p_B}\big]$, $Q_0 = \frac{M_{k+1}}{r_a/q_a}$

$A_1 = (A_0 + p_B(C_0 - E_0)(Q_0 - 1))$

$B_1 = -(A_0 - B_0 - Q_0(F_0 - D_0)p_B)$

$C_1 = p_B(D_0 - F_0) - B_0$

Then $G(\rho)$ can be reduced to a quadratic form:

$G(\rho) = A_1\rho^2 + B_1\rho + C_1$

Notice that $A_1 + B_1 + C_1 = 0$, so ρ =1 is one of G's zeros. Before we continue with the proof of Theorem 4.5.1, we first prove the following two results about the coefficients C_1 and A_1.

Lemma 4.5.1 *$C_1 > 0$, where $C_1 = p_B(D_0 - F_0) - B_0$, is defined above.*

Proof. We rewrite C_1 as:

$C_1 = \ln\big[1-\frac{p_b}{p_b+p_B}\big]p_B - \ln\big[1-\frac{p_b}{p_a+p_b+p_B}\big](p_a+p_B)$

Consider the following function $f(x)$:

$f(x) = \ln\big[1-\frac{p_b}{x+p_b+p_B}\big](x+p_B)$,

$f'(x) = \frac{\big(1+\ln\big[\frac{x+p_B}{x+p_b+p_B}\big]\big)p_b + \ln\big[\frac{x+p_B}{x+p_b+p_B}\big](x+p_B)}{x+p_b+p_B}$,

$f''(x) = \frac{p_b^2}{(x+p_B)(x+p_b+p_B)^2}$.

It is seen that f'' is always positive, so f' is increasing. At the same time, the limit of f' is 0 as t goes to infinity. So f' is negative for all x. So f is decreasing. We know that when x= 0, $C_1 = f(0) - f(0) = 0$. So for $p_a > 0$, $C_1 = f(0) - f(p_a) > 0$.

Lemma 4.5.2 *$A_1 < 0$, where $A_1 = (A_0 + p_B(C_0 - E_0)(Q_0 - 1))$, is defined above.*

Proof. We rewrite A_1 as:

$A_1 = (-1+Q_0)p_B \ln\big[1-\frac{p_a}{p_a+p_B}\big] - \ln\big[1-\frac{p_a}{p_a+p_b+p_B}\big]((-1+Q_0)p_B - p_b)$

Let $Z = Q_0 - 1$ and consider the following function $g(x)$:

$g(x) = \ln\big[1-\frac{p_a}{x+p_a+p_B}\big](-x+Zp_B)$,

$g'(x) = -\ln\left[\frac{x+p_B}{x+p_a+p_B}\right] - \frac{p_a(x-Zp_B)}{(x+p_B)(x+p_a+p_B)}$,
$g''(x) = -\frac{p_a(2(1+Z)p_B(x+p_B)+p_a(x+(2+Z)p_B))}{(x+p_B)^2(x+p_a+p_B)^2}$.

It is seen that g'' is always negative, so g' is decreasing. At the same time, the limit of g' at infinity is 0. So g' is positive for all x. So g is increasing. We know that when $x = 0$, $A_1 = g(0) - g(0) = 0$. So for $p_b > 0$, $A_1 = g(0) - g(p_b) < 0$.

Now let's go back to the proof of Theorem 4.5.1. From Lemma 4.5.1 and 4.5.2 we know that G_0 is a quadratic function concave downwards and one of its zeros is 1 and the other is negative. So it is straightforward that when $0 < \rho < 1$, $G(\rho) > 0$, and when $\rho > 1$, $G(\rho) < 0$. In other words, if $\frac{r_a}{q_a} > \frac{r_b}{q_b}$, $W_{<a,b,B>} < W_{<b,a,B>}$, and executing actions with a larger "action index" first will reduce the variance of rewards up to the first jump.

We remark that for rewards beyond the first jump the index type switching policy may not have a smaller variance. A counterexample is given below.

Example 4.5.1 *Consider an MDP with four states: states* $\{1, 2, 3\}$ *are regular states and state* 0 *is an absorbing state; see Figure 2.*

At state 1*, two actions,* a *and* b*, are available. The probability to take action* a *and* b *are* p_a *and* p_b *respectively. If* a *is taken, the process earns rewards at a rate of* r_a*, transits to state 2 deterministically and earn an instant reward of* R_a *after an exponential sojourn time with the intensity* q_a*.*

If b *is taken, the process earns rewards at a rate of* r_b*, transits to state 3 deterministically and earn an instant reward of* R_b *after an exponential sojourn time with the intensity* q_b*.*

At state 2 and 3 only one action is available. At state 2 the reward rate is r_2 *and the process transits to the absorbing state after an exponential sojourn time with the intensity* q_2*.*

At state 3, the reward rate is r_3 *and the process transits to the absorbing state after an exponential sojourn time with the intensity* q_3*. Once in the absorbing state, the process stops. The discount factor* $\alpha = 0$*.*

We consider a switching policy $\phi =< a, b >$ *and a randomized policy* $\sigma = \{a, b\}$. *For the switching policy* ϕ, *the switching epoch* S_a *is*

$S_a = -\ln(1 - p_a)/q_a$

Let T_a *be the exponential random variable with intensity* q_a. *Conditioning on whether* $T_a < S_a$ *or not the second moment of the total reward is*

$W_\phi = \int_0^{S_a} q_a e^{-q_a T_a} \left[(r_a T_a + R_a + \frac{r_2}{q_2})^2 + (\frac{r_2}{q_2})^2 \right] dT_a$

$+ P\{T_a > S_a\} \left[(r_a S_a + \frac{r_b}{q_b} + R_b + \frac{r_3}{q_3})^2 + (\frac{r_b}{q_b})^2 + (\frac{r_3}{q_3})^2) \right]$

For the randomized policy σ, *the second moment is*

$W_\sigma = p_a \left[(\frac{r_a}{q_a})^2 + (\frac{r_2}{q_2})^2 + (\frac{r_a}{q_a} + R_a + \frac{r_2}{q_2})^2) \right] + (1 - p_a) \left[(\frac{r_b}{q_b})^2 + (\frac{r_3}{q_3})^2 + (\frac{r_b}{q_b} + R_b + \frac{r_3}{q_3})^2 \right]$

The difference of the two is

$W_\phi - W_\sigma = -2(1 - p_a)(\frac{r_a}{q_a}) \left[(\frac{r_b}{q_b} + R_b + \frac{r_3}{q_3}) - (\frac{r_a}{q_a} + R_a + \frac{r_2}{q_2}) \right) \ln(1 - p_a)$

Note that $-2(1 - p_a)(\frac{r_a}{q_a}) < 0$ *and* $\ln(1 - p_a) < 0$. *Therefore the difference of* W_ϕ *and* W_σ *depends only on* $(\frac{r_b}{q_b} + R_b + \frac{r_3}{q_3}) - (\frac{r_a}{q_a} + R_a + \frac{r_2}{q_2})$, *which depends not only on the index defined above but also on the value of* $\frac{r_3}{q_3}$ *and* $\frac{r_2}{q_2}$.

By appropriately choosing $\frac{r_3}{q_3}$ *and* $\frac{r_2}{q_2}$ *we can make the indexed switching strategy arbitrarily worse than the randomized policy.*

4.6 Counterexamples for infinite horizon

We have shown in Theorem 4.4.1 that the indexed switching strategy has a smaller variance than the randomized policy, for rewards up to the first jump. However, for rewards beyond the first jump, in particular for infinite horizon, the randomized policy may outperform any switching strategy. In this section we give an example to illustrate this fact.

Example 4.6.1 *Consider a simple MDP with two states: 1 and 2. At state*

1, it either takes action a with probability p_a or actions b with probability p_b. If action a is taken, the reward rate is r_a, the instant reward at jump is R_a and the jump intensity is q_a. If action b is taken, the reward rate is r_b, the instant reward at jump is R_b and the jump intensity is q_b. At state 2, both the reward rate and instant reward at jump are zero and the jump intensity is q_2.

Consider

$r_a = 0.1, q_a = 0.9, p_a = 0.75, R_a = 0$

$r_b = 0.9, q_b = 0.7, p_b = 0.25, R_b = 0$

Discount factor $\alpha = 0.05$

Consider the randomized policy $\{a, b\}$ and two switching strategies $< a, b >$ and $< b, a >$. Due to the difficulty in computing the variances analytically we compute the variances (standard deviations) of the three policies through numerical simulation.

For a switching policy either $< a, b >$ or $< b, a >$ we compute the changing epoch using the formula defined in (2.2.1). As the matter of fact since there are only two actions the formulas can be simplified into

$$s_a = \frac{-1}{q_a + \alpha} \ln(1 - p_a),$$

$$s_b = \frac{-1}{q_b + \alpha} \ln(1 - p_b).$$

We need to generate an exponential random variable $T \sim exp(\alpha)$, which represents the life time of the process. We keep tracking the time elapsed since $t = 0$, denoted as t. If $t > T$ the process stops. Otherwise the simulation continues.

Let U represent the total discounted rewards earned for the infinite horizon under the probabilistic discounting. The following algorithm outlines the simulation steps described above:

Simulation scheme of switching policy $< a, b >$

Step 0: $t = 0, U = 0$, generate $T \sim exp(\alpha)$, compute s_a.

Step 1: Generate $t_a \sim \exp(q_a)$.

Case 1: If $t_a < s_a$:

Case 1.1: If $t + t_a < T$, compute $U = U + r_a t_a + R_a, t = t + t_a$.

Case 1.2: If $t + t_a > T$, compute $U = U + r_a(T - t)$ and stop.

Case 2: If $t_a > s_a$, generate $t_b \sim exp(q_b)$:

Case 2.1: If $t + s_a + t_b < T$, compute $U = U + r_a s_a + r_b t_b + R_b$ and $t = t + s_a + t_b$.

Case 2.2: If $t + s_a + t_b > T$, compute $U = U + r_a s_a + r_b(T - t - s_a)$ and stop.

Generate $z \sim exp(q_2)$, update $t = t + z$ and goto step 1.

Simulation scheme of randomized policy $\{a, b\}$

Step 0: $t_0 = 0, U = 0$, generate $T \sim exp(\alpha)$, compute s_a.

Step 1: Generate a uniform random number u:

Case 1: If $u < p_a$, generate $t_a \sim \exp(q_a)$:

Case 1.1: If $t + t_a < T$, compute $U = U + r_a t_a + R_a, t = t + t_a$.

Case 1.2: If $t + t_a > T$, compute $U = U + r_a(T - t)$ and stop.

Case 2: If $u > p_a$, generate $t_b \sim \exp(q_b)$:

Case 2.1: If $t + t_b < T$, compute $U = U + r_b t_b + R_b, t = t + t_b$.

Case 2.2: If $t + t_b > T$, compute $U = U + r_b(T - t)$ and stop.

Generate $z \sim exp(q_2)$, update $t = t + z$ and goto step 1.

Simulation results

We perform 10,000 trials in each simulation and for each trial the discounted total rewards obtained from applying the three policies are recorded. We then compute the sample standard deviations from the 10,000 samples. To see how good the simulated variances are close from simulation to simulation we ran the simulation 10 times and the results are shown in the table below:

Table 3.1 Simulated Standard Deviations

switching $< a, b >$	switching $< b, a >$	randomized $\{a, b\}$
7.2946	7.2168	6.5688
7.1553	6.9924	6.3554
7.3725	7.2798	6.6238
7.2775	7.2490	6.5315
7.3648	7.1746	6.5520
7.4409	7.2929	6.6385
7.2925	7.2515	6.5354
7.3518	7.1738	6.5701
7.3849	7.2655	6.6130
7.4354	7.2759	6.6596

We can see that the numerical simulation has exhibited quite good stability. The standard deviation from applying the randomized policy $\{a, b\}$ is less than either of the switching policies. This example shows that for problems beyond the first jump and in particular for the infinite horizon problems even the best switching policy may not outperform the randomized policy.

Chapter 5

Variance under multiplicative discounting

In this chapter we briefly discuss the variance under the multiplicative discounting. When the continuously compound rate is α, one unit of reward at time t is worth $e^{-\alpha t}$ at $t = 0$.

The computation of variances under the multiplicative discount is much more complicated than under the probabilistic discounting. In Appendix B we derive the simultaneous equations that can be used to solve for the first and second moments and therefore compute the variances.

Unlike the probabilistic discounting where the indexed policy has a smaller variance than the randomized policy under the multiplicative discounting even the best switching strategy may not have a smaller variance than the randomized policy. We give a counterexample below to illustrate this.

Example 5.0.2 *Consider two actions* $\{a, b\}$ *at state* 0*:*

Reward rates: $r_a = 8, r_b = 4$.

Instant rewards: $R_a = R_b = 0$.

Transition rates: $q_a = 1, q_b = 1$.

Probabilities of taking action a *and* b *at state* 0*:* $p_a = 0.1, p_b = 0.4$.

Consider the discount rewards up to the first jump. Direct calculation using

(B.2.1) shows that when $\alpha \leq 0.02$*, the switching strategy* $< a, b >$ *has a smaller variance than the randomized policy and the other switching strategy* $< b, a >$*. However, when* $\alpha \geq 0.045$*, the randomized policy outperforms both switching strategies. For example, when* $\alpha = 0.05$*, the variances of switching strategy* $< a, b >, < b, a >$ *and the randomized policy* $\{a, b\}$ *are 24.543, 29.978, and 22.165, respectively.*

Chapter 6

Application to dynamic power management

6.1 Introduction

In [46] Dynamic Power Management (DPM) for portable electronic systems was formulated as a constrained continuous-time MDP problem. A linear programming approach that minimizes the average cost was proposed. By solving the LP the authors obtained an optimal randomized policy. However, in practice it is hard to implement a randomized policy so they tried to search for the "best" nonrandomized stationary policy using either a nonlinear programming approach or a heuristic policy iteration. Finding such a policy is an NP-hard problem; see [11]. In addition, it typically has worse performance than the optimal randomized policy or may not exist even for some feasible problems [12, 13].

In [12, 13], another form of optimal policy was proposed – the so-called switching stationary strategy. The proposed strategy has two advantages compared with the "best" nonrandomized policies generated by the NLP procedure or iterative algorithm in [46]. First, it yields better performance than the "best" nonrandomized policy. This makes sense because the switching stationary strategy has the same performance as the optimal randomized policy while the "best" nonrandomized policies were at most at good as the optimal randomized policy. Second, the computation of the switching stationary

strategy is much simpler than finding the "best" nonrandomized policy. The former is P-hard while the latter is NP-hard.

The rest of this chapter is organized as follows. In section 2 we give a brief review of the continuous-time MDP model proposed in [46]. Section 3 describes how to construct the optimal switching stationary policy. Section 4 gives some numerical results compared with results obtained in [46].

6.2 Model review

Typically in a dynamic power management system there are four components: service provider (SP), high-priority service queue (HSQ), low-priority service queue (LSQ), and service requester (SR). The SR generates service requests for the SP. The SQ buffers the service requests. The SP provides service to the requests in a top-down manner. The PM monitors the states of the SR, SQ, and SP and issues state-transition commands to the SP.

The relationships between the HSQ and LSQ are:

1. Requests in the HSQ have a smaller waiting time than those in the LSQ.

2. The SP will not start serving the requests in the LSQ until it finishes all the requests in the HSQ. Therefore, the service rate of the LSQ is a function not only of the state of SP but also of the state of HSQ.

There are three states for the SP: busy, idle and sleep. As a simple example, the capacity of HSQ and LSQ are 2 and 3, respectively. By defining the number of requests in the service queue as the state, there are 3 states for HSQ: 0, 1 and 2 requests. Similarly, there are 4 states for LSQ: 0, 1, 2 and 3 requests. There are two types of requests generated by SR: high priority request or low priority request. The former demands shorter response time. Each component is modelled as a separate continuous-time Markov Decision Process. By combining the four components we have a joint process with 72 ($= 3 \times 3 \times 4 \times 2$) states.

In [46] the HSQ and LSQ processes are modelled together as a SQ model. The state set of the SQ is given by $Q = Q_{LSQ} \times Q_{HSQ}$, and the generator matrix is given by $G_{SQ}(s, r) = G_{LSQ}(s, r) \bigoplus G_{HSQ}(s, r, hq)$, where s is the state of SP, r is the state of SR, and the $\bigoplus$ operator is the tensor sum defined in [46, Definition 4.1].

The processes of SP and SQ are modelled into one joint process as well. The generator matrix is denoted as G_{SP-SQ}. The generator matrix of the system is then

$$G_{SYS}(a) = G_{SP-SQ}(a, r) \bigoplus G_{SR}$$

where a is a action taken by SP.

In a power-managed system the trade-off between operational performance and power consumption is the key. The goal is always to achieve as little power consumption as possible while keeping the delay of processing the requests under some tolerable level. In terms of a controlled MDP, this becomes a constrained Markov Decision Process, in which we have three criteria: minimizing the overall average power consumption, minimizing the delay for high-priority queue, and minimizing the delay for low-priority queue. A typical method to deal with it is to target one criterion by restricting the rest within certain levels. The model is reflected in LP2 in [46] and is restated as follows

$$\begin{array}{l} \text{Min} \sum_{i\in S}\sum_{a\in A(i)} \text{c\_pow}_{ia} x_{ia}, \\ \text{s.t.} \sum_{a\in A(i)} x_{ia} - \sum_{j\in S}\sum_{a\in A(j)} p(i,j,a)x_{ja} = 0, \text{for any} i \in S \\ \sum_{i\in S}\sum_{a\in A(i)} x_{ia} = 1 \\ \sum_{i\in S}\sum_{a\in A(i)} \text{c\_hsq}_{ia} x_{ia} \leq D_H \\ \sum_{i\in S}\sum_{a\in A(i)} \text{c\_lsq}_{ia} x_{ia} \leq D_L \\ x_{ia} \geq 0 \text{for all } i \in S, a \in A(i) \end{array} \tag{6.2.1}$$

where $\text{c\_pow}_{ia}, \text{c\_hsq}_{ia}, \text{c\_lsq}_{ia}$ are the power consumption of the system, the delay cost of the high-priority request, and the delay cost of the low-priority request during the time the system stays in state i and action a is taken. The formulas to calculate them are as follows:

$$\text{c\_pow}_{ia} = \text{pow}_i \tau_{ia} + \sum_j \text{ene}_{ij} p(i,j,a) \tag{6.2.2}$$

$$\text{c\_hsq}_{ia} = \text{hq}_i \tau_{ia} \tag{6.2.3}$$

$$\text{c\_lsq}_{ia} = \text{lq}_i \tau_{ia} \tag{6.2.4}$$

where:

pow_i is the power at state i in W;

ene_{ij} is the switching energy from state i to state j, in Joule;

$p(i,j,a)$ is the transition probability from state i to state j when action a is taken;

hq_i is the number of requests in HSQ;

lq_i is the number of requests in LSQ;

τ_{ia} is the expectation of the time that the system will be in state i if action a is chosen in this state.

We remark that x_{ia} in the LP model represents the long-run fraction of

time that the system spends in state i and when action a is taken. They are state-action probabilities defined in [12].

6.3 Switching stationary strategy

In the formulated problem there are no absorbing states, and the modelling guarantees the ergodicity of the Markov chain. [12, Theorem 2.1] guarantees the existence of the optimal switching stationary policy if the above LP is feasible.

Let $X = \{x_{ia} : i \in S, a \in A(i)\}$ be a feasible solution to the above LP. The associated randomized stationary policy σ is defined as follows

$$\begin{aligned} \sigma(a'|i) &= \frac{q(i,a')x_{ia'}}{\sum_{a\in A(i)} q(i,a)x_{ia}}, \text{if} \sum_{a\in A(i)} q(i,a)x_{ia} > 0 \\ &= a, \text{otherwise} \end{aligned} \tag{6.3.1}$$

where a is an arbitrary element of $A(i)$.

Fixing any order of the action set at state i and applying formula (2.2.1) we will obtain the switching stationary strategy.

In the power management system utilizing the switching strategy the PM knows exactly when to switch to a different action, and the policy is deterministic. Compared with the randomized policy, it is easier to implement. Its performance in terms of expected power consumption and expected delay costs on HSQ and LSQ is as good as the optimal randomized policy.

6.4 Numerical results

Using exactly the same parameters in [46, Table IV], we compute the average power consumption satisfying the delay constraints for HSQ and LSQ. The results are summarized in Table 5.1. Note that the results under the column "LP-based CTMDP policy Power (mW)" were obtained through simulation and reported in [46]. The results under the column "Switching Stationary Strategy Power (mW)" are our theoretic results.

Table 5.1 Theoretical results using switching strategies

Delay for LSQ D_L (sec)	Delay for HSQ D_H (sec)	Policy from [46] Power (mW)	Switching Stationary Strategy Power (mW)	Reduction of Power ΔP (%)
0.399	0.143	0.942	0.927	1.59
0.232	0.118	1.156	0.995	13.93
0.183	0.104	1.865	1.107	45.47
0.151	0.255	1.067	0.989	7.31

In addition to the theoretical results we conducted simulation using our switching strategies. The results are shown in Table 5.2.

Table 5.2 Simulation results using switching strategies

Delay for LSQ D_L (sec)	Delay for HSQ D_H (sec)	Policy from [46] Power (mW)	Switching Stationary Strategy Power (mW)	Reduction of Power ΔP (%)
0.399	0.143	0.942	0.913	3.08
0.232	0.118	1.156	0.984	14.88
0.183	0.104	1.865	1.235	33.78
0.151	0.255	1.067	0.976	8.53

Table 5.1 and Table 5.2 show that use of switching stationary strategy may significantly reduce the power consumption.

Chapter 7

Concluding remarks and future work

In this dissertation we deal with two definitions of discounting: multiplicative discount and probabilistic discounting through stopping times. We have shown in the chapter that the variances under the multiplicative discounting is at most as large as that under the probabilistic discounting. We also give a condition for the equality to hold: if and only if $\mathrm{Var}(J_2|\mathcal{F}_\infty) = 0$ P-a.s. This is a very general condition. We feel that we can make it more specific and expressed in terms of r_t and R_n. We conjecture that the more specific condition for the equality to hold is $r_t = 0$ almost everywhere P-a.s. and $R_n = 0$ P-a.s. We will further investigate this question.

In the dissertation we primarily focus on the probabilistic discounting mainly and only give very brief discussion of multiplicative discounting. The major reason is due to the complexity with multiplicative discounting. If time permits it is possible to investigate the multiplicative discounting in more details. Although there is not an optimal switching strategy that has a smaller variance than the randomized policy uniformly for any α it is possible to show that there exists an optimal switching strategy that has a smaller variance than the randomized policy for sufficiently small α and this critical α depends on the primitive parameters of the MDP.

We focus on the discounted total rewards. In literature there has been a lot of research on average reward per unit time. It should be very interesting

and important to study the variance of the average reward per unit time. We will ask the same question: whether there is any difference in the randomized policy and the equivalent switching strategies and if so how we can find a policy that has the smaller variance of average reward per unit time.

In the literature people also consider some other nontraditional criteria such as probability criteria. This is especially important in the situation where system performance is controlled on single trial basis and high reliability is a must, e.g., the launch of spaceship. There have been some papers devoted to the probability criteria for various rewards, see [16, 17, 56, 60].

Another important direction of future work is to investigate the application of the results in business, finance, engineering and other applicable fields. For example, we can try to expand the sizes of HSQ and LSQ in the dynamic power management to see how it is suitable for practical power management in portable electronic devices. We can also investigate the application in revenue management and inventory control.

In the dissertation we limit our policy space to randomized stationary policy and its equivalent switching stationary strategies. It is possible to investigate the variance minimization problem under the general mean-variance framework for MDPs and search for optimal policies in the space of all stationary policies.

Bibliography

[1] Altman, E. and Shwartz, A. (1991) "Markov decision problems and state-action frequencies." SIAM J. Control Optim. **29**, 786–809.

[2] Bäuerle, B. (2005). "Benchmark and mean-variance problems for insurers." Math. Meth. Oper. Res. **62**, 159–165.

[3] Baykal-Gursoy, M. and Ross, K.W. (1992) "Variability sensitive Markov decision processes." Math. Oper. Res. **17**, 558–571.

[4] Baykal-Gürsoy, M. and Gürsoy, K. (2007). "Semi-markov decision processes: Nonstandard criteria." Probability in the Engineering and Informational Sciences, **21**, 635–657.

[5] Benito, F. (1982). "Calculating the variance in Markov-processes with random reward." Trabajos Estadst Investigacin Operat. **33**, 73–85.

[6] Chung, K.J. (1989). "A note on maximal mean/standard deviation ratio in an undiscounted MDP." Oper. Res. Lett. **8**, 201–203.

[7] Chung, K.J. (1994). "Mean-variance tradeoffs in an undiscounted MDP: the unichain case." Oper. Res. **42**, 184–188.

[8] Duffie, D. and H. Richardson (1991). "Mean-variance hedging in continuous time" Ann Appl Probab. **1**, 1–15.

[9] Elton, E.J., M.J. Gruber (1975). "Finance as a dynamic process." Prentice Hall, Englewood Cliffs.

[10] Feinberg, E.A. (1994). "A generalization of "expectation equals reciprocal of intensity" to nonstationary distributions," J. Appl. Probability, **31**, 262–267

[11] Feinberg, E.A. (2000). "Constrained discounted MDP & Hamiltonian cycles" Math. Oper. Res, **26**, 130–140.

[12] Feinberg, E.A. (2002). "Optimal control of average reward constrained continuous-time finite Markov decision processes" Proceedings of the 41st IEEE Conference on Decision and Control. **4**, 3805–3810

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

[14] Feinberg, E.A. and Fei, J. (2009). "Inequality for variances of the discounted rewards." J. Appl. Probability, **46**, to appear in December 2009.

[15] Feinberg, E.A. and Shwartz, A. (1996). "Constrained discounted dynamic programming." Math. Oper. Res. **21**, 922–945.

[16] Filar, J. (1983). "Percentiles and Markov decision processes." OR Letters. **2**, 13–15.

[17] Filar, J., Krass, D. and Ross, K.W. (1995). "Percentile performance criteria for limiting average Markov decision processes." IEEE Transactions on Automatic Control. **40**, 2–9.

[18] Filar, J., Kallenberg, L. C. M. and Lee, H.-M. (1989). "Variance penalized Markov decision processes." Math. Operat. Res. **14**, 147–161.

[19] Francis, J.C. (1976). "Investments: analysis and management" McGraw-Hill, New York.

[20] Fristedt, B. and Gray, L. (1997). "A Modern Approach to Probability Theory." Birkhäuser, Boston.

[21] Grauer,R.R. and N.H. Hakansson (1993) "On the use of mean-variance and quadratic approximations in implementing dynamic investment strategies: a comparison of returns and investment policies" Management Sci. **39**, 856–871.

[22] Hakansson,N.H. (1971). "Multi-period mean-variance analysis: Toward a general theory of portfolio choice." J. Finance. **26**, 857–884.

[23] Huang, Y. and Kallenberg, L. C. M. (1994). "On finding optimal policies for Markov decision chains: a unifying framework for mean-variance-tradeoffs." Math. Operat. Res. **19**, 434–448.

[24] Hwang, C.-H. and W, A. (1997). "A predictive system shutdown method for energy saving of event-driven computation." Proceedings of the International Conference on Computer Aided Design, 28–32, Nov. 1997

[25] Jacod, J. (1975). "Multivariate point processes: predictable projections, Radon-Nikodym derivatives, representation of martingales." Z. Wahr. verw. Geb. **31**, 235–253.

[26] Jaquette, S. C. (1972). "Markov decision processes with a new optimality criterion: small interest rates." Ann. Math. Statist. **43**, 1894–1901.

[27] Jaquette, S. C. (1973). "Markov decision processes with a new optimality criterion: discrete time." Ann. Statist. **1**, 496–505.

[28] Jaquette, S.C. (1975). "Markov decision processes with a new optimality criterion: Continuous time." Ann. Statist. **3**, 547–553.

[29] Jaquette, S. C. (1976). "A utility criterion for Markov decision processes." Manag. Sci. **23**, 43–49.

[30] Kadota, Y. (1997). "A minimum average-variance in Markov decision processes." Bull. Inf. Cybernet. **29**, 83–89.

[31] Kawai, H. (1987). "A variance minimization problem for a Markov decision process." Europ. J. Operat. Res. **31**, 140–145.

[32] Kitaev, Yu. M. (1985). "Semi-Markov and jump Markov controlled models: average cost criterion." SIAM Theory Probab. Appl. **30**, 272–288.

[33] Kitaev, Yu. M., Rykov, V. V. (1995). "Controlled Queueing Systems." CRC Press, New York.

[34] Kurano, M. (1987). "Markov decision processes with a minimum-variance criterion." J. Math. Anal. Appl. **123**, 572–583.

[35] Kushner, H. J. (1984). "Approximation and weak convergence methods for random processes, with applications to stochastic systems theory." MIT Press, Cambridge, MA.

[36] Liu, J. and Zhao, X, (2004) "On average reward semi-Markov decision processes with a general multichain structure." Math. Operat. Res. **29**, 339–352.

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

[38] Mandl, P. (1974). "Estimation and control in Markov chains." Adv. in Appl. Probab. **6**, 40–60.

[39] Markowitz, H. (1952). "Portfolio selection." Journal of Finance. **7**, 77–91.

[40] Miller, B. L. (1968). "Finite state continuous time Markov decision processes with a finite planning horizon." SIAM J. Control. **6**, 266–280.

[41] Miller, B. L. (1968). "Finite state continuous time Markov decision processes with an infinite planning horizon." J. Math. Anal. Appl. **22**, 552–569.

[42] Mossin J. (1968). "Optimal multiperiod portfolio policies" J. Business. **41**, 215–229.

[43] Parthasarathy,K. R. (1967). "Probability measures on metric spaces." Academic Press.

[44] Pliska, S. R. (1997) "Introduction to mathematical finance: discrete time models" Blackwell, Oxford

[45] Puterman, M.L (1994) "Markov Decision Processes: discrete stochastic dynamic programming." Wiley, New York.

[46] Qiu, Q.R.,Wu, Q. and Pedram, M. (2001) "Stochastic modeling of a power-managed system - construction and optimization." IEEE Transactions on Computer-Aided, Design of Integrated Circuits and Systems. 20, 1200–1217.

[47] Ross, S.M. (1983) "Stochastic Processes." Johns Wiley and Sons, New York.

[48] Sobel, M. J. (1982). "The variance of discounted Markov decision processes." J. Appl. Prob. **19**, 794–802.

[49] Sobel, M. J. (1985). "Maximal mean/standard deviation ratio in an undiscounted MDP." Operat. Res. Lett. **4**, 157–159.

[50] Sobel, M. J. (1994). "Mean-variance tradeoffs in an undiscounted MDP." Oper. Res. **42**, 175–183.

[51] Samuelson, P.A. (1969) "Lifetime portfolio selection by dynamic stochastic programming" Rev. Econ. Stat. **51**, 239–246.

[52] Shiryaev, A.N. (1996). "Probability." Second edition. Springer, New York.

[53] Srivastava, M, Chandrakasan, A. and Brodersen R. (1996) "Predictive system shutdown and other architectural techniques for energy effcient programmable computation." IEEE Transactions on VLSI Systems. **4**, 42–55.

[54] Van Dijk, N.M. and Sladký, K. (2006). "On the total reward variance for continuous-time Markov reward chains." J. Appl. Probab. **43**, 1044–1052.

[55] White, D. J. (1988). "Mean, variance and probability criteria in finite Markov decision processes: A review." J. Optimization Theory Appl. **56**, 1–29.

[56] White, D. J. (1993). "Minimizing a threshold probability in discounted Markov Decision Processes." J. Math. Anal. Appl. **173**, 634–646.

[57] White, D. J. (1995). "Mean-variance analysis in infinite horizon non-discounted Markov decision processes: technical note." J. Inform. Optim. Sci. **16**, 381–386.

[58] Xiong, J. and Zhou, X.Y. (2007). "Mean-variance portfolio selection under partial information." SIAM J. Control Optim. **46**, 156–175.

[59] Yushkevich, A. A. (1980). "On reducing a jump controllable Markov model to a model with discrete time." SIAM Theory Probab. Appl. **25**, 58–69.

[60] Zhang, W.Q., Jiang, Q.Y.Zhou and Lin Y.L (1983). "Reliability-constrained Markov Decision Programming and Penalty Factor Method." J. Tsinghua Univ. **23**, 61–71.

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

[62] Zhu, S.S., Li, D., and Wang, S.Y. (2004). "Risk control over bankruptcy in dynamic portfolio selection: A generalized-variance formulation." IEEE Trans. Automat. Control. **49**, 447–457.

Appendix A

Definition of strategies

In this appendix, we define strategies and various classes of strategies for Continuous-Time MDPs. Staring from Miller [40, 41], the literature on Continuous-Time MDPs usually deals only with Markov strategies. For these strategies decisions depend only on the current time and state. Markov strategies define the corresponding stochastic processes via Kolmogorov's backward equation. Yushkevich [59], Kitaev [32] Kitaev and Rykov [33], and Feinberg [10] considered past-dependent strategies.

In this paper we follow definitions from Feinberg [13], where a general control rule, for which the choice of actions depend on past states, past jump epoches, and the current state and time, was called *a strategy*, and a particular case of a strategy, for which the choice of actions depend only on the past states and the current state was called *a policy*. While strategies can change actions between jumps, policies cannot. The simplest subclass of policies is *stationary* policies, for which decisions depend only on the current state. It is also possible to consider randomized strategies defined below and randomized policies defined in Section 2.1.

We recall that a multivariate (also called marked) point process with the state space S is a stochastic sequence $\{T_n, X_n : n \geq 0\}$ such that $0 = T_0 \leq T_1 \leq \ldots \leq T_n \leq \ldots$, $X_n \in S$, $n \geq 0$, and there is a special state $x_\infty \in S$ such that $X_n = x_\infty$ if $T_n = \infty$. The times $T_1, T_2, \ldots$ are jump epoches and $X_0, X_1, \ldots$ are the states at epoches $t \in [T_n, T_{n+1})$, $n = 0, 1, \ldots$. Let $\xi_n = T_n - T_{n-1}$ for $n \geq 1$ and $T_\infty = \lim_{n\to\infty} T_n$. Then $X_0, T_1, \ldots, X_{n-1}, T_n, X_n, T$ is the history up to time $T < T_{n+1}$.

Strategies (or, equivalently, nonrandomized strategies).

Let $\Omega^* = \cup_{n\geq 0}(S \times [0, \infty))^n$ and let $\mathcal{F}^*$ be the Borel σ-field on Ω^* induced by the Borel σ-field on $[0, \infty)$). Consider the set of all finite histories $\Omega = \{(i_0, t_1, i_1, t_2 \ldots, i_n, t) : 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq t < \infty,\ n = 0, 1, \ldots\}$. Then Ω is a measurable subset of Ω^*. We denote by $\mathcal{F} = \{B \in \mathcal{F}^* | \, b \subset \Omega\}$ the

constriction of $\mathcal{F}^*$ to Ω.

A (nonrandomized) strategy ψ is a Borel mapping from Ω to A such that $\psi(i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) \in A(i_n)$ for each $(i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) \in \Omega$. For a strategy ψ and given the history $\omega_n = i_0, t_1 \ldots t_n, i_n, t$, the joint probability distribution that the jump happens during the interval $[t, t + dt)$ and $i_{n+1} = j$ is $q(i, j, \psi(i_0, t_1, \ldots, t_n, i_n, t)dt$. However, if $t_n = \infty$ then $t_{n+k} = \infty$ for all $k > 0$ and $x_{n+k} = x_\infty$ for all $k \geq 0$.

Let $\Omega_\infty = (S \times [0, \infty])^\infty$ and $\mathcal{F}_\infty$ be the Borel σ-field induced by the Borel σ-fields on $[0, \infty]$. According to Jacod [25, Lemma 3.3], each strategy ψ and each initial distribution μ of the initial state i_0 define a unique multivariate point process on $(\Omega_\infty, \mathcal{F}_\infty)$. We denote by P_μ^ψ and E_μ^ψ the probabilities and expectations for this process. We also write P_i^ψ and E_i^ψ instead of P_μ^ψ and P_μ^ψ when $\mu(i) = 1$ for some $i \in S$.

Randomized strategies. It is also possible to consider randomized strategies. These more general objects can be defined as (nonrandomized) strategies when actions are replaced with probability distributions on the action set A.

Let $\mathcal{P}(A)$ be the set of probability measures on the Borel space A. Consider the topology of weak convergence on $\mathcal{P}(A)$. Since A is a Polish (i.e., complete separable metric) space then $\mathcal{P}(A)$ is a Polish space; see Parthasarathy [43, Theorem 6.4, Chapter II]. Since $A(i)$ are compact subsets of the Polish space A, $\mathcal{P}(A(i))$ are compact subsets of $\mathcal{P}(A)$; [43, Theorem 6.4, Chapter II].

Consider a Continuous-Time MDP $\{S, \tilde{A}, \tilde{A}(\cdot), \tilde{q}, K, \tilde{r}_k, \tilde{R}_k\}$, $k = 0, \ldots, K\}$, where $\tilde{A} = \mathcal{P}(A)$, $\tilde{A}(i) = \mathcal{P}(A(i))$, and for $i \in S$ and $\tilde{a} \in \tilde{A}(i)$

$$\tilde{q}(i, \tilde{a}, j) = \int_{A(i)} q(i, a, j)\tilde{a}(da), \tag{A.0.1}$$

$$\tilde{r}_k(i, \tilde{a}) = \int_{A(i)} r_k(i, a)\tilde{a}(da), \qquad k = 0, 1, \ldots, K, \tag{A.0.2}$$

$$\tilde{R}_k(i, \tilde{a}, j) = \frac{\int_{A(i)} R(i, a, j)q(i, a, j)\tilde{a}(da)}{\tilde{q}(i, \tilde{a}, j)}, \qquad k = 0, 1, \ldots, K, \tag{A.0.3}$$

where (A.0.1) means that jump intensities for the control $\tilde{a}$ are convex combinations of the jump intensities for the corresponding controls a and formulas (A.0.2) and (A.0.3) mean that the reward rates for the control $\tilde{a}$ are convex combinations of the reward rates for the corresponding controls a.

A strategy in a new model is called *a randomized strategy*. In other words, a randomized strategy is a measurable mapping π from Ω to $\mathcal{P}(A)$ such that $\pi(A(i_n)|i_0, t_1, \ldots, t_n, i_n, t) = 1$ for any $(i_0, t_1, \ldots, t_{n-1}, i_n, t) \in \Omega, n = 0, 1, \ldots$. Since a randomized strategy is defined as nonrandomized strategy in the cor-

responding model, it also defines a unique multivariate point process for any given initial distribution μ.

So, we can extend the notations P_μ^π and E_μ^π and the definition of the expected total discounted rewards (2.1.1) from (nonrandomized) strategies to randomized strategies. So, formula (2.1.1) applied to the new model with $\psi = \pi$ defines the expected total discounted rewards $V_k(i, \pi)$ for a randomized strategy π, where a_n and $a(t)$ elements of $\tilde{A} = \mathcal{P}(A)$. We shall follow the agreement that $a \in A$ and a probability measure on A concentrated on a are the same objects. Then $P_i^\pi = P_i^\psi$ and $V_k(i, \pi) = V_k(i, \psi)$ when the randomized policy π is nonrandomized, i.e. $\pi(\psi(\omega)|\omega) = 1$ for any $\omega \in \Omega$, where ψ is a (nonrandomized) strategy for the original Continuous-Time MDP.

Of course, the question whether randomized strategies can be implemented in a particular application depends on the applications. However, randomized strategies are convenient for mathematical considerations, they do not change the objective function, and optimal nonrandomized strategies are optimal within the class of all nonrandomized strategies.

Remark. In the above form randomized policies were defined in Feinberg [13]. This definition is equivalent to the definition in Kitaev [32] and Kitaev and Rykov [33], where, for the case of the sample space $(\Omega_\infty, \mathcal{F}_\infty)$, a randomized strategy is defined as a regular transition probability from $(\Omega, \mathcal{F})$ to A such that $\pi(A(i_n)|i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) = 1$ for each $(i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) \in \Omega$. The assumption that π is a regular transition probability means that $\pi(\cdot|i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t)$ is a probability measure on A for each $(i_0, t_1, \ldots, i_{n-1}, t_n, i_n, t) \in \Omega$ and $\pi(B|\cdot)$ is a measurable function on $(\Omega, \mathcal{F})$ for any measurable subset B of A. For a strategy ψ and a given finite history $\omega_n = i_0, t_1, \ldots, t_n, i_n t$, the joint probability distribution that the jump happens during the interval $[t, t + dt)$ and $i_{n+1} = j$ is $dt \int q(i, j, a)\pi(da|i_0, t_1, \ldots, t_n, i_n, t)$. Formula (A.0.3) for instant rewards at jump epochs was not explicitly presented in [13], were it was shown that for the expected discounted rewards it is possible to adjust r_k and set $R_k = 0$. References [32] and [33] did not consider instant rewards.

Appendix B

Derivation of variances

Consider a randomized stationary policy that has m actions to choose at state i, $\sigma = \{a_1, a_2, \ldots, a_m\}$, and one of the equivalent switching strategies defined by (2.2.1), $\phi =< a_1, a_2, \ldots, a_m >$.

$U(i)$ - total discounted reward earned when the initial state is $i, i \in S$.

ζ_k - exponential RV with intensity equal to q_k - the jump intensity when action a_k is taken.

X_1 - the next state after the jump.

Since the first moments for the randomized policy and for the switching stationary policy are the same we use $M(i) = E[U(i)], i \in S$ to represent the first moments. To differentiate between the randomized and switching policies we use W_r to represent the second moment for a randomized policy and W_s to represent the second moment for a switching policy.

In the sequel we will derive two sets of simultaneous equations, one involving the first moments and the other involving both the first and the second moments. We can solve the first set of simultaneous equations to obtain $M(i)$ first, which are the same for the switching policy and for the multiplicative discounting. With that we can solve the second set of simultaneous equations to obtain $W_r(i)$ or $W_s(i)$. The variances can then be computed as $\text{Var(i)} = W_r(i) - M^2(i)$.

B.1 Variances under probabilistic discounting

For the randomized policy, the derivation is easy. Conditioning on action a we have

$$\begin{aligned}
M(i) &= E[U(i)] = EE[U(i)|a] \\
&= \sum_{k=1}^{m} p_k E\Big[r_k\zeta_k + R(i,X_1,a_k) + U(X_1)\Big] \\
&= \sum_{k=1}^{m} p_k EE\Big[r_k\zeta_k + R(i,X_1,a_k) + U(X_1)|X_1\Big] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k) E\Big[r_k\zeta_k + R(i,j,a_k) + U(j)\Big] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k)\Big[\frac{r_k}{q_k} + R(i,j,a_k) + M(j)\Big] \\
&= \sum_{k=1}^{m} p_k \Big[\frac{r_k}{q_k} + \sum_j p(i,j,a_k)\Big(R(i,j,a_k) + M(j)\Big)\Big].
\end{aligned}$$

$$\begin{aligned}
& W_r(i) = E[U^2(i)] = EE[U^2(i)|a] \\
= & \sum_{k=1}^{m} p_k E\Big[\Big(r_k\zeta_k + R(i,X_1,a_k) + U(X_1)\Big)^2\Big] \\
= & \sum_{k=1}^{m} p_k EE\Big[\Big(r_k\zeta_k + R(i,X_1,a_k) + U(X_1)\Big)^2|X_1\Big] \\
= & \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k) E\Big[\Big(r_k\zeta_k + R(i,j,a_k) + U(j)\Big)^2\Big] \\
= & \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k)\Big[\frac{r_k^2}{q_k^2} + R^2(i,j,a_k) + W_r(j) + \frac{2r_k}{q_k}[R(i,j,a_k) + M(j)] + 2R(i,j,a_k)M(j)\Big] \\
= & \sum_{k=1}^{m} p_k \Big[\frac{r_k^2}{q_k^2} + \sum_j p(i,j,a_k)\Big(R^2(i,j,a_k) + W_r(j) + \frac{2r_k}{q_k}[R(i,j,a_k) + M(j)] \\
& +2R(i,j,a_k)M(j)\Big)\Big].
\end{aligned}$$

For the switching policy, the first moments are the same as the randomized

policy, so we simply use $M(i)$ for $M(i)$ for $i \in S$. For the second moments, conditioning on T_1, the time of the first jump, we have

$$
\begin{aligned}
W_s(i) &= E[U^2(i)] = EE[U^2(i)|T_1] \\
&= \sum_{k=1}^{m} P\{S_k - 1 < T_1 < S_k\} E\Big[\Big(\sum_{\ell=1}^{k-1} r_\ell s_\ell + r_k\zeta_k + R(i, X_1, a_k) + U(X_1)\Big)^2\Big] \\
&= \sum_{k=1}^{m} p_k EE\Big[\Big(\sum_{\ell=1}^{k-1} r_\ell s_\ell + r_k\zeta_k + R(i, X_1, a_k) + U(X_1)\Big)^2 |X_1\Big] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k) E\Big[\Big(\sum_{\ell=1}^{k-1} r_\ell s_\ell + r_k\zeta_k + R(i,j,a_k) + U(j)\Big)^2\Big] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k)\Big[(\sum_{\ell=1}^{k-1} r_\ell s_\ell)^2 + \frac{2r_k^2}{q_k^2} + R^2(i,j,a_k) + W_s(j) \\
&\quad +2[\frac{r_k}{q_k} + R(i,j,a_k) + M(j)]\sum_{\ell=1}^{k-1} r_\ell s_\ell + \frac{2r_k}{q_k}[R(i,j,a_k) + M(j)] + R(i,j,a_k)M(j)\Big)\Big] \\
&= \sum_{k=1}^{m} p_k\Big[(\sum_{\ell=1}^{k-1} r_\ell s_\ell)^2 + \frac{2r_k^2}{q_k^2} + \frac{2r_k}{q_k}\sum_{\ell=1}^{k-1} r_\ell s_\ell + \sum_j p(i,j,a_k)\Big(R^2(i,j,a_k) + W_s(j) \\
&\quad +2[R(i,j,a_k) + M(j)]\sum_{\ell=1}^{k-1} r_\ell s_\ell + \frac{2r_k}{q_k}[R(i,j,a_k) + M(j)] + R(i,j,a_k)M(j)\Big)\Big].
\end{aligned}
$$

The results are summarized in the following theorem:

Theorem B.1.1 *Under the probabilistic discounting the variances for the discounted total rewards of a randomized policy $\sigma = \{a_1, a_2, \ldots, a_m\}$, and an equivalent switching strategies defined by (2.2.1), $\phi =< a_1, a_2, \ldots, a_m >$, can be computed by solving the following simultaneous equations:*

$$
\begin{aligned}
M(i) &= \sum_{k=1}^{m} p_k \left[\frac{r_k}{q_k} + \sum_j p(i,j,a_k) \Big(R(i,j,a_k) + M(j) \Big) \right], \\
W_r(i) &= \sum_{k=1}^{m} p_k \left[\frac{r_k^2}{q_k^2} + \sum_j p(i,j,a_k) \Big(R^2(i,j,a_k) + W_r(j) + \frac{2r_k}{q_k} [R(i,j,a_k) + M(j)] \right. \\
&\quad \left. + 2R(i,j,a_k) M(j) \Big) \right], \\
W_s(i) &= \sum_{k=1}^{m} p_k \left[\Big(\sum_{\ell=1}^{k-1} r_\ell s_\ell \Big)^2 + \frac{2r_k^2}{q_k^2} + \frac{2r_k}{q_k} \sum_{\ell=1}^{k-1} r_\ell s_\ell + \sum_j p(i,j,a_k) \Big(R^2(i,j,a_k) + W_s(j) \right. \\
&\quad \left. + 2[R(i,j,a_k) + M(j)] \sum_{\ell=1}^{k-1} r_\ell s_\ell + \frac{2r_k}{q_k} [R(i,j,a_k) + M(j)] + R(i,j,a_k) M(j) \Big) \right], \\
&\quad \textit{for any} \quad i \in S.
\end{aligned}
$$

B.2 Derivation of variance under multiplicative discounting

Since the first moments are the same under the two definitions of discounting we skip the derivation for the first moments.

When the continuously discount rate is α, one unit of reward at time t is worth $e^{-\alpha t}$ unit at $t = 0$. When the reward rate is r, the total discounted reward earned during $[0, t]$ and discounted back to $t = 0$ is:

$$
U_\alpha(r,t) = \int_0^t r \exp(-\alpha t) dt = \frac{r(1 - e^{-\alpha t})}{\alpha} \tag{B.2.1}
$$

We first derive for the randomized policy by conditioning on a:

$$
\begin{aligned}
& W_r(i) = E[U^2(i)] = EE[U^2(i)|a] \\
= & \sum_{k=1}^{m} p_k E\left[\left(\frac{r_k(1-e^{\alpha\zeta_k})}{\alpha} + e^{-\alpha\zeta_k}[R(i,X_1,a_k)+U(X_1)]\right)^2\right] \\
= & \sum_{k=1}^{m} p_k EE\left[\left(\frac{r_k(1-e^{\alpha\zeta_k})}{\alpha} + e^{-\alpha\zeta_k}[R(i,X_1,a_k)+U(X_1)]\right)^2 |X_1\right] \\
= & \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k) E\left[\left(\frac{r_k(1-e^{\alpha\zeta_k})}{\alpha} + e^{-\alpha\zeta_k}[R(i,j,a_k)+U(j)]\right)^2\right] \\
= & \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k)\left[\frac{2r_k^2}{(\alpha+q_k)(2\alpha+q_k)} + \frac{q_k}{2\alpha+q_k}[R^2(i,j,a_k)+W_r(j)] + \right. \\
& \left. \frac{2r_kq_k}{(\alpha+q_k)(2\alpha+q_k)}[R(i,j,a_k)+M(j)] + \frac{2q_k}{(2\alpha+q_k)}R(i,j,a_k)M(j)\right] \\
= & \sum_{k=1}^{m} p_k\left[\frac{2r_k^2}{(\alpha+q_k)(2\alpha+q_k)} + \sum_j p(i,j,a_k)\left(\frac{q_k}{2\alpha+q_k}[R^2(i,j,a_k)+W_r(j)] + \right.\right. \\
& \left.\left. \frac{2r_kq_k}{(\alpha+q_k)(2\alpha+q_k)}[R(i,j,a_k)+M(j)] + \frac{2q_k}{(2\alpha+q_k)}R(i,j,a_k)M(j)\right)\right].
\end{aligned}
$$

For the second moments we still condition on T_1:

$$
\begin{aligned}
W_s(i) &= E[U^2(i)] = EE[U^2(i)|T_1] \\
&= \sum_{k=1}^{m} P\{S_k - 1 < T_1 < S_k\} E\Bigg[\Bigg(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + e^{-\alpha S_{k-1}} \frac{r_k(1-e^{-\alpha\zeta_k})}{\alpha} \\
&\quad + e^{-\alpha S_{k-1}} e^{-\alpha\zeta_k}[R(i,X_1,a_k) + U(X_1)]\Bigg)^2\Bigg] \\
&= \sum_{k=1}^{m} p_k EE\Bigg[\Bigg(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + e^{-\alpha S_{k-1}} \frac{r_k(1-e^{-\alpha\zeta_k})}{\alpha} \\
&\quad + e^{-\alpha S_{k-1}} e^{-\alpha\zeta_k}[R(i,X_1,a_k) + U(X_1)]\Bigg)^2 |X_1\Bigg] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k) E\Bigg[\Bigg(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + e^{-\alpha S_{k-1}} \frac{r_k(1-e^{-\alpha\zeta_k})}{\alpha} \\
&\quad + e^{-\alpha S_{k-1}} e^{-\alpha\zeta_k}[R(i,j,a_k) + U(j)]\Bigg)^2\Bigg] \\
&= \sum_{k=1}^{m} p_k \sum_j p(i,j,a_k)\Bigg[\Bigg(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha}\Bigg)^2 + \frac{2r_k^2 e^{-2\alpha S_{k-1}}}{(\alpha+q_k)(2\alpha+q_k)} + \\
&\quad \frac{q_k e^{-2\alpha S_{k-1}}}{2\alpha+q_k}[R^2(i,j,a_k) + W_s(j)] + 2e^{-\alpha S_{k-1}}[\frac{r_k}{\alpha+q_k} + \frac{q_k}{\alpha+q_k} R(i,j,a_k) \\
&\quad + \frac{q_k}{\alpha+q_k} M(j)] \sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + \frac{2r_k e^{-2\alpha S_{k-1}}}{q_k}[\frac{q_k}{\alpha+q_k} R(i,j,a_k) + \frac{q_k}{\alpha+q_k} M(j)] \\
&\quad + e^{-2\alpha S_{k-1}} \frac{q_k}{2\alpha+q_k} R(i,j,a_k) M(j)\Bigg)\Bigg] \\
&= \sum_{k=1}^{m} p_k \Bigg[\Bigg(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha}\Bigg)^2 + \frac{2r_k^2 e^{-2\alpha S_{k-1}}}{(\alpha+q_k)(2\alpha+q_k)} \\
&\quad + \frac{2r_k e^{-\alpha S_{k-1}}}{\alpha+q_k} \sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + \sum_j p(i,j,a_k)\Bigg(\frac{q_k e^{-2\alpha S_{k-1}}}{2\alpha+q_k}[R^2(i,j,a_k) \\
&\quad + W_s(j)] + \frac{2q_k e^{-\alpha S_{k-1}}}{\alpha+q_k}[R(i,j,a_k) + M(j)] \sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell(1-e^{-\alpha s_\ell})}{\alpha} + \\
&\quad \frac{2r_k e^{-2\alpha S_{k-1}}}{\alpha+q_k}[R(i,j,a_k) + M(j)] + e^{-2\alpha S_{k-1}} \frac{q_k}{2\alpha+q_k} R(i,j,a_k) M(j)\Bigg)\Bigg].
\end{aligned}
$$

The results are summarized in the following theorem:

Theorem B.2.1 *Under the multiplicative discounting the variances for the discounted total rewards of a randomized policy $\sigma = \{a_1, a_2, \ldots, a_m\}$, and an equivalent switching strategies defined by (2.2.1), $\phi =< a_1, a_2, \ldots, a_m >$, can be computed by solving the following simultaneous equations:*

$$
\begin{aligned}
M(i) &= \sum_{k=1}^{m} p_k \left[\frac{r_k}{q_k} + \sum_j p(i,j,a_k) \Big(R(i,j,a_k) + M(j) \Big) \right], \\
W_r(i) &= \sum_{k=1}^{m} p_k \left[\frac{2r_k^2}{(\alpha+q_k)(2\alpha+q_k)} + \sum_j p(i,j,a_k) \Big(\frac{q_k}{2\alpha+q_k} [R^2(i,j,a_k) + W_r(j)] + \right. \\
&\quad \left. \frac{2r_k q_k}{(\alpha+q_k)(2\alpha+q_k)} [R(i,j,a_k) + M(j)] + \frac{2q_k}{(2\alpha+q_k)} R(i,j,a_k) M(j) \Big) \right], \\
W_s(i) &= \sum_{k=1}^{m} p_k \left[\left(\sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell (1-e^{-\alpha s_\ell})}{\alpha} \right)^2 + \frac{2r_k^2 e^{-2\alpha S_{k-1}}}{(\alpha+q_k)(2\alpha+q_k)} \right. \\
&\quad + \frac{2r_k e^{-\alpha S_{k-1}}}{\alpha+q_k} \sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell (1-e^{-\alpha s_\ell})}{\alpha} + \sum_j p(i,j,a_k) \Big(\frac{q_k e^{-2\alpha S_{k-1}}}{2\alpha+q_k} [R^2(i,j,a_k) \\
&\quad + W_s(j)] + \frac{2q_k e^{-\alpha S_{k-1}}}{\alpha+q_k} [R(i,j,a_k) + M(j)] \sum_{\ell=1}^{k-1} e^{-\alpha S_{\ell-1}} \frac{r_\ell (1-e^{-\alpha s_\ell})}{\alpha} \\
&\quad \left. + \frac{2r_k e^{-2\alpha S_{k-1}}}{\alpha+q_k} [R(i,j,a_k) + M(j)] + e^{-2\alpha S_{k-1}} \frac{q_k}{2\alpha+q_k} R(i,j,a_k) M(j) \Big) \right], \\
&\quad \textit{for any} \quad i \in S.
\end{aligned}
$$