Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 181-195
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Variance Optimization for Continuous-Time Markov Decision Processes

Yaqing Fu

School of Economic, Jinan University, Guangzhou, China
Email: 13535559908@163.com

How to cite this paper: Fu, Y.Q. (2019) Variance Optimization for Continuous-Time Markov Decision Processes. *Open Journal of Statistics*, **9**, 181-195.
https://doi.org/10.4236/ojs.2019.92014

Received: March 5, 2019
Accepted: March 30, 2019
Published: April 2, 2019

Abstract

This paper considers the variance optimization problem of average reward in continuous-time Markov decision process (MDP). It is assumed that the state space is countable and the action space is Borel measurable space. The main purpose of this paper is to find the policy with the minimal variance in the deterministic stationary policy space. Unlike the traditional Markov decision process, the cost function in the variance criterion will be affected by future actions. To this end, we convert the variance minimization problem into a standard (MDP) by introducing a concept called pseudo-variance. Further, by giving the policy iterative algorithm of pseudo-variance optimization problem, the optimal policy of the original variance optimization problem is derived, and a sufficient condition for the variance optimal policy is given. Finally, we use an example to illustrate the conclusion of this paper.

Keywords

Continuous-Time Markov Decision Process, Variance Optimality of Average Reward, Optimal Policy of Variance, Policy Iteration

1. Introduction

The Postal Service Company's catalogue information system, inventory issues, and supply chain management issues are all early successful applications of the Markov decision process. Later, many real-life problems, such as sequential assignments, machine maintenance issues, and secretarial issues, can be described as dynamic Markov Decision Processes (MDP) model. They are finally solved very well by MDP [1] [2].

This paper considers the variance optimization problem of average reward in continuous-time Markov decision process. It is assumed that the state space is countable and the action space is Borel measurable space. The main purpose of

this paper is to find the policy with the minimal variance in the deterministic stationary policy class, which is different from the mean-variance criterion problem. The study of the mean-variance criterion problem is generally based on the discount criterion or the average criterion. In the literature of MDPs, many studies focus on the problem of expected reward optimization in finite stage, the discounted MDP in infinite stage and the average reward problem in infinite stage [2] [3]. By establishing the optimal equation, then the existence of optimal policy is proved, and finally the policy iteration type algorithm is used to solve the MDP problem. However, in real-life, the optimal criteria of this unconstrained optimization problem are often not unique, such as queuing system and network problems. So we introduce variance to choose the optimal strategy.

Variance is an important performance metric of stochastic systems. In financial engineering, we use the mean to measure the expected return, and the variance to measure the risk. The mean-variance problem of the portfolio can be traced back to Markowitz [4]. Then the Markowitz's mean-variance portfolio problem has been studied [5]-[13], the decision maker's expected reward is often assumed to be a constant, and then the investor chooses a policy with a given expected return to minimize this risk, we can see that the Markowitz mean-variance portfolio model is a model of maximization of return and minimization of risk. However, given expected return which may not be maximal, an optimal policy in Markowitz' mean-variance portfolio may not be optimal in the usual sense of variance minimization problems for MDPs. Moreover, more and more real-life situations such as queuing systems and networks can be described as MDPs rather than stochastic differential equations, so Markowitz's mean-variance portfolio problem should be extended to MDPs. For mean-variance problem of the MDPs, as in [14] [15] [16], we aim to obtain a variance optimal policy over a set of policies where the average reward or discounted reward is optimal, so the variance criterion can be transformed into an equivalent average or discount criterion. However, when the mean criterion is not optimal, it is not clear how to develop a policy iteration algorithm to solve the problem. For discrete-time, discount and long-run average variance criterion problem has been studied in [17] [18]. They mainly consider the variance optimization problem, and do not constrain the mean. For continuous-time, the variance of the average expected return has been defined in deterministic stationary policy. The finite-horizon expected reward is defined as below.

$$V_T(i,f) = E_i^f\left\{\int_0^T r\left(X(t),f\right)\mathrm{d}t\right\}, \tag{1.1}$$

The variance of f, $\sigma^2(i,f)$, is given by:

$$\sigma^2(i,f) = \lim_{T\to\infty}\frac{1}{T}E_i^f\left\{\int_0^T r\left(X(t),f\right)\mathrm{d}t - V_T(i,f)\right\}^2 \tag{1.2}$$

However, the variance function of average expected return of the continuous-time in this paper is given by

$$\eta_\sigma^f = \lim_{T\to\infty}\frac{1}{T}E_i^f\left\{\int_0^T\left(r\left(X(t),f\right)-\eta^f\right)^2\mathrm{d}t\right\}, \tag{1.3}$$

The long-run expected average reward is defined as below.

$$\eta^f = \lim_{T\to\infty} \frac{1}{T} E_i^f \left\{ \int_0^T r\left(X(t), f\right) \mathrm{d}t \right\}. \tag{1.4}$$

The main work of this paper is to find the iterative algorithm of the optimal policy under the variance criterion (minimum variance) in the countable state space and the Borel measurable action space. For countable state space, the reward function $r(i,a)$ may be unbounded, the expected average reward η^f, may not be infinite. To guarantee the finiteness of η^f, we will impose the following Assumption 1. Next, we use a unique invariant probability measure of Markov chain to denote the average expected return and variance. To this end, we will impose the following Assumption 2, 3, 4. Suppose that Assumptions 1, 2, 3, and 4 are satisfied. We have established a variance criterion. Under the variance criterion, we define the cost function $m(i,a) = \left(r(i,a) - \eta^f\right)^2$, where $r(i,a)$ is the system reward at the current stage with state *i* and action *a*, and η^f is the expected average reward. Obviously, the cost will be affected by future actions, so, η^f is also affected by future actions. The traditional MDPs differs from this. The cost function and state transition probability depend only on the current state and the action selected on this stage. Therefore, the conclusions in [14] [15] [16] do not apply to this model. In this paper, we define a pseudo-variance $m_\lambda(i,a) = \left(r(i,a) - \lambda\right)^2$, where λ is a given constant [17]. Obviously, the value of the pseudo-variance at current stage will not be affected by future actions. It is only related to the current state and current actions, so the pseudo-variance minimization problem is a standard MDP. In this paper, we prove the relation between variance and pseudo-variance. Unlike the literature [17], we define the deviation of the deterministic stationary policy *f* for continuous-time MDP. It is proved that the deviation function and the objective function satisfy the Poisson equation, and the uniqueness of the Poisson equation is proved. Based on this, we develop a continuous time MDP policy iterative algorithm to get the optimal strategy, and we prove the convergence of the policy iterative algorithm.

2. Model and Optimization Criteria

The control model associated with the continuous-time MDP that we are concerned with is the five-tuple

$$\left\{S, \left(A(i) \subseteq A, i \in S\right), q(.\,|\,i,a), r(i,a)\right\} \tag{2.1}$$

1) A denumerable set *S*, called the stated space, which is the set of all the states of the system under observation.

2) A Borel space *A*, called the action space. Let

$$K := \left\{(i,a)\,|\,i \in S, a \in A(i)\right\}. \tag{2.2}$$

be the set of all feasible state-action pairs.

3) The transition rates $q(j\,|\,i,a)$ which satisfy $q(j\,|\,i,a) \geq 0$ for all $(i,a) \in K$ and $j \neq i$. Moreover, we assume that the transition rates $q(j\,|\,i,a)$ are conserv-

ative, *i.e.*,

$$\sum_{j\in S} q(j\,|\,i,a) = 0 \quad \forall (i,a) \in K \tag{2.3}$$

and stable, which means that

$$q^*(i) := \sup_{a\in A(i)} q_i(a) < \infty \quad \forall i \in S \tag{2.4}$$

where $q_i(a) := -q(i\,|\,i,a) \geq 0$ for all $(i,a) \in K$. In addition, $\sum_{j\in S} q(j\,|\,i,a)$ is measurable in $a \in A(i)$ for each fixed $i, j \in S$.

4) A measurable real-valued function $r(i,a)$ on *K*. called the reward function, which is assumed to be measurable in $a \in A(i)$ for each fixed $i \in S$.

The above model is a classical continuous-time MDP model [3]. In MDP, the policies have stochastic Markov policy, stochastic stationary policy and deterministic stationary policy. This paper only considers finding the minimal variance in the deterministic stationary policy class. So we only introduce the definition of deterministic stationary policy.

Definition 1. A deterministic stationary policy is a function $f : S \to A(i)$ such that $f(i)$ is in $A(i)$ for all $i \in S$. A deterministic stationary policy is simply referred to as a stationary policy.

Let *F* be the set of all deterministic stationary policies.

For each $f \in F$, the associated transition rates are defined as

$$q(j\,|\,i,f) := q(j\,|\,i,f(i)) \quad \forall i, j \in S, \tag{2.5}$$

the reward function is given by

$$r(i,f) := r(i,f(i)) \quad \forall i \in S. \tag{2.6}$$

Under Assumption 1, the transition function $p_f(i,t,j)$ is regular [3].

Assumption 1:

a) There exist a nondecreasing function $\omega \geq 1$, on *S*, and constants $c_1 > 0$, and $b_1 > 0$, such that

$$\sum_{j\in S} \omega(j) q(j\,|\,i,a) \leq -c_1 \omega(i) + b_1 \delta_{i0},$$

for all $(i,a) \in K$, where $\delta_{00} = 1, \delta_{i0} = 0, i \neq 0$.

b) $q^*(i) \leq L_0 \omega(i)$ for all $i \in S$, with $L_0 > 0$ and $q^*(i)$ as in (2.4).

c) $|r(i,a)| \leq M\omega(i)$, for all $(i,a) \in K$, with some $M > 0$.

d) The action set $A(i)$ is compact for each $i \in S$, the functions $r(i,a)$, $q(j\,|\,i,a)$ and $\sum_{k\in S} \omega(k) q(k\,|\,i,a)$ are all continuous in $a \in A(i)$ for each fixed $i, j \in S$.

e) There exists a nonnegative function ω' on *S* and constants $c' > 0, b' > 0$, and $L_1 > 0$ such that

$$q^*(i)\omega(i) \leq L_1 \omega'(i) \text{ and } \sum_{j\in S} \omega'(j) q(j\,|\,i,a) \leq c'\omega'(i) + b' \tag{2.7}$$

For all $(i,a) \in K$, with *K* and $q^*(i)$ as in (2.2) and (2.4), respectively.

For all $f \in F$ and an arbitrary initial state $i \in S$, there is a unique probabil-

ity space $\left(\Omega, B(\Omega), P_i^f\right)$, where the probability measure P_i^f is determined by f and $p_f(i,t,j)$. E_i^f is denoted as the expectation operator P_i^f. Define the expected average reward and variance respectively.

$$\eta^f(i) = \lim_{T\to\infty} \frac{1}{T} E_i^f \left\{ \int_0^T r\left(X(t), f\right) \mathrm{d}t \right\}, \tag{2.8}$$

$$\eta_\sigma^f(i) = \lim_{T\to\infty} \frac{1}{T} E_i^f \left\{ \int_0^T \left(r\left(X(t), f\right) - \eta^f \right)^2 \mathrm{d}t \right\}. \tag{2.9}$$

Remark 1. From (2.9), we can see that the definition of η_σ^f is different from the definition of the continuous-time MDP average reward variance criterion in ([3], chapter 10), where the value of cost function will be affected by future actions, so this is not a standard MDP optimization problem.

Let's give some marks first. For any measurable function $\omega \geq 1$ on S, we define the ω-weighted supremum norm $\|.\|_\omega$ of a real-valued measurable function ω on S by

$$\|u\|_\omega := \sup_{i\in S} \frac{|u(i)|}{\omega(i)}, \tag{2.10}$$

and the Banach space $B_\omega(S) := \left\{u : \|u\|_\omega < \infty\right\}$. Similarly, we can define $B_{\omega^2}(S)$.

We will use the Markov chain invariant measure to represent Equation (2.8) and Equation (2.9). To this end, we impose the following three Assumptions (see [3]).

Assumption 2:

For each $f \in F$, the corresponding Markov process $\left\{X(t)\right\}$ with transition function $p_f(i,t,j)$ is irreducible, which means that, for any two states $i \neq j$, there exists a set of distinct states $i = i_1, \cdots, i_m$ such that

$$q(i_2 \mid i_1, f) \cdots q(j \mid i_m, f) > 0. \tag{2.11}$$

Under Assumptions 1(a) and 2, for each $f \in F$, Propositions C.11 and C.12 yield that the Markov chain $\left\{X(t)\right\}$ has a unique invariant probability measure, denoted by μ_f, which satisfies that $\mu_f(j) = \lim_{t\to\infty} p_f(i,t,j)$ (independent of $i \in S$) for all $j \in S$. Thus, by [3], we have

$$\mu_f(\omega) := \sum_{j\in S} \omega(j) \mu_f(j) \leq \frac{b_1}{c_1}, \tag{2.12}$$

which shows that the μ_f-expectation of ω (*i.e.*, $\mu_f(\omega)$) is finite. Therefore, for all $f \in F$ and $u \in B_\omega(S)$, the inequality $|u(i)| \leq \|u\|_\omega \omega(i)$ for all $i \in S$ gives that the expectation

$$\mu_f(u) := \sum_{i\in S} u(i) \mu_f(i) \tag{2.13}$$

exists and is finite.

Assumption 3:

With ω as in Assumption 1, assume the following conditions are true:

a) There exists constants $c' > 0, b' > 0$, such that

$$\sum_{j\in S} \omega^2(j) q(j \mid i, a) \leq -c_2 \omega^2(i) + b_2 \tag{2.14}$$

For all $(i,a) \in K$.

b) There exists a nonnegative function ω'' on S and constants $c'' > 0, b'' > 0$ and $L_2 \geq 0$, such that

$$q^*(i)\omega^2(i) \leq L_2\omega''(i), \quad \sum_{j\in S}\omega''(j)q(j \mid i,a) \leq -c''\omega''(i) + b'' \tag{2.15}$$

for all $(i,a) \in K$, with K and $q^*(i)$ as in (2.2) and (2.4).

Assumption 4:

a) The control model (2.1) is uniformly ω-exponentially ergodic, which means the following: there exist constants $\beta > 0$, and $L_3 > 0$ $\delta > 0$ such that

$$\sup_{f\in F}\left|E_i^f u(X(t)) - \mu_f(u)\right| \leq L_3 e^{-\beta t}\|u\|_\omega \omega(i) \tag{2.16}$$

for all $i \in S$, $u \in B_\omega(S)$, and $t \geq 0$.

b) The control model (2.1) is uniformly ω^2-exponentially ergodic, the definition as in (a)

Remark 2. Under the premise of the above assumptions, it can be known from the literature [3] that for the given $f \in F$, the average reward and variance defined by Equation (2.8) and Equation (2.9) are both a number, independent of the initial state. They can represent the expectation form of invariant measures $\boldsymbol{\mu}_f$, such that

$$\eta^f = \boldsymbol{\mu}_f \boldsymbol{r} := \sum_{i\in S} r(i,f)\mu_f(i), \tag{2.17}$$

$$\eta_\sigma^f = \boldsymbol{\mu}_f \boldsymbol{m} := \sum_{i\in S} m(i,f)\mu_f(i). \tag{2.18}$$

We denote $\boldsymbol{r}$ as an S-dimensional column vector composed by element $r(i,f)$ and $\boldsymbol{m}$ as an S-dimensional column vector composed by element $m(i,f)$

$$\text{where } m(i,f) := \left(r(i,f) - \eta^f\right)^2 \quad \forall i \in S, f \in F. \tag{2.19}$$

Our optimization goal is to select $f^* \in F$ that satisfies the following condition

$$\eta_\sigma^{f^*} = \min_{f\in F}\eta_\sigma^f \tag{2.20}$$

By (2.20), the variance minimization problem of Markov chains can be defined as below.

$$f^* = \arg\min_{f\in F}\left\{\eta_\sigma^f\right\} = \arg\min_{f\in F}\left\{\boldsymbol{\mu}_f \boldsymbol{m}\right\} \tag{2.21}$$

Remark 3. From (2.21), we see that the value η^f will be affected by future actions. There the problem (2.21) is different from standard MDP.

Even if we consider $\boldsymbol{m}$ as a cost function, we can't directly use the existing conclusions to get the optimal policy.

3. Analysis and Optimization

In this section, we will define a pseudo-variance minimization problem. By proving the relation between the pseudo-variance and the variance, the optimization problem of (2.21) is transformed into the pseudo-variance optimization problem.

Further, the optimal policy for variance optimization problem can be derived by the policy iterative algorithm for the pseudo-variance optimization problem, and we can give a sufficient condition for the variance optimal policy.

3.1. Pseudo-Variance Minimization

We define a new cost function as below.

$$m_\lambda(i,f)=\left(r(i,f)-\lambda\right)^2 \quad \forall i\in S,\ f\in F. \tag{3.1}$$

where λ is a given constant. We denote $\boldsymbol{m}_\lambda$ as an S-dimensional column vector composed by element $m_\lambda(i,f)$ and we have

$$\boldsymbol{m}_\lambda := \left(\boldsymbol{r}-\lambda \boldsymbol{I}\right)^2 \tag{3.2}$$

where $\boldsymbol{I}$ denote an S-dimensional column vector composed by element 1. We define pseudo-variance function as below

$$\eta_{\sigma\lambda}^f = \boldsymbol{\mu}_f \boldsymbol{m}_\lambda. \tag{3.3}$$

Obviously, we have

$$\eta_{\sigma\lambda}^f = \eta_\sigma^f, \text{ when } \lambda=\eta^f.$$

the pseudo-variance minimization problem of Markov chains can be defined as below.

$$f_\lambda^* = \arg\min_{f\in F}\left\{\eta_{\sigma\lambda}^f\right\} = \arg\min_{f\in F}\left\{\boldsymbol{\mu}_f\boldsymbol{m}_\lambda\right\} = \arg\min_{f\in F}\sum_{i\in S}\mu_f(i)\left(r(i,f)-\lambda\right)^2 \tag{3.4}$$

From (3.4), we can see that $\boldsymbol{m}_\lambda$ is an instant cost and it has no relation to future actions, Below, we study the relation between these two problems (2.21) and (3.4). First, we have the following lemma about the relation between $\eta_{\sigma\lambda}^f$ and η_σ^f.

Lemma 1. For all $f\in F$, the corresponding variance and the pseudo-variance has the following relation

$$\eta_{\sigma\lambda}^f = \eta_\sigma^f + \left(\eta^f-\lambda\right)^2. \tag{3.5}$$

Proof: From (3.1) and (3.3)

$$\begin{aligned}
\eta_{\sigma\lambda}^f &= \sum_{i\in S}\mu_f(i)\left(r(i,f)-\lambda\right)^2 = \sum_{i\in S}\mu_f(i)\left(r(i,f)-\eta^f+\eta^f-\lambda\right)^2\\
&= \sum_{i\in S}\mu_f(i)\left[\left(r(i,f)-\eta^f\right)^2+\left(\eta^f-\lambda\right)^2+2\left(r(i,f)-\eta^f\right)\left(\eta^f-\lambda\right)\right]\\
&= \eta_\sigma^f+\left(\eta^f\right)^2+\lambda^2-2\eta^f\lambda+2\left(\eta^f\right)^2-2\eta^f\lambda-2\left(\eta^f\right)^2+2\eta^f\lambda\\
&= \eta_\sigma^f+\left(\eta^f-\lambda\right)^2
\end{aligned}$$

The lemma is proved. □

Below we discuss how to solve the pseudo-variance minimum problem. Because (3.4) is a traditional MDP optimization problem, we can solve the problem with the policy iterative algorithm (3.4). Before using the policy iterative algorithm to solve the problem (3.4), we need to prove the existence of the pseudo-variance optimal policy. We suppose that Assumption 1, 2, 3, and 4 are all sa-

tisfied, we give the following theorems and lemmas.

Theorem 1. A pair $\left(g^*,u\right)\in\mathbb{R}\times B_{\omega^2}\left(S\right)$ is said to be a solution to the pseudo-variance of average-reward optimality equation if

$$g^* = \inf_{a\in A(i)}\left\{\left(r\left(i,a\right)-\lambda\right)^2+\sum_{j\in S}u\left(j\right)q\left(j\,|\,i,a\right)\right\} \quad \forall i\in S \tag{3.6}$$

Lemma 2. Suppose that Assumptions 1, 2, 3, and 4 are satisfied. Consider an arbitrary fixed state $i_0\in S$. Then, for all $f\in F$ and discount factors $\alpha>0$, the relative differences of the discounted-reward function η_α^f, namely,

$$u_\alpha^f\left(i\right)=\eta_\alpha^f\left(i\right)-\eta_\alpha^f\left(i_0\right) \quad i\in S \tag{3.7}$$

are uniformly ω-bounded in $\alpha>0$ and $f\in F$. More precisely, we have

$$\left\|u_\alpha^f\right\|_{\omega^2}\le\frac{L_3\left(M+|\lambda|\right)^2}{\delta}\left[1+\omega^2\left(i_0\right)\right] \quad \alpha>0,\ f\in F. \tag{3.8}$$

$$\text{where}\quad \eta_\alpha^f\left(i\right)=E_i^f\left[\int_0^\infty \mathrm{e}^{-\alpha t}\left(r\left(X\left(t\right),f\right)-\lambda\right)^2\mathrm{d}t\right]. \tag{3.9}$$

Prove: According to the literature [3], Lemma 2 can be known.

$$\text{where}\quad m_\lambda\left(i,f\right)=\left(r\left(i,f\right)-\lambda\right)^2 \quad \forall i\in S,\ f\in F.$$

Theorem 2. Suppose that Assumptions 1, 2, 3, and 4 hold. Then:

There exists a solution $\left(g^*,u\right)\in\mathbb{R}\times B_{\omega^2}\left(S\right)$ to pseudo variance of average-reward optimality equation. Moreover, the constant g^* coincides with the optimal average reward function $\eta_{\sigma\lambda}^*$, *i.e.*

$$g^*=\eta_{\sigma\lambda}^*\left(i\right) \quad \forall i\in S \tag{3.10}$$

Prove: Our assumptions ensure the existence of a policy attaining the minimization in the pseudo-variance of average-reward optimality equation, that is,

$$g^*=\left(r\left(i,f^*\right)-\lambda\right)^2+\sum_{j\in S}\overline{u}\left(j\right)q\left(j\,|\,i,f^*\right) \quad \forall i\in S,t\ge 0, \tag{3.11}$$

Therefore, Proposition 7.3 of the literature [3] gives $g^*=\eta_{\sigma\lambda}^{f^*}\left(i\right)$ $\forall i\in S$.

As a consequence, $g^*=\eta_{\sigma\lambda}^*\left(i\right)$ for every $i\in S$, and, moreover, f^* is optimal policy of pseudo-variance.

In the case where the existence of the pseudo-variance optimal policy is guaranteed, we use the policy iterative algorithm to get the optimal policy.

Suppose that Assumptions 1, 2, 3, and 4 hold, we gave the following concepts.

Definition 2. We define the bias of f as

$$h_{\sigma\lambda}^f\left(i\right)=\int_0^\infty\left[E_i^f\left(r\left(X\left(t\right),f\right)-\lambda\right)^2-\eta_{\sigma\lambda}^f\right]\mathrm{d}t \quad i\in S. \tag{3.12}$$

Assumption 4: Gives

$$\left|E_i^f\left(r\left(X\left(t\right),f\right)-\lambda\right)^2-\eta_{\sigma\lambda}^f\right|\le L_3\left(M+|\lambda|\right)^2\mathrm{e}^{-\delta t}\omega^2\left(i\right)$$

$$\left|h_{\sigma\lambda}^f\left(i\right)\right|\le\int_0^\infty L_3\left(M+|\lambda|\right)^2\mathrm{e}^{-\delta t}\omega^2\left(i\right)\mathrm{d}t=\frac{L_3\left(M+|\lambda|\right)^2\omega^2\left(i\right)}{\delta}$$

$$\sup_{f\in F}\left\|h_{\sigma\lambda}^{f}\right\|_{\omega^2}\le\frac{L_3\left(M+|\lambda|\right)^2}{\delta}.$$

So, $h_{\sigma\lambda}^{f}$ is finite and in $B_{\omega^2}(S)$. Moreover, the bias is uniformly bounded in the ω^2-norm.

Next we introduce the Poisson equation, which is one of the main results of this paper.

Theorem 3. Let $f\in F$. We say that a pair $\left(\eta_{\sigma\lambda}^{f},h_{\sigma\lambda}^{f}\right)\in\mathbb{R}\times B_{\omega^2}(S)$ is a solution to the Poisson equation for $f\in F$ if

$$\eta_{\sigma\lambda}^{f}=\left(r(i,f)-\lambda\right)^2+\sum_{j\in S}h_{\sigma\lambda}^{f}(j)q(j\mid i,f)\ \ \forall f\in F,i\in S. \tag{3.13}$$

Proof: Our assumptions (in particular, Assumptions 1 and 4) allow us to interchange the sums and integrals in the following equations:

$$\begin{aligned}
&\sum_{j\in S}h_{\sigma\lambda}^{f}(j)q(j\mid i,f)\\
&=\sum_{j\in S}\left[\int_0^\infty\left(E_j^f\left(r\left(X(t),f\right)-\lambda\right)^2-\eta_{\sigma\lambda}^{f}\right)\mathrm{d}t\right]q(j\mid i,f)\\
&=\int_0^\infty\sum_{j\in S}E_j^f\left(r\left(X(t),f\right)-\lambda\right)^2q(j\mid i,f)\mathrm{d}t-0\\
&=\int_0^\infty\sum_{j\in S}\sum_{k\in S}\left(r(k,f)-\lambda\right)^2p(j,t,k)q(j\mid i,f)\mathrm{d}t\\
&=\int_0^\infty\sum_{k\in S}\left(r(k,f)-\lambda\right)^2\sum_{j\in S}q(j\mid i,f)\,p(j,t,k)\,\mathrm{d}t\\
&=\int_0^\infty\sum_{k\in S}\left(r(k,f)-\lambda\right)^2\frac{\mathrm{d}}{\mathrm{d}t}p_f(i,t,k)\mathrm{d}t\\
&=\sum_{k\in S}\left(r(k,f)-\lambda\right)^2p_f(i,t,k)\Big|_0^\infty\\
&=\sum_{k\in S}\left(r(k,f)-\lambda\right)^2\mu_f(k)-\left(r(i,f)-\lambda\right)^2\\
&=\eta_{\sigma\lambda}^{f}-\left(r(i,f)-\lambda\right)^2
\end{aligned}$$

We have

$$\eta_{\sigma\lambda}^{f}=\left(r(i,f)-\lambda\right)^2+\sum_{j\in S}h_{\sigma\lambda}^{f}(j)q(j\mid i,f).$$

The theorem is proved. □

Finally, we should prove the uniqueness of the solution of Poisson's equation.

Theorem 4. For every $f\in F$, the solutions to the Poisson equation for f are of the form $\left(\eta_{\sigma\lambda}^{f},h_{\sigma\lambda}^{f}+z\right)$ with z any real number. Moreover, is the unique solution to the Poisson equation

$$\eta_{\sigma\lambda}^{f}=\left(r(i,f)-\lambda\right)^2+\sum_{j\in S}h_{\sigma\lambda}^{f}(j)q(j\mid i,f)\quad\forall i\in S \tag{3.14}$$

for which $\mu_f\left(h_{\sigma\lambda}^{f}\right)=0$

Prove: Suppose now that $\left(\eta_{\sigma\lambda}^{f},h_{\sigma\lambda}^{f}\right)$ and $\left(\eta_{\sigma\lambda}^{f},h'^{f}_{\sigma\lambda}\right)$ are two solutions to the Poisson equation, simultaneous transformation of both sides of Equation (3.14)

$$\begin{aligned}
\eta_{\sigma\lambda}^{f}T &= \int_0^T \sum_{k\in S}\left(r(k,f)-\lambda\right)^2 p(0,i,t,k)\mathrm{d}t \\
&\quad + \sum_{j\in S}\int_0^T h_{\sigma\lambda}^{f}(j)\sum_{k\in S} q(j\mid k,f)\,p(0,i,t,k)\mathrm{d}t \\
&= \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + \sum_{j\in S}\int_0^T h_{\sigma\lambda}^{f}(j)\sum_{k\in S} p(0,i,t,k)q(j\mid k,f)\mathrm{d}t \\
&= \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + \sum_{j\in S}\int_0^T h_{\sigma\lambda}^{f}(j)\mathrm{d}p(0,t,t,j) \\
&= \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + \sum_{j\in S} h_{\sigma\lambda}^{f}(j)\,p(0,i,t,j)\Big|_0^T \\
&= \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + \sum_{j\in S} h_{\sigma\lambda}^{f}(j)\,p(0,i,T,j) - h_{\sigma\lambda}^{f}(i) \qquad (3.15) \\
&= \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + E_i^f h_{\sigma\lambda}^{f}\left(X(T)\right) - h_{\sigma\lambda}^{f}(i)
\end{aligned}$$

Because $\left(\eta_{\sigma\lambda}^{f}, h_{\sigma\lambda}^{\prime f}\right)$ is also the solution of the Poisson equation, therefore

$$\eta_{\sigma\lambda}^{f}T = \int_0^T E_i^f\left(r(X(t),f)-\lambda\right)^2 \mathrm{d}t + E_i^f h_{\sigma\lambda}^{\prime f}\left(X(T)\right) - h_{\sigma\lambda}^{\prime f}(i) \qquad (3.16)$$

(3.15) subtract (3.16)

$$E_i^f\left[h_{\sigma\lambda}^{f}\left(X(t)\right) - h_{\sigma\lambda}^{\prime f}\left(X(t)\right)\right] = h_{\sigma\lambda}^{f}(i) - h_{\sigma\lambda}^{\prime f}(i) \quad \forall i \in S. \qquad (3.17)$$

letting $t\to\infty$, it follows from Assumption 4 that $\mu_f\left(h_{\sigma\lambda}^{f} - h_{\sigma\lambda}^{\prime f}\right) = h_{\sigma\lambda}^{f}(i) - h_{\sigma\lambda}^{\prime f}(i)$, showing that the functions $h_{\sigma\lambda}^{f}$ and $h_{\sigma\lambda}^{\prime f}$ differ by the constant $\mu_f\left(h_{\sigma\lambda}^{f} - h_{\sigma\lambda}^{\prime f}\right)$.

It remains to show that $\mu_f\left(h_{\sigma\lambda}^{f}\right) = 0$.

$$\begin{aligned}
&\mu_f\left(h_{\sigma\lambda}^{f}\right) = \sum_{i\in S} h_{\sigma\lambda}^{f}(i)\mu_f(i) \\
&= \sum_{i\in S}\int_0^\infty \left[E_i^f\left(r(X(t))-\lambda\right)^2 - \eta_{\sigma\lambda}^{f}\right]\mathrm{d}t\mu_f(i) \\
&= \sum_{i\in S}\int_0^\infty \left[E_i^f\left(r(X(t))-\lambda\right)^2 - \left(r(i,f)-\lambda\right)^2 - \sum_{j\in S} h_{\sigma\lambda}^{f}(j)q(j\mid i,f)\right]\mathrm{d}t\mu_f(i) \\
&= \sum_{i\in S}\int_0^\infty \left[E_i^f\left(r(X(t))-\lambda\right)^2\right]\mathrm{d}t\mu_f(i) - \int_0^\infty \sum_{i\in S}\left(r(i,f)-\lambda\right)^2\mu_f(i)\mathrm{d}t - 0 \\
&= \sum_{i\in S}\int_0^\infty \sum_{j\in S}\left(r(j,f)-\lambda\right)^2 p(i,t,j)\mu_f(i)\mathrm{d}t - \int_0^\infty \sum_{i\in S}\left(r(i,f)-\lambda\right)^2\mu_f(i)\mathrm{d}t \\
&= \int_0^\infty \sum_{j\in S}\left(r(j,f)-\lambda\right)^2 \sum_{i\in S}\mu_f(i)\,p(i,t,j)\,\mathrm{d}t - \int_0^\infty \sum_{i\in S}\left(r(i,f)-\lambda\right)^2\mu_f(i)\mathrm{d}t \\
&= \int_0^\infty \sum_{j\in S}\left(r(j,f)-\lambda\right)^2\mu_f(j)\mathrm{d}t - \int_0^\infty \sum_{i\in S}\left(r(i,f)-\lambda\right)^2\mu_f(i)\mathrm{d}t \\
&= 0
\end{aligned}$$

□

Remark 4. Given $f\in F$, we can determine the gain and the bias of f by solving the following system of linear equations. First, determine the i.p.m. (invariant probability measure) vas the unique nonnegative solution (by Proposition C.12) to

$$\begin{cases} \sum_{j\in S} q(j\mid i,f)\mu_f(j) = 0 \\ \sum_{j\in S}\mu_f(j) = 1 \end{cases} \qquad (3.18)$$

Then, as a consequence of lecture [2], the gain

$$\eta_{\sigma\lambda}^{f}=\sum_{j\in S}\left(r\left(j,f\right)-\lambda\right)^{2}\mu_{f}\left(j\right)\in\mathbb{R}$$

and the bias $h_{\sigma\lambda}^{f}\in B_{\omega^{2}}\left(S\right)$ of f form the unique solution to the system of linear equations

$$\begin{cases}\eta_{\sigma\lambda}^{f}=\left(r\left(i,f\right)-\lambda\right)^{2}+\sum_{j\in S}h_{\sigma\lambda}^{f}\left(j\right)q\left(j\mid i,f\right)\\ \sum_{i\in S}h_{\sigma\lambda}^{f}\left(i\right)\mu_{f}\left(i\right)=0\end{cases}\tag{3.19}$$

Proposition 1: Policy iterative algorithm.

Step 1. From $f\in F$, we can choose a arbitrary.

Step 2. (Strategy evaluation process) Determine the pseudo-variance and deviation of the stationary policy as in Remark 4.

Step 3. (Policy Improvement Process) Choose f' as an improvement policy such that

$$f'\in\arg\min_{f\in F}\left\{\left(r\left(i,f\right)-\lambda\right)^{2}+\sum_{j\in S}h_{\sigma\lambda}^{f}\left(j\right)q\left(j\mid i,f\right)\right\}.\tag{3.20}$$

Step 4. If $f'=f$, the iteration stops, it is the optimal strategy to minimize the pseudo variance, otherwise, replace f with f' and return to step 2.

Proposition 2: Convergence of the strategy iterative algorithm.

When the assumptions 1, 2, 3, 4 are established, let $f_{1}\in F$ be an arbitrary initial policy, let $f_{n}\in F$ be the sequence of policies obtained from the policy iterative algorithm. The one of the following results is hold.

1) After a finite number of policy iterations, the algorithm converges to the pseudo variance of average-reward optimal strategy.

2) as $n\to\infty$, the sequence $\eta_{\sigma\lambda}^{f_{n}}$ converges to the optimal AR function value $\eta_{\sigma\lambda}^{f^{*}}$.

3.2. Variance Minimization

The minimum pseudo-variance problem has been solved. The following theorem gives that when the pseudo-variance reaches a minimum, the variance is also minimized.

Theorem 5. For any policy $f\in F$, we compute η^{f} with (2.8), and set $\lambda=\eta^{f}$. If we obtain an improved policy f' such that $\eta_{\sigma\lambda}^{f'}\le\eta_{\sigma\lambda}^{f}$, then we have $\eta_{\sigma}^{f'}\le\eta_{\sigma}^{f}$. If $\eta_{\sigma\lambda}^{f'}<\eta_{\sigma\lambda}^{f}$, such that $\eta_{\sigma}^{f'}<\eta_{\sigma}^{f}$.

Prove: With lemma 1, we have

$$\eta_{\sigma\lambda}^{f}=\eta_{\sigma}^{f}+\left(\eta^{f}-\lambda\right)^{2}.\tag{3.21}$$

$$\eta_{\sigma\lambda}^{f'}=\eta_{\sigma}^{f'}+\left(\eta^{f'}-\lambda\right)^{2}.\tag{3.22}$$

Let (3.22) subtract (3.21) and substituting $\lambda=\eta^{f}$, we obtain

$$\eta_{\sigma}^{f'}-\eta_{\sigma}^{f}=\eta_{\sigma\lambda}^{f'}-\eta_{\sigma\lambda}^{f}-\left(\eta^{f'}-\eta^{f}\right)^{2}\tag{3.23}$$

Obviously, if $\eta_{\sigma\lambda}^{f'}\le\eta_{\sigma\lambda}^{f}$, then $\eta_{\sigma}^{f'}\le\eta_{\sigma}^{f}$; if $\eta_{\sigma\lambda}^{f'}<\eta_{\sigma\lambda}^{f}$, then $\eta_{\sigma}^{f'}<\eta_{\sigma}^{f}$.

With Theorem 3: when the pseudo-variance reaches a minimum, the variance also reaches a minimum. A sufficient condition for the variance minimization problem is obtained.

4. Examples

This section, we give an example to illustrate the conclusions of this paper.

Example 1 (Control $M/M/\infty$ of queue systems)

The system state $X(t)$ indicates the number of customers waiting at the moment (including being served), the arrival rate λ is fixed, and the service rates μ can be controlled. When the system status is $i \in S = \{0,1,\cdots\}$, the decision-maker takes an action *a* from the allowed action set $A(i)$. When the system is empty, we may impose that $A(0) := 0$. For each $i \geq 1$, let $A(i) := [\mu_1, \mu_2]$ with constants $\mu_2 > \mu_1 > 0$, $\mu \in [\mu_1, \mu_2]$, which may increase or decrease the service rate. This action incurs a cost $c(i,a)$. In addition, suppose that there is a benefit represented by $p > 0$ for each arriving customer, and then the net income of the system is

$$r(i,a) = pi - c(i,a) \tag{4.1}$$

This is a continuous time MDP model, the corresponding transition rate are given as follows.

For each $a \in [\mu_1, \mu_2]$,

$$q(0 \mid 0,0) = -q(1 \mid 0,0) := -\lambda,\ q(j \mid 0,0) = 0,\ j \geq 2, \tag{4.2}$$

For all $i \geq 1, a \in A(i) = [\mu_1, \mu_2]$,

$$q(j \mid i,a) := \begin{cases} \lambda i & \text{if } j = i+1, \\ -(\lambda + u)i + a & \text{if } j = i, \\ \mu i - a & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

Our goal is to find the existence of a variance optimal policy. To this end, we consider the following assumptions:

D_1. $\mu - \lambda > 0$.

D_2. The function $c(i,a)$ is continuous in $a \in A(i)$ for each fixed $i \in S$, and $\sup_{a \in A(i)} |c(i,a)| < \tilde{M}(i+1)$ for all $i \in S$, for some constant $\tilde{M} \geq 0$.

Proposition 3: under conditions D_1, D_2, the above controlled satisfies Assumptions 1, 2, 3, and 4. Therefore, there exists an variance optimal stationary policy.

Proof: Let $c_1 := \frac{1}{3}(\mu - \lambda)$, $b_1 := \mu_2 + \lambda$, $\omega(i) := i+1$, for all $i \in S$. Then, from (4.2) and (4.3), we have

$$\sum_{j \in S} \omega(j) q(j \mid 0,a) = \lambda \leq -c_1 \omega(0) + \mu_2 + \lambda \quad \forall a \in A(i) \tag{4.4}$$

Moreover, for all $i \geq 1, a \in [\mu_1, \mu_2]$

$$\sum_{j \in S} \omega(j) q(j \mid i,a) = -(\mu - \lambda)i + a \leq -c_1 \omega(i) + \mu_2 + \lambda. \tag{4.5}$$

which verifies Assumption 1(a).

On the other hand, by (4.2)-(4.3), we have

$$q^*(i) \le (\mu+\lambda)(i+1) = (\mu+\lambda)\omega(i), \tag{4.6}$$

So Assumption 1(b) follows.

By (4.1) and D_2, we have $|r(i,a)| \le (p+\tilde{M})\omega(i)$, for all $i \in S$, which implies Assumption 1(c).

By (4.2)-(4.3) and D_2, we see that Assumption 1(d) holds.

To verity Assumption 1(e), let

$$\omega'(i) := (i+1)(i+2), \text{ for each } i \in S \tag{4.7}$$

Then by (4.2)-(4.3) we have

$$q^*(i)\omega(i) \le (\mu+\lambda)\omega'(i) \quad \forall i \in S \tag{4.8}$$

$$\sum_{j\in S}\omega'(j)q(j\,|\,i,a) \le (4\lambda+\mu_2)\omega'(i) \quad \forall a \in [\mu_1,\mu_2], i \in S \tag{4.9}$$

which imply Assumption 1(e) with

$$L_1 := (\mu+\lambda),\ c' := 4\lambda+\mu_2,\ b' := 0.$$

Obviously, Assumption 2 follows from the description of the model.

We verity Assumption 3, by D_1 and (4.2)-(4.3), for all $i \in S$, $i \in A(i)$, we have

$$\sum_{j\in S}\omega^2(j)q(j\,|\,0,a) = 3\lambda \le -\frac{1}{2}(\mu-\lambda)\omega^2(0)+b_2, \tag{4.10}$$

(there is $b_2 > 0$, (presence is guaranteed by (D_1)))

$$\sum_{j\in S}\omega^3(j)q(j\,|\,0,a) = 7\lambda \le -(\mu-\lambda)\omega^3(0)+b_3. \tag{4.11}$$

(there is $b_3 > 0$, (presence is guaranteed by (D_1)))

For $i \ge 1, a \in A(i) = [\mu_1,\mu_2]$,

$$\begin{aligned}
&\sum_{j\in s}\omega^2(j)q(j\,|\,i,a)\\
&= -2(\mu-\lambda)i^2 - (\mu-2a-3\lambda)i + a\\
&\le -2(\mu-\lambda)(i+1)^2 + 4(\mu-\lambda)i + 2(\mu-\lambda) + (3\lambda+2a)i + a\\
&\le -\frac{1}{2}(\mu-\lambda)\omega^2(i)+b_2
\end{aligned} \tag{4.12}$$

$$\begin{aligned}
&\sum_{j\in s}\omega^3(j)q(j\,|\,i,a)\\
&= -3(\mu-\lambda)i^3 + 3(3\lambda-\mu+a)i^2 - (\mu-7\lambda-3a)i + a\\
&= -3(\mu-\lambda)(i+1)^3 + 9(\mu-\lambda)i^2 + 9(\mu-\lambda)i\\
&\quad + 3(3\lambda-\mu+a)i^2 - (\mu-7\lambda-3a)i + a\\
&\le -(\mu-\lambda)\omega^3(i)+b_3
\end{aligned} \tag{4.13}$$

$$q^*\omega^2(i) \le (\mu+\lambda)\omega''(i) \quad \forall i \in S \tag{4.14}$$

Let $\omega''(i) := \omega^3(i) = (i+1)^3, c_2 := \frac{1}{2}(\mu-\lambda)$, b_2 as above, $c'' := (\mu-\lambda)$, $b'' = b_3$ and $L_2 = \mu+\lambda$, by (4.10)-(4.14), we see that Assumption 1(d) holds.

Finally, we verify Assumption4, by (4.2)-(4.3) we have, for each fixed $f \in F$,

$$\sum_{j\geq k} q\left(j \mid i, f(i)\right) \leq \sum_{j\geq k} q\left(j \mid i+1, f(i+1)\right) \quad \forall i,k \in S, k \neq i+1$$

which, together with Proposition C.16 of [3], implies that the corresponding Markov process $X(t)$ is stochastically ordered. Thus, Assumption 4(a) follows from Proposition 7.6. Similarly, the assumption 4(b) is established.

Assumption 1, 2, 3 and 4 holds. So, the Proposition is proved.

5. Discussion and Conclusion

In this article, it defines the variance optimization problem of continuous-time Markov decision processes, which is different from the mean-variance optimization problem previously studied. By defining pseudo-variance, the deviation of the deterministic stationary policy f and the Poisson equation, a series of concepts and theorems, we prove the existence of the variance optimal strategy in the deterministic stationary policy space, and give the policy iterative algorithm to calculate optimal policy. Finally we prove the convergence of the policy iterative algorithm.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Puterman, M.L. (1994) Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons Ltd., New York. https://doi.org/10.1002/9780470316887

[2] Liu, K. and Cao, P. (2015) Theory and Application of Markov Decision Processes. Science Press, Beijing.

[3] Guo, X.P. and Hernandez-lerma, O. (2009) Continuous-Time Markov Decision Processes. Springer, Berlin, Heidelberg.

[4] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91.

[5] Bielecki, T.R., Jin, H.Q., Pliska, S.R. and Zhou, X.Y. (2005) Continuous-Time Mean-Variance Portfolio Selection with Bankruptcy Prohibition. *Mathematical Finance*, **15**, 213-244. https://doi.org/10.1111/j.0960-1627.2005.00218.x

[6] Costa, O.L.V., Maiali, A.C. and de Pinto, A.D.C. (2010) Sampled Control for Mean-Variance Hedging in a Jump Diffusion Financial Market. *IEEE Transactions on Automatic Control*, **55**, 1704-1709. https://doi.org/10.1109/TAC.2010.2046923

[7] Fu, C.P., Lari-Lavassani, A. and Li, X. (2010) Dynamic Mean-Variance Portfolio Selection with Borrowing Constraint. *European Journal of Operations Research*, **200**, 312-319. https://doi.org/10.1016/j.ejor.2009.01.005

[8] Markowitz, H.M. (1987) Mean-Variance Analysis in Portfolio Choice and Capital Markets. Basil Blackwell, Oxford, UK.

[9] Xiong, J. and Zhou, X.Y. (2007) Man-Variance Portfolio Selection under Partial Information. *SIAM Journal on Control and Optimization*, **46**, 156-175. https://doi.org/10.1137/050641132

[10] Yin, G. and Zhou, X.Y. (2004) Markowitz's Mean-Variance Portfolio Selection with Regime Switching: From Discrete-Time Models to Their Continuous-Time Limits. *IEEE Transactions on Automatic Control*, **49**, 349-360. https://doi.org/10.1109/TAC.2004.824479

[11] Zhou, X.Y. and Li, D. (2000) Continuous-Time Mean-Variance Portfolio Selection: A Stochastic LQ Framework. *Applied Mathematics and Optimization*, **42**, 19-33. https://doi.org/10.1007/s002450010003

[12] Zhou, X.Y. and Yin, G. (2003) Markowitz's Mean-Variance Portfolio Selection with Regime Switching: A Continuous-Time Model. *SIAM Journal on Control Optimization*, **42**, 1466-1482. https://doi.org/10.1137/S0363012902405583

[13] Mannor, S. and Tsitsiklis, J.N. (2013) Algorithmic Aspects of Mean-Variance Optimization in Markov Decision Processes. *European Journal of Operational Research*, **231**, 645-653. https://doi.org/10.1016/j.ejor.2013.06.019

[14] Guo, X.P. and Song, X.Y. (2009) Mean-Variance Criteria for Finite Continuous-Time Markov Decision Processes. *IEEE Transactions on Automatic Control*, **54**, 2151-2157. https://doi.org/10.1109/TAC.2009.2023833

[15] Guo, X.P., Ye, L.E. and George, Y. (2012) A Mean-Variance Optimization Problem for Discounted Markov Decision Processes. *European Journal of Operational Research*, **220**, 423-429. https://doi.org/10.1016/j.ejor.2012.01.051

[16] Ye, L.E. and Huang, X.X. (2014) Variance Optimization for Continuous-Time Markov Decision Processes. *China Science Journal*, **44**, 883-898.

[17] Xia, L. (2016) Optimization of Markov Decision Processes under the Variance Criterion. *Automatica*, **73**, 269-278. https://doi.org/10.1016/j.automatica.2016.06.018

[18] Xia, L. (2018) Mean-Variance Optimization of Discrete Time Discounted Markov Decision Processes. *Automatica*, **88**, 76-82. https://doi.org/10.1016/j.automatica.2017.11.012