Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

An Iterative Approach to Reduce the Variance of Stochastic Dynamic Systems ⋆

Li Xia *

* *Center for Intelligent and Networked Systems (CFINS), Department of Automation, TNList, Tsinghua University, Beijing 100084, China (e-mail: xial@tsinghua.edu.cn).*

Keywords: Markov decision process, variance criterion, policy iteration, gradient-based optimization, discrete event systems

Abstract: In this paper, we study the variance optimization problem in Markov decision processes (MDP). The objective is to find the optimal policy which has the minimal average variance of the system rewards. As the variance function is quadratic and the variance of rewards are correlated mutually, the associated variance minimization problem is not a linear program. The traditional approaches of classical MDP theory, which are good at solving linear problems, are inapplicable to this problem. In this paper, we define a fundamental quantity called variance potential and derive a variance difference equation which quantifies the difference of variances of Markov systems under any two policies. Based on the variance difference equation, we propose an iterative algorithm, which is similar to the policy iteration in classical MDP theory, to reduce the reward variance of Markov systems. Although this algorithm converges to a local optimum, it is very efficient compared with the traditional gradient-based algorithms. Numerical experiments demonstrate the main idea of this paper.

1. INTRODUCTION

Markov decision process (MDP) is an important theory to study the performance optimization of stochastic dynamic systems and it has been richly studied in the literature [5, 7, 16, 18]. In the literature of MDP, many studies focus on the performance optimization of MDP under the long-run average or discounted performance criteria. Much less literature studies the optimization of MDP under the *variance* criterion. However, variance is an important performance metric of stochastic systems and it reflects the risk-related factors of systems. For example, in financial engineering, the investors want to optimize their portfolios to minimize the risk of their investments while keeping their average returns above a certain amount, which is called the *mean-variance* optimization of portfolio management [13, 14, 20].

The current studies of MDP under the mean and variance criterion mainly have two categories. In the field of machine learning, studies focus on the two-objectives optimization of mean and variance [15, 19]. The optimization goal is to minimize the variance while keeping the mean performance above a certain level, or to maximize the mean performance while keeping the variance under a certain level, or to take the variance as a penalty factor and to maximize the combined objective function. Gradient-based approaches are usually applied to find the optimal policy or parameters [19]. However, these studies suffer from the intrinsic deficiencies of gradient-based methods, such as the slow convergence speed, difficulty of selecting step-size, and being trapped into a local optimum. On the other hand, in the research field of MDP, studies usually target on the variance minimization problem within an optimal policy set where the system average or discounted performance already achieves maximum [3, 16], or within a policy set where the average performance is not less than a given value [4, 17], or within a policy set with a given discounted performance [8, 9]. However, all these studies require that the variance should be minimized within a given optimal policy set. There is no study to directly minimize the variance of MDP within the entire policy space.

It is theoretically and practically meaningful to minimize the variance of stochastic systems regardless of the mean performance. For example, in a process control plant, it is important to control the reaction process steadily to reduce the quality variation of products [10, 11]. The difficulty of this problem is mainly because that the cost function of MDP under the variance criterion is *nonlinear*. Denote $r(i,a)$ as the reward of a Markov system at state i with action a adopted. The cost function of this MDP under the variance criterion is denoted as $f(i,a) = [r(i,a) - \eta]^2$, where η is the long-run average reward of this Markov system. Therefore, the cost function of this MDP is with a quadratic form, which indicates that the optimization of this MDP is a nonlinear problem. As we know, a standard MDP problem under the average criterion can be formulated as a linear program [16], $\max_{x(i,a)} \sum_{i\in\mathcal{S},a\in\mathcal{A}} r(i,a)x(i,a)$, subject to some linear constraints (i.e., transition probability conservation equations, $\sum_{i\in\mathcal{S},a\in\mathcal{A}} x(i,a) = 1$, and

⋆ This work was supported in part by the National Natural Science Foundation of China (61203039, U1301254), the National 111 International Collaboration Project (B06002), the Specialized Research Fund for the Doctoral Program of Higher Education (20120002120009).

$x(i,a) \geq 0$), where $x(i,a)$'s are optimization variables. However, for an MDP problem under the variance criterion, the associated programming formulation is

$$\min_{x(i,a)} \left\{ \sum_{i\in\mathcal{S},a\in\mathcal{A}} r^2(i,a)x(i,a) - \left[\sum_{i\in\mathcal{S},a\in\mathcal{A}} r(i,a)x(i,a) \right]^2 \right\},$$

subject to the same linear constraints as above. Obviously, this programming problem is not linear. It is a quadratic programming problem which is not necessarily convex. The traditional approaches in MDP theory, such as the policy iteration or the value iteration, do not fit this problem. This partly explains the difficulty of this variance minimization problem.

In this paper, we use the direct-comparison theory, which was proposed by X.-R. Cao [1, 2], to study the variance minimization problem of MDP. The key idea of the direct-comparison theory is the performance difference equation. Difference equation clearly describes the relation between the Markov system performance and the policies or parameters. The direct-comparison theory is widely valid for Markov systems, no matter the cost function is linear, instant, or not. For the variance minimization problem, we define a fundamental quantity called *variance potential* which quantifies the long-term accumulated deviation of reward variances. Based on the variance potential, we derive a variance difference equation, which quantifies the difference of variances of Markov systems under any two different policies. With the variance difference equation, we derive a *necessary condition* for the optimal policy under the variance criterion. An iterative algorithm similar to the policy iteration is further developed to efficiently reduce the variance of MDP. Although our algorithm also converges to a local optimum, it is usually more efficient compared with the traditional gradient-based approach. Finally, we conduct numerical experiments to demonstrate the efficiency of our approach.

2. VARIANCE CRITERION OF MARKOV DECISION PROCESSES

Consider a Markov chain $\boldsymbol{X} = \{X_t, t = 0, 1, \cdots\}$, where X_t is the system state at time epoch t. The state space $\mathcal{S}$ is assumed finite. Without loss of generality, we define the state space as $\mathcal{S} = \{1, 2, \cdots, S\}$, where S is the size of the state space. When the system is at state i, we can choose an action a from the action space $\mathcal{A}(i)$, $i \in \mathcal{S}$. For simplicity, we assume that the action spaces at different states are identical, i.e., $\mathcal{A}(i) = \mathcal{A}$, $\forall i \in \mathcal{S}$. We assume $\mathcal{A}$ is finite and $\mathcal{A} = \{a_1, a_2, \cdots, a_A\}$, where A is the size of action space $\mathcal{A}$. After an action $a \in \mathcal{A}$ is adopted at state i, the system state will transit to state j at the next time epoch with a transition probability $p^a(i,j)$, $i, j \in \mathcal{S}$. Meanwhile, the system will get an instant reward denoted as $r(i,a)$. The transition probability matrix $\boldsymbol{P}$ is an S-by-S matrix and its ith row jth column element is the transition probability $p^a(i,j)$, $i, j \in \mathcal{S}$. The reward function $\boldsymbol{r}$ is an S-dimensional column vector whose ith element is $r(i,a)$, $i \in \mathcal{S}$. In some places of this paper, we may assume that the reward is independent of the action adopted and we denote it as $r(i)$ for simplicity, $i \in \mathcal{S}$. The steady state distribution of the Markov system is denoted as an S-dimensional row vector $\boldsymbol{\pi} = (\pi(1), \pi(2), \cdots, \pi(S))$, where $\pi(i)$ is the probability that the system stays at state i, $i \in \mathcal{S}$. Obviously, we have $\boldsymbol{\pi P} = \boldsymbol{\pi}$, $\boldsymbol{Pe} = \boldsymbol{e}$, and $\boldsymbol{\pi e} = 1$, where $\boldsymbol{e}$ is an S-dimensional column vector whose elements are all 1. The long-run average performance of the Markov chain is denoted as η and we have

$$\eta = \boldsymbol{\pi r} = \lim_{T\to\infty} \frac{1}{T} E\left\{ \sum_{t=0}^{T-1} r(X_t) \right\}, \tag{1}$$

where we assume that the Markov chain is ergodic and η is independent of the initial state X_0.

According to the definition of variance in a stochastic process, we define the variance of a Markov chain as below [4, 16].

$$\eta_{var} = \lim_{T\to\infty} \frac{1}{T} E\left\{ \sum_{t=0}^{T-1} [r(X_t) - \eta]^2 \right\}. \tag{2}$$

In the optimization of MDP, we have to choose a policy which determines the action selection rule at different system states. We consider the stationary and deterministic policy. Thus, a policy $\mathcal{L}$ is a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$. $\mathcal{L}(i)$ indicates the action adopted at state i, $i \in \mathcal{S}$. The total policy space is denoted as Ψ, i.e., $\mathcal{L} \in \Psi$. Different policy $\mathcal{L}$ will affect the value of transition probability matrix, reward function, system performance, etc. We use the superscript "$\mathcal{L}$" to identify the effect of different policy, such as $\boldsymbol{P}^{\mathcal{L}}$, $\boldsymbol{r}^{\mathcal{L}}$, $\eta_{var}^{\mathcal{L}}$, etc. The optimization goal is to find the optimal policy $\mathcal{L}^*$ from the policy space Ψ, which minimizes the average variance of the Markov chain. That is,

$$\mathcal{L}^* = \arg\min_{\mathcal{L}\in\Psi} \left\{ \eta_{var}^{\mathcal{L}} \right\}. \tag{3}$$

As we discussed in Section 1, the variance criterion is nonlinear and this optimization problem is not necessarily *convex*. The traditional optimization approach of MDP theory, such as the policy iteration, cannot be directly applied to this problem. In the next section, we resort to other approaches to handle this problem.

3. ANALYSIS AND OPTIMIZATION

In this section, we derive the variance difference equation of MDP under any two different policies, which is similar to the direct-comparison theory in Markov systems [1, 2]. Based on the difference equation, we propose an iterative algorithm to efficiently reduce the variance of MDP. Some optimality properties and theorems of this variance minimization problem are also studied.

3.1 Variance Difference Equation

First, we define a fundamental quantity called *variance potential* of Markov systems.

$$g_{var}(i) = \lim_{T\to\infty} E\left\{ \sum_{t=0}^{T} [(r(X_t) - \eta)^2 - \eta_{var}] \Big| X_0 = i \right\}, \tag{4}$$

where η and η_{var} are the long-run average performance and the steady state variance defined in (1) and (2), respectively. From the above definition, we see that $g_{var}(i)$ quantifies the long-term accumulated deviations of $(r(X_t)-\eta)^2$ from the average variance η_{var} under the condition of the initial state i.

Extending the summation terms of (4) at time $t = 0$ and recursively substituting it, we have the following equation

$$g_{var}(i) = (r(i) - \eta)^2 - \eta_{var} + \sum_{j \in \mathcal{S}} p(i,j) g_{var}(j), \quad (5)$$

Furthermore, we can rewrite (5) in a matrix form as below.

$$\boldsymbol{g}_{var} = (\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} - \eta_{var} \boldsymbol{e} + \boldsymbol{P} \boldsymbol{g}_{var}, \quad (6)$$

where $\boldsymbol{g}_{var} := (g_{var}(1), g_{var}(2), \cdots, g_{var}(S))^T$ is an S-dimensional column vector, $(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot}$ means the componentwise square of each element of vector $(\boldsymbol{r} - \eta \boldsymbol{e})$, i.e., $(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} := ((r(1) - \eta)^2, (r(2) - \eta)^2, \cdots, (r(S) - \eta)^2)^T$.

With (4), we observe that the value of variance potential $\boldsymbol{g}_{var}$ can be estimated from the sample path of the Markov system under the current policy. This is because that the value of $r(X_t)$ is observed and the values of η and η_{var} can be estimated directly based on their definitions. On the other hand, we see that $\boldsymbol{g}_{var}$ can also be calculated by directly solving equation (6), since $\boldsymbol{r}$, $\boldsymbol{P}$, η, and η_{var} are either known parameters or calculable parameters. Please note, (6) is a non-deterministic equation. For a solution of $\boldsymbol{g}_{var}$ to (6), $\boldsymbol{g}_{var} + c\boldsymbol{e}$ is also a solution to (6) for any $c \in \mathcal{R}$. We can add one more equation $\boldsymbol{\pi} \boldsymbol{g}_{var} = 0$ or $\boldsymbol{g}^T_{var} \boldsymbol{e} = 0$ to make the solution of (6) deterministic.

Below, we discuss the difference of variance η_{var} of the Markov system under any two different policies $\mathcal{L}$ and $\mathcal{L}'$. For simplicity, we use the superscript "$'$" to indicate the parameters of Markov system under policy $\mathcal{L}'$. With (2), we see that the variance of Markov system with policy $\mathcal{L}$ can also be written as below.

$$\eta_{var} = \boldsymbol{\pi}(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot}. \quad (7)$$

Similarly, the variance of Markov system with policy $\mathcal{L}'$ is written as

$$\eta'_{var} = \boldsymbol{\pi}'(\boldsymbol{r}' - \eta' \boldsymbol{e})^2_{\odot}. \quad (8)$$

The goal of our analysis is to quantify the difference between η_{var} and η'_{var} with an equation. By right-multiplying $\boldsymbol{\pi}'$ on both sides of (6), we have

$$\begin{aligned} \boldsymbol{\pi}' \boldsymbol{g}_{var} &= \boldsymbol{\pi}'(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} - \boldsymbol{\pi}' \eta_{var} \boldsymbol{e} + \boldsymbol{\pi}' \boldsymbol{P} \boldsymbol{g}_{var} \\ &= \boldsymbol{\pi}'(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} - \eta_{var} + \boldsymbol{\pi}' \boldsymbol{P} \boldsymbol{g}_{var}. \end{aligned} \quad (9)$$

Substituting $\boldsymbol{\pi}' \boldsymbol{P}' = \boldsymbol{\pi}'$ and (8) into the above equation, we have

$$\begin{aligned} \eta'_{var} - \eta_{var} &= \boldsymbol{\pi}'(\boldsymbol{r}' - \eta' \boldsymbol{e})^2_{\odot} + \boldsymbol{\pi}' \boldsymbol{P}' \boldsymbol{g}_{var} - \boldsymbol{\pi}' \boldsymbol{P} \boldsymbol{g}_{var} \\ &\quad - \boldsymbol{\pi}'(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} \\ &= \boldsymbol{\pi}'(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_{var} + \boldsymbol{\pi}' \boldsymbol{r}'^2_{\odot} - \eta'^2 - \boldsymbol{\pi}'(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot}. \end{aligned} \quad (10)$$

We can further rewrite (10) as

$$\begin{aligned} \eta'_{var} - \eta_{var} &= \boldsymbol{\pi}'(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_{var} + \boldsymbol{\pi}'(\boldsymbol{r}' - \eta \boldsymbol{e})^2_{\odot} - (\eta' - \eta)^2 \\ &\quad - \boldsymbol{\pi}'(\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot}. \end{aligned} \quad (11)$$

Therefore, we obtain the following *variance difference equation* of MDP under any two different policies $\mathcal{L}$ and $\mathcal{L}'$

$$\begin{aligned} \eta'_{var} - \eta_{var} = \boldsymbol{\pi}' \Big[&(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_{var} + (\boldsymbol{r}' - \eta \boldsymbol{e})^2_{\odot} \\ &- (\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} \Big] - (\eta' - \eta)^2. \end{aligned} \quad (12)$$

The above equation gives a clear description of the relation between the variance and the policy (represented by $\boldsymbol{P}$ and $\boldsymbol{r}$) in MDP. For the current policy, we can calculate or estimate the long-run average performance η and the variance potential $\boldsymbol{g}_{var}$. Since the Markov chain is assumed ergodic, $\pi'(i)$'s are always positive for any policy and $i \in \mathcal{S}$. If we choose a new policy with proper $\boldsymbol{P}'$ and $\boldsymbol{r}'$, which makes the element of the column vector represented by the square bracket in (12) negative, then we see that $\boldsymbol{\pi}' \left[(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_{var} + (\boldsymbol{r}' - \eta \boldsymbol{e})^2_{\odot} - (\boldsymbol{r} - \eta \boldsymbol{e})^2_{\odot} \right] < 0$. Since $(\eta' - \eta)^2$ is always non-negative, we have $\eta'_{var} - \eta_{var} < 0$ and the variance of the Markov system under this new policy is reduced. This is the basic idea which we utilize to construct an iterative algorithm to reduce the variance of MDP in Subsection 3.2.

Remark 1. Variance difference equation (12) avoids the difficulty of computing the value of η' since $(\eta' - \eta)^2$ is always non-negative. Otherwise, computing η' under every possible policy $\mathcal{L}'$ is computationally cumbersome, which is equivalent to the enumerative comparison of every policy.

Based on the variance difference equation, we directly derive the following theorem.

Theorem 1. If we choose a new policy $\mathcal{L}'$ with $\boldsymbol{P}'$ and $\boldsymbol{r}'$ which satisfies $\sum_{j \in \mathcal{S}} p'(i,j) g_{var}(j) + (r'(i) - \eta)^2 \leq \sum_{j \in \mathcal{S}} p(i,j) g_{var}(j) + (r(i) - \eta)^2$ for all $i \in \mathcal{S}$, then we have $\eta'_{var} \leq \eta_{var}$. If the inequality strictly holds ($<$) for at least one $i \in \mathcal{S}$, then we have $\eta'_{var} < \eta_{var}$.

This theorem is very straightforward with (12) and we omit the detailed proof for simplicity. Based on (12), we further derive the following *necessary condition* of the optimal policy with the minimal variance.

Theorem 2. For the variance minimization problem, the optimal policy $\mathcal{L}$ with $\boldsymbol{P}$ and $\boldsymbol{r}$ must satisfy $\sum_{j \in \mathcal{S}} p'(i,j)$ $g_{var}(j) + (r'(i) - \eta)^2 \geq \sum_{j \in \mathcal{S}} p(i,j) g_{var}(j) + (r(i) - \eta)^2$ for any $i \in \mathcal{S}$ and $\mathcal{L}' \in \Psi$.

The proof of this theorem is also omitted for the limit of space.

Remark 2. The condition listed in Theorem 2 is only a *necessary* condition of optimal policy for the variance minimization problem of MDP. It is not a *sufficient* condition.

Remark 3. When the long-run average performance of all the policies in a given policy space is the same, the condition listed in Theorem 2 is a *necessary and sufficient* condition of optimal policy with minimal variance.

With Remark 3, the variance difference equation (12) can be used to thoroughly solve the following special variance minimization problem

$$\min_{\mathcal{L} \in \Psi^{\lambda}} \{\eta^{\mathcal{L}}_{var}\}, \quad \text{subject to } \eta^{\mathcal{L}} = \lambda, \quad (13)$$

where Ψ^{λ} is a policy set with the same long-run average performance as λ. Ψ^{λ} should be decomposable as $A^{\lambda}(1) \times A^{\lambda}(2) \times \cdots \times A^{\lambda}(S)$, where $A^{\lambda}(i)$ is a set of actions at state i and any policy chooses actions from $A^{\lambda}(i)$ has the same average performance λ, $i \in \mathcal{S}$. For this optimization problem (13), the variance difference equation (12) can be rewritten as below.

$$\eta'_{var}-\eta_{var} = \boldsymbol{\pi}' \left[(\boldsymbol{P}' - \boldsymbol{P})\boldsymbol{g}_{var} + (\boldsymbol{r}' - \eta\boldsymbol{e})^2_{\odot} - (\boldsymbol{r} - \eta\boldsymbol{e})^2_{\odot} \right], \quad (14)$$

where the term $(\eta' - \eta)^2$ disappears since $\eta' = \eta$ for all the policies in Ψ^{λ}. As we will discuss later in Subsection 3.2, a policy iteration type algorithm can be developed to find the global optimal solution of this optimization problem.

3.2 Iterative Optimization Algorithm

With (12) and Theorem 1, we can construct the following iterative algorithm to reduce the variance of Markov chain.

Algorithm 1. Policy iteration type algorithm to reduce the variance of Markov chain.

Initialization

- Arbitrarily choose an initial policy $\mathcal{L}^{(0)}$ from the policy space Ψ, set $l = 0$.

Policy Evaluation

- For the current policy $\mathcal{L}^{(l)}$, calculate or estimate η, η_{var}, and $\boldsymbol{g}_{var}$ based on the system sample path.

Policy Update

- Update the policy as follows:

  $$\mathcal{L}^{(l+1)}(i) = \arg\min_{a\in\mathcal{A}} \sum_{j\in\mathcal{S}} p^a(i,j) g_{var}(j) + (r(i,a) - \eta)^2,$$

  for all $i \in \mathcal{S}$ and keep $\mathcal{L}^{(l+1)}(i) = \mathcal{L}^{(l)}(i)$ if possible.

Stopping Rule

- If $\mathcal{L}^{(l+1)} = \mathcal{L}^{(l)}$, stop; Otherwise, set $l := l + 1$ and go to step 2.

We can see that the above iterative algorithm is similar to the policy iteration for long-run average performance in the classical MDP theory. Therefore, the above algorithm also possesses the similar properties of the policy iteration, such as the fast convergence speed in most of the situations.

Theorem 3. When the reward function $\boldsymbol{r}$ remains unvaried under different policies, Algorithm 1 converges to a policy with a local minimal variance in the randomized policy space.

The proof of this theorem is also omitted for the limit of space.

Remark 4. Algorithm 1 can find the global optimal solution of the optimization problem (13). This is because that the output policy of Algorithm 1 satisfies the necessary and sufficient condition of the optimal policy, as indicated by Remark 3.

For a general variance minimization problem unlike (13), although Algorithm 1 cannot be guaranteed to find the global optimal policy, Theorem 3 indicates that it converges to a local optimum when $\boldsymbol{r}$ is independent of policies. In the literature, there are some studies using the gradient-based approach to minimize the variance of MDP [12, 19]. The gradient-based approach also converges to a local optimum. However, compared with the gradient-based approach in the literature, the policy iteration type approach in Algorithm 1 usually has a much faster convergence speed, which is demonstrated by the numerical experiment in the next section.

4. NUMERICAL EXPERIMENT

In this section, we use a toy example to demonstrate the effectiveness of our approach for the variance minimization problem of Markov chains.

Consider a small Markov chain with 3 states, i.e., $\mathcal{S} = \{1, 2, 3\}$. The action space is $\mathcal{A} = \{a_1, a_2, a_3\}$. At every state, we can choose an action from the action space. A policy is represented as a 3-element row vector, such as $\mathcal{L} = [a_2, a_3, a_1]$ which selects action a_2, a_3, and a_1 at state 1, 2, and 3, respectively. Obviously, the size of the policy space is $|\Psi| = |\mathcal{A}|^{|\mathcal{S}|} = 3^3 = 27$. Different actions induce different state transition probabilities. For state 1, we have $p^{a_1}(1,:) = (0.8, 0.1, 0.1)$, $p^{a_2}(1,:) = (0.1, 0.7, 0.2)$, $p^{a_3}(1,:) = (0.1, 0.3, 0.6)$. For state 2, we have $p^{a_1}(2,:) = (0.7, 0.1, 0.2)$, $p^{a_2}(2,:) = (0.1, 0.8, 0.1)$, $p^{a_3}(2,:) = (0.1, 0.1, 0.8)$. For state 3, we have $p^{a_1}(3,:) = (0.6, 0.3, 0.1)$, $p^{a_2}(3,:) = (0.2, 0.6, 0.2)$, $p^{a_3}(3,:) = (0, 0.1, 0.9)$. The reward function is $\boldsymbol{r} = (10, 1, 2)^T$ which is unvaried under different policies.

We use the policy iteration type algorithm to minimize the variance of this Markov chain. We arbitrarily choose an initial policy from the policy space and use Algorithm 1 to reduce the variance of Markov chain. As Theorem 3 states, Algorithm 1 converges to a local optimum. For different initial policy, Algorithm 1 may converge to different local optimum. Since the policy space is small, we enumerate every initial policy and find that Algorithm 1 truly converges to two possible local optima, one is $\hat{\mathcal{L}}^* = [a_1, a_1, a_1]$ and the other is $\mathcal{L}^* = [a_3, a_3, a_3]$. Policy $\hat{\mathcal{L}}^*$ indicates that the actions at all the states are a_1 and the corresponding average performance and variance are $\hat{\eta}^* = 8$ and $\hat{\eta}^*_{var} = 13.1020$, respectively. Policy $\mathcal{L}^*$ indicates that the actions at all the states are a_3 and the corresponding average performance and variance are $\eta^* = 1.988$ and $\eta^*_{var} = 0.8294$, respectively. We can see that the average performance of $\mathcal{L}^*$ is worsen than that of $\hat{\mathcal{L}}^*$, while the variance of $\mathcal{L}^*$ is less than that of $\hat{\mathcal{L}}^*$. Therefore, $\mathcal{L}^*$ is the global optimal policy of Markov chain under the variance criterion.

From the experiment results, we observe that Algorithm 1 converges to the local optimum $\hat{\mathcal{L}}^*$ under three initial policies, as listed in Table 1. For all the other initial policies, Algorithm 1 will converge to the other local optimum $\mathcal{L}^*$ (it is also the global optimum), where some of the optimization processes are listed in Table 2. From these tables, we observe that the variance of Markov chain is strictly reduced during the optimization process, while the average performance has no regular trends. This demonstrates the effectiveness of Algorithm 1 for reducing the variance of Markov chains. From the experiment results, we observe that Algorithm 1 needs only 1 or 2 iterations to converge in most of the situations (the worst

Table 1. The optimization process of Algorithm 1 with different initial policy, which converges to the local optimum.

l	$\mathcal{L}^{(l)}$	η	η_{var}	l	$\mathcal{L}^{(l)}$	η	η_{var}	l	$\mathcal{L}^{(l)}$	η	η_{var}
0	$[a_1, a_1, a_1]$	8.0000	13.1020	0	$[a_1, a_1, a_2]$	7.4824	15.2850	0	$[a_1, a_3, a_1]$	7.1628	15.9037
1	$[a_1, a_1, a_1]$	8.0000	13.1020	1	$[a_1, a_1, a_1]$	8.0000	13.1020	1	$[a_1, a_1, a_1]$	8.0000	13.1020
				2	$[a_1, a_1, a_1]$	8.0000	13.1020	2	$[a_1, a_1, a_1]$	8.0000	13.1020

Table 2. The optimization process of Algorithm 1 with different initial policy, which converges to the global optimum.

l	$\mathcal{L}^{(l)}$	η	η_{var}	l	$\mathcal{L}^{(l)}$	η	η_{var}	l	$\mathcal{L}^{(l)}$	η	η_{var}
0	$[a_1, a_2, a_3]$	3.0000	10.0000	0	$[a_2, a_3, a_1]$	3.9350	14.8408	0	$[a_2, a_2, a_1]$	2.5368	10.5434
1	$[a_3, a_3, a_3]$	1.9886	0.8294	1	$[a_2, a_2, a_3]$	1.9524	3.4739	1	$[a_2, a_2, a_2]$	2.1348	7.9369
2	$[a_3, a_3, a_3]$	1.9886	0.8294	2	$[a_3, a_3, a_3]$	1.9886	0.8294	2	$[a_2, a_2, a_3]$	1.9524	3.4739
				3	$[a_3, a_3, a_3]$	1.9886	0.8294	3	$[a_3, a_3, a_3]$	1.9886	0.8294
								4	$[a_3, a_3, a_3]$	1.9886	0.8294

case is 3 iterations). This demonstrates the efficiency of Algorithm 1.

5. DISCUSSION AND CONCLUSION

The variance minimization problem of MDP does not fit the standard model of MDP. The traditional approaches of MDP theory, such as the policy iteration, cannot be directly applied to this problem. In this paper, we derive a variance difference equation to study this problem from a viewpoint of direct comparison. By directly comparing the reward variance and the policies or parameters of MDP, we propose a policy iteration type approach for this variance minimization problem. The optimality properties and related theorems are also derived. Compared with the traditional gradient-based approaches which are often used in the literature, our approach is more efficient.

In the future work, it is valuable to further study the policy iteration type approach to optimize the performance of MDP, considering both the average performance and the variance of system rewards. Our approach provides a promising research direction for the study of the mean-variance optimization problem. Moreover, the policy iteration type approach in this paper converges to the local optimum. How to identify the conditions with which our approach can converge to the global optimum is another future research topic.

REFERENCES

[1] X. R. Cao and H. F. Chen, "Potentials, perturbation realization, and sensitivity analysis of Markov processes," *IEEE Transactions on Automatic Control*, Vol. 42, pp. 1382-1393, 1997.

[2] X. R. Cao, *Stochastic Learning and Optimization – A Sensitivity-Based Approach*, New York: Springer, 2007.

[3] X. R. Cao and J. Zhang, "The nth-order bias optimality for multi-chain Markov decision processes," *IEEE Transactions on Automatic Control*, Vol. 53, No. 2, pp. 496-508, 2008.

[4] K. J. Chung, "Mean-variance tradeoffs in an undiscounted MDP: the unichain case," *Operations Research*, Vol. 42, No. 1, pp. 184-188, 1994.

[5] E. Feinberg and A. Schwartz, *Handbook of Markov Decision Processes: Methods and Applications*, Boston, MA: Kluwer Academic Publishers, 2002.

[6] R. C. Grinold and R. N. Kahn R N, *Active Portfolio Management*, New York: McGraw-Hill, 2000.

[7] X. Guo and O. Hernandez-Lerma, *Continuous-Time Markov Decision Processes: Theory and Applications*, Springer, 2009.

[8] X. Guo and X. Y. Song, "Mean-variance criteria for finite continuous-time Markov decision processes," *IEEE Transactions on Automatic Control*, Vol. 54, pp. 2151-2157, 2009.

[9] X. Guo, L. Ye, and G. Yin, "A mean-variance optimization problem for discounted Markov decision processes," *European Journal of Operational Research*, Vol. 220, pp. 423-429, 2012.

[10] C. A. Harrison and S. J. Qin, "Minimum variance performance map for constrained model predictive control," *Journal of Process Control*, Vol. 19, pp. 1199-1204, 2009.

[11] B. Huang, "Minimum variance control and performance assessment of time-variant processes," *Journal of Process Control*, Vol. 12, pp. 707-719, 2002.

[12] S. R. Kuindersma, R. A. Grupen, and A. G. Barto, "Variable risk control via stochastic optimization," *The International Journal of Robotics Research*, Vol. 32, pp. 806-825, 2013.

[13] H. Markowitz, "Portfolio selection," *The Journal of Finance*, Vol. 7, pp. 77-91, 1952.

[14] H. Markowitz, "Portfolio selection: efficient diversification of investments," Cowles Foundation Monograph No. 16, 1959.

[15] S. Mannor and J. N. Tsitsiklis, "Mean-variance optimization in Markov decision processes," *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: John Wiley & Sons, 1994.

[17] M. J. Sobel, "Mean-variance tradeoffs in an undiscounted MDP," *Operations Research*, Vol. 42, pp. 175-183, 1994.

[18] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*, Princeton, NJ: Princeton University Press, 2009.

[19] A. Tamar, D. D. Castro, and S. Mannor, "Policy gradients with variance related risk criteria," *Proceedings of*

the 29th International Conference on Machine Learning (ICML), Edinburgh, Scotland, 2012.

[20] X. Y. Zhou and G. Yin, "Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model," *SIAM Journal on Control and Optimization*, Vol. 42, pp. 1466-1482, 2004.