Automatica 73 (2016) 269–278

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Optimization of Markov decision processes under the variance criterion☆

Li Xia [1]

CFINS, Department of Automation, TNList, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 23 August 2015
Received in revised form
21 May 2016
Accepted 3 June 2016

Keywords:
Markov decision process
Variance criterion
Sensitivity-based optimization
Policy iteration
Policy gradient

ABSTRACT

In this paper, we study a variance minimization problem in an infinite stage discrete time Markov decision process (MDP), regardless of the mean performance. For the Markov chain under the variance criterion, since the value of the cost function at the current stage will be affected by future actions, this problem is not a standard MDP and the traditional MDP theory is not applicable. In this paper, we convert the variance minimization problem into a standard MDP by introducing a concept called pseudo variance. Then we derive a variance difference formula that quantifies the difference of variances of Markov systems under any two policies. With the difference formula, the correlation of the variance cost function at different stages can be decoupled through a nonnegative term. A necessary condition of the optimal policy is obtained. It is also proved that the optimal policy with the minimal variance can be found in the deterministic policy space. Furthermore, we propose an efficient iterative algorithm to reduce the variance of Markov systems. We prove that this algorithm can converge to a local optimum. Finally, a numerical experiment is conducted to demonstrate the efficiency of our algorithm compared with the gradient-based method widely adopted in the literature.

1. Introduction

Markov decision processes (MDPs) are widely used to model and analyze the performance optimization in stochastic dynamic systems (Chang, Fu, Hu, & Marcus, 2007; Feinberg & Schwartz, 2002; Puterman, 1994). In the literature on MDPs, many studies focus on the performance optimization under the long-run average or discounted performance criterion. Much less attention has been paid to the *variance* criterion. However, variance is an important performance metric of stochastic systems and it reflects risk-related factors. For example, in financial engineering, the risk-averse investors optimize their portfolios to minimize the risk of their investments while keeping an acceptable return, which is called *mean–variance optimization* of portfolio management (Markowitz, 1952; Zhou & Yin, 2004). In the process control of plants, the minimum variance control is used to control the reaction process steadily and reduce the quality variation of products (Harrison & Qin, 2009; Huang, 2002).

The mean–variance optimization always considers the mean and variance simultaneously. In the field of machine learning, research is conducted to develop optimization algorithms that minimize the variance while keeping the mean performance above a certain level, or maximize the mean performance while keeping the variance under a certain level, or take the variance as a penalty factor and maximize the combined objective function. Policy gradient algorithms are proposed to find a local optimum (Prashantha & Ghavamzadeh, 2013; Tamar, Castro, & Mannor, 2012). However, these studies suffer from the intrinsic deficiencies of gradient-based methods, such as the slow convergence speed, the difficulty of selecting step-size, the sensitivity to the initial point, and the possibility of being trapped into a local optimum. Mannor and Tsitsiklis (2013) further study the algorithmic complexity of a special form of mean–variance optimization in finite horizon and prove that such problem is NP-hard in some cases. In the field of MDPs, studies usually focus on the variance minimization within an optimal policy set in which the mean performance already achieves optimum (Cao & Zhang, 2008; Guo & Song, 2009; Hernandez-Lerma,

☆ This work was supported in part by the National Natural Science Foundation of China (61573206, 61203039, U1301254), the Key R&D Project of China (2016YFB0901902), and the National 111 International Collaboration Project (B06002). The material in this paper was partially presented at the 19th World Congress of the International Federation of Automatic Control (IFAC 2014), August 24–29, 2014, Cape Town, South Africa (Xia, 2014). This paper was recommended for publication in revised form by Associate Editor Bart De Schutter under the direction of Editor Christos G. Cassandras.

E-mail address: xial@tsinghua.edu.cn.

[1] Fax: +86 10 62796115.

http://dx.doi.org/10.1016/j.automatica.2016.06.018
0005-1098/

Vega-Amaya, & Carrasco, 1999). When the average or discounted performance is already optimal, the variance criterion can be transformed to an equivalent average or discounted criterion (Guo, Ye, & Yin, 2012; Huang & Chen, 2012) and the traditional MDP approaches, such as the policy iteration, are still applicable. However, it is not clear how to develop a policy iteration type algorithm to minimize the variance when the mean performance is not optimal.

In this paper, we aim to study the variance minimization problem in a discrete time MDP, regardless of the mean performance. Different from the mean–variance optimization studied in the literature, we focus only on the variance criterion. This problem is of both theoretical and engineering significance. From the theoretical viewpoint, our work can contribute to a more systematic study of MDP theory under various criteria, besides the widely used average and discounted criterion. From the application viewpoint, the variance minimization problem has practical background in engineering systems. For example, in a smart grid integrated with wind farms, a lot of wind electricity is abandoned because of its large variation (Ummels, Gibescu, Pelgrum, Kling, & Brand, 2007). Energy storage systems can be used to reduce the variation of wind electricity. It is of importance to optimize the scheduling policy such that the variation of electricity output from the storage system to the smart grid is minimized, while the wind abandonment is avoided. This problem can be modeled as a typical variance minimization problem.

The difficulty of the variance minimization problem is mainly caused by the quadratic form of the variance. We observe that the cost function under the variance criterion is $f(i,a) = (r(i,a) - \eta)^2$, where $r(i,a)$ is the system reward at the current stage with state i and action a, and η is the mean performance. A standard MDP requires that both the cost function and the state transition probability should be Markovian. The cost should be an instant cost and its value at the current stage should not be affected by future stages (see page 20 in the book of Puterman, 1994). However, in the variance minimization problem, the cost function $f(i,a)$ will be affected by future actions. This is because $r(i',a')$ at future stages will affect the value of η, consequently affect the value of $(r(i,a) - \eta)^2$ at the current stage. Therefore, the values of $f(i,a)$ at different stages are coupled and this variance minimization problem is not a standard MDP. The Bellman optimality equation that is critical for dynamic programming does not hold for this problem. The traditional MDP approaches, such as the policy iteration or the value iteration, are not applicable to this problem. The difficulty of this problem is also pointed out by Sobel (1982) that a general optimization problem considering variance metrics is not amenable to the policy iteration algorithm.

In this paper, we define a quantity called *pseudo variance* $f_\lambda(i,a) = (r(i,a) - \lambda)^2$, where λ is a given constant. Obviously, the value of the pseudo variance at the current stage will not be affected by future actions and the pseudo variance minimization problem is a standard MDP. We prove that the policy improvement for the pseudo variance also reduces the variance of Markov systems. Therefore, the original variance minimization problem is converted into a standard MDP of minimizing the pseudo variance. Then we derive a *variance difference formula* that clearly describes the relation between the variance and the policy adopted. This formula can decouple the correlation of variances at different stages, through a nonnegative term $(\eta' - \eta)^2$. It gives a new direction to study the variance minimization problem of MDP from the sensitivity viewpoint. With the variance difference formula, we derive a *necessary condition* for the optimal policy under the variance criterion. We also prove that the optimal policy with the minimal variance can be found in the deterministic policy space, which is not trivial since this problem does not fit a standard MDP formulation. For the variance criterion with mean performance as constraint, the optimal policy cannot be always achieved in the deterministic policy space (Puterman, 1994). An iterative algorithm similar to policy iteration is further developed to efficiently reduce the variance of Markov systems. This algorithm is proved to converge to a *local optimum*. Although the policy gradient approaches widely adopted in the literature also converge to a local optimum, our policy iteration algorithm has a much faster convergence speed, which is demonstrated in numerical experiments. To the best of our knowledge, this is the first work that provides a policy iteration type algorithm to minimize the variance of Markov systems. In the literature, the previous works either study the policy iteration to minimize the variance after the average or discounted performance already achieves optimum or study the policy gradient algorithm to approach to the local minimum of variance. Our approach provides a promising way to directly minimize the variance of Markov systems, regardless of the mean performance. Compared with our conference paper (Xia, 2014), this journal paper makes substantial contributions, especially on the pseudo variance and related proofs.

The remainder of the paper is organized as follows. In Section 2, we give a mathematical formulation for the variance minimization problem. In Section 3, we convert this problem into a standard MDP of minimizing the pseudo variance. The variance difference formula is further derived. Some optimality properties are also obtained. A policy iteration type algorithm is then developed to efficiently reduce the variance of Markov systems. In Section 4, we conduct numerical experiments to demonstrate the efficiency of our approach. Finally, we conclude this paper in Section 5.

2. Problem formulation

Consider a discrete time Markov chain $\boldsymbol{X} := \{X_t, t = 0, 1, \ldots\}$, where X_t is the system state at time epoch t. The state space $\mathcal{S}$ is assumed finite and we denote it as $\mathcal{S} := \{1, 2, \ldots, S\}$, 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 $\mathcal{A} := \{a_1, a_2, \ldots, a_A\}$, where A is the size of action space $\mathcal{A}$. After an action a is adopted at state i, the system state will transit to state j at the next time epoch with a transition probability $p^a(i,j)$, $i, j \in \mathcal{S}$ and $a \in \mathcal{A}$. Meanwhile, the system will obtain an instant reward denoted as $r(i,a)$. All the transition probabilities of the Markov chain compose an S-by-S matrix denoted as $\boldsymbol{P}$. For notation simplicity, we may also use $r(i)$ to replace $r(i,a)$ when the action selection rule is determined. We further denote the reward function $\boldsymbol{r}$ as an S-dimensional column vector composed by element $r(i)$, $i \in \mathcal{S}$. The steady state distribution of the Markov chain is denoted as an S-dimensional row vector $\boldsymbol{\pi} := (\pi(1), \pi(2), \ldots, \pi(S))$, where $\pi(i)$ is the probability of the system staying at state i, $i \in \mathcal{S}$. Obviously, we have $\boldsymbol{\pi P} = \boldsymbol{\pi}$, $\boldsymbol{P1} = \boldsymbol{1}$, and $\boldsymbol{\pi 1} = 1$, where $\boldsymbol{1}$ is an S-dimensional column vector with all elements being 1. The long-run average performance of the Markov chain is defined as below.

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

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

According to the definition of the variance in a stochastic process, we define the *steady state variance* of an ergodic Markov chain as below (Chung, 1994; Sobel, 1994).

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

In the literature, there exists another definition of variance that originally comes from finite stage Markov chains. For a Markov chain with T stages, the total reward cumulated at the final stage is denoted as

$$C_T := \sum_{t=0}^{T-1} r(X_t). \tag{3}$$

Note that C_T is a random variable. The variance of C_T is defined as below (Mannor & Tsitsiklis, 2013; Zhou & Yin, 2004).

$$\mathbb{VAR}[C_T] = \mathbb{E}\left\{\sum_{t=0}^{T-1} r(X_t) - \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t)\right]\right\}^2. \tag{4}$$

Based on (4), another definition of the variance of an infinite stage MDP is as below (Guo et al., 2012; Hernandez-Lerma et al., 1999).

$$\begin{aligned}\eta_{\sigma l} &:= \lim_{T\to\infty} \frac{1}{T}\mathbb{VAR}[C_T] \\ &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1} r(X_t) - \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t)\right]\right\}^2.\end{aligned} \tag{5}$$

We call $\eta_{\sigma l}$ the *limiting cumulative variance* of the Markov chain. Note that the above discussion is for an undiscounted Markov chain, the similar analysis is also valid for discounted one.

Both of these two different definitions of variances exist in the literature, but there is little discussion about their relations. In this paper, we derive that $\eta_{\sigma l}$ and $\eta_{\sigma s}$ have the following relation (details can be referred to Appendix)

$$\eta_{\sigma l} = \eta_{\sigma s} + \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0,k\neq t}^{T-1} \mathbb{COV}(r(X_t), r(X_k)), \tag{6}$$

where $\mathbb{COV}(X, Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$ is the covariance of two random variables X and Y. From the above definitions, we see that $\eta_{\sigma l}$ focuses on the fluctuation of rewards at the final stage while $\eta_{\sigma s}$ focuses more on the fluctuations during the whole process. In many practical problems, people concern the fluctuation of rewards not only at the final stage, but also during the whole process. For example, in the asset management of financial engineering, if the fluctuations of asset returns are very big during the contract period, the investment may be withdrawn by risk-averse investors before the end of the contract. The investors cannot tolerate a big risk during the asset management process, even though the variance of returns at the end of the contract is small. Moreover, the estimation variance of $\eta_{\sigma s}$ is usually much smaller than that of $\eta_{\sigma l}$. Therefore, we choose $\eta_{\sigma s}$ as the variance definition of MDP in this paper. For notation simplicity, we use η_σ to replace $\eta_{\sigma s}$ in the rest of the paper.

In the optimization of MDPs, we have to choose a policy that determines the action selection rule at different states. We consider a 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 $\boldsymbol{P}^{\mathcal{L}}$, $\boldsymbol{\pi}^{\mathcal{L}}$, $\boldsymbol{r}^{\mathcal{L}}$, $\eta_\sigma^{\mathcal{L}}$, $\eta^{\mathcal{L}}$, etc. For notation simplicity, we usually omit the superscript "$^{\mathcal{L}}$" by default and use the superscript "$'$" instead of "$^{\mathcal{L}'}$" in the rest of the paper if applicable.

The optimization objective is to find the optimal policy $\mathcal{L}^*$ from the policy space Ψ, which minimizes the steady state variance of the Markov chain. We assume that the Markov chain is *ergodic* under any policy in Ψ. The cost function under the variance criterion and policy $\mathcal{L}$ is defined as

$$f(i) := (r(i, \mathcal{L}(i)) - \eta)^2, \tag{7}$$

where η is the long-run average performance defined in (1). We denote $\boldsymbol{f}$ as an S-dimensional column vector composed by elements $f(i), i \in \mathcal{S}$. The variance minimization problem of Markov chains can be defined as below.

$$\mathcal{L}^* = \operatorname*{argmin}_{\mathcal{L}\in\Psi}\{\eta_\sigma\} = \operatorname*{argmin}_{\mathcal{L}\in\Psi}\{\boldsymbol{\pi f}\}. \tag{8}$$

From (7), we see that the value of $f(i)$ will be affected by future actions through affecting the value of η. Therefore, the cost function is not Markovian and the problem (8) is not a standard MDP. We have to develop new approaches to optimize this problem.

3. Analysis and optimization

In this section, we first define a pseudo variance minimization problem and prove the equivalence between this problem and the original problem (8). The difficulty of non-Markovian structure in (8) is avoided. Then we derive the variance difference and derivative formulas and some interesting optimality properties. Finally, we develop an iterative algorithm to efficiently reduce the variance of Markov systems.

3.1. Pseudo variance minimization

From (8), we see that the difficulty mainly comes from the fact that the value of η is unknown and affected by future actions. To handle this difficulty, we define a new cost function as below.

$$f_\lambda(i) := (r(i, \mathcal{L}(i)) - \lambda)^2, \tag{9}$$

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

$$\boldsymbol{f}_\lambda := (\boldsymbol{r} - \lambda\boldsymbol{1})_\odot^2, \tag{10}$$

where $(\boldsymbol{r} - \lambda\boldsymbol{1})_\odot^2$ means the component-wise square of vector $(\boldsymbol{r} - \lambda\boldsymbol{1})$, i.e.,

$$(\boldsymbol{r} - \lambda\boldsymbol{1})_\odot^2 := ((r(1) - \lambda)^2, \ldots, (r(S) - \lambda)^2)^T. \tag{11}$$

The long-run average performance with $\boldsymbol{f}_\lambda$ is defined as the *pseudo variance* of the Markov chain and we have

$$\eta_{\sigma,\lambda} = \boldsymbol{\pi f}_\lambda. \tag{12}$$

Obviously, the pseudo variance $\eta_{\sigma,\lambda}$ is different from the variance η_σ defined in (2) and we have

$$\eta_{\sigma,\lambda} = \eta_\sigma, \quad \text{when } \lambda = \eta. \tag{13}$$

We define the pseudo variance minimization problem in a Markov chain as below.

$$\begin{aligned}\mathcal{L}_\lambda^* &= \operatorname*{argmin}_{\mathcal{L}\in\Psi}\left\{\eta_{\sigma,\lambda}\right\} = \operatorname*{argmin}_{\mathcal{L}\in\Psi}\{\boldsymbol{\pi f}_\lambda\} \\ &= \operatorname*{argmin}_{\mathcal{L}\in\Psi}\sum_{i\in\mathcal{S}} \pi(i)(r(i, \mathcal{L}(i)) - \lambda)^2.\end{aligned} \tag{14}$$

Since $\boldsymbol{f}_\lambda$ is an instant cost and it has no relation to future actions, the above problem (14) is a standard MDP. Below, we study the relation between these two problems (8) and (14). First, we have the following lemma about the relation between $\eta_{\sigma,\lambda}$ and η_σ.

Lemma 1. *The variance and the pseudo variance of a Markov chain have the following relation*

$$\eta_{\sigma,\lambda} = \eta_\sigma + (\eta - \lambda)^2. \tag{15}$$

Proof. From the definition of the pseudo variance in (12) and (9), we have

$$\begin{aligned}\eta_{\sigma,\lambda} &= \sum_{i\in\mathcal{S}} \pi(i)(r(i,\mathcal{L}(i))-\lambda)^2 \\ &= \sum_{i\in\mathcal{S}} \pi(i)(r(i,\mathcal{L}(i))-\eta)^2 \\ &\quad + \sum_{i\in\mathcal{S}} \pi(i)\left(2r(i,\mathcal{L}(i))\eta - 2r(i,\mathcal{L}(i))\lambda - \eta^2+\lambda^2\right).\end{aligned}$$

Substituting (1) and (2) into the above equation, we have

$$\eta_{\sigma,\lambda} = \eta_\sigma + 2\eta^2 - 2\eta\lambda - \eta^2 + \lambda^2 = \eta_\sigma + (\eta-\lambda)^2. \tag{16}$$

The lemma is proved. □

Then, we discuss how to solve the pseudo variance minimization problem (14). Since (14) is a standard MDP, we can use the policy iteration to solve (14). We denote $\boldsymbol{g}_{\sigma,\lambda}$ as the *performance potential* (or called *relative value function*) of the MDP problem (14). $\boldsymbol{g}_{\sigma,\lambda}$ is an S-dimensional column vector and its element $g_{\sigma,\lambda}(i)$, $i\in\mathcal{S}$, is defined as below.

$$\begin{aligned}g_{\sigma,\lambda}(i) &:= \lim_{T\to\infty}\mathbb{E}\left\{\sum_{t=0}^{T}[f_\lambda(X_t)-\eta_{\sigma,\lambda}]\Big|X_0=i\right\} \\ &= \lim_{T\to\infty}\mathbb{E}\left\{\sum_{t=0}^{T}[(r(X_t)-\lambda)^2-\eta_{\sigma,\lambda}]\Big|X_0=i\right\}.\end{aligned} \tag{17}$$

From the above definition, we can see that $g_{\sigma,\lambda}(i)$ quantifies the long-term cumulated deviations of $f_\lambda(X_t)$ away from $\eta_{\sigma,\lambda}$, under the initial condition of $X_0=i$. By extending the summation terms of (17) at time $t=0$, we obtain the following equation after recursive institution

$$g_{\sigma,\lambda}(i) = (r(i)-\lambda)^2 - \eta_{\sigma,\lambda} + \sum_{j\in\mathcal{S}} p(i,j)g_{\sigma,\lambda}(j). \tag{18}$$

We write the above equation in matrix form and obtain

$$\boldsymbol{g}_{\sigma,\lambda} = (\boldsymbol{r}-\lambda\mathbf{1})^2_\odot - \eta_{\sigma,\lambda}\mathbf{1} + \boldsymbol{P}\boldsymbol{g}_{\sigma,\lambda}. \tag{19}$$

We can numerically compute the value of $\boldsymbol{g}_{\sigma,\lambda}$ with (19) or estimate it with (17). For more details, audience can refer to Chapter 3 of the book (Cao, 2007).

Below, we discuss the difference of pseudo variance $\eta_{\sigma,\lambda}$ of the Markov chain under any two different policies $\mathcal{L}$ and $\mathcal{L}'$. For simplicity, we use the superscript "′" to indicate the parameters of the Markov chain under policy $\mathcal{L}'$. With (10) and (12), we see that the pseudo variance of the Markov chain under policy $\mathcal{L}$ can be written as

$$\eta_{\sigma,\lambda} = \boldsymbol{\pi}(\boldsymbol{r}-\lambda\mathbf{1})^2_\odot. \tag{20}$$

Similarly, the pseudo variance of the Markov chain under policy $\mathcal{L}'$ can be written as

$$\eta'_{\sigma,\lambda} = \boldsymbol{\pi}'(\boldsymbol{r}'-\lambda\mathbf{1})^2_\odot. \tag{21}$$

Left-multiplying $\boldsymbol{\pi}'$ on both sides of (19), we obtain

$$\boldsymbol{\pi}'\boldsymbol{P}'\boldsymbol{g}_{\sigma,\lambda} = \boldsymbol{\pi}'(\boldsymbol{r}-\lambda\mathbf{1})^2_\odot - \eta_{\sigma,\lambda} + \boldsymbol{\pi}'\boldsymbol{P}\boldsymbol{g}_{\sigma,\lambda}, \tag{22}$$

where we use the fact $\boldsymbol{\pi}'\boldsymbol{P}'=\boldsymbol{\pi}'$ and $\boldsymbol{\pi}'\mathbf{1}=1$. Subtracting (22) from (21), we obtain the *difference formula* of the pseudo variance of Markov chains under any two policies $\mathcal{L}$ and $\mathcal{L}'$.

$$\eta'_{\sigma,\lambda} - \eta_{\sigma,\lambda} = \boldsymbol{\pi}'\Big[(\boldsymbol{P}'-\boldsymbol{P})\boldsymbol{g}_{\sigma,\lambda} + (\boldsymbol{r}'-\lambda\mathbf{1})^2_\odot - (\boldsymbol{r}-\lambda\mathbf{1})^2_\odot\Big]. \tag{23}$$

The difference formula is the key result of the sensitivity-based optimization theory (Cao, 2007; Cao & Chen, 1997) and it quantifies the change of the performance of Markov systems when the policy or the parameters are changed. Based on (23), we can directly construct the policy iteration to solve (14) as follows. For the current policy $\mathcal{L}$, we obtain the value of $\boldsymbol{g}_{\sigma,\lambda}$ by estimation with (17) or by computation with (19). Then we can do the policy improvement to obtain a better policy $\mathcal{L}'$ as follows. For every state $i\in\mathcal{S}$, we have

$$\mathcal{L}'(i) = \operatorname*{argmin}_{a\in\mathcal{A}}\left\{(r(i,a)-\lambda)^2 + \sum_{j\in\mathcal{S}} p^a(i,j)g_{\sigma,\lambda}(j)\right\}. \tag{24}$$

This is one round of the policy iteration. Repeat this process iteratively until the policy cannot be improved anymore. Then the output policy is optimal.

Although the above policy iteration is for (14) that is different from the original problem (8), we have the following theorem to describe the relation between these two problems.

Theorem 1. *For any policy $\mathcal{L}\in\Psi$, we compute η with (1) and set $\lambda=\eta$. If we obtain an improved policy $\mathcal{L}'$ such that $\eta'_{\sigma,\lambda}\le\eta_{\sigma,\lambda}$, then we have $\eta'_\sigma\le\eta_\sigma$. If the first inequality strictly holds ($<$), then we have $\eta'_\sigma<\eta_\sigma$.*

Proof. With Lemma 1, we have

$$\begin{aligned}\eta_{\sigma,\lambda} &= \eta_\sigma + (\eta-\lambda)^2. \\ \eta'_{\sigma,\lambda} &= \eta'_\sigma + (\eta'-\lambda)^2.\end{aligned} \tag{25}$$

Subtracting the first equation from the second one and substituting $\lambda=\eta$, we obtain

$$\eta'_\sigma - \eta_\sigma = \eta'_{\sigma,\lambda} - \eta_{\sigma,\lambda} - (\eta'-\eta)^2. \tag{26}$$

Obviously, if $\eta'_{\sigma,\lambda}\le\eta_{\sigma,\lambda}$, then $\eta'_\sigma\le\eta_\sigma$; if $\eta'_{\sigma,\lambda}<\eta_{\sigma,\lambda}$, then $\eta'_\sigma<\eta_\sigma$. The theorem is proved. □

With Theorem 1, we see that the pseudo variance minimization problem (14) is equivalent to the variance minimization problem (8) to some extent. We further have the following corollary.

Corollary 1. *For any policy $\mathcal{L}\in\Psi$, we compute η with (1) and set $\lambda=\eta$. If we do one round of the policy iteration for problem (14) and obtain a new policy $\mathcal{L}'$ as indicated by (24), then the original problem (8) is also improved under the new policy $\mathcal{L}'$, i.e., we have $\eta'_\sigma\le\eta_\sigma$.*

Proof. Substituting (25) into (23), we obtain

$$\begin{aligned}\eta'_\sigma - \eta_\sigma &= \boldsymbol{\pi}'\left[(\boldsymbol{P}'-\boldsymbol{P})\boldsymbol{g}_{\sigma,\lambda} + (\boldsymbol{r}'-\lambda\mathbf{1})^2_\odot - (\boldsymbol{r}-\lambda\mathbf{1})^2_\odot\right] \\ &\quad - (\eta'-\lambda)^2 + (\eta-\lambda)^2.\end{aligned} \tag{27}$$

Since $\lambda=\eta$, we have

$$\begin{aligned}\eta'_\sigma - \eta_\sigma &= \boldsymbol{\pi}'\left[(\boldsymbol{P}'-\boldsymbol{P})\boldsymbol{g}_{\sigma,\lambda} + (\boldsymbol{r}'-\eta\mathbf{1})^2_\odot - (\boldsymbol{r}-\eta\mathbf{1})^2_\odot\right] \\ &\quad - (\eta'-\eta)^2.\end{aligned} \tag{28}$$

When we obtain an improved policy $\mathcal{L}'$ as indicated by (24), we know that all the elements in the column vector represented by the square bracket in (28) are nonpositive. Since all the elements of $\boldsymbol{\pi}'$ are always positive, we have

$$\eta'_\sigma - \eta_\sigma \le 0 - (\eta'-\eta)^2 \le 0. \tag{29}$$

The corollary is proved. □

With Theorem 1 and Corollary 1, we see that although the original problem (8) is not a standard MDP, we can construct a standard MDP (14) to minimize the pseudo variance. When the pseudo variance is reduced by the policy iteration (24), the variance in (8) is also reduced. This indicates an effective way to minimize the variance of Markov systems. Below, we further study the variance minimization of Markov systems along this direction.

3.2. Variance difference formula

In this subsection, we derive the variance difference formula for the original problem (8). Some interesting optimality properties also follow directly.

In (28), the performance potential $\boldsymbol{g}_{\sigma,\lambda}$ can be further specified since we set $\lambda = \eta$. Similar to (17), we define a fundamental quantity called *variance potential* for the Markov chain under the variance criterion.

$$g_\sigma(i) := \lim_{T\to\infty} \mathbb{E}\left\{\sum_{t=0}^{T}[(r(X_t)-\eta)^2-\eta_\sigma]\Big|X_0=i\right\}, \tag{30}$$

where η and η_σ are the long-run average performance and the steady state variance defined in (1) and (2), respectively. Similar to (19), we also obtain a recursive equation of $\boldsymbol{g}_\sigma$ in matrix form as below.

$$\boldsymbol{g}_\sigma = (\boldsymbol{r}-\eta\mathbf{1})^2_\odot - \eta_\sigma\mathbf{1} + \boldsymbol{P}\boldsymbol{g}_\sigma, \tag{31}$$

where $\boldsymbol{g}_\sigma := (g_\sigma(1), g_\sigma(2), \ldots, g_\sigma(S))^T$ is an S-dimensional column vector.

Therefore, with the definition of $\boldsymbol{g}_\sigma$ and (28), we directly derive the *variance difference formula* of the Markov chain under any two different deterministic policies $\mathcal{L}$ and $\mathcal{L}'$

$$\eta'_\sigma - \eta_\sigma = \boldsymbol{\pi}'\Big[(\boldsymbol{P}'-\boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{r}'-\eta\mathbf{1})^2_\odot - (\boldsymbol{r}-\eta\mathbf{1})^2_\odot\Big] - (\eta'-\eta)^2. \tag{32}$$

The above formula is very fundamental to study the variance minimization problem (8). It gives a clear description of the relation between the reward variance and the policy (represented by $\boldsymbol{P}$ and $\boldsymbol{r}$). The term $(\eta'-\eta)^2$ is the most important feature of (32) as it is always nonnegative. For any policy $\mathcal{L}$, we can compute or estimate the values of η and $\boldsymbol{g}_\sigma$. Since the Markov chain is assumed ergodic, $\pi'(i)$'s are always positive. If we choose a new policy with proper $\boldsymbol{P}'$ and $\boldsymbol{r}'$ such that the elements of the column vector represented by the square bracket in (32) are negative, then we see that $\boldsymbol{\pi}'\left[(\boldsymbol{P}'-\boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{r}'-\eta\mathbf{1})^2_\odot - (\boldsymbol{r}-\eta\mathbf{1})^2_\odot\right] < 0$. Since $(\eta'-\eta)^2$ is always nonnegative, we have $\eta'_\sigma - \eta_\sigma < 0$ and the variance of the Markov chain under this new policy is reduced. This is the basic idea that we utilize later to develop an iterative algorithm to reduce the variance of Markov chains in Section 3.5.

Remark 1. In (32), $(\eta'-\eta)^2$ is always nonnegative even though the value of η' is unknown. It avoids the difficulty of computing the value of η' under every possible $\mathcal{L}' \in \Psi$, of which the computation complexity is equivalent to that of a brute-force enumeration algorithm for (8).

With (32), we can mathematically quantify the change of the reward variance of Markov chains when the policy is changed. Similar to Theorem 1 and Corollary 1, we derive the following theorem to reduce the variance of Markov chains.

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

This theorem is straightforward based on the variance difference formula (32). For simplicity, we omit the proof. Based on (32), we further derive the following *necessary condition* of the optimal policy with the minimal variance.

Theorem 3. *For the variance minimization problem (8), the optimal policy $\mathcal{L}$ with $\boldsymbol{P}$ and $\boldsymbol{r}$ must satisfy $\sum_{j\in\mathcal{S}} p'(i,j)g_\sigma(j) + (r'(i)-\eta)^2 \ge \sum_{j\in\mathcal{S}} p(i,j)g_\sigma(j) + (r(i)-\eta)^2$ for any $i \in \mathcal{S}$ and $\mathcal{L}' \in \Psi$.*

Proof. We use the contradiction method to prove this theorem. Assume the current policy $\mathcal{L}$ with $\boldsymbol{P}$ and $\boldsymbol{r}$ is optimal. If the condition in the theorem is not satisfied, it means that for some state, say state k, there exists an action a' such that $\sum_{j\in\mathcal{S}} p'(k,j)g_\sigma(j) + (r'(k)-\eta)^2 < \sum_{j\in\mathcal{S}} p(k,j)g_\sigma(j) + (r(k)-\eta)^2$. Therefore, we can construct a new policy $\mathcal{L}'$ as follows: for state k, choose action a'; for other states, choose the same action as that of the optimal policy $\mathcal{L}$. Therefore, with (32), we directly have

$$\eta'_\sigma - \eta_\sigma = \pi'(k)\Big[\sum_{j\in\mathcal{S}}(p'(k,j)-p(k,j))g_\sigma(j) + (r'(k)-\eta)^2 - (r(k)-\eta)^2\Big] - (\eta'-\eta)^2 < 0 - (\eta'-\eta)^2.$$

Therefore, $\eta'_\sigma - \eta_\sigma < 0$ and the policy $\mathcal{L}'$ is better than $\mathcal{L}$. This contradicts the assumption that $\mathcal{L}$ is the optimal policy. The theorem is proved. □

Remark 2. The condition listed in Theorem 3 is only a *necessary* condition of the optimal policy for the variance minimization problem of Markov chains. It is not a *sufficient* condition.

Remark 3. If the long-run average performance of the Markov chain under all the policies in Ψ is the same and Ψ can be represented as the product of the action set at every state, the condition listed in Theorem 3 is a *necessary and sufficient* condition of the optimal policy with the minimal variance.

In the literature, there are some studies to minimize the variance after the average or discounted performance is already maximized (Cao & Zhang, 2008; Guo & Song, 2009; Puterman, 1994). These problems satisfy the condition mentioned in Remark 3 and we can similarly use our approach to solve these problems in a straightforward way.

3.3. Variance derivative formula

Besides the variance difference formula derived in the last subsection, we further study the variance derivative formula in a randomized policy space (Schweitzer, 1968), which is used to prove the following Theorems 4 and 5. Suppose the current policy is $\mathcal{L}$ and the corresponding transition probability matrix and reward function are $\boldsymbol{P}$ and $\boldsymbol{r}$, respectively. Arbitrarily choose another policy $\mathcal{L}'$ with $\boldsymbol{P}'$ and $\boldsymbol{r}'$. Both $\mathcal{L}$ and $\mathcal{L}'$ are deterministic policies. We denote a randomized policy as $\mathcal{L}^\delta_{\mathcal{L}'}$, which means at each decision stage we adopt policy $\mathcal{L}'$ with probability δ and adopt policy $\mathcal{L}$ with probability $1-\delta$, where $0 \le \delta \le 1$. $\mathcal{L}^\delta_{\mathcal{L}'}$ is also called a *mixed policy* (Feinberg & Schwartz, 2002) that is mixed by $\mathcal{L}$ and $\mathcal{L}'$ with probability $1-\delta$ and δ, respectively. Obviously, we have $\mathcal{L}^0_{\mathcal{L}'} = \mathcal{L}$ and $\mathcal{L}^1_{\mathcal{L}'} = \mathcal{L}'$. Thus, the corresponding transition probability matrix of this randomized policy is

$$\boldsymbol{P}^\delta = \delta\boldsymbol{P}' + (1-\delta)\boldsymbol{P}. \tag{33}$$

We denote the average performance and the variance of the Markov chain under policy $\mathcal{L}^\delta_{\mathcal{L}'}$ as η^δ and η^δ_σ, respectively. With the randomized policy $\mathcal{L}^\delta_{\mathcal{L}'}$, we see that the cost function $f^\delta_\sigma(i)$ under the variance criterion is

$$f^\delta_\sigma(i) = \delta\left(r'(i)-\eta^\delta\right)^2 + (1-\delta)\left(r(i)-\eta^\delta\right)^2 \tag{34}$$

and the variance of the Markov chain is

$$\eta^\delta_\sigma = \sum_{i\in\mathcal{S}} \pi^\delta(i) f^\delta_\sigma(i) = \sum_{i\in\mathcal{S}} \pi^\delta(i)\left\{\delta\left(r'(i)-\eta^\delta\right)^2 + (1-\delta)\left(r(i)-\eta^\delta\right)^2\right\}.$$

We denote $\boldsymbol{f}_\sigma^\delta$ as an S-dimensional column vector with element $f_\sigma^\delta(i)$ and we have

$$\begin{aligned} \boldsymbol{f}_\sigma^\delta &= \delta(\boldsymbol{r}' - \eta^\delta \mathbf{1})_\odot^2 + (1-\delta)(\boldsymbol{r} - \eta^\delta \mathbf{1})_\odot^2. \\ \eta_\sigma^\delta &= \boldsymbol{\pi}^\delta \boldsymbol{f}_\sigma^\delta. \end{aligned} \tag{35}$$

The cost functions of the Markov chain under policy $\mathcal{L}$ and $\mathcal{L}'$ are written as below respectively.

$$\begin{aligned} \boldsymbol{f}_\sigma &= (\boldsymbol{r} - \eta \mathbf{1})_\odot^2. \\ \boldsymbol{f}_\sigma' &= (\boldsymbol{r}' - \eta' \mathbf{1})_\odot^2. \end{aligned} \tag{36}$$

Comparing (35) and (36), we see that $\boldsymbol{f}_\sigma^\delta$ cannot be represented as a linear combination of $\boldsymbol{f}_\sigma$ and $\boldsymbol{f}_\sigma'$, i.e., $\boldsymbol{f}_\sigma^\delta \neq \delta \boldsymbol{f}_\sigma' + (1-\delta)\boldsymbol{f}_\sigma$, because the variance cost function is nonlinear.

Premultiplying by $\boldsymbol{\pi}^\delta$ on both sides of (31), we obtain the following equation

$$\eta_\sigma = \boldsymbol{\pi}^\delta (\boldsymbol{r} - \eta \mathbf{1})_\odot^2 - \boldsymbol{\pi}^\delta \boldsymbol{P}^\delta \boldsymbol{g}_\sigma + \boldsymbol{\pi}^\delta \boldsymbol{P} \boldsymbol{g}_\sigma, \tag{37}$$

where we use the fact $\boldsymbol{\pi}^\delta \boldsymbol{P}^\delta = \boldsymbol{\pi}^\delta$ and $\boldsymbol{\pi}^\delta \mathbf{1} = 1$. Subtracting (37) from (35), we obtain (38) given in Box I, where we use $\boldsymbol{\pi}^\delta \mathbf{1} = 1$ for deriving the third equality. Furthermore, it is easy to verify that the average performance of the Markov chain under the randomized policy $\mathcal{L}_{\mathcal{L}'}^\delta$ is

$$\eta^\delta = \boldsymbol{\pi}^\delta \boldsymbol{r}^\delta = \boldsymbol{\pi}^\delta [\delta \boldsymbol{r}' + (1-\delta)\boldsymbol{r}]. \tag{39}$$

Substituting the above equation into (38), we have

$$\begin{aligned} \eta_\sigma^\delta - \eta_\sigma &= \boldsymbol{\pi}^\delta \left[(\boldsymbol{P}^\delta - \boldsymbol{P}) \boldsymbol{g}_\sigma + \delta (\boldsymbol{r}' - \eta \mathbf{1})_\odot^2 - \delta (\boldsymbol{r} - \eta \mathbf{1})_\odot^2 \right] \\ &\quad + 2\eta\eta^\delta - 2(\eta^\delta)^2 + (\eta^\delta)^2 - \eta^2 \\ &= \boldsymbol{\pi}^\delta \left[(\boldsymbol{P}^\delta - \boldsymbol{P}) \boldsymbol{g}_\sigma + \delta (\boldsymbol{r}' - \eta \mathbf{1})_\odot^2 - \delta (\boldsymbol{r} - \eta \mathbf{1})_\odot^2 \right] \\ &\quad - (\eta^\delta - \eta)^2. \end{aligned} \tag{40}$$

Substituting (33) into the above equation, we obtain the following variance difference formula for the randomized policy $\mathcal{L}_{\mathcal{L}'}^\delta$ and the deterministic policy $\mathcal{L}$

$$\begin{aligned} \eta_\sigma^\delta - \eta_\sigma &= \boldsymbol{\pi}^\delta \delta [(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_\sigma + (\boldsymbol{r}' - \eta \mathbf{1})_\odot^2 - (\boldsymbol{r} - \eta \mathbf{1})_\odot^2] \\ &\quad - (\eta^\delta - \eta)^2. \end{aligned} \tag{41}$$

Taking the differential operation with respect to δ on both sides of the above equation and letting $\delta \to 0$, we obtain the *variance derivative formula* of the Markov chain in the randomized policy space

$$\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\delta} = \boldsymbol{\pi} [(\boldsymbol{P}' - \boldsymbol{P}) \boldsymbol{g}_\sigma + (\boldsymbol{r}' - \eta \mathbf{1})_\odot^2 - (\boldsymbol{r} - \eta \mathbf{1})_\odot^2], \tag{42}$$

where we use the fact that $\boldsymbol{\pi}^\delta \to \boldsymbol{\pi}$ and $\frac{\mathrm{d}(\eta^\delta - \eta)^2}{\mathrm{d}\delta} = 2(\eta^\delta - \eta)\frac{\mathrm{d}\eta^\delta}{\mathrm{d}\delta} \to 0$ as $\delta \to 0$. The above formula quantifies the derivative of the variance of the Markov chain along a policy changing direction from $\mathcal{L}$ to $\mathcal{L}'$ in the randomized policy space.

3.4. Parametric randomized policy

In practice, the policy of Markov systems is often controlled by parameters. For example, we can control the value of $\theta_{i,a}$ that is the probability of choosing action a at state i, $i \in \mathcal{S}$ and $a \in \mathcal{A}$. Obviously, we have $0 \leq \theta_{i,a} \leq 1$ and $\sum_{a \in \mathcal{A}} \theta_{i,a} = 1$ for all $i \in \mathcal{S}$. The state transition probability is written as $p(i,j) = \sum_{a \in \mathcal{A}} p^a(i,j)\theta_{i,a}$, $i, j \in \mathcal{S}$. We call it *parametric randomized policy*. The objective is to determine the optimal values of parameters $\theta_{i,a}$'s such that the variance of the Markov chain is minimal. Tamar et al. (2012) made interesting studies on this problem from a perspective of the policy gradient method. Below, we study this problem from a new perspective with the variance difference formula (32).

For simplicity, we assume that only the probability $\theta_{i,a}$ at a particular state, say state k, is adjustable. The following analysis can be similarly applied to the case of multiple states since the variance difference formula (32) is obviously additive. Therefore, the optimization variables are $\theta_{k,a}$, $a \in \mathcal{A}$. The probabilities $\theta_{i,a}$'s at other states are fixed. We assume that the reward function $\boldsymbol{r}$ is unvaried for different policies, i.e., $\boldsymbol{r}' = \boldsymbol{r}$. With (32), we see that the variance difference of the Markov chain under two sets of parameters $\theta_{k,a}'$'s and $\theta_{k,a}$'s is

$$\begin{aligned} \eta_\sigma' - \eta_\sigma &= \pi'(k) \sum_{j \in \mathcal{S}} \sum_{a \in \mathcal{A}} p^a(k,j)(\theta_{k,a}' - \theta_{k,a}) g_\sigma(j) \\ &\quad - (\eta' - \eta)^2. \end{aligned} \tag{43}$$

With the above difference formula and the necessary condition in Theorem 3, we can derive the following theorem about the optimal parametric policy.

Theorem 4. *When the reward function is unvaried for different policies, the optimal policy with the minimal variance of Markov systems can be found in the deterministic policy space.*

Proof. For a particular state, say state k, we need to prove that the optimal action selection rule is deterministic. That is, the optimal parameter $\theta_{k,a}$ is either 0 or 1. Denote the vector of parameters $\theta_{k,a}$'s as $\boldsymbol{\theta}_k$, i.e., $\boldsymbol{\theta}_k := (\theta_{k,a_1}, \theta_{k,a_2}, \ldots, \theta_{k,a_A})$. Assume $\boldsymbol{\theta}_k^*$ is an optimal solution. With the necessary condition in Theorem 3, we have

$$\boldsymbol{\theta}_k^* \in \begin{cases} \operatorname{argmin}_{\boldsymbol{\theta}_k} \sum_{j \in \mathcal{S}} \sum_{a \in \mathcal{A}} p^a(k,j) g_\sigma^*(j) \theta_{k,a}, \\ \text{s.t.}, \sum_{a \in \mathcal{A}} \theta_{k,a} = 1, \quad \theta_{k,a} \geq 0, \ \forall a \in \mathcal{A}. \end{cases} \tag{44}$$

Since $p^a(k,j)$ is given and $g_\sigma^*(j)$ is known by computation or estimation, solving (44) is a linear programming. With the well known result of linear programming, the optimal $\boldsymbol{\theta}_k^*$ can be achieved at the vertex of the multidimensional polyhedron of $\boldsymbol{\theta}_k$, which is also the boundary of the value domain of $\boldsymbol{\theta}_k$. Since the value domain of $\theta_{k,a}$ is $[0, 1]$, the optimal $\theta_{k,a}^*$ can be 1 for a particular action a and be 0 for all the other actions, $a \in \mathcal{A}$. This means that the optimal policy can be deterministic and the randomized policy cannot dominate the deterministic policy for this variance minimization problem. The theorem is proved. □

The proof of the above theorem is concise. This is because we have the advantage of the variance difference formula and the necessary condition for this variance minimization problem. With the above theorem, we have the following remark.

Remark 4. The result in Theorem 4 is similar to that in the traditional MDP theory, where the optimal policy of a standard MDP can be found in the deterministic policy space under the average or discounted criterion.

Note that the cost function of the variance minimization problem is not Markovian and (8) is not a standard MDP. We cannot apply the result in MDP theory, such as the optimality of deterministic policy, to this variance minimization problem. However, as Theorem 4 states, we can prove that the optimality of deterministic policy is still valid for (8). With this theorem, we can focus our optimization effort on the deterministic policy, without considering the randomized policy. This can greatly reduce the optimization complexity. Note that Theorem 4 requires that the reward function $\boldsymbol{r}$ should be unvaried under different policies. Actually, this requirement can be relaxed and the optimality of

$$\begin{aligned}\eta_\sigma^\delta - \eta_\sigma &= \boldsymbol{\pi}^\delta\left[\delta(\boldsymbol{r}' - \eta^\delta\mathbf{1})_\odot^2 + (1-\delta)(\boldsymbol{r} - \eta^\delta\mathbf{1})_\odot^2\right] - \left[\boldsymbol{\pi}^\delta(\boldsymbol{r} - \eta\mathbf{1})_\odot^2 - \boldsymbol{\pi}^\delta\boldsymbol{P}^\delta\boldsymbol{g}_\sigma + \boldsymbol{\pi}^\delta\boldsymbol{P}\boldsymbol{g}_\sigma\right]\\ &= \boldsymbol{\pi}^\delta\left[(\boldsymbol{P}^\delta - \boldsymbol{P})\boldsymbol{g}_\sigma + \delta(\boldsymbol{r}' - \eta^\delta\mathbf{1})_\odot^2 - \delta(\boldsymbol{r} - \eta^\delta\mathbf{1})_\odot^2\right] + \boldsymbol{\pi}^\delta\left[(\boldsymbol{r} - \eta^\delta\mathbf{1})_\odot^2 - (\boldsymbol{r} - \eta\mathbf{1})_\odot^2\right]\\ &= \boldsymbol{\pi}^\delta\left[(\boldsymbol{P}^\delta - \boldsymbol{P})\boldsymbol{g}_\sigma + \delta(\boldsymbol{r}' - \eta\mathbf{1})_\odot^2 - \delta(\boldsymbol{r} - \eta\mathbf{1})_\odot^2\right] + \boldsymbol{\pi}^\delta\delta\left(2\boldsymbol{r}'\eta - 2\boldsymbol{r}\eta - 2\boldsymbol{r}'\eta^\delta + 2\boldsymbol{r}\eta^\delta\right) + (\eta^\delta)^2 - \eta^2 + \boldsymbol{\pi}^\delta\left(2\boldsymbol{r}\eta - 2\boldsymbol{r}\eta^\delta\right)\\ &= \boldsymbol{\pi}^\delta\left[(\boldsymbol{P}^\delta - \boldsymbol{P})\boldsymbol{g}_\sigma + \delta(\boldsymbol{r}' - \eta\mathbf{1})_\odot^2 - \delta(\boldsymbol{r} - \eta\mathbf{1})_\odot^2\right] + \boldsymbol{\pi}^\delta\left\{2\eta\left[\delta\boldsymbol{r}' + (1-\delta)\boldsymbol{r}\right] - 2\eta^\delta\left[\delta\boldsymbol{r}' + (1-\delta)\boldsymbol{r}\right]\right\} + (\eta^\delta)^2 - \eta^2. \quad (38)\end{aligned}$$

Box I.

deterministic policy is still valid. However, the proof is much more complicated and we omit it to keep this paper concise. More details can be found in another conference paper (Xia, 2016).

Based on (43), we directly obtain the variance derivative formula with respect to $\theta_{k,a}$.

$$\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\theta_{k,a}} = \pi(k)\sum_{j\in\mathcal{S}} p^a(k,j)g_\sigma(j). \quad (45)$$

In the literature, there are studies of policy gradient methods for the optimization of variance-related MDPs (Prashantha & Ghavamzadeh, 2013; Tamar et al., 2012) and they have similar forms of policy gradients to our derivative formulas (42) or (45). With these derivative formulas, we can numerically compute the performance gradients or estimate them based on sample paths. A gradient-descent algorithm can follow to update the optimization parameters. For example, with the performance gradient in (45), we can update $\theta_{k,a}^{l+1} := \theta_{k,a}^l - \beta_l\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\theta_{k,a}}$, where $\theta_{k,a}^l$ is the value of $\theta_{k,a}$ at the lth iteration and β_l is the step-size at the lth iteration, $l = 0, 1, \ldots$. With the stochastic approximation approach, the step-size satisfies the convergence condition, such as $\sum_{l=0}^\infty \beta_l = \infty$ and $\sum_{l=0}^\infty \beta_l^2 < \infty$. One example is $\beta_l = 1/l$. How to efficiently implement the online gradient estimation algorithm and the stochastic approximation algorithm has been extensively studied in the literature. In Section 4, we will implement a simple version of such gradient-based approach and compare it with our policy iteration type approach proposed in Section 3.5.

There is another derivative that is also widely used in practice. If the transition probability matrix $\boldsymbol{P}$ and the reward function $\boldsymbol{r}$ depend on a parameter θ, we denote them as $\boldsymbol{P}(\theta)$ and $\boldsymbol{r}(\theta)$, respectively. Suppose $\boldsymbol{P}$ and $\boldsymbol{r}$ are both differentiable w.r.t. θ. Similar to the above analysis, we can derive the following derivative formula based on (32)

$$\frac{\mathrm{d}\eta_\sigma(\theta)}{\mathrm{d}\theta} = \boldsymbol{\pi}(\theta)\left[\frac{\mathrm{d}\boldsymbol{P}(\theta)}{\mathrm{d}\theta}\boldsymbol{g}_\sigma(\theta) + 2(\boldsymbol{r}(\theta) - \eta(\theta)\mathbf{1})\odot\frac{\mathrm{d}\boldsymbol{r}(\theta)}{\mathrm{d}\theta}\right],$$

where $\odot$ is a component-wise product. The detailed derivation is omitted for simplicity and it deserves further studies when θ is a parameter vector.

3.5. Iterative optimization algorithm

With (32) and Theorem 2, we construct the following Algorithm 1 to reduce the variance of Markov chains.

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

Remark 5. Algorithm 1 transforms the original non-standard MDP (8) into a standard MDP with cost function $f(i,a) = (r(i,a) - \eta)^2$, where η is constant under different policies. With Corollary 1 and (32), we can see that a better policy derived from the new MDP also performs better for the original variance minimization problem (8).

Algorithm 1 (*Policy Iteration Type Algorithm to Reduce the Variance of Markov Chains*)**.**

- Initialization: Arbitrarily choose an initial policy $\mathcal{L}^{(0)}$ from the policy space Ψ, set $l = 0$.
- Evaluation: For the current policy $\mathcal{L}^{(l)}$, numerically compute or estimate η, η_σ, and $\boldsymbol{g}_\sigma$ based on their definitions (1), (2), and (31), respectively.
- Improvement: Update the policy as follows:

$$\mathcal{L}^{(l+1)}(i) = \operatorname*{argmin}_{a\in\mathcal{A}}\left\{(r(i,a) - \eta)^2 + \sum_{j\in\mathcal{S}} p^a(i,j)g_\sigma(j)\right\}, \quad (46)$$

  for all $i \in \mathcal{S}$ and keep $\mathcal{L}^{(l+1)}(i) = \mathcal{L}^{(l)}(i)$ if possible to avoid oscillations.
- Stopping Rule: If $\mathcal{L}^{(l+1)} = \mathcal{L}^{(l)}$, stop; Otherwise, let $l \leftarrow l+1$ and go to step 2.

Before we discuss the convergence of Algorithm 1, we give the following definition of the local optimum in the randomized policy space.

Definition 1. For a policy $\mathcal{L}$, if $\exists\Delta \in (0,1]$ and we always have $\eta_\sigma^{\mathcal{L}} \le \eta_\sigma^{\mathcal{L}^\delta_{\mathcal{L}'}}$, $\forall 0 < \delta < \Delta, \forall\mathcal{L}' \in \Psi$, then we say $\mathcal{L}$ is a local optimum of the variance in the randomized policy space.

With the above definition, we derive the convergence theorem for Algorithm 1 as follows.

Theorem 5. *Algorithm 1 converges to a policy with a local minimal variance in the randomized policy space.*

Proof. First, we prove that Algorithm 1 can converge in finite steps. From the policy improvement step in Algorithm 1, we see that $\eta_\sigma^{\mathcal{L}^{(l+1)}} \le \eta_\sigma^{\mathcal{L}^{(l)}}$ by applying Theorem 2. More specifically, we have $\eta_\sigma^{\mathcal{L}^{(l+1)}} < \eta_\sigma^{\mathcal{L}^{(l)}}$ when $\mathcal{L}^{(l+1)} \neq \mathcal{L}^{(l)}$. That is, the variance of the Markov chain will be strictly reduced at each iteration of Algorithm 1. Since the policy space Ψ is finite, it is obvious that Algorithm 1 can stop after finite steps.

Second, we prove that the policy converged to is a local optimum. We see that when Algorithm 1 stops, it indicates that for any $\boldsymbol{P}'$ we always have $\sum_{j\in\mathcal{S}} p'(i,j)g_\sigma(j) + (r(i,a') - \eta)^2 \ge \sum_{j\in\mathcal{S}} p(i,j)g_\sigma(j) + (r(i,a) - \eta)^2$. That is, for any $\boldsymbol{P}'$, the element of the column vector $[(\boldsymbol{P}' - \boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{r}' - \eta\mathbf{1})_\odot^2 - (\boldsymbol{r} - \eta\mathbf{1})_\odot^2]$ is always nonnegative. Since $\boldsymbol{\pi}$ is always positive, the value of the derivative formula (42) is always nonnegative. In other words, whatever we choose a policy changing direction in the randomized policy space, the derivative of the variance of the Markov chain is always nonnegative. With Definition 1 and the common knowledge of Taylor expansion, we see that the current solution is a local optimum. The theorem is proved. □

In the literature, there are some studies using the policy gradient approach to minimize the variance of MDPs (Kuindersma, Grupen, & Barto, 2013; Prashantha & Ghavamzadeh, 2013; Tamar et al., 2012). The policy gradient approach also converges to a local optimum. However, the policy iteration type approach in Algorithm 1 usually has a much faster convergence speed, which is demonstrated by numerical experiments in the next section.

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

l	$\mathcal{L}^{(l)}$	η	η_σ	l	$\mathcal{L}^{(l)}$	η	η_σ	l	$\mathcal{L}^{(l)}$	η	η_σ
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 policies, which converges to the global optimum.

l	$\mathcal{L}^{(l)}$	η	η_σ	l	$\mathcal{L}^{(l)}$	η	η_σ	l	$\mathcal{L}^{(l)}$	η	η_σ
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

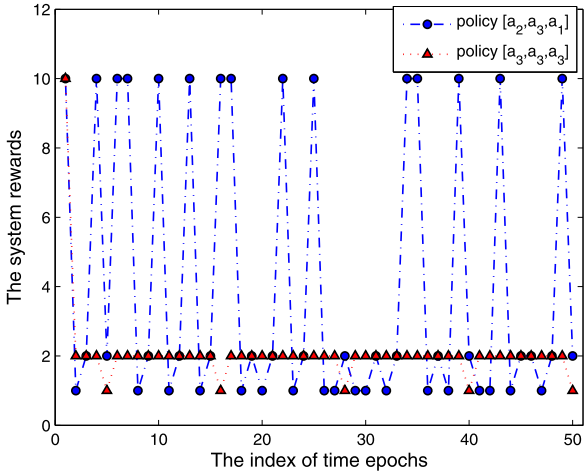

Fig. 1. The sample paths of rewards under two different policies.

4. Numerical experiment

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 choose an action from the action space. A policy is represented as a 3-element row vector, such as $\mathcal{L} = [a_2, a_3, a_1]$ that 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 $\boldsymbol{r} = (10, 1, 2)^T$ that is unvaried under different policies.

First, we illustrate that the fluctuations of the system rewards under different policies are generally different. We plot the sample paths of the system rewards under two different policies in Fig. 1. One policy is $[a_2, a_3, a_1]$ and the other one is $[a_3, a_3, a_3]$. From Fig. 1, we observe that the system reward under policy $[a_2, a_3, a_1]$ has much bigger fluctuations while the system reward under policy $[a_3, a_3, a_3]$ is more stable.

Then, we use Algorithm 1 to minimize the variance of this Markov chain. The initial policy is arbitrarily selected. As Theorem 5 states, Algorithm 1 converges to a local optimum. For different initial policies, Algorithm 1 may converge to different local optima. 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 $\hat{\eta}^*_\sigma = 13.1020$, 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 $\eta^*_\sigma = 0.8294$, respectively. We can see that the average performance of $\mathcal{L}^*$ is worse than that of $\hat{\mathcal{L}}^*$, while the variance of $\mathcal{L}^*$ is less than that of $\hat{\mathcal{L}}^*$. $\mathcal{L}^*$ is the global optimal policy of this 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 converges to the other local optimum $\mathcal{L}^*$ (it is also the global optimum). Some of the optimization processes are listed in Table 2. From these tables, we observe that the variance of the Markov chain is strictly reduced during each iteration, while the average performance has no regular trends. This demonstrates the effectiveness of Algorithm 1 for reducing the variance of Markov chains. We also observe that Algorithm 1 needs only 1 or 2 iterations to converge in most of situations (the worst case is 3 iterations). This demonstrates the fast convergence of Algorithm 1.

As a comparison, we use the gradient-based approach to minimize the variance of this Markov chain, which is widely adopted in the literature. The initial policy is arbitrarily chosen as $[a_2, a_3, a_1]$. Similar to the derivative formula (45), we compute the variance derivatives $\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\theta_{k,a}}$ for all $a \in \mathcal{A}$ and find the action a_m whose derivative is minimal, $k = 1, 2, 3$. Update the probability of selecting a_m as $\theta^{l+1}_{k,a_m} = \theta^l_{k,a_m} + \beta_l$, where the step-size is set as $\beta_l = 5/l$. Do normalization $\theta^{l+1}_{k,a} := \theta^{l+1}_{k,a} / \sum_{a\in\mathcal{A}} \theta^{l+1}_{k,a}$ to keep the sum of probability $\theta^{l+1}_{k,a}$'s over $a \in \mathcal{A}$ equal to 1. Repeating this process until the maximal gap (infinity norm $|\cdot|_\infty$) between two successive updates is no more than 0.1% in ratio. Fig. 2 illustrates the curves of the probabilities during the optimization process. Comparing with the result in the middle column of Table 2, we observe that the final optimization results of these two approaches are the same. However, the convergence speed of the gradient-based approach is much slower. This comparison also demonstrates the efficiency of our policy iteration type algorithm.

5. Discussion and conclusion

Since the cost function under the variance criterion is not Markovian, the variance minimization problem of Markov systems is not a standard MDP. The traditional approaches of MDPs, such as the policy iteration, cannot be directly applied to this problem. To avoid this difficulty, we convert this problem into a standard MDP of minimizing the pseudo variance. The variance difference formula is derived to directly compare the variances of Markov systems under any two policies. The variance difference formula

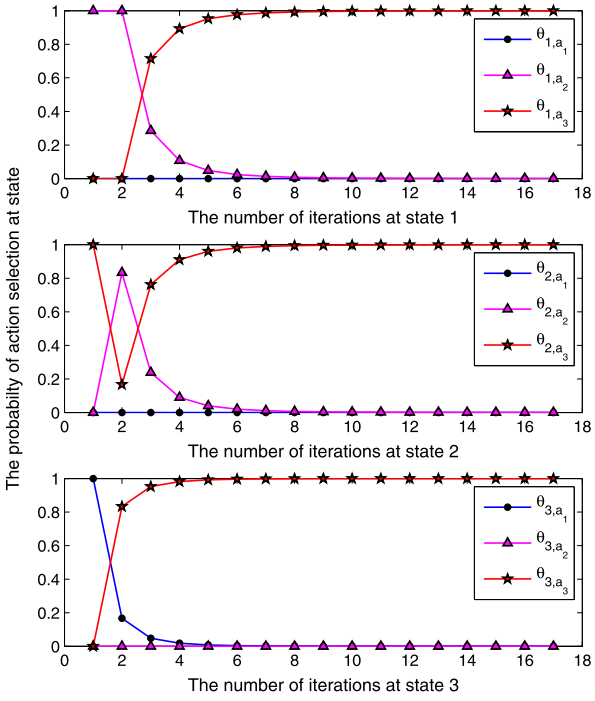

Fig. 2. The curves of $\theta_{i,a}$'s during the gradient-based optimization process.

plays a key role in this paper and it clearly describes the relation of the variance and the policy of Markov systems. Following this idea, some optimality properties are further obtained. We also develop the policy iteration type algorithm to optimize the variance of Markov systems. Compared with the policy gradient approaches widely used in the literature, our approach is demonstrated to be more efficient.

In future work, it is valuable to further study the policy iteration type approach to optimize the performance of MDPs, considering both the mean and the variance of system rewards. Our approach provides a promising way to study the mean–variance optimization problem. Moreover, the policy iteration type approach in this paper converges to a local optimum. How to identify the conditions with which our approach can converge to the global optimum is another interesting topic. The algorithmic issues about the parameter estimation and the online implementation of our algorithm also deserve further investigations.

Acknowledgments

The author would like to thank the three anonymous referees and the associate editor for their valuable comments which help improve the paper.

Appendix

Below, we give a detailed derivation of the relation formula (6) between two types of variance definitions, $\eta_{\sigma s}$ and $\eta_{\sigma l}$. Suppose the stochastic process is ergodic and it already reaches steady state. Then, we have

$$\begin{aligned}\eta_{\sigma l} &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1} r(X_t) - \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t)\right]\right\}^2\\ &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1} r(X_t) - T\eta\right\}^2\\ &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}\sum_{k=0}^{T-1} r(X_t)r(X_k) - 2T\eta\sum_{t=0}^{T-1} r(X_t) + T^2\eta^2\right\}\\ &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}\sum_{k=0}^{T-1} r(X_t)r(X_k) - T^2\eta^2\right\}\\ &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}\sum_{k=0}^{T-1}[r(X_t)r(X_k) - \eta^2]\right\}\\ &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}\sum_{k=0}^{T-1}[r(X_t) - \eta][r(X_k) - \eta]\right\}\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{T-1}\mathbb{COV}(r(X_t), r(X_k)).\end{aligned} \tag{A.1}$$

Therefore, the above equation shows that the limiting cumulative variance $\eta_{\sigma l}$ equals the average summation of covariances between rewards at one stage and all possible stages. $\eta_{\sigma l}$ quantifies the *correlations* of the system rewards at different stages. Furthermore, (A.1) can be rewritten as below.

$$\begin{aligned}\eta_{\sigma l} &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{COV}(r(X_t), r(X_t))\\ &\quad + \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0,k\neq t}^{T-1}\mathbb{COV}(r(X_t), r(X_k)).\end{aligned} \tag{A.2}$$

With the definition (2) of $\eta_{\sigma s}$, the relation formula (6) is directly obtained.

References

Cao, X. R. (2007). *Stochastic learning and optimization—a sensitivity-based approach*. New York: Springer.

Cao, X. R., & Chen, H. F. (1997). Perturbation realization, potentials, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, *42*, 1382–1393.

Cao, X. R., & Zhang, J. (2008). The nth-order bias optimality for multi-chain Markov decision processes. *IEEE Transactions on Automatic Control*, *53*, 496–508.

Chang, H. S., Fu, M. C., Hu, J., & Marcus, S. I. (2007). *Simulation-based algorithms for markov decision processes*. Springer.

Chung, K. J. (1994). Mean–variance tradeoffs in an undiscounted MDP: the unichain case. *Operations Research*, *42*, 184–188.

Feinberg, E., & Schwartz, A. (2002). *Handbook of markov decision processes: methods and applications*. Boston, MA: Kluwer Academic Publishers.

Guo, X., & Song, X. Y. (2009). Mean–variance criteria for finite continuous-time Markov decision processes. *IEEE Transactions on Automatic Control*, *54*, 2151–2157.

Guo, X., Ye, L., & Yin, G. (2012). A mean–variance optimization problem for discounted Markov decision processes. *European Journal of Operational Research*, *220*, 423–429.

Harrison, C. A., & Qin, S. J. (2009). Minimum variance performance map for constrained model predictive control. *Journal of Process Control*, *19*, 1199–1204.

Hernandez-Lerma, O., Vega-Amaya, O., & Carrasco, G. (1999). Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM Journal on Control and Optimization*, *38*, 79–93.

Huang, B. (2002). Minimum variance control and performance assessment of time-variant processes. *Journal of Process Control*, *12*, 707–719.

Huang, Y., & Chen, X. (2012). A sensitivity-based construction approach to sample-path variance minimization of Markov decision processes. In *Proceedings of the 2012 Australian control conference, AUCC'2012*, Sydney, Australia, (pp. 215–220).

Kuindersma, S. R., Grupen, R. A., & Barto, A. G. (2013). Variable risk control via stochastic optimization. *The International Journal of Robotics Research*, *32*, 806–825.

Mannor, S., & Tsitsiklis, J. N. (2013). Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, *231*, 645–653.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*, 77–91.

Prashantha, L.A., & Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems, NIPS'13*, (pp. 252–260).

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. New York: John Wiley & Sons.

Schweitzer, P. J. (1968). Perturbation theory and finite Markov chains. *Journal of Applied Probability*, *5*, 401–413.

Sobel, M. J. (1982). The variance of discounted Markov decision. *Journal of Applied Probability*, *19*, 794–802.

Sobel, M. J. (1994). Mean–variance tradeoffs in an undiscounted MDP. *Operations Research*, *42*, 175–183.
Tamar, A., Castro, D.D., & Mannor, S. (2012). Policy gradients with variance related risk criteria. In *Proceedings of the 29th international conference on machine learning, ICML*, Edinburgh, Scotland.
Ummels, B. C., Gibescu, M., Pelgrum, E., Kling, W. L., & Brand, A. J. (2007). Impacts of wind power on thermal generation unit commitment and dispatch. *IEEE Transactions on Energy Conversion*, *22*, 44–51.
Xia, L. (2014). An iterative approach to reduce the variance of stochastic dynamic systems. In *Proceedings of the 19th world congress of the international federation of automatic control, IFAC'14*, Cape Town, South Africa, August 24–29, 2014, (pp. 10487–10492).
Xia, L. (2016). Optimization of parametric policies of Markov decision processes under a variance criterion. In *Proceedings of the 13th international workshop on discrete event systems, WODES'16*, Xi'an, China, May 30-June 1, 2016.
Zhou, X. Y., & Yin, G. (2004). Markowitz's mean–variance portfolio selection with regime switching: A continuous-time model. *SIAM Journal on Control and Optimization*, *42*, 1466–1482.

Li Xia is an associate professor with the Center for Intelligent and Networked Systems (CFINS), Department of Automation, Tsinghua University, Beijing China. He received the Bachelor and the Ph.D. degree in Control Theory in 2002 and 2007 respectively, both from Tsinghua University. After graduation, he worked at IBM Research China as a research staff member (2007–2009) and at the King Abdullah University of Science and Technology (KAUST) Saudi Arabia as a postdoctoral research fellow (2009–2011). Then he returned to Tsinghua University as a faculty member in 2011. He was a visiting scholar in Stanford University, the Hong Kong University of Science and Technology, etc. He serves/served as program committee member and associate editor of a number of international conferences and journals. His research interests include the methodology research in stochastic optimization, queueing theory, Markov decision processes, and the application research in building energy, energy Internet, production systems, Internet of things, etc. He is a senior member of IEEE.