Discrete Event Dyn Syst (2018) 28:63–81
DOI 10.1007/s10626-017-0258-5

Variance minimization of parameterized Markov decision processes

Li Xia[1]

Received: 5 October 2016 / Accepted: 20 July 2017 / Published online: 26 August 2017

Abstract In this paper, we study the variance minimization problem of Markov decision processes (MDPs) in which the policy is parameterized by action selection probabilities or other general parameters. Different from the average or discounted criterion mostly used in the traditional MDP theory, the variance criterion is difficult to handle because of the non-Markovian property caused by the nonlinear (quadratic) structure of variance function. With the basic idea of sensitivity-based optimization, we derive a difference formula of the reward variance under any two parametric policies. A variance derivative formula is also obtained. With these sensitivity formulas, we obtain a necessary condition of the optimal policy with the minimal variance. We also prove that the optimal policy with the minimal variance can be found in the deterministic policy space. An iterative algorithm is further developed to efficiently reduce the reward variance and this algorithm can converge to the local optimal policy. Finally, we conduct some numerical experiments to demonstrate the main results of this paper.

Keywords Markov decision process · Parametric policy · Variance criterion · Sensitivity-based optimization · Discrete event system

This article belongs to the Topical Collection: *Special Issue on Performance Analysis and Optimization of Discrete Event Systems*
Guest Editors: Christos G. Cassandras and Alessandro Giua

✉ Li Xia
xial@tsinghua.edu.cn

[1] Center for Intelligent Networked Systems (CFINS), Department of Automation, TNList, Tsinghua University, Beijing 100084, China

1 Introduction

Markov decision processes (MDPs) are very important for the performance analysis and optimization of stochastic dynamic decision problems. The goal of MDPs is to find the optimal policy such that the expectation of system performance can be maximized, where the Bellman optimality equation plays a key role in developing the theory of MDPs.

In the literature, most of the studies on the MDP theory focus on the average or discounted criteria (Bertsekas 2012; Feinberg and Schwartz 2002; Guo and Hernandez-Lerma 2009; Puterman 1994). Much less attention has been paid to the variance criterion. The variance criterion is also an important performance metric in many practical problems. For example, in financial engineering, the variance criterion usually reflects the risk related factors. Portfolio management is a very important topic in financial engineering and it aims to reduce the variance of asset returns, thus to control the risk of assets. A key formulation for this problem is called the *mean-variance optimization*, which is proposed by H. Markowitz, the 1990 Nobel Laureate in Economics (Markowitz 1952). In the mean-variance optimization, there are two objectives that are considered together, one is the mean of rewards, and the other is the variance of rewards. The goal is to find an optimal policy such that the mean performance is maximized while the variance is lower than a given value, or the variance is minimized while the mean performance is larger than a given value. The *Pareto optimal* solutions to the mean-variance optimization compose a curve called *efficient frontier*, which gives an intuitive guide to balance the return and risk of assets from the economic viewpoint.

There are many studies on the mean-variance optimization. One of the main threads is the policy gradient approach, which is widely used by the researchers from the community of computer science while its root can be originated from the idea of perturbation analysis in Markov systems (Cao 2007; Cao and Chen 1997). The key idea of policy gradient is to derive a formula for the performance derivative with respect to (w.r.t.) the policy or system parameters (Marbach and Tsitsiklis 2001). Then the value of derivatives or gradients can be numerically computed or estimated from the system sample path (Mannor and Tsitsiklis 2011; Tamar et al. 2012). Finally, a gradient descent algorithm or stochastic approximation algorithm can follow to approach to the local optimal solution in the policy space or the parameter space. The gradient-based approach is easy to adopt in practice. However, it suffers from some intrinsic deficiencies, such as the trap of local optimum, the difficulty of selecting proper step-size, and the sensitivity to the initial point. There are also other works studying this problem from other perspectives. For example, some works study this problem by formulating it as a mathematical programming problem, where the techniques of linear and quadratic programming are used to study the problem structure (Chung 1994; Sobel 1994). Another main thread to study this problem is based on the traditional theory of MDPs. Although the variance criterion is not Markovian, we can convert the variance minimization problem into an equivalent MDP with a new performance function, at the condition that the average or discounted performance metric of the system is already maximized (Guo et al. 2012; Hernandez-Lerma et al. 1999). For other general cases, such as unbounded transition rates and state-dependent discount factors, there also exist many works to study the mean-variance optimization from the framework of MDPs (Guo et al. 2015; Huo et al. 2017).

In this paper, we study the optimization of MDPs under the variance criterion, where the policy is parameterized by some system parameters. Our goal is to find the optimal parameters such that the variance of system rewards can be minimized. Different from the mean-variance optimization introduced above, the average or discounted performance metric is not considered in our problem. The variance minimization problem has practical meanings in engineering systems. For example, for a wind farm with energy storage system illustrated in Fig. 1, we aim to schedule the power output of the whole system such that the power variation can be reduced. The power stability is very important to keep the safety of electricity grid (Ummels et al. 2007). In this problem, the reduction of power variation is more important than the improvement of utilization ratio of wind power. The stochastic process of wind power can be modeled by a Markov chain (Luh et al. 2014). The scheduling algorithm has to determine a series of values of output power to the grid at different energy storage level or wind power level. This series of values can be viewed a parametric policy and this decision problem can be modeled as a parameterized MDP. If we use the variance criterion to quantify the power variation, we can formulate this problem as a variance minimization problem of parameterized MDPs.

There exist some difficulties in this problem. The main difficulty is caused by the *nonlinear* property of the variance function. In a standard MDP model, we require that the cost function and the state transition probability should be Markovian. That is, the cost at the current stage should not be affected by the actions in future stages (see page 20 in Puterman's book (Puterman 1994)). However, in our problem, since the variance function is quadratic and it is also related to the mean performance, the associated cost function of MDPs under the variance criterion is dependent on the action selection in future stages (Xia 2016a, b). Thus, the variance function is not *additive* and it does not have Markovian property. The traditional approaches in MDP theory cannot be applied to our problem. Although the gradient-based approach is valid for this problem (Mannor and Tsitsiklis 2011; Tamar et al. 2012), it suffers from the intrinsic deficiencies as we discussed above. The other difficulty comes from the parametric policy that is parameterized by some parameters (Xia and Jia 2015). In a standard MDP model, the policy is a mapping from the state space to the action space. However, in a parameterized MDP, the policy is controlled by one or multiple parameters. We may not freely adjust the parameters at every state. For example, in an M/M/1 queue, we control the value of service rate μ to maximize the average performance of the system. The service rate μ has the same value at different system state n (queue length). This service rate control problem is a parameterized MDP. The correlation of the policy at different states makes the traditional approaches in MDP theory inapplicable to this problem. In summary, our problem is not a standard MDP and it suffers from the difficulties caused by the variance function and the parametric policy. The Bellman optimality equation does not hold for this problem. We have to resort to other approaches.

In this paper, we use the sensitivity-based optimization theory of Markov systems to study this variance minimization problem in parameterized MDPs. We discuss two types of parametric policies. The first one is to control the selection probability of every action at every state. The second one is a set of general parameters that have effects on the transition probabilities and reward functions. The first type of parametric policies is easy to handle

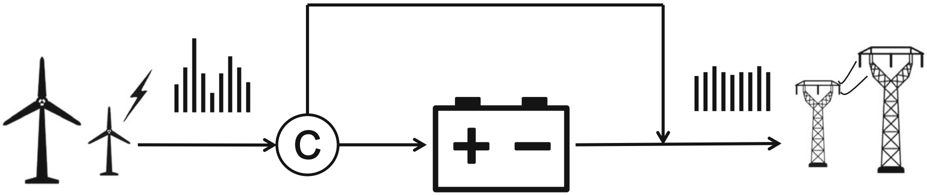

Fig. 1 Power variation reduction for wind farms and energy storage systems

since we can freely change the value of parameters at every state (no correlation among different states). The second one is difficult and we give some discussions under proper conditions. Our goal is to find the optimal values of parameters to minimize the reward variance of the Markov system. The key idea of the sensitivity-based optimization theory is the difference formula that quantifies the performance difference of Markov systems under any two different policies or parameters (Cao 2007; Cao and Chen 1997). This theory does not depend on the Bellman optimality equation and it may remain valid for a general controlled Markov system, even the problem does not fit a standard MDP model. For the parametric policy of action selecting probability, we derive a variance difference formula under any two different policies, where a nonnegative term plays an important role to alleviate the difficulties mentioned above. A derivative formula of the reward variance w.r.t. the parameter is also obtained. With these sensitivity formulas, we derive a necessary condition of the optimal policy. We also prove that the optimal policy with the minimal variance can be found in the deterministic policy space. We further develop an iterative algorithm to strictly reduce the variance of Markov systems. For the general parametric policy, we also derive the similar results as above. Compared with our previous work (Xia 2016b), this paper mainly studies the parametric policy and the reward function can be varied under different policies, which makes our results more general for parameterized MDPs.

The rest of the paper is organized as follows. Section 2 gives a mathematical formulation for the variance minimization problem of parameterized MDPs. In Section 3, we apply the sensitivity-based optimization theory to study this problem in which the parameters are the action selection probabilities. The main results of this paper are derived in this section. In Section 4, we further extend our study to a case where the parameters can be general ones. In Section 5, we conduct numerical experiments to demonstrate the main results. Finally, we conclude this paper in Section 6.

2 Problem formulation

Consider a discrete time Markov chain $\boldsymbol{X} := \{X_0, X_1, \cdots, X_t, \cdots\}$, where X_t is the system state at time t, $t = 0, 1, \cdots$. The state space is $\boldsymbol{\mathcal{S}} := \{1, 2, \cdots, S\}$ and its size is S. When the system is at state i, we can select an action a from the action space $\boldsymbol{\mathcal{A}}(i)$, where $i \in \boldsymbol{\mathcal{S}}$. For simplicity, we assume $\boldsymbol{\mathcal{A}}(i) = \boldsymbol{\mathcal{A}}$ for all $i \in \boldsymbol{\mathcal{S}}$. The main results in this paper remain valid when $\boldsymbol{\mathcal{A}}(i)$'s are different. The action space is finite and we define it as $\mathcal{A} := \{a_1, a_2, \cdots, a_A\}$, where A is the size of $\boldsymbol{\mathcal{A}}$. When an action a is adopted at state i, the system will receive a reward denoted as $r(i, a)$ and the system state will transit to the next state j with a transition probability $p(j|i, a)$, where $i, j \in \boldsymbol{\mathcal{S}}$, $a \in \boldsymbol{\mathcal{A}}$. Obviously, we have $p(j|i, a) \geq 0$ and $\sum_j p(j|i, a) = 1$. We assume that the Markov chain is ergodic and the long-run average performance of the Markov chain is defined as below.

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

where $\mathbf{A}_t$ is the action adopted at time t. The steady state distribution $\boldsymbol{\pi}$ is denoted as an S-dimensional row vector as follows.

$$\boldsymbol{\pi} := (\pi(1), \boldsymbol{\pi}(2), \cdots, \pi(S)). \tag{2}$$

The reward function $\mathbf{r}$ is denoted as an S-by-A matrix defined as below.

$$\mathbf{r} := \begin{pmatrix} r(1,a_1), & r(1,a_2), & \cdots, & r(1,a_A) \\ r(2,a_1), & r(2,a_2), & \cdots, & r(2,a_A) \\ \vdots & \vdots & \ddots & \vdots \\ r(S,a_1), & r(S,a_2), & \cdots, & r(S,a_A) \end{pmatrix}. \tag{3}$$

The steady state variance of the Markov chain is defined as below.

$$\eta_\sigma := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}[r(X_t, A_t) - \eta]^2\right\} \tag{4}$$

According to the terminology of MDPs, a policy of MDP is a sequence of action selection rules that are mapping functions from the state space (or more generally, the historical trajectory of states and actions) to the action space. However, the policy of many practical decision problems is controlled by system parameters, which is easy to adopt by practitioners. In this paper, we limit our discussion on such parametric policies and we call such decision problems *parameterized* MDPs.

There are different types of parametric policies in practice. First, we study a special case in which the controlled parameters are action selection probabilities $\boldsymbol{\theta}_{i,a}$, $i \in \boldsymbol{\mathcal{S}}, a \in \boldsymbol{\mathcal{A}}$. That is, we choose the action a at state i with probability $\theta_{i,a}$ that satisfies $\theta_{i,a} \geq 0$ and $\sum_a \theta_{i,a} = 1$ for all i. The policy is further characterized by an S-by-A matrix $\boldsymbol{\theta}$ that is defined as below.

$$\boldsymbol{\theta} := \begin{pmatrix} \theta_{1,a_1}, & \theta_{1,a_2}, & \cdots, & \theta_{1,a_A} \\ \theta_{2,a_1}, & \theta_{2,a_2}, & \cdots, & \theta_{2,a_A} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{S,a_1}, & \theta_{S,a_2}, & \cdots, & \theta_{S,a_A} \end{pmatrix}. \tag{5}$$

Therefore, different $\boldsymbol{\theta}$ represents different policy and we use the superscript $\boldsymbol{\theta}$ to denotes the corresponding notations of the Markov chain with policy $\boldsymbol{\theta}$, such as $\boldsymbol{\pi}^{\boldsymbol{\theta}}$, $\eta^{\boldsymbol{\theta}}$, $\eta_\sigma^{\boldsymbol{\theta}}$, etc. Under policy $\boldsymbol{\theta}$, the state transition probability can be written as

$$p^{\boldsymbol{\theta}}(i,j) := \sum_{a\in\boldsymbol{\mathcal{A}}} \theta_{i,a} p(j|i,a). \tag{6}$$

The transition probability matrix of the Markov chain under policy $\boldsymbol{\theta}$ is defined as below.

$$P^{\theta} := \begin{pmatrix} p^{\boldsymbol{\theta}}(1,1), & p^{\boldsymbol{\theta}}(1,2), & \cdots, & p^{\boldsymbol{\theta}}(1,S) \\ p^{\boldsymbol{\theta}}(2,1), & p^{\boldsymbol{\theta}}(2,2), & \cdots, & p^{\boldsymbol{\theta}}(2,S) \\ \vdots & \vdots & \ddots & \vdots \\ p^{\boldsymbol{\theta}}(S,1), & p^{\boldsymbol{\theta}}(S,2), & \cdots, & p^{\boldsymbol{\theta}}(S,S) \end{pmatrix}. \tag{7}$$

The value domain of $\boldsymbol{\theta}$ is a high dimensional real number space $\mathbb{R}^{S\times A}$, with the constraints $\boldsymbol{\theta} \geq 0$ and $\boldsymbol{\theta}1 = 1$, where 1 is a proper dimension column vector with all elements as 1. We denote the valid value domain of $\boldsymbol{\theta}$ as Θ that is an S-by-A dimensional polyhedron in the real number space. That is, we define

$$\Theta := \{\text{all } \boldsymbol{\theta} : \boldsymbol{\theta} \geq 0, \boldsymbol{\theta}1 = 1\}. \tag{8}$$

It is easy to verify that Θ is a *convex set*. Our goal is to find the optimal parameter $\boldsymbol{\theta}^*$ from the solution space Θ to minimize the reward variance of the Markov chain. That is, the variance minimization problem for such parametric policies is formulated as below.

$$\begin{aligned}\boldsymbol{\theta}^* &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\eta_{\sigma}^{\boldsymbol{\theta}}\right\}\\ &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\lim_{T\to\infty}\frac{1}{T}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{T-1}\left[r(X_t,A_t)-\eta^{\boldsymbol{\theta}}\right]^2\right]\right\},\end{aligned} \tag{9}$$

where $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ indicates the mathematical expectation of the Markov chain under the policy $\boldsymbol{\theta}$.

3 Main results

As we discussed above, the optimization problem described by Eq. 9 does not fit the standard model of MDPs since the variance function is quadratic. Therefore, we use the sensitivity-based optimization theory that is valid for the performance optimization of any Markov systems (Cao 2007; Cao and Chen 1997). According to the terminology of MDPs, we define the cost function of Eq. 9 under the variance criterion as below.

$$\begin{aligned}f_{\sigma}(i) &:= \sum_{a\in\boldsymbol{\mathcal{A}}}\boldsymbol{\theta}_{i,a}(r(i,a)-\eta)^2\\ &= \sum_{a\in\boldsymbol{\mathcal{A}}}\boldsymbol{\theta}_{i,a}r^2(i,a)-2\eta\sum_{a\in\boldsymbol{\mathcal{A}}}\boldsymbol{\theta}_{i,a}r(i,a)+\eta^2.\end{aligned} \tag{10}$$

For simplicity, we further define the following notations

$$\begin{aligned}\tilde{r}_2(i) &:= \sum_{a\in\boldsymbol{\mathcal{A}}}\boldsymbol{\theta}_{i,a}r^2(i,a),\\ \bar{r}(i) &:= \sum_{a\in\boldsymbol{\mathcal{A}}}\boldsymbol{\theta}_{i,a}r(i,a).\end{aligned} \tag{11}$$

In an S-dimensional column vector form, we can rewrite the above definitions as below.

$$\begin{aligned}\tilde{r}_2 &:= (\boldsymbol{\theta}\odot r_{\odot}^2)1,\\ \bar{r} &:= (\boldsymbol{\theta}\odot r)1,\end{aligned} \tag{12}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{r}$ are S-by-A matrices defined in Eqs., 5 and 3 respectively, $\mathbf{1}$ is an A-dimensional column vector with element 1, $\odot$ denotes the *Hadamard product* (componentwisely) of two vectors or matrices, i.e., for any vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with the same dimension, we define

$$\begin{aligned}\boldsymbol{a}\odot\boldsymbol{b} &:= (a_1b_1,a_2b_2,\cdots),\\ \boldsymbol{a}_{\odot}^2 &:= \boldsymbol{a}\odot\boldsymbol{a} := (a_1^2,a_2^2,\cdots).\end{aligned} \tag{13}$$

Therefore, we have the variance function as below.

$$\begin{aligned}f_{\sigma}(i) &= \tilde{r}_2(i)-2\eta\bar{r}(i)+\eta^2,\\ \boldsymbol{f}_{\sigma} &= \tilde{\boldsymbol{r}}_2-2\eta\bar{\boldsymbol{r}}+\eta^2\mathbf{1},\end{aligned} \tag{14}$$

where $\boldsymbol{f}_{\sigma}$ is an S-dimensional column vector whose element is $f_{\sigma}(i)$, $i\in\boldsymbol{\mathcal{S}}$. Obviously, we have

$$\begin{aligned}\eta_{\sigma} &= \boldsymbol{\pi}\boldsymbol{f}_{\sigma},\\ \eta &= \boldsymbol{\pi}\bar{\boldsymbol{r}}.\end{aligned} \tag{15}$$

The state transition probability of this Markov chain under parameter $\boldsymbol{\theta}$ is written as $p^{\boldsymbol{\theta}}(i,j)$ in Eq. 6. The transition probability matrix is written as $\boldsymbol{P}^{\boldsymbol{\theta}}$ in Eq. 7. Note that in some places of this paper, we omit the superscript $\boldsymbol{\theta}$ by default and denote it as P for simplicity.

The *performance potential* is a fundamental quantity defined in the sensitivity-based optimization theory. It quantifies the contribution of an initial state to the average performance of Markov systems (Cao 2007). For the variance minimization problem (9), the optimization performance is the system reward variance. Similar to the concept of performance potential, we define a quantity called *variance potential* as below.

$$g_\sigma(i) := E\left\{\sum_{t=0}^{\infty}[f_\sigma(X_t) - \eta_\sigma]\,|X_0 = i\right\}, \quad i \in \boldsymbol{\mathcal{S}}. \tag{16}$$

The above definition can be further rewritten as below.

$$g_\sigma(i) := E\left\{\sum_{t=0}^{\infty}[(r(X_t, A_t) - \eta)^2 - \eta_\sigma]\,|X_0 = i\right\}, \quad i \in \boldsymbol{\mathcal{S}}. \tag{17}$$

By extending the first summation at $t = 0$ in Eq. 16, we can further rewrite it in a matrix form as below.

$$\boldsymbol{g}_\sigma = \boldsymbol{f}_\sigma - \eta_\sigma \mathbf{1} + \boldsymbol{P}\boldsymbol{g}_\sigma \tag{18}$$

We can numerically solve the above equation to compute the value of $\boldsymbol{g}_\sigma$. We can also estimate the value of $\boldsymbol{g}_\sigma$ based on the definition (17) or other variations from a single sample path of the Markov chain. The basic idea of estimation or computation of $\boldsymbol{g}_\sigma$ is similar to the discussion of performance potentials and the details can be referred to chapter 3 of Cao's book (Cao 2007).

Suppose the parametric policy is changed from $\boldsymbol{\theta}$ to $\boldsymbol{\theta'}$. The corresponding transition probability matrix and the variance function are changed from $\boldsymbol{P}, \boldsymbol{f}_\sigma$ to $\boldsymbol{P'}, \boldsymbol{f'}_\sigma$, respectively. That is, the state transition probability under the new parameter $\boldsymbol{\theta'}$ is

$$\boldsymbol{P'}(i,j) = \sum_{a \in \boldsymbol{\mathcal{A}}} \theta'_{i,a} p(j|i,a), \quad i, j \in \boldsymbol{\mathcal{S}}. \tag{19}$$

The cost function of the Markov system under the variance criterion with the new parameter $\boldsymbol{\theta'}$ is

$$\boldsymbol{f'}_\sigma = \tilde{\boldsymbol{r}}'_2 - 2\eta'\bar{\boldsymbol{r}}' + \eta'^2\mathbf{1}, \tag{20}$$

where

$$\begin{aligned} \tilde{\boldsymbol{r}}'_2 &= (\boldsymbol{\theta'} \odot r^2_{\odot})\mathbf{1}, \\ \bar{\boldsymbol{r}}' &= (\boldsymbol{\theta'} \odot r)\mathbf{1}, \end{aligned} \tag{21}$$

and η' is the long-run average performance under the new parameter $\boldsymbol{\theta'}$. Obviously, we have

$$\begin{aligned} \eta'_\sigma &= \boldsymbol{\pi}'\boldsymbol{f'}_\sigma, \\ \eta' &= \boldsymbol{\pi}'\bar{\boldsymbol{r}}', \end{aligned} \tag{22}$$

where $\boldsymbol{\pi'}$ is the steady state distribution of the Markov chain under the new parameter $\boldsymbol{\theta'}$.

Right-multiplying $\boldsymbol{\pi'}$ on both sides of Eq. 18 and utilizing Eq. 22 and $\boldsymbol{\pi'}\boldsymbol{P'} = \boldsymbol{\pi'}$, we can derive the difference formula of the variance of Markov systems under these two sets of parameters as follows.

$$\eta'_\sigma - \eta_\sigma = \boldsymbol{\pi'}\left[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{f'}_\sigma - \boldsymbol{f}_\sigma)\right]. \tag{23}$$

The above formula can also be viewed as a direct result by applying the difference formula of the sensitivity-based optimization theory to our problem formulated in Eq. 9. To apply the above difference formula, we have to know the values of $\boldsymbol{P'}$ and $\boldsymbol{f'}_\sigma$ under any new parameter $\boldsymbol{\theta'}$. The value of $\boldsymbol{P'}$ can be directly obtained with Eq. 19. However, it is difficult to directly compute the value of $\boldsymbol{f'}_\sigma$ with Eq. 20, because the value of η' in Eq. 20 is unknown. If we compute the value of η' under every possible $\boldsymbol{\theta'}$, the computational complexity is exhaustive and it is equivalent to a brute-force enumeration for the original optimization problem (9).

Remark 1 Since the value of $\boldsymbol{f'}_\sigma$ is unknown, the associated optimization problem (9) is not a standard MDP. We cannot directly use the difference formula (23) or traditional MDP approaches to solve this problem.

Fortunately, we find new results that can avoid the above difficulty. With Eqs. 14 and 20, we have

$$\begin{aligned}
\boldsymbol{\pi'}(\boldsymbol{f'}_\sigma - \boldsymbol{f}_\sigma) &= \boldsymbol{\pi'}\left[\tilde{\boldsymbol{r}}'_2 - 2\eta'\bar{\boldsymbol{r}}' + \eta'^2\mathbf{1} - \tilde{\boldsymbol{r}}_2 + 2\eta\bar{\boldsymbol{r}} - \eta^2\mathbf{1}\right] \\
&= \boldsymbol{\pi'}\tilde{\boldsymbol{r}}'_2 - 2\eta'\eta' + \eta'^2 - \boldsymbol{\pi'}\tilde{r}_2 + \boldsymbol{\pi'}2\eta\bar{\boldsymbol{r}} - \eta^2 \\
&= \boldsymbol{\pi'}\tilde{r}'_2 - 2\eta\eta' - \boldsymbol{\pi'}\tilde{\boldsymbol{r}}_2 + \boldsymbol{\pi'}2\eta\bar{\boldsymbol{r}} - \eta^2 - \eta'^2 + 2\eta\eta' \\
&= \boldsymbol{\pi'}\left[\tilde{\boldsymbol{r}}'_2 - 2\eta\bar{\boldsymbol{r}}' - \tilde{\boldsymbol{r}}'_2 + 2\eta\bar{\boldsymbol{r}}\right] - (\eta' - \eta)^2, \qquad (24)
\end{aligned}$$

where we utilize the equality $\eta' = \boldsymbol{\pi'}\bar{\boldsymbol{r}}'$ and $\boldsymbol{\pi'}\mathbf{1} = 1$.

Substituting Eq. 24 into Eq. 23, we derive the following *variance difference formula* for the Markov system under any two different parametric policies $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$

$$\eta'_\sigma - \eta_\sigma = \boldsymbol{\pi'}\left[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + \tilde{\boldsymbol{r}}'_2 - 2\eta\bar{\boldsymbol{r}}' - \tilde{\boldsymbol{r}}_2 + 2\eta\bar{\boldsymbol{r}}\right] - (\eta' - \eta)^2. \qquad (25)$$

The above difference formula is of a general matrix form. We can further obtain a more specific form with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$. Substituting Eqs. 6, 12, 19, and 21 into Eq. 25, we have

$$\begin{aligned}
\eta'_\sigma - \eta_\sigma = &\sum_{i\in\mathcal{S}}\pi'(i)\left[\sum_{j\in\mathcal{S}}(p^{\boldsymbol{\theta'}}(i,j) - p^{\boldsymbol{\theta}}(i,j))g_\sigma(j) + \tilde{r}'_2(i) - 2\eta\bar{r}'(i) - \tilde{r}_2(i) + 2\eta\bar{r}(i)\right] \\
&-(\eta' - \eta)^2 \\
= &\sum_{i\in\mathcal{S}}\boldsymbol{\pi'}(i)\left[\sum_{j\in\mathcal{S}}\left(\sum_{a\in\mathcal{A}}\theta'_{i,a}p(j|i,a) - \sum_{a\in\mathcal{A}}\theta_{i,a}p(j|i,a)\right)g_\sigma(j)\right. \\
&\left.+\sum_{a\in\mathcal{A}}\theta'_{i,a}r^2(i,a) - 2\eta\sum_{a\in\mathcal{A}}\theta'_{i,a}r(i,a) - \sum_{a\in\mathcal{A}}\theta_{i,a}r^2(i,a) + 2\eta\sum_{a\in\mathcal{A}}\theta_{i,a}r(i,a)\right] \\
&-(\eta' - \eta)^2 \\
= &\sum_{i\in\mathcal{S}}\boldsymbol{\pi'}(i)\sum_{a\in\mathcal{A}}\left(\theta'_{i,a} - \theta_{i,a}\right)\left[\sum_{j\in\mathcal{S}}p(j|i,a)g_\sigma(j) + r^2(i,a) - 2\eta r(i,a)\right] - (\eta' - \eta)^2.
\end{aligned}$$

(26)

We further define the quantity in the square bracket of the above equation as $G(i,a)$, i.e.,

$$G(i,a) := \sum_{j\in\mathcal{S}}p(j|i,a)g_\sigma(j) + r^2(i,a) - 2\eta r(i,a). \qquad (27)$$

Note that $p(j|i,a)$ and $r(i,a)$ are given parameters, $g_\sigma(j)$ and η can be computed or estimated based the system sample path with the current policy $\boldsymbol{\theta}$. Therefore, the value of $G(i,a)$ can be computed or estimated from the sample path of the current system with $\boldsymbol{\theta}$. Actually, it should be denoted as $G^{\boldsymbol{\theta}}(i,a)$ and we omit the superscript $\boldsymbol{\theta}$ in most of situations for simplicity. Substituting Eq. 27 into Eq. 26, we obtain a specific form of variance difference formula for the Markov system under $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$ as below.

$$\eta'_\sigma - \eta_\sigma = \sum_{i\in\boldsymbol{\mathcal{S}}} \boldsymbol{\pi}'(i) \sum_{a\in\boldsymbol{\mathcal{A}}} (\theta'_{i,a} - \theta_{i,a}) G(i,a) - (\eta' - \eta)^2. \tag{28}$$

The difference formulas (25) and (28) are fundamental for the analysis of the variance minimization problem (9). We can derive useful insights and construct optimization algorithms to solve this problem. With Eq. 28 as an instance, we see that $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$ are given parameters and $G(i,a)$ is computable or estimatable based on the system sample path. Although the value of η' is unknown and computationally cumbersome under every possible $\boldsymbol{\theta'}$, the term $(\eta' - \eta)^2$ is always *nonnegative*. Since the element of $\boldsymbol{\pi'}$ is always positive, we only have to choose a proper $\boldsymbol{\theta'}$ that makes the value of $\sum_{a\in\boldsymbol{\mathcal{A}}}(\theta'_{i,a} - \theta_{i,a})G(i,a)$ negative. We will directly have $\eta'_\sigma - \eta_\sigma < 0 - (\eta' - \eta)^2 \le 0$ and the reward variance of the Markov system under the new parametric policy θ' is reduced. This is the basic idea to perform an iterative algorithm to reduce the variance of Markov systems. We will give more detailed discussions in the next section.

Remark 2 The key advantage of variance difference formulas (25) and (28) is the nonnegative term $(\eta' - \eta)^2$, which avoids the enumerative computation of η' under every possible policy.

The quadratic term $(\eta' - \eta)^2$ in Eqs. 25 and 28 is important since it is always nonnegative despite of the exact value of η'. With Eqs. 25 and 28, we can make further studies for the optimization problem (9) and derive some results that are difficult to obtain with the traditional approach in the literature. One of the direct results is the following *necessary condition* of the optimal policy of this problem.

Theorem 1 *If $\boldsymbol{\theta^*}$ is the optimal parametric policy of optimization problem* (9)*, then it has to satisfy the following necessary condition*

$$\boldsymbol{P'g^*}_\sigma + \tilde{\boldsymbol{r}}'_2 - 2\eta^* \bar{\boldsymbol{r}}' \succeq \boldsymbol{P^*g^*}_\sigma + \tilde{\boldsymbol{r}^*}_2 - 2\eta^* \bar{\boldsymbol{r}}^*, \qquad \forall\, \boldsymbol{\theta'} \in \Theta, \tag{29}$$

where $\boldsymbol{g^*}_\sigma$, η^*, $\boldsymbol{P^*}$, $\tilde{\boldsymbol{r}}^*_2$, and $\bar{\boldsymbol{r}}^*$ are the corresponding quantities of Markov system under the policy $\boldsymbol{\theta^*}$, $\succeq$ means $\ge$ componentwisely for a vector.

Proof We give the proof by contradiction. Suppose the inequality (29) does not hold, i.e., for some state, say state i, there exists parameter $\theta'_{i,a}$, $a \in \boldsymbol{\mathcal{A}}$, which makes

$$P'(i,:)\boldsymbol{g^*}_\sigma + \sum_{a\in\boldsymbol{\mathcal{A}}} \theta'_{i,a} r^2(i,a) - 2\eta^* \sum_{a\in\boldsymbol{\mathcal{A}}} \theta'_{i,a} r(i,a)$$
$$< \boldsymbol{P^*}(i,:)\boldsymbol{g^*}_\sigma + \sum_{a\in\boldsymbol{\mathcal{A}}} \theta^*_{i,a} r^2(i,a) - 2\eta^* \sum_{a\in\boldsymbol{\mathcal{A}}} \theta^*_{i,a} r(i,a). \tag{30}$$

Therefore, we construct a new policy denoted as $\boldsymbol{\theta'}$ that selects the parameter value as $\theta^*_{i,a}$ for states $j \neq i$ and selects $\theta'_{i,a}$ for state i, $a \in \boldsymbol{\mathcal{A}}$. According to the variance difference formula (25), the variance difference of Markov systems under the policy $\boldsymbol{\theta'}$ and $\boldsymbol{\theta^*}$ can be written as below.

$$\begin{aligned}\eta'_\sigma - \eta^*_\sigma &= \boldsymbol{\pi'}\left[(\boldsymbol{P'} - \boldsymbol{P^*})\boldsymbol{g^*}_\sigma + \tilde{\boldsymbol{r}}'_2 - 2\eta^*\bar{\boldsymbol{r}}' - \tilde{\boldsymbol{r}}^*_2 + 2\eta^*\bar{\boldsymbol{r}}^*\right] - (\eta' - \eta^*)^2 \\ &= \boldsymbol{\pi'}(i)\Bigg[(\boldsymbol{P'}(i,:) - \boldsymbol{P^*}(i,:))\boldsymbol{g^*}_\sigma + \sum_{a\in\boldsymbol{\mathcal{A}}}\theta'_{i,a}r^2(i,a) \\ &\quad -2\eta^*\sum_{a\in\boldsymbol{\mathcal{A}}}\theta'_{i,a}r(i,a) - \sum_{a\in\boldsymbol{\mathcal{A}}}\theta^*_{i,a}r^2(i,a) + 2\eta^*\sum_{a\in\boldsymbol{\mathcal{A}}}\theta^*_{i,a}r(i,a)\Bigg] - (\eta' - \eta^*)^2,\end{aligned} \tag{31}$$

where the second equality holds since the parameter value $\theta_{j,a}$ of policy $\boldsymbol{\theta^*}$ is the same as that of policy $\boldsymbol{\theta'}$, $j \neq i$, and the corresponding elements are eliminated. Since $\boldsymbol{\pi'}(i)$ is always positive, we substitute Eq. 30 into the above equation and have

$$\eta'_\sigma - \eta^*_\sigma < -(\eta' - \eta^*)^2 \leq 0. \tag{32}$$

Therefore, $\eta'_\sigma < \eta^*_\sigma$ and it contradicts the assumption that $\boldsymbol{\theta^*}$ is the optimal policy. The theorem is proved. □

Remark 3 With Eq. 28, the necessary condition (29) can be specifically rewritten as below.

$$\sum_{a\in\boldsymbol{\mathcal{A}}}\theta'_{i,a}G^*(i,a) \geq \sum_{a\in\boldsymbol{\mathcal{A}}}\theta^*_{i,a}G^*(i,a), \qquad \forall\, \boldsymbol{\theta'} \in \Theta,\ i \in \boldsymbol{\mathcal{S}}, \tag{33}$$

where $G^*(i,a)$ is the definition (27) under the optimal parameter $\boldsymbol{\theta^*}$, i.e., $G^*(i,a) = \sum_{j\in\boldsymbol{\mathcal{S}}} p(j|i,a)g^*_\sigma(j) + r^2(i,a) - 2\eta^* r(i,a)$.

Compared with Eq. 29, condition (33) is simpler and easy to validate in practice. The variance difference formulas (25) and (28) are the most general cases of sensitivity formula for the variance minimization problem in MDPs. Some other analogous results for this problem can be viewed as a special case of Eq. 25. Below, we discuss three different cases to introduce the analogous form of this sensitivity formula and its variations.

Case 1, deterministic policy: We consider two deterministic policies $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{L}'}$, the corresponding reward functions are denoted as $\boldsymbol{r}$ and $\boldsymbol{r'}$, respectively. With a little abuse of notations, $\boldsymbol{r}$ and $\boldsymbol{r'}$ are S-dimensional column vectors in this situation. It is easy to verify that $\tilde{\boldsymbol{r}}_2 = \boldsymbol{r}^2_\odot$, $\bar{\boldsymbol{r}} = \boldsymbol{r}$, $\tilde{\boldsymbol{r}}'_2 = \boldsymbol{r'^2}_\odot$, $\bar{\boldsymbol{r}}' = \boldsymbol{r'}$. Substituting them to Eq. 25, we obtain

$$\begin{aligned}\eta'_\sigma - \eta_\sigma &= \boldsymbol{\pi'}\left[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + \boldsymbol{r'^2}_\odot - 2\eta\boldsymbol{r'} - \boldsymbol{r}^2_\odot + 2\eta\boldsymbol{r}\right] - (\eta' - \eta)^2 \\ &= \boldsymbol{\pi'}\left[(\boldsymbol{P'} - \boldsymbol{P})g_\sigma + (\boldsymbol{r'} - \eta\boldsymbol{1})^2_\odot - (\boldsymbol{r} - \eta\boldsymbol{1})^2_\odot\right] - (\eta' - \eta)^2.\end{aligned} \tag{34}$$

This formula is exactly the same as the variance difference formula for deterministic policies in MDPs, which can be referred to Eq. 32 in our previous study (Xia 2016b).

Case 2, randomized policy: We consider a randomized policy $\boldsymbol{\mathcal{L}}^\delta_{\boldsymbol{\mathcal{L}'}}$ that adopts deterministic policy $\boldsymbol{\mathcal{L}'}$ with probability δ and adopts deterministic policy $\boldsymbol{\mathcal{L}}$ with probability $1 - \delta$, where $0 \leq \delta \leq 1$. Such a policy is also called a *mixed policy*. With Eq. 12, we can verify that $\tilde{\boldsymbol{r}}_2 = \boldsymbol{r}^2_{\odot}$, $\bar{\boldsymbol{r}} = \boldsymbol{r}$, $\tilde{\boldsymbol{r}}^{\delta}_2 = \delta \boldsymbol{r'^2}_{\odot} + (1-\delta)\boldsymbol{r}^2_{\odot}$, $\bar{\boldsymbol{r}}^{\delta} = \delta r' + (1-\delta)\boldsymbol{r}$, where $\boldsymbol{r}$ and $\boldsymbol{r'}$ are S-dimensional column vectors that are the same as those in Case 1. Substituting them to Eq. 25, we obtain

$$\begin{aligned}
\eta^{\delta}_{\sigma} - \eta_{\sigma} &= \boldsymbol{\pi}'\Big[(\boldsymbol{P}^{\delta} - \boldsymbol{P})\boldsymbol{g}_{\sigma} + \tilde{\boldsymbol{r}}^{\delta}_2 - 2\eta\bar{\boldsymbol{r}}^{\delta} - \tilde{\boldsymbol{r}}_2 + 2\eta\bar{\boldsymbol{r}}\Big] - (\eta^{\delta} - \eta)^2 \\
&= \boldsymbol{\pi}'\Big[\delta(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_{\sigma} + \delta\boldsymbol{r'^2}_{\odot} + (1-\delta)\boldsymbol{r^2}_{\odot} - 2\eta(\delta\boldsymbol{r'} + (1-\delta)\boldsymbol{r}) - \boldsymbol{r^2}_{\odot} + 2\eta\boldsymbol{r}\Big] - (\eta^{\delta} - \eta)^2 \\
&= \delta\boldsymbol{\pi}'\Big[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_{\sigma} + \boldsymbol{r'^2}_{\odot} - \boldsymbol{r^2}_{\odot} - 2\eta(\boldsymbol{r'} - \boldsymbol{r})\Big] - (\eta^{\delta} - \eta)^2 \\
&= \delta\boldsymbol{\pi}'\Big[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_{\sigma} + (\boldsymbol{r'} - \eta\boldsymbol{1})^2_{\odot} - (\boldsymbol{r} - \eta\boldsymbol{1})^2_{\odot}\Big] - (\eta^{\delta} - \eta)^2. \qquad (35)
\end{aligned}$$

Taking the derivative operation with respect to δ and letting δ go to 0, we obtain the following derivative formula

$$\frac{\mathrm{d}\eta_{\sigma}}{\mathrm{d}\delta} = \boldsymbol{\pi}\left[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_{\sigma} + (\boldsymbol{r'} - \eta\boldsymbol{1})^2_{\odot} - (\boldsymbol{r} - \eta\boldsymbol{1})^2_{\odot}\right]. \qquad (36)$$

Remark 4 Comparing Eqs. 36 and 34, we see that these two formulas are similar except that the term $-(\eta^{\delta} - \eta)^2$ disappears and $\boldsymbol{\pi'}$ is replaced by $\boldsymbol{\pi}$ in Eq. 36.

Case 3, parametric randomized policy with particular parameters: We consider a parametric randomized policy $\boldsymbol{\theta}$ in which only the parameters at a particular state, say state k, have changes. We similarly obtain the corresponding difference formula and derivative formula for the reward variance of Markov systems. Suppose that the parameters $\theta_{k,a}$ are changed to $\theta'_{k,a}$, $a \in \boldsymbol{\mathcal{A}}$ and other parameters $\theta_{i,a'}$ are fixed, $i \neq k$ and $a' \in \boldsymbol{\mathcal{A}}$. With Eq. 25, the variance difference of Markov systems in this situation is

$$\begin{aligned}
\eta'_{\sigma} - \eta_{\sigma} &= \boldsymbol{\pi'}(k)\left[\sum_{j \in \boldsymbol{\mathcal{S}}}\sum_{a \in \boldsymbol{\mathcal{A}}}(\theta'_{k,a} - \theta_{k,a})p(j|k,a)g_{\sigma}(j) + \sum_{a \in \boldsymbol{\mathcal{A}}}(\theta'_{k,a} - \theta_{k,a})r^2(k,a)\right. \\
&\qquad \left. -2\eta\sum_{a \in \boldsymbol{\mathcal{A}}}(\theta'_{k,a} - \theta_{k,a})r(k,a)\right] - (\eta' - \eta)^2 \\
&= \boldsymbol{\pi'}(k)\sum_{a \in \boldsymbol{\mathcal{A}}}(\theta'_{k,a} - \theta_{k,a})\left[\sum_{j \in \mathcal{S}}p(j|k,a)g_{\sigma}(j) + r^2(k,a) - 2\eta r(k,a)\right] - (\eta' - \eta)^2, \qquad (37)
\end{aligned}$$

where the term of square bracket can also be represented by $G(k,a)$ defined in Eq. 27. With the above difference formula, we can further derive the derivative formula of the reward variance with respect to parameter $\theta_{k,a}$ as below.

$$\begin{aligned}
\frac{\mathrm{d}\eta_{\sigma}}{\mathrm{d}\theta_{k,a}} &= \boldsymbol{\pi}(k)\left[\sum_{j \in \boldsymbol{\mathcal{S}}}p(j|k,a)g_{\sigma}(j) + r^2(k,a) - 2\eta r(k,a)\right] \\
&= \boldsymbol{\pi}(k)G(k,a). \qquad (38)
\end{aligned}$$

Note that when the reward function is independent of the action, i.e., $r(i,a) = r(i), \forall a \in \boldsymbol{\mathcal{A}}$, we can simplify the above derivative formula (38) as below.

$$\frac{\mathrm{d}\eta_{\sigma}}{\mathrm{d}\theta_{k,a}} = \boldsymbol{\pi}(k)\sum_{j \in \boldsymbol{\mathcal{S}}}p(j|k,a)g_{\sigma}(j), \qquad (39)$$

where the term $r^2(k,a) - 2\eta r(k,a)$ disappears because this term has a fixed value for different actions and it can be eliminated in Eq. 37 since $\sum\limits_{a \in \boldsymbol{\mathcal{A}}} (\theta'_{k,a} - \theta_{k,a}) = 0$.

Remark 5 The derivative formula (39) is the same as the result (45) in our previous paper (Xia 2016b) at the condition that the reward function $\boldsymbol{r}$ is unvaried under different parameters $\boldsymbol{\theta}$. Therefore, Eq. 38 is more general than the result in (Xia 2016b) and it quantifies the system derivatives when r is varied under different parameters or policies.

Therefore, with the difference formula (37) and the derivative formula (38), we can optimize the parameter $\boldsymbol{\theta}$ and reduce the reward variance of Markov systems. With Eq. 37, we observe that in order to reduce the reward variance, we have to choose $\theta'_{k,a}$'s that make the value of $\sum\limits_{a \in \boldsymbol{\mathcal{A}}} \theta'_{k,a} \left[\sum\limits_{j \in \boldsymbol{\mathcal{S}}} p(j|k,a) g_{\sigma}(j) + r^2(k,a) - 2\eta r(k,a) \right]$, i.e., $\sum\limits_{a \in \boldsymbol{\mathcal{A}}} \theta'_{k,a} G(k,a)$, as small as possible. Since $(\eta' - \eta)^2$ is always nonnegative and $\boldsymbol{\pi'}(k)$ is always positive, the above selection rule of $\theta'_{k,a}$'s will effectively reduce the variance of Markov systems. With a further analysis, we can directly derive the following theorem about this variance minimization problem in parameterized Markov systems.

Theorem 2 *For the variance minimization problem of Markov systems formulated in* Eq. 9, *the optimal policy can be found in the deterministic policy space.*

Proof Since the parametric randomized policy is more general than the deterministic policy, we only have to prove that the optimal parametric randomized policy can be found in the deterministic policy space. Therefore, we focus on the optimization of the parametric randomized policy. We study a situation in which the parameters $\theta_{k,a}$'s on a particular state k are to be optimized. With the variance difference formula (37) and the necessary condition in Theorem 1, we can directly have the following result. If $\boldsymbol{\theta^*}_k := (\theta^*_{k,a_1}, \cdots, \theta^*_{k,a_A})$ is optimal, it has to satisfy the following necessary condition

$$\boldsymbol{\theta}^*_k \in \begin{cases} \underset{\boldsymbol{\theta}_k}{\operatorname{argmin}} \sum\limits_{a \in \boldsymbol{\mathcal{A}}} \theta_{k,a} G^*(k,a), \\ \text{s.t.} \quad \sum\limits_{a \in \boldsymbol{\mathcal{A}}} \theta_{k,a} = 1, \quad \theta_{k,a} \geq 0, \ \forall a \in \boldsymbol{\mathcal{A}}. \end{cases} \tag{40}$$

In the above problem, the values of all the parameters are known except $\theta_{k,a}$'s. Obviously, the above problem is a linear program with optimization variables $\theta_{k,a}$'s, $a \in \boldsymbol{\mathcal{A}}$. According to the theory of linear programming, it is well known that the optimal solution $\theta^*_{k,a}$ can be found on the vertex of the multidimensional polyhedron of feasible solution $\theta_{k,a}$'s. From the constraints in Eq. 40, we can see that the value domain of $\theta_{k,a}$ is [0, 1]. Therefore, the optimal solution $\theta^*_{k,a}$ can be either 0 or 1, for all $k \in \boldsymbol{\mathcal{S}}$ and $a \in \boldsymbol{\mathcal{A}}$, which means that the optimal policy can be deterministic. The theorem is proved. □

Remark 6 The optimality of deterministic policy for the variance minimization problem of MDPs is similar to the analogous result in a standard MDP with discounted or average criterion.

Note that for the mean-variance optimization problem in MDPs, the optimal policy cannot be guaranteed as a deterministic policy (Chung 1994; Mannor and Tsitsiklis 2011). The mean-variance optimization problem can be viewed as a constrained optimization problem that minimizes the reward variance with a constraint of mean performance. The optimal policy may be randomized in many situations. However, as we proved in Theorem 2, the optimal policy of our problem (9) can be deterministic. This result lets us focus the optimization attention on the deterministic policy space, which greatly reduces the optimization complexity.

With the variance difference formula (28) and Theorems 1 and 2, we can further develop an iterative algorithm to reduce the reward variance of the parameterized MDP problem (9).

Algorithm 1 Iterative algorithm to reduce the reward variance of parameterized MDPs.

Initialization

- Arbitrarily choose an initial parametric policy $\theta^{(0)}$ from the policy space Θ, set $l = 0$.

Policy Evaluation

- For the current policy $\theta^{(l)}$, numerically compute or estimate η, η_σ, g_σ, and $G(i,a)$'s based on their definitions respectively.

Policy Improvement

- Update the policy as follows:

$$\theta_i^{(l+1)} = \underset{\theta_i 1=1,\ \theta_{i,a}\geq 0}{\operatorname{argmin}} \left\{ \sum_{a\in\mathcal{A}} \theta_{i,a} G(i,a) \right\}, \quad \forall i \in \mathcal{S}. \tag{41}$$

We keep $\theta_i^{(l+1)} = \theta_i^{(l)}$ if possible to avoid oscillations.

Stopping Rule

- If $\theta^{(l+1)} = \theta^{(l)}$, stop; Otherwise, let $l \leftarrow l+1$ and go to step 2.

The main procedure of Algorithm 1 is similar to the policy iteration in the traditional MDP theory. The policy improvement step (41) can be further written as below.

$$\boldsymbol{\theta}_i^{(l+1)} = (0, \cdots, 0, 1, 0, \cdots, 0), \quad \text{where } \theta_{i,a^*}^{(l+1)} = 1 \text{ with } a^* = \underset{a\in\boldsymbol{\mathcal{A}}}{\operatorname{argmin}} \left\{G(i,a)\right\}. \tag{42}$$

The above formula means that the updated policy is deterministic, which is in accordance with Theorem 2.

In Algorithm 1, we can see that the key step is to compute the value of $G(i,a)$'s at every iteration. The variance of the Markov chain will be reduced after every iteration. With Theorem 2 or Eq. 42, we know that the policies derived by Algorithm 1 are deterministic. Based on these facts, we can further prove that Algorithm 1 will converge to a *local optimum* that is defined in the randomized policy space. The similar result can also be found in our previous paper (Xia 2016b), although the targeted problem models in these two papers are different (in this paper we study the parameterized MDPs with varied reward function, while in Xia (2016b) we study deterministic policies with unvaried reward functions). The main idea to prove the local optimum can be partly motivated by Eq. 40. When Algorithm 1 stops, it indicates that $\theta^*_{k,a^*} = 1$ for $a^* = \underset{a\in\boldsymbol{\mathcal{A}}}{\operatorname{argmin}}\{G(k,a)\}$ and $\theta^*_{k,a} = 0$ for other actions $a \neq a^*$. With the derivative formula (38), it is easy to verify that the total derivative will be positive if we change the values of $\theta^*_{k,a}$'s in a small enough neighborhood, which means that the convergence point is a local optimum in the randomized policy space. We omit the proof details as the space limitation. Interested readers can refer to the proof of Theorem 5 in our previous paper (Xia 2016b).

Although currently we cannot give a specific analysis for the algorithmic complexity of Algorithm 1, we can refer to the existing results of complexity analysis for the classical policy iteration since Algorithm 1 is similar to that. For the steps 2-3 in Algorithm 1, the time-complexity for computing η, η_σ, g_σ, and $G(i,a)$'s is of complexity $O(S^3)$ approximately, since it involves solving linear equations such as Eqs. 15 and 18. The time-complexity for executing Eq. 41 is of complexity $O(SA)$, since we need $S \times A$ comparisons at most if we use Eq. 42. The iterative-complexity of the classical policy iteration is still an open question (Littman et al. 1995). It has been showed with counter examples that a simple policy iteration (update actions at only one state per iteration) may require exponential times of iterations to find the optimal policy (Melekopoglou and Condon 1990). However, the classical policy iteration usually shows a very fast convergence rate for most of small-scale problems. It is reasonable to argue that Algorithm 1 also has a good performance of convergence for many small-scale problems. For large-scale problems, we may resort to approximation techniques to reconstruct Algorithm 1, such as approximate dynamic programming (Powell 2007), neuro-dynamic programming (Bertsekas and Tsitsiklis 1996), deep neural networks (Silver et al. 2016), and other data-driven learning techniques.

4 Extension

In the previous section, we study the parametric policy in which $\theta_{i,a}$ is the probability of selecting action a at state i, $i \in \boldsymbol{\mathcal{S}}$ and $a \in \boldsymbol{\mathcal{A}}$. In this section, we study a general case in which $\boldsymbol{\theta}$ is a set of parameters that will affect the value of $\boldsymbol{P}$ and $\boldsymbol{r}$.

First, we give a problem formulation for such general parameterized MDPs. With a little abuse of notations, we denote $\boldsymbol{\theta}$ as an N-dimensional vector as below.

$$\boldsymbol{\theta} := (\theta_1, \theta_2, \cdots, \theta_N). \tag{43}$$

The change of the value of θ_n will change the values of the transition probabilities $p(i,:)$'s and the rewards $r(i)$'s for some states i's, $n = 1, 2, \cdots, N$, $i \in \boldsymbol{\mathcal{S}}$. Therefore, the whole state space $\boldsymbol{\mathcal{S}}$ can be partitioned based on the following definition.

Definition 1 $\boldsymbol{\mathcal{S}}_n$ is defined as the set of states i whose transition probabilities $p(i,:)$ and reward $r(i)$ are affected by θ_n, $n = 1, 2, \cdots, N$.

Different parameters θ_n's have different $\boldsymbol{\mathcal{S}}_n$'s. For simplicity, we consider a special case that the state sets $\boldsymbol{\mathcal{S}}_n$'s are mutually exclusive. That is, we have the following assumption

Assumption 1 The set of states $\boldsymbol{\mathcal{S}}_n$'s are mutually exclusive, i.e., $\boldsymbol{\mathcal{S}}_n \cap \boldsymbol{\mathcal{S}}_m = \varnothing$ if $n \neq m$.

With this assumption, we can see that the state space $\boldsymbol{\mathcal{S}}$ can be partitioned by the parameter $\boldsymbol{\theta}$ and every state's transition probabilities $p(i,:)$ and reward $r(i)$ are controlled by only one parameter θ_n, where $i \in \boldsymbol{\mathcal{S}}_n$. With Assumption 1, we can partition the state space $\boldsymbol{\mathcal{S}}$ into a series of subsets $\boldsymbol{\mathcal{S}}_n$'s according to θ_n's. That is, we have

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}_0 \cup \boldsymbol{\mathcal{S}}_1 \cup \cdots \cup \boldsymbol{\mathcal{S}}_N, \tag{44}$$

where $\boldsymbol{\mathcal{S}}_0$ is the set of states whose $p(i,:)$ and $r(i)$ are not affected by $\boldsymbol{\theta}$, where $i \in \boldsymbol{\mathcal{S}}_0$. In special cases, we may have $\boldsymbol{\mathcal{S}}_0 = \varnothing$. Below, we give an example of admission control in queueing networks to illustrate the above definitions.

Example 1 Consider an open Jackson network with 3 servers. The system state is $n := (n_1, n_2, n_3)$, where n_k is the number of customers at server k. We assume that the whole network has a capacity $N = 4$, i.e., the number of total customers cannot exceed 4. We conduct admission control at the entrance of the network. In specific, the newly arriving customers are admitted to enter the network with an admission probability a_n, where n is the number of total customers observed by the arriving customer, $n = 0, 1, \cdots, 4$, and $a_n \in \mathbb{R}[0,1]$. Obviously, we always have $a_4 = 0$. Therefore, this optimization problem is a parameterized MDP and the optimization parameter is $\boldsymbol{\theta} = (a_0, a_1, a_2, a_3)$. If we change the value of parameter a_1, then the transition probability and reward at the state subset $\boldsymbol{\mathcal{S}}_1 = \{(0,0,1), (0,1,0), (1,0,0)\}$ will be affected. We can easily verify that this admission control problem satisfies the above assumptions. More details about this admission control problem can be referred to our previous work (Xia 2014; Xia and Jia 2015).

Similar to the notations in Section 3, we also use $\boldsymbol{P}^{\boldsymbol{\theta}}$ and $\boldsymbol{r}^{\boldsymbol{\theta}}$ to denote the effect of $\boldsymbol{\theta}$ on the dynamics of Markov systems. The long-run average performance of the Markov system under the parameter $\boldsymbol{\theta}$ is

$$\eta^{\boldsymbol{\theta}} := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left\{\sum_{t=0}^{T-1} r^{\boldsymbol{\theta}}(X_t)\right\}. \tag{45}$$

The reward variance of the Markov system is

$$\eta_\sigma^{\boldsymbol{\theta}} := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left\{\sum_{t=0}^{T-1} \left[r^{\boldsymbol{\theta}}(X_t) - \eta^{\boldsymbol{\theta}}\right]^2\right\}. \tag{46}$$

The value domain of the parameter $\boldsymbol{\theta}$ is an N-dimensional polyhedron in real number space and we denote it as Θ, $\Theta \in \mathbb{R}^N$. Our goal is to find the optimal parameter $\boldsymbol{\theta^*}$ such that the reward variance is minimized, i.e.,

$$\boldsymbol{\theta^*} = \operatorname*{argmin}_{\theta\in\Theta} \left\{\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(r^{\boldsymbol{\theta}}(X_t) - \eta^{\boldsymbol{\theta}}\right)^2\right]\right\}. \tag{47}$$

In Section 3, we define the parametric policy $\theta_{i,a}$ with which the action selection is randomized, so the system reward is also randomized and we define the variance function as Eq. 14. In this section, the parameter is $\boldsymbol{\theta}$ and the system reward is deterministic and we denote it as $r^\theta(i)$. Therefore, we define the variance function in this parameterized MDPs as below.

$$f_\sigma^{\boldsymbol{\theta}}(i) = (r^{\boldsymbol{\theta}}(i) - \eta^{\boldsymbol{\theta}})^2. \tag{48}$$

For notation simplicity, we also omit the superscript "$^\theta$" by default and use $\boldsymbol{P'}, \boldsymbol{r'}, \eta', \eta'_\sigma$ to replace $\boldsymbol{P}^{\boldsymbol{\theta'}}, \boldsymbol{r}^{\boldsymbol{\theta'}}, \eta^{\boldsymbol{\theta'}}, \eta_\sigma^{\boldsymbol{\theta'}}$ respectively. Similar to the analysis in Section 3, we can apply the sensitivity-based optimization theory to this problem and derive the variance difference formula for this parameterized MDP when the parameter is changed from $\boldsymbol{\theta}$ to $\boldsymbol{\theta'}$.

$$\begin{aligned}
\eta'_\sigma - \eta_\sigma &= \boldsymbol{\pi}'[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{f'}_\sigma - \boldsymbol{f}_\sigma)] \\
&= \boldsymbol{\pi}'[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{r'} - \eta'\mathbf{1})^2_\odot - (\boldsymbol{r} - \eta\mathbf{1})^2_\odot] \\
&= \boldsymbol{\pi}'[(\boldsymbol{P'} - \boldsymbol{P})\boldsymbol{g}_\sigma + (\boldsymbol{r'} - \eta\mathbf{1})^2_\odot - (\boldsymbol{r} - \eta\mathbf{1})^2_\odot] - (\eta' - \eta)^2.
\end{aligned} \tag{49}$$

The above formula has the same form as Eq. 34 in which we study deterministic policies. With Assumption 1, we can further rewrite the above formula as below.

$$\eta'_\sigma - \eta_\sigma = \sum_{n=1}^{N} \sum_{i \in \boldsymbol{\mathcal{S}}_n} \boldsymbol{\pi}'(i) \left[\sum_{j \in \boldsymbol{\mathcal{S}}} (p'(i,j) - p(i,j)) g_\sigma(j) + (r'(i) - \eta)^2 - (r(i) - \eta)^2 \right] - (\eta' - \eta)^2. \tag{50}$$

Note that in the above formula, $p'(i, j)$ and $r'(i)$ are affected only by the value of θ'_n for $i \in \boldsymbol{\mathcal{S}}_n$, but $\boldsymbol{\pi'}(i)$ and η' are affected by the value of the whole parameter vector $\boldsymbol{\theta'}$.

With the variance difference formula (50), we further study the performance derivatives. Suppose that the parameter θ_k is changed to θ'_k, while other parameters θ_n remain unvaried, $n = 1, 2, \cdots, N$ and $n \neq k$. The above difference formula (50) becomes

$$\eta'_\sigma - \eta_\sigma = \sum_{i \in \boldsymbol{\mathcal{S}}_k} \boldsymbol{\pi}'(i) \left[\sum_{j \in \boldsymbol{\mathcal{S}}} (p'(i,j) - p(i,j)) g_\sigma(j) + (r'(i) - \eta)^2 - (r(i) - \eta)^2 \right] - (\eta' - \eta)^2. \tag{51}$$

Taking the derivative operation w.r.t. θ_k on the above formula, we can obtain

$$\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\theta_k} = \sum_{i \in \boldsymbol{\mathcal{S}}_k} \boldsymbol{\pi}(i) \left[\sum_{j \in \boldsymbol{\mathcal{S}}} \frac{\mathrm{d}p(i,j)}{\mathrm{d}\theta_k} g_\sigma(j) + 2(r(i) - \eta) \frac{\mathrm{d}r(i)}{\mathrm{d}\theta_k} \right], \quad k = 1, 2, \cdots, N. \tag{52}$$

In the above analysis, we assume that the parameterized MDP has the special structure defined in Assumption 1. For a general case in which the problem does not have such structures, we can conduct similar analysis and obtain the following derivative formula in a matrix form

$$\frac{\mathrm{d}\eta_\sigma}{\mathrm{d}\theta} = \boldsymbol{\pi} \left[\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\theta} \boldsymbol{g}_\sigma + 2(\boldsymbol{r} - \eta \mathbf{1}) \odot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}\theta} \right], \tag{53}$$

where $\boldsymbol{\theta}$ is a scalar parameter, $\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\theta}$ and $\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}\theta}$ are matrix and vector derivatives w.r.t. $\boldsymbol{\theta}$, respectively.

5 Numerical experiments

In this section, we conduct numerical experiments to verify the main results of this paper. Consider a Markov chain with state space $\boldsymbol{\mathcal{S}} = \{1, 2, 3\}$ and action space $\boldsymbol{\mathcal{A}} = \{a_1, a_2, a_3\}$. The transition probabilities are different under different actions. For state $i = 1$, we have $p(:|1, a_1) = (0.6, 0.2, 0.2)$, $p(:|1, a_2) = (0.2, 0.5, 0.3)$, $p(:|1, a_3) = (0.1, 0.2, 0.7)$; For state $i = 2$, we have $p(:|2, a_1) = (0.5, 0.3, 0.2)$, $p(:|2, a_2) = (0.2, 0.7, 0.1)$, $p(:|2, a_3) = (0.1, 0.1, 0.8)$; For state $i = 3$, we have $p(:|3, a_1) = (0.4, 0.2, 0.4)$, $p(:|3, a_2) = (0.1, 0.6, 0.3)$, $p(:|3, a_3) = (0.2, 0.1, 0.7)$. The system reward is varied under different actions adopted, which is different from the unvaried reward function used in our previous work (Xia 2016b). For state $i = 1$, we have $r(1, a_1) = 1$, $r(1, a_2) = 2$, $r(1, a_3) = 3$; For state $i = 2$, we have $r(2, a_1) = 5$, $r(2, a_2) = 1$, $r(2, a_3) = 3$; For state $i = 3$, we have $r(3, a_1) = 6$, $r(3, a_2) = 4$, $r(3, a_3) = 2$. The optimization parameters are the action selection probabilities at every state, as defined in Eq. 5. The goal is to find the optimal parameter $\boldsymbol{\theta^*}$ that minimizes the variance of the system rewards of this Markov chain.

By applying Algorithm 1, we conduct the policy iteration type algorithm to reduce the reward variance. We compute the value of η, η_σ, and $\boldsymbol{g_\sigma}$ under the current policy, thus we

Table 1 4 different local optima to which Algorithm 1 may converge

$\mathcal{L}$	(a_2, a_3, a_3)	(a_3, a_3, a_2)	(a_2, a_2, a_3)	(a_1, a_2, a_3)
η	2.1731	3.5267	1.5500	1.3333
η_σ	0.1431	0.2493	0.2475	0.2222

obtain the value of $G(i, a)$'s using Eq. 27. Then we use the policy improvement formula (41) or (42) to find a new policy improved. As stated in Theorem 2, the optimal policy can be found in the deterministic policy space. Therefore, we can simplify the form of parametric policy from a 3 × 3 matrix $\boldsymbol{\theta}$ to a vector $\boldsymbol{\mathcal{L}}$. For example, $\boldsymbol{\mathcal{L}} = (a_2, a_3, a_1)$ indicates that we choose action a_2 at state 1, action a_3 at state 2, and action a_1 at state 3. If in a matrix form as (5), it indicates

$$\boldsymbol{\theta} = \begin{pmatrix} 0, & 1, & 0 \\ 0, & 0, & 1 \\ 1, & 0, & 0 \end{pmatrix}.$$

We enumerate all the initial policies and find that Algorithm 1 typically converges within 1 or 2 iterations. There are 4 different policies to which Algorithm 1 may converge, as we illustrate in Table 1. These 4 policies are the local minima of this variance minimization problem. If Algorithm 1 starts with different initial policies, it may converge to different local optimum policies. The first column in Table 1, $\boldsymbol{\mathcal{L}} = (a_2, a_3, a_3)$ and $\eta_\sigma = 0.1431$, is the global minimum of this variance minimization problem.

Since this Markov chain is a small example and it only has $3^3 = 27$ different deterministic policies, we enumerate all these policies and obtain their mean and variance of system rewards. Plotting them in a 2-dimension plane, we obtain Fig. 2, where the star-point is the global optimum and the triangle-points are the local optimum. If our goal is to maximize the mean while minimize the variance, we can obtain the efficient frontier of this 2-objective optimization problem, as illustrated in Fig. 2. For these 4 solutions listed in Table 1, we can see that the first solution is dominant over the third and fourth solutions both in mean and variance.

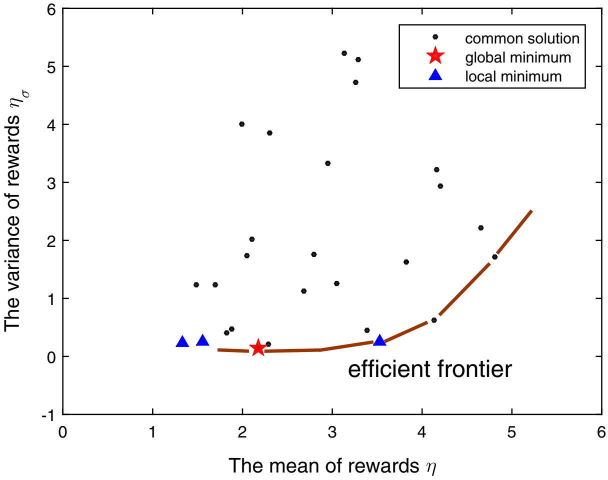

Fig. 2 The mean and variance of different policies and the efficient frontier

6 Conclusion

In this paper, we study the optimization of parameterized MDPs under the variance criterion. The variance difference formulas (25) and (28) are the key findings of this paper and the nonnegative term $(\eta' - \eta)^2$ is the key term of the variance difference formula. The sensitivity-based optimization theory provides a new perspective to study this parameterized MDP, which is different from the traditional gradient-based approach. Based on the above results, we further derive a necessary condition for the optimal parametric policy. The optimality of deterministic policy for this variance minimization problem is also proved, which can be utilized to greatly reduce the optimization complexity. Finally, we develop an iterative algorithm to efficiently reduce the variance of Markov systems and conduct numerical experiments to demonstrate the main results of this paper.

During the implementation of the optimization algorithm, one of the key problems is to efficiently compute or estimate the quantity $\boldsymbol{g_\sigma}$ or $G(i, a)$'s. This problem is similar to the computation or estimation of value functions or Q-factors in the classical MDP theory. The similar ideas, such as the approximate dynamic programming or other function approximation techniques (Bertsekas 2012), can also be considered to handle the curse of dimensionality issue in our problem. On the other hand, we consider only the variance criterion in this paper, regardless of the average criterion. How to extend our approach to the mean-variance optimization is another important topic deserving future investigations.

Acknowledgments This work was supported in part by the National Key Research and Development Program of China (2016YFB0901900), the National Natural Science Foundation of China (61573206, 61203039, U1301254), and the Suzhou-Tsinghua Innovation Leading Action Project.

References

Bertsekas DP (2012) Dynamic programming and optimal control – vol II, 4th edn. Athena Scientific, Massachusetts

Bertsekas DP, Tsitsiklis JN (1996) Neuro-dynamic programming, Athena scientific. Belmont, Massachusetts

Cao XR (2007) Stochastic learning and optimization – a sensitivity-based approach. Springer, New York

Cao XR, Chen HF (1997) Perturbation realization, potentials, and sensitivity analysis of Markov processes. IEEE Trans Autom Control 42:1382–1393

Chung KJ (1994) Mean-variance tradeoffs in an undiscounted MDP: the unichain case. Oper Res 42(1): 184–188

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

Guo X, Hernandez-Lerma O (2009) Continuous-time Markov decision processes. Springer, Theory and Applications

Guo X, Huang X, Zhang Y (2015) On the first passage g-mean-variance optimality for discounted continuous-time Markov decision processes. SIAM J Control Optim 53(3):1406–1424

Guo X, Ye L, Yin G (2012) A mean-variance optimization problem for discounted Markov decision processes. Eur J Oper Res 220:423–429

Hernandez-Lerma O, Vega-Amaya O, Carrasco G (1999) Sample-path optimality and variance-minimization of average cost Markov control processes. SIAM J Control Optim 38:79–93

Huo H, Zou X, Guo X (2017) The risk probability criterion for discounted continuous-time Markov decision processes. Discrete Event Dynamic Systems: Theory and Applications

Littman ML, Dean TL, Kaelbling LP (1995) On the complexity of solving Markov decision problems. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.

Luh PB, Yu Y, Zhang B, Litvinov E, Zheng T, Zhao F, Zhao J, Wang C (2014) Grid integration of intermittent wind generation: a Markovian approach. IEEE Trans Smart Grid 5(2):732–741

Marbach P, Tsitsiklis JN (2001) Simulation-based optimization of Markov reward processes. IEEE Trans Autom Control 46:191–209

Markowitz H (1952) Portfolio selection. J Financ 7:77–91

Mannor S, Tsitsiklis JN (2011) Mean-variance optimization in Markov decision processes. In: Proceedings of the 28th international conference on machine learning. Bellevue, WA, USA

Melekopoglou M, Condon A (1990) On the complexity of the policy iteration algorithm for stochastic games. Technical Report CS-TR-90-941, Computer Sciences Department, University of Wisconsin Madison

Powell WB (2007) Approximate dynamic programming: solving the curses of dimensionality. Wiley

Puterman ML (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley, New York

Silver D, Huang A, Maddison CJ et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489

Sobel MJ (1994) Mean-variance tradeoffs in an undiscounted MDP. Oper Res 42:175–183

Tamar A, Castro DD, Mannor S (2012) Policy gradients with variance related risk criteria. In: Proceedings of the 29th international conference on machine learning (ICML). Edinburgh, Scotland

Ummels BC, Gibescu M, Pelgrum E, Kling WL, Brand AJ (2007) Impacts of wind power on thermal generation unit commitment and dispatch. IEEE Trans Energy Convers 22:44–51

Xia L (2014) Event-based optimization of admission control in open queueing networks. Discrete Event Dynamic Systems: Theory and Applications 24(2):133–151

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 (WODES2016). Xi'an, China, May 30-June 1, pp 332–337

Xia L (2016) Optimization of Markov decision processes under the variance criterion. Automatica 73: 269–278

Xia L, Jia QS (2015) Parameterized Markov decision process and its application to service rate control. Automatica 54:29–35

Li Xia is an associate professor at the Center for Intelligent and Networked Systems (CFINS), Department of Automation, Tsinghua University, Beijing China. He received the B.S. degree 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 in 2011. He was a visiting scholar at Stanford University, the Hong Kong University of Science and Technology, etc. He serves/served as an associate editor and program committee member of a number of international journals and conferences. His research interests include the methodology research in stochastic learning and optimization, queueing theory, Markov decision processes, reinforcement learning, and the application research in building energy, energy Internet, industrial Internet, Internet of things, etc. He is a senior member of IEEE.