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).

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