Automatica 88 (2018) 76–82

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Mean–variance optimization of discrete time discounted Markov decision processes☆

Li Xia

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

ARTICLE INFO

Article history:
Received 13 September 2016
Received in revised form 2 June 2017
Accepted 28 August 2017
Available online 22 December 2017

Keywords:
Markov decision process
Mean–variance optimization
Variance criterion
Sensitivity-based optimization

ABSTRACT

In this paper, we study a mean–variance optimization problem in an infinite horizon discrete time discounted Markov decision process (MDP). The objective is to minimize the variance of system rewards with the constraint of mean performance. Different from most of works in the literature which require the mean performance already achieve optimum, we can let the discounted performance equal any constant. The difficulty of this problem is caused by the quadratic form of the variance function which makes the variance minimization problem not a standard MDP. By proving the decomposable structure of the feasible policy space, we transform this constrained variance minimization problem to an equivalent unconstrained MDP under a new discounted criterion and a new reward function. The difference of the variances of Markov chains under any two feasible policies is quantified by a difference formula. Based on the variance difference formula, a policy iteration algorithm is developed to find the optimal policy. We also prove the optimality of deterministic policy over the randomized policy generated in the mean-constrained policy space. Numerical experiments demonstrate the effectiveness of our approach.

1. Introduction

The mean–variance optimization is an important problem in stochastic optimization and its origin can go back to the pioneering work by H. Markowitz, 1990 Nobel Laureate in Economics, on the modern portfolio management (Markowitz, 1952). In financial engineering, the mean indicates the return of assets and the variance indicates the risk of assets. The objective of the mean–variance optimization is to find an optimal policy such that the mean and the variance of system rewards are optimized in tradeoff and the efficient frontier (a curve comprised of *Pareto optima*) is obtained.

The mean–variance optimization is first proposed in a static optimization form in Markowitz's original paper (Markowitz, 1952), in which the decision variables are the investment percentage of securities and the securities returns are described as random variables with known means and variances. Then, the mean–variance optimization is further studied in a dynamic optimization form and Markov decision processes (MDPs) are widely used as an important analytical model. The difficulty of this problem mainly comes from the *non-additiveness* of the variance criterion which makes the principle of *consistent choice* in dynamic programming invalid (Sobel, 1982; Xia, 2016). Such invalidness means that the optimal action selection during $[t+1, \infty)$ may be not optimal for the action selection during $[t, \infty)$. In the literature, there are different ways to study the mean–variance optimization. Many works studied the variance minimization of MDPs after the mean performance is already maximized (Guo & Song, 2009; Hernandez-Lerma, Vega-Amaya, & Carrasco, 1999; Huang & Chen, 2012). For such problem, the variance minimization problem can be transformed to another standard MDP under an equivalent average or discounted criterion. There are also studies that use the policy gradient approach to study the mean–variance optimization when the policy is parameterized (Prashantha & Ghavamzadeh, 2013; Tamar, Castro, & Mannor, 2012).

Sobel (1982) gave an early study on the mean–variance optimization in a discrete time discounted Markov chain, but no optimization algorithm was presented in that paper. Chung (1994) and Sobel (1994) studied the variance minimization problem in a discrete time Markov chain with the constraint that the long-run average performance is larger than a given constant. This problem was transformed to a sequence of linear programming problems, which may have concerns of computation efficiency since the number of sequential problems may be large. Guo, Huang, and Zhang (2015) and Huo, Zou, and Guo (2017) studied the mean–variance optimization problem in a continuous time Markov chain with unbounded transition rates and state–action dependent discount factors, where the performance is accumulated until a

☆ The material in this paper was not presented at any conference. 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.

https://doi.org/10.1016/j.automatica.2017.11.012
0005-1098/

certain state is reached. There are certainly numerous other excellent works about the mean–variance optimization in the literature. However, most of the works in the literature either require a condition of optimal mean performance or reformulate the problem as variations of mathematical programming. Although linear programming may be used to study the mean–variance optimization in some cases, it does not utilize the structure of Markov systems and the efficiency is not satisfactory. Policy iteration is a classical approach in dynamic programming and it usually has a high convergence efficiency. There is little work to study the mean–variance optimization using policy iteration, at the condition that the mean performance equals a given value.

In this paper, we study a mean–variance optimization problem in an infinite horizon discrete time discounted Markov chain. The objective is to find the optimal policy with the minimal variance of rewards from the policy set in which the discounted performance equals a given constant. The motivation of this problem can be explained with a financial example. People may not always choose an asset portfolio with the maximal expected return, since a portfolio with big return usually has big risk (quantified by variance). People always like to seek a portfolio with minimal risk and acceptable return. The solution with minimal risk and fixed return is called *Pareto optimum*. All the Pareto optimal solutions compose a curve called *Pareto frontier*, or *efficient frontier* in financial engineering.

The difficulty of such mean–variance optimization problem mainly comes from two aspects. The first one is the difficulty caused by the non-additiveness of the variance criterion, which makes the mean–variance optimization not a standard MDP and policy iteration is not applicable directly. Another difficulty comes from the fact that the policy set with a fixed mean performance usually has no satisfactory structure, such as that described later in Theorem 1. For example, the policy set whose long-run average performance equals a given constant may not be decomposable as that in Theorem 1. Without such property, dynamic programming and policy iteration cannot be used for these problems.

In this paper, we use the sensitivity-based optimization theory to study this nonstandard MDP problem. For the policy set in which the discounted performance equals a given constant, we prove that this policy set is decomposable on the action space and the action can be chosen independently at every state. A difference formula is derived to quantify the variance difference under any two feasible policies. The original variance minimization problem with constraints is transformed to a standard unconstrained MDP under an equivalent discounted criterion with a new discount factor β^2 and a new reward function, where β is the discount factor of the original Markov chain. With this equivalent MDP, we prove the existence of the optimal policy for this mean–variance optimization problem. A policy iteration algorithm is developed to find the optimal policy with the minimal variance. The optimality of deterministic policy is also proved, compared with randomized policies generated in the mean-constrained policy space. Finally, we conduct a numerical experiment to demonstrate the effectiveness of our approach. The efficient frontier of this numerical example is also analyzed.

This paper is a continued work compared with our previous papers (Xia, 2016, 2017), which aim to minimize the variance of the long-run average performance of the Markov chain without considering the constraint of mean performance. The targeted models in these papers are different, so are the main results. To the best of our knowledge, this is the first paper that develops a policy iteration algorithm to minimize the variance of a discrete time discounted Markov chain at the condition of any given discounted performance.

2. Problem formulation

We consider a finite MDP in discrete time. X_t is denoted as the system state at time t, $t = 0, 1, \ldots$. The state space is finite and denoted as $\mathcal{S} = \{1, 2, \ldots, S\}$, where S is the size of the state space. We only consider the deterministic and stationary policy d which is a mapping from the state space to the action space. If the current state is i, the policy d determines to choose an action a from a finite action space $\mathcal{A}(i)$ and a system reward $r(i, a)$ is obtained. The system will transit to a new state j with a transition probability $p(j|i, a)$ at the next time epoch, where $i, j \in \mathcal{S}$ and $a \in \mathcal{A}(i)$. Obviously, we have $\sum_{j\in\mathcal{S}} p(j|i, a) = 1$. Since d is a mapping in the state space, we have $a = d(i)$ and $d(i) \in \mathcal{A}(i)$ for all $i \in \mathcal{S}$. We define the policy space $\mathcal{D}$ as the family of all deterministic stationary policies. For each $d \in \mathcal{D}$, $\boldsymbol{P}(d)$ is denoted as a transition probability matrix and its (i, j)th element is $p(j|i, d(i))$, and $\boldsymbol{r}(d)$ is denoted as a column vector and its ith element is $r(i, d(i))$. We assume that the Markov chain is ergodic for any policy in $\mathcal{D}$. The discount factor of the MDP is β, $0 < \beta < 1$. For initial state i, the discounted performance of the MDP under policy d is defined as below.

$$J(d, i) := \mathbb{E}_i^d \left[\sum_{t=0}^{\infty} \beta^t r(X_t, d(X_t)) \right], \quad i \in \mathcal{S}, \tag{1}$$

where E_i^d is an expectation operator of the Markov chain at the condition that the initial state is i and the policy is d. $\boldsymbol{J}(d)$ is an S-dimensional column vector and its ith element is $J(d, i)$. The variance of the discounted Markov chain is defined as below.

$$\sigma^2(d, i) := \mathbb{E}_i^d \left[\left(\sum_{t=0}^{\infty} \beta^t r(X_t, d(X_t)) \right) - J(d, i) \right]^2, \ i \in \mathcal{S}. \tag{2}$$

We observe that $\sigma^2(d, i)$ quantifies the variance of the limiting random variable $\sum_{t=0}^{\infty} \beta^t r(X_t, d(X_t))$. $\boldsymbol{\sigma}^2(d)$ is the variance vector of the discounted Markov chain and its ith element is $\sigma^2(d, i)$.

Denote $\boldsymbol{\lambda}$ as a given mean reward function on $\mathcal{S}$. That is, $\boldsymbol{\lambda}$ is an S-dimensional column vector and its ith element is denoted as $\lambda(i)$, $i \in \mathcal{S}$. The set of all feasible policies with which the discounted performance of the Markov chain equals $\boldsymbol{\lambda}$ is defined as below.

$$\mathcal{D}_{\boldsymbol{\lambda}} := \{\text{all } d \in \mathcal{D} | J(d, i) = \lambda(i), \text{ for all } i \in \mathcal{S}\}. \tag{3}$$

Note that $\mathcal{D}$ is a deterministic stationary policy set, so is $\mathcal{D}_{\boldsymbol{\lambda}}$. In this paper, we do not consider randomized stationary policies. The optimality of deterministic policies will be studied in the next section, see Theorem 5. It is easy to see that the policy set $\mathcal{D}_{\boldsymbol{\lambda}}$ may be empty if the value of $\boldsymbol{\lambda}$ is not chosen properly. In this paper, we assume that $\mathcal{D}_{\boldsymbol{\lambda}}$ is not empty, which is similar to the assumption in Markowitz's mean–variance portfolio problem (Markowitz, 1952; Zhou & Yin, 2003). For a given discounted performance vector $\boldsymbol{\lambda}$, $\mathcal{D}_{\boldsymbol{\lambda}}$ may contain more than one policy. The objective is to find an optimal policy from $\mathcal{D}_{\boldsymbol{\lambda}}$ such that the variance of the Markov chain is minimized. The mathematical formulation is written as below.

$$\min_{d\in\mathcal{D}_{\boldsymbol{\lambda}}} \left\{\sigma^2(d, i)\right\}, \quad \text{for all } i \in \mathcal{S}. \tag{4}$$

That is, we aim to find an optimal policy among all feasible policies whose discounted performance is equal to a given constant vector $\boldsymbol{\lambda}$, such that the variance of discounted rewards is minimized. We denote such a *mean–variance optimal policy* as $d^*_{\boldsymbol{\lambda}}$. The existence of the solution $d^*_{\boldsymbol{\lambda}}$ to the problem (4) is not guaranteed because the minimization in (4) is over every state $i \in \mathcal{S}$, i.e., (4) can be viewed as a multi-objective optimization problem. For a general multi-objective optimization problem, it is possible that no solution can dominate all the other solutions on the value of every dimension of objective function. In the next section, we will discuss the existence of such optimal policy $d^*_{\boldsymbol{\lambda}}$ and develop an optimization algorithm to find it. Moreover, we have the following remarks.