European Journal of Operational Research 231 (2013) 645–653

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Algorithmic aspects of mean–variance optimization in Markov decision processes

Shie Mannor [a,*], John N. Tsitsiklis [b]

[a] Department of Electrical and Engineering, Technion, Haifa 32000, Israel
[b] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 12 June 2011
Accepted 12 June 2013
Available online 24 June 2013

Keywords:
Markov processes
Dynamic programming
Control
Complexity theory

ABSTRACT

We consider finite horizon Markov decision processes under performance measures that involve both the mean and the variance of the cumulative reward. We show that either randomized or history-based policies can improve performance. We prove that the complexity of computing a policy that maximizes the mean reward under a variance constraint is NP-hard for some cases, and strongly NP-hard for others. We finally offer pseudopolynomial exact and approximation algorithms.

1. Introduction

The classical theory of Markov decision processes (MDPs) deals with the maximization of the cumulative (possibly discounted) expected reward, to be denoted by W. However, a risk-averse decision maker may be interested in additional distributional properties of W. In this paper, we focus on the case where the decision maker is interested in both the mean and the variance of the cumulative reward (e.g., trying to optimize the mean subject to a variance constraint or vice versa), and we explore the associated computational issues.

Risk aversion in MDPs is of course an old subject. In one approach, the focus is on the maximization of $\mathbb{E}[U(W)]$, where U is a concave utility function. Problems of this type can be handled by state augmentation (e.g., Bertsekas, 1995), namely, by introducing an auxiliary state variable that keeps track of the cumulative past reward. In a few special cases, e.g., with an exponential utility function, state augmentation is unnecessary, and optimal policies can be found by solving a modified Bellman equation (Chung & Sobel, 1987). (The exponential utility function is often viewed as a surrogate for trading off mean and variance, on the basis of a single tunable parameter. The difficulty of solving mean–variance optimization problems—which is the focus of this paper—does provide some support for using a surrogate criterion, more amenable to exact optimization.) Another interesting case where optimal policies can be found efficiently involves a "one-switch utility functions" (the sum of a linear and an exponential) Liu and Koenig (2005), or piecewise linear utility functions with a single break point (Liu & Koenig, 2006).

In another approach, the objective is to optimize a so-called coherent risk measure (Artzner, Delbaen, Eber, & Heath, 1999), which turns out to be equivalent to a robust optimization problem: one assumes a family of probabilistic models and optimizes the worst-case performance over this family. In the multistage case (Riedel, 2004), problems of this type can be difficult (Le Tallec, 2007), except for some special cases (Iyengar, 2005; Nilim & El Ghaoui, 2005) that can be reduced to Markov games (Shapley, 1953).

Mean–variance optimization lacks some of the desirable properties of approaches involving coherent risk measures or risk-sensitive utility functions (e.g., exponential utility functions) and sometimes leads to counterintuitive policies. Bellman's principle of optimality does not hold, and as a consequence, a decision maker who has received unexpectedly large rewards in the first stages, may actively seek to incur losses in subsequent stages in order to keep the variance small. Counterintuitive and seemingly "irrational" behavior (i.e., incompatible with expected utility maximization) can even arise in static problems under a mean–variance formulation: for example, under a variance constraint, one may prefer to forgo a profit which is guaranteed to be positive but has a positive variance. Nevertheless, mean–variance optimization is a common approach in financial decision making e.g., (Luenberger, 1997), especially for static (one-stage) problems. Consider, for example, a fund manager who is interested in the 1-year

* Corresponding author. Tel.: +972 4 8293284; fax: +972 4 8295757.
E-mail address: shie@ee.technion.ac.il (S. Mannor).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.06.019

performance of the fund whose investment strategies will be judged according to the mean and variance of the return. Assuming that the manager is allowed to undertake periodic re-balancing actions in the course of the year, one obtains a Markov decision process with mean–variance criteria, and it is important to know the least possible variance achievable under a set target for the mean return. While the applicability of the financial strategies arising from mean–variance optimization in multi-period fund management can be debated (due to the "irrational" aspects mentioned above), mean–variance optimization is definitely a meaningful objective in various engineering contexts. Consider, for example, an engineering process whereby a certain material is deposited on a surface. Suppose that the primary objective is to maximize the amount deposited, but that there is also an interest in having all manufactured components be similar to each other; this secondary objective can be addressed by keeping the variance of the amount deposited small. In general, the applicability of the formulations studied in this paper will depend on the specifics of a particular application.

Mean–variance optimization problems resembling ours have been studied in the literature. For example, (Guo, Ye, & Yin, 2012) consider a mean–variance optimization problem, but subject to a constraint on the vector of expected rewards starting from each state, which results in a simpler problem, amenable to a policy iteration approach. Collins (1997) provides an apparently exponential-time algorithm for a variant of our problem, and Tamar, Di-Castro, and Mannor (2012) propose a policy gradient approach that aims at a locally optimal solution. Expressions for the variance of the discounted reward for stationary policies were developed in Sobel (1982). However, these expressions are quadratic in the underlying transition probabilities, and do not lead to convex optimization problems. Similarly, much of the earlier literature (see Kawai (1987), Huang & Kallenberg (1994) for a unified approach) on the problem provides various mathematical programming formulations. In general, these formulations either deal with problems that differ qualitatively focusing on the variation of reward from its average (Filar, Kallenberg, & Lee, 1989; White, 1992) from ours or are nonconvex, and therefore do not address the issue of polynomial-time solvability which is our focus. Indeed, we are not aware on any complexity results on mean–variance optimization problems. We finally note some interesting variance bounds obtained by Arlotto, Gans, and Steel (2013).

Motivated by considerations such as the above, this paper deals with the computational complexity aspects of mean–variance optimization. The problem is not straightforward for various reasons. One is the absence of a principle of optimality that could lead to simple recursive algorithms. Another reason is that, as is evident from the formula $\mathrm{var}(W) = \mathbb{E}[W^2] - (\mathbb{E}[W])^2$, the variance is not a linear function of the probability measure of the underlying process. Nevertheless, $\mathbb{E}[W^2]$ and $\mathbb{E}[W]$ are linear functions, and as such can be addressed simultaneously using methods from multicriteria or constrained Markov decision processes (Altman, 1999). Indeed, we will use such an approach in order to develop pseudopolynomial exact or approximation algorithms. On the other hand, we will also obtain various NP-hardness results, which show that there is little hope for significant improvement of our algorithms.

The rest of the paper is organized as follows. In Section 2, we describe the model and our notation. **We also define various classes of policies and performance objectives of interest. In Section 3, we compare different policy classes and show that performance typically improves strictly as more general policies are allowed. In Section 4, we establish NP-hardness results for the policy classes we have introduced. Then, in Sections 5 and 6, we develop exact and approximate pseudopolynomial time algorithms. Unfortunately, such algorithms do not seem possible for some of the more restricted classes of policies, due to strong NP-completeness results established in Section 4. Finally, Section 7 contains some brief concluding remarks.

2. The model

In this section, we define the model, notation, and performance objectives that we will be studying. Throughout, we focus on finite horizon problems.[1]

2.1. Markov decision processes

We consider a Markov decision process (MDP) with finite state, action, and reward spaces. An MDP is formally defined by a sextuple $\mathcal{M} = (T, \mathcal{S}, \mathcal{A}, \mathcal{R}, p, g)$ where:

(a) T, a positive integer, is the time horizon;
(b) $\mathcal{S}$ is a finite collection of states, one of which is designated as the initial state;
(c) $\mathcal{A}$ is a collection of finite sets of possible actions, one set for each state;
(d) $\mathcal{R}$ is a finite subset of $\mathbb{Q}$ (the set of rational numbers), and is the set of possible values of the immediate rewards. We let $K = \max_{r\in\mathcal{R}}|r|$.
(e) $p : \{0,\ldots,T-1\} \times \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{Q}$ describes the transition probabilities. In particular, $p_t(s'|s,a)$ is the probability that the state at time $t+1$ is s', given that the state at time t is s, and that action a is chosen at time t.
(d) $g : \{0,\ldots,T-1\} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to \mathbb{Q}$ is a set of reward distributions. In particular, $g_t(r|s,a)$ is the probability that the immediate reward at time t is r, given that the state and action at time t is s and a, respectively.

With few exceptions (e.g., for the time horizon T), we use capital letters to denote random variables, and lower case letters to denote ordinary variables. The process starts at the designated initial state. At every stage $t = 0,1,\ldots,T-1$, the decision maker observes the current state S_t and chooses an action A_t. Then, an immediate reward R_t is obtained, distributed according to $g_t(\cdot|S_t,A_t)$, and the next state S_{t+1} is chosen, according to $p_t(\cdot|S_t,A_t)$. Note that we have assumed that the possible values of the immediate reward and the various probabilities are all rational numbers. This is in order to address the computational complexity of various problems within the standard framework of digital computation. Finally, we will use the notation $x_{0:t}$ to indicate the tuple $(x_0,\ldots,x_t)$.

2.2. Policies

We will use the symbol π to denote policies. Under a *deterministic policy* $\pi = (\mu_0,\ldots,\mu_{T-1})$, the action at each time t is determined according to a mapping μ_t whose argument is the history $H_t = (S_{0:t},A_{0:t-1},R_{0:t-1})$ of the process, by letting $A_t = \mu_t(H_t)$. We let Π_h be the set of all such history-based policies. (The subscripts are used as a mnemonic for the variables on which the action is allowed to depend.) We will also consider *randomized* policies. Intuitively, at each point in time, the policy can pick an action at random, with the probability of each action determined by the current information (which is H_t as well as the outcomes of earlier randomizations). Randomness can always be simulated by using an independent uniform random variable as the seed, which leads to the following formal definition. We assume that there is

[1] Negative complexity results are straightforward to extend to the more general case of infinite horizon problems. Also, some of the positive results, such as the approximation algorithms of Section 6, can be extended to the infinite horizon discounted case; this is beyond the scope of this paper.