A SOLUTION FOR THE VARIANCE-PENALIZED MARKOV DECISION PROBLEM BASED ON PARAMETRIC LINEAR PROGRAMMING

Lodewijk C.M. Kallenberg, Leiden

Abstract: Considered is the Markov decision process with finite state and action spaces and with average expected reward, appropriately modified to include a penalty for the variability in the stream of rewards. It is shown that a pure and stationary policy maximizes this criterion. A parametric linear program is formulated, from which the optimal policy can be obtained.

1. Introduction

In /1/ the variance-penalized Markov decision problem has been introduced. For more background and relations with other work we refer to the introduction of /1/. In this paper we present a solution procedure for the variance-penalized Markov decision problem based on parametric linear programming. This approach is also related to the papers /2/ and /3/. In /4/ we will extend these results to a unifying framework for Markov decision problems with mean-variance trade-offs.

A discrete Markov decision process with finite state and action spaces is observed at discrete time points \( t = 1, 2, \ldots \). The state space is denoted by \( E = \{1, 2, \ldots, N\} \), and \( A(i) \) is the action space in state \( i \), \( i \in E \).

At any time point \( t \) the system is in one of the states and an action has to be chosen by the decision maker. If the system is in state \( i \) and action \( a \) is chosen, then an immediate reward \( r_{ia} \) is earned and the process moves to a state \( j \in E \) with transition probability \( p_{iaj} \), where \( p_{iaj} \geq 0 \) and \( \Sigma_j p_{iaj} = 1 \).

A policy \( R \) is a sequence of decision rules, one rule for each time point. If the rules are identical and nonrandomized the policy is called pure and deterministic. Such a policy is denoted by a vector \( f \), where \( f(i) \) is the action chosen in state \( i \).

Let \( X_t \) be the state at time \( t \) and \( Y_t \) be the action at time \( t \). Denote by \( P_R(X_t = j, Y_t = a | X_1 = i) \) the conditional probability that at time \( t \) the state is \( j \) and the action taken is \( a \), given that the initial state is \( i \) and the decision maker uses policy \( R \).
Let \( \beta = (\beta_1, \beta_2, \ldots, \beta_N) \) be a given initial distribution, i.e. \( \beta_i \) is the probability that \( X_1 = i \), where \( \beta_i \geq 0 \) for all \( i \in E \) and \( \Sigma_i \beta_i = 1 \).

For any policy \( R \) and initial distribution \( \beta \), we define the average expected reward over the infinite horizon, shortly the \( \beta \)-average reward for policy \( R \), by

\[
\phi(\beta, R) := \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \beta_i \cdot E_R(r_{X_t} \gamma_t | X_1 = i)
\]

From (1) it follows that

\[
\phi(\beta, R) = \liminf_{T \to \infty} \sum_{j,a} x_{ja}^T(R) r_{ja},
\]

where \( x_{ja}^T(R) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \beta_i \cdot P(R(X_t = j, Y_t = a | X_1 = i), a \in A(j), j \in E, t \in N \).

We shall say that \( R^* \) is a \( \beta \)-average optimal policy if \( \phi(\beta, R^*) \geq \phi(\beta, R) \) for every policy \( R \). It is well-known that a \( \beta \)-average optimal pure and deterministic policy exists /5/.

Let \( X(R) \) denote the set of vector-limit points of the sequence \( (x^T(R), t \in N) \). Define \( C := \{ R | X(R) \text{ is a singleton} \} \). It is known /6/ that \( C \) contains the stationary policies. If we denote by \( x(R) \) the unique element of \( X(R) \) for any \( R \in C \), then from (2) it follows

\[
\phi(\beta, R) = \sum_{j,a} x_{ja}(R) r_{ja}
\]

The long-run variance of a policy \( R \) will be defined by

\[
V(\beta, R) := \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \beta_i \cdot E_R([r_{X_t} \gamma_t - \phi(\beta, R)]^2 | X_1 = i)
\]

For \( R \in C \), we obtain

\[
V(\beta, R) = \sum_{j,a} x_{ja}(R)r_{ja}^2 - [\sum_{j,a} x_{ja}(R)r_{ja}]^2
\]

The variance-penalized Markov decision problem is the problem of maximizing

\[
\psi_*(\beta, R) := \phi(\beta, R) - \lambda \cdot V(\beta, R),
\]

where \( \lambda \geq 0 \) is the penalty for the variance and the maximization is taken over the policies of \( C \). Hence, for \( R \in C \), we have

\[
\psi_*(\beta, R) = \sum_{j,a} x_{ja}(R)[r_{ja}^2 - \lambda r_{ja}^2] + \lambda \cdot [\sum_{j,a} x_{ja}(R)r_{ja}]^2
\]

A policy \( R^* \in C \) is called variance-penalized optimal if \( \psi_*(\beta, R^*) \geq \psi_*(\beta, R) \) for all \( R \in C \).

**Assumption:** \( r_{ia} \geq 0 \) for all \( a \in A(i) \) and \( i \in E \).

This assumption is without loss of generality; if the rewards \( r_{ia} \) are replaced by \( r_{ia} - c \), where \( c := \min_{j,b} r_{jb} \), then the rewards are nonnegative, \( V(\beta, R) \) is unchanged and \( \phi(\beta, R) \) is replaced by \( \phi(\beta, R) - c \).