

Computational Complexity in Markov Decision Theory

John N. Tsitsiklis

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

Cambridge, MA 02139, U.S.A.

jnt@mit.edu

1. Introduction.

Markov Decision Processes are the standard formulation of the problem of sequential decision making in an uncertain environment. As such, they arise in a multitude of contexts, from economics to engineering. We provide an overview of the current theoretical understanding of the computational issues that they present, with an emphasis on discrete (finite-state, finite-action, discrete-time) formulations. We discuss both perfectly and imperfectly observed problems, provide a few comments on continuous counterparts, and conclude with a discussion of some nonstandard formulations, such as robust control, and mean-variance criteria.

2. Perfectly observed MDPs.

A (finite and stationary) Markov Decision Problem is specified by a finite state space, a finite action space, a (possibly infinite) time horizon, a discount factor $\alpha \in [0, 1]$, transition probabilities $p_{ij}(u)$, and one-stage costs $g(i, u)$. At a typical time t , the state is equal to some i , an action u is applied, and the cost $g(i, u)$ is incurred. The next state is chosen at random and is equal to j with probability $p_{ij}(u)$. The objective is to find a *policy* (a possibly time-dependent mapping from states into actions) that minimizes a criterion such as expected (possibly discounted) total or expected average cost per stage.

This problem is well-known to be polynomial-time solvable, through a reduction to linear programming. However, there are two major outstanding open problems: whether a strongly polynomial algorithm is possible, and whether the classical *policy iteration* algorithm runs in polynomial time.

Even though MDPs can be solved in time which increases polynomially in the number of states, many problems of practical interest involve a very large number of states, while the problem data (e.g., the transition probabilities) are succinctly described, in terms of a small number of parameters. (Prominent examples here are the multi-armed bandit problem, and many problems in the control of queueing networks.) An important question is whether such problems can be solved in time polynomial in the size of the problem description (e.g., the number of parameters). It turns out that the multi-armed bandit problem is a rare case of a polynomial-time solvable problem, whereas standard queueing network problems have provably exponential complexity.

3. Problems with continuous state spaces.

When the state space is continuous (as opposed to finite), any discussion of computational complexity needs to be prefaced by the specification of an appropriate model of

computation. In one model, one assumes that the problem data are given by an oracle that can provide, at request, information on the value of various input functions, at specified points. A lower bound on the problem's complexity (number of steps required to provide an answer within some desired accuracy) can then be obtained by lower bounding the number of necessary queries. For several classes of continuous MDPs, such lower bounds turn out to be tight, and indicate an exponential complexity increase with the dimension of the state space.

Another interesting point of contact between finite MDPs and continuous problems arises through the important observation that certain natural ways of discretizing *deterministic* continuous (optimal control) problems results in computational problems with the structure of (stochastic) finite-state MDPs. This analogy between optimal control and dynamic programming has proved fruitful, resulting in the development of efficient numerical methods for certain classes of Hamilton-Jacobi equations ("fast marching").

4. Imperfectly observed problems

Unfortunately, the control of partially observable MDPs (POMDPs) is a rather difficult problem, even when the time horizon is finite (PSPACE-complete). Even the problem of open-loop control of MDPs (the case of no observations) is NP-hard.

5. Nonstandard criteria

Much recent activity has centered on the study of MDPs under nonstandard formulations. In a *robust* formulation, rather than starting with an exact model of the underlying MDP, one assumes that the true model lies in a given family of models (in the simplest case, there is a choice between two alternative transition probability matrices), and tries to optimize a suitable criterion (e.g., the worst-case or the average performance over all possible models). Such robust formulations are special cases of POMDPs; one might hope that they are simple enough special cases to admit efficient solutions. However, most formulations of this type are NP-hard or worse, with only a few exceptions that can be reformulated as Markov games and to which dynamic programming algorithms can be applied.

Another choice of "risk-sensitive" performance criteria involves mean-variance trade-offs (e.g., minimizing the expected cost, subject to a constraint on the variance of the incurred cost). Unfortunately, most formulations of this type are also NP-hard.