Synchronous and Asynchronous Stochastic Control* to a Terminal State

by

Dimitri P. Bertsekas and John N. Tsitsiklis**

Abstract

We consider a classical Markovian decision problem where for each node of a graph, we must choose a probability distribution over the set of successor nodes so as to reach a certain destination node with minimum expected cost. The costs of transition between successive nodes can be positive as well as negative. We prove natural generalizations of the standard results for the deterministic shortest path problem, and we extend the corresponding theory for undiscounted finite state Markovian decision problems by removing the usual restriction that costs are either all nonnegative or all nonpositive. We also discuss various implementations of the successive approximation algorithm in a serial and a parallel asynchronous computational environment.

* Supported by the National Science Foundation under Grant DDM-8903385, the Army Research Office under Grant DAAL03-86-K-0171, and a Presidential Young Investigator Award to the second author.

** Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Cambridge, Mass. 02139.

Summary

Given a directed graph with nodes $1, 2, \ldots, n$ and with a length (or cost) assigned to each arc, the (deterministic) shortest path problem is to select at each node $j \neq 1$, a successor node $\mu(j)$ so that $(j, \mu(j))$ is an arc, and the path formed by a sequence of successor nodes starting at any node i terminates at node 1 and has minimum length (i.e. minimum sum of arc lengths), over all paths that start at i and terminate at 1.

The stochastic shortest path problem is a generalization whereby at each node, we must select a probability distribution over all possible successor nodes, out of a given set of probability distributions. For a given selection of distributions and for a given origin node, the path traversed as well as its length are now random, but we wish that the path leads to node 1 with probability one and has minimum expected length. Note that if every feasible probability distribution assigns probability one to a single successor node, we obtain the deterministic shortest path problem.

It is possible to analyze the stochastic shortest path problem by using the general theory of Markovian decision problems [2], [3], [6], [8], [9], [12]. This theory, however, applies only when the arc costs are either all nonnegative or all nonpositive (corresponding to the classical positive and negative dynamic programming models [5], [10]). On the other hand, the existing theory of the (deterministic) shortest path problem allows arc lengths that can be negative as well as positive. As a result, an analysis of the stochastic shortest path problem that generalizes the known results of its deterministic counterpart cannot be inferred from Markovian decision theory, and is not available at present. The purpose of this paper is to provide such an analysis. In particular, we allow arc lengths that are negative as well as positive.

In our analysis, we require a condition that generalizes the positive cycle condition for the deterministic shortest path problem (every cycle must have positive length). We also require that the available probability distributions at each state satisfy a connectivity condition analogous to the one for the deterministic shortest path problem (every node is connected to the destination node 1 with a path). These conditions are formulated using the notion of a *proper stationary policy*, that is, a policy that leads to node 1 with probability one, regardless of the initial node. The results that we prove are as strong as those for discounted Markovian decision problems. In particular, we show that:

(a) The optimal cost vector is the unique solution of Bellman's equation.

(b) The successive approximation method converges to the optimal cost vector for an arbitrary starting vector.

(c) The policy iteration algorithm yields an optimal stationary policy starting from an arbitrary proper policy.

Despite the strength of our results, our assumptions do not imply that the corresponding dynamic programming mapping is a contraction (unlike the situation in discounted problems), unless all policies are proper.

To put the contribution of the present paper in perspective, we provide a survey of earlier work. Our problem was first formulated by Eaton and Zadeh [7] who called it a problem of *pursuit*. They were motivated by a problem of intercepting in minimum expected time a target that moves randomly among a finite number of states. They showed how to formulate such a problem as one with a stationary target (i.e., a destination in a shortest path context) by viewing as state the pair of pursuer and target positions. Eaton and Zadeh [7] introduced the notion of a proper policy and assumed that at each state except the destination, the one–stage expected cost is positive, and the set of controls is finite. Within this context, they showed the results (a), (b), and (c) outlined above. The analysis of Eaton and Zadeh was replicated and streamlined in the text by Pallu de la Barriere [9], and in the text by Derman [6], who refers to the problem as the *first passage* problem. Derman remarks that the finite–horizon, finite–state Markovian decision problem is a special case. Veinott [11] shows that the dynamic programming mapping is a contraction if all stationary policies are proper. Kushner [8] improves on the results of Eaton and Zadeh by allowing the set of controls at each state to be infinite while imposing a compactness assumption, essentially our Assumption 2 of the next section. Kushner [8] also analyzes problems in which the state space is countable and illustrates some of the associated pathologies. Whittle [12] considers related problems under the name *transient programming*. Whittle investigates cases involving infinite state and control spaces under uniform boundedness conditions on the expected termination time; his results have the same flavor as the contraction result of Veinott [11]. The text by the first author [2] strengthens the earlier finite–state, finite–control results by weakening the positive cost assumption; costs are instead assumed nonnegative, and existence of an optimal proper policy is assumed, rather than implied by the positivity of the costs.

One main result of the present paper dispenses with the cost nonnegativity assumption, assuming instead that all improper policies yield a cost of $+\infty$ for some initial state, and establishing a stronger connection with the theory of deterministic shortest path problems. Furthermore, we allow the set of controls at each state to be infinite; this introduces substantial technical complications.

A second result relates to the computation of optimal policies in a parallel asynchronous setting. We show that such policies can be computed by parallel asynchronous Dynamic Programming starting from arbitrary initial conditions. We relate this results to the theory of real-time learning and

control as recently discussed by Barto et. al. [1]

References

[1] Barto, A. G., Bradtke, S. J., and Singh, S. P., "Real Time Learning and Control using Asynchronous Dynamic Programming," Computer and Information Science Tech. Report 91-57, Univ. of Massachusetts at Amherst, Aug. 1991.

[2] Bertsekas, D. P., *Dynamic Programming: Deterministic and Stochastic Models*, Prentice–Hall, Englewood Cliffs, N.Y., 1987.

[3] Bertsekas, D. P., and Shreve, S. E., *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, N.Y., 1978.

[4] Bertsekas, D. P., and Tsitsiklis, J. N., *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, N.Y., 1989.

[5] Blackwell, D., "Positive Dynamic Programming", *Proc. of 5th Berkeley Symp. Math., Statist., and Probability*, Vol. 1, 1965, pp. 415–418.

[6] Derman, C., *Finite State Markovian Decision Processes*, Academic Press, N.Y., 1970.

[7] Eaton, J. H., and Zadeh, L. A., "Optimal Pursuit Strategies in Discrete State Probabilistic Systems", *Trans. ASME Ser. D, J. Basic Eng.*, Vol. 84, 1962, pp. 23–29.

[8] Kushner, H., *Introduction to Stochastic Control*, Holt, Rinehart, and Winston, N.Y., 1971.

[9] Pallu de la Barriere, R., *Optimal Control Theory*, Saunders, Phila., 1967.

[10] Strauch, R., "Negative Dynamic Programming", *Ann. Math. Statistics*, Vol. 37, 1966, pp. 871–890.

[11] Veinott, A. F., Jr., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria", *Ann. Math. Statistics*, Vol. 40, 1969, pp. 1635–1660.

[12] Whittle, P., *Optimization over Time*, Wiley, N.Y., Vol. 2, 1983.