MATHEMATICS OF OPERATIONS RESEARCH
Vol 16, No 3, August 1991
Printed in U S A

AN ANALYSIS OF STOCHASTIC SHORTEST PATH PROBLEMS*†

DIMITRI P. BERTSEKAS AND JOHN N. TSITSIKLIS

We consider a stochastic version of the classical shortest path problem whereby 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.

1. Introduction. 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 [1], [2], [6], [10], [16]. This theory, however, applies only when the arc costs are either nonnegative or all nonpositive (corresponding to the classical positive and negative dynamic programming models [4], [13]). 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

*Received October 14, 1988; revised February 12, 1990.
AMS 1980 subject classification. Primary: 90C47.
IAOR 1973 subject classification. Main: Programming: Markov Decision.
OR/MS Index 1978 subject classification. Primary: 117 Dynamic Programming/Markov.
Key words. Shortest path, dynamic programming, Markovian decision problems, first passage, policy iteration.
†Supported by the National Science Foundation under Grant NSF-ECS-8519058, the Army Research Office under Grant DAAL03-86-K-0171, and a Presidential Young Investigator Award to the second author with matching funds from IBM, Inc. and Du Pont, Inc.

0364-765X/91/1603/0580/$01.25

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 [8] 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 [8] 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 [11], 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 [15] shows that the dynamic programming mapping is a contraction if all stationary policies are proper (see the remark preceding our Proposition 1 in §3). Kushner [10] 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 [10] also analyzes problems in which the state space is countable and illustrates some of the associated pathologies. Whittle [16] considers related problems under the name *transient programming*. (With three different names already introduced for the same problem, we feel only slightly guilty for introducing the fourth name "stochastic shortest path".) 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 [15]. The text by the first author [1] 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 as in Proposition 3 of §3, rather than implied by the positivity of the costs. Our main result of the present paper (Proposition 2 in §3) 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 somewhat simpler version of our result, where the set of controls at each state is assumed finite, is included in our recent text [3].

There is a class of interesting problems that is closely connected with the stochastic shortest path problem, but is not covered by our results. This is the class of optimal stopping problems investigated by Dynkin [7], and Grigelionis and Shiryaev [9], and further considered in several texts [6], [10], [12], [16]. Here, a state-dependent cost is

incurred only when invoking a stopping action that drives the system to the destination; all costs are zero prior to stopping. Eventual stopping is a requirement here, so to formulate such a stopping problem as a Markovian decision problem, it is necessary to make the stopping costs negative (by adding a negative constant to all stopping costs if necessary), providing an incentive to stop. We then come under the framework of this paper but with Assumption 1 of the next section violated because the improper policy that never stops, while nonoptimal, does not yield infinite cost for any starting state. Unfortunately, this seemingly small relaxation of our assumptions invalidates our results as will be shown in §3 (cf. the example of Figure 3).

2. Problem formulation. We have a discrete-time dynamic system with n states denoted $1, 2, \ldots, n$. At each state i, we are given a set of decisions or controls $U(i)$. If the state is i and control u is chosen at some time, the cost incurred is $c_i(u)$; the system then moves to state j with given probability $p_{ij}(u)$. In specific applications, it may be natural to model the cost of a transition from state i to state j as a scalar $a_{ij}(u)$ that also depends on j (cf. the stochastic shortest path problem context discussed in the previous section). We may reduce this case to the case discussed here by viewing $c_i(u)$ as an expected cost, given by $c_i(u) = \sum_{j=1}^{n} p_{ij}(u)a_{ij}(u)$.

Consider the set of functions

$$M = \{\mu \mid \mu(i) \in U(i), i = 1, \ldots, n\}.$$

A sequence $\{\mu_0, \mu_1, \ldots\}$ with $\mu_t \in M$ for all t, is called a *policy*, and if μ_t is the same for all t, it is called a *stationary policy*.

Let $P(\mu)$ be the transition probability matrix corresponding to $\mu \in M$, that is, the matrix with elements

$$[P(\mu)]_{ij} = p_{ij}(\mu(i)), \qquad i, j = 1, \ldots, n.$$

Let also

$$c(\mu) = \begin{pmatrix} c_1(\mu(1)) \\ \vdots \\ c_n(\mu(n)) \end{pmatrix}.$$

For any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$\Pr(\text{State is } j \text{ at time } t \mid \text{Initial state is } i, \text{ and } \pi \text{ is used})$$

$$= [P(\mu_0)P(\mu_1) \cdots P(\mu_{t-1})]_{ij}.$$

Therefore, if $x_i(\pi)$ is the expected cost corresponding to initial state i and policy $\pi = \{\mu_0, \mu_1, \ldots\}$, and $x(\pi)$ is the vector with coordinates $x_1(\pi), \ldots, x_n(\pi)$, we have

$$x(\pi) = \sum_{t=0}^{\infty} [P(\mu_0)P(\mu_1) \cdots P(\mu_{t-1})]c(\mu_t), \tag{1}$$

assuming the above series converges. When the above series is not known to converge, we use the definition

$$x(\pi) = \liminf_{k \to \infty} \sum_{t=0}^{k} [P(\mu_0)P(\mu_1) \cdots P(\mu_{t-1})]c(\mu_t), \tag{2}$$

where the lim inf is taken separately for each coordinate sequence. For a stationary policy $\{\mu, \mu, \ldots\}$, the corresponding cost vector is denoted by $x(\mu)$. We define the optimal expected cost starting at state i as

$$x_i^* = \inf_{\pi} x_i(\pi), \quad \forall\, i. \tag{3}$$

We say that the policy π^* is *optimal* if

$$x_i(\pi^*) = \inf_{\pi} x_i(\pi), \quad \forall\, i. \tag{4}$$

It is convenient to introduce the mappings $T_\mu\colon \Re^n \mapsto \Re^n$ and $T\colon \Re^n \mapsto \Re^n$ defined by

$$T_\mu(x) = c(\mu) + P(\mu)x, \qquad \mu \in M, \tag{5}$$

$$T(x) = \inf_{\mu \in M} [c(\mu) + P(\mu)x] = \inf_{\mu \in M} T_\mu(x). \tag{6}$$

A straightforward calculation verifies that for all $t \geqslant 1$, μ, and x, the t-fold composition of the mapping T_μ is given by

$$T_\mu^t(x) = P(\mu)^t x + \sum_{k=0}^{t-1} P(\mu)^k c(\mu). \tag{7}$$

Furthermore, for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$x(\pi) = \lim_{t \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_t})(x^0),$$

where $x^0 = (0, \ldots, 0)$ is the zero vector.

We note two basic properties of T and T_μ. The first is that T and T_μ are *monotone* in the sense

$$x \leqslant x' \quad \Rightarrow \quad T(x) \leqslant T(x'), \tag{8}$$

$$x \leqslant x' \quad \Rightarrow \quad T_\mu(x) \leqslant T_\mu(x'), \quad \forall\, \mu \in M, \tag{9}$$

where the above inequalities are meant to hold separately for each coordinate, that is, we write $x \leqslant y$ if $x_i \leqslant y_i$ for all i. The second basic property is that for all $x \in \Re^n$, scalars r, integers $t > 0$, and functions $\mu, \mu_1, \ldots, \mu_t \in M$, we have

$$T^t(x + re) = T^t(x) + re, \quad T_\mu^t(x + re) = T_\mu^t(x) + re, \tag{10}$$

$$(T_{\mu_1} T_{\mu_2} \cdots T_{\mu_t})(x + re) = (T_{\mu_1} T_{\mu_2} \cdots T_{\mu_t})(x) + re, \tag{11}$$

where e is the vector $(1, 1, \ldots, 1)$.

We say that a stationary policy $\{\mu, \mu, \ldots\}$ is *proper* if $\lim_{t \to \infty} [P(\mu)^t]_{i1} = 1$ for all i; otherwise, we say that the policy is *improper*. We refer to a function $\mu \in M$ as proper or improper, if the corresponding stationary policy is proper or improper, respectively. We introduce the following assumptions:

Assumption 1. State 1 is absorbing and cost-free, that is, $p_{11}(u) = 1$, and $c_1(u) = 0$ for all $u \in U(1)$. Furthermore, there exists at least one proper stationary policy, and

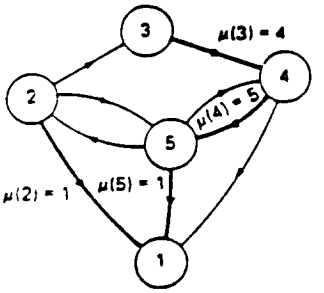

FIGURE 1. Viewing a (deterministic) shortest path problem as a special case of the problem of §2. The figure shows the paths corresponding to a stationary policy $\{\mu, \mu, \ldots\}$. The control $\mu(i)$ associated with a node $i \neq 1$ is a successor node of i. The policy shown is proper because the path from every state leads to the destination.

each improper stationary policy yields infinite cost for at least one initial state, that is, for each improper $\mu \in M$, there is a state i such that $\lim_{k\to\infty}[\sum_{t=0}^{k} P(\mu)^t c(\mu)]_i = \infty$.

Assumption 2. For all states i, the set $U(i)$ is a compact subset of a metric space, the function $c_i(\cdot)$ is lower-semicontinuous over $U(i)$, and the functions $p_{ij}(\cdot)$, $j = 1, \ldots, n$, are continuous over $U(i)$. (This is true in particular if $U(i)$ is a finite set.)

An important implication of Assumption 2 is that the infimum over $u \in U(i)$ in the definition (6) of the mapping T is attained. Otherwise stated, for every x in the n-dimensional Euclidean space $\Re^n$, there exists a $\mu \in M$ such that $T_\mu(x) = T(x)$.

The deterministic shortest path problem is an important example of a dynamic programming problem where the above assumptions are natural. Here we are given a directed graph with nodes $1, 2, \ldots, n$, and a length a_{ij} for each arc (i, j). We view the nodes as the states of a Markov chain. The stationary policies $\{\mu, \mu, \ldots\}$ correspond to assigning to each node $i \neq 1$ a (single) neighbor node $\mu(i) = j$; the cost of the transition is the length $a_{i\mu(i)}$ of arc $(i, \mu(i))$ (see Figure 1). Node 1 is the destination and is viewed as an absorbing and cost-free state for all μ. It is seen that the usual *connectivity assumption* (there is a path connecting every node with node 1) is equivalent to the existence of at least one proper policy. An improper policy, by definition, is one for which there exists an initial state $i \neq 1$ such that starting at i, the sequence of generated successor nodes does not contain node 1 and cycles indefinitely. The associated cost starting from i is infinite if and only if the corresponding cycle has positive length. It can be seen therefore that Assumption 1 is satisfied if and only if the connectivity assumption together with the *positive cycle assumption* (every cycle not containing node 1 has positive length) hold. Consider now a proper policy. For every starting node, the generated path of successor nodes eventually reaches node 1 and terminates there, in which case, the total cost is equal to the length of the path. It follows that under Assumption 1, optimal stationary policies are those that correspond to shortest paths.

3. Main results. Since the cost associated with the absorbing state 1 is zero, all operations of interest for our problem take place in the subspace $X = \{x \in \Re^n | x_1 = 0\}$. In particular under Assumption 1, T and T_μ map X into itself as can be seen from the definitions (5) and (6). An important property of a proper policy $\{\mu, \mu, \ldots\}$ is that the associated mapping T_μ is a contraction mapping over X with respect to some

weighted maximum norm, that is, a norm of the form

$$\|x\|_\infty^w = \max_{i=1,\ldots,n} \frac{|x_i|}{w_i},$$

where $w \in \Re^n$ is a vector with positive coordinates. This contraction property follows from classical results on nonnegative matrices, and is also a special case of the following proposition, which shows that T is a contraction when all stationary policies are proper. The proof of this proposition is given by Veinott [15, Lemma 3] and is attributed to A. J. Hoffman. Another proof which is short and constructive was given recently by P. Tseng in [14]. These proofs assume that the set $U(i)$ is finite but can be extended for the case where Assumption 2 holds instead.

PROPOSITION 1. *Let Assumptions* 1 *and* 2 *hold, and assume that all stationary policies are proper. Then the mapping* T *of equation* (6) *is a contraction mapping over the subspace* $X = \{x \in \Re^n | x_1 = 0\}$ *with respect to some weighted maximum norm.*

Given a proper policy $\{\mu, \mu, \ldots\}$, we can consider the variation of the problem where $\{\mu, \mu, \ldots\}$ is the only available policy, i.e. $U(i) = \{\mu(i)\}$ for all i. Application of Proposition 1 then shows that T_μ is a contraction with respect to some weighted maximum norm. Given the contraction property of the mappings T and T_μ, one can infer a number of strong results regarding the associated Markovian decision problem, essentially those available for discounted problems. In particular, under the assumptions of Proposition 1 we have (see [5] and [2]):

(a) x^* and $x(\mu)$, $\mu \in M$, are the unique fixed points of T and T_μ, $\mu \in M$, respectively.

(b) We have $T^t(x) \to x^*$ and $T_\mu^t(x) \to x(\mu)$, $\mu \in M$, for all $x \in \Re^n$. Furthermore the rate of convergence is geometric.

If there exist some improper policies, the mapping T need not be a contraction over X with respect to any norm. This is shown by the example of Figure 2 for which Assumptions 1 and 2 hold. Despite this fact, the following proposition proves in effect the same results as when we have a contraction property. (The only essential consequence of the fact that T is not a contraction is that a geometric convergence rate for the sequence $\{T^t(x)\}$ cannot be shown.)

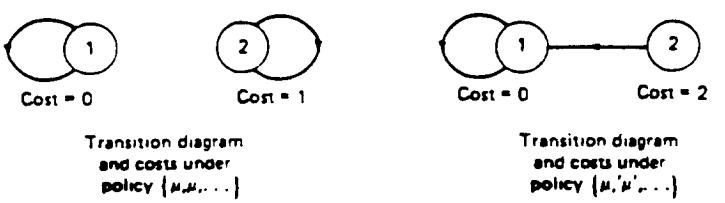

FIGURE 2. Example problem where Assumptions 1 and 2 are satisfied, but the mapping T is not a contraction mapping over the subspace $X = \{x \in \Re^2 | x_1 = 0\}$. Here $M = \{\mu, \mu'\}$ with transition probabilities and costs as shown. The mapping T over the set $X = \{(x_1, x_2) | x_1 = 0\}$ is given by

$$[T(x)]_1 = 0,$$

$$[T(x)]_2 = \min\{1 + x_2, 2\}.$$

Thus, for $x = (0, x_2)$ and $x' = (0, x_2')$ with $x_2 < 1$ and $x_2' < 1$ we have

$$|[T(x)]_2 - [T(x')]_2| = |(1 + x_2) - (1 + x_2')| = |x_2 - x_2'|.$$

Therefore, T is not a contraction mapping over X with respect to any norm.

PROPOSITION 2. *Let Assumptions* 1 *and* 2 *hold. Then*:

(a) *The optimal cost vector* x^* *is the unique fixed point of* T *within the subspace* $X = \{x \in \Re^n | x_1 = 0\}$.

(b) *For every* $x \in X$, *there holds*

$$\lim_{t \to \infty} T^t(x) = x^*. \tag{12}$$

(c) *A stationary policy* $\{\mu^*, \mu^*, \dots\}$ *is optimal if and only if*

$$T_{\mu^*}(x^*) = T(x^*). \tag{13}$$

Furthermore there exists an optimal (*stationary*) *proper policy.*

PROOF. The proof is based on the following three lemmas. Our proof of the third lemma is quite long and has been relegated to an appendix.

LEMMA 1. *Let Assumption* 1 *hold*:

(a) *If* μ *is proper, then* $x(\mu)$ *is the unique fixed point of* T_μ *within* X. *Furthermore,* $\lim_{t \to \infty} T_\mu^t(x) = x(\mu)$ *for all* $x \in X$.

(b) *If* $x \geqslant T_\mu(x)$ *for some* $x \in X$, *then* μ *is proper.*

PROOF. (a) By Proposition 1, T_μ is a contraction mapping over X when μ is proper. As remarked earlier, this implies the result. (b) If $x \in X$ and $x \geqslant T_\mu(x)$, then by the monotonicity of T_μ,

$$x \geqslant T_\mu^t(x) = P(\mu)^t x + \sum_{k=0}^{t-1} P(\mu)^k c(\mu), \quad \forall\, t \geqslant 1. \tag{14}$$

If $\{\mu, \mu, \dots\}$ were improper, then some subsequence of $\sum_{k=0}^{t-1} P(\mu)^k c(\mu)$ would have a coordinate that tends to infinity, thereby contradicting the above inequality. Q.E.D.

LEMMA 2. *Let Assumption* 1 *hold. The mapping* T *of equation* (6) *is continuous over* X.

PROOF. A standard calculation (e.g. [1, p. 185]) shows that

$$\|T(x) - T(x')\|_\infty \leqslant \|x - x'\|_\infty, \quad \forall\, x, x' \in \Re^n,$$

where $\|\cdot\|_\infty$ is the maximum norm in $\Re^n$ ($\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$). The continuity of T follows. Q.E.D.

LEMMA 3. *Let Assumptions* 1 *and* 2 *hold. Suppose that* $\{\mu^k\} \subset M$ *is a sequence such that each* μ^k *is proper and* $\mu^k \to \mu$ *for some* $\mu \in M$. *Then*:

(a) *If* μ *is proper, then* $\liminf_{k \to \infty} x(\mu^k) \geqslant x(\mu)$.

(b) *If* μ *is improper, there exists some* i *such that* $\{x_i(\mu^k)\}$ *is unbounded above.*

PROOF. See the Appendix.

We now return to the proof of Proposition 2. We first show that T has at most one fixed point within X. Indeed, if x and x' are two fixed points in X, then we select μ and μ' such that $x = T(x) = T_\mu(x)$ and $x' = T(x') = T_{\mu'}(x')$; this is possible because of Assumption 2. By Lemma 1(b), we have that μ and μ' are proper, and Lemma 1(a) implies that $x = x(\mu)$ and $x' = x(\mu')$. We have $x = T^t(x) \leqslant T_{\mu'}^t(x)$ for all $t \geqslant 1$, and by Lemma 1(a), we obtain $x \leqslant \lim_{t \to \infty} T_{\mu'}^t(x) = x(\mu') = x'$. Similarly, $x' \leqslant x$, showing that $x = x'$ and T has at most one fixed point within X.

We next show that T has a fixed point within X. Let $\{\mu, \mu, \ldots\}$ be a proper policy (there exists one by Assumption 1). Choose $\mu' \in M$ such that $T_{\mu'}(x(\mu)) = T(x(\mu))$. Then we have $x(\mu) = T_{\mu}(x(\mu)) \geqslant T_{\mu'}(x(\mu))$. By Lemma 1(b), μ' is proper, and by the monotonicity of $T_{\mu'}$ and Lemma 1(a), we obtain

$$x(\mu) \geqslant \lim_{t \to \infty} T_{\mu'}^t(x(\mu)) = x(\mu'). \tag{15}$$

Continuing in the same manner, we construct a sequence $\{\mu^k\} \subset M$ such that each μ^k is proper and

$$x(\mu^k) \geqslant T(x(\mu^k)) \geqslant x(\mu^{k+1}), \qquad \forall\, k = 0, 1, \ldots . \tag{16}$$

By the compactness of $U(i)$ (cf. Assumption 2), there is a subsequence $\{\mu^k\}_{k \in K}$ converging to some $\mu \in M$. From Lemma 3(b) and equation (16), it is seen that μ must be proper, and using Lemma 3(a), we obtain

$$\liminf_{k \to \infty,\, k \in K} x(\mu^k) \geqslant x(\mu).$$

Since by equation (16), $\{x(\mu^k)\}$ is monotonically nonincreasing, it follows that the entire sequence $\{x(\mu^k)\}$ converges to some x_∞ and we have

$$x_\infty \geqslant x(\mu). \tag{17}$$

Therefore x_∞ has finite coordinates, and by taking the limit in equation (16), and by using the continuity of T (cf. Lemma 2), we obtain $x_\infty = T(x_\infty)$. Thus, x_∞ is the unique fixed point of T within X. In fact we can show that this fixed point is equal to $x(\mu)$. Indeed, from equation (16) we have

$$T_\mu(x(\mu^k)) \geqslant T(x(\mu^k)) \geqslant x_\infty,$$

and by taking the limit we obtain $T_\mu(x_\infty) \geqslant x_\infty$. This implies that $x(\mu) = \lim_{t \to \infty} T_\mu^t(x_\infty) \geq x_\infty$, which combined with equation (17) yields $x(\mu) = x_\infty$.

Next we show that the unique fixed point of T within X is equal to the optimal cost vector x^*, and that $T^t(x) \to x^*$ for all $x \in X$. The construction of the preceding paragraph provides a proper μ such that $T(x(\mu)) = x(\mu)$. We will show that $T^t(x) \to x(\mu)$ for all $x \in X$, and that $x(\mu) = x^*$. Let Δ be the vector with coordinates

$$\Delta_i = \begin{cases} 0, & \text{if } i = 1 \\ \delta, & \text{if } i \neq 1, \end{cases} \tag{18}$$

where $\delta > 0$ is some scalar, and let x^Δ be the vector in X satisfying $T_\mu(x^\Delta) = x^\Delta - \Delta$. To see that there is a unique such vector, note that the equation $x^\Delta = c(\mu) + \Delta + P(\mu)x^\Delta$ $(= T_\mu(x^\Delta) + \Delta)$ has a unique solution within X because μ is proper, and thus the mapping on the right side of the equation is a contraction. Since x^Δ is the cost vector corresponding to μ for $c(\mu)$ replaced by $c(\mu) + \Delta$, we have $x^\Delta \geqslant x(\mu)$. We have

$$x(\mu) = T(x(\mu)) \leqslant T(x^\Delta) \leqslant T_\mu(x^\Delta) = x^\Delta - \Delta \leqslant x^\Delta.$$

Using the monotonicity of T and the previous relation, we obtain

$$x(\mu) = T^t(x(\mu)) \leqslant T^t(x^\Delta) \leqslant T^{t-1}(x^\Delta) \leqslant x^\Delta, \quad \forall\, t \geqslant 1.$$

Hence, $T^t(x^\Delta)$ converges to some $\tilde{x} \in X$, and by continuity of T (cf. Lemma 2), we must have $\tilde{x} = T(\tilde{x})$. By the uniqueness of the fixed point of T shown earlier, we must have $\tilde{x} = x(\mu)$. It is also seen that

$$x(\mu) - \Delta = T(x(\mu)) - \Delta \leqslant T(x(\mu) - \Delta) \leqslant T(x(\mu)) = x(\mu).$$

Thus, $T^t(x(\mu) - \Delta)$ is monotonically increasing and bounded above. Similarly, as earlier, it follows that $\lim_{t\to\infty} T^t(x(\mu) - \Delta) = x(\mu)$. For any $x \in X$, we can find $\delta > 0$ such that $x(\mu) - \Delta \leqslant x \leqslant x^\Delta$. By the monotonicity of T, we then have

$$T^t(x(\mu) - \Delta) \leqslant T^t(x) \leqslant T^t(x^\Delta), \quad \forall\, t \geqslant 1,$$

and since $\lim_{t\to\infty} T^t(x(\mu) - \Delta) = \lim_{t\to\infty} T^t(x^\Delta) = x(\mu)$, it follows that $\lim_{t\to\infty} T^t(x) = x(\mu)$. To show that $x(\mu) = x^*$, take any policy $\pi = \{\mu_0, \mu_1, \ldots\}$. We have

$$(T_{\mu_0} \cdots T_{\mu_{t-1}})(x^0) \geqslant T^t(x^0),$$

where x^0 is the zero vector. Taking the limit inferior in the preceding inequality, we obtain $x(\pi) \geqslant x(\mu)$, so $\{\mu, \mu, \ldots\}$ is an optimal policy and $x(\mu) = x^*$.

To prove part (c), we note that if $\{\mu^*, \mu^*, \ldots\}$ is optimal, then $x(\mu^*) = x^*$ and μ^* is proper, so $T_{\mu^*}(x^*) = T_{\mu^*}(x(\mu^*)) = x(\mu^*) = x^* = T(x^*)$. Conversely, if $x^* = T(x^*) = T_{\mu^*}(x^*)$, it follows from Lemma 1(b) that μ^* is proper, and by using Lemma 1(a), we obtain $x^* = x(\mu^*)$. Therefore, $\{\mu^*, \mu^*, \ldots\}$ is optimal. Because by Assumption 2, there exists $\mu^* \in M$ such that $T(x^*) = T_{\mu^*}(x^*)$, it follows that there exists an optimal proper policy. Q.E.D.

We now discuss the effects of relaxing some of our assumptions. The following example shows that the compactness Assumption 2 is essential for our results.

EXAMPLE 1. Consider a problem with two states, 1 and 2. State 1 is absorbing and cost-free. In state 2 we must choose a control u from the interval $(0, 1]$; we incur a cost $-u$, and then move to state 1 with probability u or stay in state 2 with probability $1 - u$. We note that Assumption 1 is satisfied and that all stationary policies are proper. However, Assumption 2 is violated. The mapping T takes the form

$$[T(x)]_1 = 0, \quad [T(x)]_2 = \inf_{u\in(0,1]} [-u + (1 - u)x_2].$$

It can be seen that we have $x = T(x)$ for all $x \in X$ with $x_2 \leqslant -1$. Hence Proposition 2 does not hold. Furthermore T cannot be a contraction mapping, so Proposition 1 does not hold.

The next example, given in Figure 3, illustrates the sensitivity of our results to seemingly minor changes in Assumption 1. In particular, here there is a (nonoptimal) improper policy that yields finite cost for all initial states (rather than infinite cost for some initial state), and T has multiple fixed points. This example depends on the presence of a negative cost $c_t(u)$. When the costs $c_t(u)$ are all nonnegative, similar results as the ones of Proposition 2 are known under the following version of Assumption 1, which allows improper policies with finite cost for all initial states:

Assumption 1′. State 1 is absorbing and cost-free and all costs $c_t(u)$ are nonnegative. Furthermore, there exists an optimal proper policy.

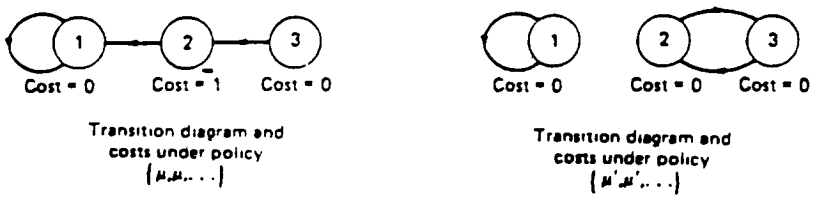

FIGURE 3. Example where Proposition 2 fails to hold when Assumption 1 is violated. This example is in effect a deterministic shortest path problem involving a cycle with zero length. There are two stationary policies, $\{\mu, \mu, \ldots\}$ and $\{\mu', \mu', \ldots\}$ with transition probabilities and costs as shown. The equation $x = T(x)$ over the subspace $X = \{(x_1, x_2, x_3) | x_1 = 0\}$ is given by

$$x_1 = 0,$$

$$x_2 = \min\{-1, x_3\},$$

$$x_3 = x_2,$$

and is satisfied by any vector of the form $x = (0, \delta, \delta)$ with $\delta \leqslant -1$. Here the proper policy $\{\mu, \mu, \ldots\}$ is optimal and the corresponding optimal cost vector is $x^* = (0, -1, 1)$ The difficulty is that there is the nonoptimal improper policy $\{\mu', \mu', \ldots\}$ that has finite (zero) cost for all initial states.

The following result is essentially given in [1, §6.4]:

PROPOSITION 3. *Let Assumptions 1′ and 2 hold. Then:*

(a) *The optimal cost vector x^* is the unique fixed point of T within the set $X^+ = \{x \in \Re^n | x \geqslant 0, x_1 = 0\}$.*

(b) *For every $x \in X^+$, there holds $\lim_{t \to \infty} T^t(x) = x^*$.*

(c) *$\{\mu^*, \mu^*, \ldots\}$ is optimal if and only if $T_{\mu^*}(x^*) = T(x^*)$.*

PROOF. Part (a) is given in [1, p. 256]. (Note that finiteness of $U(i)$ is assumed in [1], but the extension to the more general case where Assumption 2 holds is straightforward.) Part (c) is a standard result for undiscounted Markovian decision problems with $c_i(u) \geqslant 0$ ([1, p. 216]). Part (b) is shown as in the proof of Proposition 2, by establishing the inequality

$$T^t(x^0) \leqslant T^t(x) \leqslant T^t(x^\Delta), \quad \forall\, x \in X^+, t = 0, 1, \ldots,$$

where x^0 is the zero vector and x^Δ is a suitable vector of the form (18). Q.E.D.

It is essential to use X^+, rather then X in the results of Proposition 3. To see this, consider the example of Figure 3 but change the costs $c_i(u)$ so that they are now all equal to zero. Then, both the proper policy $\{\mu, \mu, \ldots\}$ and the improper $\{\mu', \mu', \ldots\}$ are optimal, so Assumption 1′ holds. However, it can be seen that all vectors of the form $x = (0, \delta, \delta)$ with $\delta \leqslant 0$ are fixed points of T. Note also that, in the example of Figure 3, we have $c_i(u) \leqslant 0$ for all i and $u \in U(i)$. Therefore an assumption analogous to Assumption 1′ with all costs nonpositive, instead of nonnegative, does not imply that x^* is the unique fixed point of T within the set $X^- = \{x \in \Re^n | x \leqslant 0, x_1 = 0\}$, even if all improper policies are nonoptimal.

We finally note that if the optimal cost vector x^* is known to be finite, the results of Proposition 2 can be shown with the compactness Assumption 2 replaced by the following weaker assumption:

Assumption 2′. The set $U(i)$ is a subset of some metric space and the set

$$\left\{ u \in U(i) \,\middle|\, c_i(u) + \sum_{j=1}^{n} p_{ij}(u) x_j \leqslant \alpha \right\} \tag{19}$$

is compact for all i, $x \in \Re^n$, and $\alpha \in \Re$. Furthermore, the function $c_i(\cdot)$ is lower-semicontinuous over $U(i)$, and the functions $p_{ij}(\cdot)$, $j = 1, \ldots n$, are continuous over $U(i)$.

Briefly, under Assumptions 1 and 2′ it can be seen that the infimum in the definition of $T(x)$ (cf. equation (6)) is attained, that Lemmas 1 and 2 hold, and that the proof of Lemma 3 also goes through. A sequence $\{\mu^k\}$ can be constructed as in the proof of Proposition 2 with $x(\mu^k)$ monotonically nonincreasing and converging to some x_∞, which is finite since $x(\mu^k) \geqslant x^*$. It can be seen that $\{\mu^k(i)\}$ lies in the set

$$\left\{ u \in U(i) \middle| c_i(u) + \sum_{j=1}^{n} p_{ij}(u)(x_\infty)_j \leqslant x_i(\mu^0) \right\},$$

which is compact by Assumption 2′. Therefore, we can extract a subsequence $\{\mu^k\}_{k \in K}$ converging to some proper μ. The remainder of the proof of Proposition 2 goes through with no changes.

4. Constructing an optimal proper policy. We now show how an optimal proper policy can be obtained through the successive approximation and policy iteration methods.

Consider first the successive approximation method. Here, starting from some $x \in X$ we compute $T^k(x)$ for $k = 1, 2, \ldots$, as well as functions $\mu^k \in M$ such that

$$T_{\mu^k}\left[T^{k-1}(x)\right] = T^k(x), \quad \forall\, k = 1, 2, \ldots$$

or equivalently

$$(20)\quad c_i(\mu^k(i)) + \sum_{j=1}^{n} p_{ij}(\mu^k(i))\left[T^{k-1}(x)\right]_j \leqslant c_i(u) + \sum_{j=1}^{n} p_{ij}(u)\left[T^{k-1}(x)\right]_j,$$

$$\forall\, k, i, u \in U(i).$$

(We can select such μ^k because of Assumption 2.) From Proposition 2(b) we have $\lim_{k \to \infty} T^k(x) = x^*$. Consider any subsequence $\{\mu^k\}_{k \in K}$ that converges to some $\mu \in M$; since $\mu^k(i) \in U(i)$ for all k and $U(i)$ is compact (by Assumption 2), there exists at least one such subsequence. By taking limit inferior in equation (20) as $k \to \infty$, $k \in K$, we see that

$$c_i(\mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) x_j^* \leqslant c_i(u) + \sum_{j=1}^{n} p_{ij}(u) x_j^*, \qquad \forall\, i, u \in U(i).$$

It follows that $T_\mu(x^*) = T(x^*) = x^*$, so by Proposition 2(c), the policy $\{\mu, \mu, \ldots\}$ obtained from the successive approximation method is optimal.

The policy iteration algorithm is based on the construction used in the proof of Proposition 2 to show that T has a fixed point. In the typical iteration of this algorithm, given a proper policy $\{\mu, \mu, \ldots\}$ and the corresponding cost vector $x(\mu)$, one obtains a new proper policy $\{\mu', \mu', \ldots\}$ satisfying the equation $T_{\mu'}(x(\mu)) = T(x(\mu))$, or, equivalently,

$$\mu'(i) = \arg \min_{u \in U(i)} \left[c_i(u) + \sum_{j=1}^{n} p_{ij}(u) x_j(\mu) \right], \qquad i = 2, 3, \ldots, n.$$

It was shown in equation (15) that $x(\mu') \leqslant x(\mu)$. It can be seen also that strict inequality $x_i(\mu') < x_i(\mu)$ holds for at least one state i, if μ is nonoptimal; otherwise we would have $x(\mu) = T(x(\mu))$ and by Proposition 2(b), μ would be optimal. Therefore, the new policy is strictly better if the current policy is nonoptimal. When the sets $U(i)$ are finite, the number of stationary policies is also finite, and it follows that the policy iteration algorithm terminates after a finite number of iterations with an optimal proper policy. Under the more general Assumption 2, the argument of the proof of Proposition 2 following equation (15) shows that if $\{\mu^k\}$ is the generated sequence by the policy iteration algorithm, then every limit point of $\{\mu^k\}$ is an optimal proper policy.

Appendix: Proof of Lemma 3. We introduce some terminology. We say that *state j is reachable from state i under a transition probability matrix P* if there exists a positive integer t such that the ijth element of the matrix P^t, denoted $[P^t]_{ij}$, is positive. This is equivalent to the existence of a sequence of positive probability transitions starting at i and ending at j. (In fact, it is seen that this sequence need not contain more than $n-1$ transitions.) A policy $\{\mu, \mu, \ldots\}$ is proper if and only if state 1 is reachable from all other states under $P(\mu)$ (or equivalently, if and only if $[P(\mu)^n]_{i1} > 0$ for all i).

A set of states S is said to be an *ergodic class* under a transition probability matrix if every state in S is reachable from every other state in S under this matrix and no state outside S is reachable from a state in S. States that do not belong to any ergodic class are referred to as *transient*. Note that, under Assumption 1, the set $\{1\}$ consisting of just state 1 is an ergodic class under all $P(\mu)$, $\mu \in M$. Furthermore, μ is improper if and only if under $P(\mu)$, there exists an ergodic class other than $\{1\}$.

For any proper policy μ, we have $x(\mu) = c(\mu) + P(\mu)x(\mu)$, from which, using the fact $x_1(\mu) = 0$, we obtain

$$(I - P(\mu) + E)x(\mu) = c(\mu), \tag{21}$$

where I is the identity, and E is the matrix with all elements in the first column equal to one and all other elements equal to zero. Since μ is proper, we have

$$E = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P(\mu)^k$$

and from a well-known result (e.g. [1, p. 337]), it follows that $I - P(\mu) + E$ is invertible. Therefore,

$$x(\mu) = (I - P(\mu) + E)^{-1} c(\mu) \tag{22}$$

for all proper μ. From this equation and the definition of $x(\mu)$ it also follows that

$$(I - P(\mu) + E)^{-1} c = \left(\sum_{t=0}^{\infty} P(\mu)^t \right) c \tag{23}$$

for all vectors c with $c_1 = 0$.

PROOF OF PART (a). Since $\mu^k \to \mu$ and $c_i(\cdot)$ is lower semicontinuous, we have for every $\epsilon > 0$

$$c_i(\mu^k(i)) \geqslant c_i(\mu(i)) - \epsilon$$

for all i and all k sufficiently large. Therefore,

$$\left(\sum_{t=0}^{\infty} P(\mu^k)^t\right)c(\mu^k) \geqslant \left(\sum_{t=0}^{\infty} P(\mu^k)^t\right)c(\mu) - \epsilon\left(\sum_{t=0}^{\infty} P(\mu^k)^t\right)e,$$

where $e = (0, 1, 1, \ldots, 1)'$. Using equations (22) and (23), we see that

$$\begin{aligned} x(\mu^k) &= \left(I - P(\mu^k) + E\right)^{-1} c(\mu^k) \\ &\geqslant \left(I - P(\mu^k) + E\right)^{-1} c(\mu) - \epsilon\left(I - P(\mu^k) + E\right)^{-1} e. \end{aligned}$$

Since μ is proper and $P(\cdot)$ is continuous, by taking the limit inferior as $k \to \infty$ and by using also equation (22), we obtain

$$\liminf_{k\to\infty} x(\mu^k) \geqslant x(\mu) - \epsilon\left(I - P(\mu) + E\right)^{-1} e.$$

Since ϵ is an arbitrary positive number, we can take the limit as $\epsilon \to 0$, and it follows that

$$\liminf_{k\to\infty} x(\mu^k) \geqslant x(\mu).$$

PROOF OF PART (b). Let S be the set consisting of state 1 together with the states from which state 1 is reachable under $P(\mu)$. Let $\bar{S}$ be the complementary set of states, $\bar{S} = \{i | i \notin S\}$. Since μ is improper, it is seen that $\bar{S}$ is the union of a nonempty subset $\bar{S}_E$ consisting of ergodic classes and a (possibly empty) subset $\bar{S}_T$ of transient states from which only states in $\bar{S}_E$ are reachable under $P(\mu)$. We claim that

$$x_i(\mu) = \infty, \quad \forall\, i \in \bar{S}. \tag{24}$$

To show this, assume the contrary, i.e. that there exists some $i \in \bar{S}$ with $x_i(\mu) < \infty$. Let

$$C = \max_i \left| c_i(\mu(i)) \right|.$$

We distinguish two cases:

(a) $i \in \bar{S}_E$: In this case, let E_i be the ergodic class of i. Then for all states $j \in E_i$, we have $x_j(\mu) < \infty$, since

$$x_j(\mu) \leqslant x_i(\mu) + CE\{T_{ji}\} < \infty,$$

where T_{ji} is the (random) number of transitions to reach i for the first time starting from j and $E\{T_{ji}\}$ is its expected value under $P(\mu)$. Consider now any proper $\bar{\mu}$ and let

$$\hat{\mu}(s) = \begin{cases} \mu(s), & \text{if } s \in E_i, \\ \bar{\mu}(s), & \text{otherwise.} \end{cases}$$

Then it is seen that there are two ergodic classes under $P(\hat{\mu})$, namely E_i and $\{1\}$, and since $x_s(\hat{\mu}) = x_s(\mu) < \infty$ for all $s \in E_i$, it follows that $x_s(\hat{\mu}) < \infty$ for all $s = 1, \ldots, n$. This contradicts Assumption 1 since $\hat{\mu}$ is improper.

(b) $i \in \bar{S}_T$: In this case, we have

$$x_i(\mu) \geqslant \min_{j \in \bar{S}_E} \{x_j(\mu)\} - CE\{T_i\}, \tag{25}$$

where T_i is the (random) number of transitions to reach a state in $\bar{S}_E$ for the first time starting from i. Since $x_j(\mu) = \infty$ for all $j \in \bar{S}_E$ (as shown earlier) and $E\{T_i\} < \infty$ by the definition of $\bar{S}_T$ and $\bar{S}_E$, equation (25) yields a contradiction.

We have thus completed the proof of equation (24). This equation can be written as

$$\lim_{T \to \infty} \sum_{t=0}^{T-1} \sum_{j=1}^{n} \left[P(\mu)^t\right]_{ij} c_j(\mu(j)) = \infty, \quad \forall\, i \in \bar{S}.$$

It follows that there exist a $T > 0$ and an $A > 0$ such that

$$\sum_{t=0}^{T-1} \sum_{j=1}^{n} \left[P(\mu)^t\right]_{ij} c_j(\mu(j)) > A, \quad \forall\, i \in \bar{S}.$$

Since $P(\mu^k) \to P(\mu)$ and $\liminf_{k \to \infty} c_j(\mu^k(j)) \geqslant c_j(\mu(j))$ (cf. the lower semicontinuity of $c_j(\cdot)$), it follows that for all sufficiently large k we have

$$\sum_{t=0}^{T-1} \sum_{j=1}^{n} \left[P(\mu^k)^t\right]_{ij} c_j(\mu^k(j)) \geqslant A, \quad \forall\, i \in \bar{S}. \tag{26}$$

Note that we can choose the integer T arbitrarily large; for reasons that will become apparent shortly (cf. equation (29)), we will choose $T \geqslant n$.

We now write for all i and k

$$x_i(\mu^k) = \sum_{t=0}^{T-1} \sum_{j=1}^{n} \left[P(\mu^k)^t\right]_{ij} c_j(\mu^k(j)) + \sum_{j \in \bar{S}} \left[P(\mu^k)^T\right]_{ij} x_j(\mu^k) + \sum_{j \in S,\, j \neq 1} \left[P(\mu^k)^T\right]_{ij} x_j(\mu^k), \tag{27}$$

where we have used the property $x_1(\mu^k) = 0$. Let

$$B = T \min_j c_j(\mu(j)) - \epsilon,$$

where ϵ is some positive number, and let

$$m^k = \min_{j \in S} x_j(\mu^k), \quad \bar{m}^k = \min_{j \in \bar{S}} x_j(\mu^k).$$

Fix some $i \in S$. Using equation (27) we have for all k large enough

$$x_i(\mu^k) \geq B + \delta_i^k m^k + \epsilon_i^k \bar{m}^k, \quad \text{where} \tag{28}$$

$$\delta_i^k = \sum_{j \in S,\, j \neq 1} \left[P(\mu^k)^T\right]_{ij}, \quad \epsilon_i^k = \sum_{j \in \bar{S}} \left[P(\mu^k)^T\right]_{ij}.$$

Note that

$$\delta_i^k + \epsilon_i^k < 1, \quad \forall\, k, i \in S, \tag{29}$$

because μ^k is proper and $T \geqslant n$. Let $i(k)$ be such that $x_{i(k)}(\mu^k) = m^k$. Then from equation (28) we obtain

$$m^k \geq B + \delta_{i(k)}^k m^k + \epsilon_{i(k)}^k \overline{m}^k. \tag{30}$$

Furthermore, in view of the definition of the set S and the continuity of $P(\cdot)$, we have

$$\begin{aligned} \limsup_{k\to\infty} \delta_{i(k)}^k &\leqslant \max_{i\in S,\, i\neq 1} \lim_{k\to\infty} \delta_i^k = \max_{i\in S,\, i\neq 1} \sum_{j\in S,\, j\neq 1} \left[P(\mu)^T\right]_{ij} \\ &= 1 - \min_{i\in S,\, i\neq 1} \left[P(\mu)^T\right]_{i1} < 1. \end{aligned} \tag{31}$$

Fix some $j \in \bar{S}$. Using equations (26) and (27) we have

$$x_j(\mu^k) \geqslant A + \alpha_j^k \overline{m}^k + \beta_j^k m^k, \quad \text{where}$$

$$\alpha_j^k = \sum_{s\in\bar{S}} \left[P(\mu^k)^T\right]_{js}, \quad \beta_j^k = \sum_{s\in S,\, s\neq 1} \left[P(\mu^k)^T\right]_{js}.$$

Note that the continuity of P implies that

$$\lim_{k\to\infty} \alpha_j^k = \sum_{s\in\bar{S}} \left[P(\mu)^T\right]_{js} = 1, \quad \lim_{k\to\infty} \beta_j^k = \sum_{s\in S,\, s\neq 1} \left[P(\mu)^T\right]_{js} = 0, \tag{32}$$

because policy μ is improper, and because of the definition of the sets S and $\bar{S}$. Let $j(k)$ be such that $x_{j(k)}(\mu^k) = \overline{m}^k$. Then

$$\overline{m}^k \geqslant A + \alpha_{j(k)}^k \overline{m}^k + \beta_{j(k)}^k m^k. \tag{33}$$

From equation (30) and the fact $1 - \delta_{i(k)}^k > 0$ [cf. equation (29)], we have

$$m^k \geqslant \frac{B}{1 - \delta_{i(k)}^k} + \frac{\epsilon_{i(k)}^k}{1 - \delta_{i(k)}^k} \overline{m}^k.$$

By using this relation in equation (33), we obtain

$$\overline{m}^k \geqslant A + \alpha_{j(k)}^k \overline{m}^k + \beta_{j(k)}^k \frac{\epsilon_{i(k)}^k}{1 - \delta_{i(k)}^k} \overline{m}^k - \beta_{j(k)}^k \frac{B}{1 - \delta_{i(k)}^k} \quad \text{or}$$

$$\overline{m}^k \left(1 - \alpha_{j(k)}^k - \beta_{j(k)}^k \frac{\epsilon_{i(k)}^k}{1 - \delta_{i(k)}^k}\right) \geqslant A - \beta_{j(k)}^k \frac{B}{1 - \delta_{i(k)}^k}. \tag{34}$$

Now in view of equation (29), we have

$$\frac{\epsilon_{i(k)}^k}{1 - \delta_{i(k)}^k} < 1,$$

and we also have

$$\alpha^k_{j(k)} + \beta^k_{j(k)} < 1, \quad \forall\, k,$$

because μ^k is proper. Therefore, the scalar

$$1 - \alpha^k_{j(k)} - \beta^k_{j(k)} \frac{\epsilon^k_{i(k)}}{1 - \delta^k_{i(k)}}$$

is positive and we can divide with it in equation (34) without reversing the inequality sign. We obtain

$$\overline{m}^k \geqslant \left(A - \beta^k_{j(k)} \frac{B}{1 - \delta^k_{i(k)}} \right) \Bigg/ \left(1 - \alpha^k_{j(k)} - \beta^k_{j(k)} \frac{\epsilon^k_{i(k)}}{1 - \delta^k_{i(k)}} \right). \tag{35}$$

For sufficiently large k, the numerator in the right side of the above inequality is positive and bounded away from zero because $\limsup_{k \to \infty} \delta^k_{i(k)} < 1$ [cf. equation (31)], $\beta^k_{j(k)} \to 0$ [cf. equation (32)], and $A > 0$, while the denominator is positive by the preceding argument and converges to zero since $\alpha_{j(k)} \to 1$ and $\beta^k_{j(k)} \to 0$ [cf. equation (32)]. It follows therefore from equation (35) that $\lim_{k \to \infty} \overline{m}^k = \infty$, which proves the result. Q.E.D.

References

[1] Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, NJ.

[2] _____ and Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York.

[3] _____ and Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ.

[4] Blackwell, D. (1965). Positive Dynamic Programming. *Proc 5th Berkeley Sympos. Math., Statist., and Probability*. Vol. 1, 415–418.

[5] Denardo, E. V. (1967). Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Rev.* **9** 165–177.

[6] Derman, C. (1970). *Finite State Markovian Decision Processes*, Academic Press, New York.

[7] Dynkin, E. B. (1963). The Optimum Choice of the Instant for Stopping a Markov Process. *Soviet Math. Dokl.* **150** 238–240.

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

[9] Grigelionis, R. I. and Shiryaev, A. N. (1966). On Stefan's Problem and Optimal Stopping Rules for Markov Processes. *Theor. Probab. Appl.* **11** 541–558.

[10] Kushner, H. (1971). *Introduction to Stochastic Control*. Holt, Rinehart, and Winston, New York.

[11] Pallu de la Barriere, R. (1967). *Optimal Control Theory*. Saunders, Philadelphia.

[12] Shiryaev, A. N. (1978). *Optimal Stopping Problems*. Springer-Verlag, New York.

[13] Strauch, R. (1966). Negative Dynamic Programming. *Ann. Math. Statist.* **37** 871–890.

[14] Tseng, P. (1990). Solving H-Horizon Stationary Markov Decision Problems in Time Proportional to $\log(H)$. *Oper. Res. Lett.*, **9** 287–297.

[15] Veinott, A. F., Jr. (1969). Discrete Dynamic Programming with Sensitive Discount Optimality Criteria. *Ann. Math. Statist.* **40** 1635–1660.

[16] Whittle, P. (1983). *Optimization over Time*. Wiley, New York.

LABORATORY FOR INFORMATION AND DECISION SYSTEMS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139