# Monotonically improving limit-optimal strategies in finite state decision processes

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

Memorandum COSOR 84-15

# MONOTONICALLY IMPROVING LIMIT-OPTIMAL STRATEGIES IN FINITE STATE DECISION PROCESSES

by

Theodore P. Hill and Jan van der Wal

Eindhoven, The Netherlands

December 1984

# MONOTONICALLY IMPROVING LIMIT-OPTIMAL STRATEGIES IN FINITE STATE DECISION PROCESSES

by

Theodore P. Hill[1)]
University of Hawaii

and

Jan van der Wal
Eindhoven University of Technology

Abstract

In every finite-state leavable gambling problem and in every finite-state Markov decision process with discounted, negative or positive reward criteria there exists a Markov strategy which is monotonically improving and optimal in the limit along every history. An example is given to show that for the positive and gambling cases such strategies cannot be constructed by simply switching to a "better" action or gamble at each successive return to a state.

AMS Subject Classification (1980): primary 60G40, 90C40;
secondary 90C39

Key words and phrases: gambling problem, Markov decision process, strategy, stationary strategy, monotonically improving strategy. limit-optimal strategy.

---

[1)] Research partially supported by a NATO postdoctoral grant, and NSF Grant DMS-84-01604.

## § 1. Introduction.

Suppose you are in a casino with a number of dollars you wish to gamble. You may quit whenever you please, and your objective is to find a strategy which will maximize the probability that you reach some goal, say $ 1000. In formal gambling-theoretic terminology, since there are only a finite number of dollars in the world, and since you may quit and leave whenever you wish, this is a *finite-state leavable gambling problem* [4], and the classical results of Dubins and Savage [4, Theorem 3.9.2.] says that for each $\varepsilon > 0$ there is always a stationary strategy which is uniformly $\varepsilon$-optimal. That is, there is always a strategy for betting in which the bet you place at each play depends only on your current fortune; and using this strategy your expected fortune at the time you quit gambling is within $\varepsilon$ of the most you could expect under any strategy. In general, *optimal* stationary strategies do not always exist, even in finite-state leavable gambling problems [4, Example 3.9.2.] although they do if the number of bets available for each fortune is also finite [4, Theorem 3.9.1.], an assumption which certainly does not hold in a casino with an *oddsmaker* (someone who will let you bet any amount on practically any future event - he simply sets odds he considers favourable to the house). An $\varepsilon$-optimal stationary strategy is by definition quite good, but it does have the disadvantage that it is not getting any *better*, and in general always remains $\varepsilon$ away from optimal at some states.

The purpose of this paper is to introduce the notion of a strategy which is monotonically improving and optimal in the limit, and to prove that such strategies exist in all finite-state leavable gambling problems and in all finite-state Markov decision processes with positive, negative, and discounted pay-offs; in fact even *Markov* strategies [6] with these properties are shown to exist. The questions of whether monotonically improving limit-optimal (MILO) strategies exist in non-leavable finite-state gambling problems, in finite-state average reward Markov decision processes, or in countable state problems (with various pay-offs) are left open.

This paper is organized as follows : Section 2 contains preliminaries including notation and the definition of MILO strategies; and Section 3, 4 and 5 establish the existence of MILO strategies in the discounted, negative, and positive dynamic programming cases, respectively. The existence of MILO strategies in finite-state leavable gambling problems follows from the corresponding result for the positive case.

## § 2. MILO Strategies.

A finite state *Markov decision process* [8,12] can be characterized as a quadruple (S.A,p,r) where : S is a finite set representing the state space; A is a function which associates to each $i \in S$ a non-empty set A(i) (the actions available at state i); p is the trasition probability function with $p_{ij}^{a}$ the probability of a transition to j when in state i action a is taken and r is a function from S to $\mathbb{R}$, where r(i) represents the reward incurred at state i.

(As in [9], the main results in this paper carry over easily to the case where r depends on the action as well as state.)

A *strategy* is a function $\pi$ from partial histories $(i_0,i_1,\ldots,i_n)$ to actions satisfying $(i_0,i_1,\ldots,i_n) \in A(i_n)$ for all $i_0,i_1,\ldots. \in S$ and all n. The collection of all strategies will be denoted by $\Pi$, and M and F will be the sets of all *Markov* and *stationary* strategies respectively (see e.g. [3, 4, 8, 12] for formal definitions). The *conditional strategy* $\pi$ given that partial history $i_0,\ldots,i_n$ has occurred is denoted by $\pi[i_0,i_1,\ldots,i_n]$. For each initial state i, a strategy $\pi$ induces a probability measure $P_{i,\pi}$ on the Borel sigma algebra of subsets of $S^\infty$ (S endowed with the discrete topology), expectation with respect to $P_{i,\pi}$ is denoted $E_{i,\pi}$ . $X_n$ is a random variable denoting the state at time n.

The *value of a strategy* $\pi$ is, for the discounted case

$$v_\beta(\pi) = E_\pi \sum_{n=0}^{\infty} \beta^n r(X_n)$$

and for the positive and negative (i.e., $r \geqslant 0$ and $r \leqslant 0$) cases,

$$v(\pi) = E_\pi \sum_{n=0}^{\infty} r(X_n),$$

where omitting the argument i means that a column vector notation is being used.

Similarly, the *value of state* i is, for the discounted case

$$v_\beta^*(i) = \sup_{\pi\in\Pi} v_\beta(i,\pi)$$

and for the positive and negative cases

$$v^*(i) = \sup_{\pi\in\Pi} v(i,\pi).$$

Assumption : Throughout this paper, it will be assumed that $-\infty < v^* < \infty$.

The question of existence of optimal or nearly optimal strategies of various types (e.g., stationary, Markov) has been studied extensively, for example in [1, 2, 3, 4, 8, 9, 12]. However, as mentioned above, even a stationary strategy which is only $\varepsilon$-optimal is not getting any better, and in general remains $\varepsilon$ away from optimal at some states. Thus it seems natural to ask if there exist strategies which are steadily improving, and optimal in the limit.

Definition 2.1. : A strategy $\pi$ is (everywhere) *monotonically improving* (MI) if for all $i \in S$, all $(i_0,i_1,\ldots) \in S^\infty$, and all $n > 0$,

$$v_\beta(i,\pi[i_0,\ldots,i_n,i_{n+1}]) \geqslant v_\beta(i,\pi[i_0,\ldots,i_n]) \qquad \text{(discounted case);}$$

and

$$v(i,\pi[i_0,\ldots,i_n,i_{n+1}]) \geqslant v(i,\pi[i_0,\ldots,i_n]) \qquad \text{(positive and negative cases).}$$

A strategy is (everywhere) *limit-optimal* (LO) if for all $i \in S$ and all $(i_0,i_1,\ldots) \in S^\infty$,

$$\lim_{n\to\infty} v_\beta(i,\pi[i_0,\ldots,i_n]) = v^*_\beta(i) \qquad \text{(discounted case);}$$

and

$$\lim_{n\to\infty} v(i,\pi[i_0,\ldots,i_n]) = v^*(i) \qquad \text{(positive and negative cases).}$$

Remarks.

(*i*) The notion of a MILO strategy does not require the introduction of an 'external parameter' $\varepsilon$, in contrast to most other notations of 'good' strategies.

(*ii*) *Strictly* monotonically improving strategies do not exist in general (for example in problems where there is only one strategy).

(*iii*) Every stationary strategy is monotonically improving (in a trivial sense, since only weak inequality was required).

(*iv*) A stationary strategy is MILO if and only if it is optimal.

(*v*) Optimal stationary strategies do not always exist, (even in finite-state leavable gambler's problems, recall), so in general MILO strategies, if they exist, *must be non-stationary.*

(*vi*) A strategy is limit optimal if and only if it is 'eventually' persistently $\varepsilon$-optimal [5] for all $\varepsilon > 0$.

(*vii*) An optimal strategy need not be MILO (simply because it need not be conditionally optimal, or even good), as the following examples shows.

Example 2.1. $S = \{1,2,3\}$, $A = \{1,2\}$. State 3 is absorbing, $r(3) = 0$, $p(3,a,3) = 1$. State 2 is reflecting, $r(2) = 1$, $p(2,a,3) = 1$. In state 1, $r(1) = 0$ and $p(1,1,2) = 1$ and $p(1,2,3) = 1$. The strategy which uses action 1 initially at each state, and then uses action 2 at all later times and states is optimal, yet not MILO (although in a rather trivial sense).

(*viii*) A MILO strategy may be bad initially (consider Example 2.1. with action 2 used at time 1 and action 1 thereafter at all states). Of course, one may easily obtain an arbitrarily good MILO strategy from *any* MILO strategy, by just 'starting' it late.

(*ix*) MILO strategies need not exist if r is unbounded (and hence S necessarily infinite), even in leavable gambling problems with countable state spaces (and countably additive gambles).

In fact the following example even shows that limit-optimal strategies need not exist in positive dynamic programming problems (under either the additive notion of $\varepsilon$-optimality given above, or the multiplicative notion used in [5].

Example 2.2. (Modification of an example of Dubins and Sudderth [5, Example 1]). $S = \{0,\pm1,\pm2,\ldots\}$, $A = \{1,2,3\}$; $r(m) = 0$ if $m \geqslant 0$, $r(m) = 2^{-m} - 1$ if $m < 0$; $p(m,1,m) = 1$, $p(m,2,-m) = 1$, $p(m,3,0) = \frac{1}{2} = p(m,3,m+1)$ if $m > 0$, $p(m,a,0) = 1$ for all $m \leqslant 0$. As in [5] it may be shown that no strategy is persistently $\frac{1}{2}$-optimal at state 1, so (via Remark ($\nu\iota$) it is easy to see that no strategy is limit-optimal (or even limit-optimal on a set of histories of positive measure).
Whether MILO strategies exist for unbounded r if a multiplicative notion of $\varepsilon$-optimality is used (as in [5]) is not known to the authors; the proof given below depends very heavily on the finiteness of the state space.

($\chi$) MILO strategies cannot always be constructed by simply switching to a 'better' action at each successive return to a state - one is forced to use some action for extremely long periods, then discard them for actions to be used even longer, and so on.

Example 2.3. (Modification of Example 3 in [7]). $S = \{1,2,3\}$, $A = \{1,2,3,\ldots\}$. State 3 is absorbing, $r(3) = 0$, $p(3,a,3) = 1$. State 2 is reflecting $r(2) = 1$, $p(2,a,3) = 1$.
In state 1, $r(1) = 0$ and $p(1,a,2) = 2^{-a}$, $p(1,a,3) = 3^{-a}$, $p(1,a,1) = 1-2^{-a}-3^{-a}$. Clearly $v^*(1) = 1$, and by Ornstein's result (Proposition 5.3. below) there is a stationary strategy with value $1-\varepsilon$ at state 1;

to find it simply choose an action a satisfying $2^{-a}/(2^{-a}+3^{-a}) \geq 1 - \varepsilon$, and always use action 'a' at state 1. But such a strategy is not limit-optimal, and hence not MILO. The stationary strategy using action a+1 at state 1 is strictly better than the one using a, so in some sense a+1 is a 'better' action than a, but a MILO strategy (the existence of which is guaranteed by Theorem 5.1.) cannot be constructed simply by switching to 'better' actions each time one remains at state 1, for the following reason. Suppose $\pi$ is a strategy which uses no action at 1 more than N times. Then $\pi$ is, and remains, less than $\varepsilon$-optimal, i.e.,

$$v(1,\pi) \leq 1 - \varepsilon \text{ and } v(1,\pi[1,1,\ldots,1]) \leq 1 - \varepsilon,$$

where Prob. (never leave state 1) $\geq \varepsilon := \prod_{a=1}^{\infty} (1-2^{-a}-3^{-a})^N > 0$.

## § 3. Discounted Dynamic Programming.

The main purpose of this section is to prove the following theorem.

Theorem 3.1. In every finite-state discounted Markov decision process a monotonically improving limit-optimal Markov strategy exists.

Recall that $\beta \in (0,1)$, and let $L_\beta(f)v = r + \beta P(f)v$, where f is any map from S into A satisfying $f(i) \in A(i)$, P(f) is the transition matrix corresponding to f, and, as before, omission of the argument i means that vector notation is being used.

The proof of Theorem 3.1. will use several lemmas, the first of which is just the optimality equation of Bellman.

Lemma 3.1. $\sup_{f\in F} L_\beta(f)v^*_\beta = v^*_\beta$.

Lemma 3.2. Let $\pi = (f_0,f_1,\ldots)(f_k \in F)$ be a Markov strategy satisfying $L_\beta(f_k)v^*_\beta \geqslant v^*_\beta - \varepsilon_k e$ for each $f_k$. Then

$$v_\beta(\pi) \geqslant v^*_\beta - \sum_{k=0}^{\infty} \varepsilon_k e.$$

(Here 'e' represents the vector $(1,1,\ldots,1)^T$).

Proof.
$$\begin{aligned} v_\beta(\pi) &= \lim_{k\to\infty} L_\beta(f_0)L_\beta(f_1)\ldots L_\beta(f_k)v^*_\beta \\ &\geqslant \lim_{k\to\infty} L_\beta(f_0)L_\beta(f_1)\ldots L_\beta(f_{k-1})(v^*_\beta - \varepsilon_k e) \\ &\geqslant \lim_{k\to\infty} L_\beta(f_0)\ldots L_\beta(f_{k-1})v^*_\beta - \varepsilon_k e \\ &\geqslant \ldots \geqslant \lim_{k\to\infty} [v^*_\beta - (\varepsilon_0 + \ldots + \varepsilon_k)e] \\ &= v^*_\beta - \sum_{k=0}^{\infty} \varepsilon_k e. \end{aligned}$$

■

For a Markov strategy $\pi = (f_0,f_1,\ldots)$, let $\pi^{(k)}$ denote the Markov strategy $\pi^{(k)} = (f_k,f_{k+1},\ldots)$. Then a Markov strategy $\pi$ is (β-discounted) monotonically improving if $v_\beta(\pi^{(k+1)}) \geqslant v_\beta(\pi^{(k)})$, and is limit-optimal if : $\lim_{k\to\infty} v_\beta(\pi^{(k)}) = v^*_\beta$.

Recall that action 'a' is called (β-discounted) *conserving* in state i [4,8,12] if $r(i) + \beta\sum_j p^a_{ij} v^*_\beta(j) = v^*_\beta(i)$. Let $S_0 \subset S$ be the set of all states in which a conserving action exists, and for each $i \in S_0$, let a(i) be a conserving action in state i.

Further, let $F_0 \subset F$ be the subset of policies with $f(i) = a(i)$ for all $i \in S_0$. (Since the $a(i)$ need not be unique, $F_0$ need also not be unique).

Lemma 3.3. $\sup_{f \in F_0} v_\beta(f) = v_\beta^*$.

Proof. Choose $\varepsilon > 0$ and let $f \in F_0$ be such that $L_\beta(f)v_\beta^* \geq v_\beta^* - \varepsilon e$, which is possible by Lemma 3.1. and the definition of $F_0$. Then for the stationary strategy $f$, $v_\beta(f) \geq v_\beta^* - \varepsilon(1-\beta)^{-1}e$. Since $\varepsilon$ was arbitrary,

$$\sup_{f \in F_0} v_\beta(f) \geq \sup_{\varepsilon > 0} (v_\beta^* - \varepsilon(1-\beta)^{-1}e) = v_\beta^*. \qquad \blacksquare$$

Proof of Theorem 3.1.

If $S_0 = S$, then by Lemma 3.3. it is easy to see that any stationary strategy $f \in F_0$ is *optimal*, and hence by remark *(iv)* in § 1, MILO. Suppose $S_0 \neq S$, let $\varepsilon_0 = 1$, and for $k = 0,1,2,\ldots$ pick a policy $f_k \in F_0$ and define numbers $\delta_k$ and $\varepsilon_{k+1} > 0$ such that

(i) $L_\beta(f_k)v_\beta^* \geq v_\beta^* - \varepsilon_k e$;

(ii) $\delta_k = \min_{i \in S \setminus S_0} (v_\beta^* - L_\beta(f_k)v_\beta^*)(i)$; and

(iii) $\varepsilon_{k+1} = \delta_k/2$.

It will now be shown that the Markov strategy $\pi = (f_0, f_1, \ldots)$ is MILO. To see that $\pi$ is LO, observe first that $\delta_k \leq \varepsilon_k$ for all $k$, which, by (iii), implies that $\varepsilon_{k+1} \leq \varepsilon_k/2 \leq \ldots \leq 2^{-(k+1)}\varepsilon_0 = 2^{-(k+1)}$, and that

$$\varepsilon_{k+j} \leq 2^{-j+1}\varepsilon_{k+1} = 2^{-j}\delta_k.$$

Then (i) and Lemma 3.2. imply that

$$v_\beta(\pi^{(k)}) \geqslant v_\beta^* - \sum_{j=k}^{\infty} \varepsilon_j e \geqslant v_\beta^* - \sum_{j=k}^{\infty} 2^{-j} e = v_\beta^* - 2^{-k+1} e,$$

which shows that $\pi$ is limit-optimal (by the observation following the proof of Lemma 3.2.).

To show that $\pi$ is MI, the states in $S_0$ and in $S \backslash S_0$ will be treated separately.

*Case 1.* $i \in S \backslash S_0$. Then

$$v_\beta(i,\pi^{(k)}) \leqslant v_\beta^*(i) - \delta_k \leqslant v_\beta(i,\pi^{(k+1)}),$$

since

$$v_\beta(i,\pi^{(k+1)}) \geqslant v_\beta^*(i) - \sum_{j=k+1}^{\infty} \varepsilon_j$$

and

$$\delta_k \geqslant \sum_{j=k+1}^{\infty} \varepsilon_j.$$

*Case 2.* $i \in S_0$. Define t to be the hitting time (first entry time of $S \backslash S_0$. Then

$$\begin{aligned} v_\beta(i,\pi^{(k)}) &= E_{i,\pi^{(k)}}\Big[\sum_{j=0}^{t-1} \beta^j r(X_j) + \beta^t v_\beta(X_t,\pi^{(k+t)})\Big] \\ &\leqslant E_{i,\pi^{(k)}}\Big[\sum_{j=0}^{t-1} \beta^j r(X_j) + \beta^t v_\beta(X_t,\pi^{(k+t+1)})\Big] \\ &= E_{i,\pi^{(k+1)}}\Big[\sum_{j=0}^{t-1} \beta^j r(X_j) + \beta^t v_\beta(X_t,\pi^{(k+t+1)})\Big] \\ &= v_\beta(i,\pi^{(k+1)}), \end{aligned}$$

where the inequality follows easily from *Case 1* since $x_t \in S \backslash S_0$ for $t < \infty$ and since $\beta^t = 0$ if $t = \infty$, and the second equality follows since $\pi^{(k)}$ and $\pi^{(k+1)}$ agree up to time t.

Together, *Cases 1* and *2* imply that $\pi$ is monotonically improving.

## § 4. Negative Dynamic Programming.

Theorem 4.1. In every finite-state negative (dynamic programming) Markov decision process a monotonically improving limit-optimal Markov strategy exists.

Recall that $r \leqslant 0$, and let $L(f)v = r + P(f)v$.

The proof of Theorem 4.1. is an exact parallel of that of Theorem 3.1. and only the statements of the key steps will be given.

Lemma 4.1. $\sup_{f \in F} L(f)v^* = v^*$.

Lemma 4.2. Let $\pi = (f_0, f_1, \ldots)$ be a Markov strategy satisfying $L(f_k)v^* \geqslant v^* - \varepsilon_k e$ for all $f_k \in F$. Then $v(\pi) \geqslant v^* - \sum_{k=0}^{\infty} \varepsilon_k e$.

Action 'a' is called (negative dynamic programming) conserving in state i [8,12] if $r(i) + \sum_j p_{ij}^a v^*(j) = v^*(i)$. Let $S_0$, $a(i)$ and $F_0$ be the analogs of those in Section 3. Further, let $M_0$ be the subset of all Markov strategies using policies in $F_0$ only i.e.,

$$M_0 = \{\pi = (f_0, f_1, \ldots): f_k \in F_0 \text{ for all } k = 0,1,2,\ldots\}.$$

Lemma 4.3. $\sup_{\pi \in M_0} v(\pi) = v^*$.

The construction of a negative dynamic programming MILO strategy $\pi = (f_0, f_1, \ldots)$ is then essentially the same as for the discounted case.

## § 5. Positive Dynamic Programming.

Recall that for positive dynamic programming, it is assumed that $r \geqslant 0$ and that $v^*(i) < \infty$ for all $i \in S$. The main purpose of this section is to prove the following theorem.

Theorem 5.1. In every finite-state positive (dynamic programming) Markov decision process a monotonically improving limit-optimal Markov strategy exists.

The essential difference between this case and the discounted and negative cases is that in those cases, one may select any conserving action at a state, and use that action always when the process is in that state, without sacrificing any optimality ... such is not the case in general for conserving actions in positive dynamic programming problems. Therefore a somewhat different argument is needed.

The proof of Theorem 5.1. is based on several propositions and lemmas. Again, let $L(f)v = L^1(f)v = r + P(f)v$ and for $m > 1$, let $L^m(f)v = L(f)L^{m-1}(f)v$ and let $v_m(\pi) = E_\pi \sum_{n=0}^{m} r(X_n)$.

Proposition 5.2. (Ornstein [9], Proposition B). If there is an optimal strategy at each $i \in S$, then there is an $f \in F$ with $v(i,f) = v^*(i)$ for all $i \in S$.

Proposition 5.3. (Blackwell [2], Ornstein [9]). If S is finite, then for each $\varepsilon > 0$ there is an $f \in F$ with both $v(i,f) \geq v^*(i) - \varepsilon$ and $v(i,f) \geq (1-\varepsilon)v^*(i)$ for all $i \in S$.

For the remainder of this section, (S,A,p,r) characterizes a fixed Markov decision process with $|S| < \infty$.

Let $B := \{i \in S: v^*(i) = 0\}$;

$C := \{i \in S:$ there is a $\pi \in \Pi$ with $v(i,\pi) = v^*(i)\}$;

$D := \{i \in S: |A(i)| = 1$ ; and

$T := S \setminus D$.

Further it is assumed that there is no state outside D in which the action set can be restricted to a singleton without changing $v^*$, that is,

(1) If $i \in T$ then for each $a \in A(i)$, $\sup\{v(i,\pi): \pi \in \Pi$ with $\pi(i) = \pi(i_0,\ldots,i_n,i) = a$ for all $i_0,\ldots,i_n \in S\} < v^*(i)$.

First it will be shown that this assumption can be made without any loss of generality.

To see this, first observe that an easy modification of Proposition 5.2. implies the existence of a stationary strategy f which is optimal for all states in C; so for all $i \in C$ the action set A(i) can be reduced to the singleton $\{f(i)\}$.

On S\C there still may be states in which the action set can be restricted to a singleton. Pick one of those states and reduce its action set to such a singleton. Continuing, one ends up with a Markov decision process satisfying (1) which has the same value as the original one, and any MILO strategy in the restricted problem is MILO in the original one. Note however that this construction of D need not be unique.

Example 5.4. $S = \{1,2,3,4\}$; $A(3) = A(4) = \{1\}$, $p(3,1,4) = p(4,1,4) = 1$; $A(1) = A(2) = \{0,1,2,\ldots\}$, $p(1,0,2) = p(2,0,1) = 1$, $p(1,n,3) = 1 - p(1,n,4) = p(2,n,3) = 1 - p(2,n,4) = 1 - n^{-1}$; $r(1) = r(2) = r(4) = 0$, $r(3) = 1$. Clearly $v^*(1) = v^*(2) = 1$, and action 0 is good in both states 1 and 2, but not simultaneously. The construction of D may lead to either $\{1,3,4\}$ or $\{2,3,4\}$.

Lemma 5.5. $\emptyset \neq B \subset C \subset D$.

Proof. $B = \emptyset$ together with $|S| < \infty$ implies the existence of a $\delta > 0$ so that $v^*(i) \geq \delta$ for all $i \in S$, which by Proposition 5.3. would imply the existence of a stationary strategy f with $v_n(i,f) \geq \delta/2$ for all $i \in S$. But then $v_{nk}(i,f) \geq k\delta/2$ for all k and $i \in S$, so $v(i,f) = \infty$ for all i which contradicts $v^* < \infty$; hence $B \neq \emptyset$. On B all strategies are optimal so $B \subset C$; and the conclusion $C \subset D$ follows as in the first part of the justification of (1). ∎

Let $t_E$ be the hitting time of $E \in S$.

Lemma 5.6. There exist policies $f_1, f_2, \ldots \in F$ satisfying

(i) $f_n(i) = f_1(i)$ for all $i \in D$ and all n;

(ii) $v(i,f_n) \geq (1 - 2^{-n})v^*(i)$ for all $i \in S$;

(iii) $P_{i,f_n}(t_\beta < \infty) = 1$ for all $i \in S$ and all $n$; and

(iv) $P_{ij}^{f_n(i)} > 0 \Leftrightarrow P_{ij}^{f_1(i)} > 0$ for all $i, j \in S$ and all $n$.

Proof. Conclusion (i) follows since on D there is only one policy, and (ii) follows from Proposition 5.3. For (iii), fix n and let k be so large that $v_k(f_n) \geqslant \delta v^*$ for some $\delta > 0$.
Clearly $v_k(f_n) + P^k(f_n)v^* \leqslant v^*$, so $P^k(f_n)v^* \leqslant v^* - v_k(f_n) \leqslant (1 - \delta)v^*$.
But this implies $\lim_{m\to\infty} P^m(f_n)v^* = \lim_{j\to\infty} P^{kj}(f_n)v^* \leqslant \lim_{j\to\infty} (1 - \delta)^j v^* = 0$, which proves (iii). Finally, (iv) follows by taking subsequences of the finite probability transition matrices. ∎

Corollary 5.7. Let $f_1, f_2, \ldots$ be as in Lemma 5.6. Then all policies $f \in F$ with $f(i) = f_{k_i}(i)$ for some $k_i$ satisfy $P^n(f)v^* \to 0$ as $n \to *$.

Proof. Immediate from Lemma 5.6. (iv) and (iii), and the definition of B. ∎

Definition 5.8. For any two Markov strategies $\pi_1 = (g_1, g_2, \ldots)$ and $\pi_2 = (h_1, h_2, \ldots)$ the strategy $\pi_1^k \circ \pi_2$ denotes the Markov strategy $(g_1, g_2, \ldots, g_k, h_1, h_2, \ldots)$.

Lemma 5.9. Let $f_1, f_2, \ldots,$ be as in Lemma 5.6. and let $\pi, \tilde{\pi} \in \Pi$. Then

(i) $v_k(i,\pi) \uparrow v(i,\pi)$ as $k \to \infty$ for all $i \in S$;

(ii) $v_k(i,\pi) \leqslant v(i, \pi^k \circ \tilde{\pi})$ for all $k \geqslant 0$, $i \in S$ and all $\tilde{\pi} \in \Pi$; and

(iii) $L^k(f_n)v(\tilde{\pi}) \to v(f_n)$ as $k \to \infty$ for all $\tilde{\pi} \in \Pi$ and all $n$.

Proof. (i) follows from the assumption that $r \geqslant 0$ and the monotone convergence theorem, (ii) by $r \geqslant 0$ and the definitions of $v_k$ and $\overset{k}{\pi} \circ \tilde{\pi}$, and (iii) from Lemma 5.6. (iii). ∎

Lemma 5.10. Fix $f \in F$ and $\pi \in \Pi$. If $L(f)v(\pi) \leqslant v(\pi)$ then $L^k(f)v(\pi)$ is non-increasing in k. Similarly, if $L(f)v(\pi) \geqslant v(\pi)$, then $L^k(f)v(\pi)$ is non-decreasing in k.

Proof. Immediate from the monotonicity of $L(f)$. ∎

Lemma 5.11. For each $n \geqslant 1$ there is an $m > n$ satisfying $L(f_n)v(f_m) < v(f_m)$ on T.

Proof. Fix $n \geqslant 1$ and suppose, by way of contradiction, that for each $m > n$ there is an $i_m \in T$ satisfying $v(i_m, \overset{1}{f_n} \circ f_m) \geqslant v(i_m, f_m)$. Since $|T| < \infty$, this implies there is an $\hat{i} \in T$ so that for infinitely many m

$$v(\hat{i}, \overset{1}{f_n} \circ f_m) \geqslant v(\hat{i}, f_m) \quad . \tag{2}$$

It will now be shown that $\hat{i}$ must be in D, thereby contradicting the fact that $T \cap D = \emptyset$. To show $\hat{i} \in D$, it is enough to show that for each $\varepsilon > 0$ and each $i \in S$ there is a strategy $\hat{\pi} \in \Pi$ using only $f_n(\hat{i})$ at $\hat{i}$ which is $\varepsilon$-optimal at i (since then, without loss of generality, $A(\hat{i}) = \{f_n(\hat{i})\}$, and so by (1) $\hat{i} \in D$). Fix $\varepsilon > 0$.

By Lemma 5.6. (ii) it is possible to choose $\hat{m}$ satisfying (2) and also $v(i, f_{\hat{m}}) \geqslant v^*(i) - \varepsilon$ for all $i \in S$. Define $\hat{f} \in F$ by $\hat{f}(\hat{i}) = f_n(\hat{i})$, and $\hat{f}(i) = f_{\hat{m}}(i)$ for $i \neq \hat{i}$.

By Lemma 5.10 $\hat{f}$ satisfies

(3) $\quad L^N(\hat{f})v(f_{\hat{m}}) \geqslant v(f_{\hat{m}})$ for all $N \geqslant 1$.

By Lemmas 5.6 (iv) and 5.9 (iii), (3), and the $\varepsilon$-optimality of $f_{\hat{m}}$ for all $i \in S$ follows $v(i,\hat{f}) = \lim_{N\to\infty} v(i,\hat{f}^N \circ f_{\hat{m}}) \geqslant v(i,f_{\hat{m}}) \geqslant v^*(i) - \varepsilon$ for all $i \in S$. Since $\hat{f}$ uses only $f_n(\hat{i})$ at $\hat{i}$, this completes the contradiction. ∎

For the remainder of this section let $f_1, f_2, \ldots$ be a sequence of policies satisfying Lemma 5.6 and also

(4) $\quad L(f_{m-1})v(f_m) < v(f_m)$ on T for all m.

Let $\delta_1, \delta_2, \ldots > 0$ satisfy both

(5a) $\quad \delta_m < 2^{-m}$; and

(5b) $\quad L(f_{m-1})v(f_m) \leqslant v(f_m) - 2\delta_m v^*$ on T,

where (5b) is possible by (4).

Next, let $n_1 < n_2 < \ldots$ be sufficiently large integers such that

(6a) $\quad v(f_m) - v_{n_m}(f_m) \leqslant \delta_m v^*$; and

(6b) $\quad P^{n_m}(f_m)v^* \leqslant \delta_m v^*$,

where (6a) is possible by Lemma 5.9 (iii), and (6b) by Corollary 5.7.

Lemma 5.12. If $n_1, n_2, \ldots$ satisfy (6a - b), then for all $m > 1$,

$$L(f_{m-1})L^{n_m}(f_m)\hat{v} \leqslant L^{n_m}(f_m)\hat{v} \text{ on } T \text{ for all } 0 \leqslant \hat{v} \leqslant v^* .$$

Proof. On T,

$$\begin{aligned} L(f_{m-1})L^{n_m}(f_m)\hat{v} &\leqslant L(f_{m-1})[v(f_m) + P^{n_m}(f_m)v^*] \\ &\leqslant L(f_{m-1})v(f_m) + P(f_{m-1})\delta_m v^* \\ &\leqslant L(f_{m-1})v(f_m) + \delta_m v^* \\ &\leqslant v(f_m) - 2\delta_m v^* + \delta_m v^* \\ &\leqslant v_{n_m}(f_m) + \delta_m v^* - 2\delta_m v^* + \delta_m v^* \\ &= v_{n_m}(f_m) \leqslant L^{n_m}(f_m)\hat{v} , \end{aligned}$$

where the first inequality follows since $\hat{v} \geqslant v^*$ and since $r \leqslant 0$; the second by (6b); the fourth by (5b); the fifth by (6a); and the last inequality since $\hat{v} \geqslant 0$. ∎

It will now be shown that $\pi = (f_1^{n_1} \circ f_2^{n_2} \circ f_3^{n_3} \circ \ldots)$ is MILO; first two more lemmas are needed.

(Note that (5a) has not yet been used; it will be needed for Lemma 5.14.)

Lemma 5.13. For all m, all $j < m$, and all n,

(i) $L(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v^* \leqslant L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v^*$; and

(ii) $L^{n+1}(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v^* \leqslant L^{n}(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_{n_m})v^* .$

Proof. Fix m. Clearly $L(f_m)v^* \leqslant v^*$, so

$$L^{n+1}(f_m)v^* \leqslant L^n(f_m)v^* \quad \text{for all } n . \tag{7}$$

First consider (i) for $j = m - 1$. That (i) holds on T follows from Lemma 5.12. On D, $f_m = f_{m-1}$ (Lemma 5.6 (i)), so by (7)

$$L(f_{m-1})L^{n_m}(f_m)v^* = L^{n_m+1}(f_m)v^* \leqslant L^{n_m}(f_m)v^* \quad \text{on D} .$$

Combining the results on T and D yields (i) for $j = m - 1$. Next, (ii) for $j = m - 1$ follows by the order preserving property of $L(f)(\cdot)$. By induction one easily establishes both results for all $j < m$; only the argument for the case $j = m - 2$ will be given. On T, (i) holds again by Lemma 5.12, and on D $f_{m-2} = f_{m-1}$, so using (ii) for $j = m - 1$,

$$\begin{aligned} L(f_{m-2})L^{n_{m-1}}(f_{m-1})L^{n_m}(f_m)v^* &= L^{n_{m-1}+1}(f_{m-1})L^{n_m}(f_m)v^* \\ &\leqslant L^{n_{m-1}}(f_{m-1})L^{n_m}(f_m)v^* . \end{aligned}$$

Then (ii) with $j = m - 2$ is immediate from (i) for $j = m - 2$ as before. ∎

Lemma 5.14. For all j and all n

$$\lim_{m\to\infty} L^n(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v^* = v(f_j^n \circ f_{j+1}^{n_{j+1}} \circ f_{j+2}^{n_{j+2}} \circ \ldots).$$

Proof. For all j, all n, and all m,

$$\begin{aligned} &L^n(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v^* \\ &\geqslant L_n(f_j)L^{n_{j+1}}(f_{j+1}) \ldots L^{n_m}(f_m)v(f_{m+1}^{n_{m+1}} \circ f_{m+2}^{n_{m+2}} \circ \ldots) \\ &= v(f_j^n \circ f_{j+1}^{n_{j+1}} \circ \ldots) , \end{aligned}$$

so

$$\liminf_{m\to\infty} L^{n}(f_j)L^{n_{j+1}}(f_{j+1}) \dots L^{n_m}(f_m)v^* \geqslant v(f_j^{\,n} \circ f_{j+1}^{\,n_{j+1}} \circ \dots) .$$

To obtain the reverse inequality, note that by Lemma 6.6 (ii), (5a) and (6a) one has

$$(8) \qquad v(f_{m+1}^{\,n_{m+1}} \circ f_{m+2}^{\,n_{m+2}} \circ \dots) \geqslant v_{n_{m+1}}(f_{m+1}) \geqslant v(f_{m+1}) - 2^{-m-1}v^*$$

$$\geqslant (1 - 2^{-m-1})v^* - 2^{-m-1}v^* = (1 - 2^{-m})v^* .$$

Thus

$$v(f_j^{\,n} \circ f_{j+1}^{\,n_{j+1}} \circ \dots)$$

$$= L^{n}(f_j)L^{n_{j+1}}(f_{j+1}) \dots L^{n_m}(f_m)\, v(f_{m+1}^{\,n_{m+1}} \circ f_{m+2}^{\,n_{m+2}} \circ \dots)$$

$$\geqslant L^{n}(f_j)L^{n_{j+1}}(f_{j+1}) \dots L^{n_m}(f_m)v^* - 2^{-m}v^* ,$$

so

$$\limsup_{m\to\infty} L^{n}(f_j)L^{n_{j+1}}(f_{j+1}) \dots L^{n_m}(f_m)v^* \leqslant v(f_j^{\,n} \circ f_{j+1}^{\,n_{j+1}} \circ \dots) ,$$

which completes the proof. ∎

Proof of Theorem 5.1. The strategy $\pi = (f_1^{\,n_1} \circ f_2^{\,n_2} \circ \dots)$ is MILO. To see this, first note that by Lemmas 5.13 and 5.14, for all j and n

$$v(f_j^{\,1} \circ f_{j+1}^{\,n_{j+1}} \circ f_{j+2}^{\,n_{j+2}} \circ \dots) \leqslant v(f_{j+1}^{\,n_{j+1}} \circ f_{j+2}^{\,n_{j+2}} \circ \dots) \quad \text{and}$$

$$v(f_j^{\,n+1} \circ f_{j+1}^{\,n_{j+1}} \circ f_{j+2}^{\,n_{j+2}} \circ \dots) \leqslant v(f_j^{\,n} \circ f_{j+1}^{\,n_{j+1}} \circ f_{j+2}^{\,n_{j+2}} \circ \dots) ,$$

so $\pi$ is MI. That $\pi$ is also LO is immediate from (8). ∎

By paralleling the proof of Theorem 5.1 in a gambling-theoretic framework, or simply by rewriting a finite-state leavable gambling problem as a finite-state total reward problem, one has the following stronger version of Theorem 1 of [7].

Corollary 5.15. In every finite-state leavable gambling problem there is a monotonically improving limit-optimal strategy.

Acknowledgement

The first author is grateful to the Mathematics Department of the University of Leiden for its hospitality and technical assistance during the academic year 1982-83, and to the Mathematics Department of the Eindhoven University of Technology for invitations for several visits that year.

Bibliography.

[1] Blackwell, D., Discounted dynamic programming, Ann. Math. Statist. 36 226-235 (1965).

[2] Blackwell, D., Positive dynamic programming, Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, 415-418 (1967).

[3] Demko, S. and Hill, T., Decision processes with total-cost criteria, Ann. Prob. 9 293-301 (1981).

[4] Dubins, L. and Savage, L., How to gamble if you must, (Inequalities for Stochastic Processes) Dover, New York (1976).

[5] Dubins, L. and Sudderth, W., Persistently $\varepsilon$-optimal strategies, Math. Operations Res. 2, 125-134 (1977).

[6] Hill, T., On the existence of good Markov strategies, Trans. Amer. Math. Soc. 247, 157-176 (1979).

[7] Hill, T., Monotonically improving limit-optimal gambling strategies, Technical Report TWI 83-30, Univ. of Leiden (1983).

[8] Hordijk, A., Dynamic programming and Markov potential theory, Math. Centre Tract 51, Mathematisch Centrum, Amsterdam (1974).

[9] Ornstein, D., On the existence of stationary optimal strategies, Proc. Amer. Math. Soc. 20 563-569 (1969).

[10] Sudderth, W., On measurable gambling problems, Ann. Math. Statist. 42, 260-269 (1971).

[11] Sudderth, W., On the Dubins and Savage characterization of optimal strategies, Ann. Math. Statist. 43, 498-507 (1972).

[12] Van der Wal, J., Stochastic dynamic programming, Mathematical Centre Tract No. 139, Center for Mathematics and Computer Science (1981).