THE EXISTENCE OF GOOD MARKOV STRATEGIES FOR DECISION PROCESSES WITH GENERAL PAYOFFS

Theodore P. HILL
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA

Victor C. PEStIEN
Department of Mathematics and Computer Science, University of Miami, Coral Gables, FL 33124, USA

For countable-state decision processes (dynamic programming problems), a general class of objective functions is identified for which it is shown that good Markov strategies always exist. This class includes product and lim inf rewards, as well as practically all the classical dynamic programming expected payoff functions.

dynamic programming • Markov decision process • Markov strategy

1. Introduction

A fundamental question in the theory of decision processes is whether decision rules (strategies) which depend only on the current state or only on the current time and state yield as high rewards as strategies which take the whole past into account. For many types of objective functions such as average reward payoffs, the first type of strategy (called stationary) is often much inferior to more general strategies, whereas for most common objective functions the second type (called Markov strategies) have been found to be as good as general strategies (cf. [1, 3, 5, 7, 8, 10, 11, 13, 14, 17, 21]).

The main purpose of this paper is to prove the existence of good Markov strategies for a large class of objective functions which includes product and lim inf payoffs as well as practically all of the classical dynamic programming expected payoff functions. One of the key new ideas is the use of a randomized Markov strategy which depends not only on a given non-randomized strategy, but on the (product) reward function as well, in contrast to the usual randomized Markov average which depends only on the original strategy [7, 8, 11, 14, 19].
This paper is organized as follows: Section 2 contains the preliminary definitions and averaging results; Section 3 contains the definition of product-reward dependent randomized Markov strategies, and their application in establishing the existence of good randomized Markov strategies in a large class of objective functions \( \mathcal{W} \); Section 4 establishes the existence of good non-randomized Markov strategies in \( \mathcal{W} \); and Section 5 establishes the existence of good (non-randomized) Markov strategies for the expected lim inf reward criterion.

2. Countable-state decision processes and randomized strategies

A countable-state decision process is a pair \((X, \Gamma)\), where \(X\) is a countable set and \(\Gamma\) associates to each point \(x\) in \(X\) a non-empty collection \(\Gamma(x)\) of probability measures on \(X\). (In classical dynamic programming terminology, \(X\) represents the state space and \(\Gamma(x)\) the actions available at \(x\); in the gambling theory terminology of Dubins and Savage [6] (upon which most of the notation used in this paper is based), \((X, \Gamma)\) is simply a gambling house with a countable number of fortunes and countably-additive gambles). A strategy is a function from the finite sequences in \(X\) (including the empty sequence "\(\emptyset\)") to the probability measures on \(X\), and the same symbol, \(\sigma\), will be used to denote both a strategy and the probability measure it generates on the Borel product sigma-algebra \(\mathcal{B}^X\) on \(X^\infty\) (\(X\) endowed with the discrete topology). A strategy \(\sigma\) in \(\Gamma\) at \(x\) is a strategy such that \(\sigma(\emptyset) \in \Gamma(x)\) and \(\sigma(x_1, \ldots, x_n) \in \Gamma(x_n)\) for all \(x_1, \ldots, x_n \in X\) and all \(n \geq 1\). A strategy \(\sigma\) is Markov [10] if \(\sigma(x_1, \ldots, x_n) = \sigma(x'_1, \ldots, x'_n)\) whenever \(x'_n = x_n\), and is stationary if \(\sigma(x_1, \ldots, x_n) = \sigma(x'_1, \ldots, x'_m)\) whenever \(x'_m = x_n\). The conditional strategy given \(x_1, \ldots, x_n\), \(\sigma[x_1, \ldots, x_n]\), is defined by

\[
\sigma[x_1, \ldots, x_n](x'_1, \ldots, x'_m) = \sigma(x_1, \ldots, x_n, x'_1, \ldots, x'_m).
\]

\(X_n\) denotes the state of the process at time \(n\) and can be regarded as the \(n\)th-coordinate projection map on \(X^\infty\). Thus \(X_n\) is a random variable on the probability space \((X^\infty, \mathcal{B}^X, \sigma)\). In addition, \(\delta(x)\) denotes the Dirac delta measure at \(x\), and \(I_A\) the indicator function of the set \(A\).

Definition 2.1. The discrete randomized closure of \(\Gamma\), \(\hat{\Gamma}\), is the function

\[
\hat{\Gamma}(x) = \left\{ \sum_{i=1}^{\infty} p_i \gamma_i : \gamma_i \in \Gamma(x), p_i > 0, \sum_{i=1}^{\infty} p_i = 1 \right\}.
\]

Thus \(\hat{\Gamma} \supset \Gamma\), and in general the gambler has more strategies available in \(\hat{\Gamma}\) than in \(\Gamma\), namely the "randomized averages" of his originally available strategies. One may think of constructing a strategy in \(\hat{\Gamma}\) by use of independent lotteries: when one is at state \(x\), he may select the lottery (i.e. the \(\{p_i\}\) of his choice, and use that lottery to determine which gamble in \(\Gamma(x)\) he selects (see the proof of Proposition 2.2). Clearly \(\hat{\Gamma}(x)\) is convex for all \(x\), and in general is strictly larger than the convex hull of \(\Gamma(x)\).
The first proposition in this section says that every Markov strategy \( \hat{\sigma} \) in \( \hat{\Gamma} \) may be expressed as an average (expected value) of Markov strategies in \( \Gamma \). Aside from some notational differences, the proposition is similar to theorems of Krylov [14, Theorem 1] and Fainberg [7, Theorem 1]. Because of the changes in notation, a proof is provided here.

**Proposition 2.2.** Let \( \hat{\sigma} \) be a strategy in \( \hat{\Gamma} \) at \( x \). Then there exists a probability triple \((\Omega, \mathcal{A}, \mu)\) and a collection of strategies \( \{\sigma_\omega : \omega \in \Omega\} \) in \( \Gamma \) at \( x \) satisfying

\[
\hat{\sigma}(B) = \int \sigma_\omega(B) \, d\mu(\omega) \quad \text{for all } B \in \mathcal{B}^x.
\]  

Moreover, if \( \hat{\sigma} \) is a Markov strategy, then \( \{\sigma_\omega : \omega \in \Omega\} \) can be chosen to be Markov.

**Proof.** The argument will be provided for the case where \( \hat{\sigma} \) is Markov; the demonstration in the non-Markov case is essentially the same. Fix a Markov strategy \( \hat{\sigma} \) in \( \hat{\Gamma} \) at \( x \), and without loss of generality let \( X = \{1, 2, 3, \ldots\} \). Enlarging the underlying probability space if necessary, embed the "state process" \( X = (X_1, X_2, X_3, \ldots) \) in a larger process \( (X_1, Y_1, X_2, Y_2, \ldots) \) where: the conditional distribution of \( Y_i \) given \( X_1, Y_1, \ldots, Y_{i-1}, X_i \) is uniform on \([0, 1]\) (so the \( \{Y_i\} \) are i.i.d. uniform \([0, 1]\)), and \( Y_i \) is independent of \( X_1, \ldots, X_i \); and the conditional distribution of \( X_{i+1} \) given \( X_1, Y_1, \ldots, X_i, Y_i \) is \( \gamma_{i,k} \) on the set

\[
\{X_i = j\} \cap \{Y_i \in [a_{i,j,k-1}, a_{i,j,k})\},
\]

where \( \sum_{k-1}^\infty p_{i,j,k} \gamma_{i,j,k} \) is the gamble in \( \hat{\Gamma} \) which \( \hat{\sigma} \) uses if in state \( j \) at time \( i \), and

\[
a_{i,j,k} = p_{i,j,1} + \cdots + p_{i,j,k}.
\]

(Recall that both \( p_{i,j,k} > 0 \) and \( \gamma_{i,j,k} \in I^r(j) \) for all \( i, j, k \).) Observe that the distribution of \( X \) in the embedded process is the same as the distribution of \( X \) under the strategy \( \hat{\sigma} \), that is, \( P(X \in B) = \hat{\sigma}(B) \) for all \( B \in \mathcal{B}^x \). Since the random vectors \( X \) and \( Y = (Y_1, Y_2, \ldots) \) take values in Borel spaces, it follows [2, Theorem 4.34] that there exists a regular conditional distribution for \( X \) given \( Y \). That is, there is a function \( \sigma_\omega(\cdot) \) on \([0, 1]^x \times \mathcal{B}^y \) such that for fixed \( \omega \in [0, 1]^x \), \( \sigma_\omega(\cdot) \) is a probability measure on \( (X^x, \mathcal{B}^y) \), and for fixed \( B \in \mathcal{B}^y \), \( \sigma_\omega(\cdot)(B) \) is a version of \( P(X \in B \mid Y) \). Moreover, for each \( \omega \) there is a natural Markov strategy \( \sigma_\omega \) in \( \Gamma \) at \( x \) associated with the measure \( \sigma_\omega \); if \( \omega = (r_1, r_2, \ldots) \) then \( \sigma_\omega \) is the strategy which in state \( j \) at time \( i \) uses \( \gamma_{i,j,k} \in I^r(j) \), where \( k \) is determined by \( a_{i,j,k-1} = a_{i,j,k} \). Let \( \Omega \) be \([0, 1]^x, \mathcal{A} \) the product-Borel sigma algebra, and \( \mu \) the product of countably many copies of Lebesgue measure on \([0, 1] \). The conclusion (2.1) then follows, since

\[
\hat{\sigma}(B) = P(X \in B) = \int P(X \in B \mid Y) \, dP_Y = \int \sigma_\omega(B) \, d\mu(\omega).
\]

In contrast to the conclusion of Proposition 2.2 pertaining to Markov strategies, it is not always possible to write a stationary strategy in \( \hat{\Gamma} \) as the average of stationary strategies in \( \Gamma \).
Example 2.3. Let $X = \{a, b, c\}$; let $I'(a) = \{\delta(b), \delta(c)\}$, and $I'(b) = I'(c) = \delta(a)$. Let $\hat{\sigma}$ be the stationary strategy in $\hat{\Gamma}$ at $b$ which uses $(\delta(b) + \delta(c))/2$ always at state "a", and let $B = \{X_2 = b$ and $X_4 = c\}$. Then $\hat{\sigma}(B) = \frac{1}{2}$, but $\sigma(B) = 0$ for all (i.e. both) stationary strategies $\sigma$ in $\Gamma$ at $b$.

Proposition 2.4. Let $\hat{\sigma}$ be a strategy in $\hat{\Gamma}$ at $x$, and let $f : X \to \mathbb{R}$ be measurable and $\hat{\sigma}$-integrable (i.e. $\int |f| \, d\hat{\sigma} < \infty$). Then there exist strategies $\sigma_1$ and $\sigma_2$ in $\Gamma$ at $x$ with

$$\int f \, d\sigma_1 \leq \int f \, d\hat{\sigma} \leq \int f \, d\sigma_2.$$  

Moreover, if $\hat{\sigma}$ is Markov, then $\sigma_1$ and $\sigma_2$ can be chosen to be Markov.

Proof. By Proposition 2.2 there is a probability space $(\Omega, \mathcal{A}, \mu)$ and a collection of (Markov) strategies $\{\sigma_\omega\}$ in $\Gamma$ at $x$ satisfying (2.1). It then follows routinely (for indicator functions, then for simple functions, then for limits of simple functions) using the dominated convergence theorem that

$$\int f \, d\sigma = \int \left( \int f \, d\sigma_\omega \right) \, d\mu(\omega).$$

Since $\mu$ is a probability measure, this even implies there is a set of $\omega$ of positive $\mu$-measure for which $\int f \, d\sigma_\omega \geq \int f \, d\sigma$. \qed

An immediate corollary to Proposition 2.4 is that in decision processes with bounded payoffs $f$, one may do just as well with non-randomized (pure) strategies as with randomized ones.

Corollary 2.5. Let $f : X \to \mathbb{R}$ be bounded and measurable. Then

(i) $\sup \left\{ \int f \, d\sigma : \sigma \text{ in } \Gamma \text{ at } x \right\} = \sup \left\{ \int f \, d\hat{\sigma} : \hat{\sigma} \text{ in } \hat{\Gamma} \text{ at } x \right\}$, and

(ii) $\inf \left\{ \int f \, d\sigma : \sigma \text{ in } \Gamma \text{ at } x \right\} = \inf \left\{ \int f \, d\hat{\sigma} : \hat{\sigma} \text{ in } \hat{\Gamma} \text{ at } x \right\}$.

Equality is not always attainable in the conclusion of Proposition 2.4, as the following easy example shows.

Example 2.6. Let $X = \{a, b, c\}$; let $I'(a) = \{\delta(b), \delta(c)\}$, $I'(b) = \{\delta(b)\}$, and $I'(c) = \{\delta(c)\}$; and define $f$ by $f(x_1, x_2, \ldots) = 1$ if $x_1 = b$, and $= 0$ otherwise. If $\hat{\sigma}$ is the stationary strategy in $\hat{\Gamma}$ at "a" given by $\hat{\sigma}(\emptyset) = (\delta(b) + \delta(c))/2$, then $\int f \, d\hat{\sigma} = \frac{1}{2}$, but $\int f \, d\sigma$ is $0$ or $1$ for all $\sigma$ in $\Gamma$. 

3. Randomized Markov strategies

Assume that \((X, \Gamma)\) is a countable-state decision process.

**Definition 3.1.** For each positive integer \(n\), let \(r_n\) be a non-negative real-valued function with domain \(X\) and let \(\mathcal{S}_\Gamma\) be the collection of all strategies in \(\Gamma\). Let \(\mathcal{W}_1\) be the set of all functions \(w_1: \mathcal{S}_\Gamma \rightarrow [-\infty, +\infty] = \mathbb{R}\) which are of the form

\[
w_1(\sigma) = g\left( E_\sigma[r_1(X_1)], E_\sigma[r_2(X_2)], \ldots, E_\sigma[r_n(X_n)], \ldots \right),
\]

and let \(\mathcal{W}_2\) be the set of all functions \(w_2: \mathcal{S}_\Gamma \rightarrow [-\infty, +\infty]\) of the form

\[
w_2(\sigma) = g\left( E_\sigma[r_1(X_1)], E_\sigma[r_1(X_1)r_2(X_2)], \ldots, E_\sigma\left[ \prod_{j=1}^{n} r_j(X_j) \right], \ldots \right),
\]

where \(g\) is any function having domain \((\mathbb{R}^+)\) and taking values in \(\mathbb{R}\). Let \(\mathcal{W}_1 \cup \mathcal{W}_2\) be denoted by \(\mathcal{W}\).

Thus \(\mathcal{W}_1\) consists of those payoff functions which depend only on expected one-stage rewards, while \(\mathcal{W}_2\) includes those payoffs which depend on successive expected product rewards.

The following example lists some typical functions in \(\mathcal{W}\), both some standard dynamic programming objectives and several non-standard ones.

**Example 3.2.** (a) **total reward.** Then \(w \in \mathcal{W}_1\), where

\[
w(\sigma) = E_\sigma\left[ \sum_{n=1}^{\infty} r_n(X_n) \right].
\]

(In particular, if \(\beta > 0\) and \(r_n = \beta^{n-1}r\), then \(w(\sigma)\) is "total discounted reward"; if \(r_n = 0\) for \(n \geq N\), \(w(\sigma)\) is "total finite horizon reward".)

(b) **average reward.** Then \(w \in \mathcal{W}_1\), where

\[
w(\sigma) = \limsup_{n \to \infty} \frac{1}{n} E_\sigma\left[ \sum_{j=1}^{n} r_j(X_j) \right].
\]

(c) **exponential (product) reward.** Then \(w \in \mathcal{W}_1\), where

\[
w(\sigma) = \lim_{n \to \infty} \exp E_\sigma\left[ \sum_{j=1}^{n} r_j(X_j) \right] = \lim_{n \to \infty} \left[ \prod_{j=1}^{n} \exp E_\sigma[r_j(X_j)] \right].
\]

(d) **supremum reward.** Then \(w \in \mathcal{W}_1\), where

\[
w(\sigma) = \sup_{n} E_\sigma r_n(X_n).
\]

(e) **average periodic reward.** Let \(p\) be a positive integer. Then \(w \in \mathcal{W}_1\), where

\[
w(\sigma) = \limsup_{n \to \infty} E_\sigma\left[ \frac{1}{p \sum_{j=pn+1}^{(n+1)p-1}} r_j(X_j) \right].
\]
(In particular, if \( p = 1 \), then \( w(\sigma) \) is “lim sup reward”.)

(f) expected product reward. Then \( w \in \mathcal{W}_2 \), where

\[
w(\sigma) = \limsup_{n \to \infty} E_\sigma \left[ \prod_{j=1}^{n} r_j(X_j) \right].
\]

(In particular, if \( A \subset X \) and \( r_n = I_A \) for each \( n \), then \( w(\sigma) \) is the probability of staying in \( A \) forever.)

(g) maximum deviation. If \( p \) is a constant, then \( w \in \mathcal{W}_1 \), where

\[
w(\sigma) = \limsup_{n \to \infty} | E_\sigma [r_n(X_n)] - p |.
\]

(h) maximum variation (in expected performance). Then \( w \in \mathcal{W}_1 \), where

\[
w(\sigma) = \sup_{n \to \infty} \left( \max_{j \leq n} E_\sigma [r_j(X_j)] - \min_{j \leq n} E_\sigma [r_j(X_j)] \right).
\]

Two objective functions which are not in \( \mathcal{W} \) are the classical gambling-theoretic payoff \( w(\sigma) = E_\sigma [\limsup_{n \to \infty} r(X_n)] \), and the product reward payoff \( w(\sigma) = E_\sigma [r(X_1) \cdot r(X_2) \cdot r(X_3)] \) for the case where \( r \) may take negative and positive values. The question of adequacy of Markov strategies for the gambling-theoretic payoff is answered affirmatively, if \( X \) is finite, in Hill [10]; for the product reward problem, Markov strategies are not in general adequate (Example 3.7 below).

As preparation for the fundamental definition of product-dependent randomized Markov strategy, some notation is needed. For each positive integer \( n \), let \( r_n : X \to \mathbb{R} \) be a non-negative function and let \( \sigma \) be a strategy. If \( q = (x_1, x_2, \ldots, x_n) \) is a sequence (or partial history) of length \( n \) in \( X \), let \( \xi(\sigma, q) \) denote the product \( r_1(x_1)r_2(x_2) \cdots r_n(x_n)P_\sigma(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \).

That is,

\[
\xi(\sigma, q) = \left[ \prod_{j=1}^{n} r_j(x_j) \right] \sigma(\emptyset)(\{x_1\}) \times \sigma(x_1)(\{x_2\})\sigma(x_1x_2)(\{x_3\}) \cdots \sigma(x_1x_2 \cdots x_{n-1})(\{x_n\}).
\]

It is easy to see that

\[
E_\sigma \left[ \prod_{j=1}^{n} r_j(X_j) \right] = \sum_{q: X^n} \xi(\sigma, q).
\]

**Definition 3.3.** The product-dependent randomized Markov strategy for the strategy \( \sigma \) and the rewards \( r_1, r_2, r_3, \ldots \) is the strategy \( \hat{\sigma} \) constructed as follows: let \( \hat{\sigma}(\emptyset) = \sigma(\emptyset) \), and for each \( x \) in \( X \), let \( \hat{\sigma}(x) = \sigma(x) \). For any partial history \( q \) of length \( n \) such that \( n > 1 \) and \( q_n = x_n \), let

\[
\hat{\sigma}(q) = \frac{\sum_{p \in X^n} \xi(\sigma, px_n) \sigma(px_n)}{\sum_{p \in X^n} \xi(\sigma, px_n)} \quad \text{if} \quad \sum_{p \in X^n} \xi(\sigma, px_n) > 0,
\]

and \( \hat{\sigma}(\emptyset) = \sigma(\emptyset) \) otherwise.
\[ \hat{\sigma}(q) = \sigma(x_n) \quad \text{if } \xi(\sigma, px_n) = 0 \text{ for each } p \in X'^{-1}. \]

Notice that the transition probability (gamble) that \( \hat{\sigma} \) uses at \( x_n \) is a mixture of gambles that \( \sigma \) uses at \( x_n \), weighted with respect to the products \( \xi(\sigma, px_n) \). The assumption that each \( r_j \) is non-negative guarantees that \( \hat{\sigma}(q) \geq 0 \). In the case where \( \xi(\sigma, px_n) = 0 \) for each \( p \in X'^{-1} \), the definition of \( \hat{\sigma}(q) \) is rather arbitrary; \( \sigma(x_n) \) was chosen for convenience to ensure that \( \hat{\sigma} \) is Markov and in \( \hat{\Gamma} \).

**Remark.** In the special case where \( r_n \equiv 1 \) for each \( n \), the strategy \( \hat{\sigma} \) in Definition 3.3 has the property that

\[ \hat{\sigma}(X_{n+1} \in G \mid X_n = x) = \sigma(X_{n+1} \in G \mid X_n = x) \]

for each subset \( G \) of \( X \), each positive integer \( n \), and each \( x \in X \). Such a randomized Markov average \( \hat{\sigma} \) has been used often in the literature of dynamic programming. (For example, see Fainberg [7, 8], or Strauch [19]).

The product-dependent randomized Markov strategy \( \hat{\sigma} \) yields the same product reward as the original strategy \( \sigma \), as seen below:

**Proposition 3.4.** If \( \sigma \) is a strategy, \( r_n : X \to \mathbb{R} \) is a non-negative function for each \( n \), and if \( \hat{\sigma} \) is the product dependent randomized Markov strategy for \( \sigma \) and \( r_1, r_2, r_3, \ldots \), then for each positive integer \( n \),

\[ E_n \left[ \prod_{j=1}^{n} r_j(X_j) \right] = E_{\hat{\sigma}} \left[ \prod_{j=1}^{n} r_j(X_j) \right]. \]

**Proof.** It follows from Definition 3.3 (since \( \sum_{p \in X'^{-1}} \xi(\sigma, px_n) = 0 \) if and only if each summand is zero) that for all \( n \geq 1 \), all \( x_n \in X \), and all \( q \in X'^n \) such that \( q_n = x_n \),

\[ \hat{\sigma}(q) = \sum_{p \in X'^{-1}} \xi(\sigma, px_n) = \sum_{p \in X'^{-1}} \xi(\sigma, px_n) \sigma(px_n). \]

The plan is to show by induction that for all \( n \geq 2 \),

\[ \sum_{q \in X'^n} \xi(\sigma, q) \sigma(q) = \sum_{q \in X'^n} \xi(\hat{\sigma}, q) \hat{\sigma}(q). \]

Then, evaluating each side of the measure equality (3.4) at \( (X_n = x_n) \), multiplying both sides by \( r_n(x_n) \), and summing over all \( x_n \) in \( X \), the desired equality (3.2) will be obtained for all \( n \geq 2 \).

To prove (3.4) for \( n = 2 \), observe that for each \( x \) in \( X \),

\[ \xi(\sigma, x) = r_1(x) \sigma(\emptyset)(\{x\}) = r_1(x) \hat{\sigma}(\emptyset)(\{x\}) = \xi(\hat{\sigma}, x). \]

For the induction step, calculate as follows:

\[ \sum_{q \in X'^n} \xi(\sigma, q) \sigma(q) = \sum_{x_n \in X} \left[ \sum_{p \in X'^{-1}} \xi(\sigma, px_n) \sigma(px_n) \right] \]

\[ = \sum_{x_n \in X} \left[ \sum_{p \in X'^{-1}} \xi(\sigma, px_n) \hat{\sigma}(px_n) \right]. \]
using (3.3) and observing that $\hat{\sigma}(px_n)$ does not depend on $p$. Then, with aid of the induction hypothesis,

$$
\sum_{q \in X^n} \xi(\sigma, q)\sigma(q) = \sum_{x_n \in X} r_n(x_n) \left[ \sum_{p \in X^{n-1}} \xi(\sigma, p)\sigma(p)(\{x_n\})\hat{\sigma}(px_n) \right]
$$

$$
= \sum_{x_n \in X} r_n(x_n) \left[ \sum_{p \in X^{n-1}} \xi(\hat{\sigma}, p)\hat{\sigma}(p)(\{x_n\})\hat{\sigma}(px_n) \right]
$$

$$
= \sum_{q \in X^n} \xi(\hat{\sigma}, q)\hat{\sigma}(q),
$$

and (3.4) is proved; the relation (3.2) follows, as long as $n \geq 2$. For the case where $n = 1$, (3.2) holds because $\hat{\sigma}(\emptyset) = \sigma(\emptyset)$. $\square$

The following theorem, the main result in this section, guarantees the existence of good randomized Markov strategies for all countable-state decision processes with objective function in $\mathcal{W}$. It will also serve as a stepping stone to Theorem 4.2, which asserts the existence of good non-randomized Markov strategies for many of the objective functions in $\mathcal{W}$.

**Theorem 3.5.** Let $(X, I')$ be a countable-state decision process, and suppose $w \in \mathcal{W}$. Then for each $x \in X$ and each $\sigma$ in $I'$ at $x$, there is a randomized Markov strategy $\hat{\sigma}$ in $I'$ at $x$ with $w(\hat{\sigma}) = w(\sigma)$.

**Proof.** Let $x \in X$ and let $\sigma$ be in $I'$ at $x$. In case $w \in \mathcal{W}_1$, let $\hat{\sigma}$ be the randomized Markov strategy described in the remark following Definition 3.3. The relation (3.1) guarantees that

$$
E_{\sigma}[r_n(X_n)] = E_{\hat{\sigma}}[r_n(X_n)]
$$

(3.5)

for each $n \geq 1$, and hence that $w(\sigma) = w(\hat{\sigma})$, proving the theorem. In case $w \in \mathcal{W}_2$, let $\hat{\sigma}$ be the product-dependent randomized Markov strategy for $\sigma$ and $r_1, r_2, r_3, \ldots$. The proof is completed by applying Proposition 3.4. $\square$

**Remarks.** The portion of Theorem 3.5 which pertains to $\mathcal{W}_1$ is a special case of results of Derman and Strauch [5, Theorem 2] and of Hordijk [11, Theorem 13.2].

In Definition 3.1, each function $r_n$ was assumed non-negative, and in the case where $w \in \mathcal{W}_2$, the proof of Theorem 3.5 did use this assumption. However, in the case where $w \in \mathcal{W}_1$, the only purpose of the hypothesis $r_n \geq 0$ was to guarantee the existence of each expectation $E_{\sigma}[r_n(X_n)]$. Thus for the $\mathcal{W}_1$ case, the non-negativity assumption on $r_n$ may be weakened.

Actually, in the $\mathcal{W}_1$ case, the proof of Theorem 3.5 depended only upon the fact that the distributions of $X_n$ under $\sigma$ and $\hat{\sigma}$ are the same, and not upon the integrals in (3.5). Therefore its conclusion would hold for a much larger class than $\mathcal{W}_1$, specifically, objective functions which depend only on the distributions of the
random variables, such as functions of the medians, supports, or variances. But for the application of this theorem in the proof of the existence of good non-randomized Markov strategies (Theorem 4.4), integration (via averages of probability measures and Fatou's Lemma) plays an important role.

The $W_1$ case of Theorem 3.5 could also be extended to reward functions of the form $r_n(X_n, X_{n+1})$, since the distribution of $(X_n, X_{n+1})$ is the same under $\sigma$ as under the $\hat{\sigma}$ described in (3.1). Even further, the reward could depend on the gamble or action as well, with rewards of the form $r_n(X_n, \gamma, X_{n+1})$. Such reward functions have been investigated before (cf. Ornstein [15] or Schäl and Sudderth [18]).

This section concludes with three examples of decision processes where the payoff functions are not in $W$, and where, in contrast to the setting of Theorem 3.5, there are non-Markov strategies in $I$ which yield substantially larger payoffs than any Markov strategies in $\hat{I}$. In the first example, the payoff is the expected maximum reward over times 1, 2, and 3 (recall the maximum expected reward $3.2(d)$ is in $W$). In the second example, the payoff is the product of rewards over times 1, 2, and 3, but negative rewards are allowed. The third example shows that it is not possible to extend Theorem 3.5 to include payoff functions $w$ which can be written as a sum of two payoffs, one from $W_1$ and one from $W_2$.

Example 3.6. Let $X = \{a, b, c, d, e\}$; $r_1 = r_2 = r_3 = r$, where $r(a) = r(c) = r(d) = 0$, $r(b) = 5$, $r(e) = 10$; $I'(a) = \{\frac{1}{2}\delta(b) + \frac{1}{2}\delta(c)\}$, $I'(b) = I'(c) = I'(e) = \{\delta(d)\}$, and $I'(d) = \{\delta(b), \gamma\}$, where $\gamma = \frac{1}{2}\delta(e) + \frac{1}{2}\delta(a)$; and $w(\sigma) = E_{\omega}[\max\{r(X_1), r(X_2), r(X_3)\}]$. Let $\sigma_A$ be any strategy in $I$ at "a" which, when in state $d$ at time 2, uses $\gamma$ if $X_1 = b$ and uses $\delta(b)$ if $X_1 = c$. (That is, $\sigma_A(b, d) = \gamma$ and $\sigma_A(c, d) = \delta(b)$). It is easily verified that $w(\sigma_A) = 3^5$, while for any Markov strategy $\hat{\sigma}$ in $\hat{I}$ at "a", $w(\hat{\sigma}) = 5$.

Example 3.7. Let $X = \{a, b, c, d, e, f\}$, and $r_1 = r_2 = r_3 = r$, where $r(a) = r(b) = r(d) = r(e) = 1$, and $r(c) = r(f) = -1$. Let $I'(a) = \{\frac{1}{2}\delta(b) + \frac{1}{2}\delta(c)\}$, $I'(b) = I'(c) = I'(e) = I'(f) = \{\delta(d)\}$, and $I'(d) = \{\delta(e), \delta(f)\}$. Define the expected product reward objective by $w(\sigma) = E_{\omega}[r(X_1) \cdot r(X_2) \cdot r(X_3)]$ for $\sigma \in \mathcal{I}$. If $\sigma_A$ is a strategy in $I'$ at "a" such that $\sigma_A(b, d) = \delta(e)$ and $\sigma_A(c, d) = \delta(f)$, then $w(\sigma_A) = +1$, but for any Markov strategy $\hat{\sigma}$ in $\hat{I}$ at "a", $w(\hat{\sigma}) = 0$.

Example 3.8. Let $(X, I')$ be as in Example 3.6, let $r_1 = r_2 = r_3 = r_4 = r$, where $r(a) = r(c) = 0$, $r(b) = r(d) = 1$, $r(e) = 5$, and let $w(\sigma) = E_{\omega}[r(X_1) \cdot r(X_2) \cdot r(X_3)] + E_{\omega}[r(X_4)]$. A calculation shows that for the non-Markov strategy $\sigma_A$ defined in Example 3.6, $w(\sigma_A) = 5$, while for any Markov strategy $\hat{\sigma}$ in $\hat{I}$ at "a", $w(\hat{\sigma}) < 1.5$.

If the reward functions $r_n$ are non-negative, then the expected product reward objective $w$ lies in $W$ (Example 3.2(f)), and by Theorem 3.5, good randomized Markov strategies do exist. It follows from Theorem 4.2 in the next section that even good non-randomized Markov strategies exist for such an objective.
Denardo and Rothblum [4] use linear programming to compute optimal policies for problems which have exponential utility functions and which satisfy certain transience conditions. In [12] and [13], Kreps studies the existence of optimal (non-Markov) strategies in problems with finite action spaces and general objective functions and the existence of good strategies which are Markov or stationary with respect to certain attached "summary spaces". Furukawa and Iwamoto [9] prove the existence of $\varepsilon$-optimal stationary strategies for decision problems which have multiplicative payoffs and which satisfy certain monotonicity and Lipschitz conditions. The multiplicative payoff is also used by Rothblum in [16].

As pointed out in [9], the multiplicative payoff often arises naturally in problems where the objective is to maximize system reliability in a device with components in series.

4. Existence of good Markov strategies in $\Gamma$

The previous section established that if the objective function $w$ lies in $\mathcal{W}$ then for any strategy there is a randomized Markov strategy with an equivalent payoff. In this section, Theorems 4.1 and 4.2 demonstrate that for a large class of objective functions within $\mathcal{W}$, it is even possible to find a non-randomized Markov strategy whose payoff is at least as good as the payoff for the original strategy. For those objective functions in $\mathcal{W}$ which are in $\mathcal{W}'$, the conclusions of Theorems 4.1 and 4.2 are similar to those of theorems of Fainberg [8, Section 4]. It is assumed throughout that $(X, \Gamma)$ is a countable-state decision process.

**Theorem 4.1.** If $x \in X$, $w \in \mathcal{W}$, and $w$ is of the form

$$w(\sigma) = \int f \, d\sigma \quad \text{for some bounded, measurable } f : X^\Gamma \to \mathbb{R},$$

then for any strategy $\sigma_\Lambda$ in $\Gamma$ at $x$ there exist Markov strategies $\sigma_M$ and $\sigma_{M'}$ in $\Gamma$ at $x$ such that $w(\sigma_M) \geq w(\sigma_\Lambda) \geq w(\sigma_{M'})$.

**Proof.** Suppose $\sigma_\Lambda$ is in $\Gamma$ at $x$ and $w \in \mathcal{W}$. By Theorem 3.5, there is a randomized Markov strategy $\hat{\sigma}$ in $\hat{\Gamma}$ at $x$ with $w(\hat{\sigma}) = w(\sigma_\Lambda)$. Using Proposition 2.4, there exists a Markov strategy $\sigma_M$ in $\Gamma$ at $x$ such that

$$w(\sigma_M) = \int f \, d\sigma_M \geq \int f \, d\hat{\sigma} = w(\hat{\sigma}) = w(\sigma_\Lambda).$$

The $\sigma_{M'}$ conclusion follows similarly. $\square$

It was shown earlier (Theorem 3.5) that if $w : \mathcal{S}_\Gamma \to \mathbb{R}$ is in $\mathcal{W}$, then for any strategy $\sigma$ there is a randomized Markov strategy $\hat{\sigma}$ with $w(\hat{\sigma}) = w(\sigma)$. By imposing some convexity restrictions, the following key theorem is obtained, which asserts the
existence of good non-randomized Markov strategies. (Of course, by enlarging the state space sufficiently, any problem can be transformed into one where even good "stationary" strategies always exist, but the essence of "stationary" and "Markov" for the original problem is lost under the transformation.)

**Theorem 4.2.** Suppose $r_n : X \to \mathbb{R}^+$ are bounded and $g_n : (\mathbb{R}^+) \to \mathbb{R}$ are convex functions, and that $w : \mathcal{F} \to [-\infty, \infty)$ is defined by $w = \limsup G_n$, where

(i) $G_n(\sigma) = g_n(E_\sigma[r_1(X_1)], \ldots, E_\sigma[r_n(X_n)])$, or

(ii) $G_n(\sigma) = g_n\left(E_\sigma[r_1(X_1)], E_\sigma[r_1(X_1)r_2(X_2)], \ldots, E_\sigma\left[\prod_{j=1}^n r_j(X_j)\right]\right)$.

For each $x \in X$, if

$$\sup\{G_n(\sigma) : n \geq 1 \text{ and } \sigma \text{ is in } \Gamma \text{ at } x\} < \infty,$$

then for any strategy $\sigma_\Lambda$ in $\Gamma$ at $x$ there is a (non-randomized) Markov strategy $\sigma_M$ in $\Gamma$ at $x$ such that

$$w(\sigma_M) \geq w(\sigma_\Lambda).$$

**Proof.** Let $\sigma_\Lambda$ be in $\Gamma$ at $x$. By Theorem 3.5, there is a randomized Markov strategy $\hat{\sigma}$ in $\hat{\Gamma}$ at $x$ such that $w(\hat{\sigma}) = w(\sigma_\Lambda)$. Apply Proposition 2.2 to obtain a family $\{\sigma_\omega\}$ of Markov strategies in $\Gamma$ at $x$ such that for all $n$,

$$E_\omega[r_n(X_n)] = \int \left[ \int r_n(X_n) \, d\sigma_\omega \right] \, d\mu(\omega)$$

and

$$E_\omega\left[\prod_{j=1}^n r_j(X_j)\right] = \int \left[ \int \prod_{j=1}^n r_j(X_j) \, d\sigma_\omega \right] \, d\mu(\omega).$$

Then by the convexity of $g_n$ and by Jensen's inequality, it follows that

$$G_n(\hat{\sigma}) \leq \int G_n(\sigma_\omega) \, d\mu(\omega)$$

for all $n$. Since $\sup_{\omega,n}\{G_n(\sigma_\omega)\} < \infty$, Fatou's Lemma applies, and

$$w(\hat{\sigma}) = \limsup G_n(\hat{\sigma}) \leq \limsup \int G_n(\sigma_\omega) \, d\mu(\omega)$$

$$\leq \limsup \int G_n(\sigma_\omega) \, d\mu(\omega)$$

$$= \int w(\sigma_\omega) \, d\mu(\omega).$$

Since $\mu$ is a probability measure, the relation above guarantees that there exists $\omega$ (in fact, a set of $\omega$'s of positive measure) with

$$w(\sigma_\omega) \geq w(\hat{\sigma}) = w(\sigma_\Lambda).$$
For \( x \in X \), let \( W(x) = \sup\{w(\sigma) : \sigma \text{ is in } \Gamma \text{ at } x\} \). A strategy \( \sigma_A \) in \( \Gamma \) at \( x \) is \textit{optimal} if \( w(\sigma_A) = W(x) \); for \( \epsilon > 0 \), \( \sigma_A \) is \( \epsilon\)-\textit{optimal} if \( w(\sigma_A) > W(x) - \epsilon \). The following corollary is immediate.

\textbf{Corollary 4.3.} If the hypotheses of Theorem 4.2 are satisfied, then for each \( \epsilon > 0 \) and \( x \in X \), there is a Markov \( \epsilon\)-optimal strategy in \( \Gamma \) at \( x \). If also there is an optimal strategy in \( \Gamma \) at \( x \), then there exists a Markov optimal strategy in \( \Gamma \) at \( x \).

\textbf{Remarks.} If one replaces "\( \limsup \)" by "\( \sup \)" in the definition of \( w \) in Theorem 4.2, the conclusion still holds; to see this, let \( h_n = \max_{j \leq n} g_j \), and observe that the \( h_n \) are convex and that \( \limsup h_n = \lim h_n = \sup g_n \).

The convexity assumption in Theorem 4.2 is not as restrictive as it might seem at first glance; observe that all the payoffs in Example 3.2 except (f) are of this form (in fact, in (a), (b), (d), and (e), the \( g_n \) are even \( \text{linear} \)).

If the functions \( g_n \) in the statement of Theorem 4.2 are not convex, there may be no Markov strategy \( \sigma_M \) for which \( w(\sigma_M) \geq w(\sigma_A) \). Indeed, \( w(\sigma_A) \) could be strictly larger than the supremum of \( w(\sigma_M) \) over Markov strategies \( \sigma_M \) available at \( x \), as the next example illustrates.

\textbf{Example 4.4.} Let \( X = \{a, b, c, d, e, f\} \) and \( r_n = r \) for each \( n \), where \( r(e) = 1 \) and \( r = 0 \) otherwise. Let \( I'(a) = \{\delta(b), \delta(c)\} \), \( I'(b) = I'(c) = I'(e) = I'(f) = \{\delta(d)\} \), and \( I'(d) = \{\delta(e), \delta(f)\} \). Define \( w \) on \( \mathcal{F}_I \) by

\[
w(\sigma) = \begin{cases} 1 & \text{if } E_n[r(X)] = \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}
\]

In the notation of Theorem 4.2, the function \( g_n : \mathbb{R}^n \to \mathbb{R} \) \( (n \geq 3) \) is defined by

\[
g_n(y_1, \ldots, y_n) = \begin{cases} 1 & \text{if } y_j = \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}
\]

Notice that \( g_n \) is not convex, and that if \( \sigma_A \) is a strategy in \( \Gamma \) at \( a \) which satisfies \( \sigma_A(b, d) = \delta(e) \) and \( \sigma_A(c, d) = \delta(f) \), then \( w(\sigma_A) = 1 \). However, for any Markov strategy \( \sigma_M \) in \( \Gamma \) at \( a \), \( E_{\sigma_M}(X) = 1 \) or 0, so \( w(\sigma_M) = 0 \). (Of course, as Theorem 3.5 implies, there does exist a \textit{randomized} Markov strategy \( \hat{\sigma} \) in \( \Gamma \) at \( a \) such that \( w(\hat{\sigma}) = 1 \).)

Practically no condition, including the convexity of the \( g_n \)'s, is \textit{necessary} for the conclusion of Theorem 4.2 to hold, as can be seen by looking at any decision process where \( I'(x) \) has only one element for each \( x \) (so there is only one strategy, and that is even stationary) and an arbitrary payoff function.

\section{5. Good Markov strategies for the \textit{lim inf} objective}

The results of the earlier sections will now be used to show that for the expected \textit{lim inf} payoff, it is possible to find Markov strategies which are nearly optimal.
In the classical gambler’s problem of Dubins and Savage [6], applied to the special case of a countable-state decision process, the payoff associated with a strategy is

$$E_{\sigma} \left[ \limsup_{n \to \infty} u(X_n) \right],$$

where $u : X \to \mathbb{R}$, the utility function, is bounded.

Sudderth [20], working from the Dubins and Savage framework, investigated the analogous payoff

$$E_{\sigma} \left[ \liminf_{n \to \infty} u(X_n) \right]$$

and established sufficient conditions for the existence of good stationary strategies under such a payoff. In particular, he showed that if the state space $X$ is finite and $I(x)$ is finite for each $x \in X$, then optimal stationary strategies exist. In a general countable-state decision problem, however, good stationary strategies need not exist for the expected lim inf objective, as can be seen by examining Example 3.9.2 of [6]. It is proved below that by allowing Markov strategies instead of only stationary ones, near-optimality can be achieved.

Let $(X, I)$ be a countable-state decision process and let $u : X \to \mathbb{R}$ be bounded. Define $W : X \to \mathbb{R}$ by

$$W(x) = \sup \{ E_{\sigma} \left[ \liminf_{n \to \infty} u(X_n) \right] : \sigma \text{ is a strategy in } I \text{ at } x \}.$$

**Theorem 5.1.** For each $\varepsilon > 0$ and each $x \in X$, there exists a Markov strategy $\sigma_M$ in $I$ at $x$ such that

$$E_{\sigma_M} \left[ \liminf_{n \to \infty} u(X_n) \right] > W(x) - \varepsilon.$$

**Proof.** The theorem is easily reduced to the case where $u$ is non-negative and takes only finitely many values. Next, let $x \in X$, $\varepsilon > 0$, and $u$ have values $c_1 < c_2 < \cdots < c_l$. Let $\delta$ be $\varepsilon$ or

$$\min \{ c_{j+1} - c_j : 1 \leq j \leq l-1 \},$$

whichever is smaller. Let $\sigma$ be a strategy in $I$ at $x$ such that

$$E_{\sigma} \left[ \liminf_{n \to \infty} u(X_n) \right] > W(x) - \frac{\delta \varepsilon}{36 c_1}. \tag{5.1}$$

According to Lemma 1 of Sudderth [20], $W(x)$, $W(X_1)$, $W(X_2), \ldots$ is a supermartingale which converges $\sigma$-almost surely,

$$\sigma[\liminf_{n \to \infty} u(X_n) > \liminf_{n \to \infty} W(X_n)] = 0, \tag{5.2}$$

and

$$W(x) \geq E_{\sigma} \left[ \liminf_{n \to \infty} W(X_n) \right] \geq E_{\sigma} \left[ \liminf_{n \to \infty} u(X_n) \right]. \tag{5.3}$$
Use relations (5.1) and (5.3) to obtain

\[
E_u[\liminf_{n \to \infty} W(X_n) - \liminf_{n \to \infty} u(X_n)] < \frac{\delta \varepsilon}{36 c_i}.
\] (5.4)

Then by (5.2), (5.4), and Markov's inequality,

\[
\sigma \left[ \liminf_{n \to \infty} W(X_n) - \liminf_{n \to \infty} u(X_n) \leq \frac{\delta}{6} \right] > 1 - \frac{\varepsilon}{6 c_i}.
\] (5.5)

For each \(m \geq 1\), let

\[
D_m = \left[ \inf_{j \geq m} u(X_j) = \liminf_{n \to \infty} u(X_n) \land \sup_{j \geq m} |W(X_j) - \liminf_{n \to \infty} u(X_n)| \leq \frac{\delta}{3} \right].
\]

From (5.5), the almost-sure convergence of \(W(x), W(X_1), W(X_2), \ldots\), and the assumption that \(u\) takes only finitely many values, there exists a positive integer \(N\) such that

\[
\sigma(D_N) > 1 - \frac{\varepsilon}{36 c_i}.
\] (5.6)

For each \(k, 1 \leq k \leq l\), let

\[
A_k = \left\{ x \in X : u(x) \geq c_k \land c_k - \frac{\delta}{3} < W(x) < c_k + \frac{\delta}{3} \right\}
\]

and

\[
B^N_k = \{(x_1, x_2, \ldots) \in X^\infty : \text{for all } n \geq N, x_n \in A_k\}.
\]

Notice that the sets \(\{B^N_k\}\) are disjoint and that

\[
D_N \subset \bigcup_{k=1}^l B^N_k.
\] (5.7)

For each positive integer \(k\), let \(\hat{\sigma}_k\) be the product-dependent randomized Markov strategy for \(\sigma\) and the rewards \(r_{1,k}, r_{2,k}, r_{3,k}, \ldots\), where each \(r_{j,k} : X \to \mathbb{R}\) is defined by

\[
r_{j,k} = \begin{cases} 1, & \text{for } j < N, \\ l_{A_k} & \text{for } j \geq N. \end{cases}
\] (5.8)

Now define the randomized Markov strategy \(\hat{\sigma}\) in \(\hat{\Gamma}\) at \(x\) as follows: if \(q\) is a partial history of length less than \(N\), let

\[
\hat{\sigma}(q) = \hat{\sigma}_1(q),
\]

and observe that for such \(q\) and for all \(j\), \(\hat{\sigma}_j(q) = \hat{\sigma}_1(q)\). Further, if \(q\) is a partial
history of length $m$, where $m \geq N$, let

$$
\hat{\sigma}(q) = \begin{cases} 
\hat{\sigma}_k(q) & \text{if } q_m \in A_k, \\
\hat{\sigma}_1(q) & \text{if } q_m \in X \setminus \left( \bigcup_{k=1}^{t} A_k \right). 
\end{cases}
$$

(The definition of $\hat{\sigma}(q)$ above is quite arbitrary in the case where $q_m$ does not lie in any of the sets $A_k$.) With aid of (5.8) and Proposition 3.4,

$$
\hat{\sigma}(B_k^N) = E_r \left[ \prod_{j=1}^{\infty} r_{j,k}(X_j) \right] = E_r \left[ \prod_{j=1}^{\infty} r_{j,k}(X_j) \right] 
$$

for each $k \geq 1$. Then calculate:

$$
E_r[\liminf_{n \to \infty} u(X_n)] \geq \sum_{k=1}^{t} \int_{H_k^N} \liminf_{n \to \infty} u(X_n) \ dm \geq \sum_{k=1}^{t} c_k \hat{\sigma}(B_k^N) \geq \sum_{k=1}^{t} c_k \sigma(B_k^N) 
$$

$$
\geq \sum_{k=1}^{t} \int_{H_k^N} \liminf_{n \to \infty} W(X_n) \ dm \frac{e}{3} 
$$

$$
\geq E_r[\liminf_{n \to \infty} W(X_n)] - \frac{2e}{3} > W(x) - e.
$$

(The next-to-last inequality used (5.6) and (5.7), and the last inequality used (5.1) and (5.3).)

Finally, use the above sequence of inequalities together with Proposition 2.4 to obtain a (non-randomized) Markov strategy $\sigma_M$ in $I'$ at $x$ which satisfies the desired relation. □

If the state space $X$ is finite, the conclusion of Theorem 5.1 is a special case of a result of Hill [10], who showed that if the function $f$ in (4.1) is both shift-invariant and permutation-invariant, then good Markov strategies exist.

Acknowledgement

The authors are grateful to the referees for several useful suggestions and comments.

References


