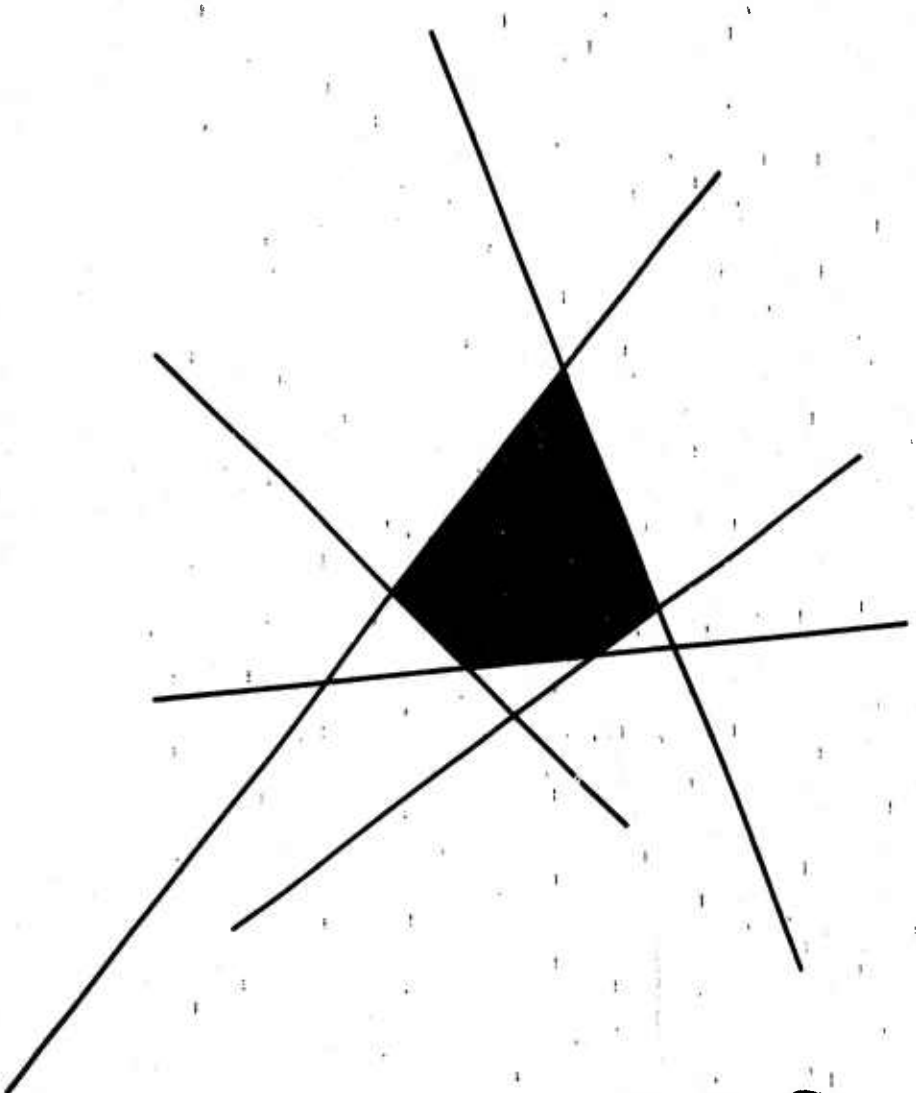
ORC 72-24
SEPTEMBER 1972

DYNAMIC PROGRAMMING AND GAMBLING MODELS

by

SHELDON M. ROSS



OPERATIONS
RESEARCH
CENTER

COLLEGE OF ENGINEERING
UNIVERSITY OF CALIFORNIA • BERKELEY

DYNAMIC PROGRAMMING AND GAMBLING MODELS

by

Sheldon M. Ross
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

SEPTEMBER 1972 ORC 72-24

This research has been partially supported by the U.S. Army Research Office-Durham under Contract DA-31-124-ARO-D-331 and the Office of Naval Research under Contract N00014-69-A-0200-1036 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) University of California, Berkeley	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3 REPORT TITLE

DYNAMIC PROGRAMMING AND GAMBLING MODELS

4 DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Research Report

5 AUTHOR(S) (First name, middle initial, last name)

Sheldon M. Ross

6 REPORT DATE	7a. TOTAL NO OF PAGES	7b. NO OF REFS
September 1972	29	13

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
DA-31-124-ARO-D-331 b. PROJECT NO 20014501B14C c. d.	ORC 72-24 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11 SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Also supported by the Office of Naval Research under Contract N00014-69-A-0200-1036.	U.S. Army Research Office-Durham Box CM, Duke Station Durham, North Carolina 27706

13 ABSTRACT

SEE ABSTRACT.

ii

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Positive Dynamic Programming						
Negative Dynamic Programming						
Gambling Model						
Red-Black Model						
Work-Gambling Model						
Counterexamples						

ABSTRACT

In this paper we formulate and obtain optimal gambling strategies for certain gambling models. We do this by setting these models within the framework of dynamic programming (also referred to as Markovian decision processes) and then utilize results in this field.

In Section 2 we present some dynamic programming results. In particular we review and expand upon two of the main results in dynamic programming. Loosely put these results are:

(i) In problems in which one is interested in maximizing nonnegative rewards, a policy is optimal if and only if its expected return satisfies the optimality equation, and

(ii) in problems in which one is interested in minimizing nonnegative costs, the policy determined by the optimality equation is optimal.

In Section 3 we show how the results of Section 2 may be applied in some simple gambling models. In particular we consider the situation where an individual may bet any integral amount not greater than his fortune and he will win this amount with probability p or lose it with probability $1 - p$. It is shown that if $p \geq 1/2$ then the timid strategy (always bet 1 dollar) both maximizes the probability of ever reaching any preassigned fortune, and also stochastically maximizes the time until the bettor becomes broke. Also, if $p < 1/2$ then the timid strategy while not stochastically maximizing the playing time does maximize the expected playing time.

In Section 4 we consider the same model but with the additional structure that the bettor need not gamble but may instead elect to work for some period of time. His goal is to minimize the expected time until his fortune reaches some preassigned goal. We show that if $p < 1/2$ then (i) always working is optimal, and (ii) among those strategies that only allow working when the bettor is broke it is the bold strategy that is optimal.

In Section 5 we return to the general dynamic programming model and consider the problem of determining "good" subclasses of policies. Two counterexamples are presented.

DYNAMIC PROGRAMMING AND GAMBLING MODELS

by

Sheldon M. Ross

1. INTRODUCTION AND SUMMARY

In this paper we formulate and obtain optimal gambling strategies for certain gambling models. We do this by setting these models within the framework of dynamic programming (also referred to as Markovian decision processes) and then utilize results in this field.

In Section 2 we present some dynamic programming results. In particular we review and expand upon two of the main results in dynamic programming. Loosely put these results are:

(i) In problems in which one is interested in maximizing nonnegative rewards, a policy is optimal if and only if its expected return satisfies the optimality equation, and

(ii) in problems in which one is interested in minimizing nonnegative costs, the policy determined by the optimality equation is optimal.

In Section 3 we show how the results of Section 2 may be applied in some simple gambling models. In particular we consider the situation where an individual may bet any integral amount not greater than his fortune and he will win this amount with probability p or lose it with probability $1 - p$. It is shownthat if $p \geq 1/2$ then the timid strategy (always bet 1 dollar) both maximizes the probability of ever reaching any preassigned fortune, and also stochastically maximizes the time until the bettor becomes broke. Also, if $p < 1/2$ then the timid strategy while not stochastically maximizing the playing time does maximize the expected playing time.

In Section 4 we consider the same model but with the additional structure that the bettor need not gamble but may instead elect to work for some period of time. His goal is to minimize the expected time until his fortune reaches some preassigned goal. We show that if $p < 1/2$ then (i) always working is optimal, and (ii) among those strategies that only allow working when the bettor is broke it is the bold strategy that is optimal.

In Section 5 we return to the general dynamic programming model and consider the problem of determining "good" subclasses of policies. Two counterexamples are presented.

2. SOME DYNAMIC PROGRAMMING PRELIMINARIES

Consider a process that is observed at discrete time points to be in one of the possible states $0,1,2, \ldots$. After observing the state, one of a finite number of possible actions must be chosen. If the process is in state i and action a is chosen then (i) we obtain a reward $R(i,a)$; and (ii) the next state of the process is chosen according to the Markov transition probabilities $\{P_{ij}(a), i,j \geq 0\}$.

If $R(i,a) \geq 0$ for all i , a , then we say that we are in the positive case. A policy is any rule for choosing actions, and for each policy f we define $V_f(i)$ to be the total expected reward earned if the initial state is i and policy f is employed. Also, let $V(i) = \sup_f V_f(i)$. The following equation, known as the optimality equation, is easily established in the positive case (see, for example, page 121 of [11]).

$$V(i) = \max_a \left\{ R(i,a) + \sum_j P_{ij}(a) V(j) \right\}. \tag{1}$$

The major result about the positive case that we shall use is the following.

Proposition 1:

Assume $R(i,a) \geq 0$. The policy f is optimal, i.e., $V_f(i) = V(i)$ $\forall$ i , if and only if its return function $V_f(i)$ satisfies the optimality equation.

This proposition was originally proven by Blackwell [1] in a more general setting than the above. A simple proof is as follows.

Proof of Proposition 1:

If f satisfies the optimality equation then it follows that using f is better (in the expected reward sense) than using any other policy for exactly one

stage and then switching to f. But repeating this argument after the first stage shows that f is better than using any other policy for exactly 2 stages and then switching to f. By induction, it follows that f is better than using any policy for n stages and then switching to f. But as $R(i,a) \geq 0$ it follows that the expected return obtained from time $n+1$ onward is nonnegative. Hence, the expected return from f is greater than the n-stage return from any other policy. The result now follows by letting $n \to \infty$.

Q.E.D.

Remark:

The above proof shows that it is not necessary to require that $R(i,a) \geq 0$. A weaker sufficient condition would be that $V_f(i) \geq 0$. In fact an even weaker sufficient condition would be that

$$\liminf_n E_g[V_f(X_n) \mid X_o = i] \geq 0 \quad \forall \quad i, \forall \, g$$

when X_n is the state of the process at time n.

If $R(i,a) \leq 0$ for all i, a, then we say that we are in the negative case. In this case it is more natural to define $C(i,a) = -R(i,a)$, so as to minimize nonnegative costs as opposed to maximizing nonpositive rewards. Letting V_* denote the infimum of the total expected cost incurred under a policy, it is again easy to establish the optimality equation which now takes the following form.

$$V_*(i) = \min_a \left\{ C(i,a) + \sum_j P_{ij}(a) V_*(j) \right\} . \tag{2}$$

The following proposition was originally proven by Strauch [12].

Proposition 2:

Assume $C(i,a) \geq 0$, and let f be a policy which, when the process is in state i, selects an action minimizing the right-hand side of the optimality equation (2). Then f is optimal.

Proof:

Proposition 2 is proven by noting that if f is determined by the optimality equation (2) then we can get within $\varepsilon/2$ of V_* if we use f for exactly one stage and then switch to a policy within $\varepsilon/2$ of the optimal. Repeating this argument n times shows that we can get within $\frac{\varepsilon}{2} + \frac{\varepsilon}{2^2} + \ldots + \frac{\varepsilon}{2^n}$ of V_* if we use f for n stages and then switch to a policy within $\varepsilon/2^n$ of the optimal. However, as costs are nonnegative it thus follows that the n-stage cost under f is smaller than $V_* + \varepsilon$ and, as ε is arbitrary, the result follows by letting $n \to \infty$.

Q.E.D.

Remark:

The above proof shows that it is not necessary to require that $C(i,a) \geq 0$. It is sufficient for $V_*(i) \geq 0$; and a weaker sufficient condition would be for

$$\liminf_n E_f[V_*(X_n) \mid X_o = i] \geq 0 \quad \forall \quad i .$$

Unfortunately, Proposition 1 is not true in the negative case; nor is Proposition 2 in the positive. A simple counterexample to Proposition 2 in the positive case which is due to Strauch [12] is the following: The states are given by the positive integers, and when in state i we have the choice of either accepting a terminal reward $1 - 1/i$ or else receiving no reward and going to state $i + 1$. Clearly an optimal policy does not exist and hence Proposition 2 could not be valid. From the remark following the proof of Proposition 2 it does, however,

follow in the positive case that if f is chosen by the optimality equation (1) then f is optimal if

$$E_f[V(X_n) \mid X_o = i] \to 0 \quad \text{as} \quad n \to \infty$$

The following counterexample shows that Proposition 1 is not necessarily true in the negative case.

Counterexample:

There are two states and two actions.

$$C(1,1) = 0 \quad C(1,2) = 1 \quad C(2,i) = 0 \quad i = 1, 2$$

$$P_{1,1}(1) = 1 \quad P_{1,2}(2) = 1 \quad P_{2,2}(i) = 1 \quad i = 1, 2$$

Let f be the policy that always chooses action 2. Then $V_f(1) = 1$, $V_f(2) = 0$ and

$$V_f(1) \leq C(1,1) + V_f(1)$$

$$V_f(2) \leq C(2,1) + V_f(2) .$$

Hence V_f satisfies the optimality equation but is obviously not optimal.

One sufficient condition under which Proposition 1 will be valid in the negative case is that the number of stages in our problem be bounded. That is, suppose that there exists a stopped state having the property that once the process enters that state it can never leave it and all costs incurred while in that state are 0. Then it follows from the proof of Proposition 1 that if the time until the process first enters the stopped state is, with probability 1, bounded, for each initial state and for each policy, then the proposition is valid.

A second sufficient condition requires the notion of a stationary policy. We say that a policy is stationary if the action it chooses at any time is a deterministic function of the state of the process at that time. If f is a stationary policy, we define $f(i)$ to be the action f chooses when the process is in state i. We are now ready for

Proposition 3:

Assume that $C(i,a) \geq 0$. If the state and action spaces are both finite, and if f is a stationary policy such that

$$V_f(i) < C(i,a) + \sum_j P_{ij}(a)V_f(j) \quad \forall \quad a \neq f(i) \ , \ \forall \ i \tag{3}$$

where $V_f(i)$ is the total expected cost incurred under f, then f is optimal.

Proof:

We introduce a discount factor α, $0 < \alpha < 1$, and define $V_f^\alpha(i)$ to be the total expected discounted cost incurred under policy f. Since $C(i,a) \geq 0$, it follows from Lebesque's monotone convergence theorem that $\lim_{\alpha \to 1} V_f^\alpha(i) = V_f(i)$. Hence, since the state and action spaces are both finite it follows from (3) that

$$V_f^\alpha(i) < C(i,a) + \alpha \sum_j P_{ij}(a)V_f^\alpha(j) \quad \forall \quad a \neq f(i) \ , \ \forall \ i$$

for all α sufficiently near 1. But, as is well known, if the expected discounted cost from a policy satisfies the discounted optimality equation then that policy is discount optimal. Hence f is α-discount optimal for all α near 1. Therefore, for any policy g, $V_f^\alpha(i) \leq V_g^\alpha(i)$ for α near 1 and the result follows by letting $\alpha \to 1$.

Q.E.D.

A Technical Remark

In the above formulation we have assumed a countable state space and a finite action space. In the general setting of arbitrary state and action spaces these results are more difficult to prove. The main difficulty lies in establishing the optimality equation

$$V(x) = \sup_a \left\{ R(x,a) + \int V(y) dP(y \mid x,a) \right\} .$$

The reason for this difficulty is that it is not easy, in general, to prove that $V(x)$ is a measurable function. However, assuming that $R(x,a)$ is a measurable function and that the probability transition density $P(\cdot \mid x,a)$ is a regular conditional probability measure (this implies that $g(x) \equiv \int h(y) dP(y \mid x,a)$ is a measurable function whenever $h(x)$ is), then we can easily establish the measurability of $V(x)$ in the positive case as long as the action space is countable. This is shown by defining

$$V_1(x) = \sup_a R(x,a)$$

$$V_{n+1}(x) = \sup_a \left\{ R(x,a) + \int V_n(y) dP(y \mid x,a) \right\} .$$

As the supremum of a countable number of measurable functions is itself measurable, it follows by induction that $V_n(x)$ is a measurable function. Also, as $V_n(x)$ is the optimal expected return function for an n-stage problem it follows, in the positive case, that $\lim_n V_n(x) = V(x)$. [For any policy f , since the n-stage return under f is less than V_n , we have that $V_f(x) \leq \lim_n V_n(x)$. On the other hand, since rewards are nonnegative we may, for any ε , define a policy f_n such that $V_{f_n}(x) \geq V_n(x) - \varepsilon$. Hence, $V(x) \leq \lim_n V_n(x) \leq \overline{\lim} V_{f_n}(x) + \varepsilon \leq V(x) + \varepsilon$]. The measurability of V now follows from the fact that the limit of a countable

number of measurable functions is itself measurable.

In the negative case it is not necessarily true that $\lim_n V_n(x)$ is equal to $V(x)$. However, if in addition to assuming that the action space is countable we also assume that the one-stage costs are bounded then we can again easily establish the measurability of $V(x)$. This is done by introducing a discount factor α, $0 < \alpha < 1$. Define $V_\alpha(x)$ to be the optimal discounted cost function. By the same argument as given in the positive case it follows that $V_\alpha(x)$ is measurable. (Since costs are bounded the discount factor assures us that the optimal n-stage discounted cost function will converge to the optimal infinite stage discounted cost function.) The measurability of $V(x)$ now follows since, by Lebesque's monotone convergence theorem, $V(x) = \lim_{n\to\infty} V_{\frac{n-1}{n}}(x)$.

As long as the action space is countable, and $V(x)$ is measurable then Propositions 1 and 2 go through exactly as before. For the analysis in the most general cases the interested reader should consult Blackwell [1], [2], and Strauch [12].

3. THE RED-BLACK GAMBLING MODEL

The red-black gambling model is concerned with the following situation. An individual enters a gambling casino that allows any bet of the following form: If you have a fortune of i units then you are allowed to bet any positive integral amount less than or equal to i. Furthermore if you bet j then you either win j with probability p or lose j with probability $q \equiv 1 - p$.

The first problem we shall consider is that of maximizing the probability that an individual will attain a fortune of N before going broke. This problem fits the framework of positive dynamic programming since if we suppose that a terminal reward of 1 is earned if we ever reach state N and all other rewards are zero, then the expected total reward equals the probability of ever reaching state N. In order to determine an optimal policy we first note that if our present fortune is i then it would never pay to bet more than $N - i$. Hence, from proposition 1 it follows that a policy f will be optimal if and only if its return satisfies

$$V_f(i) \geq pV_f(i + k) + qV_f(i - k) \quad , \quad 0 < i < N \ , \quad k \leq \min(i, N - i) \tag{4}$$

Define the timid strategy to be that strategy which always bets 1. Under this strategy the game becomes the classic gamblers' ruin model and $U(i)$, the probability of reaching N before going broke when you start with i, $0 < i < N$, is given by

$$U(i) = \begin{cases} \dfrac{1 - (q/p)^i}{1 - (q/p)^N} & p \neq 1/2 \\ i/N & p = 1/2 \end{cases} \tag{5}$$

Theorem 1:

If $p \geq 1/2$ the timid strategy maximizes the probability of ever attaining

a fortune of N.

Proof:

If $p = 1/2$ then $U(i) = i/N$ trivially satisfies (4). When $p > 1/2$ we must show that

$$\frac{1 - (q/p)^i}{1 - (q/p)^N} \geq p\left[\frac{1 - (q/p)^{i+k}}{1 - (q/p)^N}\right] + q\left[\frac{1 - (q/p)^{i-k}}{1 - (q/p)^N}\right]$$

or equivalently that

$$(q/p)^i \leq p(q/p)^{i+k} + q(q/p)^{i-k}$$

or

$$1 \leq p[(q/p)^k + (p/q)^{k-1}]$$

Note that the above holds for $k = 1$ and the result will be proven if we can show that $f(x) \equiv \left(\frac{q}{p}\right)^x + \left(\frac{p}{q}\right)^{x-1}$ is an increasing function of x for $x \geq 1$ when $p > 1/2$. This however follows immediately upon differentiation. Q.E.D.

Theorem 1 seems to be one of those results that are well-known but never seem to have been specifically proven in the literature. Of course, timid play was known to be optimal among the class of strategies that always bet a fixed amount at each stage.

Define the bold strategy to be the strategy which, if our present fortune is i

$$\text{bets} \quad i \quad \text{if} \quad i \leq N/2$$

$$\text{bets} \quad N - i \quad \text{if} \quad i > N/2$$

In [6] Dubins and Savage have shown that the bold strategy maximizes the probability of ever attaining a fortune of N when $p \leq 1/2$. Their approach was similar in that they proved this result by showing that the return from the

bold strategy satisfies the optimality equation (4) whenever $p \leq 1/2$. However, as opposed to the timid strategy case it is not possible to easily obtain an exact expression for the return from the bold strategy and Dubins and Savage had to resort to a quite ingenious proof to establish (4).

Thus when $p \leq 1/2$ bold play is optimal while if $p \geq 1/2$ then it is timid play that is optimal. (When $p = 1/2$ it follows from Martingale Theory that any strategy that never bets to strictly exceed N is optimal.) Also by regarding your losses as the winnings of your opponent it follows that if $p \leq 1/2$ then your worst possible strategy is the timid strategy and if $p \geq 1/2$ then your worst possible strategy is the bold one (assuming, of course, that you would never consider betting more than $N - i$ when your present fortune is i).

Suppose now that our objective is not to reach some preassigned goal but rather is to maximize our playing time. We now show that if $p \geq 1/2$ then timid play stochastically maximizes our playing time. That is, for each n, the probability that we will be able to play n or more times before going broke is maximized by the timid strategy.

Theorem 2:

If $p \geq 1/2$ then timid play stochastically maximizes our playing time.

Proof:

By assuming that a reward of 1 is attained if we are able to play at least n times, we see that this problem also fits the framework of the positive case. Hence we must show that starting with i, it is better to play timidly than it is to make an initial bet of k, $k \leq i$, and then play timidly. However this follows since, by Theorem 1, the timid strategy maximizes the probability that we will get to $i + k$ before $i - k$ and it takes *at least* one unit of time. More formally, letting $U_n(i)$ denote the probability that we will

be able to play at least n times given that our initial fortune is i and we play timidly, we obtain by conditioning on the time T that our fortune reaches either $i - k$ or $i + k$ and the value X that is reached we obtain

$$U_n(i) = E[U_{n-T}(X)]$$

$$\geq E[U_{n-1}(X)]$$

$$= U_{n-1}(i + k)P\{X = i + k\} + U_{n-1}(i - k)P\{X = i - k\}$$

$$\geq pU_{n-1}(i + k) + qU_{n-1}(i - k)$$

The first inequality follows from the fact that $U_n(i)$ is a decreasing function of n and $T \geq 1$, while the second inequality follows since $P\{X = i + k\} \geq p$ by Theorem 1 and $U_{n-1}(i + k) \geq U_{n-1}(i - k)$. Q.E.D.

In [9] Molenaar and Van Der Velde considered a gambling casino that accepted any (k, c) bet when k and c are integers. A (k, c) bet would win k with probability $\frac{c}{k + c}$ and would lose c with probability $\frac{k}{k + c}$. Note that these are all fair bets in the sense that the expected gain is zero. They proved, by a concavity argument, that the timid strategy (always bet (1,1)) stochastically maximizes the bettors playing time before going broke. This result, however, also easily follows by our approach since playing timidly is better than making any initial (k, c) bet and then following this initial bet with timid play. This is true since under timid play we would also reach $i + k$ before $i - c$ with probability $\frac{c}{k + c}$ and the amount of time until reaching either value is *at least* one. (Here, of course, i is the bettors initial fortune.)

It turns out that if $p < 1/2$ then the timid strategy does not stochastically maximize our playing time. For suppose $p = .1$ and we start with an initial fortune of 2. The probability that we will be able to play at least 5 games if we play timidly is $1 - (.9)^2 - 2(.9)^3(.1) = .0442$. On the other hand if

we bet 2 initially and then play timidly then the probability of playing at least 5 games is .1 . It is however true that timid play maximizes our expected playing time.

Theorem 3:

If $p < 1/2$ then timid play maximizes our expected playing time.

Proof:

Let $U(i)$ denote the expected number of bets made before we go broke given that we start with i and always bet 1 . To calculate $U(i)$, let X_j denote your winnings on the j^{th} bet and let T denote the number of bets you make before going broke. Then, since $\sum_{j=1}^{T} X_j \equiv -i$ we have by Wald's equation that

$$-i = EX\ ET$$

or

$$U(i) = ET = \frac{i}{1 - 2p}$$

Since maximizing our expected playing time falls under the positive case (we receive a reward of 1 each time that we are able to continue playing) the result follows since

$$U(i) \geq 1 + pU(i + k) + qU(i - k)\ ,\ 1 \leq k \leq i$$

follows by direct verification. Q.E.D.

Theorem 3 remains true in more general gambling models. Consider a gambling casino that allows you to make any bet such that, when your present fortune is i , the outcome of the bet is an integer valued random variable X satisfying

(i) $X \geq -i$ with probability 1

(ii) $|X| \geq 1$ with probability 1

(iii) $EX \leq \alpha - 1$

where α is some fixed positive number less than 1. It follows now in the same manner as in Theorem 3 that the timid strategy which always bets . to either win or lose 1 with respective probabilities $\alpha/2$ and $1 - \alpha/2$ maximizes your expected playing time. This is true since $U(i) = \frac{i}{1 - \alpha}$ is easily shown to satisfy

$$U(i) \geq 1 + EU(i + X)$$

whenever X satisfies (i), (ii), and (iii).

The above also shows that playing in an unfair game with a minimum bet will eventually break you and in a finite expected time (compare with Breiman [4] p. 101).

4. A GAMBLING-WORK MODEL

In this section we consider the following variation of the red-black gambling model. We again suppose that a bettor whose fortune is i may bet any amount j, $j \leq i$ and win or lose j with respective probabilities p and $q \equiv 1 - p$. However, we now suppose that the bettor need not place any bet at all but rather may elect to work. If he decides to work then he works for c units of time and earns 1 dollar. Assuming that each gamble takes 1 unit of time the problem is to determine a strategy that minimizes the expected time until our worker-gambler attains a fortune of N dollars. (The worker-gambler must work when he is broke.)

Theorem 4:

If $p \leq 1/2$ then the strategy of always working minimizes the expected time until a fortune of N is attained.

Proof:

The expected time to reach N if we start from i and always work is given by $U(i) = (N - i)c$. Since the above is clearly a problem of minimizing nonnegative costs we shall apply Proposition 3. Hence, we need show that

$$U(i) < 1 + EU(i + X)$$

or equivalently

$$(N - i)c < 1 + cE(N - i - X)$$

or

$$0 < 1 - cEX$$

which follows since EX, the expected gain of a bet, is negative. In fact, since

$EX \leq 2p - 1$, the above shows that always working is optimal for all values of p such that $p < \frac{c+1}{2c}$.

Q.E.D.

Thus for $p \leq 1/2$ the optimal strategy is always to work. However, let us now consider this same problem but under the assumption that the gambler will only work when broke and thus will gamble at all other times. Under this condition we shall show that it is the bold strategy that is optimal.

Let us suppose that each time the gambler reaches a fortune of N he gives that amount away and then starts to play again. That is after reaching N he then works for c units of time and then starts gambling again with a fortune of 1 dollar. Let us say that a cycle is completed each time the gambler's fortune reaches 0 or N . Letting X_i denote the time of the ith cycle and letting T denote the number of cycles that it takes our gambler to reach a fortune of N we have that the expected time until the gambler reaches N is given by $E\left[\sum_{i=1}^{T} X_i\right]$. Now if the gambler is initially broke and he employs a stationary strategy then the random variables $X_1, X_2, \ldots$ are independent and identically distributed. Hence, by Wald's equation the expected time until the gambler first reaches N is given by $EXET$. However, T is a geometric random variable with mean $1/\alpha$ where α is the probability that starting with 1 the gambler will reach N before 0. Hence, by the Dubins-Savage result it follows, since $p \leq 1/2$, that ET is minimized by the bold strategy. Therefore, as we know by Proposition 2 that an optimal stationary strategy exists, we can show that the bold strategy is optimal if we can show that it minimizes EX . That is we need to show in the original red-black model that the bold strategy minimizes the expected time until the gambler reaches a fortune of either 0 or N . In fact, we shall establish this by proving the stronger result that the bold strategy minimizes $E[\min(X,n)]$ for all n . That is, if the bettor is allowed to play at most n stages and if

he stops before this if he ever reaches 0 or N then the strategy minimizing his playing time is the bold one. The reason for considering this modified problem is that it is a problem with a bounded number of stages and thus we would only need show that the expected playing time under the bold strategy satisfies the optimality equation.

Following Dubins and Savage (Chapter 5 of [6]) we first generalize the model as follows: We suppose that the initial fortune may be any number in $(0,1)$ and that the bettor stops playing either when his fortune reaches 0 or 1 or when he has already played n times. We shall refer to this as the n-stage red-black model.

Define the bold strategy to be the strategy which, if the bettors present fortune is r and he is allowed to bet

$$\text{bets} \quad r \quad \text{if} \quad r \leq 1/2$$

$$\text{bets} \quad 1 - r \quad \text{if} \quad r > 1/2 \ .$$

Let $U_n(r)$ denote the bettors expected playing time in the n-stage red-black model if the initial fortune is r and the bold strategy is employed. By conditioning upon the outcome of the first play we obtain

$$U_n(r) = \begin{cases} 1 + pU_{n-1}(2r) & 0 \leq r \leq 1/2 \\ 1 + qU_{n-1}(2r - 1) & 1/2 < r < 1 \end{cases} \tag{7}$$

$$U_n(1) = U_n(0) = 0 \ , \ U_o(r) = 0 \ , \ 0 < r < 1 \ .$$

Theorem 5:

In the n-stage red-black model, among those strategies which, when the bettor's fortune is r, never bet an amount greater than $1 - r$, the bold strategy minimizes the bettor's expected playing time.

Proof:

Assume first that $p \geq 1/2$. It suffices to prove that

$$U_n(r) \leq 1 + pU_{n-1}(r + s) + qU_{n-1}(r - s) \tag{8}$$

for all $0 \leq s \leq r$, $s \leq 1 - r$. A number of the form $\frac{i}{2^k}$ where i and k are nonnegative integers such that $i \leq 2^k$ will be said to be of order at most k . For example 0 and 1 are of order at most 0; 0, 1/2, 1 are of order at most 1, etc. We first show, by induction, that (8) holds for all n whenever r and s are of order at most k .

Since $U_n(0) = U_n(1) = 0$ it follows that (8) holds for all n whenever r and s are both of order 0 . So assume that (8) holds for all r and s of order at most k and suppose that r and s are of order at most $k + 1$. We first note that if x is of order at most $k + 1$ then

$2x$ is of order at most k when $x \leq 1/2$

$2x - 1$ is of order at most k when $x > 1/2$.

There are four cases we need consider.

Case 1: $r + s \leq 1/2$, $s \leq r$

In this case we have by (7) that

$$U_n(r) - pU_{n-1}(r + s) - qU_{n-1}(r - s) - 1$$

$$= 1 + pU_{n-1}(2r) - p - p^2U_{n-2}(2r + 2s) - q - qpU_{n-2}(2r - 2s) - 1$$

$$= p[U_{n-1}(2r) - pU_{n-2}(2r + 2s) - qU_{n-2}(2r - 2s)] - 1 .$$

But $2r$ and $2s$ are both of order at most k and so the above is nonpositive by the induction hypothesis.

Case 2: $r - s \geq 1/2$

The proof is just as in Case 1, except that the second functional equation of (7) is used instead of the first.

Case 3: $r \leq 1/2 \leq r + s$, $s \leq r$

From (7) we have that

$$\begin{aligned} &U_n(r) - pU_{n-1}(r + s) - qU_{n-1}(r - s) - 1 \\ &= 1 + pU_{n-1}(2r) - p - pqU_{n-2}(2r + 2s - 1) - q - qpU_{n-2}(2r - 2s) - 1 . \end{aligned} \tag{9}$$

Now, since $r \geq s$, it follows that $2r \geq r + s \geq 1/2$ and thus

$$U_{n-1}(2r) = 1 + qU_{n-2}(4r - 1) . \tag{10}$$

Also, since $2r - 1/2 \leq 1/2$ we also have that

$$U_{n-1}(2r - 1/2) = 1 + pU_{n-2}(4r - 1) . \tag{11}$$

Thus from (10) and (11) we obtain that

$$pU_{n-1}(2r) = p + q[U_{n-1}(2r - 1/2) - 1] .$$

Inserting this into (9) yields that (9) is equal to

$$\begin{aligned} &p + qU_{n-1}(2r - 1/2) - q - pqU_{n-2}(2r + 2s - 1) - qpU_{n-2}(2r - 2s) - 1 \\ &= q[U_{n-1}(2r - 1/2) - pU_{n-2}(2r + 2s - 1) - pU_{n-2}(2r - 2s) - 1] + p - 1 . \end{aligned} \tag{12}$$

Now, if $s \geq 1/4$ then, since $p \geq q$, we have that (12) is less than or equal to

$$q[U_{n-1}(2r - 1/2) - pU_{n-2}(2r + 2s - 1) - qU_{n-2}(2r - 2s) - 1] + p - 1$$

which is nonpositive by the induction hypothesis since both $2r - 1/2$ and $2s - 1/2$ are both of order at most k . On the other hand, if $s < 1/4$ then since

$p \geq q$ we have that (12) is less than or equal to

$$q[U_{n-1}(2r - 1/2) - pU_{n-2}(2r - 2s) - qU_{n-2}(2r + 2s - 1) - 1] + p - 1$$

which is nonpositive by the induction hypothesis since $2r - 1/2$ and $1/2 - 2s$ are both of order at most k.

Case 4: $r - s \leq 1/2 \leq r$

The proof for this case is similar to that of the preceding ones and the induction is completed.

Thus, when $p \geq 1/2$, we have proven (8) whenever r and s are binary rationals. (That is, numbers of the form $i/2^k$.) By considering a second player whose initial fortune is $1 - r$, and by viewing the winnings of the first player as the losses of the second one and vice-versa, it follows since both players play the same amount of time that this result is also true when $p \leq 1/2$.

It thus remains to establish (8) when r and s are not binaries. We will do this by showing that $U_n(r)$ is continuous at r whenever r is not a binary rational. Then by letting $\{r_j, j \geq 1\}$ and $\{s_j, j \geq 1\}$ be sequences of binaries such that $r_j \to r$ and $s_j \to s$ it follows by continuity that (8) holds for all $0 \leq s \leq r$, $s \leq 1 - r$. The following lemma thus completes the proof.

Lemma 6:

$U_n(r)$ is continuous at r whenever r is not a binary rational.

Proof:

Let B denote the set of binary rationals, and define

$$s^- = \underset{\substack{\{r_j, r\}, n \\ r_j \to r^- \\ r \notin B}}{\text{supremum}} \limsup_j |U_n(r_j) - U_n(r)|$$

$$s^+ = \underset{\substack{\{r_j, r\}, n \\ r_j \to r^+ \\ r \notin B}}{\text{supremum}} \limsup_j |U_n(r_j) - U_n(r)| .$$

That is, s^- and s^+ measure the worst possible discontinuity of any of the functions $U_n(r)$ when r is not binary. Note that since there is a positive probability of at least $\min(p,1-p)$ that play will end at each stage when the bold strategy is employed, it follows that $U_n(r) \leq \frac{1}{\min(p,1-p)}$. Hence s^- and s^+ are both finite. Note also that $U_n(r)$ is continuous at all $r \notin B$ if and only if $s^- = s^+ = 0$. We show this by contradiction. Assume, for instance, that $s^- > 0$. Let n, $\{r_j\}$, r, be such that

$$r_j \to r^- , \quad r \notin B , \quad \limsup_j |U_n(r_j) - U_n(r)| > s^- \max(p,q) .$$

There are two cases.

Case 1: $r < 1/2$

In this case

$$U_n(r_j) = 1 + pU_{n-1}(2r_j)$$

$$U_n(r) = 1 + pU_{n-1}(2r) .$$

Therefore,

$$\limsup_j |U_{n-1}(2r_j) - U_{n-1}(2r)| = \limsup_j |U_n(r_j) - U_n(r)|/p$$

$$> s^- p/p$$

which is a contradiction since $2r_j \to 2r^-$ and $2r \notin B$.

Case 2: $r > 1/2$

In this case for all j sufficiently large (so that $r_j > 1/2$)

$$U_n(r_j) = 1 + qU_{n-1}(2r_j - 1)$$

$$U_n(r) = 1 + qU_{n-1}(2r - 1)$$

implying that

$$\limsup_j |U_{n-1}(2r_j - 1) - U_{n-1}(2r - 1)| = \limsup_j |U_n(r_j) - U_n(r)|/q$$

$$> s^- \frac{q}{q}$$

which is a contradiction since $2r_j - 1 \to (2r - 1)^-$ and $2r - 1 \notin B$.

Hence s^- must be 0. A similar argument holds for s^+ and hence the Lemma 1 and thus Theorem 5, are proven.

Q.E.D.

Applying Theorem 5 to the red-black gambling-work model yields the following.

Theorem 6:

In the Red-Black Gambling-Work model if the gambler will only work when his fortune is zero then the bold strategy minimizes the expected time until he reaches his goal.

Proof:

This theorem has already been proven (see the remarks following Theorem 4) when the gambler's initial fortune is zero. For an arbitrary initial fortune the proof is exactly as before; we let X_i denote the length of the ith cycle and T the number of required cycles. Then $X_1, X_2, \ldots$ are independent (though X_1 has a different distribution from the others). From Wald's equation the expected time until the gambler reaches N is $E \sum_{i=1}^{T} X_i = E \sum_{i=1}^{T} EX_i$. By the Dubins-Savage result, ET is minimized by the bold strategy and by Theorem 5 EX_i is minimized, for all i , also by the bold strategy.

Q.E.D.

We have not obtained any general results for the work-gambling model when $p > 1/2$. There are, however, reasonable conjectures that may or may not be true. For instance if we are always free to either work or gamble then it seems reasonable that the optimal strategy could be chosen so as to have the following 3-region structure

Optimal Strategy

Work	Gamble	Work
0	Fortune	N

Th is we should work when our fortune is either near 0 or near N and gamble otherwise (of course, any of these regions may be vacuous). For example, suppose that c, the unit of work time, is 1. Then it is obvious that it is optimal to work when our fortune is quite small (for instance 1) or quite large (for instance N - 1), and the conjecture is that you should gamble with an in-between fortune. Of course, the amount gambled remains to be determined.

When the gambler is only permitted to work when broke, the problem is to determine how much he is to gamble at each fortune. In [3] Breiman has shown that in the pure gambling red-black model if one is allowed to bet any fraction of his fortune (and not just integral amounts) then the strategy that *asymptotically* minimizes his expected time to reach some preassigned goal is the Kelly strategy which always bets the fixed fraction $p - q$ of your fortune. Of course Breiman's model does not allow you to work when broke and thus rules out any such strategy as the bold one (which would have an infinite expected time). Nevertheless, assuming that we change the model so as to allow us to bet any fraction of our fortune then it may turn out that the Kelly strategy would remain, in some sense, asymptotically optimal.

5. Some Counterexamples in Dynamic Programming

In this section, we return to the general dynamic programming model consisting of a countable state space and a finite action space, and in which the objective is to maximize nonnegative rewards (i.e., the positive case). In most instances when a specific model (such as the gambling models of Sections 3 and 4) is analyzed within this framework, it turns out that we need only consider stationary policies. However, this is not always the case (though Blackwell [2] did show that if an optimal policy exists then there is a stationary optimal policy) and it is worthwhile to try to determine some subclass of policies such that for any arbitrary policy there necessarily exists a policy within this subclass which performs at least as well. We now define certain subclasses of policies.

A policy is said to be

(1) *stationary*, if the action it chooses at any time is a deterministic function of the state at that time.

(2) *randomized stationary*, if its action at any time is a randomized function of the state at that time.

(3) *Markov* or *memoryless*, if its action at time t is a deterministic function of the state at time t and t.

(4) *randomized Markov* or *randomized memoryless*, if its action at time t is a randomized function of the state at time t and t.

It follows from results presented by Derman and Strauch [5] that if the initial state is fixed then we need never go outside the class of randomized memoryless policies. That is, for any policy f and initial state i, there exists a randomized memoryless policy f' such that

$$V_{f'}(i) \geq V_f(i) .$$

They prove this by showing that the class of randomized memoryless policies is large enough so that, for any policy f, there exists a randomized memoryless policy f' such that

$$P_f\{X_t = j, a_t = a \mid X_0 = i\} = P_{f'}\{X_t = j, a_t = a \mid X_0 = i\}$$

which, of course, implies that $V_f(i) = V_{f'}(i)$.

We now show, by counterexample, that we cannot generally restrict attention to either the class of randomized stationary policies or to the class of memoryless policies.

The first counterexample shows that we cannot always restrict attention to the memoryless policies.

Example 1:

Let the states be given by $0,1,1',2,2', \ldots$. State 0 is an absorbing state and once entered can never be left, i.e.,

$$P_{00} = 1 .$$

In state $n, n > 0$, there are 2 possible actions having respective transition probabilities

$$P_{n,n+1}(1) = P_{n,n'}(2) = 1 , \quad n > 0 .$$

In state $n', n > 0$, there is a single available action, having transition probabilities

$$P_{n',(n-1)'} = 1 \quad n > 1$$

$$P_{1',0} = 1 \quad n = 1 .$$

The rewards depend only on the state and are given by

$$R(n) = 0 \qquad n \geq 0$$

$$R(n') = 1 \qquad n > 0 .$$

Suppose the initial state is state 1. It is easy to see that under any memoryless rule the total expected reward will be finite. However the randomized stationary policy which, when in state n, selects action 1 with probability a_n and action 2 with probability $1 - a_n$ has an infinite expected return when the a_n are chosen so that

$$\prod_{i=1}^{n} a_i \to 0 \quad \text{as} \quad n \to \infty$$

and

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} a_i = \infty$$

for the first condition implies that a primed state will eventually be reached with probability 1 while the second condition implies that the expected number of this first primed state is infinite.

The second example shows that we cannot always restrict attention to the randomized stationary policies.

Example 2:

The states are given by $1,2,3, \ldots, \infty$. In state n there are 2 possible actions having respective transition probabilities

$$P_{n,n+1}(1) = 1 \qquad 1 \leq n < \infty$$

$$P_{n,1}(2) = \alpha_n = 1 - P_{n,\infty}(2) \qquad 1 \leq n < \infty .$$

State ∞ is an absorbing state, i.e.,

$$P_{\infty\infty} = 1 .$$

The rewards depend only on the state and are given by

$$R(1) = 1 \quad .$$

$$R(n) = 0 \qquad n = 2,3, \ldots, \infty .$$

The values α_n are chosen to satisfy

$$\prod_{n=1}^{\infty} \alpha_n > 0 , \qquad \alpha_n < 1 \quad \text{all} \quad n . \tag{13}$$

Suppose that the initial state is state 1. It is easy to see that under any randomized stationary policy the expected number of visits to state 1 is a geometric random variable with finite means, hence the total expected return is finite. However, consider the policy which on its nth return to state 1 chooses action 1 n times and then chooses action 2. Since, by (13) this policy has a positive probability of visiting state 1 infinitely often, it has an infinite expected return.

REFERENCES

[1] Blackwell, D., "Positive Dynamic Programming," PROCEEDINGS OF THE FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, Vol. 1, University of California Press, (1967).

[2] Blackwell, D., "On Stationary Strategies," Royal Statistical Society Journal, Series A, (1970).

[3] Breiman, L. "Optimal Gambling Systems for Favorable Games," FOURTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, University of California Press.

[4] Breiman, L., PROBABILITY, Addison-Wesley, (1968).

[5] Derman, C. and R. Strauch, "A Note on Memoryless Rules for Controlling Sequential Control Processes," Ann. Math. Statist., Vol. 37, pp. 276-279, (1966).

[6] Dubins, L. and L. Savage, HOW TO GAMBLE IF YOU MUST, McGraw Hill, (1965).

[7] Epstein, R., THEORY OF GAMBLING AND STATISTICAL LOGIC, Academic Press (1967).

[8] Freedman, D., "Timid Play is Optimal," Ann. Math. Statist., Vol. 38, pp. 1281-1284, (1967).

[9] Molenaar, W. and E. A. Van Der Velde, "How to survive a Fixed Number of Fair Bets," Ann. Math. Statist., Vol. 38, pp. 1278-1281, (1967).

[10] Ornstein, D., "On the Existence of Stationary Optimal Strategies," Proceedings of Amer. Math. Soc., Vol. 20, pp. 563-569, (1969).

[11] Ross, S., APPLIED PROBABILITY MODELS WITH OPTIMIZATION APPLICATIONS, Holden-Day, (1970).

[12] Strauch, R., "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 171-889, (1966).

[13] Thorp, E., "Optimal Gambling Systems for Favorable Games," Review of International Statistical Institute, Vol. 37, No. 3, (1969).