The Annals of Statistics
1979, Vol. 7, No. 5, 1086–1105

BERNOULLI ONE-ARMED BANDITS—ARBITRARY DISCOUNT SEQUENCES

BY DONALD A. BERRY[1] AND BERT FRISTEDT[2]

University of Minnesota

Each of two arms generate an infinite sequence of Bernoulli random variables. At each stage we choose which arm to observe based on past observations. The parameter of the left arm is known; that of the right arm is a random variable. There are two conflicting desiderata: to observe a success at the present stage and to obtain information useful for making future decisions. The payoff is α_m for a success at stage m. The objective is to maximize the expected total payoff. If the sequence $(\alpha_1, \alpha_2, \cdots)$ is *regular* an observation of the left arm should always be followed by another of the left arm. A rather explicit characterization of optimal strategies for regular sequences follows from this result. This characterization generalizes results of Bradt, Johnson, and Karlin (1956) who considered α_m equal to 1 for $m \leq n$ and 0 for $m > n$ and of Bellman (1956) who considered $\alpha_m = \alpha^{m-1}$ for $0 \leq \alpha < 1$.

1. Introduction. We have two mechanisms $\mathcal{R}$ and $\mathcal{L}$, called the *right arm* and the *left arm*, each of which generates Bernoulli random variables having parameters ρ and λ, respectively.

Making an observation on an arm is called a *pull*. We are to pull $\mathcal{R}$ or $\mathcal{L}$ at each of an infinite number of stages. After pulling an arm we may pull that arm again or we may switch and pull the other arm. Which arm we pull at any stage, say stage m, may depend on the pulls and resulting observations at stages 1 through $m - 1$ (vacuous if $m = 1$). A *strategy*, often denoted by τ, with or without subscript, is a function that assigns to each finite history of pulls and observations the symbol $\mathcal{R}$ or $\mathcal{L}$ denoting the arm to be pulled at the next stage.

The objective is to maximize the expected value of $\sum_{m=1}^{\infty} \alpha_m Z_m$ where Z_m equals 1 or 0 according as a success or failure is observed at stage m and $\mathbf{A} = (\alpha_1, \alpha_2, \cdots)$, called a *discount sequence*, is a nonincreasing sequence of nonnegative numbers such that $\sum_{m=1}^{\infty} \alpha_m < \infty$. Any τ yielding the maximum is *optimal*. The random variable $\sum_{m=1}^{\infty} \alpha_m Z_m$ is called the *payoff* and its expected value under a strategy τ is called the *expected payoff under* τ.

Let X_i and Y_j denote, respectively, the observations on the ith pull of $\mathcal{R}$ and jth pull of $\mathcal{L}$. For convenience we assume that the sequences $(X_1, X_2, \cdots)$ and $(Y_1, Y_2, \cdots)$ are nonterminating, whether or not we, in fact, make infinitely many pulls on each arm.

Received October 1977; revised March 1978.
[1]Supported in part by NIH Grant 5R01-G, 22234-01.
[2]Supported in part by NSF Grant 74-05786 A02 and NSF Grant 78-02694.
AMS 1970 *subject classifications.* Primary 62L05; secondary 62L15.
Key words and phrases. One-armed bandit, sequential decisions, optimal stopping, two-armed bandit, regular discounting, Bernoulli bandit.

We hypothesize that the two sequences $(X_1, X_2, \cdots)$ and $(Y_1, Y_2, \cdots)$ are independent of each other. Further, λ is a known constant and $(Y_1, Y_2, \cdots)$ is a sequence of independent random variables. Also, given the random variable ρ, $(X_1, X_2, \cdots)$ is a sequence of independent random variables. Therefore, the unconditional finite-dimensional distributions of $(X_1, X_2, \cdots)$ are invariant under permutations of the subscripts; that is, $(X_1, X_2, \cdots)$ is a sequence of *exchangeable* random variables.

Let R denote the distribution function, and also the distribution measure, of ρ. The "information" present about $\mathfrak{R}$ initially is given by R. At any stage, if s successes and f failures have been obtained on $\mathfrak{R}$, the information present about $\mathfrak{R}$ is given by $\sigma^s\varphi^f R$, where $\sigma^s\varphi^f R$ is absolutely continuous with respect to R, and

$$\frac{d\sigma^s\varphi^f R}{dR}(x) = \frac{x^s(1-x)^f}{\int_{[0,1]} u^s(1-u)^f R(du)} = \frac{x^s(1-x)^f}{E[\rho^s(1-\rho)^f]}.$$

(In case the support of R is a subset of $\{0, 1\}$ some values of (s, f) are not possible.) Rather than write the conditional expectation $E(\rho^2 | X_1 + \cdots + X_{s+f} = s)$, for instance, we shall write $E(\rho^2 | \sigma^s\varphi^f R)$. In particular, we have the notation $E(\rho | R)$ in which the dependence of the expectation on the underlying probability structure is made explicit.

For R, λ, and a discount sequence $\mathbf{A}$, let $V(\mathbf{A}, R, \lambda)$ denote the maximum over the set of all strategies, or supremum in case the maximum does not exist, of the expected payoff. The maximal expected payoff is at least that of a strategy which pulls the same arm at every stage and is no larger than if ρ is known at the outset. That is,

$$[\lambda \vee E\rho]\Sigma_{m=1}^{\infty}\alpha_m \leqslant V(\mathbf{A}, R, \lambda) \leqslant E(\lambda \vee \rho)\Sigma_{m=1}^{\infty}\alpha_m. \tag{1.1}$$

In Theorem 5.3 we assert that V is a continuous function (of 3 variables). It is easy to see that V is a convex function of each of its three variables by considering an optimal strategy (or ϵ-optimal strategies) for the convex combination; these considerations also show that V is not linear. If $\alpha_i \leqslant \beta_i$ for each i, then $V(\mathbf{A}, R, \lambda) \leqslant V(\mathbf{B}, R, \lambda)$ where $\mathbf{A} = (\alpha_1, \alpha_2, \cdots)$ and $\mathbf{B} = (\beta_1, \beta_2, \cdots)$. The monotonicity of V as a function of R is studied in Section 3. Arguments similar to, but easier than, those used there show $V(\mathbf{A}, R, \cdot)$ to be a nondecreasing function.

For the function V, and other functions, we shall often suppress one or more of the dependent variables. For instance, we may write $V(R)$ in a discussion where both $\mathbf{A}$ and λ are fixed. We use the term "$(\mathbf{A}, R, \lambda)$-bandit" to indicate the particular situation under consideration and abbreviate this to "R-bandit" or "for (R, λ)" when there can be no confusion.

The decision of which arm to pull is governed by two possibly conflicting desiderata: obtaining a success at that stage and obtaining information about $\mathfrak{R}$ that can be used for making decisions in the future. The decision problem is not easy to resolve because these two desiderata cannot be separated and considered one at a time.

We shall need the following result in various proofs where, after an inductive argument, a limiting procedure is needed. Dynamic programming is applicable, though possibly cumbersome, for the discount sequence $(\alpha_1, \cdots, \alpha_n, 0, \cdots)$ mentioned there.

LEMMA 1.1. *Suppose that, for each n, τ_n is an optimal strategy for R, λ, and the discount sequence $(\alpha_1, \cdots, \alpha_n, 0, \cdots)$. Then for R, λ, and the discount sequence $(\alpha_1, \alpha_2, \cdots)$ there is an optimal strategy τ that through any stage m agrees with at least one τ_n (in fact, infinitely many τ_n).*

PROOF. Define τ recursively. At stage m define it so that it agrees with infinitely many of the τ_n through stage m. An easy limiting argument shows τ to be optimal. □

PROPOSITION 1.1. *For any bandit there is an optimal strategy that is independent of observations obtained by pulling $\mathfrak{L}$.*

The existence of an optimal strategy is a consequence of Lemma 1.1. That it may be chosen to be independent of observations on $\mathfrak{L}$ is an immediate consequence of the following proposition, whose proof we omit, and which says that an optimal strategy need only depend on the history through the current distribution of ρ (conditioned by the history). This principle, familiar to researchers in sequential decision theory, is valid in a variety of contexts; ours is a special case of the one discussed by DeGroot (1970, Section 14.5).

PROPOSITION 1.2. *Let τ be an optimal strategy for the $((\alpha_1, \alpha_2, \cdots), R, \lambda)$-bandit. For any history having positive probability under τ and containing exactly s successes and f failures on $\mathfrak{R}$ for stages 1 through $m-1 (s+f \leqslant m-1)$ it is optimal to pull an arm at stage m if and only if it is optimal to pull that arm at stage 1 for the $((\alpha_m, \alpha_{m+1}, \cdots), \sigma^s\varphi^f R, \lambda)$-bandit.*

REMARK. When appropriate we shall say which results can be generalized to bandit problems in which the distributions of the X_i's and Y_j's are arbitrary—not necessarily Bernoulli. Then we can view ρ as a random distribution rather than as a random variable; conditioned on ρ, the sequence $(X_1, X_2, \cdots)$ is independent, the conditional distribution of each X_i being ρ. The measure R is a probability measure on the space of probability measures on the real line. Reasonable hypotheses are that R is supported by a set of distributions ρ, each of which has a finite mean μ_ρ, and that $\int |\mu_\rho| R(d\rho) < \infty$. While Lemma 1.1 applies only for discrete variables, with appropriate natural changes, Propositions 1.1 and 1.2 are true in this more general context.

The problem as described has been called the "*two-armed bandit* with one arm known." All such problems considered in the literature known to us are "optimal stopping problems": one only need consider strategies which never follow a pull of $\mathfrak{L}$ with a pull of $\mathfrak{R}$. Such strategies are determined by the stage at which $\mathfrak{L}$ is first pulled. Therefore, such problems have also been called "*one-armed bandits*." We

show in Section 2 (Theorem 2.1) that this term is appropriate if and only if the discount sequence **A** is *regular*.

DEFINITION 1.1. A discount sequence $(\alpha_1, \alpha_2, \cdots)$ is *regular* if, for each m, $\gamma_m\gamma_{m+2} \leqslant \gamma_{m+1}^2$ where $\gamma_p = \sum_{i=p}^{\infty}\alpha_i$.

The sequence $(\gamma_1, \gamma_2, \cdots)$ is always nonincreasing. If $(\alpha_1, \alpha_2, \cdots)$ is regular then

$$\gamma_{m+2}/\gamma_{m+1} \leqslant \gamma_{m+1}/\gamma_m$$

whenever $\gamma_{m+1} > 0$, that is, the sequence $(\gamma_2/\gamma_1, \gamma_3/\gamma_2, \cdots)$ is also nonincreasing.

Bradt, Johnson and Karlin (Section 4, 1956) consider the regular discount sequence $(1, \cdots, 1, 0, \cdots)$. Much of Sections 2, 3 and 4 is a generalization of their work. Indeed, Theorem 2.1 allows us to use many of their methods for the more general situation considered here. Many of the current results also generalize results of Bellman (1956) who considers the regular sequence $(1, \alpha, \alpha^2, \cdots)$ for $0 \leqslant \alpha < 1$.

In Section 3 we explore relations between optimal strategies and the maximal expected payoffs for two bandits $(\mathbf{A}, R_1, \lambda)$ and $(\mathbf{A}, R_2, \lambda)$ on the basis of an order relation between R_1 and R_2. These relations are exploited in Section 4 to obtain a characterization of optimal strategies (Theorem 4.3). This characterization is then used to obtain explicit formulas for optimal strategies in various special cases. In Section 5 we obtain an explicit sufficient condition for the optimality of $\mathfrak{R}$ (Theorem 5.1) and also an explicit necessary condition for its optimality (Theorem 5.2).

Two references on the finite horizon, two-armed bandit, deserve mention here. Fabius and van Zwet (1970) characterize Bayes strategies and admissible strategies in the general case of dependent arms. Berry (1972) finds explicit solutions for many cases in which the arms are independent and proves the stay-on-a-winner rule for independent arms.

2. Regular sequences and one-armed bandits. Since pulls of $\mathfrak{R}$ give information about $\mathfrak{R}$ it is a common phenomenon that optimal strategies require switches from $\mathfrak{R}$ to $\mathfrak{L}$. The next example shows that $\mathfrak{L}$ may be uniquely optimal initially and $\mathfrak{R}$ uniquely optimal at the second stage. This example can serve as an aid in understanding the proof of Theorem 2.1.

EXAMPLE 2.1. Let $\mathbf{A} = (4, 1, 1, 0, \cdots)$, $\lambda = .6$, and $R(\{0\}) = R(\{1\}) = .5$. If $\mathfrak{R}$ is pulled initially and an optimal strategy is followed thereafter, the expected payoff is

$$.5(4 + 1 + 1) + .5(.6)(1 + 1) = 3.6.$$

If $\mathfrak{L}$ is pulled initially and an optimal strategy is followed thereafter, the expected payoff is

$$.6(4) + .5(1 + 1) + .5(.6) = 3.7.$$

Therefore, *the* optimal strategy is: "pull $\mathfrak{L}$, then $\mathfrak{R}$—pull $\mathfrak{R}$ again on a success and switch back to $\mathfrak{L}$ on a failure."

In this example the discount sequence **A** is not regular. The next proposition says that if the discount sequence is not regular there are always examples of the same sort as Example 2.1 and if the sequence is regular there are no such examples. Accordingly, the term "one-armed bandit" is appropriate precisely when the sequence is regular since the problem can then be reduced to deciding when to stop pulling $\mathfrak{R}$.

THEOREM 2.1. *The following statements are equivalent*:

(i) *For every R and λ there is an optimal strategy under which every pull of $\mathfrak{L}$ is followed by another pull of $\mathfrak{L}$*;

(ii) *The discount sequence is regular.*

REMARK. Theorem 2.1 can be generalized to accommodate distributions on $\mathfrak{R}$ other than Bernoulli.

PROOF. Part I, (ii) $\Rightarrow$ (i). Let $\mathfrak{S}_n$ denote the set of all regular discount sequences $(\alpha_1, \alpha_2, \cdots)$ satisfying the condition $\alpha_{n+1} = 0$. The proof is by induction on n. Clearly, (i) holds for every member of $\mathfrak{S}_1$. Assume it holds for every member of $\mathfrak{S}_{n-1}$. Let $\mathbf{A} = (\alpha_1, \alpha_2, \cdots) \in \mathfrak{S}_n$; then $(\alpha_2, \alpha_3, \cdots) \in \mathfrak{S}_{n-1}$. Therefore, if it is optimal to pull $\mathfrak{R}$ initially, then, by the inductive hypothesis, there is an optimal continuation which never switches back to $\mathfrak{R}$ after a switch to $\mathfrak{L}$. If it is optimal to pull $\mathfrak{L}$ initially, then the inductive hypothesis applies immediately to show (i) unless the corresponding optimal strategy has the form τ^*: "pull $\mathfrak{L}$ initially, pull $\mathfrak{R}$ at stages $2, \cdots, N$, and pull $\mathfrak{L}$ subsequently."

The stage N is random with $P(N > 1) = 1$, it may be infinite with positive probability, and it may depend on the history of pulls and observations. Since, by Proposition 1.1, we may assume that τ^* does not depend on the observations on $\mathfrak{L}$, $\{N \geqslant m\}$ is, for each m, measurable with respect to the σ-field generated by the outcomes of the pulls of $\mathfrak{R}$ at stages 2 through m, that is, the σ-field generated by $(X_1, \cdots, X_{m-1})$. We may assume, with no loss of generality, that, under τ^*, if s successes and $f = m - s - 1$ failures have been obtained with $\mathfrak{R}$ at stages 2 through m, then

$$N = m \Rightarrow E(\rho | \sigma^s \varphi^f R) \leqslant \lambda. \tag{2.1}$$

We show that there is a strategy τ which starts with $\mathfrak{R}$ and is at least as good as τ^*. We choose τ by modifying τ^* in the following way: "pull $\mathfrak{R}$ initially and imitate τ^* subsequently by pulling the indicated arm one stage earlier."

Even though $\alpha_{n+1} = 0$, the notational conventions $\Sigma_{m=\infty}^{\infty} \alpha_m = \alpha_\infty = \gamma_\infty = 0$ will be useful. Under τ^* the expected payoff equals

$$E\left(\lambda\alpha_1 + \Sigma_{m=2}^{N} X_{m-1}\alpha_m + \lambda\Sigma_{m=N+1}^{\infty}\alpha_m\right), \tag{2.2}$$

which, since τ^* is optimal, is no smaller than $\lambda\gamma_1$. Hence,

$$\Sigma_{m=2}^{\infty} E\big[(X_{m-1} - \lambda)\mathbf{1}_{\{N>m\}}\big]\alpha_m = E\Sigma_{m=2}^{N}(X_{m-1} - \lambda)\alpha_m \geqslant 0. \tag{2.3}$$

The expected payoff under τ equals

$$E\big(\Sigma_{m=2}^{N} X_{m-1}\alpha_{m-1} + \lambda\Sigma_{m=N+1}^{\infty}\alpha_{m-1}\big). \tag{2.4}$$

We subtract (2.2) from (2.4) to obtain

$$\Sigma_{m=2}^{\infty} E\big[(X_{m-1} - \lambda)\mathbf{1}_{\{N>m\}}\big](\alpha_{m-1} - \alpha_m). \tag{2.5}$$

We shall use (2.3) to show that (2.5) is nonnegative by showing that

$$\Sigma_{m=2}^{\infty} b_m\alpha_m \geqslant 0 \Rightarrow \Sigma_{m=2}^{\infty} b_m(\alpha_{m-1} - \alpha_m) \geqslant 0, \tag{2.6}$$

where

$$b_m = E\big[(X_{m-1} - \lambda)\mathbf{1}_{\{N>m\}}\big].$$

We write the sums in (2.6) as follows:

$$\Sigma_{m=2}^{\infty} b_m\alpha_m = \Sigma_{m=2}^{\infty} b_m(\gamma_m - \gamma_{m+1}) = b_2\gamma_2 + \Sigma_{m=2}^{\infty}(b_{m+1} - b_m)\gamma_{m+1};$$

$$\Sigma_{m=2}^{\infty} b_m(\alpha_{m-1} - \alpha_m) = b_2\alpha_1 + \Sigma_{m=2}^{\infty}(b_{m+1} - b_m)\alpha_m.$$

The truth of (2.6) follows from two facts. The first is immediate for regular discount sequences: $\gamma_{m+1} = 0$ or $\alpha_1/\gamma_2 \leqslant \alpha_m/\gamma_{m+1}$, $m = 2, 3, \cdots$. The second is that the sequence $(b_1, b_2, \cdots)$ is nondecreasing. To show this write

$$\begin{aligned} b_{m+1} - b_m &= E\big[(\lambda - X_m)\mathbf{1}_{\{N=m\}}\big] + E\big[(X_m - X_{m-1})(1 - \mathbf{1}_{\{N<m\}})\big] \\ &= E\big[(\lambda - X_m)\mathbf{1}_{\{N=m\}}\big] \geqslant 0. \end{aligned}$$

We have used the exchangeability of the X_i's and the fact that $\{N < m\}$ is measurable with respect to the σ-field generated by $(X_1, \cdots, X_{m-2})$ for the second equality and (2.1) for the inequality.

We have proved (i) for any discount sequence belonging to $\cup_{n=1}^{\infty}\mathcal{S}_n$. That (i) holds for every regular discount sequence follows from Lemma 1.1.

Part II, (i) $\Rightarrow$ (ii). Suppose that

$$\gamma_M\gamma_{M+2} > \gamma_{M+1}^2 \tag{2.7}$$

for some M. We shall prove the result by finding a pair (R, λ) for which there is a strategy τ that follows a pull of $\mathcal{L}$ with a pull of $\mathcal{R}$ (on a history that has positive probability under τ) and which is strictly better than every strategy that does not.

Choose $\varepsilon > 0$ so that

$$\frac{\varepsilon\gamma_{M+1}}{\gamma_{M+1} - \gamma_{M+2}} \leqslant \frac{\alpha_M - \varepsilon\alpha_{M+1}}{2\alpha_M - (1 + \varepsilon)\alpha_{M+1}} \tag{2.8}$$

and

$$\frac{\gamma_{M+1} - \varepsilon\gamma_{M+2}}{2\gamma_{M+1} - (1 + \varepsilon)\gamma_{M+2}} \leqslant \frac{1}{1 + \varepsilon}. \tag{2.9}$$

Inequality (2.8) is always possible since $\alpha_{M+1} = \gamma_{M+1} - \gamma_{M+2} > 0$, a consequence of $\gamma_{M+2} > 0$. Select R so that the probability that $\rho = 0$ and $\rho = 1 - \varepsilon$ are both $\frac{1}{2}$

after $M-1$ failures on $\mathfrak{R}$; that is,

$$(\varphi^{M-1}R)(\{0\}) = (\varphi^{M-1}R)(\{1-\varepsilon\}) = \tfrac{1}{2}.$$

This can be accomplished by setting

$$R(\{1-\varepsilon\}) = \frac{1}{1+\varepsilon^{M-1}} = 1 - R(\{0\}).$$

For reasons that will soon be made clear, we choose λ so that

$$\frac{[1-\varepsilon][\alpha_M - \varepsilon\alpha_{M-1}]}{2\alpha_M - (1+\varepsilon)\alpha_{M+1}} < \lambda < \frac{[1-\varepsilon][\gamma_{M+1} - \varepsilon\gamma_{M+2}]}{2\gamma_{M+1} - (1-\varepsilon)\gamma_{M+2}}. \tag{2.10}$$

That λ can be so chosen is a consequence of (2.7).

Defined thus, $\lambda < 1-\varepsilon$. Therefore, since $(\sigma R)(\{1-\varepsilon\}) = 1$, we may restrict attention to strategies under which a success with $\mathfrak{R}$ is followed indefinitely by pulls of $\mathfrak{R}$. Among such strategies the only ones having the property that no pull of $\mathfrak{L}$ is followed by a pull of $\mathfrak{R}$ are the strategies τ_J, $J = 0, 1, \cdots, \infty$: "pull $\mathfrak{R}$ at the first J stages; thereafter pull $\mathfrak{R}$ or $\mathfrak{L}$ according as a success was or was not obtained at one or more of the first J stages." Let τ denote the strategy: "pull $\mathfrak{R}$ at the first $M-1$ stages; continue indefinitely with $\mathfrak{R}$ thereafter if a success was obtained at one or more of the first $M-1$ stages, and, if not, pull $\mathfrak{L}$ at stage M, pull $\mathfrak{R}$ at stage $M+1$, and thereafter pull $\mathfrak{R}$ or $\mathfrak{L}$ indefinitely according as a success was or was not obtained at stage $M+1$."

We shall show that τ is better than every τ_J. A straightforward calculation shows that τ is better than τ_{M-1} if and only if the second inequality in (2.10) holds. That τ_{M-1} is at least as good as τ_J for $J < M-1$ follows from $\lambda < (1-\varepsilon)/(1+\varepsilon)$, a consequence of (2.9) and the second inequality of (2.10). A calculation involving the first inequality in (2.10) shows τ to be better than τ_M, which is, according to a calculation using $\lambda > (1-\varepsilon)\varepsilon\gamma_{M+1}/(\gamma_{M+1} - \gamma_{M+2})$ (see (2.8) and (2.10)), at least as good as τ_J for $J > M$. (Depending on $\mathbf{A}$, τ may be optimal; we have only shown that it is better than every τ_J.) □

On the basis of (i), and, therefore, (ii), in Theorem 2.1 we can generalize results of Bradt, Johnson and Karlin (Section 4, 1956) who considered the special case where, for some n, $\alpha_m = 1$ or $\alpha_m = 0$ according as $m \leqslant n$ or $m > n$. The next theorem is the first step in this program.

THEOREM 2.2. *For each probability distribution R on $[0, 1]$ and each regular discount sequence $\mathbf{A}$, not identically 0, there exists a $\Lambda(\mathbf{A}, R) \in [0, 1]$ such that the only optimal initial actions are "pull $\mathfrak{R}$ if $\lambda \leqslant \Lambda(\mathbf{A}, R)$" and "pull $\mathfrak{L}$ if $\lambda \geqslant \Lambda(\mathbf{A}, R)$."*

REMARK. Theorem 2.2 can be generalized to accommodate distributions R other than Bernoulli.

PROOF. Let λ_1 and λ_2 be two values of λ with $\lambda_2 < \lambda_1$. Suppose that an optimal strategy specifies a pull of $\mathfrak{R}$ at the first stage when $\lambda = \lambda_1$. We want to show that no optimal strategy specifies a pull of $\mathfrak{L}$ at the first stage when $\lambda = \lambda_2$.

Let τ_1 be an optimal strategy when $\lambda = \lambda_1$ that pulls $\mathcal{R}$ at stage 1. Let N denote the last stage at which $\mathcal{R}$ is pulled using τ_1; $N = \infty$ if there is no such stage. Since τ_1 is no worse than pulling $\mathcal{L}$ at every stage,

$$V(R, \lambda_1) \geqslant \lambda_1 \Sigma_1^\infty \alpha_m.$$

Let $V^*(R, \lambda_2)$ denote the expected payoff from using τ_1 when $\lambda = \lambda_2$; clearly,

$$V(R, \lambda_2) \geqslant V^*(R, \lambda_2).$$

We have:

$$0 \leqslant V(R, \lambda_1) - \lambda_1 \Sigma_1^\infty \alpha_m = E\big[(\rho - \lambda_1)\Sigma_1^N \alpha_m\big] \tag{2.11}$$

$$< E\big[(\rho - \lambda_2)\Sigma_1^N \alpha_m\big] = V^*(R, \lambda_2) - \lambda_2 \Sigma_1^\infty \alpha_m,$$

and, therefore,

$$V(R, \lambda_2) > \lambda_2 \Sigma_1^\infty \alpha_m. \tag{2.12}$$

Strict inequality holds in (2.11), and therefore in (2.12), since, under τ_1, $P(N \geqslant 1) = 1$. If $\mathcal{L}$ is an optimal first pull when $\lambda = \lambda_2$ then, from Theorem 2.1, an optimal strategy is to pull $\mathcal{L}$ at every stage. But inequality (2.12) shows that this strategy cannot be optimal. □

The function defined in Theorem 2.2 completely determines the set of optimal strategies when the discount sequence is regular. If $\lambda < \Lambda(\mathbf{A}, R)$ then $\mathcal{R}$ is uniquely optimal initially; if $\lambda > \Lambda$ then $\mathcal{L}$ is uniquely optimal; and if $\lambda = \Lambda$ then both $\mathcal{R}$ and $\mathcal{L}$ are optimal initially. At the second stage λ is similarly compared to $\Lambda((\alpha_2, \alpha_3, \cdots), R)$, $\Lambda((\alpha_2, \alpha_2, \cdots), \sigma R)$, or $\Lambda((\alpha_2, \alpha_3, \cdots), \varphi R)$ according as $\mathcal{L}$ was pulled initially, $\mathcal{R}$ was pulled initially yielding success, or $\mathcal{R}$ was pulled initially yielding failure; and so on for subsequent stages.

A consequence of Theorem 2.1 is that

$$\Lambda((\alpha_2, \alpha_3, \cdots), R) \leqslant \Lambda((\alpha_1, \alpha_2, \cdots), R)$$

for all R and regular $(\alpha_1, \alpha_2, \cdots)$. The continuity of Λ is asserted in Theorem 5.3. A "weak type of convexity" holds for it with respect to each of its variables. If $p \in [0, 1)$, then

$$\Lambda(\mathbf{A}, pR_1 + (1 - p)R_2) \leqslant \Lambda(\mathbf{A}, R_1) \vee \Lambda(\mathbf{A}, R_2);$$

strict inequality is possible. A similar statement holds for a convex combination $p\mathbf{A}_1 + (1 - p)\mathbf{A}_2$ of regular discount sequences provided that $p\mathbf{A}_1 + (1 - p)\mathbf{A}_2$ is also regular. The easy proofs of these assertions depend on the convexity properties of V mentioned in Section 1. It can also be proved that the function

$$c \to \Lambda((\alpha_1, \cdots, \alpha_n, c\alpha_{n+1}, c\alpha_{n+2}, \cdots), R)$$

is nondecreasing. Monotonicity with respect to R is studied in Section 3.

3. Comparing bandits with common left arm. Theorem 2.2 indicates that the desire to pull $\mathcal{L}$ is not decreased by an increase in λ. It seems reasonable to expect a similar result regarding a pull of $\mathcal{R}$ and the random variable ρ. One way of

"increasing" a random variable is to make its distribution function smaller. This concept is embodied in the following definition.

DEFINITION 3.1. A distribution function R_1, is *to the right of* a distribution function R_2 if $R_1(x) \leqslant R_2(x)$ for every x.

REMARK. If two random variables are defined with distribution functions R_1 and R_2 with R_1 to the right of R_2, then the first random variable is said to be *stochastically larger than* the second.

The following examples show that pulling $\mathcal{R}$ may be optimal for the (R_2, λ)-bandit but not for the (R_1, λ)-bandit even though R_1 is to the right of R_2.

EXAMPLE 3.1. Let $\mathbf{A} = (1, \alpha, \alpha^2, \alpha^3, \cdots)$ and $R_2(\{0\}) = R_2(\{\frac{3}{4}\}) = R_1(\{\frac{1}{4}\}) = R_1(\{\frac{3}{4}\}) = \frac{1}{2}$. Clearly, R_1 is to the right of R_2. It will be seen (by application of Examples 4.2 and 4.4) that for these two bandits,

$$\Lambda(\mathbf{A}, R_2) = \frac{3(4 - \alpha)}{4(8 - 5\alpha)}$$

and

$$\Lambda(\mathbf{A}, R_1) = \frac{\left[1 + (1 - 3\alpha^2/4)^{\frac{1}{2}}\right] - 3\alpha/4}{2\left[1 + (1 - 3\alpha^2/4)^{\frac{1}{2}}\right] - 2\alpha}.$$

As $\alpha \uparrow 1$, both expressions approach $\frac{3}{4}$ and their respective derivatives approach 1 and $\frac{3}{2}$. Therefore,

$$\Lambda(\mathbf{A}, R_1) < \Lambda(\mathbf{A}, R_2)$$

for α sufficiently large. For such an α, select $\lambda \in (\Lambda(R_1), \Lambda(R_2))$. Then, for the $(\mathbf{A}, R_1, \lambda)$-bandit it is optimal to pull $\mathcal{L}$ at every stage and, therefore, $V(\mathbf{A}, R_1, \lambda) = \lambda/(1 - \alpha)$. For the $(\mathbf{A}, R_2, \lambda)$-bandit $\mathcal{R}$ is uniquely optimal initially, and so $V(\mathbf{A}, R_2, \lambda) > \lambda/(1 - \alpha)$.

EXAMPLE 3.2. The distributions R_1 and R_2 in Example 3.1 provide a counterexample in the finite horizon case as well. Let $\alpha_1 = \cdots = \alpha_{12} = 1$ and $\alpha_{13} = \cdots = 0$ and $\lambda = \frac{2}{3}$. Then $\mathcal{L}$ is uniquely optimal for $(\mathbf{A}, R_1, \lambda)$ and $\mathcal{R}$ is uniquely optimal initially for $(\mathbf{A}, R_2, \lambda)$.

While R_1 is to the right of R_2 in these examples, this inequality is not necessarily preserved after $\mathcal{R}$ is pulled. In fact, in these examples success on $\mathcal{R}$ reverses the inequality; σR_2 is to the right of σR_1. The next definition strengthens the concept of "to the right of" to include all possible measures derived from R_1 and R_2 by pulls on $\mathcal{R}$.

DEFINITION 3.2. Let R_1 and R_2 denote distribution functions on $[0, 1]$. R_1 is *strongly to the right of* R_2 if $\sigma^s\varphi^f R_1$ is to the right of $\sigma^s\varphi^f R_2$ for every pair (s, f) of nonnegative integers for which $\sigma^s\varphi^f R_1$ and $\sigma^s\varphi^f R_2$ are defined.

The following two easy lemmas are given without proof. The first lemma says that the probability of success is at least as large on R_1 as on R_2 when R_1 is to the right of R_2. The second says that "strongly to the right" is preserved by a pull of $\mathfrak{R}$.

LEMMA 3.1. *If R_1 is to the right of R_2 then $E(\rho|R_1) \geqslant E(\rho|R_2)$.*

LEMMA 3.2. *For any R, σR is strongly to the right of φR. If R_1 is strongly to the right of R_2 then σR_1 is strongly to the right of both σR_2 and φR_2 and φR_1 is strongly to the right of φR_2.*

Though we will apply the next theorem only for $R_1 = \sigma R$ and $R_2 = \varphi R$ we will prove it in its full generality. (For a version of Theorem 3.1 for many-armed bandits see (Berry and Fristedt, 1979).)

THEOREM 3.1. *Suppose the distribution function R_1 is strongly to the right of R_2. Then, for each regular discount sequence* **A**, *not identically* 0, *and each* $\lambda \in [0, 1]$, $V(\mathbf{A}, R_1, \lambda) \geqslant V(\mathbf{A}, R_2, \lambda)$ *and* $\Lambda(\mathbf{A}, R_1) \geqslant \Lambda(\mathbf{A}, R_2)$.

PROOF. As in Part I of the proof of Theorem 2.1, let

$$\mathcal{S}_n = \{\mathbf{A}: \mathbf{A} \text{ is regular and } \alpha_{n+1} = 0\}.$$

The major part of the proof is the use of induction on n to prove that $V(\mathbf{A}, R_1) \geqslant V(\mathbf{A}, R_2)$ for every $\mathbf{A} \in \cup_{n=1}^{\infty} \mathcal{S}_n$ and every R_1 and R_2 for which R_1 is strongly to the right of R_2.

By Lemma 3.1, $V(\mathbf{A}, R_1) \geqslant V(\mathbf{A}, R_2)$ for $\mathbf{A} \in \mathcal{S}_1$ since $V(\mathbf{A}, R_i) = E(\rho|R_i)$. For the remainder of the induction assume $\mathbf{A} \in \mathcal{S}_n$.

Write $\mathbf{A}^{(1)} = (\alpha_2, \alpha_3, \cdots)$. By the induction hypothesis and the fact that $\mathbf{A}^{(1)} \in \mathcal{S}_{n-1}$,

$$V(\mathbf{A}^{(1)}, T_1) \geqslant V(\mathbf{A}^{(1)}, T_2) \tag{3.1}$$

whenever T_1 is strongly to the right of T_2.

If $\mathfrak{L}$ is optimal for the $(\mathbf{A}, R_2)$-bandit, then $V(\mathbf{A}, R_2) = \lambda\gamma_1 \leqslant V(\mathbf{A}, R_1)$. Accordingly, we assume that $\mathfrak{R}$ is initially optimal for $(\mathbf{A}, R_2)$ and obtain (cf. Proposition 1.2)

(3.2)

$$V(\mathbf{A}, R_2) = \alpha_1 E(\rho|R_2) + E(\rho|R_2) V(\mathbf{A}^{(1)}, \sigma R_2) + [1 - E(\rho|R_2)] V(\mathbf{A}^{(1)}, \varphi R_2).$$

One strategy for $(\mathbf{A}, R_1)$ is to pull $\mathfrak{R}$ initially and thereafter act optimally. Thus,

(3.3)

$$V(\mathbf{A}, R_1) \geqslant \alpha_1 E(\rho|R_1) + E(\rho|R_1) V(\mathbf{A}^{(1)}, \sigma R_1) + [1 - E(\rho|R_1)] V(\mathbf{A}^{(1)}, \varphi R_1).$$

Subtracting (3.2) from (3.3) yields

$$
\begin{aligned}
V(\mathbf{A}, R_1) - V(\mathbf{A}, R_2) \geqslant\ & \alpha_1\big[E(\rho|R_1) - E(\rho|R_2)\big] \\
& + E(\rho|R_2)\big[V(\mathbf{A}^{(1)}, \sigma R_1) - V(\mathbf{A}^{(1)}, \sigma R_2)\big] \\
& + \big[1 - E(\rho|R_1)\big]\big[V(\mathbf{A}^{(1)}, \varphi R_1) - V(\mathbf{A}^{(1)}, \varphi R_2)\big] \\
& + \big[E(\rho|R_1) - E(\rho|R_2)\big]\big[V(\mathbf{A}^{(1)}, \sigma R_1) - V(\mathbf{A}^{(1)}, \varphi R_2)\big].
\end{aligned} \tag{3.4}
$$

By Lemma 3.1, the first term on the right side of (3.4) is nonnegative; it represents the immediate advantage at stage 1 of the R_1-bandit over the R_2-bandit provided that $\mathfrak{R}$ is pulled at stage 1. By Lemma 3.2, (3.1) is valid for these choices of (T_1, T_2): $(\sigma R_1, \sigma R_2)$, $(\varphi R_1, \varphi R_2)$, and $(\sigma R_1, \varphi R_2)$; hence, the last three terms in (3.4) are nonnegative and the induction is complete. (The last three terms in (3.4) represent the advantage for stages after the first of the R_1-bandit over the R_2-bandit when $\mathfrak{R}$ is pulled initially for both. The term weighted by $E(\rho|R_2)$ is the advantage resulting from an initial success with both bandits. The term weighted by $1 - E(\rho|R_1)$ is the advantage resulting from an initial failure with both bandits. The last term is weighted by the probability of an initial success with the R_1-bandit and an initial failure with the R_2-bandit. This interpretation requires that the two bandits be placed on a common probability space so that an initial success with the R_2-bandit entails an initial success with the R_1-bandit as well.)

That $V(\mathbf{A}, R_1) \geqslant V(\mathbf{A}, R_2)$ for every regular $\mathbf{A}$ follows from Lemma 1.1.

The second conclusion, $\Lambda(\mathbf{A}, R_1) \geqslant \Lambda(\mathbf{A}, R_2)$, follows from the first conclusion and Theorem 2.1. □

REMARK. A slight modification of the proof of Theorem 3.1 shows that $V(\mathbf{A}, R_1, \lambda) \geqslant V(\mathbf{A}, R_2, \lambda)$ even if $\mathbf{A}$ is not regular.

4. The nature of optimal strategies. Bradt et al. (1956) prove a *stay-with-a-winner-rule* for the finite horizon one-armed bandit. The following theorem generalizes that result to arbitrary regular discount sequences. It says that whenever a success is obtained with $\mathfrak{R}$ while using an optimal strategy, then a pull of $\mathfrak{R}$ is optimal at the next stage as well—uniquely optimal unless the prior distribution R concentrates its mass on λ.

THEOREM 4.1. *Suppose that* $\mathbf{A} = (\alpha_1, \alpha_2, \cdots)$ *is a regular discount sequence with* $\alpha_2 > 0$. *Let* $\mathbf{A}^{(1)} = (\alpha_2, \alpha_3, \cdots)$. *Then, for all* R,

$$
\Lambda(\mathbf{A}, R) \leqslant \Lambda(\mathbf{A}^{(1)}, \sigma R), \tag{4.1}
$$

with equality only if R *is a one-point distribution.*

PROOF. We will show that $\lambda \leqslant \Lambda(\mathbf{A}, R) \Rightarrow \lambda \leqslant \Lambda(\mathbf{A}^{(1)}, \sigma R)$; that is, if $\mathfrak{R}$ is optimal for $(\mathbf{A}, R, \lambda)$, then it is also optimal for $(\mathbf{A}^{(1)}, \sigma R, \lambda)$. Consider two cases:

(i) $\lambda < E(\rho|\sigma R)[= E(\rho^2|R)/E(\rho|R)]$,

(ii) $\lambda \geqslant E(\rho|\sigma R)$.

In case (i) we assume $\Lambda(\mathbf{A}^{(1)}, \sigma R) < \lambda$. Then $V(\mathbf{A}^{(1)}, \sigma R) = \gamma_2\lambda$ by Theorem 2.1. But the expected payoff of the strategy that pulls $\mathfrak{R}$ at every stage for $(\mathbf{A}^{(1)}, \sigma R)$ is $\gamma_2 E(\rho|\sigma R) > V(\mathbf{A}^{(1)}, \sigma R)$, which is a contradiction.

In case (ii) assume $\lambda > \Lambda(\mathbf{A}^{(1)}, \sigma R)$. By Theorem 3.1, since σR is strongly to the right of φR for all R, $\lambda > \Lambda(\mathbf{A}^{(1)}, \varphi R)$ as well. Theorem 2.1 applies to show that $\mathfrak{L}$ is optimal at all stages subsequent to the first, independent of the initial result. Therefore, $V(\mathbf{A}, R) = \alpha_1 E(\rho|R) + \gamma_2\lambda$. But the expected payoff of the strategy that pulls $\mathfrak{L}$ at every stage, including the first, is $\gamma_1\lambda = \alpha_1\lambda + \gamma_2\lambda \geqslant \alpha_1 E(\rho|\sigma R) + \gamma_2\lambda > \alpha_1 E(\rho|R) + \gamma_2\lambda$, provided R is not one-point. □

The complementary result,

$$\Lambda(\mathbf{A}^{(1)}, \varphi R) \leqslant \Lambda(\mathbf{A}, R) \tag{4.2}$$

holds for all R and every regular $\mathbf{A}$ for which $\alpha_2 \neq 0$. For, in view of Theorem 3.1, $\Lambda(\mathbf{A}^{(1)}, \varphi R) \leqslant \Lambda(\mathbf{A}^{(1)}, R)$, and, in view of Theorem 2.1, $\Lambda(\mathbf{A}^{(1)}, R) \leqslant \Lambda(\mathbf{A}, R)$. Inequalities (4.1) and (4.2) generalize results of Bellman (1956, Theorem 2) who considered $\mathbf{A} = (1, \alpha, \alpha^2, \cdots)$.

The next theorem specifies a class of strategies among which there is at least one optimal strategy when $\mathbf{A}$ is regular. It is an immediate consequence of the preceding development.

THEOREM 4.2. *Suppose that the discount sequence* $\mathbf{A}$ *is regular and not identically* 0. *If* $\lambda \geqslant \Lambda(\mathbf{A}, R)$ *it is optimal to pull* $\mathfrak{L}$ *at every stage. If* $\lambda \leqslant \Lambda(\mathbf{A}, R)$ *there is a sequence* $K = (k_1, k_2, \cdots)$, $k_i \in \{0, 1, \cdots, \infty\}$, *for which the following strategy is optimal. Pull* $\mathfrak{R}$ *until* i *failures have occurred,* $i = 1, 2, \cdots$, *then switch to* $\mathfrak{L}$ *at the next stage (and all subsequent stages) if and only if the total number of successes obtained thus far is less than* $k_1 + \cdots + k_i$.

Not all optimal strategies are described in Theorem 4.2. We will now mention several other possibilities. If $\alpha_m = 0$, then every continuation from stage m is optimal. Similarly, if R is concentrated at λ, then every strategy is optimal. The only interesting optimal strategies not described in Theorem 4.2 occur when $\gamma_M\gamma_{M+2} = \gamma_{M+1}^2 > 0$ for some M. From Part II of the proof of Theorem 2.1 we see that it may also be optimal to pull $\mathfrak{R}$ at stage $M - 1$, $\mathfrak{L}$ at stage M, and $\mathfrak{R}$ at stage $M + 1$.

Henceforth, besides the strategy "pull $\mathfrak{L}$ at every stage," we will only consider strategies that correspond to a sequence K. The next theorem gives an expression for the "break-even" value of λ.

THEOREM 4.3. *For every* R *and regular discount sequence* $\mathbf{A}$ *not identically* 0,

$$\Lambda(\mathbf{A}, R) = \max_K \left\{ \frac{E_K \Sigma_{m=1}^N X_m \alpha_m}{E_K \Sigma_{m=1}^N \alpha_m} \right\} \tag{4.3}$$

where E_K *denotes expectation with respect to the strategy corresponding to the sequence* K *and* N *equals the (random) number (possibly* $+\infty$*) of pulls of* $\mathfrak{R}$.

Moreover, if $\lambda = \Lambda(\mathbf{A}, R)$ *then those* K*'s which correspond to optimal strategies are exactly those for which the maximum in* (4.3) *is attained.*

PROOF. The proof of Theorem 4.1 in (Bradt et al. 1956) applies in this more general setting with evident modifications. □

While (4.3) is an expression for the break-even value of λ for an arbitrary one-armed bandit, it is not easy to apply. The number of sequences K that have to be checked, even using a clever searching algorithm, can be very large. The next four examples are applications of the theorems in this section for three kinds of restrictions on the distribution R in which the solutions can be given explicitly.

EXAMPLE 4.1. Suppose $\mathbf{A}$ is a regular discount sequence and the support of R is a subset of $[0, \lambda] \cup \{1\}$. Once a failure occurs with $\mathcal{R}$, a switch to $\mathcal{L}$ is optimal since $\varphi\sigma^s R([0, \lambda]) = 1$ for all s. According to Theorem 4.2 (or Theorems 2.1 and 4.1) we need only compare two strategies: "pull $\mathcal{L}$ at every stage" and "pull $\mathcal{R}$ until it fails (if ever) and $\mathcal{L}$ thereafter." The result of this comparison is that $\mathcal{L}$ is optimal if $\lambda \geqslant \lambda^*$ and $\mathcal{R}$ is optimal initially if $\lambda \leqslant \lambda^*$, where λ^* (a quantity that appears as λ^*_∞ in Theorem 5.1) is defined by:

$$\lambda^* = \frac{\sum_{m=1}^{\infty} \alpha_m E\rho^m}{\sum_{m=1}^{\infty} \alpha_m E\rho^{m-1}},$$

it being understood that $0^0 = 1$. The calculations of expected payoffs yield:

$$V(R, \lambda) = \lambda\gamma_1 \quad \text{if} \quad \lambda \geqslant \lambda^*$$
$$= \lambda\gamma_1 + \sum_{m=1}^{\infty} \alpha_m E(\rho^m - \lambda\rho^{m-1}) \quad \text{if} \quad \lambda \leqslant \lambda^*.$$

It follows that $\Lambda(\mathbf{A}, R) = \lambda^*$ if the support of R is a subset of $[0, \lambda^*] \cup \{1\}$. If on the other hand, the support of R contains a member of $(\lambda^*, 1)$, then the condition $\lambda \leqslant \lambda^*$ is incompatible with the condition that the support of R is a subset of $[0, \lambda] \cup \{1\}$ and, hence, all that can be immediately concluded about Λ is that the support of R contains a member of $(\Lambda, 1)$. However, from using $K = (\infty, 0, 0, \cdots)$ in Theorem 4.3 it follows that $\Lambda \geqslant \lambda^*$ with equality holding if and only if the support of R is a subset of $[0, \lambda^*] \cup \{1\}$. Accordingly, characterizing optimal strategies using only Theorem 4.3 avoids calculating expected payoffs under the two strategies. The necessary and sufficient condition for $\Lambda = \lambda^*$ appears again in Theorem 5.1.

EXAMPLE 4.2. Suppose $\mathbf{A}$ is a regular discount sequence, the support of R is a subset of $\{0\} \cup [\lambda, 1]$, and $R(\{0\}) > 0$. Among those strategies described in Theorem 4.2, the only ones we need consider are τ_M, $M = 0, 1, \cdots, \infty$, where: τ_0 denotes "pull $\mathcal{L}$ at every stage"; τ_M for $0 < M < \infty$ denotes the strategy given by $K = (0, \cdots, 0, 1, 0, \cdots)$, the Mth coordinate equalling 1; and τ_∞ denotes the strategy "pull $\mathcal{R}$ at every stage" given by $K = (0, 0, \cdots)$. The expected payoff under τ_M is

$$E\{\rho\gamma_1 + (1 - \rho)^M(\lambda - \rho)\gamma_{M+1}\}. \tag{4.4}$$

Since, as a function of M, (4.4) is continuous at ∞, it achieves a maximum. We shall show that (4.4) increases to its maximum and then decreases, the possibilities of a maximum at 0 or at ∞ or at more than one value of M not being excluded. We assume that the expected payoff under τ_{M+1} is larger than the expected payoff under τ_M and show that the expected payoff under τ_M is larger than the expected payoff under τ_{M-1}. That is, we show

$$E\{(\rho - \lambda)(1 - \rho)^M(\alpha_{M+1} + \rho\gamma_{M+2})\} > 0$$
$$\Rightarrow E\{(\rho - \lambda)(1 - \rho)^{M-1}(\alpha_M + \rho\gamma_{M+1})\} > 0.$$

This follows since, for regular discount sequences,

$$\frac{(1 - x)^M(\alpha_{M+1} + x\gamma_{M+2})}{(1 - x)^{M-1}(\alpha_M + x\gamma_{M+1})}$$

is a decreasing function of x.

Set the expected payoff under τ_M (from (4.4)) equal to the expected payoff under τ_{M-1}. Solving for λ gives, for $0 < M < \infty$,

$$\lambda_M = \frac{E\big[\rho(1 - \rho)^{M-1}(\alpha_M + \rho\gamma_{M+1})\big]}{E\big[(1 - \rho)^{M-1}(\alpha_M + \rho\gamma_{M+1})\big]}. \tag{4.5}$$

Define $\lambda_0 = 1$. The sequence $\{\lambda_M\}$ is monotonic and $\lim_{M\to\infty}\lambda_M = 0$. The above argument shows that, except for arbitrary continuations from stage m when $\alpha_m = 0$, the only optimal strategies of the type described in Theorem 4.2 are as follows: If $\lambda_M \geqslant \lambda \geqslant \lambda_{M+1}$ then pull $\mathcal{R}$ at each of the first M stages and thereafter pull $\mathcal{L}$ or $\mathcal{R}$ according as all failures occur or not in the first M stages. The maximal expected payoff is

$$V(\mathbf{A}, R, \lambda) = \gamma_1 E\rho + \gamma_{M+1}E\big((\lambda - \rho)(1 - \rho)^M\big).$$

Of course, if $\lambda = 0$ then τ_∞ is optimal; moreover $V(\mathbf{A}, R, 0) = \gamma_1 E\rho$. The preceding statements do not imply that the condition $\lambda_{M-1} \geqslant \lambda \geqslant \lambda_M$ for some M is necessarily compatible with the support of R being a subset of $\{0\} \cup [\lambda, 1]$. If these are compatible for $M = 1$ then our arguments show $\Lambda = \lambda_1$. If the support of R contains a member of $(0, \lambda_1)$, then, according to Theorem 4.3, $\Lambda > \lambda_1$.

If, instead of appealing to Theorem 4.2, Theorem 4.3 is applied directly, in case the support of R is a subset of $\{0\} \cup [\lambda_1, 1]$, we can conclude only that $\Lambda = \lambda_1$ or the support of R contains a member of $(0, \Lambda)$. However, we can rule out the latter possibility in the following way. Define $R_w = (1 - w)R + w\delta_0$, where δ_0 denotes a unit mass at 0, and use Theorem 3.1, Theorem 5.3, and (4.5) to deduce that $\Lambda(\mathbf{A}, R_w)$ is a decreasing, continuous function of w and λ_1 is, when defined with R_w playing the role of R, a strictly decreasing function of w.

The distributions R considered in this example are shown in Theorem 5.1 to be extreme in the sense that $\Lambda \geqslant \lambda_1$ for arbitrary R. In Theorem 5.1, λ_1 is denoted λ_1^*.

EXAMPLE 4.3. For some $c > 0$ and $\alpha \in [0, 1]$ suppose that $\mathbf{A}$ is the regular discount sequence $(c, c\alpha, c\alpha^2, \cdots)$. Also, suppose that R is supported by $\{\rho_1, \rho_2\}$ with $\rho_1 < \rho_2$ and let $R(\{\rho_2\}) = r$. Since $\mathbf{A}$ is proportional to $(c\alpha^n, c\alpha^{n+1}, \cdots)$ for all n,

$$\Lambda(\mathbf{A}, R) = \Lambda((c\alpha^n, c\alpha^{n+1}, \cdots), R), \qquad n > 0.$$

Therefore we suppress the dependence of Λ on $\mathbf{A}$. Since $\sigma^s\varphi^f R$ depends on R only through r (for ρ_1 and ρ_2 fixed), we write $\Lambda(r)$ for $\Lambda(R)$ and $\sigma^s\varphi^f r$ for $(\sigma^s\varphi^f R)(\{\rho_2\})$.

From Theorem 3.1 the function $r \to \Lambda(r)$ is nondecreasing. Therefore, when $\lambda = \Lambda(r)$ an optimal strategy is to pull $\mathfrak{R}$ at stage 1 and to keep pulling it until $\sigma^s\varphi^f r < r$, where s and f denote the current numbers of successes and failures, and then switch permanently to $\mathfrak{L}$. When $\lambda = \Lambda(r)$ it is also optimal to pull $\mathfrak{L}$ at stage 1 and thereafter. The expected payoff of the latter strategy is $\lambda\gamma_1$. To calculate the expected payoff of the former strategy directly we introduce some additional notation.

For $\rho, x \in [0, 1]$ let $S_n(\rho, x)$ denote the nth partial sum of a random walk, each of whose individual steps equals

$$1 - x \text{ with probability } \rho$$
$$-x \text{ with probability } 1 - \rho.$$

Let

$$N(\rho, x) = \inf\{n: S_n(\rho, x) < 0\}, \tag{4.6}$$

possibly $+\infty$. For $\alpha \in [0, 1)$, let

$$g(\alpha, \rho, x) = 1 - E\alpha^N, \tag{4.7}$$

where the dependence on ρ and x has been suppressed on the right side of (4.7). By a theorem of E. Sparre Andersen (Feller 1971, XII (7.3)),

$$g(\alpha, \rho, x) = \exp\left(-\Sigma_{n=1}^{\infty}\frac{\alpha^n}{n}P\{S_n < 0\}\right). \tag{4.8}$$

Using Bayes's theorem,

$$\sigma^s\varphi^f r = \left[\frac{\rho_1^s(1-\rho_1)^f(1-r)}{\rho_2^s(1-\rho_2)^f r} + 1\right]^{-1}.$$

Now define

$$x = \frac{\log[(1-\rho_1)/(1-\rho_2)]}{\log[(1-\rho_1)/(1-\rho_2)] + \log[\rho_2/\rho_1]}. \tag{4.9}$$

The following four statements are equivalent:

$$\sigma^s\varphi^f r < r;\ \rho_2^s(1-\rho_2)^f < \rho_1^s(1-\rho_1)^f;$$
$$\frac{s}{s+f} < x;\ (1-x)s + (-x)f < 0.$$

Accordingly, either $N(\rho_1, x)$ or $N(\rho_2, x)$, defined by (4.6), is the last stage at which $\mathcal{R}$ is pulled. Thus, the expected payoff under the first optimal strategy given above is

$$(1-r)E\left\{\rho_1\Sigma_{m=1}^{N(\rho_1,x)}c\alpha^{m-1}+\lambda\Sigma_{m=N(\rho_1,x)+1}^{\infty}c\alpha^{m-1}\right\}$$
$$+rE\left\{\rho_2\Sigma_{m=1}^{N(\rho_2,x)}c\alpha^{m-1}+\lambda\Sigma_{m=N(\rho_2,x)+1}^{\infty}c\alpha^{m-1}\right\}.$$

Equating this to $\lambda\gamma_1$ and solving gives

$$\Lambda(\mathbf{A}, R)=\frac{(1-r)\rho_1 g(\alpha,\rho_1,x)+r\rho_2 g(\alpha,\rho_2,x)}{(1-r)g(\alpha,\rho_1,x)+rg(\alpha,\rho_2,x)}, \tag{4.10}$$

where g is defined by (4.7), or by (4.8).

EXAMPLE 4.4. In addition to the conditions in Example 4.3 suppose that $\rho_2=1-\rho_1$. Then (4.9) gives $x=\frac{1}{2}$, and Feller (1968, XI (3.6)) gives the explicit formula:

$$g\left(\alpha,\rho,\tfrac{1}{2}\right)=1-\frac{1-\left[1-4\alpha^2\rho(1-\rho)\right]^{\frac{1}{2}}}{2\rho\alpha}.$$

With $r_2=r$ and $r_1=1-r$, (4.10) becomes

$$\Lambda(\mathbf{A}, R)=\frac{\left[\rho_1 r_1+\rho_2 r_2\right]\left[1+\left(1-4\alpha^2\rho_1\rho_2\right)^{\frac{1}{2}}\right]-2\alpha\rho_1\rho_2}{1+\left(1-4\alpha^2\rho_1\rho_2\right)^{\frac{1}{2}}-2\alpha(\rho_1 r_1+\rho_2 r_2)}.$$

5. Bounds for Λ. In this section we obtain lower and upper bounds for Λ, the break-even value of λ. We then prove that Λ and V are continuous.

Theorem 5.1 gives a sequence of lower bounds for Λ. This sequence may be increasing, constant, or decreasing, but, more typically, it increases to a maximum and then decreases. The quantity λ_∞^* defined below in (5.1) is also considered in Example 4.1 where it is denoted λ^*.

THEOREM 5.1. *Suppose that* $\mathbf{A}=(\alpha_1, \alpha_2, \cdots)$ *is a regular discount sequence, not identically* 0. *Then for each* $k \in \{1, 2, \cdots, \infty\}$, $\Lambda(\mathbf{A}, R) \geqslant \lambda_k^*$ *where*

$$\lambda_k^*=\frac{\Sigma_{m=1}^{\infty}\alpha_m E\rho^{m\wedge(k+1)}}{\Sigma_{m=1}^{\infty}\alpha_m E\rho^{(m-1)\wedge k}}. \tag{5.1}$$

In case $\alpha_m>0$ *for every* m, $\Lambda(\mathbf{A}, R)=\lambda_k^*$ *if and only if the support of* R *is a subset of* $\{0, 1\}$ *or* $k=1$ *and the support of* R *is a subset of* $\{0\}\cup[\lambda_1^*, 1]$ *or* $k=\infty$ *and the support of* R *is a subset of* $[0, \lambda_\infty^*]\cup\{1\}$. *In general, the sequence* $(\lambda_1^*, \lambda_2^*, \cdots)$ *achieves a maximum on a subinterval of* $\{1, 2, \cdots, \infty\}$.

PROOF. The inequality $\Lambda \geqslant \lambda_k^*$ follows from the use of $K=(k, 0, 0, \cdots)$ in (4.3). If $k=1$ and the support of R is a subset of $\{0\}\cup[\lambda_1^*, 1]$, then, according to the discussion in Example 4.2, $\Lambda=\lambda_1^*$. The other conditions for $\Lambda=\lambda_k^*$ follow easily from Theorem 4.3.

To prove the last assertion of the theorem we note the continuity at $k = \infty$, and, hence, that a maximum is attained. We will show

$$\lambda^*_{k-1} - \lambda^*_k \geqslant 0 \Rightarrow \lambda^*_k - \lambda^*_{k+1} \geqslant 0,$$

or equivalently,

$$(5.2)\quad E\{(\rho - \lambda^*_k)\Sigma\alpha_m\rho^{(m-1)\wedge(k-1)}\} \geqslant 0 \Rightarrow E\{(\rho - \lambda^*_k)\Sigma\alpha_m\rho^{(m-1)\wedge(k+1)}\} \geqslant 0.$$

Subtracting

$$E\{(\rho - \lambda^*_k)\Sigma\alpha_m\rho^{(m-1)\wedge k}\} = 0$$

from both of the expressions in (5.2) reduces the problem to showing that

$$(5.3)\qquad E\{(\rho - \lambda^*_k)\rho^{k-1}(1-\rho)\} \geqslant 0$$

implies

$$(5.4)\qquad E\{(\rho - \lambda^*_k)\rho^{k}(1-\rho)\} \geqslant 0.$$

Subtracting λ^*_k times the left side of (5.3) from the left side of (5.4) yields

$$E\{(\rho - \lambda^*)^2\rho^{k-1}(1-\rho)\} \geqslant 0. \qquad \square$$

The next theorem gives an upper bound for Λ. The result is not as readily applied as any of the lower bounds in Theorem 5.1 since the solution of equation (5.5) will usually require iteration.

THEOREM 5.2. *Suppose that* $\mathbf{A} = (\alpha_1, \alpha_2, \cdots)$ *is a regular discount sequence, not identically* 0. *Then* $\Lambda(\mathbf{A}, R)$ *is less than or equal to the unique solution in* [0, 1] *of the equation*

(5.5)

$$\lambda(\alpha_1 + \gamma_2 E\rho) - (\alpha_1 E\rho + \gamma_2 E\rho^2) + \Sigma_{m=3}^{\infty}\alpha_m E[\rho(\lambda - \rho)^+(1 - \rho^{m-2})] = 0.$$

PROOF. The left side of (5.5) is nonpositive when $\lambda = 0$, nonnegative when $\lambda = 1$, and continuous and strictly increasing as a function of λ. Accordingly, we need only show that the left side of (5.5) is nonpositive when $\lambda = \Lambda$.

We fix λ equal to Λ. Then, since one optimal strategy is "pull $\mathfrak{L}$ at every stage," the maximal expected payoff is $\lambda\gamma_1$. This expected payoff can also be achieved by an initial pull of $\mathfrak{R}$ and optimal pulls at stages 2, 3, $\cdots$. According to (4.2) it is optimal to pull $\mathfrak{L}$ at stages 2, 3, $\cdots$ in case a failure is obtained with $\mathfrak{R}$ at stage 1, and, according to the stay-with-a-winner rule (Theorem 4.1), it is optimal to keep pulling $\mathfrak{R}$ until a failure is obtained. Let τ denote this latter strategy. The conditional payoff using τ is

(5.6)

$$E_\tau(\Sigma_1^\infty\alpha_m Z^m|\rho) = (1-\rho)\lambda\gamma_2 + \Sigma_{n=1}^{\infty}\rho^n(1-\rho)[\Sigma_{m=1}^{n}\alpha_m + E_\tau(\Sigma_{m=n+2}^{\infty}\alpha_m Z_m|\rho)]$$

$$\leqslant (1-\rho)\lambda\gamma_2 + \Sigma_{n=1}^{\infty}\rho^n(1-\rho)[\Sigma_{m=1}^{n}\alpha_m + (\rho \vee \lambda)\gamma_{n+2}]$$

if $\rho < 1$ and it equals γ_1 if $\rho = 1$. The maximum of ρ and λ occurs in (5.6) for the

same reason it occurs in (1.1). In view of (5.6),

$$\lambda\gamma_1 \leqslant \lambda\gamma_2(1 - E\rho) + E\Sigma_{n=1}^{\infty}\rho^n(1 - \rho)\left[\Sigma_{m=1}^{n}\alpha_m + (\rho \vee \lambda)\gamma_{n+2}\right].$$

The result follows from straightforward calculations that are omitted. □

At stage 1, $\mathcal{R}$ is optimal if $\lambda \leqslant \lambda_k^*$ for some k and $\mathcal{L}$ is optimal if the left side of (5.5) is nonnegative.

EXAMPLE 5.1. Suppose the support of R is a subset of $\{0, 1\}$. Then Λ equals every lower bound in Theorem 5.1 and the upper bound in Theorem 5.2—namely, $\gamma_1 E\rho/(\alpha_1 + \gamma_2 E\rho)$, which is also obtainable as a special case of either Example 4.1 or Example 4.2.

EXAMPLE 5.2. Suppose that $\alpha_1 = \cdots = \alpha_n = 1$ and $\alpha_{n+1} = 0$. We give the lower bound offered by Bradt et al. (1956) via their special case of Theorem 5.1:

$$\Lambda(\mathbf{A}, R) \geqslant \frac{\Sigma_{m=1}^{n} E\rho^{m \wedge (k+1)}}{\Sigma_{m=1}^{n} E\rho^{(m-1)\wedge k}},$$

which for $k \geqslant n - 1$ becomes

$$\Lambda(\mathbf{A}, R) \geqslant \frac{E\{\rho(1 - \rho^n)/(1 - \rho)\}}{E\{(1 - \rho^n)/(1 - \rho)\}}.$$

Equation (5.5) for an upper bound on Λ becomes

$$\lambda\left[1 + (n - 1)E\rho\right] - \left[E\rho + (n - 1)E\rho^2\right] + E\left\{\rho(\lambda - \rho)^+\left[n - 2 - \rho(n - 1) + \rho^{n-1}\right]/(1 - \rho)\right\} = 0.$$

EXAMPLE 5.3. Suppose that $\mathbf{A} = (1, \alpha, \alpha^2, \cdots)$. The lower bounds for Λ described in Theorem 5.1 are

$$\lambda_k^* = \frac{E\left\{\rho\left[(1 - \alpha^k\rho^k)(1 - \alpha\rho)^{-1} + \alpha^k\rho^k(1 - \alpha)^{-1}\right]\right\}}{E\left\{(1 - \alpha^k\rho^k)(1 - \alpha\rho)^{-1} + \alpha^k\rho^k(1 - \alpha)^{-1}\right\}}.$$

In particular,

$$\Lambda \geqslant \lambda_\infty^* = \frac{\psi(\alpha) - 1}{\alpha\psi(\alpha)},$$

where the generating function ψ is defined by $\psi(\alpha) = E\{(1 - \alpha\rho)^{-1}\}$. Equation (5.5) for an upper bound becomes:

$$\lambda\left[(1 - \alpha) + \alpha E\rho\right] - \left[(1 - \alpha)E\rho + \alpha E\rho^2\right] + \alpha^2 E\left[\rho(1 - \rho)(\lambda - \rho)^+ (1 - \alpha\rho)^{-1}\right] = 0.$$

At various places in the paper we have discussed functional properties of V and Λ. The continuity of Λ given in the next theorem has been referred to in Example 4.2. We use the l^1 topology on the space of regular discount sequences; the distance between $\mathbf{A} = (\alpha_1, \alpha_2, \cdots)$ and $\mathbf{B} = (\beta_1, \beta_2, \cdots)$ is $|\mathbf{A} - \mathbf{B}| =$

$\Sigma_{m=1}^{\infty}|\alpha_m - \beta_m|$. We use the topology induced by convergence in distribution for the space of distributions of ρ.

THEOREM 5.3. *Both V and Λ are (jointly) continuous functions.*

PROOF. It is easy to see that $|V(\mathbf{A}, R, \lambda) - V(\mathbf{B}, R, \lambda)| < \varepsilon$ whenever $|\mathbf{A} - \mathbf{B}| < \varepsilon$ and $|V(\mathbf{A}, R, \lambda_1) - V(\mathbf{A}, R, \lambda_2)| < \varepsilon$ whenever $|\lambda_1 - \lambda_2| < \varepsilon/|\mathbf{A}|$. We can complete the proof that V is continuous by showing that $V(\mathbf{A}, \cdot, \lambda)$ is continuous for each $\mathbf{A}$ and λ.

As in the proof of Theorem 2.1 we prove it for $\mathbf{A} \in \mathcal{S}_n$ by induction on n and use Proposition 1.1 to prove it for a general regular $\mathbf{A}$. For the induction step we fix R_0 and consider a variable R close to R_0. If it is optimal to pull $\mathcal{L}$ for both bandits then $V(R) = V(R_0)$. If it is optimal to pull $\mathcal{R}$ initially for both bandits then

$$\begin{aligned}|V(\mathbf{A}, R) - V(\mathbf{A}, R_0)| \leqslant \alpha_1 &|E(\rho|R) - E(\rho|R_0)| \\ &+ |E(\rho|R)V(\mathbf{A}^{(1)}, \sigma R) - E(\rho|R_0)V(\mathbf{A}^{(1)}, \sigma R_0)| \\ &+ |E(1-\rho|R)V(\mathbf{A}^{(1)}, \varphi R) - E(1-\rho|R_0)V(\mathbf{A}^{(1)}, \varphi R_0)|.\end{aligned}$$

Let $\varepsilon > 0$, N_σ a neighborhood of σR_0, and N_φ a neighborhood of φR_0. By choosing R in an appropriate neighborhood of R_0 one can ensure $|E(\rho|R) - E(\rho|R_0)| < \varepsilon$, $\sigma R \in N_\sigma$, and $\varphi R \in N_\varphi$. If it is optimal to pull $\mathcal{R}$ initially for the R-bandit but $\mathcal{L}$ for the R_0-bandit then $V(\mathbf{A}, R) \geqslant \lambda\gamma_1$, but not much larger, for the above calculation shows that it is not much larger than the expected payoff for the R_0-bandit when $\mathcal{R}$ is pulled initially and an optimal strategy followed thereafter. Similar reasoning holds in case it is optimal to pull $\mathcal{R}$ initially for R_0 but not for R. The continuity of V is proved.

Suppose that $(\mathbf{A}_n, R_n) \to (\mathbf{A}_0, R_0)$ as $n \to \infty$, and, for some $\varepsilon > 0$ and every n, $\Lambda(\mathbf{A}_n, R_n) \geqslant \Lambda(\mathbf{A}_0, R_0) + \varepsilon$. Let $\lambda_0 = \Lambda(\mathbf{A}_0, R_0)$ and $\lambda = \lambda_0 + \varepsilon$. Clearly, $V(\mathbf{A}_0, R_0, \lambda_0) = \lambda_0\Sigma_{m=1}^{\infty}\alpha_{m,0}$ and $V(\mathbf{A}_n, R_n, \lambda) \geqslant \lambda\Sigma_{m=1}^{\infty}\alpha_{m,n}$ where $\alpha_{m,n}$ denotes the mth coordinate of $\mathbf{A}_n$. Since it is optimal to pull $\mathcal{R}$ initially for both $(\mathbf{A}_n, R_n, \lambda)$ and $(\mathbf{A}_n, R_n, \lambda_0)$,

$$V(\mathbf{A}_n, R_n, \lambda) - V(\mathbf{A}_n, R_n, \lambda_0) \leqslant \varepsilon\Sigma_{m=2}^{\infty}\alpha_{m,n}.$$

Hence,

$$V(\mathbf{A}_n, R_n, \lambda_0) - V(\mathbf{A}_0, R_0, \lambda_0) \geqslant \varepsilon\alpha_{1,n} - \lambda_0|\mathbf{A}_n - \mathbf{A}_0|$$

for every n, which contradicts the continuity of V. Similar reasoning yields a contradiction from the hypothesis $\Lambda(\mathbf{A}_n, R_n) \leqslant \Lambda(\mathbf{A}_0, R_0) - \varepsilon$. □

REFERENCES

BELLMAN, R. (1956). A problem in the sequential design of experiments. *Sankhya* **16** 221–229.

BERRY, D. A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.

BERRY, D. A. and FRISTEDT, B. (1979). Bernoulli k-armed bandits. Univ. Minnesota, School of Statistics, Technical Report.

BRADT, R. N., JOHNSON, S. M. and KARLIN, S. (1956). On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27** 1060–1070.

DEGROOT, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
FABIUS, J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41** 1906–1916.
FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications 1*, (*3rd Edition*). Wiley, New York.
FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications 2*, (*2nd Edition*). Wiley, New York.

DEPARTMENT OF THEORETICAL STATISTICS
UNIVERSITY OF MINNESOTA
270 VINCENT HALL
MINNEAPOLIS, MINNESOTA 55455

SCHOOL OF MATHEMATICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455