The Annals of Statistics
1988, Vol. 16, No. 1, 254–264

THE TWO-ARMED BANDIT WITH DELAYED RESPONSES[1]

BY STEPHEN G. EICK

AT&T Bell Laboratories

A general model for a two-armed bandit with delayed responses is introduced and solved with dynamic programming. One arm has geometric lifetime with parameter θ, which has prior distribution μ. The other arm has known lifetime with mean κ. The response delays completely change the character of the optimal strategies from the no delay case; in particular, the bandit is no longer a stopping problem. The delays also introduce an extra parameter p into the state space. In clinical trial applications, this parameter represents the number of patients previously treated with the unknown arm who are still living. The value function is introduced and investigated as p, μ and κ vary. Under a regularity condition on the discount sequence, there exists a manifold in the state space such that both arms are optimal on the manifold, arm x is optimal on one side and arm y on the other. Properties of the manifold are described.

1. Introduction. Consider a clinical trial in which patients arrive sequentially at times $0, 1, \ldots, n-1$ ($n = \infty$ is allowed). Each patient receives one of two irreversible treatments, say x and y. The first patient is treated at time 0. When the second patient arrives at time 1, treatment assignment is based on the information available at that time; it is known that either the first patient has survived to time 1 or not. When the third patient arrives at time 2, it is known whether the second patient has survived and also whether a first patient who had survived until time 1 has also survived until time 2, et cetera. As the trial progresses, information about the relative treatment effectiveness accrues. The objective is to assign treatments to maximize the total patient survival time, possibly discounting for future successes. This is an example of a bandit problem with delayed responses.

Bandit problems have been studied extensively in the statistical literature. Authors making significant contributions include Robbins (1952), Bradt, Johnson and Karlin (1956), Bellman (1956), Feldman (1962), Gittins and Jones (1974), Rodman (1978), Bather (1981) and Berry and Fristedt (1985). However, when applied to clinical trials, all papers in the bandit literature assume that the previous patient lifetimes are known before the next patient is treated. For clinical trials this assumption is unrealistic because it is infeasible to wait for the first patient to respond before treating the second. The inability to account for response delays is cited frequently as one of the problems in using adaptive

Received February 1986; revised May 1987.

[1]Research partially conducted while the author was a student at the University of Minnesota and partially supported under NSF Grants DMS-83-01450 and DMS-83-19924 and the University of Minnesota Statistics Alumni Fellowship.

AMS 1980 *subject classifications.* Primary 62L05; secondary 62L15.

Key words and phrases. Two-armed bandits, delayed responses, randomized clinical trials, dynamic programming, optimal strategies.

strategies in clinical trials (see Armitage, [(1985), page 22] and Simon (1977)). I address the problem of maximizing the expected total patient lifetime, possibly discounting future patients, when treatment assignment is based on partial information, the censored lifetimes, rather than the exact lifetimes.

I assume that patients treated with x have conditionally i.i.d. geometric lifetimes: $X_1, X_2, \ldots, X_n$, given $\theta \in (0,1)$, have probability mass function $(1-\theta)\theta^t$, $t = 0, 1, \ldots, \infty$. I take a Bayesian approach and assume θ is random with prior distribution μ. The sufficient statistics are the number of patient time period successes S and patient failures F. The lifetimes of the patients treated with y are independent, and independent of the X's, with known expectation κ: $E[Y_i] = \kappa$, $i = 1, \ldots, n$.

For each j, either X_j or Y_j can be observed but not both. Using treatment x initially provides information about θ, which may be useful for treating future patients. However, $E[X|\mu]$ may be less than κ, in which case a patient treated with x has a shorter life expectancy than one treated with y. This conflict between effective treatment and gathering information characterizes bandits more generally [Berry and Fristedt (1985)].

The major results in this paper concern the value of this bandit, the expected discounted patient lifetime when the best treatment allocation is used, the dynamic programming solution for finite horizon bandits, those in which only finitely many patients are treated, and the structure of the optimal strategy for a particular class of prior distributions $(s, f)\mu$ extending the beta distribution. In Section 2 I define the delayed response bandit state space, which summarizes all information available when the current patient must be treated. I also define the discount sequence, which determines the relative weights of patients in the trial, the value function, and fix the concepts with an elementary example. An interesting result in Section 2 is that, in general, the delayed response bandit is not a stopping problem. In Section 3 I solve the finite horizon bandit with dynamic programming and use the result to prove the value function is monotone in the prior distribution parameters s, f and also in κ. In Section 4 I describe the optimal strategy under a regularity condition on the discount sequence. I show there is a manifold in the state space so that arm x is optimal on one side, arm y on the other and both arms are optimal on the manifold. In Section 5 I extend the result to both arms unknown.

2. The state space. The state of a bandit summarizes all information available when the next patient is to be treated. In the current setting the state consists of the tuple $((s, f)\mu, p; \kappa; A)$. I refer to a bandit in state $((s, f)\mu, p; \kappa; A)$ as the $((s, f)\mu, p; \kappa; A)$ bandit. The first element $(s, f)\mu$ is the distribution of θ conditioned by the sufficient statistics s and f. When $s = f = 0$, $(0,0)\mu = \mu$ and, in general,

$$d(s, f)\mu = \theta^s(1-\theta)^f \, d\mu / b(s, f), \tag{1}$$

where

$$b(s, f) = \int_0^1 \theta^s(1-\theta)^f \, d\mu(\theta). \tag{2}$$

I assume that μ is not concentrated at a single point and that μ assigns no mass to $\{0,1\}$. The parameters s and f are allowed to be continuous but restricted such that $(s, f)\mu$ is defined and $E[X_j|(s, f)\mu] < \infty$. A necessary and sufficient condition for the pair μ and (s, f) to be considered is $b(s, f-1) < \infty$.

The second element in the state is p, the number of patients previously treated with x whose lifetimes are censored when the current patient is treated. These patients form an information bank; information accrues as they respond, either positively or negatively. The third element in the state is κ, the expected lifetime of patients treated with y, which I assume known. Successes and failures from patients treated with y are not included in the state since they cannot affect κ.

The discount sequence A is the final component in the state. I consider general discounting and allow $A = (\alpha_1, \alpha_2, \dots)$ to be an arbitrary summable sequence of nonnegative numbers. After n patients have been treated, the discount sequence for the bandit presenting itself is $A^{(n)} = (\alpha_{n+1}, \alpha_{n+2}, \dots)$. This discount sequence is obtained from A by deleting the first n elements of A. The horizon of A is $\inf\{i\colon \alpha_j = 0,\ j > i\}$. If this set is empty, then A is said to have an infinite horizon. It is often convenient to work with finite horizon discount sequences. The horizon n truncation of A is

$$A_n = (\alpha_1, \alpha_2, \dots, \alpha_n, 0, 0, \dots).$$

When $\alpha_j = \alpha^{j-1}$, A is said to be the geometric discount sequence with factor α, and when $\alpha_j = 1$ for $j = 1, \dots, n$ and $\alpha_j = 0$ for $j > n$, A is said to be the uniform discount sequence with horizon n. The jth tail mass of A is

$$\gamma_j = \sum_{j}^{\infty} \alpha_i.$$

A strategy for the $(\mu, p; \kappa; A)$ bandit indicates which treatment to use at each stage in the trial, depending on past treatments and the patient lifetimes censored at the present time. The worth of a strategy is the expected discounted total patient lifetime when the strategy is followed,

$$W(\tau) = E_\tau\left[\sum_{1}^{\infty} \alpha_j Z_j\right],$$

where Z_j is X_j if τ indicates x at time $j-1$ or Y_j if τ indicates y. The value of the $(\mu, p; \kappa; A)$ bandit is the supremum of the worths,

$$V = V(\mu, p; \kappa; A) = \sup_{\tau} W(\tau).$$

A strategy τ is optimal if $W(\tau) = V$. An arm is optimal if there exists an optimal strategy which indicates it initially; an optimal arm is the first selection of an optimal strategy. The worth of selecting x and then following an optimal strategy given the result is

$$V^{(x)} = V^{(x)}(\mu, p; \kappa; A) = \sup\{W(\tau)|\tau \text{ indicates } x \text{ initially}\}.$$

An analogous definition holds for $V^{(y)}$. Then arm x is optimal if and only if $V^{(x)} = V$ and similarly for arm y. Both arms are optimal if $V^{(x)} = V^{(y)} = V$.

To illustrate the states, consider a trial in which three patients are treated, each receiving equal weight. The initial state is $(\mu, 0; \kappa; A)$, where $A = (1, 1, 1, 0, 0, \ldots)$. Suppose the patient arriving at time 0 is treated with x. The state at time 1 is random depending on whether or not the first patient survives to time 1. If the patient does survive, then at time 1 the state is $((1, 0)\mu, 1; \kappa; A^{(1)})$. There has been one success, no failures, one patient is in the information bank and two patients remain to be treated, $A^{(1)} = (1, 1, 0, 0, \ldots)$.

Now suppose that the patient arriving at time 1 is treated with y and both patients survive to time 2. Then the state is $((2, 0)\mu, 1; \kappa; A^{(2)})$. Two successes have been observed on the patient treated with x at time 0, and no failures have been observed. The information bank still contains one x-observation; the y-observation is not in the information bank because the distribution of y is known. One patient remains to be treated, $A^{(2)} = (1, 0, 0, \ldots)$.

Suppose the third and final patient is given treatment x, and that patients 2 and 3 do not survive until time 3 while patient 1 does. Then the state is $((3, 1)\mu, 1; \kappa; A^{(3)})$, where $A^{(3)} = (0, 0, \ldots)$. A zero discount sequence indicates trial completion.

A classical result for immediate response bandits with arm y known is that for regular discounting (regular discounting extends both geometric and uniform discounting), the bandit is an optimal stopping problem [see Berry and Fristedt (1985), page 92]. The optimal strategy indicates arm x N times, where N is the random stopping time, and then indicates y at every subsequent stage. Delayed response bandits are not stopping problems. Simple examples show that the optimal strategy may indicate x and switch to y with patients in the information bank. If these patients survive sufficiently long, arm x eventually will become optimal.

3. The value function. The value function satisfies the *fundamental equation of dynamic programming*,

$$V(\mu, p; \kappa, A) = V^{(x)}(\mu, p; \kappa; A) \vee V^{(y)}(\mu, p; \kappa; A), \tag{3}$$

where

$$V^{(x)}(\mu, p; \kappa; A) = \alpha_1 E[X|\mu] + E\big[V\big((S^{(x)}, F^{(x)})\mu, P^{(x)}; \kappa; A^{(1)}\big)|\mu\big], \tag{4}$$

$$V^{(y)}(\mu, p; \kappa; A) = \alpha_1 \kappa + E\big[V\big((S^{(y)}, F^{(y)})\mu, P^{(y)}; \kappa; A^{(1)}\big)|\mu\big] \tag{5}$$

and $S^{(z)}$, $F^{(z)}$ and $P^{(z)}$ are the random number of successes, failures and bank size at time 1 when arm z was selected at time 0. The notation $a \vee b = \max\{a, b\}$.

When A has horizon $n < \infty$, (3)–(5) can be used to calculate V recursively. The starting points are all possible states for which the discount sequence $A^{(n-1)}$ has horizon 1,

$$V\big(\mu, p; \kappa; A^{(n-1)}\big) = \alpha_n\{E[X|\mu] \vee \kappa\}.$$

The main result presented in this section is that $V((s, f)\mu, p; \kappa; A)$ is nondecreasing in s, p and κ and nonincreasing in f.

THEOREM 1. *Suppose μ is not a one-point measure and $\mu(\{0,1\}) = 0$. Then $V((s, f)\mu, p; \kappa; A)$ is continuous in s, f and κ, nondecreasing in s, p and κ, nonincreasing in f and convex in κ.*

The proof of Theorem 1 uses the following lemma and a stochastic ordering of the prior distributions $(s, f)\mu$ induced on the state space, which I subsequently define. The lemma is a standard result in the bandit literature. See, for example, Berry and Fristedt [(1985), Theorem 2.6.1].

LEMMA. *For any $(\mu, p; \kappa; A)$ bandit,*

$$\begin{aligned} &V(\mu, p; \kappa, A_n) + \gamma_{n+1}\{E[X|\mu] \vee \kappa\} \\ &\quad \le V(\mu, p; \kappa; A) \\ &\quad \le V(\mu, p; \kappa; A_n) + \gamma_{n+1}E\left[\frac{\theta}{1-\theta} \vee \kappa \middle| \mu\right]. \end{aligned} \tag{6}$$

A random variable Z_1 with distribution ν_1 is *stochastically larger* than Z_2, with distribution ν_2 if, for all t,

$$P\{Z_1 \ge t|\nu_1\} \ge P\{Z_2 \ge t|\nu_2\}, \tag{7}$$

with strictness if (7) holds with inequality for some t. If g is a nondecreasing function, then $g(Z_1)$ is stochastically larger than $g(Z_2)$ and furthermore, if $E[g(Z_2)]$ exists, then $E[g(Z_1)] \ge E[g(Z_2)]$. The prior distributions $(s, f)\mu$ induce a stochastic ordering on θ; $(s, f)\mu$ is stochastically increasing in s and decreasing in f with strictness if μ is not supported at a single point.

PROOF OF THEOREM 1. First note that $V(\mu, p_1; \kappa; A) \ge V(\mu, p_2; \kappa; A)$ when $p_1 \ge p_2$. This is immediate since any information gained from the extra patients in the bank of the $(\mu, p_1; \kappa; A)$ bandit can be ignored.

I show the finite horizon case by induction and then extend to infinite horizons using the previous lemma. When A has horizon 1, the result is immediate. Suppose the theorem holds for all horizons $m < n$. Let A have horizon n. Consider $V^{(x)}$,

$$\begin{aligned} &V^{(x)}((s, f)\mu, p; \kappa; A) \\ &\quad = \alpha_1 E[X|(s, f)\mu] \\ &\qquad + E\left[V\left((S^{(x)} + s, F^{(x)} + f)\mu, P^{(x)}; \kappa; A^{(1)}\right)|(s, f)\mu\right], \end{aligned} \tag{8}$$

where the horizon of $A^{(1)}$ is $n - 1$. The first term of (8) is nondecreasing in s and nonincreasing in f since the family of prior distributions $(s, f)\mu$ for θ is stochastically ordered. Writing

$$S^{(x)} = P^{(x)} \quad \text{and} \quad F^{(x)} = p + 1 - P^{(x)}$$

and substituting into the second term on the right-hand side of (8),

$$V\big((S^{(x)}+s, F^{(x)}+f)\mu, P^{(x)}; \kappa; A^{(1)}\big) \tag{9}$$
$$= V\big((P^{(x)}+s, p+1-P^{(x)}+f)\mu, P^{(x)}; \kappa; A^{(1)}\big).$$

The monotonicity for this term follows by induction and the stochastic ordering of $P^{(x)}$. The conclusion for κ also follows since, by induction, (9) is convex and, therefore, the second term in (8) is a finite weighted sum of convex functions. A similar argument applies to $V^{(y)}$. The inductive step is complete since $V = V^{(x)} \vee V^{(y)}$. □

Extension to monotone likelihood ratio. The proof of Theorem 1 actually shows that $V(\mu, p; \kappa; A)$ is nondecreasing when μ is allowed to vary within any class of priors with a monotone likelihood ratio. Thus

$$V(\mu, p; \kappa; A) \leq V(\nu, p; \kappa; A)$$

if $d\nu/d\mu$ is nondecreasing.

4. Optimal strategies. Theorem 1 describes the value of the $((s, f)\mu, p; \kappa; A)$ bandit but gives little insight into the optimal strategies. The main result in this section is Theorem 3. It shows, under a regularity condition on the discount sequence, that for each p there exists a manifold in $(s, f; \kappa)$ space such that arm x is optimal on one side of the manifold and arm y is optimal on the other. Both arms are optimal on the manifold. This reduces the decision problem to finding the manifold. This manifold is the zero of the Δ function.

Then Δ is the difference in worths between pulling arm x and proceeding optimally given the result and pulling arm y and proceeding optimally,

$$\Delta(\mu, p; \kappa; A) = V^{(x)}(\mu, p, \kappa; A) - V^{(y)}(\mu, p; \kappa; A).$$

The sign of Δ determines the optimal initial selection. A large positive value of Δ indicates that x is strongly preferred to y; Δ is the amount lost if arm y is selected initially even if an optimal strategy is followed thereafter.

There is a special relationship between $\Delta(\mu, p; \kappa; A_1) = \alpha_1\{E[X|\mu] - \kappa\}$ and the myopic strategies; those which maximize the lifetime of the current patient at each stage in the trial. A strategy is myopic if it indicates arm x when $\Delta(\mu, p; \kappa; A_1)$ is nonnegative and arm y when $\Delta(\mu, p; \kappa; A_1)$ is nonpositive.

As $p \to \infty$, it is easy to show that $\Delta(\mu, p; \kappa; A) \to \Delta(\mu, p; \kappa; A_1)$. The intuition behind this result is that for an arbitrarily large bank size, complete information about θ will be available one time period later. Therefore, the first selection should maximize the lifetime of the current patient.

Theorem 2 develops recursive formulas for Δ. When $n = 2$, the formulas are the delayed response analogue of a standard result for classical bandits. Besides being interesting in its own right, Theorem 2 will be used to prove Theorem 3.

Theorem 2 decomposes Δ into three parts. The first term is a multiple of the expected lifetime difference between the arms, the second is the expected

difference in the positive and negative parts of Δ at times 1 to $n-1$ and the third is the difference in value functions averaged over the states at time n.

Let $S^{(z_1)}$ be the random number of successes at time 1 when z_1 is selected at time 0 and $S^{(z_1 z_2 \cdots z_k)}$ be the random number of successes at time k when arm z_i is selected at time $i-1$, $i = 1, \ldots, k$. Similar definitions apply to $F^{(z_1 \cdots z_k)}$ and $P^{(z_1 \cdots z_k)}$. Let τ_k be the length k tuple $(xy \cdots y)$ and σ_k be the length k tuple $(yx \cdots x)$. Then S^{τ_k} is the random number of successes at time k, given that x was selected at time 0 and y at times 1 through $k-1$, and S^{σ_k} is the random number of successes at time k, given that y was selected at time 0 and x at times 1 through $k-1$.

THEOREM 2. *For all μ, p, κ and $n \geq 2$ ($n = \infty$ allowed),*

$$
\begin{aligned}
\Delta(\mu, p; \kappa; A) = {} & \left\{\alpha_1 - \sum_{j=2}^{n} \alpha_j\right\}\{E[X|\mu] - \kappa\} \\
& + \sum_{j=1}^{n-1} E\left[\Delta^+\left((S_j^{\tau_j}, F_j^{\tau_j})\mu, P_j^{\tau_j}, \kappa; A^{(j)}\right)\right. \\
& \qquad \left. - \Delta^-\left((S_j^{\sigma_j}, F_j^{\sigma_j})\mu, P_j^{\sigma_j}; \kappa; A^{(j)}\right)|\mu\right] \\
& + E\left[V\left((S_n^{\tau_n}, F_n^{\tau_n})\mu, P_n^{\tau_n}; \kappa; A^{(n)}\right)\right. \\
& \qquad \left. - V\left((S_n^{\sigma_n}, F_n^{\sigma_n})\mu, P_n^{\sigma_n}; \kappa; A^{(n)}\right)|\mu\right].
\end{aligned} \tag{10}
$$

PROOF. I prove the case $n = 2$. The result for arbitrary finite n follows by iterating the argument and the proof for $n = \infty$ follows by approximation. Write

$$
\begin{aligned}
\Delta(\mu, p; \kappa; A) = {} & \alpha_1\{E[X|\mu] - \kappa\} \\
& + E\left[V\left((S^{(x)}, F^{(x)})\mu, P^{(x)}; \kappa; A^{(1)}\right)\right. \\
& \qquad \left. - V\left((S^{(y)}, F^{(y)})\mu, P^{(y)}; \kappa; A^{(1)}\right)|\mu\right].
\end{aligned} \tag{11}
$$

In (11), replace V by $\Delta^+ + V^{(y)}$ and $-V$ by $-\Delta^- - V^{(x)}$:

$$
\begin{aligned}
\Delta(\mu, p; \kappa; A) = {} & \alpha_1\{E[X|\mu] - \kappa\} \\
& + E\left[\Delta^+\left((S^{(x)}, F^{(x)})\mu, P^{(x)}; \kappa; A^{(1)}\right)\right. \\
& \qquad \left. + V^{(y)}\left((S^{(x)}, F^{(x)})\mu, P^{(x)}; \kappa; A^{(1)}\right)|\mu\right] \\
& - E\left[\Delta^-\left((S^{(y)}, F^{(y)})\mu, P^{(y)}; \kappa; A^{(1)}\right)\right. \\
& \qquad \left. + V^{(x)}\left((S^{(y)}, F^{(y)})\mu, P^{(y)}; \kappa; A^{(1)}\right)|\mu\right].
\end{aligned} \tag{12}
$$

In the first expectation on the right-hand side of (12), write $V^{(y)}$ as $\alpha_2\kappa + V$ and in the second, write $V^{(x)}$ as $\alpha_2 E[X|(S^{(y)}, F^{(y)})\mu] + V$. Then (10), with $n = 2$, follows because

$$
E\left[E\left[X|(S^{(y)}, F^{(y)})\mu\right]|\mu\right] = E[X|\mu]. \qquad \square
$$

Theorem 3 shows for discount sequences satisfying $\alpha_j \geq \gamma_{j+1}$ (see Section 2 for notation) that Δ is nondecreasing in s and nonincreasing in f and κ. This proves

the existence of the boundary manifold described at the beginning of this section and describes the optimal strategy in the following sense. For fixed s, f and p, there exists a κ^* ($\kappa^* = \infty$ allowed) such that arm x is optimal if and only if $\kappa \leq \kappa^*$. Similarly, for fixed s, κ and p, there exists an f^* ($f^* = +$ or $-\infty$ allowed) such that arm x is optimal if and only if $f \leq f^*$, and for fixed f, κ and p, there exists an s^* ($s^* = +$ or $-\infty$ allowed) such that arm x is optimal if and only if $s \geq s^*$. However, the s^*, f^* and κ^* are very difficult to calculate. In particular, these results hold for geometric discounting $A = (1, \alpha, \alpha^2, \ldots)$ when $\alpha \leq \frac{1}{2}$.

I conjecture that a similar result holds for all nonincreasing discount sequences. I have verified this conjecture for geometric discounting under the additional restriction $p = 0$.

THEOREM 3. *Suppose the discount sequence A satisfies*

$$\alpha_j \geq \gamma_{j+1} \tag{13}$$

for $j = 1, 2, \ldots$. Then for all μ and p, $\Delta((s, f)\mu, p; \kappa; A)$ is nondecreasing in s and nonincreasing in f and κ. Furthermore

$$\Delta((s, f)\mu, p + 1; \kappa; A) \geq \Delta((s, f + 1)\mu, p; \kappa; A). \tag{14}$$

If μ is not concentrated at a single point and there is strict inequality in (13) *for $j = 1$, then Δ is increasing in s and decreasing in f and κ. In this case, there is strict inequality in* (14).

The proof of Theorem 3 will be developed gradually in Lemmas A, B and C. Lemma A shows that Theorem 3 holds under the additional assumption that A has finite horizon. Lemma B extends the result to infinite horizons and Lemma C proves "strictness."

LEMMA A. *Theorem 3 holds when the horizon of A is finite.*

REMARK. The proof of this lemma depends on the following observation. The distributions of θ, $P_j^{\tau_j}$ and $P_j^{\sigma_j}$ are stochastically increasing in s and decreasing in f; for any w and p, $P\{\theta \geq w|(s, f)\mu\}$, $P\{P_j^{\tau_j} \geq p|(s, f)\mu\}$ and $P\{P_j^{\sigma_j} \geq p|(s, f)\mu\}$ are nondecreasing in s and nonincreasing in f.

PROOF OF LEMMA A. Proceed by induction. When A has horizon 1, the results are trivial. In this case $A = A_1$ and (14) follows since $\Delta((s, f)\mu, p; \kappa; A_1)$ does not depend on p and is nonincreasing in f. Assume the result holds for all horizons $m < n$ and that A has horizon n. Consider (10),

$$\begin{aligned}\Delta((s, f)\mu, p; \kappa; A) &= (\alpha_1 - \gamma_2)\{E[X|(s, f)\mu] - \kappa\} \\ &\quad + \sum_{j=1}^{n-1} E\left[\Delta^+\left((s + S_j^{\tau_j}, f + F_j^{\tau_j})\mu, P_j^{\tau_j}; \kappa; A^{(j)}\right)\right. \\ &\quad \left. - \Delta^-\left((s + S_j^{\sigma_j}, f + F_j^{\sigma_j})\mu, P_j^{\sigma_j}; \kappa; A^{(j)}\right)|(s, f)\mu\right].\end{aligned} \tag{15}$$

The first term on the right-hand side of (15) is nondecreasing in s and nonincreasing in both f and κ from (13).

Consider the jth term in the sum:

$$E\left[\Delta^+\left((s+S_j^{\tau_j}, f+F_j^{\tau_j})\mu, P_j^{\tau_j}; \kappa; A^{(j)}\right)\right. \tag{16}$$
$$\left.-\Delta^-\left((s+S_j^{\sigma_j}, f+F_j^{\sigma_j})\mu, P_j^{\sigma_j}; \kappa; A^{(j)}\right)\middle|(s, f)\mu\right].$$

I show the first term in (16) in nondecreasing in s and nonincreasing in f and κ; the second is similar. For notational ease, let $S_j = S_j^{\tau_j}$, $F_j = F_j^{\tau_j}$ and $P_j = P_j^{\tau_j}$. Then $S_j = S_{j-1} + P_j$ and $F_j = p + 1 - P_j$, where $S_{j-1} = P_1 + \cdots + P_{j-1}$. Thus

$$\Delta^+\left((s+S_j, f+F_j)\mu, P_j; \kappa; A^{(j)}\right)$$
$$= \Delta^+\left((s+S_{j-1}+P_j, f+p+1-P_j)\mu, P_j; \kappa; A^{(j)}\right).$$

By induction,

$$\Delta^+\left((s+S_{j-1}+p_j, f+p+1-p_j)\mu, p_j; \kappa; A^{(j)}\right) \tag{17}$$

is nondecreasing in both s and S_{j-1} and nonincreasing in both f and κ. Also (14) implies that (17) is nondecreasing in p_j. However, the conditional distribution of P_j given P_{j-1} and θ is binomial and hence stochastically nondecreasing in θ and P_{j-1}. Therefore,

$$E\left[\Delta^+\left((s+S_{j-1}+P_j, f+p+1-P_j)\mu, P_j; \kappa; A^{(j)}\right)\middle|\theta, S_{j-1}, P_{j-1}\right]$$

is nondecreasing in θ, S_{j-1}, P_{j-1} and s and nonincreasing in f and κ. But S_{j-1} and P_{j-1} are stochastically nondecreasing in θ, whence

$$E\left[\Delta^+\left((s+S_{j-1}+P_j, f+p+1-P_j)\mu, P_j; \kappa; A^{(j)}\right)\middle|\theta\right] \tag{18}$$

is nondecreasing in both s and θ and nonincreasing in f and κ. Finally, since the distribution of θ is $(s, f)\mu$, the expectation of (18) with respect to $(s, f)\mu$ is nondecreasing in s and nonincreasing in f and κ.

To complete the induction, I show that each term in (16) satisfies (14). Let * denote a random variable from the $((s, f)\mu, p+1; \kappa; A)$ bandit as opposed to the $((s, f+1)\mu, p; \kappa; A)$ bandit. Consider the first term in (16); the second is analogous. Write the jth difference of Δ^+ functions as

$$\Delta^+\left((s+S_j^*, f+F_j^*)\mu, P_j^*; \kappa; A^{(j)}\right)$$
$$-\Delta^+\left((s+S_j, f+1+F_j)\mu, P_j; \kappa; A^{(j)}\right) \tag{19}$$
$$= \Delta^+\left((s+P_1^*+\cdots+P_j^*, f+p+2-P_j^*)\mu, \kappa; P_j^*; A^{(j)}\right)$$
$$-\Delta^+\left((s+P_1+\cdots+P_j, f+p+2-P_j)\mu, \kappa; P_j; A^{(j)}\right).$$

Then the random number of failures in the respective terms of the right-hand side of (19) is $f+p+2-P_j^*$ and $f+p+2-P_j$. But P_j^* is stochastically larger than P_j and $P_1^*+\cdots+P_j^*$ is stochastically larger than $P_1+\cdots+P_j$. The result follows by induction using (14). □

The next step in the proof of Theorem 3 is to extend Lemma A to infinite horizons.

LEMMA B. *Theorem* 3 *holds when the horizon of A is infinite.*

PROOF. This is immediate from Lemma A since Δ is continuous in A. □

The proof of Theorem 3 is completed by proving the strictness assertion.

LEMMA C. *If* μ *is not concentrated at a single point and there is strict inequality in* (13) *for* $j = 1$, *then* Δ *is increasing in* s *and decreasing in* f *and* κ. *Furthermore*, (14) *holds with strict inequality.*

PROOF. Strictness in (13) and the hypothesis on μ implies the first term in (15) is increasing in s, decreasing in f and κ and satisfies (14) with strictness. □

Extension to monotone likelihood ratio. The proof of Theorem 3 actually shows that $\Delta(\mu, p; \kappa; A)$ is nondecreasing when μ is allowed to vary within any class of priors with a monotone likelihood ratio. Thus

$$\Delta(\mu, p; \kappa; A) \leq \Delta(\nu, p; \kappa; A)$$

if $d\nu/d\mu$ is nondecreasing.

The following corollary provides a condition when an optimal strategy is to indicate y at all stages.

COROLLARY. *Suppose A is geometric with* $\alpha \leq \frac{1}{2}$. *Assume arm y is optimal at time* 0 *in the* $(\mu, p; \kappa; A)$ *bandit and all p patients in the information bank fail. Then an optimal strategy is to indicate arm y at all subsequent times.*

PROOF. Since y is optimal initially, then $0 \geq \Delta(\mu, p; \kappa; A)$. Since $\alpha \leq \frac{1}{2}$, the regularity condition of Theorem 3 is satisfied. From (14),

$$\begin{aligned} 0 \geq \alpha\Delta(\mu, p; \kappa; A) &= \Delta(\mu, p; \kappa; A^{(1)}) \\ &\geq \Delta((0,1)\mu, p-1; \kappa; A^{(1)}) \geq \cdots \geq \Delta((0,p)\mu, 0; \kappa; A^{(1)}). \end{aligned}$$

Then arm y is optimal in the $((0, p)\mu, 0; \kappa; A^{(1)})$ bandit. When arm y is selected and the bank size is zero, the state of the bandit presenting itself at the next stage differs from the current state only by a multiple of the discount sequence. This does not effect the optimal arm and so arm y continues to be optimal. □

5. Extensions to both arms x and y unknown. There is an interesting extension of Theorem 3 when observations on both arms x and y are random with prior distributions. As before, assume that $X_1, X_2, \ldots, X_n$ given $\theta \in (0,1)$ have i.i.d. geometric lifetimes with probability mass function $(1-\theta)\theta^t$, $t = 0, 1, 2, \ldots$, where θ is random with prior distribution μ. Now assume $Y_1, Y_2, \ldots, Y_n$ given λ also have i.i.d. geometric lifetimes with probability mass function $(1-\lambda)\lambda^s$, $s = 0, 1, 2, \ldots$, where λ is random with prior distribution ν. Assume θ is independent of λ. When $\nu = \delta_\lambda$ is concentrated at a single point λ, this setting is equivalent to that considered previously with $\kappa = \lambda/(1-\lambda)$. In the new setting, the state space is six-dimensional, $((s_x, f_x)\mu, p_x; (s_y, f_y)\nu, p_y; A)$, where

s_z, f_z and p_z are the number of z-patient successes, failures and bank size for $z = x$ or y. Theorem 3 extends to this setting with the same proof.

THEOREM 4. *Suppose the discount sequence A satisfies* (13) *for* $j = 1, 2, \ldots$. *Then for all* μ, p_x, ν *and* p_y, $\Delta((s_x, f_x)\mu, p_x; (s_y, f_y)\nu, p_y; A)$ *is nondecreasing in* s_x *and* f_y *and nonincreasing in* f_x *and* s_y. *Furthermore,*

$$\Delta\big((s_x, f_x)\mu, p_x + 1; (s_y, f_y)\nu, p_y; A\big) \geq \Delta\big((s_x, f_x + 1)\mu, p_x; (s_y, f_y)\nu, p_y; A\big) \tag{20}$$

and

$$\Delta\big((s_x, f_x)\mu, p_x; (s_y, f_y)\nu, p_y + 1; A\big) \leq \Delta\big((s_x, f_x)\mu, p_x; (s_y, f_y + 1)\nu, p_y; A\big). \tag{21}$$

When (13) holds with strictness for $j = 1$, then if μ is not concentrated at a single point, (20) holds with strictness, and if ν is not concentrated at a single point, (21) holds with strictness.

When the conditions of Theorem 4 hold, there is a partitioning of the state space similar to that described in Section 4. For each fixed p_x and p_y, there exists a boundary manifold in $(s_x, f_x; s_y, f_y)$ space such that arm x is optimal on one side, arm y on the other and both arms optional on the manifold.

Acknowledgments. I wish to thank Donald A. Berry and an unknown referee for their many helpful suggestions and encouragement.

REFERENCES

ARMITAGE, P. (1985). The search for optimality in clinical trials. *Internat. Statist. Rev.* **53** 15–24.

BATHER, J. A. (1981). Randomized allocations of treatments in sequential experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **43** 265–292.

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

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

BRADT, R. N., JOHNSON, S. M. and KARLIN, S. (1956). On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27** 1060–1074.

FELDMAN, D. (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33** 847–856.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (J. Gani, ed.) 241–266. North-Holland, Amsterdam.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535.

RODMAN, L. (1978). On the many-armed bandit problem. *Ann. Probab.* **6** 491–498.

SIMON, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics* **33** 743–749.

AT&T BELL LABORATORIES
ROOM 3F-511A
CRAWFORDS CORNER ROAD
HOLMDEL, NEW JERSEY 07733-1988