Strategies for the Two-Armed Bandit Problem With Delayed Responses

by

Stephen G. Eick*
University of Minnesota

Technical Report No. 457
December 1985

---

* Partially supported by the University of Minnesota Statistics Alumni Fellowship. Adapted in part from the author's dissertation.

Summary

Either of two stochastic processes (arms) is selected for observation at each of times $0,\cdots,n-1$. Each observation is the survival time of the experimental unit and the observations are in real time. Arm x is conditionally geometric with parameter $\theta \in (0,1)$ which is random with prior μ. Arm y is known to have mean κ. The arm observed at time j can depend on the previous selections and observations, but the observations are censored at time j. The objective is to maximize the expected sum of the observations, possibly discounting over time. 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 1 is optimal on one side, and arm 2 on the other. Properties of the manifold are investigated.

1. Introduction

1.1. The One-armed Bandit with Delayed Responses

The one-armed bandit with delayed responses is introduced by Eick (1985) as a model for sequential clinical trials. Patients arrive sequentially at times $0,1,2,\cdots,n-1$ ($n = \infty$ is allowed) and must receive one of two irreversible treatments, say x or y. When the next patient arrives the treatment assignment is based on the information available at that time, however, the previous patients' lifetimes are censored. This assumption differs from the classical approach taken by Bradt, Johnson, and Karlin (1956), Bellman (1956), Feldman (1962), Gittins and Jones (1974), Rodman (1978), Bather (1981), and Berry and Fristedt (1985) in which, when applied to clinical trials, it is assumed that

all patients respond immediately. The objective is to assign treatments to maximize the expected total patient survival time, possibly discounting future patients.

I assume that patients treated with x have conditionally i.i.d. geometric lifetimes: $X_1, X_2, \cdots, X_n$ given $\theta \in (0,1)$ have probability mass function $(1-\theta)\theta^t$, $t = 0,1,\cdots,\infty$. 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 expected lifetime of patients treated with y is known to be κ: $E[Y_i] = \kappa$, $i = 1,\cdots,n$.

The <u>state</u> 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 $((s,f)\mu,p;\kappa;A)$. 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

$$d(s,f)\mu = \theta^s(1-\theta)^f d\mu / b(s,f) \tag{1.1}$$

where

$$b(s,f) = \int_0^1 \theta^s(1-\theta)^f d\mu(\theta). \tag{1.2}$$

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. There is a simple relationship between S, F, and the information bank. The observed patient time period successes S is the sum of the information bank size over all previous times and the observed failures F is the number of patients treated with x minus the information bank size at the current time. The third element in the state is κ. Successes, failures, and patients treated with y are not included since they cannot affect κ.

The discount sequence A is the final component in the state. I consider general discounting and allow $A = (\alpha_1,\alpha_2,\cdots)$ to an arbitrary summable sequence of nonnegative numbers. After n patients have been treated the discount sequence for the bandit presenting itself is $A^{(n)} = (\alpha_{n+1},\alpha_{n+2},\cdots)$. This discount sequence is obtained from A by deleting the first n elements of A. The horizon of A is $\inf\{i: \alpha_j = 0, j > i\}$. 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 $A_n = (\alpha_1,\alpha_2,\cdots,\alpha_n,0,0,\cdots)$. The jth tail mass of A is

$$\gamma_j = \sum_j^\infty \alpha_i.$$

By assumption $\gamma_j \longrightarrow 0$ as $j \longrightarrow \infty$.

A strategy for the $(\mu,p;\kappa;A)$-bandit indicates with 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:

$$W(\tau) = E_\tau[\sum_1^\infty \alpha_j Z_j],$$

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 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

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

An analogous definition holds for $V^{(y)}$. Then arm x is optimal if and only if $V^{(x)} = V$ and similarly for arm y. When A has finite horizon the optimal strategy and value can be calculated recursively using the dynamic programming equation. See Eick (1985).

1.2. Summary of Results

This paper investigates the optimal initial selection as a function of the state. In Section 2.1, I consider $\Delta = V^{(x)} - V^{(y)}$; $\Delta \geq 0$ when arm x is optimal

and $\Delta \leqq 0$ when arm y is optimal. Theorem 2.1 presents an exact formula for Δ when A has horizon 2. Theorem 2.5 expresses Δ as the life expectancy difference between a single selection on arm x and arm y, plus the difference in value functions when an extra patient is added to the information bank. Section 2.4 develops recursive formulas for Δ involving the discount sequence.

Theorem 3.1 is the main result in this paper. Under a regularity condition on the discount sequence, it says that Δ is monotone in s, f, and κ. For fixed p this reduces the decision problem to finding the manifold in $(s,f;\kappa)$-space on which Δ vanishes.

2. The Δ-Function

2.1. Definition of Δ

The delta function, Δ, is the difference in worths between pulling arm x and proceeding optimally given the result versus pulling arm y and proceeding optimally:

$$\Delta(\mu,p;\kappa;A) = V^{(x)}(\mu,p,\kappa;A) - V^{(y)}(\mu,p;\kappa;A). \qquad (2.1)$$

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.

2.2. The Δ-function for Horizon 2 Bandits

When A is the horizon 2 uniform an explicit formula for $\Delta(\mu,p;\kappa;A)$ exists. For notational ease let

$$g(s,f) = E[X|(s,f)\mu],$$

$$a(p,j) = \binom{p}{j} E[\theta^j(1-\theta)^{p-j}|\mu],$$

for $0 \leq j \leq p$. Then $a(p,j)$ is the probability that j of p patients survive to the next time period and $g(s,f)$ is the life expectancy of the next patient treated with arm x when the sufficient statistics are s and f.

Theorem 2.1. Suppose A is uniform with horizon 2. Then for all μ, p and κ,

$$\Delta(\mu,p;\kappa;A) = \Delta(\mu,p;\kappa;A_1) + \begin{cases} 0 & \text{if } \kappa \leq g(0,p+1), \\ \dfrac{\binom{p}{k}}{\binom{p+1}{k}} a(p+1,k)(\kappa - g(k,p+1-k)) & \text{if } g(k,p+1-k) \leq \kappa \leq g(k,p-k), \\ \dfrac{\binom{p}{k}}{\binom{p+1}{k+1}} a(p+1,k+1)(g(k+1,p-k) - \kappa) & \text{if } g(k,p-k) \leq \kappa \leq g(k+1,p-k), \\ 0 & \text{if } g(p+1,0) \leq \kappa. \end{cases}$$

The proof of Theorem 2.1 is easy and so is omitted.

For the general finite horizon case, $\Delta(\mu,p;\kappa;A) - \Delta(\mu,p;\kappa;A_1)$ is nonnegative, piecewise linear, and has compact support.

2.3. General Properties of Δ

The Δ-function is continuous in s, f, κ and A since V is continuous (Eick, 1985). The next proposition says that $\Delta(\mu,p;\kappa,A) \longrightarrow \Delta(\mu,p;\kappa;A_1)$ as $p \longrightarrow \infty$. This shows, for example, that the myopic selection is optimal for sufficiently large p.

Proposition 2.2. For all μ, p, κ and A, as $p \longrightarrow \infty$, $\Delta(\mu,p;\kappa;A) \longrightarrow \Delta(\mu,p;\kappa;A_1)$.

Proof: Proposition 2.2 follows from Theorem 4.4 of Eick (1985). □

The intuitive justification of Proposition 2.2 is that for sufficiently large p complete information will be available at the next stage since by the law of large numbers the fraction of patients surviving converges to θ. The initial patient should be treated to maximize his or her life expectancy. Surprisingly, the convergence in Proposition 2.2 is not monotone in p. There exist examples where $\Delta(\mu,p;\kappa;A) < 0$ for all odd p and $\Delta(\mu,p;\kappa;A) > 0$ for all even $p \leq M$, where M can be arbitrarily large.

The following theorem expresses Δ in terms of $V^{(x)}$ and $V^{(y)}$ as p varies. Part (a) says that $\Delta(\mu,p;\kappa;A) = \Delta(\mu,p;\kappa;A_1)$ plus the increase in value when arm y is selected initially and p is changed to p+1. Part (b) is a similar result for arm x.

Theorem 2.3. The following hold for all μ, κ and A.

(a) For arbitrary p:

$$\Delta(\mu,p;\kappa;A) = \Delta(\mu,p;\kappa;A_1) + V^{(y)}(\mu,p+1;\kappa;A) - V^{(y)}(\mu,p;\kappa;A). \qquad (2.2)$$

(b) For $p \geq 1$:

$$\Delta(\mu,p;\kappa;A) = \Delta(\mu,p;\kappa;A_1) + V^{(x)}(\mu,p;\kappa;A) - V^{(x)}(\mu,p-1;\kappa;A). \qquad (2.3)$$

Proof. Let $P^{(x)}$ be the random bank size at time 1 when arm x is selected at time 0 and $P^{(y)}$ when arm y is selected. Then when x is selected at time 0 the state at time 1 is $[(P^{(x)},p+1-P^{(x)})\mu;P^{(x)};\kappa;A^{(1)}]$ and when y is selected at time 0 the state at time i is $[(P^{(y)},p-P^{(y)})\mu;P^{(y)};\kappa;A^{(1)}]$. The distribution of $P^{(x)}$ given θ is binomial with sample size $p + 1$. There are p patients initially alive and the patient treated at time 0 receives treatment x. When arm y is selected at time 0, the number of patients alive at time 1 in the $(\mu,p+1;\kappa;A)$-bandit is also conditionally binomial with sample size $p + 1$. Excluding the stage 1 lifetimes, $V^{(y)}(\mu,p+1;\kappa;A)$ and $V^{(x)}(\mu,p;\kappa;A)$ are equal. This is so since the states of both bandits have identical distributions at time 1. The difference in stage one lifetimes is $-\Delta(\mu,p;\kappa;A_1)$. This shows

$$V^{(y)}(\mu,p+1;\kappa;A) + \Delta(\mu,p;\kappa;A_1) = V^{(x)}(\mu,p;\kappa;A). \qquad (2.4)$$

Equation (2.2) is obtained by substituting (2.4) into (2.1). The derivation of (2.3) is similar. □

Theorem 2.3 is particularly interesting because it decomposes $\Delta(\mu,p;\kappa;A)$ into $\Delta(\mu,p;\kappa;A_1)$ for the myopic selection plus an information factor. This factor is nonnegative and represents the increase in value due to a better allocation of future treatments which could be obtained if an extra patient were

in the x information bank.

2.4. Recursive Formulas for Δ

The upcoming three theorems develop recursive formulas for Δ. Theorem 2.4 is the delayed response analogue of a standard result for classical bandits (Berry, 1972). Theorems 2.5 and 2.6 are delayed response extensions of Theorem 2.4. These results are peculiar to the delayed response setting and have no classical bandit analogue. Besides being interesting in its own right, Theorem 2.6 will be used to prove the forthcoming Theorem 3.1.

Theorem 2.4 decomposes Δ into three parts. The first term is related to life expectancy difference between the arms, the second is the expected difference in the positive and negative parts of Δ at time 1, and the third is the difference in value functions averaged over the states at time 2.

Let $S^{(x)}$ be the random number of successes at time 1 when x is selected at time 0 and $S^{(xy)}$ be the random number of successes at time 2 when y is then selected at time 1. Similar definitions apply to $S^{(y)}$, $S^{(yx)}$, $F^{(x)}$, $F^{(xy)}$, $F^{(y)}$, $F^{(yx)}$, $P^{(x)}$, $P^{(xy)}$, $P^{(y)}$, and $P^{(yx)}$.

Theorem 2.4. For all μ, p, κ and A, the delta function satisfies:

$$\begin{aligned}\Delta(\mu,p;\kappa;A) &= (\alpha_1-\alpha_2)\{E[X|\mu]-\kappa\} \\ &+ E[\Delta^+(S^{(x)},F^{(x)})_{\mu},P^{(x)};\kappa;A^{(1)}) - \Delta^-((S^{(y)},F^{(y)}_{\mu};P^{(y)};\kappa;A^{(1)})] \\ &+ E[V((S^{(xy)},F^{(xy)}_{\mu},P^{(xy)};\kappa;A^{(2)}) \\ &\qquad - V((S^{(yx)},F^{(yx)})_{\mu},P^{(yx)};\kappa;A^{(2)})]. \qquad (2.5)\end{aligned}$$

Proof. Write

$$\Delta(\mu,p;\kappa;A) = \alpha_1\{E[X|\mu]-\kappa\}$$
$$+ E[V((S^{(x)},F^{(x)})_\mu,P^{(x)};\kappa;A^{(1)}) - V((S^{(y)},F^{(y)})_\mu,P^{(y)};\kappa;A^{(1)})]. \quad (2.6)$$

In (2.6), replace V by $\Delta^+ + V^{(y)}$ and -V by $-\Delta^- - V^{(x)}$:

$$\Delta(\mu,p;\kappa;A) = \alpha_1\{E[X|\mu]-\kappa\}$$
$$+ E[\Delta^+((S^{(x)},F^{(x)})_\mu,P^{(x)};\kappa;A^{(1)}) + V^{(y)}((S^{(x)},F^{(x)})_\mu,P^{(x)};\kappa;A^{(1)})]$$
$$- E[\Delta^-((S^{(y)},F^{(y)})_\mu;P^{(y)};\kappa;A^{(1)}) + V^{(x)}((S^{(y)},F^{(y)})_\mu,P^{(y)};\kappa;A^{(1)})]. \quad (2.7)$$

In the first expectation on the right-hand side of (2.7) write $V^{(y)}$ as $\kappa + V$ and in the second, write $V^{(x)}$ as $E[X|((S^{(y)},F^{(y)})_\mu] + V$. Then (2.5) follows from the linearity of the expectation. □

Let $\tau = (xyy\cdots)$ be a deterministic strategy which indicates x initially and y at every subsequent stage independently of the accumulating data. Let $\sigma = (yxx\cdots)$ also be deterministic indicating y initially then x at all subsequent stages. Let S_j^τ, F_j^τ, and P_j^τ be the random number of successes, failures, and bank size at time j when following τ and similarly for σ. Then the decomposition of Δ in Theorem 2.4 can be extended to n time periods.

Theorem 2.5. For all μ, p, and κ, the delta function satisfies:

$$\Delta(\mu,p;\kappa;A) = \{\alpha_1 - \sum_{j=2}^{n} \alpha_j\}\{E[X|\mu]-\kappa\}$$

$$+ \sum_{j=1}^{n-1} E[\Delta^+((S_j^\tau,F_j^\tau)\mu,P_j^\tau;\kappa;A^{(j)}) - \Delta^-((S_j^\sigma,F_j^\sigma)\mu,P_j^\sigma;\kappa;A^{(j)}|\mu]$$

$$+ E[V((S_n^\tau,F_n^\tau)\mu,P_n^\tau;\kappa;A^{(n)}) - V((S_n^\sigma,F_n^\sigma)\mu,P_n^\sigma;\kappa;A^{(n)})|\mu]. \qquad (2.8)$$

Proof. Iterate the replacement of V by $\Delta^+ + V^{(y)}$ and -V by $-\Delta^- - V^{(x)}$ in the right hand side of (2.7). Equation (2.8) follows after n-1 iterations. □

Theorem 2.6 provides a recursive formula for Δ when A has finite horizon.

Theorem 2.6. For all μ, p, and κ, the delta function satisfies:

$$\Delta(\mu,p;\kappa;A) = (\alpha_1-\gamma_2)\{E[X|\mu]-\kappa\}$$

$$+ \sum_{j=1}^{\infty} E[\Delta^+((S_j^\tau,F_j^\tau)\mu,P_j^\tau;\kappa;A^{(j)}) - \Delta^-((S_j^\sigma,F_j^\sigma)\mu,P_j^\sigma;\kappa;A^{(j)})]. \qquad (2.9)$$

Proof. Let $n \longrightarrow \infty$, in (2.8). Then (2.9) follows from the bound in Theorem 4.1 of Eick (1985). □

3. Characterization Theorems

3.1. Monotonicity Results

The upcoming Theorem 3.1 shows for discount sequences satisfying $\alpha_j \geq \gamma_{j+1}$ (see Section 1.1 for notation) that Δ is nondecreasing in s and nonincreasing in f and κ. This characterizes the optimal strategy in the following sense. For

fixed s, f, and p there exists a κ^* such that arm x is optimal if and only if $\kappa \leq \kappa^*$. Similarly, for fixed s, κ and p there exists an f^* such that arm x is optimal if and only if $f \leq f^*$ and for fixed f, κ, and p there exists an s^* 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, $A = (1,\alpha,\alpha^2,\cdots)$, when $\alpha \leq 1/2$.

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 3.1. Suppose the discount sequence A satisfies

$$\alpha_j \geq \gamma_{j+1} \tag{3.1}$$

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

$$\Delta((s,f)\mu,p+1;\kappa;A) \geq \Delta((s,f+1)\mu,p;\kappa;A). \tag{3.2}$$

If μ is not concentrated at a single point and there is strict inequality in (3.1) for j = 1, then Δ is increasing in s and decreasing in f and κ. Furthermore, in this case there is strict inequality in (3.2).

The proof of Theorem 3.1 will be developed gradually in Lemmas 3.2, 3.3, and 3.4. Lemma 3.2 shows that Theorem 3.1 holds under the additional assumption that A has finite horizon. Lemma 3.3 extends the result to infinite horizons and Lemma 3.4 proves "strictness."

Lemma 3.2. Theorem 3.1 holds when the horizon of A is finite.

Remark. The proof of this lemma depends on the following observation which is proved in Theorem 6.4 and Lemma 6.6 of Eick (1985). Let τ and σ be as defined in Section 2.4. The distributions of θ, P_j^τ, and P_j^σ are stochastically monotone: for any w and p, $P\{\theta \geq w|(s,f)\mu\}$, $P\{P_j^\tau \geq p|(s,f)\mu\}$, and $P\{P_j^\sigma \geq p|(s,f)\mu\}$ are nondecreasing in s and nonincreasing in f.

Proof of Lemma 3.2. Proceed by induction. When A has horizon 1 the results are trivial. In this case $A = A_1$ and (3.2) 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 (2.9):

$$\Delta((s,f)\mu,p;\kappa;A) = (\alpha_1-\gamma_2)\{E[X|(s,f)\mu]-\kappa\}$$

$$+ \sum_{j=1}^{n-1} E[\Delta^+((s+S_j^\tau,f+F_j^\tau)\mu,P_j^\tau;\kappa;A^{(j)}) - \Delta^-((s+S_j^\sigma,f+F_j^\sigma)\mu,P_j^\sigma;\kappa;A^{(j)})]. \quad (3.3)$$

The first term on the right-hand side of (3.3) is nondecreasing in s and nonincreasing in f and κ.

Consider the jth term in the sum:

$$E[\Delta^+((s+S_j^\tau,f+F_j^\tau)\mu,P_j^\tau;\kappa;A^{(j)}) - \Delta^-((s+S_j^\sigma,f+F_j^\sigma)\mu,P_j^\sigma;\kappa;A^{(j)})]. \quad (3.4)$$

I show the first term in (3.4) is nondecreasing in s and nonincreasing in f and κ; the second is similar. For notational ease let $S_j = S_j^\tau$, $F_j = F_j^\tau$, and $P_j = P_j^\tau$. Write $S_j = S_{j-1} + P_j$ and $F_j = p + 1 - P_j$, where $S_{j-1} = P_1 + \cdots + P_{j-1}$. Then

$$\Delta^+((s + S_j, f + F_j)\mu, P_j; \kappa; A^{(j)}) =$$

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

By induction,

$$\Delta^+((s + s_{j-1} + p_j, f + p + 1 - p_j)\mu, p_j; \kappa; A^{(j)}) \qquad (3.5)$$

is nondecreasing in s, s_{j-1}, and nonincreasing in f and κ. Also (3.2) implies that (3.5) is nondecreasing in p_j. However, P_j, conditional on P_{j-1} and θ, is binomial and hence stochastically nondecreasing in θ and P_{j-1}. Therefore

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

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

$$E[\Delta^+((s + S_{j-1} + P_j, f + p + 1 - P_j)\mu, P_j; \kappa; A^{(j)}) | \theta] \qquad (3.6)$$

is nondecreasing in s and θ, and nonincreasing in f and κ. Finally, since θ is stochastically monotone the expectation of (3.6) when $\theta \sim (s,f)\mu$ is nondecreasing in s and nonincreasing in f and κ.

To complete the induction I show that each term in (3.4) satisfies

$$\Delta((s,f)\mu, p+1; \kappa; A) - \Delta((s,f+1)\mu, p; \kappa; A) \geq 0.$$

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 (3.4); the second is

analogous. Write the jth difference of Δ^+ functions as

$$\Delta^+((s+S_j^*, f+F_j^*)\mu, P_j^*; \kappa; A^{(j)}) - \Delta^+((s+S_j, f+1+F_j)\mu, P_j; \kappa; A^{(j)})$$

$$= \Delta^+((s + P_1^* + \cdots + P_j^*, f + p + 2 - P_j^*)\mu, \kappa; P_j^*; A^{(j)})$$

$$- \Delta^+((s + P_1 + \cdots + P_j, f + p + 2 - P_j)\mu, k; P_j; A^{(j)}). \qquad (3.7)$$

Then the random number of failures in the respective terms on the right-hand side of (3.7) 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 (3.2). □

The next step in the proof of Theorem 3.1 is to extend Lemma 3.2 to infinite horizons.

Lemma 3.3. Theorem 3.1 holds when the horizon of A is infinite.

Proof. This is immediate from Lemma 3.2 since Δ is continuous in A. □

The proof of Theorem 3.1 is completed by proving the strictness assertion.

Lemma 3.4. If μ is not concentrated at a single point and there is strict inequality in (3.1) for $j = 1$, then Δ is increasing in s and decreasing f and κ. Furthermore, (3.2) holds with strict inequality.

Proof. Strictness in (3.1) and the hypothesis on μ implies the first term in (3.3) is increasing in s, decreasing in f and κ, and satisfies (3.2) with

strictness. □

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

Corollary 3.5. Suppose A is geometric with $\alpha \leq 1/2$. Assume arm y is optimal at stage 1 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 stages.

Proof. Since y is optimal initially, then $0 \geq \Delta(\mu,p;\kappa;A)$. Since $\alpha \leq 1/2$, the regularity condition of Theorem 3.2 is satisfied. From (3.2):

$$0 \geq \alpha\Delta(\mu,p;\kappa;A) = \Delta(\mu,p;\kappa;A^{(1)})$$

$$\geq \Delta((0,1)\mu,p-1;\kappa;A^{(1)}) \geq \cdots \geq \Delta((0,p)\mu,0;\kappa;A^{(1)}).$$

Then arm y is optimal in the $((0,p)\mu,0;\kappa;A^{(1)})$-bandit. When arm y is selected 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. □

4. Discussion

In this paper I characterize optimal strategies for two-armed delayed response bandits. Under a regularity condition on the discount sequence Theorem 3.1 shows that $\Delta((s,f)\mu,p;\kappa;A)$ is nondecreasing in s and nonincreasing in f and κ for each fixed p. This partitions the state space into connected regions

where x is optimal and where y is optimal. On the boundary, both x and y are optimal. The decision regions are determined by the sign of Δ and the boundary by the manifold where Δ vanishes. For each fixed p and A, Δ defines a different manifold in $(s,f;\kappa)$-space. When x is optimal and s is increased, x remains optimal from the monotonicity of Δ. Similarly, when y is optimal and f or κ is increased, y remains optimal. Equation (3.2) is a relationship between the manifolds for p and p + 1.

Acknowledgment

I wish to thank Donald A. Berry for his many helpful suggestions.

REFERENCES

Bather, J.A. (1981). Randomized allocations of treatments in sequential experiments (with discussion). J.R. Statist. Soc. B. 43:265-292.

Bellman, R. (1956). A problem in sequential design of experiments. Sankhya A 16:221-229.

Berry, D.A. (1972). A Bernoulli two-armed bandit. Ann. Math. Statist. 43:872-897.

Berry, D.A. and Fristedt, B. (1985). Bandit Problems: Sequential Allocation of Experiments. Chapman-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. 27:1060-1070.

Eick, S.G. (1985). Two-armed bandits with delayed responses. Univ. of Minnesota Statistics Tech. Rep. No. 456.

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. Progress in Statistics (ed. by J. Gani, et al.), pp. 241-266. North-Holland, Amsterdam.

Kumar, P.R., and Seidman, T.I. (1981). On the optimal solution of the one-armed bandit adaptive control problem. IEEE Trans. Autom. Control. 26:1176-1189.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bull. Amer. Math. Soc. 58:527-536.

Rodman, L. (1978). On the many-armed bandit problem. Ann. of Prob. 6:491-498.

Whittle, P. (1980). Multi-armed bandits and the Gittins Index. J.R. Statist. Soc. B 42:143-149.