Two-armed Bandits with Delayed Responses

by

Stephen E. Eick*

Technical Report No. 456
December 1985

---

* Partially supported under NSF grants DMS 8301450 (D.A. Berry, principal investigator) and DMS8319924 (M.L. Eaton, principal investigator) and the University of Minnesota Statistics Alumni Fellowship. Adapted in part from the author's dissertation.

Summary

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 mean lifetime λ. The response delays completely change the character of 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 investigated as a function of p, μ, and κ.

1. Introduction

Consider a clinical trial in which patients arrive sequentially at times 0, 1,$\cdots$, n - 1 (n = ∞ 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 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 relative treatment effectiveness accrues. The objective is to assign treatments to maximize total patient survival time, possibly discounting for future successes.

Bandit problems have been studied extensively in the statistical literature. Authors making significant contributions include Robbins (1952),

Bradt, Johnson, and Karlin (1956), Bellman (1956), Gittins and Jones (1974), 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 strategies in clinical trials (see Armitage, 1985, p.22, and Simon, 1977). I address the problem of maximizing the expected total patient lifetime when treatment assignment is based on partial information, the censored lifetimes, rather than the exact lifetimes.

There are two arms, x and y. I assume that arm x is unknown: the lifetimes of patients treated with x are conditionally iid given unknown parameters. On the other hand, arm y is known: the lifetimes of patients treated with y are identical and known, or what is equivalent in the current setting, are random variables with a common, known mean. I take a Bayesian approach and assume that the unknown parameters are themselves random variables. As the trial proceeds, Bayes' theorem is used to update the prior distribution. This approach facilitates allowing treatment assignments to depend on the accumulating data.

1.1. Distributional Assumptions

Let Z_j be the lifetime of the patient treated at time $j - 1$. If this patient receives treatment x, or in other words if arm x is pulled at stage j, then $Z_j = X_j$. If, however, this patient receives treatment y then $Z_j = Y_j$. I

assume that $X_1, \cdots, X_n$ given $\theta \in (0,1)$ are conditionally iid geometric random variables with probability mass function

$$(1-\theta)\theta^t, \quad t = 0,1,2,\cdots. \tag{1.1}$$

This is consistent with the assumption that the probability of a patient surviving any particular time period is constant and equals θ, and that, given θ, the time periods are independent within and across patients. The patients treated with y have known expected lifetime κ:

$$E[Y_j] = \kappa, \quad j = 0,1,2,\cdots.$$

The random variable θ has prior distribution μ. The conditional expected lifetime of a patient treated with x is $E[X|\theta] = \theta/(1-\theta)$. I restrict consideration to those μ for which $E[X|\mu] = E[\theta/(1-\theta)|\mu] < \infty$. One consequence of this restriction is that $\{\theta = 1\}$ is a μ-null event.

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 smaller life expectancy than one treated with y. This conflict between effective treatment and gathering information characterizes bandits more generally (Berry and Fristedt 1985).

Based on (1.1), a sufficient statistic for θ is (S,F), where S is the number of patient survival periods for those treated with x and F is the number of failures observed on x. I denote the conditional distribution of θ given $(S,F) = (s,f)$ by $(s,f)\mu$. When $s = f = 0$, $(0,0)\mu = \mu$. Temporarily, $(s,f)\mu$ is

defined only for the integer pairs (s,f) which occur with positive probability under μ. However, in Chapter 5 I extend the domain of $(s,f)\mu$ to continuous arguments.

1.2. Summary of Results

The major results in this paper concern the value of the bandit, which is the expected discounted patient lifetime when the best treatment allocation is used. I define the value in Chapter 2. In Chapter 3 I use dynamic programming to calculate the value of finite horizon bandits, those in which only finitely many patients are treated. An interesting result in Section 3.2 is that in general the delayed response bandit is not a stopping problem. The bandit state summarizes all information available when the current patient must be treated. Chapter 4 investigates the value as a function of the state for fixed μ. Chapter 5 extends μ to a family of distributions which generalize the beta family. Chapter 6 considers the value as a function of the distribution for μ in this class.

2. Notation

2.1. The Discount Sequence

The discount sequence, $A = (\alpha_1,\alpha_2,\cdots)$, is a summable sequence of nonnegative numbers. It determines the weights associated with patients yet to be treated and incorporates the unknown aspects of the number of patients yet to be treated (see Berry and Fristedt, 1985, Ch. 3). The horizon of A is the index of the last nonzero element in A: horizon $A = \inf\{j:\alpha_i = 0 \text{ for all } i>j\}$. If

this set is empty the horizon of A is defined to be ∞.

Some important discount sequences are

$$A = (1,1,\cdots,1,0,0,\cdots), \tag{2.1}$$

$$A = (1,\alpha,\alpha^2,\cdots), \tag{2.2}$$

$$A = (1,\alpha,\alpha^2,\cdots, \alpha^{n-1},0,0,\cdots). \tag{2.3}$$

Discount sequence (2.1) is called the uniform. This sequence models a trial in which the number of patients treated is precisely known. In the previous example the discount sequence was uniform with horizon 3.

Sequence (2.2) is the geometric with factor α. For the sequence to be summable, $\alpha < 1$. The uniform and geometric are the most common discount sequences in the literature. Both (2.1) and (2.2) are special cases of (2.3) with $\alpha = 1$ and $n = \infty$, respectively.

After j patients have been treated the appropriate discount sequence for the bandit presenting itself at that time is

$$A^{(j)} = (\alpha_{j+1},\alpha_{j+2},\cdots).$$

The discount sequence $A^{(j)}$ is derived from A by deleting the first j elements in A. For the geometric discount sequence,

$$A^{(j)} = \alpha^j A,$$

which differs from the original sequence by a positive multiple. This property often makes the geometric more tractable than the other discount sequences.

In many situations infinite horizon discount sequences are much more difficult to work with than finite horizon sequences. The horizon n truncation of A,

$$A_n = (\alpha_1, \alpha_2, \cdots, \alpha_n, 0, \cdots),$$

is often a convenient approximation to A for n large.

The jth tail mass of A is the sum of the discount sequence $A^{(j-1)}$:

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

By assumption, $\gamma_j < \infty$ for every j.

2.2. The State Space

The bandit state consists of three components. The first component is the pair (μ,p) where μ is the current distribution of θ and p is the number of patients which have been treated with x and are still surviving. These patients form an "information bank." Information accrues with time from this "bank" as patients respond, positively or negatively. The second component is κ. This component does not include patients treated with y because they cannot change κ. The third component is the discount sequence A. A bandit with state $(\mu,p;\kappa;A)$ is called the $(\mu,p;\kappa;A)$-bandit.

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 $A = (1,1,1,0,0,\cdots)$. 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 state is $((1,0)\mu,1;\kappa;A^{(1)})$. There has been one success, $S = 1$, no failures, $F = 0$, one patient is in the information bank, $P = 1$, and two patients remain to be treated, $A^{(1)} = (1,1,0,0,\cdots)$.

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, $S = 2$, and no failures have been observed, $F = 0$. The information bank still contains one x-observation, $P = 1$; the y-observation is not in the information bank because the distribution of y is known. One patient remains to be treated, $A^{(2)} = (1,0,0,\cdots)$.

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

2.3. Strategies

A strategy or policy τ is a function defined on the state space, $\tau:\{(\mu,p;\kappa;A)\} \longrightarrow \{x,y\}$, indicating which treatment to use when the current state is $(\mu,p;\kappa;A)$. The treatment indicated for any patient can depend on past selections, the censored results, and the future number of patients to be treated. In the previous example it was known at time 1 that the first patient had received treatment x and had survived one time period; the second selection

(y in the example) can depend on this information.

Recall that Z_j is the lifetime of the patient treated at time j - 1. The worth of a strategy τ is the expected discounted patient lifetime,

$$W(\tau) = E_\tau[\textstyle\sum_1^\infty \alpha_j Z_j], \tag{2.4}$$

where Z's in (2.4) are determined by a strategy τ. The objective is to choose a strategy τ which maximizes (2.4).

2.4. An Example

The example in this Section illustrates the state space, strategies, and their worths in a simple, but concrete setting.

Suppose the discount sequence for the $(\mu,p;\kappa;A)$-bandit is $A = (2,1,0,\cdots)$. Assume θ has density $2(1-u)1_{(0,1)}(u)du$, and assume $\kappa = 1$. Let τ be a strategy indicating x for the first patient. This patient's expected lifetime is

$$E[X] = E[E[X|\theta]] = \int_0^1 \{\theta/(1-\theta)\}2(1-\theta)d\theta = 1.$$

Let $S^{(x)}$ $F^{(x)}$ and $P^{(x)}$ denote the random number of successes, failures, and information bank size at time 1 given that x was selected at time 0. There are two possibilities for the bandit state depending on whether the first patient survives or fails:

$$\left((S^{(x)},F^{(x)})\mu,P^{(x)};\kappa;A^{(1)}\right) = \left((1,0)\mu,1;\kappa;A^{(1)}\right) \tag{2.5}$$

or

$$\left((S^{(x)},F^{(x)})\mu,P^{(x)};\kappa;A^{(1)}\right) = \left((0,1)\mu,0;\kappa;A^{(1)}\right), \qquad (2.6)$$

with probabilities 1/3 and 2/3, respectively. Given a success the conditional expected lifetime of the next patient treated with x is 2 and given a failure the conditional expected lifetime is 1/2.

At time 1 the second patient must be treated. There are two possible selections for both of the possible states at time 1. Let τ_z be the strategy indicating z at stage 2 regardless of the result from stage 1, where z = x or y. The remaining two possibilities are τ_{xy} which indicates x if (2.5) and y if (2.6) and τ_{yx} which indicates y if (2.5) and x if (2.6).

Then the worths of the various strategies are

$$W(\tau_x) = 2 + (1/3)2 + (2/3)(1/2) = 3,$$

$$W(\tau_y) = 2 + \kappa = 3,$$

$$W(\tau_{xy}) = 2 + (1/3)2 + (2/3)\kappa = 10/3,$$

$$W(\tau_{yx}) = 2 + (1/3)\kappa + (2/3)(1/2) = 8/3.$$

The best among those strategies which indicate x initially is τ_{xy}. When a success is observed initially the strategy τ_{xy} indicates x again. This is intuitively plausible since a success on x suggests that θ is large. Conversely, when a failure is observed on x, θ is likely to be small. In this case τ_{xy} indicates y for the second patient.

Strategies τ_x and τ_y ignore the result of the first treatment. The worst

strategy is τ_{yx}. When an initial success is observed on x it indicates y. When an initial failure is observed on x it indicates x again. The initial success on x suggests the superiority of x and so indicating y after a success on x has dubious merit. Similarly, indicating x after an initial failure is unreasonable.

The other possible initial selection is y. Let σ be a strategy indicating y initially. In this case the state at time 1 is $(\mu,0;\kappa;A^{(1)})$ and there are two possible selections. Let σ_x indicate x at time 1 and σ_y indicate y. Then

$$W(\sigma_x) = 2 + 1 = 3$$

$$W(\sigma_y) = 2 + 1 = 3.$$

The strategy maximizing the worth over all strategies is τ_{xy}. It is not surprising that this strategy indicates arm x initially. Since $E[X|\mu] = \kappa$, treating the first patient with x has the same expected lifetime as y but also provides information about θ. This result is true in general for two-armed bandits with one arm known.

2.5. The Sate for the Strategy τ

The distribution of the state at a fixed point in time depends on the past selections. For any strategy τ, let S_j^τ, F_j^τ, and P_j^τ be the random number of arm x successes, failures, and patients who are then in the bank at time j when following τ. Then the state at time j when following τ is $((S_j^\tau,F_j^\tau)\mu,P_j^\tau;\kappa;A^{(j)})$. When $j = 1$, I suppress j and write $S^{(z)}$, $F^{(z)}$, and $P^{(z)}$, where the initial selection is $z = x$ or y.

There is a simple relationship between S_j^τ, F_j^τ, and P_j^τ. Since P_j^τ is the number of patients surviving from time $j-1$ to j, S_j^τ is the sum of all P_i^τ for $i \leq j$. Similarly F_j^τ is the total number of patients treated with x, including any initially in the information bank, minus P_j^τ. The geometric lifetime assumption implies that P_j^τ is conditional on P_{j-1}^τ and θ has a binomial distribution.

The following lemma summarizes these relationships. Let

$$I_j^\tau = \begin{cases} 1 \text{ if } \tau \text{ indicates x at time j,} \\ 0 \text{ if } \tau \text{ indicates y at time j.} \end{cases}$$

Lemma 2.1. At time j the random variables S_j^τ, F_j^τ, and P_j^τ satisfy

$$S_j^\tau = \sum_{i=1}^{j} P_i^\tau,$$

$$F_j^\tau = p + \sum_{i=1}^{j-1} I_i^\tau - P_j^\tau,$$

$$\{P_j^\tau | P_{j-1}^\tau, \theta\} \sim \text{bin}[P_{j-1}^\tau + I_{j-1}^\tau, \theta]. \tag{2.7}$$

2.6. The Value of a Bandit

The value of the $(\mu,p;\kappa;A)$-bandit is the supremum of the worth of at over all strategies τ:

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

where $W(\tau)$ is defined at (2.4). The supremum over strategies that indicate x

initially is

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

and $V^{(y)}$ is defined analogously. A strategy τ which attains V is said to be optimal. Arm z is optimal if there exists an optimal strategy τ which indicates that arm initially, that is, if $V^{(z)} = V$.

An intuitively reasonable strategy is one that always indicates the arm which currently has the larger expected lifetime. Such a strategy is said to be myopic. In the example in Section 2.4 the strategies τ_{xy}, σ_x, and σ_y are all myopic. However, only τ_{xy} is optimal. When A has horizon 1 a myopic strategy is optimal. In this case the value function does not depend on p. More generally, as Section 2.4 shows, a myopic strategy is optimal only in the most special settings.

3. The Dynamic Programming Solution

3.1. The Fundamental Equation of Dynamic Programming

The value function satisfies the fundamental equation of dynamic programming:

$$V(\mu,p;\kappa;A) = V^{(x)}(\mu,p;\kappa;A) \vee V^{(y)}(\mu,p;\kappa;A), \tag{3.1}$$

where

$$V^{(x)}(\mu,p;\kappa;A) = \alpha_1 E[X|\mu] + E[V((S^{(x)},F^{(x)}\mu,P^{(x)};\kappa;A^{(1)})|\mu], \tag{3.2}$$

$$V^{(y)}(\mu,p;\kappa;A) = \alpha_1 \kappa + E[V((S^{(y)},F^{(y)},\mu,P^{(y)};\kappa;A^{(1)})|\mu]. \tag{3.3}$$

When A has horizon $n < \infty$, (3.1), (3.2), and (3.3) can be used to calculate V recursively. The starting points are all possible states for which the discount sequence is $A^{(n-1)}$, which has horizon 1:

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

Calculating (3.2) and (3.3) by first conditioning on θ and using (2.7) leads to:

$$V^{(x)}(\mu,p;\kappa;A) = \alpha_1 E[X|\mu] +$$

$$\sum_{j=0}^{p+1} \binom{p+1}{j} E[\theta^j(1-\theta)^{p+1-j}|\mu]V((j,p+1-j)\mu,j;\kappa;A^{(1)}), \tag{3.4}$$

$$V^{(y)}(\mu,p;\kappa;A) = \alpha_1\kappa +$$

$$\sum_{j=0}^{p} \binom{p}{j} E[\theta^j(1-\theta)^{p-j}|\mu]V((j,p-j)\mu,j;\kappa;A^{(1)}). \tag{3.5}$$

The difference in binomial weighting terms between (3.4) and (3.5) is due to the different number of possible successes. For $V^{(x)}$ there are $p + 1$ patients in the information bank for arm x who can survive to time 1, but for $V^{(y)}$ only p can survive on arm x.

A consequence of the upcoming Theorem 6.7 is that the value terms in (3.4) and (3.5) are ordered: for $j = 0,\cdots,p$,

$$V((j,p+1-j)\mu,j;\kappa;A^{(1)}) \leq V((j,p-j)\mu,j;\kappa;A^{(1)})$$
$$\leq V((j+1,p-j)\mu,j+1;\kappa;A^{(1)}). \tag{3.6}$$

The intuition behind (3.6) is that observing one less failure or one more success increases the conditional expected lifetime of patients treated with x.

Arm x is optimal if and only if $V^{(x)} \geq V^{(y)}$. This holds when the weighted average of the p + 1 terms $V((j,p+1-j)\mu,j;\kappa;A)$ exceeds that of the p terms $V((j,p-j)\mu,j;\kappa;A)$ by more than $\kappa - E[X|\mu]$, the initial life expectancy difference. So arm x is optimal when the information gained from treating the first patient with x leads to future allocations that make up for the loss due to treating the first patient with x instead of with y.

3.2. Stopping Problems

In the two-armed bandit with arm y known one might expect the optimal strategy to indicate arm x initially if ever. For there is more opportunity to take advantage of whatever is learned if arm x is pulled sooner rather than later. This result is true in some generality in the classical bandit (Berry and Fristedt, 1979, Section 2). When it is true there is an optimal strategy which indicates x until stage N and then indicates y at all subsequent stages. The stopping stage N is random and can be 0 or ∞ with positive probability.

Surprisingly, simple examples show that this result is not true in the current setting with $p > 0$. The intuitive reason is that arm y can be selected initially while waiting for patients in the information bank of arm x to respond. The next theorem says that such a counterexample is not possible when $p = 0$ and the discount sequence A is either geometric or uniform.

Theorem 3.2. Suppose $p = 0$ and either (i) A is geometric with discount factor α, or (ii) A is uniform with horizon n. Then arm x is optimal initially, if ever.

Proof. I prove case (i) only; case (ii) is proved by assuming that arm y is optimal and interchanging the first x selection with the initial selection on y. Suppose y is uniquely optimal in the $(\mu,0;\kappa;A)$-bandit. If y is selected at stages 1 to n for $n \geq 1$, then the $(\mu,0;\kappa;\alpha^n A)$-bandit presents itself. But the optimal selections in the $(\mu,0;\kappa;A)$- and $(\mu,0;\kappa;\alpha^n A)$-bandits are identical because the respective discount sequences differ only by a positive multiple. So if x is optimal for the first time at some stage in the future, it is also optimal initially. □

4. Properties of the Value Function

4.1. A Bound on the Value Function

The following theorem provides upper and lower bounds for the value of the $(\mu,p;\kappa;A)$-bandit. I use this theorem to extend finite horizon results to infinite horizons. For the upper bound the value is less than it would be if the experimenter were acting optimally for the $(\mu,p;\kappa;A_n)$-bandit and were to be told the value θ at stage $n + 1$. For the lower bound the value exceeds that of the $(\mu,p;\kappa;A_n)$-bandit plus a correction for stages $n + 1$ to ∞.

Theorem 4.1. For any $(\mu,p;\kappa;A)$-bandit,

$$V(\mu,p;\kappa;A_n) + \gamma_{n+1}\{E[\tfrac{\theta}{1-\theta}|\mu] \vee \kappa\} \leq V(\mu,p;\kappa;A)$$

$$\leq V(\mu,p;\kappa;A_n) + \gamma_{n+1}E[\tfrac{\theta}{1-\theta} \vee \kappa|\mu]. \qquad (4.1)$$

Proof. This is a standard result in the bandit literature. See, for example, Berry and Fristedt (1985), Theorem 2.6.1. □

An easy consequence of Theorem 4.1 is that $V(\mu,p;\kappa;A)$ is continuous in A when A has the ℓ_1 topology.

4.2. The Value as a Function of κ

The next theorem says that $V(\mu,p;\kappa;A)$ is an increasing function of κ. This result is very intuitive since the expected lifetime of any patient treated with y increases as κ increases.

Theorem 4.2. The value function $V(\mu,p;\kappa;A)$ is a continuous, convex, nondecreasing function of κ for all μ, p, and A.

Proof. Continuity follows from convexity since all convex functions are continuous. The proof of convexity and monotonicity is divided into two parts. When the horizon of A is finite, convexity follows by induction using (3.4), (3.5), and (3.1). When the horizon is infinite the result follows by approximating $V(\mu,p;\kappa;A)$ with $V(\mu,p;\kappa;A_n)$. □

An argument similar to the proof of Theorem 4.2 shows that $V(\mu,p;\kappa;A)$ is actually piecewise linear in κ when the horizon of A is finite.

4.3. Value as a Function of p

The next result is that $V(\mu,p;\kappa;A)$ is nondecreasing as a function of p. The proof, which is omitted, depends on the following idea: if the number of patients in the information bank, p, is increased, the experimenter can ignore the additional information and do at least as well.

Theorem 4.3. The value function $V(\mu,p;\kappa;A)$ is nondecreasing in p for all μ, κ, and A.

As $p \longrightarrow \infty$, the value of the $(\mu,p;\kappa;A)$-bandit converges to the expected lifetime of the first patient treated plus the value from stage 2 on if the experimenter were to be told θ at stage 2.

Theorem 4.4. For all μ, κ, and A, As $p \longrightarrow \infty$,

$$V(\mu,p;\kappa;A) \longrightarrow \alpha_1[E[X|\mu] \vee \kappa] + \gamma_2 E[\tfrac{\theta}{1-\theta} \vee \kappa|\mu].$$

Proof. I will show that

$$V^{(x)} \longrightarrow \alpha_1 E[X|\mu] + \gamma_2 E[\tfrac{\theta}{1-\theta} \vee \kappa|\mu]. \tag{4.4}$$

An analogous result holds for $V^{(y)}$.

Given θ, the number of patients who survive to time 1 is binomial with sample size p and probability of success θ. The sufficient statistics for θ at time 1 are $S^{(x)}$ and $F^{(x)}$. As $p \longrightarrow \infty$, the posterior distribution of $(\theta|S^{(x)},F^{(x)})$ converges weakly to the true value θ^*, say. If $\text{supp}(\mu) \subset [0,1-t]$

for some $t > 0$, then $\theta/(1-\theta)$ is bounded on the support of μ. Hence

$$E[\tfrac{\theta}{1-\theta} \vee \kappa | (S^{(x)},F^{(x)})\mu] \longrightarrow E[\tfrac{\theta}{1-\theta} \vee \kappa | \delta_{\theta^*}] = \tfrac{\theta^*}{1-\theta^*} \vee \kappa \qquad (4.5)$$

by the weak convergence theorem. Under this assumption, (4.4) follows from (4.5). The result for general μ (with finite life expectancy as required in the introduction) follows by approximation. □

A consequence of Theorem 4.4 is that for sufficiently large p the initial selection should maximize the life expectancy of the first patient treated. That is, for this case the myopic strategy is optimal.

5. Prior Distributions

This section introduces a family of prior distributions which generalize the beta distribution. I extend the prior μ to a family of distributions $(s,f)\mu$ where $\mu = (0,0)\mu$ and s and f are continuous. I continue to assume that μ is not a one-point measure and I also assume in this chapter that $\mu(\{0,1\}) = 0$.

5.1. Prior Distributions of the Form $(s,f)\mu$

Define $(s,f)\mu$ by

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

where

$$b(s,f) = \int_0^1 \theta^s(1-\theta)^f d\mu(\theta). \qquad (5.1)$$

The parameters s and f are real numbers restricted so that $(s,f)\mu$ exists and $E[X|(s,f)\mu] < \infty$. For nonnegative, integral s and f this definition agrees with the previous definition of $(s,f)\mu$ as the conditional distribution of $(\theta|s,f)$. The dominating measure μ may be any nondegenerate measure on (0,1) such that (5.1) is finite. This family of prior distributions is natural in view of (3.4) and (3.5).

The parameters s and f are interpreted as the prior number of successes and failures on arm x. The prior $(s,f)\mu$ would be the posterior distribution of θ if s successes and f failures were observed when $\theta \sim \mu$.

Definition 5.1. Let $\Omega = \Omega(\mu)$ be the set of all (s,f) for which $(s,f)\mu$ is defined, and for which $E[X|(s,f)\mu] < \infty$. Then Ω is the consideration region for μ.

In the sequel I assume that $(s,f) \in \Omega$. The next proposition characterizes the consideration region.

Proposition 5.2. The pair $(s,f) \in \Omega$ if and only if $b(s,f-1) < \infty$.

This class of distributions generalizes the beta distribution which is the subject of Example 5.3.

Example 5.3. Let

$$d\mu(\theta) = \theta^{-1}(1-\theta)^{-1}1_{(0,1)}(\theta)d\theta,$$

then

$$b(s,f) = \int_0^1 \theta^{s-1}(1-\theta)^{f-1}d\theta = be(s,f)$$

and $(s,f)\mu$ has a beta distribution with parameters (s,f) and density

$$d(s,f)\mu(\theta) = \theta^{s-1}(1-\theta)^{f-1}1_{(0,1)}d\theta/b(s,f).$$

The beta function, $b(s,f)$, is finite whenever both s and f are positive. The mean lifetime on x is

$$E[\theta/(1-\theta)|s,f] = b(s+1,f-1)/b(s,f) = s/(f-1),$$

which is finite if and only if $f > 1$. In this case the consideration region is $\Omega = (0,\infty)x(1,\infty)$. □

5.2. The Value in Terms of b-functions

Theorem 5.6 expresses the value of the $((s,f)\mu,p;\kappa;A)$-bandit in terms of maxima of linear combinations of b-functions. The decomposition is interesting but I do not exploit it in the sequel.

Theorem 5.6. When the horizon of A is finite, $b(s,f)V((s,f)\mu,p;\kappa;A)$ is the maximum of linear combinations of b-functions.

Proof. The proof follows from induction on the horizon of A. □

6. Stochastic Monotonicity

An important property of the prior distribution $(s,f)\mu$ is a stochastic

ordering of the bandit state space. I use this ordering to prove that $V((s,f)\mu,p;\kappa;A)$ is nondecreasing in s and nonincreasing in f.

Definition 6.1. A random variable W with respect to $(s,f)\mu$ is said to be stochastically monotone if $P\{W \geq w|(s,f)\mu\}$ is nondecreasing in s and nonincreasing in f for every w. The relation is strict if there exists a w such that there is strict increase in s and decrease in f.

Stochastic monotonicity is preserved under monotone transformations. This is stated formally in the following lemma.

Lemma 6.2. Let g be a nondecreasing function. Suppose W is stochastically monotone. Then g(W) is stochastically monotone and strictly monotone provided W is strictly monotone and g is strictly increasing. Furthermore, if $E[g(W)|(s,f)\mu] < \infty$ then $E[g(W)|(s,f)\mu]$ is continuous and nondecreasing in s and nonincreasing in f.

6.1. Applications of Stochastic Monotonicity

The next theorem shows that both θ and X are stochastically monotone. Its proof depends on the following lemma which expresses the partial derivative with respect to s of $E[g(\theta)|(s,f)\mu]$ in terms of the covariance between $g(\theta)$ and $\log(\theta)$.

Lemma 6.3. Let g be a measurable function defined on (0,1) such that $E[g(\theta)|s,f] < \infty$ for $(s,f) \in \Omega$. For (s,f) in the interior of Ω,

$$\frac{\partial}{\partial s}E[g(\theta)|(s,f)\mu] = \mathrm{cov}\{\log(\theta), g(\theta)|(s,f)\mu\}. \qquad (6.1)$$

Proof. This is a standard consequence of the dominated convergence theorem. □

The stochastic monotonicity of θ and X follows.

Theorem 6.4. Suppose μ is not a one-point measure and $\mu(\{0,1\}) = 1$. Then Both θ and X are strictly stochastically monotone.

Proof. First it will be shown that θ is strictly stochastically monotone. Fix $q > 0$ and let

$$g(\theta) = \underset{\sim}{1}_{\{\theta \geq q\}} = \underset{\sim}{1}_{\{\log(\theta) \geq \log(q)\}}$$

Since $g(\theta)$ is a nondecreasing function of $\log(\theta)$, $\mathrm{cov}(g(\theta),\log(\theta)) \geq 0$. By (6.1),

$$\partial E[g(\theta)|s,f]/\partial s = \mathrm{cov}(g(\theta),\log(\theta)) \geq 0.$$

Therefore $P\{\theta \geq q|s,f\}$ is a nondecreasing function of s. Strictness follows since μ is assumed not to be concentrated at a single point. The argument that θ is a decreasing function of f is similar.

The stochastic monotonicity of θ implies that of X. For any nonnegative integer x,

$$P\{X \geq x|s,f\} = E[P\{X \geq x|\theta\}|s,f] = E[\textstyle\sum_{j \geq x}^{\infty}(1-\theta)\theta^j|s,f] = E[\theta^x|s,f],$$

which is the expectation of an increasing function of θ. □

Recall that P_n^τ is the number of patients in the information bank at time n when following strategy τ. The next theorem says that P_n^τ is stochastically monotone for strategies which ignore the accumulating data. I define these strategies next.

Definition 6.5. A strategy τ is deterministic if

$$\tau = (z_1, z_2, \cdots)$$

where each z_n is either x or y independently of any past selections and their results.

Lemma 6.6. For every n and deterministic strategy τ, P_n^τ is stochastically monotone.

Proof. I will show that

$$P\{P_n^\tau \geq x|\theta\} \tag{6.2}$$

is nondecreasing in θ. Assuming this,

$$P\{P_n^\tau \geq x|(s,f)\mu\} = E[P\{P_n^\tau \geq x|\theta\}|(s,f)\mu],$$

is the expectation of a nondecreasing function of θ. The result then follows from Lemma 6.2.

It remains to prove that (6.2) is nondecreasing in θ. Proceed by induction. Fix n = 1 and consider the distribution of P_1^τ given θ:

$$P\{P_1^\tau \geq x|\theta\} = \begin{cases} P\{\text{bin}(p,\theta) \geq x|p,\theta\} \text{ if } z_1 = y, \\ P\{\text{bin}(p+1,\theta) \geq x|p,\theta\} \text{ if } z_1 = x. \end{cases} \tag{6.3}$$

In both cases (6.3) is increasing in θ for $0 < x < p$.

The result for general n follows by conditioning on P_{n-1}^τ and using induction. □

6.2. The Value as a Function of (s,f)

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

Theorem 6.7. Suppose μ is not a one-point measure and $\mu(\{0,1\}) = 0$. For all p, κ, and A, $V((s,f)\mu,p;\kappa;A)$ is continuous as a function of (s,f), and is nondecreasing in s and nonincreasing in f.

Proof. I prove the finite horizon case by induction and then extend to infinite horizons by approximating with (4.1). When A has horizon 1 the result is immediate. Suppose the theorem holds when the horizon is $m < n$. Let A have horizon n. Consider $V^{(x)}$:

$$V^{(x)}((s,f)\mu,p;\kappa;A) = \alpha_1 E[X|(s,f)\mu] + E[V((S^{(x)}+s,F^{(x)}+f)\mu,P^{(x)};\kappa;A^{(1)})|(s,f)\mu], \tag{6.4}$$

where the horizon of $A^{(1)}$ is n - 1. Theorem 6.4 applies to the first term on the right-hand side of (6.4). Writing

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

and substituting into the second term on the right-hand side of (6.4),

$$V((S^{(x)}+s,F^{(x)}+f)\mu,P^{(x)};\kappa;A^{(1)}) =$$

$$V((P^{(x)}+s,p+1-P^{(x)}+f)\mu,P^{(x)};\kappa;A^{(1)}). \qquad (6.5)$$

The result for the second term follows from (6.5) by induction and Lemma 6.6. A similar argument applies to $V^{(y)}$. The inductive step is complete since $V = V^{(x)} \vee V^{(y)}$. □

The following corollary characterizes $V^{(x)}$ and $V^{(y)}$ as functions of (s,f) and κ.

Corollary 6.8. Suppose μ is not a one-point measure and $\mu(\{0,1\}) = 0$. The functions $V^{(x)}((s,f)\mu,p;\kappa;A)$ and $V^{(y)}((s,f)\mu,p;\kappa;A)$ are continuous in s, f, and κ, nonincreasing in s and nondecreasing in s and κ. Furthermore, provided $\alpha_1 \neq 0$, $V^{(x)}((s,f)\mu,p;\kappa;A)$ is increasing in s and decreasing in f and $V^{(y)}((s,f)\mu,p;\kappa;A)$ is decreasing in κ.

7. Discussion

In this paper I present a model for the two-armed bandit with delayed responses. The response delays introduce a new parameter p into the state space which changes the character of the solution; it is no longer a stopping problem. The important results are Theorems 4.2, 4.3, and 6.7 which show $V((s,f)\mu,p;\kappa;A)$ is nondecreasing in s, p, and κ and nonincreasing in f. These results describe the value as a function of the state but provide little insight into the optimal

strategies. In a future paper I will characterize the optimal strategies by describing the optimal arm as a function of the state.

Acknowledgment

I would like to thank Donald A. Berry for his many helpful comments.

REFERENCES

Armitage, P. (1985). The search for optimality in clinical trials. Int. Statist. Rev. 53, 1-13.

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. (1979). Bernoulli one-armed bandits--Arbitrary discount sequences. Ann. Statist. 7:1086-1105.

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.

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.

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

Simon, R. (1977). Adaptive treatment assignment methods and clinical trials. Biometrics 33:743-744.