
DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 804

March 1987

Bernoulli bandits with covariates

by

Murray K. Clayton¹
University of Wisconsin

¹Research supported in part by U.S. Army Research Office Grant DAAG29-80-C-0041 and University of Wisconsin Graduate School Grant 160701.

AMS 1980 subject classifications. Primary 62L05; Secondary 62L15.

Keywords and phrases: Sequential decisions, one-armed bandits, two-armed bandits, logit transformation.

SUMMARY

Sequential selections are to be made from two stochastic processes, or "arms", each yielding Bernoulli responses. At each stage the arm selected depends on previous observations. The objective is to maximize the expected number of successes in the first n selections. The probability of success for a given selection depends on a covariate through a logistic transformation. For one arm, this transformation is completely known; for the other, it depends on an unknown parameter. Optimal strategies are developed in terms of a break-even value for the covariate: it is optimal to observe the arm with unknown parameter if the covariate is less than the break-even value. Other properties of optimal strategies are related to those for non-covariate models.

1. Introduction

A bandit problem involves sequential selections or "pulls" from a number of stochastic processes (or "arms", machines, treatments, etc.). The available processes have unknown characteristics, so learning can take place as the processes are observed. As in Bradt, Johnson, and Karlin (1956), we shall restrict our attention to the class of finite horizon Bernoulli bandits, in which the responses are Bernoulli random variables and the goal is to maximize the expected sum of the first n observations. Berry and Fristedt (1985) discuss this and many other forms of bandit models.

The bandit model has been proposed as a model for a clinical trial. The arms represent treatments, and the goal is to allocate treatments sequentially so as to maximize the expected total number of successes. However, in a typical bandit problem, the observations made on a given arm are assumed to be exchangeable (see, for example, Berry and Fristedt, 1979). In a clinical trial, this implies that all subjects receiving the same treatment have the same (marginal) probability of success. In this paper we extend this notion by supposing that, for a given arm, the probability of success for a given subject depends both on the treatment being used and on a covariate that can encode relevant characteristics affecting the chance of success. This might include such things as the general health of the subject, age and sex of the subject, and so on.

To describe our model formally, we begin by assuming that there are two arms. Let X_i and Y_i denote the results from arms 1 and 2 respectively, at stage i : for $i \leq n$ exactly one of the pair (X_i, Y_i) is actually observed. We also assume that prior to making the i th observation we can observe a covariate S_i .

We assume that functions ρ and λ exist such that $P(X_i=1|\rho(S_i)) = \rho(S_i)$ and $P(Y_i=1|\lambda(S_i)) = \lambda(S_i)$. In what follows, we shall assume that after the $i-1^{\text{st}}$ pull and prior to the i^{th} pull, only the covariate values S_1, \dots, S_i are known. Informally, we obtain no information about the $i+1^{\text{st}}$ and later subjects until after the i^{th} subject has been treated.

Subjects are to be treated sequentially, and past information can be used in deciding how to proceed. In particular, the arm selected for observation at the i^{th} selection depends on the previous selections, the previous results, the previous covariate values, and the value of the covariate for the subject about to be treated. A decision procedure or strategy specifies which arm to select based on this information. The worth of a strategy is defined in the usual way as the expectation of the sum of the first n observations for all possible histories resulting from that strategy. A strategy is optimal if it yields the maximal expected sum. An arm is said to be optimal if it is the first selection of some optimal strategy.

Many possibilities exist for describing the relationship between ρ, λ , and the covariate. We choose a linear-logistic model: $\rho(s) = \exp\{\alpha s\}/(1+\exp\{\alpha s\})$ and $\lambda(s) = \exp\{c s\}/(1+\exp\{c s\})$. When necessary, we shall write these as $\rho(\alpha, s)$ and $\lambda(c, s)$, making the dependence on α and c explicit. We shall assume that the characteristics of arm 2 are known in the sense that c is a known constant. We assume that α is unknown and following a Bayesian approach, we suppose that prior information regarding α can be given by a probability distribution R . Finally, we assume that, prior to their observation, the covariate values are unknown, and that they are i.i.d. with a known distribution function G . This implies that, while the covariate value for a given subject is unknown until the subject arrives for treatment, the distribution of possible covariate values is known. We shall use lower case letters $s, t, s_1, s_2 \dots$ etc. to denote

observed values of the covariate. In the situation where a subject's covariate value, s , is known, but the subject has not yet been treated, we shall refer to s as the "current" covariate value.

Although we have introduced our model in a clinical trial setting, one can describe industrial and other settings where it would be equally applicable. For convenience, we shall continue to use the clinic trial setting in describing our results.

A special case arises when G is degenerate at a point. In that case, the covariate values are the same for all subjects, and the situation presented here becomes equivalent to the bandit of Bradt, et al. (1956). When comparing bandit models, we shall refer to the model set out in this paper as the "covariate" bandit model, and to the model of Bradt et al. (1956) as a "standard" bandit model.

The information about arm 1 is described by the distribution measure on α . Initially, this is given by R . As we make observations on arm 1, we can describe the posterior distribution of α given observations as follows: If successes on arm 1 have been observed when the covariate values were s_1, \dots, s_j , and if failures occurred when the covariate values were t_1, \dots, t_k , then the posterior measure on α is given by

$$\sigma_{s_1} \sigma_{s_2} \dots \sigma_{s_j} \phi_{t_1} \phi_{t_2} \dots \phi_{t_k} R$$

where

$$\frac{d \sigma_{s_1} \sigma_{s_2} \dots \sigma_{s_j} \phi_{t_1} \phi_{t_2} \dots \phi_{t_k} R}{dR} = \frac{\rho(s_1)\rho(s_2)\dots\rho(s_j)(1-\rho(t_1))(1-\rho(t_2))\dots(1-\rho(t_k))}{E[\rho(s_1)\rho(s_2)\dots\rho(s_j)(1-\rho(t_1))(1-\rho(t_2))\dots(1-\rho(t_k)) | R]}$$

This is an extension of the notation of Berry and Fristedt (1979 p. 1087). Note that order is immaterial here: for example, $\sigma_s \phi_t R = \phi_t \sigma_s R$. For notational convenience we

shall sometimes refer to $\sigma_{s_1} \sigma_{s_2} \dots \sigma_{s_j} \phi_{t_1} \phi_{t_2} \dots \phi_{t_k} R$ as $\underline{h}R$, \underline{h} denoting the previous history of successes, failures, and their corresponding covariate values.

Throughout this paper we use the notation $E(\cdot | R)$ to denote expectation over α with respect to the distribution R . When there can be no confusion, we omit R from the notation.

By an " (s, n, R, c) bandit" we mean a bandit for which the current covariate value is s , the number of subjects to treat is n , the distribution on α is R , and the parameter on arm 2 is c . (We shall regard G as fixed throughout, and suppress it from the notation.) In notation similar to Berry (1972) and Clayton and Berry (1985), let $W_n^i(s, R, c)$ be the worth of selecting arm i initially in the (s, n, R, c) bandit. The worth of proceeding optimally in the (s, n, R, c) bandit is $\max\{W_n^1(s, R, c), W_n^2(s, R, c)\}$, which we denote by $W_n(s, R, c)$. The expected worth, before s is observed, is then

$$(1.1) \quad \int W_n(s, R, c) dG(s) = W_n(R, c).$$

Note that for $n > 1$ we have the usual dynamic programming equations:

$$(1.2) \quad W_n^1(s, R, c) = E(\rho(s) | R) + E(\rho(s) | R) W_{n-1}(\sigma_s R, c) \\ + [1 - E(\rho(s) | R)] W_{n-1}(\phi_s R, c)$$

and

$$(1.3) \quad W_n^2(s, R, c) = \lambda(s) + W_{n-1}(R, c).$$

Together with the evident condition that $W_0(R, c) = 0$, the above equations give a recursion for determining $W_n(s, R, c)$ and $W_n(R, c)$.

One further quantity that we shall use in describing optimal strategies is the difference:

$$(1.4) \quad \Delta_n(s, R, c) = W_n^1(s, R, c) - W_n^2(s, R, c).$$

The sign of Δ_n indicates the optimal arm at any stage: if $\Delta_n(s, R, c) > 0$ then arm 1 is optimal initially; if $\Delta_n(s, R, c) < 0$ then arm 2 is optimal initially; and if $\Delta_n(s, R, c) = 0$ then either arm is optimal. Suppose an optimal pull has been made. If arm 2 has been observed, and if the current covariate value is

t , then we are faced with a $(t, n-1, R, c)$ bandit, and therefore observe arm 1 or arm 2 according to the sign of $\Delta_{n-1}(t, R, c)$. If, instead, arm 1 was observed on the first pull then at the next stage we pull arm 1 or arm 2 according to the sign of $\Delta_{n-1}(t, \sigma_s R, c)$ or $\Delta_{n-1}(t, \phi_s R, c)$.

The problem described here is a form of two-armed bandit with one arm known. In the special case of a standard bandit, this problem has been described as a "one-armed" bandit since it is a stopping problem: for the standard bandit an optimal strategy can always be found for which any pull of arm 2 is optimally followed by another pull of arm 2. Examples of bandits satisfying such a condition are contained in Bradt, et al. (1956), Berry and Fristedt (1979), Clayton and Berry (1985), Clayton and Witmer (1986) and others. As we shall see below, such a characterization cannot be applied in general to the covariate bandit.

Although the bandit model has been discussed extensively (see Berry and Fristedt, 1985) relatively little has been written on the incorporation of covariates in bandit models. A notable exception is that of Woodroffe (1979), who studied a bandit model that incorporated covariates and yielded normally distributed observations. For that bandit model Woodroffe derived second order approximations to optimal strategies and investigated their behavior. This contribution should be regarded as important twice over: first, for introducing a covariate bandit model, and second, for a further discussion of bandits other than Bernoulli. (See Clayton and Berry (1985), for another example of a non-Bernoulli bandit.) However, as Simons (1986) has commented, results derived for normal models are difficult to apply to the Bernoulli setting.

As we shall see below, the Bernoulli covariate model behaves in a more complicated fashion than either the standard bandit or Woodroffe's Gaussian covariate model. This is due in part to the fact that, if we view the Bernoulli covariate bandit as a Markov decision problem, then an important component of the state

space is R , the distribution on α . Except when G has a finite support, this implies that the state space is effectively infinite dimensional. As a consequence, the explicit determination of optimal strategies is difficult, unless n is small or the support of G is on a small number of points. A similar issue arises when a Dirichlet process prior is used in a sequential decision problem (Clayton, 1985, Clayton and Berry, 1985).

In the remainder of this paper we shall focus on properties of optimal strategies in the covariate bandit and on the relationship between the standard bandit and the covariate bandit. In Section 2 we investigate some basic monotonicity properties of the covariate bandit and discuss stopping rules. In Section 3 we discuss further the properties of optimal strategies in terms of a break-even value. Section 4 contains some further comments.

2. Properties of optimal strategies:

In this section we begin to describe some of the properties of optimal strategies. As with most bandits, the current model can be seen as an attempt to reconcile two conflicting goals: (1) to obtain information about R ; and (2) to maximize the chance of success for each pull. Such a conflict arises, for example, when $E\rho(s) < \lambda(s)$ at a particular stage. A pull on arm 2 will have a greater chance of success for the current subject, but a pull of arm 1 may yield information about R that will have a benefit when making future pulls. On the other hand, if R is such that $E[\rho(s)|\underline{h}_R] > \lambda(s)$ for all histories \underline{h} , then arm 1 will always be optimal, and likewise, if $E[\rho(s)|\underline{h}_R] < \lambda(s)$, for all \underline{h} , then arm 2 will always be optimal. Such situations can arise, for example, if $P(\alpha > c) = 1$ or $P(\alpha < c) = 1$.

It is possible to prove, using equations (1.1), (1.2), (1.3), induction, and Jensen's inequality, that

$$\begin{aligned}
 (2.1) \quad & n \max \{ \int \lambda(s) dG(s), \int E\rho(s) dG(s) \} \\
 & \leq n \int \max \{ \lambda(s), E\rho(s) \} dG(s) \\
 & \leq W_n(R, c) \\
 & \leq n E \int \max \{ \lambda(s), \rho(s) \} dG(s).
 \end{aligned}$$

In words, $W_n(R, c)$ is bounded above by what would be the expected utility if α were known at the outset. $W_n(R, c)$ is bounded below by the worth of a strategy that, for each subject, takes the covariate into account but ignores any posterior information gained about α during the trial. Finally, this latter worth and $W_n(R, c)$ are both bounded below by the worth of a strategy that pulls the same arm throughout the trial.

As mentioned above, the standard bandit is a stopping problem, insofar as an optimal pull of arm 2 can be followed by another optimal pull of arm 2. This need not be the case for the covariate bandit, as the following example shows.

Example 2.1: Suppose $c = 0$, and $P(\alpha=-1) = P(\alpha=1) = 1/2 = P(S=-5) = P(S=5)$.

Then it is easy, but tedious, to calculate:

$$W_1^1(-5, R) = .0102, W_1^2(-5, R) = .0067, W_2^1(5, R) = 1.4916, W_2^2(5, R) = 1.4951.$$

Hence, arm 2 is optimal when $n = 2$ and $S = +5$. However, arm 1 is optimal when $n = 1$ and $S = -5$. This result reflects the intuition that when the current covariate value is sufficiently small, arm 1 is more attractive since there is some potential that α will be +1. However, when S is large, arm 1 becomes less attractive, since there is a risk that α will be -1.////

Although the covariate bandit is not a stopping problem in the usual sense, it does satisfy a weak stopping rule property, as follows:

Theorem 2.1: If, for the (s,n,R,c) bandit arm 2 is uniquely optimal for all s , then there exists an optimal strategy for which it is optimal to pull arm 2 for the (s,m,R,c) bandit, $m < n$, for any s .

Proof: The proof is a generalization of the proof of a similar result in Bradt, et al. (1956). Suppose, to the contrary, that there exists an $m' < n$ such that arm 1 is uniquely optimal under the strategy τ for the (s',m',R,c) bandit. Consider an (s',n,R,c) bandit that follows τ for the first m' pulls, and then pulls arm 2 for the remaining $n-m'$ pulls. This has a worth that is no less than $W_n(s',R,c)$. But this contradicts the unique optimality of arm 2 for the s',n,R,c bandit.////

Another property of the standard bandit is the "stay on a winner" property: if an optimal pull of arm 1 yields a success, then it is optimal to pull arm 1 again. This need not hold, as shown in the next example.

Example 2.2: Suppose $c = 0$, $P(\alpha=-1) = P(\alpha=5) = \frac{1}{2}$ and $P(S=1) = .99 = 1 - P(S=3)$.

Then $\Delta_2(3,R) = .00125$ but $\Delta_1(3,\sigma_3R) = -.00869$. That is, with $S = 3$ and two observations to take, arm 1 is optimal. However, if a pull of arm 1 under such circumstances yields a success, a subsequent pull of arm 1 when $S = 3$ will not be optimal.////

Although no simple stay-on-a-winner rule exists, a weak form of stay-on-a-winner does exist, as follows.

Theorem 2.2: Suppose in the (s,n,R,c) bandit that an initial pull of arm 1 is uniquely optimal and that a success obtains. Then there exists an s' in the support of G such that a pull of arm 1 is optimal for the $(s',n-1,\sigma_s R)$ bandit.

Proof: Suppose to the contrary, that arm 1 is not optimal for any s' in the support of G for the $(s', n-1, \sigma_s R, c)$ bandit. Then by Theorem 2.1 arm 2 is optimal for the remaining $n-1$ pulls, and thus

$$W_{n-1}(s', \sigma_s R, c) = \lambda(s') + (n-2) \int \lambda(s) dG(s).$$

Hence $W_{n-1}(\sigma_s R, c) = (n-1) \int \lambda(s) dG(s)$. It follows from Theorem 2.3 below and equation (2.1) that

$$W_{n-1}(\sigma_s R, c) = W_{n-1}(R, c) = W_{n-1}(\phi_s R, c) = (n-1) \int \lambda(s) dG(s).$$

Moreover, since a pull of arm 2 is optimal for all pulls after the first, it follows that $E\rho(s_1) < \lambda(s_1)$ for any s_1 in the support of G . Finally, by equation (1.2) this implies that

$$\begin{aligned} W_n^1(s, R, c) &= E\rho(s) + E\rho(s)W_{n-1}(\sigma_s R, c) + [1-E\rho(s)]W_{n-1}(\phi_s R, c) \\ &= E\rho(s) + W_{n-1}(R, c) \\ &\leq \lambda(s) + W_{n-1}(R, c) \\ &= W_n^2(s, R, c), \end{aligned}$$

contradicting the fact that arm 1 is uniquely optimal for the first pull.////

While the covariate bandit and the standard bandit share the stopping rule and stay-on-a-winner properties in only a weakened sense, both bandits have several monotonicity properties in common. For example, it is easy to prove by induction, using (1.2) and (1.3) that $W_n(s, R, c)$ is nondecreasing in c .

We can also develop a monotonicity result for arm 1. This contains the finite horizon version of Theorem 3.1 of Berry and Fristedt (1979) as a special case.

Definition 2.1: For any two random variables X and X' with distribution functions F and F' respectively, we say that the distribution of X' is "to the right of" the distribution of X if $F(b) \geq F'(b)$ for all b . As noted in Marshall and Olkin, (1979), this condition is equivalent to the condition that $Eg(X') \geq Eg(X)$ for any nondecreasing g such that the expectations exist.////

Note that if the distribution of α' is to the right of the distribution of α , then the distribution of $\rho(\alpha',s)$ is to the right of the distribution of $\rho(\alpha,s)$, for all s . In addition, if the distribution of $\rho(\alpha',s)$ is to the right of the distribution of $\rho(\alpha,s)$ for some s , then it is easy to show that this must hold for all s , and that the distribution of α' is to the right of the distribution of α .

Definition 2.2: In an extension of a notion of Berry and Fristedt (1979), if R' and R are measures for α , we define R' to be "strongly to the right" of R if $\underline{h}R'$ is to the right of $\underline{h}R$ for all histories \underline{h} .////

Given these definitions, we have the following:

Theorem 2.3: If R' is strongly to the right of R then

$W_n(R',c) \geq W_n(R,c)$ and for all s , $W_n^1(s,R',c) \geq W_n^1(s,R,c)$.

Proof: This is immediately true by induction for W_n^2 . Consider W_n^1 .

$$\begin{aligned} W_n^1(s,R',c) - W_n^1(s,R,c) &= E(\rho(s) | R') + E(\rho(s) | R') W_{n-1}(\sigma_s R', c) \\ &\quad + (1 - E\rho(s) | R') W_{n-1}(\phi_s R', c) \\ &\quad - E(\rho(s) | R) - E(\rho(s) | R) W_{n-1}(\sigma_s R, c) \\ &\quad - (1 - E\rho(s) | R) W_{n-1}(\phi_s R, c) \\ &= E(\rho(s) | R') - E(\rho(s) | R) \end{aligned}$$

$$\begin{aligned}
& + E(\rho(s) | R) [W_{n-1}(\sigma_{s'} R', c) - W_{n-1}(\sigma_s R, c)] \\
& + (1 - E\rho(s) | R') [W_{n-1}(\phi_{s'} R', c) - W_{n-1}(\phi_s R, c)] \\
& + (E(\rho(s) | R') - E(\rho(s) | R)) [W_{n-1}(\sigma_{s'} R', c) - W_{n-1}(\phi_s R, c)].
\end{aligned}$$

Since R' is to the right of R , $E(\rho(s) | R') > E(\rho(s) | R)$. Note that $\sigma_{s'} R'$ is strongly to the right of $\sigma_s R$, and both of these are strongly to the right of $\phi_s R$. Also, $\phi_{s'} R'$ is strongly to the right of $\phi_s R$. By induction, each of the quantities in square brackets is nonnegative. The rest of the proof follows by definition of $W_n(s, R, c)$ and equation (1.1).////

Lemma 2.1: If $s_1 < s_2$ then

- (i) $\sigma_{s_1} R$ is to the right of $\sigma_{s_2} R$ and
- (ii) $\phi_{s_2} R$ is to the right of $\phi_{s_1} R$.

Proof: We prove part (i); the proof of part (ii) is similar. It will suffice to show that, for all b , $P(\alpha \leq b | \sigma_{s_1} R) \leq P(\alpha \leq b | \sigma_{s_2} R)$, or equivalently, that

$$(2.2) \quad \int_{(-\infty, b]} \rho(s_1) dR / E\rho(s_1) \leq \int_{(-\infty, b]} \rho(s_2) / E\rho(s_2).$$

Writing the dependence of ρ on α explicitly, (2.2) is equivalent to

$$(2.3) \quad \begin{aligned}
0 & \leq \int I_{(-\infty, a]}(\alpha) \rho(\alpha, s_2) R(d\alpha) / \int I_{(a, \infty)}(\alpha') \rho(\alpha', s_1) R(d\alpha') \\
& - \int I_{(-\infty, a]}(\alpha) \rho(\alpha, s_1) R(d\alpha) / \int I_{(a, \infty)}(\alpha') \rho(\alpha', s_2) R(d\alpha'),
\end{aligned}$$

where I_A is the indicator function of the set A . By Tonelli's theorem, the right side of (2.3) is

$$(2.4) \quad \int \int I_{(-\infty, a]}(\alpha) I_{(a, \infty)}(\alpha') [\rho(\alpha', s_1) \rho(\alpha, s_2) - \rho(\alpha', s_2) \rho(\alpha, s_1)] R(d\alpha') R(d\alpha).$$

However, if $\alpha \leq a < \alpha'$, the integrand in (2.4) is nonnegative, and if $\alpha \leq a < \alpha'$ fails, the integrand is zero.////

An immediate consequence of Lemma 2.1 and Theorem 2.3 is the following.

Theorem 2.4: For all n, R, c, t , and i , each of $W_n(\sigma_s R, c)$, $W_n^i(t, \sigma_s R, c)$, and $W_n(t, \sigma_s R, c)$ are nonincreasing in s and each of $W_n(\phi_s R, c)$, $W_n^i(t, \phi_s R, c)$, and $W_n(t, \phi_s R, c)$ are nondecreasing in s .

Theorem 2.4 and parts (i) - (iv) of Proposition 2.1 below are related to the "information" obtained by a particular pull. A success observed on arm 1 when s is large is relatively uninformative, since $\lim_{s \rightarrow \infty} \rho(s) = 1$ for any α . However, a success observed when s is large in magnitude and negative is potentially quite informative: it suggests that α itself is large. The reverse situation arises when we observe a failure on arm 1. Bounds on the "information" available in a pull are given by parts (ii) and (iii) of the Proposition. Part (v) of Proposition 2.1 suggests that as the current covariate value grows large in magnitude, we become indifferent to the choice of arms for the next pull.

Proposition 2.1: Define $\sigma_{-\infty} R$ by $\frac{d\sigma_{-\infty} R}{dR} = \frac{e^{-\alpha}}{\epsilon e^{-\alpha}}$ and $\phi_{+\infty} R$ by $\frac{d\phi_{+\infty} R}{dR} = \frac{e^{-\alpha}}{\epsilon e^{-\alpha}}$.

For all n , for $i = 1$ and 2 , for all R and for all c ,

$$(i) \quad \lim_{s \rightarrow \infty} W_n(\sigma_s R, c) = \lim_{s \rightarrow -\infty} W_n(\phi_s R, c) = W_n(R, c)$$

$$(ii) \quad \lim_{s \rightarrow -\infty} W_n(\sigma_s R, c) = W_n(\sigma_{-\infty} R, c)$$

$$(iii) \quad \lim_{s \rightarrow \infty} W_n(\phi_s R, c) = W_n(\phi_\infty R, c)$$

$$(iv) \quad \lim_{s \rightarrow \infty} W_n(t, \sigma_s R, c) = W_n(t, R, c) = \lim_{s \rightarrow -\infty} W_n(t, \phi_s R, c)$$

$$(v) \quad \lim_{t \rightarrow \pm\infty} W_n^1(t, R, c) = \lim_{t \rightarrow \pm\infty} W_n^2(t, R, c)$$

Proof: We prove some parts of the proposition. The remainder follow similarly. First we prove the first half of part (iv), using induction. The result is easy when $n = 1$. Note that

$$\begin{aligned} W_n^1(t, \sigma_s R, c) &= E(\rho(t) | \sigma_s R) \\ &+ E[\rho(t) | \sigma_s R] W_{n-1}(\sigma_t \sigma_s R, c) \\ &+ [1 - E\rho(t) | \sigma_s R] W_{n-1}(\phi_t \sigma_s R, c). \end{aligned}$$

By the induction hypothesis, $\lim_{s \rightarrow \infty} W_{n-1}(\sigma_s \sigma_t R, c) = W_{n-1}(\sigma_t R, c)$. Also it is easy to show that $\lim_{s \rightarrow \infty} E[\rho(t) | \sigma_s R] = E\rho(t)$. So

$$\lim_{s \rightarrow \infty} W_n^1(t, \sigma_s R, c) = E\rho(t) + E\rho(t) W_{n-1}(\sigma_t R, c) + [1 - E\rho(t)] W_{n-1}(\phi_t R, c) = W_n^1(t, R, c).$$

To prove part (i), use part (iv), noting that

$$\begin{aligned} \lim_{s \rightarrow \infty} W_n(\sigma_s R, c) &= \lim_{s \rightarrow \infty} \int W_n(t, \sigma_s R, c) dt \\ &= \int \lim_{s \rightarrow \infty} W_n(t, \sigma_s R, c) dt \\ &= \int W_n(t, R, c) dt = W_n(R, c). \end{aligned}$$

The second equality above follows from the dominated convergence theorem and equation (2.1). To prove part (v), note that

$$\begin{aligned}
 \lim_{t \rightarrow \infty} W_n^1(t, R, c) &= \lim_{t \rightarrow \infty} \{E\rho(t) + E\rho(t)W_{n-1}(\sigma_t R, c) + [1-E\rho(t)]W_{n-1}(\phi_t R, c)\} \\
 &= 1 + \lim_{t \rightarrow \infty} W_{n-1}(\sigma_t R, c) \\
 &= 1 + W_{n-1}(R, c) \\
 &= \lim_{t \rightarrow \infty} W_n^2(t, R, c) .////
 \end{aligned}$$

3. The function Δ_n .

As mentioned above, the function Δ_n can be used to determine optimal strategies. As must also be evident, the determination of Δ_n is nontrivial. In this section we explore certain properties of Δ_n and discuss their implications for determining properties of the optimal strategy.

Note that it is useful and proper to consider Δ_n as a function of s for all real s , even though the support of G might be on some proper subset of the reals. Of course, in using Δ_n to determine optimal strategies, attention will be restricted to those s in the support of G .

It is easy to show, by induction, and using (1.2), (1.3), and the definition of Δ_n (1.4) that Δ_n is a continuous function of s and c . From Proposition 2.1(v) it is evident that $\lim_{s \rightarrow \pm\infty} \Delta_n(s, R, c) = 0$. This fact is illustrated in Figure 1, where Δ_n is plotted as a function of s for $n = 1, \dots, 6$ and for R and G such that $P(\alpha=-1) = P(\alpha=1) = \frac{1}{2} = P(S=-1) = P(S=1)$. We note from Figure 1 that Δ_n has at most one root in s . We now set about a proof of that fact for the case $n = 1$, and derive a weaker result when $n > 1$.

Theorem 3.1. If R is not degenerate at a point, then $\Delta_1(s, R, c)$ has at most one root in s .

Remark: If R is degenerate at c , then $\Delta_1(s, R, c) = 0$ for all s . If R is degenerate at a point other than c , then $\Delta_1(s, R, c)$ has no roots in s .

Proof: Without loss of generality, we can assume $c = 0$. Let $s_1 < s_2$. We show that: (a) if $\Delta_1(s_2, R, 0) > 0$ then $\Delta_1(s_1, R, 0) > 0$. A similar approach shows that: (b) if $\Delta_1(s_1, R, 0) < 0$ then $\Delta_1(s_2, R, 0) < 0$. Parts (a) and (b) complete the proof.

To proceed with (a), note first that $\Delta_1(s, R, 0) = \lambda(s)d(s)$, where $d(s) = E[(e^\alpha - 1)/(1 + e^{\alpha s}) | R]$. Since $\lambda(s) > 0$ for all s , it will suffice to show that $\Delta_1(s_2, R, 0) > 0$ implies $d(s_1) - d(s_2) > 0$. Next, note that $\Delta_1(s_2, \sigma_{s_1} R, 0) - \Delta_1(s_2, R, 0) = E[\rho(s_1)\rho(s_2) | R]/E[\rho(s_1) | R] - E[\rho(s_2) | R]$, and if R is not degenerate at a point, this difference is strictly positive. Finally, some algebra shows that

$$d(s_1) - d(s_2) = \frac{(e^{s_2 - s_1} - 1)E\rho(s_1)}{\lambda(s_2)} \Delta_1(s_2, \sigma_{s_1} R, 0).$$

Consequently, $\Delta_1(s_2, R, 0) > 0$ implies $\Delta_1(s_2, \sigma_{s_1} R, 0) > 0$ which in turn implies $d(s_1) - d(s_2) > 0$, as required.////

The result of Theorem 3.1 may be restated in an alternative form, which we note as a Corollary.

Corollary 3.1: In the $(s_1, 1, R, c)$ bandit, there exists a quantity

$\Sigma_1 = \Sigma_1(R, c) \in [-\infty, \infty]$ such that a pull of arm 1 is optimal if $s_1 < \Sigma_1$; a pull of arm 2 is optimal if $s_1 > \Sigma_1$; and either is optimal if $s_1 = \Sigma_1$.

Remark: If $\Sigma_1 = +\infty$, then a pull of arm 1 is optimal for all s , and if $\Sigma_1 = -\infty$ then a pull of arm 2 is optimal for all s . So, for example, if $P(\alpha > c) = 1$ then $\Sigma_1 = +\infty$.

Theorem 3.2: If R' is strongly to the right of R , then $\Delta_1(s, R', c) > \Delta_1(s, R, c)$ and $\Sigma_1(R', c) > \Sigma_1(R, c)$. Also, $\Delta_1(s, R, c)$ is decreasing in c ; $\Sigma_1(R, c)$ is nonincreasing in c .

Proof: This is an immediate consequence of Theorem 2.3 and the definitions of Δ_1 and Σ_1 .////

We conjecture that, for all n , there exists a quantity Σ_n with properties similar to Σ_1 ; namely, in the (s_n, n, R, c) bandit, it is optimal to pull arm 1 if $s_n < \Sigma_n(R, c)$ and it is optimal to pull arm 2 if $s > \Sigma_n(R, c)$. In this sense $\Sigma_n(R, c)$ would be a "break-even value" for the covariate. If Σ_n were to exist for $n > 2$, then a complete determination of an optimal strategy could be given in terms of Σ_n . Our next result and its corollary are partial results in that direction.

Theorem 3.3: For all n , $\Delta_n(s, R, c) > \Delta_1(s, R, c)$.

Proof: From (1.2), (1.3), and (1.4),

$$\begin{aligned} \Delta_{n+1}(s, R, c) - \Delta_1(s, R, c) &= E\rho(s)W_n(\sigma_S R, c) + [1-E\rho(s)]W_n(\phi_S R, c) \\ (3.1) \qquad \qquad \qquad &- W_n(R, c). \end{aligned}$$

The right side of (3.1) may be written as

$$\begin{aligned} &\int [E\rho(s)W_n(t, \sigma_S R, c) + [1-E\rho(s)]W_n(t, \phi_S R, c) \\ (3.2) \qquad \qquad \qquad &- W_n(t, R, c)] dG(t) \end{aligned}$$

We show that the integrand in (3.2) is always nonnegative by induction. For $n = 1$, there are two cases. If $W_1(t, R, c) = \lambda(t)$, then the result is clear, since $W_1(t, \sigma_S R, c) > \lambda(t)$ and $W_1(t, \phi_S R, c) > \lambda(t)$.

If $W_1(t, R, c) = E\rho(t)$, then we note that $W_1(t, \sigma_S R, c) \geq E[\rho(t) | \sigma_S R]$ and $W_1(t, \phi_S R, c) \geq E[\rho(t) | \phi_S R]$, whence the integrand in (3.2) is bounded below by $E[\rho(s)]E[\rho(t) | \sigma_S R] + [1 - E\rho(s)]E[\rho(t) | \phi_S R] - E\rho(t) = 0$.

For the induction step we assume that the integrand in (3.2) is nonnegative for $n = m$. Again, we distinguish two cases:

First, when $W_{m+1}(t, R, c) = W_{m+1}^2(t, R, c)$ then we have

$$\begin{aligned}
 & E\rho(s)W_{m+1}(t, \sigma_S R, c) + [1 - E\rho(s)]W_{m+1}(t, \phi_S R, c) \\
 & > E\rho(s)W_{m+1}^2(t, \sigma_S R, c) + [1 - E\rho(s)]W_{m+1}^2(t, \phi_S R, c) \\
 & = E\rho(s)[\lambda(t) + W_m(\sigma_S R, c)] + [1 - E\rho(s)][\lambda(t) + W_m(\phi_S R, c)] \\
 & = \lambda(t) + E\rho(s)W_m(\sigma_S R, c) + [1 - E\rho(s)]W_m(\phi_S R, c) \\
 & > \lambda(t) + W_m(R, c) \\
 & = W_{m+1}(t, R, c).
 \end{aligned}$$

The first inequality above follows by definition of W_{m+1} . The second inequality follows by the induction hypothesis. In the second case, if

$W_{m+1}(t, R, c) = W_{m+1}^1(t, R, c)$ then following (1.2), the integrand in (3.2) can be written as the sum of three components, A, B, and C, say, where

$$A = E\rho(s)E[\rho(t) | \sigma_S R] + E[1 - \rho(s)]E[\rho(t) | \phi_S R] - E\rho(t)$$

$$B = E\rho(s)E[\rho(t) | \sigma_S R]W_m(\sigma_t \sigma_S R, c) + E[1 - \rho(s)]E[\rho(t) | \phi_S R]W_m(\sigma_t \phi_S R, c)$$

$$- E\rho(t)W_m(\sigma_t R, c)$$

and

$$C = E\rho(s)E[1 - \rho(t) | \sigma_S R]W_m(\phi_t \sigma_S R, c) + E[1 - \rho(s)]E[1 - \rho(t) | \phi_S R]W_m(\phi_t R, c)$$

$$- E[1 - \rho(t)]W_m(\phi_t R, c).$$

Some algebra shows that $A = 0$. Consider the expression B . We may write

$$\begin{aligned} B &= E\rho(t)\rho(s)W_m(\sigma_s\sigma_tR,c) + E\rho(t)(1-\rho(s))W_m(\phi_s\sigma_tR,c) - E\rho(t)W_m(\sigma_tR,c) \\ &= E\rho(t)\{E(\rho(s)|\sigma_tR)W_m(\sigma_s\sigma_tR,c) + E(1-\rho(s)|\sigma_tR)W_m(\phi_s\sigma_tR,c) - W_m(\sigma_tR,c)\} \end{aligned}$$

which is nonnegative by the induction hypothesis. Likewise, the induction hypothesis may be used to prove that $C \geq 0$, and the theorem now follows.////

Corollary 3.2: For any n,R,c , there exists a $\Sigma'_n(R,c)$ such that if $s_n < \Sigma'_n(R,c)$ then arm 1 is optimal in the (s_n,n,R,c) bandit. Moreover, if we take $\Sigma'_1(R,c) = \Sigma_1(R,c)$, then $\Sigma'_n(R,c) \geq \Sigma'_1(R,c)$.

Let $\hat{\Sigma}_n(R,c)$ be the largest $\Sigma'_n(R,c)$ such that the property given in Corollary 3.2 holds. (In particular $\hat{\Sigma}_1(R,c) = \Sigma'_1(R,c) = \Sigma_1(R,c)$.) Then $\hat{\Sigma}_n(R,c)$ is a weak form of "break-even value" for the covariate s_n in the following sense: if $s_n < \hat{\Sigma}_n(R,c)$ then arm 1 is optimal in the (s_n,n,R,c) bandit. Noting that $\rho(s)$ and $\lambda(s)$ are increasing in s , this says that, in terms of a clinical trial, when subjects have covariates such that the probability of success is small, then it is better to use the experimental, unknown treatment (arm 1). Our conjecture that $\Sigma_n(R,c)$ exists for all n is equivalent to the conjecture that $s_n > \hat{\Sigma}_n(R,c)$ implies that arm 2 is optimal in the (s_n,n,R,c) bandit. This certainly holds for the example in Figure 1, and we have shown it to hold in other examples not provided here. The proof of such a result in general seems particularly elusive, however. The following example represents a partial result regarding the existence of $\Sigma_n(R,c)$ for $n = 1$.

Example 3.2: Suppose $c = 0$, $P(\alpha=1) = p = 1 - P(\alpha=-1)$, and $P(S=-1) = P(S=1) = 1/2$. We demonstrate the existence of Σ_2 in this case. For brevity, we suppress c from the notation. It will suffice to show that

if $W_2(-1,R) = W_2^2(-1,R)$, then $W_2(1,R) = W_2^2(1,R)$, and if $W_2(1,R) = W_2^1(1,R)$ then $W_2(-1,R) = W_2^1(-1,R)$. We prove these simultaneously by contradiction.

Specifically, suppose that $W_2(-1,R) = W_2^2(-1,R)$ and that $W_2(1,R) = W_2^1(1,R)$.

If $W_2(-1,R) = W_2^2(-1,R)$ then it follows that $\Delta_2(-1,R) < 0$, where $\Delta_1(-1,R) < 0$ by Theorem 3.1. But, but Theorem 3.2, $\Delta_1(-1,R) < 0$ implies that $\Delta_1(-1, \phi_{-1}R) < 0$

and $\Delta_1(-1, \phi_1R) < 0$. From Theorem 3.1, we thus have $\Delta_1(1,R) < 0$, $\Delta_1(1, \phi_{-1}R) < 0$,

and $\Delta_1(1, \phi_1R) < 0$. Since $W_2(1,R) = W_2^1(1,R)$ by assumption, it must be that

$\Delta_1(-1, \sigma_1R) \geq 0$. There are now two cases: $\Delta_1(1, \sigma_1R) \geq 0$ and $\Delta_1(1, \sigma_1R) < 0$.

The second case can be shown to be impossible. Therefore,

$$W_2^1(1,R) = E\rho(1) + 1/2E\rho(1)\rho(1) + 1/2E\rho(1)\rho(-1) \\ + [1-E\rho(1)][1/2\lambda(1) + 1/2\lambda(-1)].$$

On the other hand, $W_2^2(1,R) = \lambda(1) + 1/2\lambda(1) + 1/2\lambda(-1)$, whence

$$\Delta_2(1,R) = E\rho(1) - \lambda(1) \\ + 1/2E[\rho(1)(\rho(1)-\lambda(1) + \rho(-1)-\lambda(-1))].$$

Similarly, $W_2^2(-1,R) = \lambda(-1) + 1/2\lambda(1) + 1/2\lambda(-1)$ and

$W_2^1(-1,R) = E\rho(-1) + 1/2E[\rho(-1)(\rho(1) - \lambda(1) + \rho(-1) - \lambda(-1))]$, since, by

Lemma 2.1 and Theorem 3.2 $\Delta_1(s, \sigma_{-1}R) > \Delta_1(s, \sigma_1R) > 0$. It follows that

$$\Delta_2(-1,R) = E\rho(-1) - \lambda(-1) \\ + 1/2E[\rho(-1)(\rho(1)-\lambda(1) + \rho(-1)-\lambda(-1))].$$

Some tedious algebra shows that

$$\Delta_2(1,R) > 0 \text{ iff } p > .5883$$

while

$$\Delta_2(-1,R) > 0 \text{ iff } p > .5694.$$

This leads to the sought after contradiction.////

4. Additional Comments

As mentioned, Corollaries 3.1 and 3.2 are directed toward the description of optimal strategies in terms of a break-even value for the current covariate value.

In noncovariate bandit treatments, optimal policies have been described in terms of a break-even value for arm 2. (See, for example, Bradt, et al. 1956, Berry and Fristedt, 1979, Clayton and Berry, 1985, and Berry and Fristedt, 1985. Gittens and Jones, 1974, have used such indexes in describing multi-armed bandits.) A comparable break-even value for arm 2 of the covariate bandit is a quantity $C_n(s,R)$ which would characterize optimal strategies as follows: if $c < C_n(s,R)$ then a pull of arm 1 is optimal, if $c > C_n(s,R)$ then a pull of arm 2 is optimal, and if $c = C_n(s,R)$ then either arm is optimal. We conjecture that such a quantity $C_n(s,R)$ exists for all n,s , and R . Indeed, it is easy to see that $C_1(s,R)$ exists for all s and R ; it is the root in c of the equation $\Delta_1(s,R,c) = 0$. The next result describes a situation in which $C_2(s,R)$ exists:

Proposition 4.1: If $\Delta_1(s,R,c) - \int \Delta_1(t,R,c)dG(t)$ is decreasing in c , then $C_2(s,R)$ exists.

Remark: The hypothesis of the proposition is equivalent to the requirement that $\int \lambda(t,c)(1-\lambda(t,c))dG(t) < \lambda(s,c)(1-\lambda(s,c))$.

Proof: We prove that, under the hypothesis of the proposition, $\Delta_2(s,R,c)$ is decreasing in c . The existence of $C_2(s,R)$ as a root in c of $\Delta_2(s,R,c) = 0$ then follows from the fact that $\lim_{c \rightarrow -\infty} \Delta_2(s,R,c) = E\rho(s)$ and $\lim_{c \rightarrow \infty} \Delta_2(s,R,c) = E\rho(s) - 1$.

If we let $\Delta_1^+ = \max\{\Delta_1, 0\}$ and $\Delta_1^- = -\min\{\Delta_1, 0\}$ then (1.1), (1.2), (1.3) and (1.4) can be used to show that

$$\begin{aligned} \Delta_2(s,R,c) &= \Delta_1(s,R,c) - \int \Delta_1(t,R,c)dG(t) \\ &\quad + \int [E\rho(s)\Delta_1^+(t,\sigma_s R,c) + [1-E\rho(s)]\Delta_1^+(t,\phi_s R,c) \\ &\quad - \Delta_1^-(t,R,c)]dG(t). \end{aligned}$$

The desired result now follows from the fact that $\Delta_1(t,R,c)$ is decreasing in c for all R .////

The method of proof used for Proposition 4.1 can be generalized to show that if $\Delta_1(s,R,c) - n\int \Delta_1(t,R,c)dG(t)$ is decreasing in c and if $\Delta_m(s,R,c)$ is decreasing in c for all s,R , and $m < n$, then $\Delta_n(s,R,c)$ is decreasing in c and $C_n(s,R)$ exists. Such hypotheses are undesirably strong, however.

We conclude with some comments regarding other covariate models. A natural extension of the model presented in this paper would allow the slope of the logistic transformation to vary. That is, we might have

$\rho(s) = \exp\{\alpha + \beta s\} / (1 + \exp\{\alpha + \beta s\})$ with α and β both unknown, and with a similar definition for $\lambda(s)$.

The following example shows that an index for s may not exist for such a covariate model.

Example 4.1. Suppose $\lambda(s) = e^s/(1+e^s)$ and $\rho(s) = e^{\beta s}/(1+e^{\beta s})$, with $P(\beta=0) = \frac{1}{2} = P(\beta=10)$. Then $E\rho(s) - \lambda(s)$ is positive, and so arm 1 is favored, if $s < -1.0986$ or if $0 < s < 1.0986$, and $E\rho(s) - \lambda(s)$ is negative, and arm 2 is favored, elsewhere.////

Although an index may not exist for the covariate models described in the paragraph preceding Example 4.1, we conjecture that "index sets" exist, as follows: for any distribution R on (α, β) there exists a set $\Omega_n(R)$ such that arm 1 is optimal if $s \in \Omega_n(R)$. We conjecture that a number of results similar to those developed in the previous sections will hold. For example, if R and R' are two distributions on (α, β) such that $E[g(\rho(s)) | \underline{h}R] < E[g(\rho(s)) | \underline{h}R']$ for all nondecreasing g , for all s , and all histories \underline{h} , then we expect that $\Omega_n(R)$ is a subset of $\Omega_n(R')$.

Acknowledgements: The help of Jooho Lee with computing, and the use of the U.W. Madison Department of Statistics Research Computer, are both greatly appreciated. Don Berry and I had several conversations about this problem some time ago. Those too were very helpful.

REFERENCES

- Berry, D.A. (1972). A bernoulli two-armed bandit. Ann. Math. Statist. 43, 871-897.
- 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, New York.
- 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.
- Clayton, M.K. (1985). A Bayesian nonparametric sequential test for the mean of a population. Ann. Statist. 13, 1129-1139.
- Clayton, M.K. and Berry, D.A. (1985). Bayesian nonparametric bandits. Ann. Statist. 13, 1523-1534.
- Clayton, M.K. and Witmer, J.A. (1987). Two-stage bandits. University of Wisconsin-Madison, Department of Statistics Technical Report.
- Gittens, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. In Progress in Statistics (eds. J. Gani et al.) pp. 241-266, North-Holland, Amsterdam.
- Marshall, A.W. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Applications. Academic Press, New York.
- Simons, G. (1986). Bayes rules for a clinical-trials model with dichotomous responses. Ann. Statist. 14, 954-970.
- Woodroffe, M. (1979). A one-armed bandit problem with a concomitant variable. J. Amer. Statist. Assoc. 74, 799-806.

Department of Statistics
University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

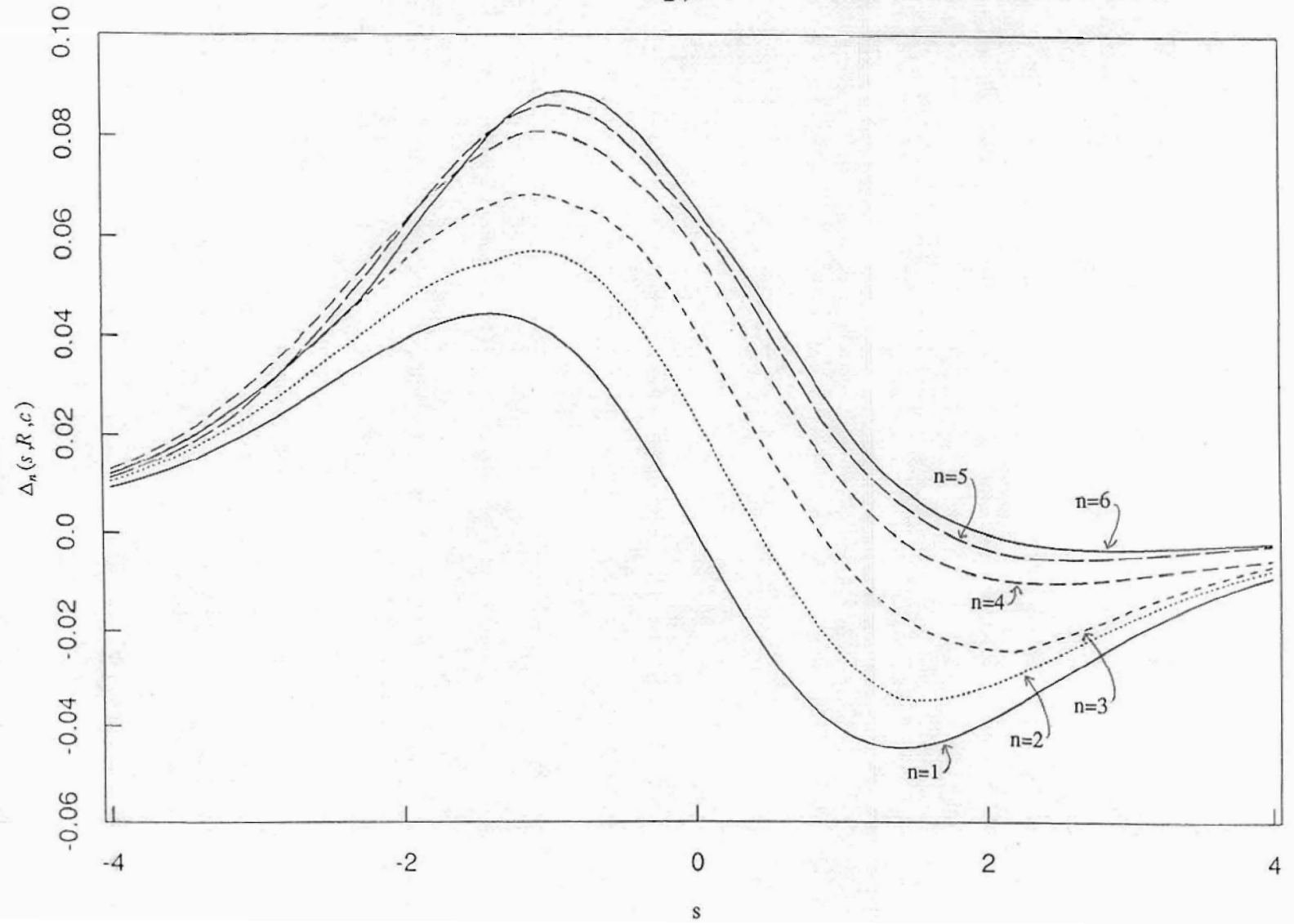


Figure 1. $\Delta_n(s, R, c)$ for various n and s . Here $P(\alpha=-1) = P(\alpha=1) = 0.5 = P(S=-1) = P(S=1)$.