The Annals of Statistics
1985, Vol. 13, No. 4, 1523–1534

BAYESIAN NONPARAMETRIC BANDITS

BY MURRAY K. CLAYTON[1] AND DONALD A. BERRY[2]

University of Wisconsin and University of Minnesota

Sequential selections are to be made from two stochastic processes, or "arms." At each stage the arm selected for observation depends on past observations. The objective is to maximize the expected sum of the first n observations. For arm 1 the observations are identically distributed with probability measure P, and for arm 2 the observations have probability measure Q; P is a Dirichlet process and Q is known. An equivalent problem is deciding sequentially when to stop sampling from an unknown population. Optimal strategies are shown to continue sampling if the current observation is sufficiently large. A simple form of such a rule is expressed in terms of a degenerate Dirichlet process which is related to P.

1. Introduction. A bandit problem involves sequential selections 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 restrict consideration to the discrete-time setting in which the objective is to maximize the expected sum of the first n observations, where n is known. This is a special case of the more general setting—not considered here—of Berry and Fristedt (1979) in which infinitely many observations may be taken and future observations are discounted.

The arm selected for observation at any time depends on the previous selections and results. A decision procedure or *strategy* specifies which arm to select at any stage for every history of previous selections and observations. The worth of a strategy is defined in the usual way as the average of the sums 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.

We assume 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 assume that the vector $(X_1, \ldots, X_n)$ is independent of $(Y_1, \ldots, Y_n)$. In addition, we assume that given the unknown probability measure (p.m.) P, the random variables $X_1, \ldots, X_n$ are independent and identically distributed with p.m. P, while the random variables $Y_1, \ldots, Y_n$ are independent and identically distributed with *known* p.m. Q. Since the objective is to maximize the expected sum of the observations and since no data can change the

Received February 1984; revised May 1985.
[1]Research partially supported by a University of Guelph Research Advisory Board Grant.
[2]Research partially supported by National Science Foundation Grant MCS8301450.
AMS 1980 *subject classifications*. Primary 62L05; Secondary 62L15
Key words and phrases. Sequential decisions, nonparametric decisions, optimal stopping, one-armed bandits, two-armed bandits, Dirichlet bandits.

information concerning Q, it is sufficient to assume that all the Y_is are equal to the mean of Q, call it λ.

The p.m. P is unknown and so, following a Bayesian approach, we take P to be random and assume that prior information regarding P is given by its probability distribution. Much of the bandit literature assumes the arms to be Bernoulli (Bradt et al., 1956 and Berry, 1972 are examples) in which case the support of P is contained in the set of those p.m.s which concentrate their mass on $\{1, 0\}$. We want a distribution for P which has large support and which yields analytically manageable posterior distributions for P conditional on observations from P. Following Ferguson (1973), we assume that P is a Dirichlet process with parameter α, where α is a bounded nonnull measure on the reals $\mathbb{R}$ with finite first moment. Let $M = \alpha(\mathbb{R})$ and $F(x) = \alpha(-\infty, x]/M$. With these definitions, F is the prior mean (in distribution function form) for P in the sense that it is the expectation of $P(X \leq x)$, and the total measure M may be interpreted as the weight of the prior in terms of sample number (Ferguson 1973, page 223). We shall frequently use MF in place of α. The prior mean μ for an observation from arm 1 is the mean of F.

The important special case $M = 0$ gives rise to an improper Dirichlet process. By a Dirichlet process with parameter $0 \cdot F$, we mean a process which generates observations $X_1, X_2, \ldots, X_k$, such that $X_1 = X_2 = \cdots = X_k$ almost surely and X_1 has distribution function F. In such a situation, one selection of arm 1 yields complete information about P. In a sense described in detail by Sethuraman and Tiwari (1982), as M tends to zero the Dirichlet process with parameter MF tends to the process with parameter $0 \cdot F$ defined here. For another application of improper Dirichlet processes see Clayton (1983).

The parameter α summarizes the prior information about P. Conditional on observations $X_1, \ldots, X_k$, the measure P is a Dirichlet process with parameter $\alpha + \sum_1^k \delta_{X_i}$, where δ_x gives mass 1 to x (Ferguson, 1973, Theorem 1). Let X be a generic observation from arm 1. The conditional expectation of a function $g(X)$ given $X_1, \ldots, X_k$ can be computed for each $\alpha + \sum_1^k \delta_{X_i}$ using Theorem 3 of Ferguson (1973). We shall denote this expectation by $E[g(X)|\alpha + \sum_1^k \delta_{X_i}]$, and shall delete the measure from the notation when there can be no confusion. Note in particular that $E(X|\alpha) = E(X|F) = \mu$.

Besides the Bernoulli, the only other bandit that is discussed in detail in the literature is the normal. Notable examples are Fahrenholtz (1982) and Chernoff (1968); the latter considers a continuous-time version using Wiener processes.

The Dirichlet process is more flexible in many ways than either the Bernoulli or the normal. Actually, the Dirichlet model encompasses the Bernoulli with a beta prior: If $\alpha = a\delta_1 + b\delta_0$ then $X_1, \ldots, X_n$ are distributed as if they were, given ρ, independent Bernoulli observations with parameter ρ, where ρ has a beta distribution with parameters a and b. The improper Dirichlet prior $0 \cdot (a\delta_1 + b\delta_0)$ corresponds to a two-point prior for ρ on $\{1, 0\}$.

An advantage of the Dirichlet process model is that, with respect to the topology of convergence in distribution, the support of P is the set of all distributions whose supports are contained in the support of α (Ferguson, 1973, Proposition 3). This provides an essentially nonparametric approach allowing us

to model those situations in which the responses can take on values in a specified set. In particular, as opposed to the Bernoulli model, we can model responses which are other than 0–1; in contrast with the normal model, we have more liberty in modeling the marginal distributions for the observations and so, for example, we can limit the possible outcomes to be other than the real line.

It is clear from (1.1) and (1.2), and true generally in bandit problems (Berry and Fristedt, 1985), that strategies can be specified by initial selections. For example, if arm 1 is selected at stage 1 then at stage 2 the strategy must specify the *first* selection for a new bandit in which α is replaced with $\alpha + \delta_x$ and n is decreased by 1.

Ideally, we would like to give an explicit specification of an optimal selection for any α, λ, and n. However, this is next to impossible unless n is very small or n is moderate and the support of α contains a small number of points. Instead, we give partial characterizations of optimal strategies. Section 3 deals with a monotonicity property of optimal strategies: Roughly, the larger the observations from arm 1, the greater the inclination to continue selecting that arm. A specific version of this notion is a form of "stay-with-a-winner" rule: There exists a quantity $b_n(\alpha)$ such that, if arm 1 is optimal initially, and if $X_1 = x$ is observed, then arm 1 is optimal again provided $x \geq b_n(\alpha)$. In Section 3 we also indicate that a "stay-with-a-winner/switch-on-a-loser" strategy is optimal, and we give an example showing that such a result may not apply in models with structure different from the Dirichlet.

Since optimal strategies are difficult to determine, we may wish to find nearly optimal strategies which are easy to describe. It seems desirable for such rules to have the stay-with-a-winner property mentioned above. However, the exact determination of $b_n(\alpha)$ requires actually finding an optimal strategy. Accordingly, in Section 4 we present an easily determined upper bound for $b_n(\alpha)$ which may be used to approximate an optimal rule. (We shall evaluate such an approximate rule in a future paper.) We compute this upper bound in two examples and give a conjecture for a lower bound for $b_n(\alpha)$.

Using notation similar to that of Berry (1972), let $W_n(\alpha, \lambda)$ be the expected payoff of an optimal strategy, where α, λ, and n are as described above. Let $W_n^i(\alpha, \lambda)$ be the expected payoff attained by selecting arm i initially and then proceeding optimally. We then have $W_n = W_n^1 \vee W_n^2$. For $n \geq 1$ the following relations are evident:

$$W_n^1(\alpha, \lambda) = \mu + E\left[W_{n-1}(\alpha + \delta_X, \lambda) | \alpha\right], \tag{1.1}$$

$$W_n^2(\alpha, \lambda) = \lambda + W_{n-1}(\alpha, \lambda). \tag{1.2}$$

Note that (1.1) and (1.2) hold for $\alpha = 0 \cdot F$ with the convention that the parameter $0 \cdot F + \delta_x$ equals δ_x.

Together with the evident initial conditions $W_0^i(\alpha, \lambda) = 0$ for $i = 1, 2$, (1.1) and (1.2) give a recursion for determining $W_n(\alpha, \lambda)$. Some properties of W_n follow from the straightforward extension of results in Berry and Fristedt (1979). For example, $W_n(\alpha, \lambda)$ is nondecreasing in both n and λ, and is continuous in λ. Further properties of W_n are given in Section 2. Repeated application of (1.1) and

(1.2) gives all optimal strategies by keeping track of whether the various W_{n-j} are equal to W^1_{n-j} or W^2_{n-j}.

Note that $W_{n-j}(\alpha + \sum_1^j \delta_{x_i}, \lambda)$ is measurable and integrable for $j = 1, \ldots, n-1$, and for $x_1, \ldots, x_j$ in the support of α. Measurability follows from (1.1) and (1.2) and the fact that the integral of measurable functions is measurable (Billingsley, 1979, Theorem 18.3). Integrability follows from

$$(1.3)\qquad (n-j)\lambda \vee \frac{M\mu + \sum_1^j \delta_{x_i}}{M+j} \le W_{n-j}\left(\alpha + \sum_1^j \delta_{x_i}, \lambda\right) \le \frac{M}{M+j} E(X \vee \lambda | \alpha) + \frac{1}{M+j} \sum_1^j (x_i \vee \lambda).$$

These inequalities correspond to the intuitive notion that the maximum expected payoff is at least that of a strategy in which the same arm is selected at every stage, and at most that for the case in which P is known at the outset.

The problem described here is a finite-horizon two-armed bandit with one arm known. The special case of Theorem 2.1 of Berry and Fristedt (1979) in which the discount sequence has a finite horizon applies in the current more general setting to show that we have described an optimal stopping problem. So, we need not consider strategies in which a selection of arm 2 is followed by a selection of arm 1. Consequently, to determine an optimal strategy, we need only determine the stage at which arm 2 is first selected, if ever. Problems of this sort are referred to as "one-armed bandits" (Berry, 1985). If $\lambda = 0$ this description is especially fitting since a selection of arm 2 in that case has no effect on the sum. Hence, when $\lambda = 0$ the problem is to select at most n observations from a population, such that the actual number of observations taken is decided upon sequentially and the goal is to maximize the expected sum of the observations taken. (We could always assume $\lambda = 0$ simply by subtracting λ from every observation made from arms 1 and 2. However, in terms of describing strategies it is easier to let λ be arbitrary, and so we shall work in that more general setting.)

2. Properties of optimal strategies. In this section we describe some properties of optimal strategies and their expected payoffs. A useful tool is the "break-even value" for λ of Bradt et al. (1956, Lemma 4.2) and Berry and Fristedt (1979, Theorem 2.2). As indicated by Berry and Fristedt, their result applies more generally than the Bernoulli and includes the current setting. We state the appropriate version without proof.

THEOREM 2.1. *For each α and n there exists a $\Lambda_n(\alpha)$ such that the only optimal initial actions are, select arm 1 if $\lambda \le \Lambda_n(\alpha)$ and select arm 2 if $\lambda \ge \Lambda_n(\alpha)$.*

At the second stage, we compare λ to $\Lambda_{n-1}(\alpha)$ or $\Lambda_{n-1}(\alpha + \delta_{x_1})$ according as arm 2 was selected initially or arm 1 was selected and $X_1 = x_1$ observed; and so

on for subsequent stages. In general, we can specify an optimal strategy by evaluating $\Lambda_j(\alpha + \sum_1^{n-j}\delta_{x_i})$ for each value of m and for each possible set of outcomes $x_1, \ldots, x_j$.

Since the problem is an optimal stopping problem it follows that $\lambda > \Lambda_n(\alpha)$ implies $\lambda > \Lambda_{n-1}(\alpha)$. That is, $\Lambda_n(\alpha)$ is nondecreasing in n. Another consequence of the optimal stopping nature of this problem is that $\lambda \geq \Lambda_n(\alpha)$ implies $W_n(\alpha, \lambda) = n\lambda$, while $\lambda < \Lambda_n(\alpha)$ implies $W_n(\alpha, \lambda) > n\lambda$. This gives a characterization of $\Lambda_n(\alpha)$ which lets us translate properties of W_n into properties of $\Lambda_n(\alpha)$. Namely:

LEMMA 2.1. *For $n \geq 1$ and for all α, $\Lambda_n(\alpha)$ is the smallest λ such that $W_n(\alpha, \lambda) - n\lambda \leq 0$.*

We mentioned in Section 1 that $W_n(\alpha, \lambda)$ is nondecreasing in λ. One might also expect the expected payoff to increase if we add a constant to each observation from arm 1. This is a special case of a more general notion:

DEFINITION 2.1. The distribution function F^* is *to the right of F* if $F^*(x) \leq F(x)$, $x \in \mathbb{R}$.

The following lemma is Proposition 17.A.1 in Marshall and Olkin (1979):

LEMMA 2.2. *If F^* is to the right of F and if g is nondecreasing, then $E[g(X)|F] \leq E[g(X)|F^*]$ whenever both expectations exist.*

We now set out to prove that $W_n(MF, \lambda)$ increases when F moves to the right. First, it is necessary to prove a special case.

PROPOSITION 2.1. *For all F and $M \geq 0$, and for $k > 0$, $W_n(MF + k\delta_z, \lambda)$ is nondecreasing in z.*

REMARK. If $z < z'$ then the distribution function form of $MF + k\delta_{z'}$ is $(MF + k\delta_{z'})/(M + k)$ and is to the right of the distribution function form of $MF + k\delta_z$.

PROOF. By induction, using (1.1) and (1.2). □

The next result indicates that, given the observation $X_1 = z$, one's inclination to select arm 1 should increase with z.

COROLLARY 2.1. *For all $M \geq 0$, for all F, and for $n \geq 1$, $\Lambda_n(MF + \delta_z)$ is nondecreasing in z.*

PROOF. Let $z < z'$. By Proposition 2.1,

$$W_n(MF + \delta_z, \Lambda_n(MF + \delta_{z'})) \leq W_n(MF + \delta_{z'}, \Lambda_n(MF + \delta_{z'})),$$

and so by Lemma 2.1,

$$W_n(MF + \delta_z, \Lambda_n(MF + \delta_{z'})) - n\Lambda_n(MF + \delta_{z'}) \leq 0.$$

But Lemma 2.1 then implies $\Lambda_n(MF + \delta_z) \leq \Lambda_n(MF + \delta_{z'})$. □

We now generalize Proposition 2.1.

PROPOSITION 2.2. *Fix $M \geq 0$, $n \geq 1$ and λ. If F^* is to the right of F, then*

$$W_n(MF, \lambda) \leq W_n(MF^*, \lambda).$$

PROOF. By induction. Assume the result holds for $n = m - 1$. By (1.1),

$$\begin{aligned} W_m^1(MF^*, \lambda) &- W_m^1(MF, \lambda) \\ &= E[X|F^*] - E[X|F] + E[W_{m-1}(MF^* + \delta_X, \lambda)|F^*] \\ &\quad - E[W_{m-1}(MF + \delta_X, \lambda)|F] \\ &\geq E[W_{m-1}(MF^* + \delta_X, \lambda)|F] - E[W_{m-1}(MF + \delta_X, \lambda)|F] \geq 0. \end{aligned}$$

The first inequality holds by Proposition 2.1 and Lemma 2.2; the second inequality holds by the induction hypothesis and the fact that $(MF^* + \delta_x)/(M + 1)$ is to the right of $(MF + \delta_x)/(M + 1)$. The remainder of the proof follows easily. □

COROLLARY 2.2. *For all $n > 1$ and $M > 0$, $\Lambda_n(MF) \leq \Lambda_n(MF^*)$ when F^* is to the right of F.*

PROOF. Follows by Lemma 2.1 and Proposition 2.2. □

REMARK. Results similar to Proposition 2.2 and Corollary 2.2 were proved by Berry and Fristedt (1979, Theorem 3.1) for the Bernoulli model.

3. Break-even observations. In this section we show that optimal strategies continue with arm 1 whenever it yields a sufficiently large observation. This can be viewed as a generalization of the stay-with-a-winner rule considered for the finite-horizon Bernoulli one-armed bandit in Bradt et al. (1956) and for the Bernoulli one-armed bandit with a *regular* discount sequence in Berry and Fristedt (1979). [See Berry (1985) or Berry and Fristedt (1985) for other references.] For the Bernoulli bandit, such a rule says that if it is optimal to select arm 1 initially, and if a success is obtained, then it is optimal to select arm 1 again. The analogous result for the Dirichlet bandit is given in Theorem 3.1.

We first show that some observations increase and others decrease the desirability of arm 1 as reflected by Λ. This gives a very weak stay-with-a-winner rule.

PROPOSITION 3.1. *Given α and $n \geq 2$, there exist x' and x'' such that*

$$\Lambda_{n-1}(\alpha + \delta_{x'}) \leq \Lambda_n(\alpha) \leq \Lambda_{n-1}(\alpha + \delta_{x''}). \tag{3.1}$$

Moreover, if the support of α is bounded above by U, we can take $x'' = U$; if the

support of α is bounded below by L, then we can take $x' = L$.

PROOF. That x'' exists follows since

$$\lim_{x\to\infty} \Lambda_n(\alpha + \delta_x) \geq \lim_{x\to\infty} \Lambda_1(\alpha + \delta_x) = \lim_{x\to\infty} (M\mu + x)/(M+1) = \infty.$$

That x' exists follows from $\lim_{x\to-\infty}\Lambda_n(\alpha + \delta_x) = -\infty$, a consequence of Lemma 2.1 and the fact that $\lim_{x\to-\infty}W_n^1(\alpha + \delta_x, \lambda) < n\lambda$.

Suppose now that the support of α is bounded above by U. To show that (3.1) is satisfied with $x'' = U$ we adapt the proof of Theorem 4.1 in Berry and Fristedt (1979). Namely, we suppose $\lambda \leq \Lambda_n(\alpha)$ and show $\lambda \leq \Lambda_{n-1}(\alpha + \delta_U)$. There are two cases: (i) $\lambda \leq E(X|\alpha + \delta_U)$ and (ii) $\lambda > E(X|\alpha + \delta_U)$. In case (i), $\lambda \leq E(X|\alpha + \delta_U) = \Lambda_1(\alpha + \delta_U) \leq \Lambda_{n-1}(\alpha + \delta_U)$ since Λ_n is nondecreasing in n. In case (ii), suppose, to the contrary, that $\lambda > \Lambda_{n-1}(\alpha + \delta_U)$. Then $\lambda \geq \Lambda_{n-1}(\alpha + \delta_x)$ for $x \leq U$ by Corollary 2.1, and so, for any outcome on arm 1, Theorem 2.1 applies to show that arm 2 is optimal for the remaining selections. Consequently,

$$\begin{aligned} W_n(\alpha, \lambda) &= \mu + (n-1)\lambda \leq (M\mu + U)/(M+1) + (n-1)\lambda \\ &= E(X|\alpha + \delta_U) + (n-1)\lambda < n\lambda, \end{aligned}$$

which is impossible since $W_n(\alpha, \lambda) \geq n\lambda$ in view of (1.3).

Finally, if the support of α is bounded below by L, then we have

$$\Lambda_{n-1}(\alpha + \delta_L) \leq \Lambda_{n-1}(\alpha) \leq \Lambda_n(\alpha). \tag{3.2}$$

The first inequality in (3.2) follows from Corollary 2.2 since the normalized form of α is to the right of the normalized form of $\alpha + \delta_L$. The second inequality in (3.2) follows since Λ is nondecreasing in n. □

The next two lemmas will be used in the proof of Theorem 3.1.

LEMMA 3.1. *For all $n \geq 1$, for $k = 0, 1, 2, \ldots$ and for $x_1, x_2, \ldots, x_k$ given,*
(i) *$W_n(\alpha + \sum_1^k \delta_{x_i} + \delta_z, \lambda)$ is jointly continuous in z and λ; and*
(ii) *$W_n^j(\alpha + \delta_z, \lambda)$ is jointly continuous in z and λ for $j = 1, 2$.*

PROOF. Part (i) follows by a straightforward induction from (1.1) and (1.2). Part (ii) follows from part (i). □

LEMMA 3.2. *For α and n given, $\Lambda_n(\alpha + \delta_x)$ is continuous in x.*

PROOF. Fix x_0 and note that $\Lambda_n(\alpha + \delta_x)$ is the unique root in λ of $W_n^1(\alpha + \delta_x, \lambda) - W_n^2(\alpha + \delta_x, \lambda) = 0$. In view of Lemma 3.1 (ii),

$$\begin{aligned} 0 &= \lim_{x\to x_0} \left[W_n^1(\alpha + \delta_x, \Lambda_n(\alpha + \delta_x)) - W_n^2(\alpha + \delta_x, \Lambda_n(\alpha + \delta_x))\right] \\ &= W_n^1\Big(\alpha + \delta_{x_0}, \lim_{x\to x_0} \Lambda_n(\alpha + \delta_x)\Big) - W_n^2\Big(\alpha + \delta_{x_0}, \lim_{x\to x_0} \Lambda_n(\alpha + \delta_x)\Big). \end{aligned}$$

By uniqueness, $\Lambda_n(\alpha + \delta_{x_0}) = \lim_{x\to x_0}\Lambda_n(\alpha + \delta_x)$. □

THEOREM 3.1. *Given α and $n \geq 2$, there exists a unique $b_n(\alpha)$ such that*

$$\Lambda_{n-1}(\alpha + \delta_x) \geq \Lambda_n(\alpha) \quad \text{if } x \geq b_n(\alpha)$$

and

$$\Lambda_{n-1}(\alpha + \delta_x) \leq \Lambda_n(\alpha) \quad \text{if } x \leq b_n(\alpha).$$

PROOF. Fix α and $n \geq 2$. In view of Corollary 2.1, to show existence it suffices to demonstrate that there exists a quantity b such that $\Lambda_n(\alpha) = \Lambda_{n-1}(\alpha + \delta_b)$. But this follows directly from Proposition 3.1, another application of Corollary 2.1, and Lemma 3.2.

We show that $b_n(\alpha)$ is unique by contradiction. Suppose $b < b'$ and that

$$\lambda = \Lambda_n(\alpha) = \Lambda_{n-1}(\alpha + \delta_b) = \Lambda_{n-1}(\alpha + \delta_{b'}).$$

Then since an initial selection of arm 2 is optimal,

$$W_{n-1}(\alpha + \delta_b, \lambda) = (n-1)\lambda = W_{n-1}(\alpha + \delta_{b'}, \lambda).$$

However, since arm 1 is also optimal, by (1.1) and (1.2) we have

$$\begin{aligned} W_{n-1}(\alpha + \delta_{b'}, \lambda) - W_{n-1}(\alpha + \delta_b, \lambda) &= (b' - b)/(M+1) \\ &\quad + E\left[W_{n-2}(\alpha + \delta_{b'} + \delta_X, \lambda) | \alpha + \delta_{b'}\right)] \\ &\quad - E\left[W_{n-2}(\alpha + \delta_b + \delta_X, \lambda) | \alpha + \delta_b)\right] \\ &> 0. \end{aligned}$$

□

We call $b_n(\alpha)$ a "break-even observation" for an obvious reason: It gives the stay-with-a-winner property previously mentioned. A proof similar to that of Theorem 3.1 yields another break-even observation, $c_n(\alpha, \lambda)$, which results in a still stronger property of optimal strategies:

THEOREM 3.2. *Given α, λ, and $n \geq 2$, there exists a unique $c_n(\alpha, \lambda)$ such that*

$$\Lambda_{n-1}(\alpha + \delta_x) \geq \lambda \quad \text{if } x \geq c_n(\alpha, \lambda) \quad \text{and} \quad \Lambda_{n-1}(\alpha + \delta_x) \leq \lambda \quad \text{if } x \leq c_n(\alpha, \lambda).$$

This result gives a "stay-with-a-winner/switch-on-a-loser" rule: If arm 1 is selected, optimally or not, and $X_1 = x$ is observed, then arm 1 is optimal if $x \geq c_n(\alpha, \lambda)$; arm 2 is optimal if $x \leq c_n(\alpha, \lambda)$; and both are optimal if $x = c_n(\alpha, \lambda)$. Hence after an optimal initial selection, optimal strategies are completely determined by $c_j(\alpha + \sum_1^{n-j} \delta_{x_i}, \lambda)$ for the various possible outcomes. In this sense c plays a role similar to that of Λ.

It is easy to show from Theorems 2.1, 3.1, and 3.2 that if $\lambda \leq \Lambda_n(\alpha)$, then $c_n(\alpha, \lambda) \leq c_n(\alpha, \Lambda_n(\alpha))$ and, moreover, $c_n(\alpha, \Lambda_n(\alpha)) = b_n(\alpha)$. However, if the support of α has upper bound U, then it need not be the case that $c_n(\alpha, \lambda) \leq U$. Likewise, if the support of α has lower bound L, it need not hold that $c_n(\alpha, \lambda) \geq L$. (Easy counterexamples are given by taking $\alpha = \delta_\mu$ and $\lambda > \mu$ in the first case and $\lambda < \mu$ in the second.)

If the p.m. P is not a Dirichlet process, then $b_n(\alpha)$ and $c_n(\alpha, \lambda)$ need not exist, as the next example shows.

EXAMPLE 3.1. Let $P = \delta_7$ with probability $\frac{1}{2}$ and $P = (\frac{1}{2})(\delta_0 + \delta_{10})$ with probability $\frac{1}{2}$. Let $\lambda = 6$. For any $n \geq 2$ arm 1 is optimal initially, and also for the second stage if $X_1 = 7$, but not if $X_1 = 0$ or 10. □

Our comments in Section 1 imply that the explicit determination of $c_n(\alpha, \lambda)$ is nearly impossible unless n is small or n is moderate and α has only a few support points. While the same is true for $b_n(\alpha)$, the latter is a simpler quantity (for example, it has only one argument) and we can find a useful upper bound for $b_n(\alpha)$. We now turn to this task.

4. A bound for the break-even observation $b_n(\alpha)$. In this section we find an upper bound for $b_n(\alpha)$, thus suggesting easily determined strategies that may perform well. Any quantity x'' given in Proposition 3.1 is an upper bound for $b_n(\alpha)$. Unfortunately, beyond knowing that it exists, we have little guidance in checking to see if a particular number (save an upper bound U of the support of α) qualifies for x''.

We present a series of results which show that $b_n(\alpha) \leq \Lambda_n(0 \cdot F)$. It is easy to show that $\Lambda_n(0 \cdot F) \leq U$ when U exists, so $\Lambda_n(0 \cdot F)$ is never worse than U in bounding $b_n(\alpha)$. Moreover, $\Lambda_n(0 \cdot F)$ is easy to compute. When $M = 0$, a single observation from arm 1 yields complete information about P, and so

$$W_n(0 \cdot F, \lambda) = \max\{\mu + (n-1)E[(X \vee \lambda)|\alpha], n\lambda\}.$$

Thus, by Lemma 2.1, $\Lambda_n(0 \cdot F)$ is the smallest λ that satisfies

$$\max\{\mu + (n-1)E[(X - \lambda)^+] - \lambda, 0\} = 0,$$

where $a^+ = a \vee 0$. It follows that $\Lambda_n(0 \cdot F)$ uniquely satisfies

$$\Lambda_n(0 \cdot F) = \mu + (n-1)E(X - \Lambda_n(0 \cdot F))^+. \tag{4.1}$$

To show $b_n(\alpha) \leq \Lambda_n(0 \cdot F)$, we first prove a seemingly unrelated result. Let x and γ be fixed values, $x \geq \gamma$. While it is true that $W_n(\alpha + \delta_x, \lambda) \geq W_n(\alpha + \delta_\gamma, \lambda)$, a bandit with Dirichlet process parameter $\alpha + \delta_x$ is not preferred to a bandit with parameter $\alpha + \delta_\gamma$ when $(x - \gamma)/(M + 1)$ is added to each observation from the latter bandit. More specifically,

LEMMA 4.1. *For all $\gamma, \alpha, \lambda, k > 0$, and for $n \geq 1$, if $x \geq \gamma$ then*

$$\frac{nk(x - \gamma)}{M + k} + W_n(\alpha + k\delta_\gamma, \lambda) - W_n(\alpha + k\delta_x, \lambda) \geq 0. \tag{4.2}$$

PROOF. We proceed by induction. The case $n = 1$ is straightforward.

For the induction step, suppose the lemma is true when $n = m - 1$. Then we have two cases. In case (i), $\lambda \geq \Lambda_m(\alpha + k\delta_x) \geq \Lambda_m(\alpha + k\delta_\gamma)$. Then the left side of (4.2) is $mk(x - \gamma)/(M + k) + m\lambda - m\lambda$, which is nonnegative.

In case (ii), $\Lambda_m(\alpha + k\delta_x) \geq \lambda$. Now, $W_m(\alpha + k\delta_x, \lambda) = W_m^1(\alpha + k\delta_x, \lambda)$. It suffices to prove

$$\frac{mk(x - \gamma)}{M + k} + W_m^1(\alpha + k\delta_\gamma, \lambda) - W_m^1(\alpha + k\delta_x, \lambda) \geq 0, \tag{4.3}$$

since the left side of (4.3) is a lower bound for the left side of (4.2) when $n = m$. But using (1.1), the left side of (4.3) can be shown to equal

$$\frac{mk(x-\gamma)}{M+k} + \frac{k(\gamma - x)}{M+k} + \frac{M}{M+k}E\left[W_{m-1}(\alpha + k\delta_\gamma + \delta_X, \lambda)|\alpha\right]$$
$$+ \frac{k}{M+k}W_{m-1}(\alpha + (k+1)\delta_\gamma, \lambda) - \frac{M}{M+k}E\left[W_{m-1}(\alpha + k\delta_x + \delta_X, \lambda)|\alpha\right]$$
$$- \frac{k}{M+k}W_{m-1}(\alpha + (k+1)\delta_x, \lambda).$$

Straightforward manipulation shows that this equals

$$\frac{M}{M+k}E\left\{\left[\frac{(m-1)k(x-\gamma)}{M+k+1} + W_{m-1}(\alpha + \delta_X + k\delta_\gamma, \lambda) - W_{m-1}(\alpha + \delta_X + k\delta_x, \lambda)\right]\right\}$$
$$+ \frac{k}{M+k}\left[\frac{(m-1)(k+1)(x-\gamma)}{M+k+1} + W_{m-1}(\alpha + (k+1)\delta_\gamma, \lambda) - W_{m-1}(\alpha + (k+1)\delta_x, \lambda)\right]. \tag{4.4}$$

But the quantities in square brackets in (4.4) are nonnegative by the induction hypothesis. □

LEMMA 4.2. *For $n \geq 2$, for all $\alpha = MF$, and for all λ,*

$$W_n(\alpha, \lambda) \leq \left[\lambda \vee \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})\right] + W_{n-1}(\alpha + \delta_{\Lambda_n^0}, \lambda) \tag{4.5}$$

where $\Lambda_n^0 = \Lambda_n(0 \cdot F)$.

PROOF. There are two cases. In case (i), $\lambda > \Lambda_n(\alpha)$. Then $W_n(\alpha, \lambda) = n\lambda$ and (4.5) is immediate. In case (ii), $\lambda \leq \Lambda_n(\alpha)$. Now, $W_n(\alpha, \lambda) = W_n^1(\alpha, \lambda)$, and since Λ_n is nondecreasing in n, it will suffice to prove

$$W_n^1(\alpha, \lambda) \leq \Lambda_1(\alpha + \delta_{\Lambda_n^0}) + W_{n-1}(\alpha + \delta_{\Lambda_n^0}, \lambda). \tag{4.6}$$

Now by (4.1),

$$\Lambda_1(\alpha + \delta_{\Lambda_n^0}) = \frac{M\mu + \Lambda_n^0}{M+1} = \mu + (n-1)E\left[(X - \Lambda_n^0)^+\right]/(M+1).$$

Using this and (1.2), we see that (4.6) is equivalent to $E[f(X)] \geq 0$, where

$$f(x) = \frac{(n-1)(x - \Lambda_n^0)^+}{M+1} + W_{n-1}(\alpha + \delta_{\Lambda_n^0}, \lambda) - W_{n-1}(\alpha + \delta_x, \lambda).$$

However, when $x \leq \Lambda_n^0$, $f(x) \geq 0$ by Proposition 2.1, while if $x > \Lambda_n^0$, $f(x) \geq 0$ by Lemma 4.1. □

THEOREM 4.1. *For all α and for $n \geq 2$, $b_n(\alpha) \leq \Lambda_n(0 \cdot F)$.*

PROOF. Let $\Lambda_n^0 = \Lambda_n(0 \cdot F)$ and take $\lambda = \lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})$ in (4.5). Then

$$W_n\big(\alpha, \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})\big) \leq \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0}) + W_{n-1}\big(\alpha + \delta_{\Lambda_n^0}, \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})\big)$$
$$= n\Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})$$

by Theorem 2.1. But by Lemma 2.1, $\Lambda_n(\alpha) \leq \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})$. The desired result now follows by the definition of $b_n(\alpha)$ and Corollary 2.1. □

Theorem 4.1 gives an easily described strategy that has the stay-with-a-winner property: If arm 1 is optimal initially in a bandit with n observations and parameter MF, and if $X_1 \geq \Lambda_n(0 \cdot F)$ is observed, then arm 1 is used again. This second selection of arm 1 is optimal, but selecting arm 2 when $X_1 < \Lambda_n(0 \cdot F)$ may not be optimal.

EXAMPLE 4.1. If $\alpha = M(a\delta_1 + b\delta_0)$, $0 < a < 1$, $a + b = 1$, then $\Lambda_n[0 \cdot (a\delta_1 + b\delta_0)] = na/[(n-1)a + 1]$. Since $\Lambda_n^0 \in (0,1)$ in this case, Theorem 4.1 gives the traditional stay-with-a-winner rule for the Bernoulli bandit. □

We conclude this section with tables of Λ_n and b_n for two examples. Let H denote the distribution function of a continuous uniform random variable on $[0,1]$, and let Φ denote the standard normal distribution function. In Table 4.1

TABLE 4.1
The quantities $\Lambda_n = \Lambda_n(MF)$ and $b_n = b_n(MF)$

a. $F = \Phi$

M	Λ_2	b_2	Λ_3	b_3	Λ_4	b_4
0	0.276	0.276	0.436	0.436	0.549	0.549
0.1	0.251	0.276	0.400	0.424	0.505	0.529
0.5	0.184	0.276	0.300	0.388	0.383	0.470
1	0.138	0.276	0.228	0.359	0.295	0.421
5	0.046	0.276	0.079	0.276	0.105	0.284
10	0.028	0.276	0.043	0.248	0.058	0.238
100	0.003	0.276	0.005	0.214	0.007	0.182

b. $F = H$

M	Λ_2	b_2	Λ_3	b_3	Λ_4	b_4
0	0.586	0.586	0.634	0.634	0.667	0.667
0.1	0.578	0.586	0.623	0.630	0.654	0.661
0.5	0.557	0.586	0.592	0.619	0.617	0.644
1	0.543	0.586	0.570	0.610	0.590	0.630
5	0.514	0.586	0.524	0.584	0.532	0.589
10	0.508	0.586	0.513	0.576	0.518	0.574
100	0.501	0.586	0.501	0.565	0.502	0.554

values of $\Lambda_n(MF)$ and $b_n(MF)$ are given for $n = 2, 3, 4$; $F = H$ and Φ; and $M = 0, .1, .5, 1, 5, 10, 100$.

It is not hard to show that $b_2(MF) = \Lambda_2(0 \cdot F)$ for all $M \geq 0$ and F, and that $b_n(0 \cdot F) = \Lambda_n(0 \cdot F)$ for $n \geq 2$ and all F. These facts are reflected in Table 4.1. As well, it is straightforward to prove that, for all nondegenerate F, $\Lambda_2(MF)$ is strictly decreasing in M. Table 4.1 suggests, and we conjecture, that $\Lambda_n(MF)$ is strictly decreasing in M for all $n \geq 2$ when F is nondegenerate. This reflects the intuitive notion that the less known about an arm, the more appealing it is since there is more information to be gained when selecting it.

For moderate M, $\Lambda_n(0 \cdot F)$ seems to be a reasonable upper bound for $b_n(MF)$. The effect of this in designing nearly optimal strategies remains to be seen. In Table 4.1, $\Lambda_n(\alpha) \leq b_n(\alpha)$. We conjecture that this holds for all n and α.

Acknowledgments. We thank Bert Fristedt for helpful discussions and a referee for suggestions which improved the paper.

REFERENCES

BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York.

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

BERRY, D. A. (1985). One- and two-armed bandit problems. In *Encyclopedia of Statistical Sciences* **6** (S. Kotz and N. L. Johnson, eds.). Wiley, New York.

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.

CHERNOFF, H. (1968). Optimal stochastic control. *Sankhyā Ser. A* **30** 221–252.

CLAYTON, M. K. (1983). Bayes sequential sampling for choosing the better of two populations. Unpublished thesis, University of Minnesota.

FAHRENHOLTZ, S. K. (1982). Normal Bayesian two-armed bandits. Unpublished thesis, Iowa State University.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

MARSHALL, A. W. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York.

SETHURAMAN, J. and TIWARI, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics III* **2** (S. S. Gupta and J. O. Berger, eds.). Academic Press, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455