The Annals of Statistics
1994, Vol. 22, No. 3, 1212–1221

TWO-ARMED DIRICHLET BANDITS WITH DISCOUNTING[1]

BY MANAS K. CHATTOPADHYAY

Gallup Organization

Sequential selections are to be made from two independent stochastic processes, or "arms." At each stage we choose which arm to observe based on past selections and observations. The observations on arm i are conditionally i.i.d. given their marginal distribution P_i which has a Dirichlet process prior with parameter α_i, $i = 1, 2$. Future observations are discounted: at stage m, the payoff is a_m times the observation Z_m at that stage. The discount sequence $A_n = (a_1, a_2, \dots, a_n, 0, 0, \dots)$ is a nonincreasing sequence of nonnegative numbers, where the "horizon" n is finite. The objective is to maximize the total expected payoff $E(\Sigma_1^n a_i Z_i)$. It is shown that optimal strategies continue with an arm when it yields a sufficiently large observation, one larger than a "break-even observation." This generalizes results of Clayton and Berry, who considered two arms with one arm known and assumed $a_m = 1 \;\forall\, m \le n$.

1. Introduction. A bandit problem involves sequential selections from $k\,(\ge 2)$ stochastic processes (or "arms," machines, treatments etc.). We restrict consideration to discrete time and two independent arms. Each of the arms generates an infinite sequence of random variables. An observation on a particular sequence is made by selecting the corresponding arm. The mth member of a sequence is observed if the corresponding arm is selected at stage m. Future observations are discounted with payoff at stage m equal to a_m times the observation Z_m at that stage. Assume $0 < a_n \le \cdots \le a_1$; $A_n = (a_1, a_2, \dots, a_n, 0, 0, \dots)$ is called a discount sequence. The *horizon* n of the discount sequence A_n equals $\inf\{r\colon a_m = 0,\ \forall\, m > r\}$. Let $A_n^k = (a_{k+1}, a_{k+2}, \dots, a_n, 0, 0, \dots)$ for $k < n$.

The selection of a process for observation at any time depends on the previous selections and results. A decision procedure (or strategy) specifies which process to select at any time for every history of previous selections and observations. The objective is to maximize the expected value of $\sum_{m=1}^{n} a_m Z_m$. Any strategy yielding the maximum expected payoff is called optimal. An arm is optimal if it is the first selection of some optimal strategy. An arm is uniquely optimal if only that arm is optimal (i.e., no other arm is optimal).

Let X_i and Y_i denote the results from arms 1 and 2, respectively, at stage i for $i = 1, 2, \dots, n$. At any stage, one of the pair (X_i, Y_i) is actually observed. The two arms are independent and hence the random vector $(X_1, X_2, \dots, X_n)$ is independent of $(Y_1, Y_2, \dots, Y_n)$. Moreover, given the unknown probability measure

Received July 1991; revised September 1993.

[1] Supported in part by NSF Grants DMS-88-03087 and DMS-89-11548.

AMS 1991 *subject classifications.* Primary 62L05; secondary 62C10.

Key words and phrases. Sequential decisions, two-armed bandits, one-armed bandits, Dirichlet bandits, Dirichlet process prior.

P_i, $i = 1, 2$, the random variables $X_1, X_2, \dots, X_n$ are i.i.d. with probability measure P_1, while the random variables $Y_1, Y_2, \dots, Y_n$ are i.i.d. with probability measure P_2.

Following a Bayesian approach, we take each P_i, $i = 1, 2$, to be random having a Dirichlet process prior [Ferguson (1973)]. Below we give the definition of Dirichlet process prior and some of its properties as introduced by Ferguson (1973).

DEFINITION 1. Let $\mathcal{R}$ be a set and $\mathcal{B}$ a σ-field of subsets of $\mathcal{R}$. Let α be a nonnull finite measure (nonnegative and finitely additive) on $(\mathcal{R}, \mathcal{B})$. A random probability P on $(\mathcal{R}, \mathcal{B})$ is a Dirichlet process on $(\mathcal{R}, \mathcal{B})$ with parameter α, denoted $P \in D(\alpha)$, if, for every $k = 1, 2, \dots$ and measurable partition $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_k$ of $\mathcal{R}$, the joint distribution of the random probabilities $(P(\mathcal{B}_1), P(\mathcal{B}_2), \dots, P(\mathcal{B}_k))$ is Dirichlet with parameters $(\alpha(\mathcal{B}_1), \alpha(\mathcal{B}_2), \dots, \alpha(\mathcal{B}_k))$.

We now assume that $\mathcal{R}$ denotes the real line, $\mathcal{B}$ denotes the σ-field of Borel sets and $M = \alpha(\mathcal{R}) < \infty$. Then $F(x) = \alpha(-\infty, x]/M$ is the distribution function corresponding to α. If $P \in D(\alpha)$ and $A_0 \in \mathcal{B}$, then $E(P(A_0)) = \alpha(A_0)/M$. If $P \in D(\alpha)$ and $X_1, X_2, \dots, X_n$ is a sample of size n from P, then the conditional distribution of P given $X_1, X_2, \dots, X_n$ is a Dirichlet process with parameter $\alpha + \sum_1^n \delta_{X_i}$, where δ_x assigns mass 1 at x [Ferguson (1973), Theorem 1]. Using these properties, F is the prior mean for P in the sense that it is the expectation of $P((-\infty, x])$, and the total measure M may be interpreted as the degree of faith in the prior guess $F(t)$ of the unknown $F_0(t) = P((-\infty, t])$ [Ferguson (1973), page 223]. Another advantage of using a Dirichlet process prior (with parameter α) for an unknown distribution P is that, with respect to the topology of convergence in distributions, the support of P is the set of all distributions whose supports are contained in the support of α [Ferguson (1973), Proposition 3].

In our problem involving two unknown arms, we assume P_i to be random having a Dirichlet process prior with parameter α_i, a bounded nonnull measure on the reals $\mathcal{R}$ with finite first moment ($i = 1, 2$). Let $F_i(x) = \alpha_i(-\infty, x]/M_i$, so that F_i is the distribution function corresponding to α_i, $i = 1, 2$. The conditional expectation of any function $h(X)$ of X, an observation from arm 1, given $X_1, X_2, \dots, X_k$ can be computed for each $\alpha_1 + \sum_1^k \delta_{X_i}$ using Theorem 3 of Ferguson (1973). This expectation will be denoted by $E[h(X) \mid \alpha_1 + \sum_1^k \delta_{X_i}]$. Note in particular that

$$E(X \mid \alpha_1) = E \int x\, dP_1(x) = [\alpha_1(\mathcal{R})]^{-1} \int x\, d\alpha_1(x) = \int x\, dF_1(x) = \mu_1$$

and

$$E(Y \mid \alpha_2) = E \int y\, dP_2(y) = [\alpha_2(\mathcal{R})]^{-1} \int y\, d\alpha_2(y) = \int y\, dF_2(y) = \mu_2.$$

So $E(X \mid \alpha_1)$ is the unconditional expectation of X with distribution P_1 having a Dirichlet process prior with parameter α_1.

Let $(\alpha_1, \alpha_2; A_n)$ denote the two-armed bandit problem where arm i has a Dirichlet process prior with parameter α_i, $i = 1, 2$, and A_n is a nonincreasing discount sequence of finite horizon n. Computational problems make it difficult to give an explicit specification of an optimal strategy. We therefore give partial characterizations of optimal strategies by proving the existence of "break-even" observations. There exists a break-even observation $c(\alpha_1, \alpha_2; A_n)$ such that if arm 1 is selected initially, optimally or not, and $X_1 = x$ is observed, then arm 1 is optimal (though not necessarily uniquely optimal) at the next stage if $x \geq c(\alpha_1, \alpha_2; A_n)$, and arm 2 otherwise. This is like a "stay-with-a-winner/switch-on-a-loser" rule. The value of c, however, changes from $c(\alpha_1, \alpha_2; A_n)$ to $c(\alpha_1 + \delta_x, \alpha_2; A_n^1)$ at the next stage. There exists another kind of break-even observation $b(\alpha_1, \alpha_2; A_n)$ such that if arm 1 is selected initially, optimally or not, and $X_1 = x$ is observed, then the expected advantage (disadvantage) of choosing arm 1 over arm 2 in the remaining $(n - 1)$-horizon problem is not smaller (greater) than its initial value if $x \geq b(\alpha_1, \alpha_2; A_n)$; it is not greater (smaller) than its initial value if $x \leq b(\alpha_1, \alpha_2; A_n)$; and it remains unchanged if $x = b(\alpha_1, \alpha_2; A_n)$. So if arm 1 is optimal initially and $X_1 = x$ is observed, then arm 1 is optimal again provided $x \geq b(\alpha_1, \alpha_2; A_n)$. This is like a "stay-with-a-winner" rule, where "win" means obtaining a large observation. In the non-Dirichlet case, such break-even observations may not exist. Example 2.1 shows that a large observation can be quite distasteful in terms of desirability of the arm producing that observation. In Section 2, we prove the existence of break-even observations and present some numerical results. Our results generalize those of Clayton and Berry (1985), who consider the case where one of the two arms is known, the distribution of the unknown arm has a Dirichlet process prior and the discount sequence is finite horizon uniform (i.e., $a_m = 1 \; \forall \; m \leq n$). Such a "one-armed bandit" is an optimal stopping problem where it is not necessary to consider strategies in which selection of the known arm is followed by selection of the unknown arm. Consequently, to determine an optimal strategy, it is only necessary to find the stage at which the known arm is first selected, if ever. However, our problem involving two unknown arms is no longer a stopping problem, and the discount sequence A_n is any nonincreasing discount sequence of horizon n.

2. Break-even observations. In this section we prove that optimal strategies continue with an arm that yields a sufficiently large observation. The existence of an optimal strategy is proved in the general setting by Berry and Fristedt [(1985), Lemma 2.3.1]. For $(\alpha_1, \alpha_2; A_n)$ bandit, we use notation similar to that of Berry (1972): $W_\tau(\alpha_1, \alpha_2; A_n)$ is the expected payoff under strategy τ; $W(\alpha_1, \alpha_2; A_n)$ is the expected payoff under an optimal strategy; $W^i(\alpha_1, \alpha_2; A_n)$ is the expected payoff of a strategy starting with arm i and then proceeding optimally ($i = 1, 2$); $\Delta(\alpha_1, \alpha_2; A_n)$ is the expected advantage of choosing arm 1 over arm 2; $\Delta^+(\alpha_1, \alpha_2; A_n) = \max(0, \Delta(\alpha_1, \alpha_2; A_n))$; $\Delta^-(\alpha_1, \alpha_2; A_n) = \min(0, \Delta(\alpha_1, \alpha_2; A_n))$.

Using the above notation,

$$W^1(\alpha_1,\alpha_2;A_n) = a_1\mu_1 + E\Big[W\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big)\mid\alpha_1\Big], \tag{2.1}$$

$$W^2(\alpha_1,\alpha_2;A_n) = a_1\mu_2 + E\Big[W\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big)\mid\alpha_2\Big], \tag{2.2}$$

$$W(\alpha_1,\alpha_2;A_n) = \max\big(W^1(\alpha_1,\alpha_2;A_n),W^2(\alpha_1,\alpha_2;A_n)\big), \tag{2.3}$$

$$\begin{aligned}\Delta(\alpha_1,\alpha_2;A_n) &= W^1(\alpha_1,\alpha_2;A_n) - W^2(\alpha_1,\alpha_2;A_n)\\ &= -\Delta(\alpha_2,\alpha_1;A_n),\end{aligned}$$

$$W^1(\alpha_1,\alpha_2;A_n) = W(\alpha_1,\alpha_2;A_n) + \Delta^-(\alpha_1,\alpha_2;A_n), \tag{2.4}$$

$$W^2(\alpha_1,\alpha_2;A_n) = W(\alpha_1,\alpha_2;A_n) - \Delta^+(\alpha_1,\alpha_2;A_n). \tag{2.5}$$

Using (2.4) and (2.5) in (2.1) and (2.2), we get

$$W^1(\alpha_1,\alpha_2;A_n) = a_1\mu_1 + E\left[\left\{W^2\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big) + \Delta^+\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big)\right\}\Big|\,\alpha_1\right],$$

$$W^2(\alpha_1,\alpha_2;A_n) = a_1\mu_2 + E\left[\left\{W^1\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big) - \Delta^-\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big)\right\}\Big|\,\alpha_2\right].$$

Therefore,

$$\begin{aligned}\Delta(\alpha_1,\alpha_2;A_n) &= W^1(\alpha_1,\alpha_2;A_n) - W^2(\alpha_1,\alpha_2;A_n)\\ &= \Big[a_1\mu_1 + E\Big\{W^2\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big)\mid\alpha_1\Big\}\Big]\\ &\quad - \Big[a_1\mu_2 + E\Big\{W^1\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big)\mid\alpha_2\Big\}\Big]\\ &\quad + E\Big\{\Delta^+\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big)\mid\alpha_1\Big\} + E\Big\{\Delta^-\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big)\mid\alpha_2\Big\}.\end{aligned} \tag{2.6}$$

Using arguments similar to those in Berry and Fristedt (1985), $a_1\mu_1 + E\{W^2(\alpha_1+\delta_X,\alpha_2;A_n^1)\mid\alpha_1\}$ is the expected worth of selecting arm 1 first and arm 2 second and then continuing optimally. Similarly, $a_1\mu_2 + E\{W^1(\alpha_1,\alpha_2+\delta_Y;A_n^1)\mid\alpha_2\}$ is the expected worth of selecting arm 2 first and arm 1 second and then continuing optimally. Hence the first minus the second is $(a_1 - a_2)(\mu_1 - \mu_2)$. Using this in (2.6) gives

$$\begin{aligned}\Delta(\alpha_1,\alpha_2;A_n) &= (a_1-a_2)(\mu_1-\mu_2) + E\Big[\Delta^+\big(\alpha_1+\delta_X,\alpha_2;A_n^1\big)\mid\alpha_1\Big]\\ &\quad + E\Big[\Delta^-\big(\alpha_1,\alpha_2+\delta_Y;A_n^1\big)\mid\alpha_2\Big].\end{aligned} \tag{2.7}$$

The next result implies that, given a preliminary observation (from arm 1) $X_1 = x$, the advantage of choosing arm 1 over arm 2 increases with increase in x.

PROPOSITION 2.1. *For all* α_1, α_2, $k > 0$ *and nonincreasing discount sequence* A_n, $\Delta(\alpha_1 + k\delta_x, \alpha_2; A_n)$ *is nondecreasing in x.*

PROOF (By induction). For $n = 1$,

$$\Delta(\alpha_1 + k\delta_x, \alpha_2; A_1) = a_1\left(\frac{M_1\mu_1 + kx}{M_1 + k} - \mu_2\right)$$

is nondecreasing in x since $a_1 \geq 0$. Assume the monotonic property to be true for $n = m - 1$. By (2.7),

$$\begin{aligned} \Delta(\alpha_1 + k\delta_x, \alpha_2; A_m) &= (a_1 - a_2)\left(\frac{M_1\mu_1 + kx}{M_1 + k} - \mu_2\right) \\ &\quad + \frac{M_1}{M_1 + k}E\left[\Delta^+\left(\alpha_1 + k\delta_x + \delta_Z, \alpha_2; A_m^1\right) \mid \alpha_1\right] \\ &\quad + \frac{k}{M_1 + k}\Delta^+\left(\alpha_1 + (k+1)\delta_x, \alpha_2; A_m^1\right) \\ &\quad + E\left[\Delta^-\left(\alpha_1 + k\delta_x, \alpha_2 + \delta_Y; A_m^1\right) \mid \alpha_2\right]. \end{aligned} \tag{2.8}$$

The first term on the right-hand side of (2.8) is clearly nondecreasing in x. For the second term, observe that, for any fixed z and $x_1 \leq x_2$,

$$\Delta\left(\alpha_1 + \delta_z + k\delta_{x_1}, \alpha_2; A_m^1\right) \leq \Delta\left(\alpha_1 + \delta_z + k\delta_{x_2}, \alpha_2; A_m^1\right)$$

by induction hypothesis since A_m^1 is a nonincreasing discount sequence of horizon $m - 1$. The nondecreasing property of the third and fourth terms follows again from the induction hypothesis. □

We now prove the continuity of $\Delta(\alpha_1 + k\delta_x, \alpha_2; A_n)$ as a function of x.

PROPOSITION 2.2. *For all* $\alpha_1, \alpha_2, k > 0$ *and nonincreasing discount sequence* A_n, $\Delta(\alpha_1 + k\delta_x, \alpha_2; A_n)$ *is a continuous function of x.*

See the Appendix for the proof of Proposition 2.2.

The following proposition will be used to prove Theorems 2.1 and 2.2.

PROPOSITION 2.3. *Given* $\alpha_1, \alpha_2, k > 0$ *and any nonincreasing discount sequence* A_n *of horizon* $n (< \infty)$,

$$\lim_{x \to \infty} \Delta(\alpha_1 + k\delta_x, \alpha_2; A_n) = \infty, \qquad \lim_{x \to -\infty} \Delta(\alpha_1 + k\delta_x, \alpha_2; A_n) = -\infty.$$

PROOF. It is enough to show that, for any increasing sequence $\{x_m\}$ tending to ∞, $\Delta(\alpha_1 + k\delta_{x_m}, \alpha_2; A_n) \to \infty$ as $m \to \infty$. This result follows by induction. For $n = 1$,

$$\Delta\left(\alpha_1 + k\delta_{x_m}, \alpha_2; A_1\right) = a_1\left(\frac{M_1\mu_1 + kx_m}{M_1 + k} - \mu_2\right) \to \infty \quad \text{as } m \to \infty.$$

At the induction step, the result follows using (2.8) by proceeding as in Proposition 2.2. The proof of $\lim_{x\to-\infty}\Delta(\alpha_1+k\delta_x,\alpha_2;A_n)=-\infty$ follows similarly by considering a sequence $\{x_m\}$ decreasing to $-\infty$. □

The next theorem proves the existence of "break-even observation" $b(\alpha_1, \alpha_2;A_n)$.

THEOREM 2.1. *Given $\alpha_1,\alpha_2,n\geq 2$ and any nonincreasing discount sequence A_n of finite horizon n, there exists a break-even observation $b(\alpha_1,\alpha_2;A_n)$ such that*

$$\Delta\big(\alpha_1+\delta_x,\alpha_2;A_n^1\big)\geq\Delta(\alpha_1,\alpha_2;A_n)\quad \text{if } x\geq b(\alpha_1,\alpha_2;A_n),$$

and

$$\Delta\big(\alpha_1+\delta_x,\alpha_2;A_n^1\big)\leq\Delta(\alpha_1,\alpha_2;A_n)\quad \text{if } x\leq b(\alpha_1,\alpha_2;A_n).$$

PROOF. By Propositions 2.2. and 2.3 (with $k=1$), there exists, by the intermediate value property of continuous functions, a quantity b such that $\Delta(\alpha_1+\delta_b,\alpha_2;A_n^1)=\Delta(\alpha_1,\alpha_2;A_n)$. It is easy to see that this $b(\alpha_1,\alpha_2;A_n)$ satisfies the conditions of the theorem. □

If arm 1 is optimal initially, then $b(\alpha_1,\alpha_2;A_n)$ gives the stay-with-a-winner property with the value of b changing from $b(\alpha_1,\alpha_2;A_n)$ to $b(\alpha_1+\delta_x,\alpha_2;A_n^1)$ at the next stage. If the observation on arm 1 is greater than $b(\alpha_1,\alpha_2;A_n)$, then the advantage of choosing arm 1 over arm 2 is at least as much as it was at the previous stage. The next theorem proves the existence of another break-even observation $c(\alpha_1,\alpha_2;A_n)$ giving stronger properties of optimal strategies.

THEOREM 2.2. *Given $\alpha_1,\alpha_2,n\geq 2$ and any nonincreasing discount sequence A_n of finite horizon n, there exists a break-even observation $c(\alpha_1,\alpha_2;A_n)$ such that*

$$\Delta\big(\alpha_1+\delta_x,\alpha_2;A_n^1\big)\geq 0\quad \text{if } x\geq c(\alpha_1,\alpha_2;A_n),$$

and

$$\Delta\big(\alpha_1+\delta_x,\alpha_2;A_n^1\big)\leq 0\quad \text{if } x\leq c(\alpha_1,\alpha_2;A_n).$$

PROOF. The proof is similar to the proof of Theorem 2.1: there exists a c such that $\Delta(\alpha_1+\delta_c,\alpha_2;A_n^1)=0$. □

The observation $c(\alpha_1,\alpha_2;A_n)$ gives a stay-with-a-winner/switch-on-a-loser rule. If the observation x on arm 1 is greater than c, then arm 1 is optimal at the next stage, and arm 2 otherwise. Example 2.1 shows that such break-even observations do not exist in general.

EXAMPLE 2.1. Let

$$P_1=\begin{cases}\delta_8, & \text{w.p. } \frac12,\\ \frac12(\delta_0+\delta_{10}), & \text{w.p. } \frac12,\end{cases}\qquad P_2=\begin{cases}\delta_5, & \text{w.p. } \frac12,\\ \frac12(\delta_6+\delta_8), & \text{w.p. } \frac12.\end{cases}$$

TABLE 1
The quantity $\Delta_3 = \Delta(\alpha_1, \alpha_2; A_3)$, *where* $F_1 = U[0, 1]$, $F_2 = U[0.01, 1.01]$

	M_2				
M_1	0	1	5	10	100
0	−0.010	0.074	0.119	0.120	0.120
1	−0.093	−0.010	0.049	0.060	0.062
5	−0.129	−0.065	−0.010	0.005	0.019
10	−0.130	−0.071	−0.023	−0.010	0.005
100	−0.130	−0.071	−0.029	−0.019	−0.010

Let $\Delta(x, P_1, P_2; A_n^1)$ be the advantage of choosing arm 1 over arm 2 after the initial observation x on arm 1. Let $n = 2$ and $a_2 = 1$. Then

$$\Delta\left(8, P_1, P_2; A_2^1\right) = a_2(8 - 6) = 2$$

and

$$\Delta\left(10, P_1, P_2; A_2^1\right) = a_2(5 - 6) = -1.$$

We conclude this section with tables of $\Delta(\alpha_1, \alpha_2; A_n)$, $b(\alpha_1, \alpha_2; A_n)$ and $c(\alpha_1, \alpha_2; A_n)$, for $n = 3$ and for some choice of F_1 and F_2. In Table 1, $F_1 = \text{Uniform}[0, 1]$, $F_2 = \text{Uniform}[0.01, 1.01]$ and it gives values of $\Delta(\alpha_1, \alpha_2; A_3)$. Table 2 gives values of $b(\alpha_1, \alpha_2; A_3)$ and $c(\alpha_1, \alpha_2; A_3)$. In both tables, each of M_1 and M_2 takes values $0, 1, 5, 10$ and 100 and the discount sequence is $A_3 = (1, 1, 1, 0, 0, \ldots)$.

Computer programs to generate Tables 1 and 2 are written in FORTRAN. In Table 1, numerical integration is carried out using subroutine package under CMLIB. For Table 2, equations like $\Delta(\alpha_1 + \delta_b, \alpha_2; A_3^1) = \Delta(\alpha_1, \alpha_2; A_3)$ and $\Delta(\alpha_1 + \delta_c, \alpha_2; A_3^1) = 0$ are solved using bisection routines.

In Table 1, $\Delta(M_1F_1, M_2F_2; A_3)$ is an increasing function of M_2 for fixed F_1, F_2 and M_1. Again $\Delta(M_1F_1, M_2F_2; A_3)$ is a decreasing function of M_1 for fixed F_1, F_2 and M_2. So the less known about an arm, the more appealing it is, for there is more information to be gained by selecting it; $\Delta(M_1F_1, M_2F_2; A_3) = \mu_1 - \mu_2$, whenever $M_1 = M_2$, that is, when the "information value" of both the arms

TABLE 2
The quantities $(b(\alpha_1, \alpha_2; A_3), c(\alpha_1, \alpha_2; A_3))$, *where* $F_1 = U[0, 1]$, $F_2 = U[0.01, 1.01]$

	M_2				
M_1	0	1	5	10	100
0	(0.589, 0.596)	(0.607, 0.553)	(0.629, 0.524)	(0.630, 0.518)	(0.630, 0.511)
1	(0.536, 0.677)	(0.567, 0.582)	(0.599, 0.519)	(0.610, 0.496)	(0.615, 0.485)
5	(0.420, 1.075)	(0.434, 0.799)	(0.536, 0.588)	(0.564, 0.538)	(0.593, 0.495)
10	(0.426, 1.554)	(0.382, 1.081)	(0.464, 0.721)	(0.517, 0.625)	(0.580, 0.537)
100	(0.514, 10.174)	(−0.202, 5.842)	(−0.002, 2.954)	(0.194, 2.297)	(0.500, 1.507)

is the same. In Table 2, if $\Delta(\alpha_1, \alpha_2; A_n) > 0$, then $b(\alpha_1, \alpha_2; A_n) > c(\alpha_1, \alpha_2; A_n)$; otherwise, $b(\alpha_1, \alpha_2; A_n) \le c(\alpha_1, \alpha_2; A_n)$. We notice that if the support of α_1 is bounded above by U and below by L (in Table 2, $U = 1.0$ and $L = 0.0$), then $L \le c(\alpha_1, \alpha_2; A_n) \le U$ need not be true. For example, in Table 2, $c(\alpha_1, \alpha_2; A_3) = 10.174$ when $M_1 = 100, M_2 = 0$. The reason is that one arm may be so much better initially than the other that even the best possible observation would not make the other arm worth selecting at the next stage. After an initial optimal selection, optimal selections can be made at any subsequent stage by computing the c value at that stage for given outcomes prior to that stage. Hence optimal strategies can be completely determined. For example, consider the $(\alpha_1, \alpha_2; A_3)$ bandit where $F_1 = U[0, 1], F_2 = U[0.01, 1.01]$ and $A_3 = (1, 1, 1, 0, 0, \ldots)$. Let $M_1 = 1$ and $M_2 = 10$. From Table 1, $\Delta(\alpha_1, \alpha_2; A_3) = 0.06 > 0$. Hence arm 1 is optimal initially. So we select arm 1 and suppose the observation is $X_1 = 0.58$. This is bigger than $c(\alpha_1, \alpha_2; A_3) = 0.496$ (from Table 2). Hence arm 1 is optimal again at the next stage. Suppose the observation on arm 1 at this stage is $X_2 = 0.42$, which is smaller than $c(\alpha_1 + \delta_{0.58}, \alpha_2; A_3^1) = 0.45$. Hence arm 2 is optimal at the last stage. However, when n is large or the support of α_1 or α_2 is large, then finding c becomes extremely difficult computationally.

APPENDIX

Proof of Proposition 2.2 (By induction on n). For $n = 1$,

$$\Delta(\alpha_1 + k\delta_x, \alpha_2; A_1) = a_1\left(\frac{M_1\mu_1 + kx}{M_1 + k} - \mu_2\right)$$

is continuous in x. Assume the result holds for $n = r - 1$. It is enough to show that $\Delta(\alpha_1 + k\delta_{x_m}, \alpha_2; A_r) \to \Delta(\alpha_1 + k\delta_x, \alpha_2; A_r)$ for any increasing or decreasing sequence $\{x_m\}$ converging to x. By (2.8),

$$\begin{aligned}\Delta\big(\alpha_1 + k\delta_{x_m}, \alpha_2; A_r\big) &= (a_1 - a_2)\left(\frac{M_1\mu_1 + kx_m}{M_1 + k} - \mu_2\right)\\ &\quad + \frac{M_1}{M_1 + k}E\Big[\Delta^+\big(\alpha_1 + k\delta_{x_m} + \delta_Z, \alpha_2; A_r^1\big) \mid \alpha_1\Big]\\ &\quad + \frac{k}{M_1 + k}\Delta^+\big(\alpha_1 + (k+1)\delta_{x_m}, \alpha_2; A_r^1\big)\\ &\quad + E\Big[\Delta^-\big(\alpha_1 + k\delta_{x_m}, \alpha_2 + \delta_Y; A_r^1\big) \mid \alpha_2\Big].\end{aligned} \tag{A.1}$$

Let $\{x_m\}$ be an increasing sequence converging to x. For the sequence of nonnegative functions $f_m(z) = \Delta^+(\alpha_1 + k\delta_{x_m} + \delta_z, \alpha_2; A_r^1)$, we have the following:

(i) $0 \le f_1(z) \le f_2(z) \le \cdots$ (by Proposition 2.1);

(ii) $\Delta^+(\alpha_1 + k\delta_{x_m} + \delta_z, \alpha_2; A_r^1) \to \Delta^+(\alpha_1 + k\delta_x + \delta_z, \alpha_2; A_r^1)$ for every z by the induction hypothesis [using the continuity of $\Delta^+(\alpha_1 + \delta_z + k\delta_x, \alpha_2; A_r^1)$ as a function of x].

Using the monotone convergence theorem (MCT),

$$\text{(A.2)}\qquad \lim_{m\to\infty} E\left[\Delta^{+}\left(\alpha_1+k\delta_{x_m}+\delta_Z,\alpha_2;A_r^1\right)\mid\alpha_1\right]=E\left[\Delta^{+}\left(\alpha_1+k\delta_x+\delta_Z,\alpha_2;A_r^1\right)\mid\alpha_1\right].$$

By the induction hypothesis,

$$\text{(A.3)}\qquad \lim_{m\to\infty} \Delta^{+}\left(\alpha_1+(k+1)\delta_{x_m},\alpha_2;A_r^1\right)\to\Delta^{+}\left(\alpha_1+(k+1)\delta_x,\alpha_2;A_r^1\right).$$

Again note that $\Delta(\alpha_2+\delta_y,\alpha_1+k\delta_x;A_r^1)=-\Delta(\alpha_1+k\delta_x,\alpha_2+\delta_y;A_r^1)$ is continuous in x (by the induction hypothesis), establishing the continuity of $\Delta^{+}(\alpha_2+\delta_y,\alpha_1+k\delta_x;A_r^1)$ in x. For the sequence of functions $g_m(y)=\Delta^{+}(\alpha_2+\delta_y,\alpha_1+k\delta_{x_m};A_r^1)$, we have the following:

(i) $g_m(y)\to\Delta^{+}(\alpha_2+\delta_y,\alpha_1+k\delta_x;A_r^1)$, for every y as $m\to\infty$ (using continuity);
(ii) $|g_m(y)|\le|\Delta^{+}(\alpha_2+\delta_y,\alpha_1+k\delta_{x_1};A_r^1)|$ $\forall\ m$ (by Proposition 2.1).

[The measurability and integrability of $\Delta(\alpha_2+\delta_y,\alpha_1+k\delta_{x_m};A_r^1)$ for any m as a function of y follows from Theorem 18.3 of Billingsley (1979) and Theorem 2.5.1 of Berry and Fristedt (1985).]

Using the Lebesgue dominated convergence theorem (LDCT), we get

$$\text{(A.4)}\qquad \begin{aligned}&\lim_{m\to\infty} E\left[\Delta^{-}\left(\alpha_1+k\delta_{x_m},\alpha_2+\delta_Y;A_r^1\right)\mid\alpha_2\right]\\ &\quad=-\lim_{m\to\infty} E\left[\Delta^{+}\left(\alpha_2+\delta_Y,\alpha_1+k\delta_{x_m};A_r^1\right)\mid\alpha_2\right]\\ &\quad=-E\left[\Delta^{+}\left(\alpha_2+\delta_Y,\alpha_1+k\delta_x;A_r^1\right)\mid\alpha_2\right]\\ &\quad=E\left[\Delta^{-}\left(\alpha_1+k\delta_x,\alpha_2+\delta_Y;A_r^1\right)\mid\alpha_2\right].\end{aligned}$$

Using (A.2), (A.3) and (A.4) in (A.1), we get

$$\text{(A.5)}\qquad \Delta\left(\alpha_1+k\delta_{x_m},\alpha_2;A_r\right)\to\Delta(\alpha_1+k\delta_x,\alpha_2;A_r)\quad\text{as } m\to\infty,$$

for any increasing sequence $\{x_m\}$ converging to x. For any decreasing sequence $\{x_m\}$ converging to x, the proof is similar by using LDCT on $\{f_m(z)\}$ and MCT on $\{g_m(y)\}$. □

Acknowledgments. I wish to thank Professor D. A. Berry for his invaluable help and guidance during the preparation of my Ph.D. thesis, from which this paper is adapted. I also thank the referees for their useful comments and suggestions which have led to improvements of this paper.

REFERENCES

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

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

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

CLAYTON, M. K. and BERRY, D. A. (1985). Bayesian nonparametric bandits. *Ann. Statist.* **13** 1523–1534.

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

GALLUP ORGANIZATION
ONE CHURCH STREET, SUITE 900
ROCKVILLE, MARYLAND 20850