Bayesian Nonparametric Bandits

by

Murray K. Clayton*
and
Donald A. Berry**

Technical Report No. 427

* University of Guelph. Research partially supported by a University of Guelph Research Advisory Board Grant.

** University of Minnesota. Research partially supported by National Science Foundation Grant MCS8301450.

Abstract

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. A stay-with-a-winner rule is defined in this setting and shown to be optimal. A simple form of such a rule is expressed in terms of a degenerate Dirichlet process.

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. In this paper, we follow Bradt, Johnson, and Karlin (1956) and restrict consideration to the discrete time setting in which the objective is to maximize the expected sum of the first n observations . 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 probability measure (p.m.) P, the random variables $X_1, \ldots, X_n$ 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 information concerning Q, it is sufficient to assume that all the Y_i's are equal to the mean of Q, call it λ.

The p.m. P is unknown and, following the Bayesian approach, we take P to be random and assume that prior information regarding P can be expressed by its probability distribution. Much of the bandit literature assumes the arms to be Bernoulli (Bradt, et al. 1956, Berry 1972, 1984) in which case the support of P is contained in the set of those p.m.'s which concentrate

their mass on {0,1}. 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. The parameter α is a bounded non-null measure on the reals, R, with finite first moment. Let $M = \alpha(R)$ 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, p. 223). The prior mean, μ, for an observation from arm 1 is the mean of F. We shall frequently use MF to denote the parameter α.

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 $X_1, X_2, \ldots, X_n$, such that $X_1 = X_2 = \ldots = X_n$ almost surely and X_1 has distribution measure F. In such a situation, one pull of arm 1 yields complete information about P. In a sense described in detail by Sethuraman and Tiwari (1983), 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 $\alpha + \Sigma_1^k \delta_{X_i}$ 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 $\alpha + \Sigma_1^k \delta_{X_i}$ using Theorem 3 of Ferguson (1973). We shall denote this expectation by $E[g(X) \mid \alpha + \Sigma_1^k \delta_{X_i}]$, and shall delete the measure from the notation when appropriate. Note in particular that $E(X \mid \alpha) = \mu$.

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 $n \geq 2$ the following relations are evident:

$$W_n^1(\alpha,\lambda) = \mu + E[W_{n-1}(\alpha + \delta_X, \lambda) \mid \alpha]$$
$$W_n^2(\alpha,\lambda) = \lambda + W_{n-1}(\alpha,\lambda) \quad . \tag{1.1}$$

Together with the evident initial conditions $W_1^1(\alpha,\lambda) = \mu$ and $W_1^2(\alpha,\lambda) = \lambda$, equations (1.1) give a recursion for determining $W_n(\alpha,\lambda)$. In addition, repeated application of (1.1) gives all optimal strategies if one keeps track of whether the various W_{n-j} are equal to W_{n-j}^1 of W_{n-j}^2. Note that $W_{n-j}(\alpha + \Sigma_1^j \delta_{x_i}, \lambda)$ 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 the fact that "the integral of measurable functions is measurable"

(Billingsley 1979, Theorem 18.3). Integrability follows from

$$\lambda \vee \frac{M\mu + \Sigma_1^j \delta_{x_i}}{M + j} \leq W_{n-j}(\alpha + \Sigma_1^j \delta_{x_i}, \lambda)$$

$$\leq \frac{M}{M + j} E(X \vee \lambda | \alpha) + \frac{1}{M + j} \Sigma_i^j (x_i \vee \lambda) .$$

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 pulled at every stage, and at most that for the case in which P is known at the outset. Note that (1.1) holds for $\alpha = 0 \cdot F$, with the convention that parameter $0 \cdot F + \delta_x$ equals δ_x.

Some of the 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. In that section we also begin describing the properties of optimal strategies. In Section 3 we show that a "stay-with-a-winner" rule is optimal. In Section 4 we give several examples and some miscellaneous results.

The problem described here is a finite horizon two-armed bandit with one arm known. A straightforward generalization of Theorem 2.1 of Berry and Fristedt (1979) shows that in fact we have described an "optimal stopping problem." That is, we need not consider strategies which follow a pull of arm 2 by a pull of arm 1, and so we need only determine the stage at which arm 2 is

first pulled, if ever. Problems of this sort are referred to as "one-armed bandits" (Berry 1984). If $\lambda = 0$, this description is especially fitting, since a pull of arm 2 in that case has no effect on the sum. Consequently, we could consider the problem with $\lambda = 0$ to be one in which at most n observations are to be taken from a population, such that the actual number of observations taken is decided upon sequentially and the goal is to maximize the expected total of the observations taken.

As mentioned, the majority of the literature on bandit problems deals with Bernoulli bandits. The only other model discussed in detail in the literature is for normally distributed observations. Recent work includes that of Fahrenholz (1982) and a continuous time version using Wiener processes (Chernoff, 1968).

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

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" of Bradt, et al. (1956, Lemma 4.2) and Berry and Fristedt (1979, Theorem 2.2). As indicated by Berry and Fristedt, their result can be generalized to include the current model. 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 "pull arm 1 if $\lambda \leq \Lambda_n(\alpha)$" and "pull arm 2 if $\lambda \geq \Lambda_n(\alpha)$."

Optimal strategies are completely determined by Λ_n. If $\lambda < \Lambda_n(\alpha)$, then arm 1 is uniquely optimal initially; if $\lambda > \Lambda_n$ then arm 2 is uniquely optimal; and if $\lambda = \Lambda_n$ then both arms are optimal initially. At the second stage we compare λ to $\Lambda_{n-1}(\alpha)$

or $\Lambda_{n-1}(\alpha + \delta_{x_1})$ accordingly as arm 2 was pulled initially or arm 1 was pulled and $X_1 = x_1$ was observed; and so on for subsequent stages. 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.

An easy consequence of this result gives some flavor of bandit problems more generally when comparing a known with an unknown arm. Namely, if $E(X|\alpha) \geq \lambda$ then arm 1 is optimal initially. This is easy to see by comparing pulls of arm 1 exclusively with pulls of arm 2. Moreover, if $E(X|\alpha) = \lambda$, $n \geq 2$, F is not concentrated on one point, and $M < \infty$, or if $E(X|\alpha) > \lambda$, then arm 1 is uniquely optimal initially.

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 if $\lambda < \Lambda_n(\alpha)$ then $W_n(\alpha,\lambda) > n\lambda$. This give a characterization of $\Lambda_n(\alpha)$ which lets us easily 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. We 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 <u>to the right of</u> F if $F'(x) \leq F(x)$, $x \in R$. If X has distribution function F and if X' has distribution function F', then we say X' is <u>stochastically larger</u> than X.

The following lemma appears as 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. It is necessary to first to prove a special case.

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

Remark: Let $z < z'$. Note that the distribution function form of $MF + k\delta_{z'}$ is $(MF + k\delta_{z'})/(M + k)$, and this is to the right of the distribution function form of $MF + k\delta_z$.

Proof: By induction. Let $z < z'$.

For the case $n = 1$ we have

$$W_1(MF + k\delta_z, \lambda) = \frac{M\mu + kz}{M + k} \vee \lambda$$
$$\leq \frac{M\mu + kz'}{M + k} \vee \lambda$$
$$= W_1(MF + k\delta_{z'}, \lambda).$$

For the induction step, suppose the result is true for $n = m - 1$. By (1.1), the induction hypothesis easily gives

$$W_m^2(MF + k\delta_z, \lambda) \leq W_m^2(MF + k\delta_{z'}, \lambda)$$

while, also by (1.1),

$$W_m^1(MF + k\delta_z, \lambda) = \frac{M\mu + kz}{M + k}$$
$$+ \frac{M}{M + k} E[W_{m-1}(MF + k\delta_z + \delta_X, \lambda) | F]$$
$$+ \frac{k}{M + k} W_{m-1}(MF + (k + 1)\delta_z, \lambda)$$
$$\leq W_m^1(MF + k\delta_{z'}, \lambda),$$

by the induction hypothesis.

Since $W_m = W_m^1 \vee W_m^2$, we have $W_m(MF + k\delta_z, \lambda) \leq W_m(MF + k\delta_{z'}, \lambda)$ and the proposition follows. □

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 $\Lambda_n(MF + \delta_z) \leq \Lambda_n(MF + \delta_{z'})$. □

Remark: Corollary 2.1 says that, given the observation $X_1 = z$, our inclination to pull arm 1 increases with increased z.

We next prove a more general result.

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. The case $n = 1$ follows by Lemma 2.2. For the induction step assume the proposition holds for $n = m - 1$. Then, by (1.1), we immediately have

$$W_m^2(MF,\lambda) \leq W_m^2(MF',\lambda).$$

Also 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 the distribution function form of $MF' + \delta_x$ is to the right of the distribution function form of $MF + \delta_x$. □

Corollary 2.2: For all $n \geq 1$ and $M \geq 0$, $\Lambda_n(MF) \leq \Lambda_n(MF')$ 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. Stay-with-a-winner rules.

In this section we prove a stay-with-a-winner rule. This was proved 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 (1984) for other references. For the Bernoulli bandit such a rule says that if it is optimal to pull arm 1 initially, and if a success is obtained, then it is optimal to pull arm 1 again. For the Dirichlet bandit, the stay-with-a-winner rule has the form: "If arm 1 is optimal initially, and if the resulting observation is sufficiently large, then it is optimal to pull arm

1 again."

Proposition 3.1: Given α and $n \geq 2$, there exist points x' and x'' such that $\wedge_{n-1}(\alpha + \delta_{x'}) \leq \wedge_n(\alpha) \leq \wedge_{n-1}(\alpha + \delta_{x''})$. 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: x'' exists since

$$\lim_{x \to \infty} \wedge_n(\alpha + \delta_x) \geq \lim_{x \to \infty} \wedge_1(\alpha + \delta_x) = \infty.$$

The inequality follows since $\wedge_n$ is nondecreasing in n, the equality follows since $\wedge_1(\alpha + \delta_x) = (M\mu + x)/(M + 1)$. The existence of x'' follows from the fact that

$$\lim_{x \to -\infty} \wedge_n(\alpha + \delta_x) = -\infty.$$

This can be proved using 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 $x'' = U$ works we adapt the proof of Theorem 4.1 in Berry and Fristedt (1979). Namely, we suppose $\lambda \leq \wedge_n(\alpha)$, and show $\lambda \leq \wedge_{n-1}(\alpha + \delta_U)$. We have two cases: (i) $\lambda < E(X|\alpha + \delta_U)$ and (ii) $\lambda \geq E(X|\alpha + \delta_U)$. In case (i), $\lambda < E(X|\alpha + \delta_U) = \wedge_1(\alpha + \delta_U)$ $\leq \wedge_{n-1}(\alpha + \delta_U)$ since $\wedge$ 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, irrespective of the outcome of the initial pull of arm 1, it is optimal by Theorem 2.1 to pull arm 2 on the remaining pulls. Consequently,

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

the worth of pulling arm 1 n times.

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

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

The first inequality in (3.1) 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 (3.1) follows since Λ is nondecreasing in n. □

The above result gives one form of a stay-with-a-winner rule. Namely, if arm 1 is optimal initially, and if $X_1 = x''$ is observed, then arm 1 is optimal again. A less stringent version of the stay-with-a-winner rule exists. We show this by showing that $\Lambda_n(\alpha + \delta_x)$ is continuous in x. This follows from the next lemma.

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) $W_n(\alpha + \Sigma_1^k \delta_{x_i} + \delta_z, \lambda)$ is jointly continuous in z and λ;

(ii) For $j = 1, 2$, $W_n^j(\alpha + \delta_z, \lambda)$ is jointly continuous in z and λ.

Proof: We prove part (i) by induction. Fix z_0, λ_0 and let $z' < z_0 < z''$, $\lambda' < \lambda_0 < \lambda''$. We show continuity at z_0 and λ_0. Note first that, for $n \geq 1$, for $z \in (z', z'')$ and $\lambda \in (\lambda', \lambda'')$,

$$(3.2) \quad W_n(\alpha + \Sigma_1^k \delta_{x_i} + \delta_{z'}, \lambda') \leq W_n(\alpha + \Sigma_1^k \delta_{x_i} + \delta_z, \lambda) \leq W_n(\alpha + \Sigma_1^k \delta_{x_i} + \delta_{z''}, \lambda'')$$

by Proposition 2.1 and the fact that $W_n(\cdot, \lambda)$ is nondecreasing in λ.

The case $n = 1$ being easy, suppose part (i) of Lemma 3.1 holds for $n = m - 1$. By (1.1), (3.2), and Lebesgue's dominated convergence theorem,

$$(3.3) \quad \lim_{(z,\lambda)\to(z_0,\lambda_0)} W_m^1(\alpha + \Sigma_1^k \delta_{x_i} + \delta_z, \lambda) = \frac{M\mu + \Sigma_1^k \delta_{x_i} + z_0}{M + k + 1} + \frac{M}{M + k + 1} E\Big[\lim_{(z,\lambda)\to(z_0,\lambda_0)} W_{m-1}(\alpha + \Sigma_1^k \delta_{x_i} + \delta_z + \delta_X, \lambda) \,|\, F\Big] + \frac{1}{M + k + 1} \cdot \Sigma_{j=1}^1 \lim_{(z,\lambda)\to(z_0,\lambda_0)} W_{m-1}^k(\alpha + \Sigma_{i=1}^k \delta_{x_i} + \delta_z + \delta_{x_j}, \lambda)$$

$$+ \frac{1}{M + k + 1} \lim_{(z,\lambda)\to(z_0,\lambda_0)} W_{m-1}(\alpha + \Sigma_{i=1}^{k}\delta_{x_i} + 2\delta_z, \lambda)$$

$$= W_m^1(\alpha + \Sigma_1^k \delta_{x_i} + \delta_{z_0}, \lambda_0).$$

The last equality follows by the induction hypothesis. As well, the induction hypothesis and (1.1) immediately give

$$\lim_{(z,\lambda)\to(z_0,\lambda_0)} W_m^2(\alpha + \Sigma_{i=1}^{k}\delta_{x_i} + \delta_z, \lambda) \tag{3.4}$$

$$= W_m^2(\alpha + \Sigma_{i=1}^{k}\delta_{x_i} + \delta_{z_0}, \lambda_0) .$$

Part (i) now follows since $W_m = W_m^1 \vee W_m^2$.

Part (ii) follows from part (i) and equations (3.3) and (3.4) if $n \geq 2$. The case $n = 1$ is easy. □

Hence,

$$0 = \lim_{x\to x_0} [W_n^1(\alpha + \delta_x, \wedge_n(\alpha + \delta_x)) - W_n^2(\alpha + \delta_x, \wedge_n(\alpha + \delta_x))]$$

$$= W_n^1(\alpha + \delta_{x_0}, \lim_{x\to x_0} \wedge_n(\alpha + \delta_x))$$

$$- W_n^2(\alpha + \delta_{x_0}, \lim_{x\to x_0} \wedge_n(\alpha + \delta_x)).$$

The second equality follows from Lemma 3.1 (ii). By uniqueness,

$$\Lambda_n(\alpha + \delta_{x_0}) = \lim_{x \to x_0} \Lambda_n(\alpha + \delta_x). \quad \square$$

Theorem 3.1: Given α and $n \geq 2$, there exists a unique $b = b_n(\alpha)$ such that

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

Proof: Existence follows from Proposition 3.1 and the fact that $\Lambda_{n-1}(\alpha + \delta_x)$ is continuous (Lemma 3.2) and nondecreasing in x (Corollary 2.1).

We show b is unique by contradiction. Suppose $b < b'$ and suppose

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

Then since an initial pull of arm 2 is optimal,

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

However, since an initial pull of arm 1 is also optimal, by (1.1) we have

$$\begin{aligned} W_{n-1}(\alpha + \delta_{b'}, \lambda) - W_{n-1}(\alpha + \delta_b, \lambda) &= (b' - b)/(M + 1) \\ &\quad + EW_{n-2}(\alpha + \delta_{b'} + \delta_X, \lambda \mid \alpha + \delta_{b'}) \\ &\quad - EW_{n-2}(\alpha + \delta_b + \delta_X, \lambda \mid \alpha + \delta_b) \\ &> 0. \quad \square \end{aligned}$$

Remark: A proof similar to that of Theorem 3.1 shows that there

exists a $c = c_n(\alpha)$ such that $\lambda = \Lambda_{n-1}(\alpha + \delta_c)$. Therefore, given a pull (optimal or not) of arm 1 resulting in $X_1 = x$, it is uniquely optimal in the second stage to pull arm 1 if $x > c$ and arm 2 if $x < c$; if $x = c$ either arm is optimal.

The 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 determining x'' unless the support of α is bounded above by U, in which case $x'' = U$.

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

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

and so, by Lemma 2.1, $\Lambda_n(0,F)$ is the smallest λ to satisfy

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

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

$$\Lambda_n(0,F) = \mu + (n-1)E(X - \Lambda_n(0,F))^+. \tag{3.5}$$

To show $b_n(\alpha) \leq \Lambda_n(0,F)$, 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 3.3: For all γ, for all $k > 0$, for all α, for $n \geq 1$, and for all λ, if $x \geq \gamma$, then

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

Proof: 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:

(i) $\lambda \geq \Lambda_m(\alpha + k\delta_x) \geq \Lambda_m(\alpha + k\delta_\gamma)$

in which case

$$D_m = \frac{mk(x - \gamma)}{M + k} + m\lambda - m\lambda \geq 0,$$

and

(ii) $\Lambda_m(\alpha + k\delta_x) \geq \lambda$.

Here, $W_m(\alpha + k\delta_x, \lambda) = W_m^1(\alpha + k\delta_x, \lambda)$, and it suffices to prove

$$\frac{mk(x - \gamma)}{M + k} + W_m^1(\alpha + k\delta_\gamma, \lambda) - W_m^1(\alpha + k\delta_x, \lambda) \geq 0 \tag{3.6}$$

since the left hand side of (3.6) is a lower bound for D_m. But the left hand side of (3.6) is

$$(3.7)\qquad \frac{mk(x-\gamma)}{M+k} + W_m^1(\alpha + k\delta_\gamma, \lambda) - W_m^1(\alpha + k\delta_x, \lambda)$$

$$= \frac{mk(x-\gamma)}{M+k} + \frac{M\mu + k\gamma}{M+k} + E[W_{m-1}(\alpha + k\delta_\gamma + \delta_X, \lambda) \mid \alpha + k\delta_\gamma]$$

$$- \frac{M\mu + kx}{M+k} - E[W_{m-1}(\alpha + k\delta_x + \delta_X, \lambda) \mid \alpha + k\delta_x]$$

$$= \frac{mk(x-\gamma)}{M+k} + \frac{k(\gamma - x)}{M+k}$$

$$+ \frac{M}{M+k} E[W_{m-1}(\alpha + k\delta_\gamma + \delta_X, \lambda) \mid]$$

$$+ \frac{k}{M+k} W_{m-1}(\alpha + (k+1)\delta_\gamma, \lambda)$$

$$- \frac{M}{M+k} E[W_{m-1}(\alpha + k\delta_x + \delta_X, \lambda) \mid \alpha]$$

$$+ \frac{k}{M+k} W_{m-1}(\alpha + (k+1)\delta_x, \lambda) .$$

Some manipulation shows this to equal

$$(3.8)\qquad \frac{M}{M+k} E\{[\frac{(m-1)k(x-\gamma)}{M+k+1} + W_{m-1}(\alpha + \delta_X + k\delta_\gamma, \lambda)$$

$$- W_{m-1}(\alpha + \delta_X + k\delta_x, \lambda)]\}$$

$$+ \frac{k}{M+k} [\frac{(m-1)(k+1)(x-\gamma)}{M+k+1} + W_{m-1}(\alpha + (k+1)\delta_\gamma, \lambda)$$

$$- W_{m-1}(\alpha + (k+1)\delta_x, \lambda)].$$

But the quantities in square brackets in (3.8) are nonnegative by the induction hypothesis. □

Suppose we are given a choice between n pulls of a bandit with Dirichlet process parameter α, or a single observation of known value and n - 1 pulls of a bandit with Dirichlet process

parameter $\alpha + \delta_z$. The latter is preferred if this observation and z are large enough.

Lemma 3.4: For $n \geq 2$, for all α = MF, and for all λ,

$$(3.9) \qquad W_n(\alpha,\lambda) \leq [\lambda \vee \wedge_{n-1}(\alpha + \delta_{\wedge_n^0})] + W_{n-1}(\alpha + \delta_{\wedge_n^0},\lambda)$$

where $\wedge_n^0 = \wedge_n(0,F)$.

Proof: There are two cases.

Case (i): $\lambda > \wedge_n(\alpha)$. In this use $W_n(\alpha,\lambda) = n\lambda$, and (3.9) holds immediately.

Case (ii): $\lambda \leq \wedge_n(\alpha)$. In this case we must prove

$$(3.10) \qquad W_n^1(\alpha,\lambda) \leq (\lambda \vee \wedge_{n-1}(\alpha + \delta_{\wedge_n^0})) + W_{n-1}(\alpha + \delta_{\wedge_n^0},\lambda).$$

Since $\wedge_1(\alpha + \delta_{\wedge_n^0}) \leq \wedge_{n-1}(\alpha + \delta_{\wedge_n^0})$, to prove (3.10) it will suffice to prove

$$(3.11) \qquad W_n^1(\alpha,\lambda) \leq \wedge_1(\alpha + \delta_{\wedge_n^0}) + W_{n-1}(\alpha + \delta_{\wedge_n^0},\lambda).$$

Now by (3.5), $\wedge_1(\alpha + \delta_{\wedge_n^0}) = \dfrac{M\mu + \wedge_n^0}{M + 1}$

$$= \mu + (n-1)E(X-\wedge_n^0)^+/(M + 1).$$

Also by (1.1), $W_n^1(\alpha,\lambda) = \mu + EW_{n-1}(\alpha + \delta_X,\lambda)$.

Therefore (3.11) is equivalent to

$$(3.12)\qquad EW_{n-1}(\alpha + \delta_X, \lambda) \le \frac{(n-1)E(X-\Lambda_n^0)^+}{M+1} + W_{n-1}(\alpha + \delta_{\Lambda_n^0}, \lambda),$$

and (3.12) is equivalent to $E[f(X)] \ge 0$, 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 $x \le \Lambda_n^0$, $f(x) \ge 0$ by Proposition 2.1, while if $x > \Lambda_n^0$, $f(x) \ge 0$ by Lemma 3.3. □

Theorem 3.2: For all α and for $n \ge 2$, $b_n(\alpha) \le \Lambda_n(0,F)$.

Proof: Let $\Lambda_n^0 = \Lambda_n(0,F)$ and take $\lambda = \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0})$ in (3.9):

$$\begin{aligned} W_n(\alpha, \Lambda_n(\alpha + \delta_{\Lambda_n^0})) &\le \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0}) \\ &\quad + W_{n-1}(\alpha + \delta_{\Lambda_n^0}, \Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0}) \\ &= n\Lambda_{n-1}(\alpha + \delta_{\Lambda_n^0}) \end{aligned}$$

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

Theorem 3.2 gives an easily described stay-with-a-winner rule: if arm 1 is optimal initially, and if $X_1 \ge \Lambda_n(0,F)$ is observed, then arm 1 is again optimal.

Example 3.1: The Bernoulli model with a beta prior. If $\alpha = M(q\delta_0 + p\delta_1)$, $0 < p < 1$, $p + q = 1$, then $\Lambda_n[0; (q\delta_0 + p\delta_1)] = np/((n-1)p + 1)$. Since $\Lambda_n^0 \in (0,1)$ in this case, Theorem 3.2 gives the traditional stay-with-a-winner rule for the Bernoulli bandit. □

By Corollary 2.1 and Theorem 3.1, the Dirichlet process bandit is "monotone": if arm 1 is initially optimal and if $X_1 = x$ is observed, then there is a unique b such that arm 1 is again optimal for any $x \geq b$. The next example shows that such monotonicity need not hold if P is not a Dirichlet process.

Example 3.2 (Non-Dirichlet): Suppose that $P = \delta_1$ with probability 1/2 and $P = (1/2)(\delta_0 + \delta_{10})$ with probability 1/2. Then, using an obvious extension of notation, $\Lambda_2 = 3$. Assuming $1 < \lambda < 3$, arm 1 is optimal initially and optimal for the second pull if $X_1 = 0$ or 10, but not if $X_1 = 1$. □

4. Examples and Comments

In this section we present some examples and suggest some easy to use but suboptimal strategies.

Let U 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 values of $\Lambda_n(MF)$ and $b_n(MF)$ are given for $n = 2, 3, 4$, $F = U$ and Φ, and $M = 0, .1,$

.5, 1, 5, 10, 100.

It is not hard to show that $b_2(MF) = \Lambda_2(0,F)$ for all $M \geq 0$ and F, and that $b_n(0,F) = \Lambda_n(0,F)$ for $n \geq 2$ and all F. These facts are reflected in Table 4.1. As well, it is straightforward to prove

$$(4.1) \qquad \Lambda_n(0,F) \geq \Lambda_n(MF) \geq \mu = \lim_{M\to\infty} \Lambda_n(MF)$$

for $n \geq 1$, for all F, and for $M \geq 0$. A stronger version of (4.1) can be given in the case $n = 2$: 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. Roughly speaking, this reflects the intuitive notion that the less known about arm 1, the more promising is a pull on it.

One difficulty with the use of the Dirichlet process in modeling is that the calculation of quantities like Λ_n and b_n can involve numerical multiple integration, a process which is typically expensive. (Of course, this problem exists when any other model for continuous data is used.) To deal with this, suppose that F has compact support S. Let $F_1, F_2, \ldots$ be a sequence of distribution functions whose supports are contained in S and which converge to F. It is possible to show that $\Lambda_n(MF_k)$ and $b_n(MF_k)$ converge to $\Lambda_n(MF)$ and $b_n(MF)$, respectively, as $k\to\infty$ (cf. Christensen (1983)). This suggests a strategy which

may be quite good: choose a discrete F_k sufficiently close to F and act as if α were MF_k instead of MF. Some preliminary work suggests that this is particularly useful when M is large. We conjecture that this approach will give good results even when the support of F is unbounded, if F_k is chosen appropriately. The advantage of proceeding in this manner is that expectations can be computed easily as sums instead of integrals.

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	.276	.276	.436	.436	.549	.549
.1	.251	.276	.400	.424	.505	.529
.5	.184	.276	.300	.388	.383	.529
1	.138	.276	.228	.359	.295	.421
5	.046	.276	.079	.276	.105	.284
10	.028	.276	.043	.248	.058	.238
100	.003	.276	.005	.214	.007	.182

b. $F = U$

M	Λ_2	b_2	Λ_3	b_3	Λ_4	b_4
0	.586	.586	.634	.634	.667	.667
.1	.578	.586	.623	.630	.654	.661
.5	.557	.586	.592	.619	.617	.644
1	.543	.586	.570	.610	.590	.630
5	.514	.586	.524	.584	.532	.589
10	.508	.586	.513	.576	.518	.574
100	.501	.586	.501	.565	.502	.554

REFERENCES

Billingsley, P. (1979). Probability and Measure. John Wiley and Sons, New York.

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

Berry, D. A. (1984). "One- and two-armed bandit problems" in: Encyclopedia of Statistical Sciences, Vol. 5 (Kotz, S., and Johnson, N. L., eds.), John Wiley and Sons, New York.

Berry, D. A. and Fristedt, B. (1979). Bernoulli one-armed bandits--arbitrary discount sequences. Ann. Statist. 7 1086-1105.

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. Sankhya A 30 221-252.

Christensen, R. (1983). Searching for the lowest price when the unknown distribution of prices is modeled with a Dirichlet process. Unpublished thesis, University of Minnesota.

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 vol. 2 (Gupta, S. S. and Berger, J. O., eds.), Academic Press, New York.