The Annals of Statistics
1992, Vol. 20, No. 3, 1625–1636

THE LEARNING COMPONENT OF DYNAMIC ALLOCATION INDICES

BY JOHN GITTINS AND YOU-GAN WANG

University of Oxford

For a multiarmed bandit problem with exponential discounting the optimal allocation rule is defined by a dynamic allocation index defined for each arm on its space. The index for an arm is equal to the expected immediate reward from the arm, with an upward adjustment reflecting any uncertainty about the prospects of obtaining rewards from the arm, and the possibilities of resolving those uncertainties by selecting that arm. Thus the learning component of the index is defined to be the difference between the index and the expected immediate reward. For two arms with the same expected immediate reward the learning component should be larger for the arm for which the reward rate is more uncertain. This is shown to be true for arms based on independent samples from a fixed distribution with an unknown parameter in the cases of Bernoulli and normal distributions, and similar results are obtained in other cases.

1. Introduction. The classical multiarmed bandit problem models sequential clinical trials of alternative treatments as Bernoulli processes, or *arms*, with different unknown success probabilities. The arms are *pulled* one at a time and the aim is to achieve a high overall success rate by at each stage selecting the next arm to be pulled on the basis of current information about the different success probabilities. The problem will be discussed in the Bayesian discounted setting introduced by Bellman (1956), and since then considered by several authors, in particular by Berry and Fristedt (1985), who give an extensive review. Our investigation extends to arms defined by sequences of independent normal and negative exponentially distributed random variables with unknown parameter values. The main application to date of these more general bandit problems has been to the selection of compounds for testing in research leading to the formulation of new drugs [see Bergman and Gittins (1985) and Gittins (1989), which we will refer to as G].

In selecting the arm to be pulled next there are two objectives: to obtain a high reward from this pull and to acquire information which may be used to make it more likely that subsequent pulls are on the whole profitable. It is the tension between these two objectives which makes multiarmed bandit problems interesting. Here we investigate the relationship between the importance of acquiring further information and the amount of information which is already available.

Received July 1990; revised October 1991.
AMS 1980 *subject classifications*. 62C10, 90C40, 93E20.
Key words and phrases. Dynamic allocation index, Gittins index, multiarmed bandit, target processes.

For a multiarmed bandit problem with exponentially discounted rewards (geometrically discounted rewards for a discrete-time problem), the optimal strategy is to select the arm for which a certain index, which depends on the state of the arm, is the largest (Gittins and Jones, 1974). The properties of these dynamic allocation (or Gittins) indices have been investigated by several author and are described in G.

In this paper we shall be interested in bandit sampling processes (G, Chapters 6 and 7). A *bandit sampling process* is a sequence of independent random variables $X_1, X_2, \ldots$, which are identically distributed with an unknown distribution belonging to a parametric family of distributions $\mathscr{D}$. The prior distribution Π for the parameter θ of $\mathscr{D}$ becomes transformed into successive posterior distributions in the standard Bayesian fashion as observations take place [e.g., see Barra (1981)]. The current distribution for θ may be regarded as the current state of the bandit sampling process. When Π belongs to a parametric family of distributions which is closed under sampling the state of the process is defined by the parameter values for the current distribution for θ.

Each observation on a bandit sampling process is assumed to take one unit of time, and to yield a reward which depends on the observed value. The reward at time t ($t = 0, 1, 2, \ldots$) is multiplied by the discount factor a^t, where $0 \le a \le 1$. For an initial state Π and stopping time defined in terms of the filtration $\{\mathscr{F}_t\colon t \ge 1\}$, where $\mathscr{F}_t$ denotes the σ field $\sigma\{X_i\colon i \le t\}$, let

$$(1)\quad R_\tau(\Pi) = E\sum_{t=0}^{\tau-1} a^t r(X_t), \qquad W_\tau(\Pi) = E\sum_{t=0}^{\tau-1} a^t, \qquad \nu_\tau(\Pi) = \frac{R_\tau(\Pi)}{W_\tau(\Pi)},$$

where $r(X_t)$ is the expected reward associated with the observation X_t. Write also

$$(2)\qquad \nu(\Pi) = \sup_\tau \nu_\tau(\Pi), \qquad \nu^T(\Pi) = \sup_{\tau \le T} \nu_\tau(\Pi),$$

for any fixed $T \ge 1$, the second supremum being over stopping times τ for which $P(\tau \le T) = 1$. Thus $\nu(\Pi)$ is the maximum average reward rate up to a stopping time. It follows that if the alternative to continued sampling is to receive rewards at a constant rate λ, and these rewards are also discounted by the factor a^t at time t, then it is optimal to sample if and only if $\nu(\Pi) \ge \lambda$.

A family of alternative sampling processes each with the same discount constant a, and exactly one of which must be sampled at each time, depending on the current states of the alternative processes, forms a *multiarmed bandit*. The object is to sample so as to maximize the expected total reward.

Since a given reward is more valuable the earlier that it occurs, because of the discounting, a first guess at an optimal policy might be always to sample from the process yielding the largest immediate expected reward, which is to say from the process with the largest current value of $\nu^1(\Pi)$. This policy is often called the *myopic* policy, because of its lack of long-range vision. It ignores the contribution to the total reward which may arise from a reduction of uncertainty about the corresponding parameter θ. Theorem 3.6 of G shows

that the optimal policy actually is always to sample from the process with the largest current value of $\nu(\Pi)$, which is therefore defined to be the *dynamic allocation index* for the process in state Π. We define the difference $\nu(\Pi) - \nu^1(\Pi)$ to be the *learning component* of the index. Theorem 4.1 of G shows that the learning component is an increasing function of the discount factor a. This is not surprising as the larger the value of a the more important are the results of future sampling, and hence the greater is the value of gathering information which enables future sampling to be done more effectively. The main contribution of the present paper is to show that for two important cases the learning component is a decreasing function of the amount of information already available about the parameter θ. Again, this is as might have been expected.

Part (i) of the following theorem is Theorem 3.4(iii) of G; part (ii) is a simple extension of this result.

THEOREM 1. (i) $\nu_\tau(\Pi) = \nu(\Pi)$ *if* $\tau = \inf\{t\colon \nu(\Pi_t) < \nu_\tau(\Pi)\}$, *where* Π_t *is the state of the process after* t *observations from the state* Π, *and* $\tau = \infty$ *if* $\nu(\Pi_t) \geq \nu(\Pi)$ *for all* t.

(ii) $\nu_{\tau(T)}(\Pi) = \nu^T(\Pi)$ *if* $\tau(T) = \min[T, \min\{t\colon \nu^{T-t}(\Pi_t) < \nu^T(\Pi)\}]$.

[*Thus* $\tau(T) \leq T$ *and is such that the maximum expected reward rate which can be achieved between* $\tau(T)$ *and* T *is less than* $\nu^T(\Pi)$.]

In this paper, two particular types of sampling process will be considered. For a *reward process* the reward received if X_{t+1} is observed at time t is $a^t X_{t+1}$. For a *target process* the discount factor is 1, and the reward is equal to 1 on the first occasion when X_{t+1} exceeds some threshold or target value L, and is otherwise equal to 0. For a target process there is an additional state C, the *completion state*, signifying that the target has been reached, and so no further reward is possible. With a family of alternative target processes the aim is to minimize the expected number of observations until the target is first reached. This objective turns out (G, Section 6.2) to be achieved by using dynamic allocation indices defined in terms of the reward structure just described.

For a Bernoulli process observations are independently equal to 1 or 0 with probability θ and $1 - \theta$. The conjugate prior distributions for θ are beta distributions, with probability density functions of the form

$$\frac{\Gamma(n)}{\Gamma(\alpha)\Gamma(n-\alpha)}\theta^{\alpha-1}(1-\theta)^{n-\alpha-1}, \qquad 0 \leq \theta \leq 1,$$

where α and $n - \alpha$ are both positive. The state of the process is given by the parameters (α, n), so that, for example, the dynamic allocation index $\nu(\Pi)$ may be written as $\nu(\alpha, n)$. Each observation increases n by 1 and increases α by the value of the observation (either 1 or 0). The expected value of the next observation in the state (α, n) is α/n, the expected value of θ, and for a reward process this is also equal to the expected value $\nu^1(\alpha, n)$ of the next

reward. Our first result (in Section 2) is that for a Bernoulli reward process the learning component $\nu(\lambda n, n) - \nu^1(\lambda n, n)$ of the index is a decreasing function of n for any fixed positive λ. This is as expected, because the uncertainty about θ, and hence the probability of useful further information about θ, both decrease as n increases.

Our second main result (in Section 3) is for sampling processes drawn from continuous distributions. This theorem is applied to normally and negative exponentially distributed reward and target processes. For the normal reward process the conclusion once again is that the learning component of the index decreases as the number of observed values increases. In the other case the results are similar, but cannot be interpreted in precisely this way.

2. Bernoulli reward processes. This is the classical Bayesian multi-armed bandit problem. From the results given by Bellman (1956) it quickly follows that

$$\nu(\alpha, n+1) < \nu(\alpha, n) < \nu(\alpha+1, n+1),$$

although his paper pre-dates the idea of using ν as a dynamic allocation index. These inequalities lead to a play-the-winner rule, since an optimal switch between arms can only occur when a failure (i.e., a 0) is observed.

LEMMA 1. *$\nu^1(\alpha, n) \le \nu^1(\alpha', n')$ and $\nu^T(\alpha, n) \le \nu^T(\alpha', n')$ for any integer $T \ge 2$, if $\alpha/n \le \alpha'/n'$, $n' \le n$, and at least one of these inequalities is strict.*

PROOF. Let $\{X_t\colon t = 1, 2, \dots\}$ be the observations starting from the state (α, n) and $\tau(T) = \min[T, \min\{t\colon \nu^{T-t}(\alpha + S_t, n+t) < \nu^T(\alpha, n)\}]$. where $S_t = \sum_{i \le t} X_i$, so that $\nu_{\tau(T)}(\alpha, n) = \nu^T(\alpha, n)$ [Theorem 1(ii)]. From the definitions (1) and (2) the following results for $T = 1$ and 2 are easily obtained.

$$T = 1: \qquad \nu^1(\alpha, n) = \alpha/n \le \alpha'/n' = \nu^1(\alpha', n'),$$

$$T = 2: \qquad \tau(2) = \begin{cases} 1, & \text{if } X_1 = 0, \\ 2, & \text{if } X_1 = 1, \end{cases}$$

so that

$$\nu^2(\alpha, n) = \frac{1 + a(\alpha+1)/(n+1)}{1 + a\alpha/n} \frac{\alpha}{n}.$$

It follows that $\nu^T(\alpha, n) < \nu^T(\alpha', n')$ when $T = 2$.

We will prove by induction that this is true for any $T \ge 2$. Thus suppose $\nu^m(\alpha, n) < \nu^m(\alpha', n')$ for $m = 2, \dots, T-1$, and for any (α, n) and (α', n') satisfying the conditions of the lemma. We must show that $\nu^T(\alpha, n) < \nu^T(\alpha', n')$. The proof proceeds by defining a Bernoulli process $\{Z_t\colon t = 1, 2, \dots\}$ starting from the state (α', n') and which dominates the X_t process in a suitable way.

Note first that for $t = 1, 2, \ldots, T-1$ (when $T > 2$) it follows from the inductive hypothesis that

$$\text{if } (\alpha + S_t)/(n+t) \le \alpha/n, \text{ then } \nu^{T-t}(\alpha + S_t, n+t) \le \nu^{T-t}(\alpha, n),$$

with strict inequality except when $t = T-1$. Since clearly $\nu^{T-t}(\alpha, n) \le \nu^T(\alpha, n)$ with strict inequality when $t = T-1$, it follows from the characterization of $\tau(T)$ given at the beginning of this proof that if $\tau(T) > t$, then $(\alpha + S_t)/(n+t) > \alpha/n$, so that

$$\text{if } \tau(T) > t, \text{ then } S_t/t > \alpha/n. \tag{3}$$

A sufficient description of the probability space P_t on which the random variable X_t is defined ($t = 1, 2, \ldots$) is generated by the set of probabilities and conditional probabilities

$$P(X_1 = 1), P(X_2 = 1|X_1), P(X_3 = 1|X_1, X_2), \ldots,$$
$$P(X_t = 1|X_1, X_2, \ldots, X_{t-1}).$$

For $1 \le r \le t$, we have

$$P(X_r = 1|X_1 = x_1, X_2 = x_2, \ldots, X_{r-1} = x_{r-1}) = (\alpha + s^x_{r-1})/(n + r - 1),$$

where $s^x_{r-1} = \sum_{i=1}^{r-1} x_i$.

We now define the Bernoulli random variables Y_t, Z_t on $\{\tau(T) \ge t\}$, using a refinement P'_t of P_t ($t = 1, 2, \ldots$). The definition is recursive, $S^z_t = \sum_{i=1}^{t} Z_i$ and lowercase x_i, y_i, z_i and s^z_t denote the values taken by the random variables X_i, Y_i, Z_i and S^z_t. Like P_t, P'_t is generated by a set of probabilities and conditional probabilities:

$$P(X_1 = 1) = \frac{\alpha}{n},$$

$$P(Y_1 = 1) = \frac{\alpha'/n' - P(X_1 = 1)}{P(X_1 = 0)},$$

$$Z_1 = X_1 + (1 - X_1)Y_1;$$

$$P(X_2 = 1|x_1, y_1) = P(X_2 = 1|s^x_1) = \frac{\alpha + s^x_1}{n+1},$$

$$P(Y_2 = 1|x_1, y_1) = P(Y_2 = 1|s^z_1) = \frac{(\alpha' + s^z_1)/(n'+1) - P(X_2 = 1|s^z_1)}{P(X_2 = 0|s^z_1)},$$

$$Z_2 = X_2 + (1 - X_2)Y_2;$$

$$\vdots$$

$$P(X_t = 1|x_1, y_1, \ldots, x_{t-1}, y_{t-1}) = P(X_t = 1|s^x_{t-1}) = \frac{\alpha + s^x_{t-1}}{n + t - 1},$$

$$P(Y_t = 1|x_1, y_1, \ldots, x_{t-1}, y_{t-1}) = P(Y_t = 1|s^x_{t-1})$$
$$= \frac{(\alpha' + s^z_{t-1})/(n' + t - 1) - P(X_t = 1|s^z_{t-1})}{P(X_t = 0|s^z_{t-1})},$$

$$Z_t = X_t + (1 - X_t)Y_t.$$

Clearly this definition leaves the conditional probabilities $P(X_t = 1|X_2, \ldots, X_{t-1})$ unaltered, so that P_t' is indeed a refinement of P_t. For it to be a valid definition we must also have

$$0 \le P(Y_t = 1|s_{t-1}^z) \le 1, \quad \text{when } s_{t-1}^z/(t-1) \ge \alpha/n.$$

The second of these inequalities is a trivial consequence of the fact that $(\alpha' + s_{t-1}^z)/(n' + t - 1) \le 1$. The first one holds because

$$\begin{aligned}
&P(X_t = 1|S_{t-1}^z = r) \\
&\quad = \frac{P(X_t = 1, S_{t-1}^z = r)}{P(S_{t-1}^z = r)} \\
&\quad = \sum_{i=0}^{r} \frac{P(X_t = 1, S_{t-1}^x = i, S_{t-1}^z = r)}{P(S_{t-1}^z = r)} \\
&\quad = \sum_{i=0}^{r} P(X_t = 1|S_{t-1}^x = i, S_{t-1}^z = r)\frac{P(S_{t-1}^x = i, S_{t-1}^z = r)}{P(S_{t-1}^z = r)} \\
&\quad = \sum_{i=0}^{r} P(X_t = 1|S_{t-1}^x = i)\frac{P(S_{t-1}^x = i, S_{t-1}^z = r)}{P(S_{t-1}^z = r)},
\end{aligned}$$

since $P(X_t = 1|\, x_1, y_1, x_2, y_2, \ldots, x_{t-1}, y_{t-1})$ depends only on s_{t-1}^x, so that

$$\begin{aligned}
P(X_t = 1|S_{t-1}^z = r) &\le P(X_t = 1|S_{t-1}^x = r)\sum_{i=0}^{r} \frac{P(S_{t-1}^x = i, S_{t-1}^z = r)}{P(S_{t-1}^z = r)} \\
&= P(X_t = 1|S_{t-1}^x = r) \\
&= \frac{\alpha + r}{n + t - 1} \le \frac{\alpha' + r}{n' + t - 1}.
\end{aligned}$$

The last inequality here is a consequence of the three inequalities $r/(t-1) > \alpha/n$ [from (3)], $\alpha/n \le \alpha'/n'$ and $n \ge n'$. Next note that, on $\{\tau(T) \ge t\}$,

$$\begin{aligned}
P(Z_t = 1|s_{t-1}^z) &= P(X_t = 1|s_{t-1}^z) + P(Y_t = 1|s_{t-1}^z)P(X_t = 0|s_{t-1}^z) \\
&= (\alpha' + s_{t-1}^z)/(n' + t - 1).
\end{aligned}$$

This is the required condition for $\{Z_t\colon t = 1, 2, \ldots\}$ to be a Bernoulli process starting from the state (α', n').

By definition

$$\begin{aligned}
R_{\tau(T)}(\alpha, n) &= \mathbf{E}\left[\sum_{1 \le t \le T} I\{t \le \tau(T)\}X_t a^{t-1}\right] \\
&< \mathbf{E}\left[\sum_{1 \le t \le T} I\{t \le \tau(T)\}Z_t a^{t-1}\right] = R_{\tau(T)}(\alpha', n')
\end{aligned}$$

and $W_{\tau(T)}(\alpha, n) = W_{\tau(T)}(\alpha', n')$. Thus $\nu^T(\alpha, n) = \nu_{\tau(T)}(\alpha, n) < \nu_{\tau(T)}(\alpha', n')$.

Finally we note that $\nu^T(\alpha', n') = \sup_{0 < \tau \le T} \nu_\tau(\alpha', n')$, so that $\nu_{\tau(T)}(\alpha', n') \le \nu^T(\alpha', n')$, and hence $\nu^T(\alpha, n) < \nu^T(\alpha', n')$. □

Note. The fact that the event $\{\tau(T) = t\}$ is measurable with respect to $\sigma(\{X_i\colon\ i \le t\})$ [and therefore with respect to $\sigma\{(X_i, Y_i\colon\ i \le t\})$] rather than with respect to $\sigma(\{Z_i\colon\ i \le t\})$ does not upset the above proof. In terms of the Z process this means that $\tau(T)$ is a randomised stopping time, and $\sup_{\tau \le T} \nu_\tau(\alpha', n')$ is unchanged if the set of stopping times over which the supremum is taken is extended to include randomized stopping times. This may be shown, for example, along the lines of the proof of Lemma 3.2 of G [also cf. Chow, Robbins and Siegmund (1971), pages 110–112].

THEOREM 2. *For a Bernoulli reward process,* $\nu(\alpha, n) < \le (\alpha', n')$ *when* $\alpha/n \le \alpha'/n'$, $n' \le n$, *and at least one of these inequalities is strict.*

PROOF. Suppose τ is such that $\nu_\tau(\alpha, n) = \nu(\alpha, n)$. Let

$$R = R_\tau(\alpha, n) = \sum_{t=1}^{\infty} a^{t-1}\mathbf{E}\big[X_t I\{\tau \ge t\}\big],$$

$$W = W_\tau(\alpha, n) = \sum_{t=1}^{\infty} a^{t-1}P\{\tau \ge t\},$$

$$R(T) = \sum_{t=1}^{T} a^{t-1}\mathbf{E}\big[X_t I\{\tau \ge t\}\big], \qquad W(T) = \sum_{t=1}^{T} a^{t-1}P\{\tau \ge t\}.$$

Thus, since $R/W = \nu(\alpha, n)$ and $R(T)/W(T) \le \nu^T(\alpha, n)$,

$$\begin{aligned}\nu(\alpha, n) - \nu^T(\alpha, n) &\le R/W - R(T)/W(T) \le \{R - R(T)\}/W(T)\\ &\le R - R(T) \le a^T/(1 - a).\end{aligned}$$

Hence we have $\lim_{T\to\infty} \nu^T(\alpha, n) = \nu(\alpha, n)$. Thus it follows from Lemma 1 that $\nu(\alpha, n) \le \nu(\alpha', n')$. Strict inequality holds because

$$\begin{aligned}\nu^T(\alpha', n') - \nu^T(\alpha, n) &\ge \big\{R_{\tau(T)}(\alpha', n') - R_{\tau(T)}(\alpha, n)\big\}/W_{\tau(T)}(\alpha, n)\\ &\ge (1 - a)\mathbf{E}\Big[\sum_{1\le t\le T} I\{t \le \tau(T)\}(Z_t - X_t)a^{t-1}\Big]\\ &\ge a(1 - a)E\big[I\{X_1 = 1\}(Z_2 - X_2)\big]\\ &= a(1 - a)(\alpha/n)P(X_2 = 0, Y_2 = 1|X_1 = 1),\end{aligned}$$

which is positive and independent of T. □

Since $\nu^1(n\lambda, n) = \lambda$, the learning component of the dynamic allocation index may be written as $\nu(n\lambda, n) - \lambda$. From Theorem 2 it follows that this is a decreasing function of n for any fixed λ $(0 \le \lambda \le 1)$.

Now suppose ρ is a constant $(0 < \rho \le 1)$ and $\alpha(n)$ is defined by the equation $\nu(\alpha(n), n) = \rho$ $(n > 0)$.

COROLLARY 1. *$\alpha(n)/n$ is increasing in n.*

PROOF. Suppose that $\nu(\alpha(n_1), n_1) = \nu(\alpha(n_2), n_2) = \rho$ and that $n_1 < n_2$. It follows from Theorem 2 that $\nu(\alpha(n_1), n_1) > \nu(n_2 n_1^{-1}\alpha(n_1), n_2) < \nu(\alpha(n_2), n_2)$. Hence $n_2 n_1^{-1}\alpha(n_1) < \alpha(n_2)$; that is, $\alpha(n_1)/n_1 < \alpha(n_2)/n_2$. □

We also have $\nu(\alpha(n), n) - \alpha(n)/n = O(1/n)$ [see Wang (1991)]. Thus $\lim_{n\to\infty} \alpha(n)/n = \rho$, and each isoindex curve in the space (α, n) is therefore asymptotic to a line through the origin with a slope equal to the index value.

3. Sampling processes with continuous distributions. An analogous result to Theorem 2 for continuous distributions is first given. This is then applied to normal and negative exponentially distributed reward and target processes.

In this section all densities are with respect to Lebesgue measure, possibly in a space of dimension greater than 1. Suppose $\mathscr{D}$ is a parametric family of distributions with densities $f(\cdot|\theta)$ with common support and that the sample sum $S_n = \sum_{t\le n} X_t$ is a sufficient statistic. If θ has the probability density $\pi_0(\theta)$, the posterior density for θ is of the form $\pi(\theta|X_1, X_2, \ldots, X_n) \propto \pi_0(\theta)\prod_{1\le i\le n} f(X_i|\theta)$. Since S_n is sufficient it follows from the Neyman factorization theorem that $\pi(\theta|X_1, X_2, \ldots, X_n)$ may be written in the form $\pi(\theta|S_n, n)$, because it depends on $(X_1, X_2, \ldots, X_n)$ only as a function of (S_n, n). The state of the corresponding bandit sampling process may therefore be represented by (S_n, n), suppressing the dependence on the common prior density π_0. Write $f(x|S_n, n)$ for the conditional density $\int f(x|\theta)\pi(\theta|S_n, n)\,d\theta$ and $F(x|S_n, n)$ for the corresponding distribution function.

Suppose now that the sampling process starts from the state (Σ, n), and suppose initially that the process is not a target process, to avoid the complication of the additional completion state C. Let $S_0 = 0$, $S_t = \sum_{i\le t} X_i$ and define inductively

$$X'_{t+1} = G_t\{F(X_{t+1}|\Sigma + S_t, n + t)\}, \qquad t = 0, 1, 2, \ldots,$$

where $G_t(\cdot)$ is the inverse function of $F(\cdot|\Sigma' + S'_t, n' + t)$, so that

$$F(G_t(b)|\Sigma' + S'_t, n' + t) = b, \qquad 0 \le b \le 1,$$

and $S'_0 = 0$, $S'_t = \sum_{i\le t} X'_i$. Thus the process $\{X'_t: t = 1, 2, \ldots\}$ is another realization of the same sampling process, this time starting from the state (Σ', n'). Moreover we have defined a (1-1) correspondence between realizations starting from (Σ, n) and (Σ', n'), so that the filtration $\{\mathscr{F}_t: t \ge 1\}$, where $\mathscr{F}_t$ denotes the σ field $\sigma(\{X_i: i \le t\})$, and the corresponding filtration defined in terms of the primed process are identical, and we may therefore, in particular, regard a stopping time which is defined in terms of one process as applying to both.

Denote the expected undiscounted reward from taking a further sample from the process when it is in state (S, m) by $r(S, m)$. Our analogue of

Theorem 2 holds under the following conditions:

1. $S'_t \geq S_t$ when $\Sigma/n \leq \Sigma'/n'$, $n \geq n'$ and $S_i/i \geq \Sigma/n$ for $1 \leq i \leq t$.
2. $r(S, m) \leq r(T, n)$ if $S/m = T/n$ and $m \geq n$.
3. $r(S, m)$ is an increasing function of S.
4. $\nu(S, m) < \infty$ for all (S, m).
5. $\sup_{M \geq 0} \mathbf{E}|r(S + X_1 + X_2 + \cdots + X_M, m + M)| < \infty$ for all (S, m), where M is a nonrandom positive integer.

Note. For target processes, conditions 4 and 5 are always satisfied.

THEOREM 3. *For a bandit sampling process with no completion state and satisfying conditions* 1–5,

$$\nu(\Sigma, n) \leq \nu(\Sigma', n') \text{ if } \Sigma/n \leq \Sigma'/n' \text{ and } n \geq n'. \tag{4}$$

We first prove the following lemma.

LEMMA 2. *Under the conditions of Theorem* 3,

$$\nu^T(\Sigma, n) \leq \nu^T(\Sigma', n'), \qquad T = 1, 2, \ldots .$$

PROOF. This proceeds by induction on T. We have

$$\nu^1(\Sigma, n) = r(\Sigma, n) \leq r(\Sigma', n') = \nu^1(\Sigma', n'),$$

so the lemma holds for $T = 1$.

Now suppose the lemma is true for $T \leq V - 1$, where $V \geq 2$. Let $\tau(V)$ $(\leq V)$ be the stopping time with respect to the filtration $\{\mathcal{F}_t\colon t \geq 1\}$ such that $\nu_{\tau(V)}(\Sigma, n) = \nu^V(\Sigma, n)$ and which is defined by Theorem 1(ii). Thus on $\{\tau(V) > t\}$,

$$\nu^{V-t}(\Sigma + S_t, n + t) \geq \nu^V(\Sigma, n) \geq \nu^{V-t}(\Sigma, n),$$

the second inequality following from the fact that if $A \subset B \subset \Re$, then $\sup(B) \geq \sup(A)$. From the inductive hypothesis it therefore follows that

$$(\Sigma + S_t)/(n + t) \geq \Sigma/n,$$

so that $S_t/t \geq \Sigma/n$ $(1 \leq t \leq V - 1)$.

Now using conditions 1, 2 and 3, we have

$$\begin{aligned} R_{\tau(V)}(\Sigma, n) &= \mathbf{E}\left[\sum_{0 \leq t < V} I\{t < \tau(V)\} r(\Sigma + S_t, n + t) a^t\right] \\ &= \mathbf{E}\left[\sum_{0 \leq t < V} I\{t < \tau(V)\} r(\Sigma' + S'_t, n' + t) a^t\right] = R_{\tau(V)}(\Sigma', n'). \end{aligned}$$

Trivially

$$W_{\tau(V)}(\Sigma, n) = W_{\tau(V)}(\Sigma', n'),$$

so that

$$\nu^V(\Sigma', n') = \sup_\tau (\Sigma', n') \geq \nu_{\tau(V)}(\Sigma', n') = R_{\tau(V)}(\Sigma', n')/W_{\tau(V)}(\Sigma', n')$$

$$\geq R_{\tau(V)}(\Sigma, n)/W_{\tau(V)}(\Sigma, n) = \nu^V(\Sigma, n).$$

Thus the lemma is also true for $T = V$. This completes the induction and the proof of the lemma. □

To complete the proof of the theorem it is sufficient to show that $\lim_{T\to\infty} \nu^T(\Sigma, n) = \nu(\Sigma, n)$. This follows along similar lines to the corresponding stage in the proof of Theorem 2, using condition 5.

COROLLARY 2. *Theorem* 3 *remains true for a target process, in which case condition* 3 *is required only for states which can be reached from at least one of the initial states* (Σ, n) *and* (Σ', n') *without first reaching the target.*

The only differences between the proofs of the theorem and the corollary are that for a target process $\tau(T) \leq \sigma$, and that in the proof of the lemma where the stopping time $\tau(T)$ is applied to the primed process it is replaced by $\tau'(T) = \min[\sigma', \tau(T)]$, where σ and σ' are the times taken to reach the target for the unprimed and primed processes, respectively. Note that $R_{\tau'(T)}(\Sigma', n') = R_{\tau(T)}(\tau', n')$ and $W_{\tau'(T)}(\Sigma', n') \leq W_{\tau(T)}(\Sigma', n')$.

EXAMPLE 1 (Normal reward and target processes). Suppose that the observations are normally distributed with mean μ (unknown) and variance σ^2 (known). The conjugate prior densities for μ are of the form $(2\pi\sigma^2 n^{-1})^{-1/2} \exp\{-n(\mu - \Sigma n^{-1})^2/(2\sigma^2)\}$. When n is an integer this is the posterior density after observations $x_1, x_2, \ldots, x_n$ with total Σ starting from an improper uniform prior density over the real line [e.g., see Raiffa and Schlaifer (1961)]. Thus (Σ, n) identifies the state.

Condition 1 of Theorem 3 may by shown to hold. For the normal reward process

$$r(\Sigma, n) = \int_{-\infty}^{+\infty} x f(x|\Sigma, n)\, dx = \Sigma/n.$$

For the normal target process, invariance properties under changes of location and scale mean that we may assume the target to be 0 ($L = 0$) and the variance to be 1 without loss of generality (G, Theorem 6.21). Therefore, $r(\Sigma, n) = P(X \geq 0)$, where X is normally distributed with mean Σ/n and variance $1 + n^{-1}$, so that

$$r(\Sigma, n) = \Phi\left[\Sigma\{n(1+n)\}^{-1/2}\right],$$

where $\Phi(\cdot)$ is the standard normal distribution function. Conditions 2–5 of Theorem 3 easily follow for the reward process, as do condition 2, and the

restricted version (see Corollary 2) of condition 3 for the target process provided $\Sigma' \leq 0$.

For the reward process it follows from Theorem 3 that $\nu(\lambda n, n)$ is a decreasing function of n for fixed λ. Since the learning component may be written as $\nu(\lambda n, n) - \lambda$ it follows that this too is a decreasing function of n for fixed λ. For the target process Corollary 2 gives the result that $\nu(\lambda n, n)$ is a decreasing function of n for fixed λ provided $\lambda < 0$ (which it normally will be if the target has not been reached). Since $r(n\lambda, n)$ is a decreasing function of n for fixed λ this result may not be expressed in terms of the learning component.

EXAMPLE 2 (Exponential reward and target processes). For an exponential process, $f(x|\theta) = \theta e^{-\theta x}$. The conjugate prior densities for θ are of the form $(\Gamma(n))^{-1}\Sigma^n\theta^{n-1}e^{-\theta\Sigma}$. When n is an integer this is the posterior density after observations $x_1, x_2, \ldots, x_n$ with total Σ starting from an improper prior density proportional to θ^{-1}. The conditional density of the next observation is

$$f(x|\Sigma, n) = \int_0^\infty f(x|\theta)\pi(\theta|\Sigma, n)\, d\theta = n\Sigma^n/(\Sigma + x)^{n+1},$$

and thus (Σ, n) identifies the state of the process. When $n \geq n'$,

$$X'_{t+1} = (\Sigma' + S'_t)\left\{1 + \frac{X_{t+1}}{\Sigma + S_t}\right\}^{(n+t)/(n'+t)} - (\Sigma + S'_t)$$

$$\geq \frac{(\Sigma' + S'_t)(n + t)}{(\Sigma + S_t)(n' + t)} X_{t+1}.$$

It follows that condition 1 is satisfied and $X'_t \geq X_t$. For the exponential reward process,

$$r(\Sigma, n) = \int_0^\infty xf(x|\Sigma, n)\, dx = \frac{\Sigma}{n - 1},$$

and conditions 2–5 are easily checked. Invariance arguments show that $\nu(\Sigma, n) = \Sigma n^{-1}\nu(1, n)$ (G, Theorem 6.11), so it follows from Theorem 3 that $\nu(1, n)$ is a decreasing function of n.

For the exponential target process ($L = 1$),

$$r(\Sigma, n) = \int_1^\infty f(x|\Sigma, n)\, dx = \left(\frac{\Sigma}{\Sigma + 1}\right)^n,$$

and conditions 2 and 3 are easily checked, so from Corollary 2 it follows that $\nu(\lambda n, n)$ is a decreasing function of n for any fixed positive value of λ.

For both reward and target processes $r(n\lambda, n)$ is decreasing as a function of n for fixed λ, so our results may not be expressed in terms of the learning component.

4. A related conjecture. The most widely studied bandit problem, apart from the case of geometric discounting, is the undiscounted case with a finite horizon N. Suppose two independent Bernoulli arms in states (α, n) and (α', n') are available. It has been shown [Berry (1972)] that, as with geometric discounting, a play-the-winner rule maximizes the sum of the first N observations. Berry (1972) conjectured that when $\alpha/n \leq \alpha'/n'$ and $n > n'$, the optimal strategy is to pull the arm which is in state (α', n'). This conjecture is similar to our Theorem 2, and a similar heuristic argument indicates that it is true. Unfortunately the proof of Theorem 2 depends on the optimality of an index policy, and this does not hold when the discounting is not geometric. Berry's conjecture remains unproven as far as we know.

Acknowledgment. The authors are pleased to acknowledge the helpful comments of a referee, which have led to an improved paper.

REFERENCES

BARRA, J. R. (1981). *Mathematical Basis of Statistics*. Academic, New York.

BELLMAN, R. E. (1956). A problem in the sequential design of experiments. *Sankhyā Ser. A* **30** 221–252.

BERGMAN, S. W. and GITTINS, J. C. (1985). *Statistical Methods for Pharmaceutical Research Planning*. Dekker, New York.

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 & Hall, London.

CHOW, Y. S, ROBBINS, H. and SIEGMUND, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

GITTINS, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley, Chichester.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation for the sequential design of experiments. In *Progress in Statistics* (J. Gani, K. Savkadi and I. Vincze, eds.) 241–266. North-Holland, Amsterdam.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Harvard Business School.

WANG, Y. G. (1991). Gittins indices and constrained allocation in clinical trials. *Biometrika* **78** 101–111.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1 SOUTH PARKS ROAD
OXFORD OX1 3TG
ENGLAND