The Annals of Statistics
1987, Vol. 15, No. 3, 1091–1114

ADAPTIVE TREATMENT ALLOCATION AND THE MULTI-ARMED BANDIT PROBLEM[1]

BY TZE-LEUNG LAI

Columbia University

A class of simple adaptive allocation rules is proposed for the problem (often called the "multi-armed bandit problem") of sampling $x_1, \ldots, x_N$ sequentially from k populations with densities belonging to an exponential family, in order to maximize the expected value of the sum $S_N = x_1 + \cdots + x_N$. These allocation rules are based on certain upper confidence bounds, which are developed from boundary crossing theory, for the k population parameters. The rules are shown to be asymptotically optimal as $N \to \infty$ from both Bayesian and frequentist points of view. Monte Carlo studies show that they also perform very well for moderate values of the horizon N.

1. Introduction and summary. Let Π_j, $j = 1, \ldots, k$, denote statistical populations (treatments, manufacturing processes, etc.) specified, respectively, by univariate density functions $f(x; \theta_j)$ with respect to some nondegenerate measure ν, where $f(\cdot\,; \cdot)$ is known and the θ_j are unknown parameters belonging to some set Θ. Assume that $E_\theta|X| = \int_{-\infty}^{\infty} |x| f(x; \theta)\, d\nu(x) < \infty$ for all $\theta \in \Theta$. How should we sample $x_1, \ldots, x_N$ sequentially from the k populations in order to maximize, in some sense, the expected value of the sum $S_N = x_1 + \cdots + x_N$? This is the classical "multi-armed bandit problem," with specified horizon N, in the statistics and engineering literature. The name derives from an imagined slot machine with $k \geq 2$ arms. When an arm is pulled, the player wins a random reward. For each arm j, there is an unknown probability distribution Π_j of the reward, and the player's problem is to choose N successive pulls on the k arms so as to maximize the total expected reward. The problem is prototypical of a general class of adaptive control and design problems in which there is a fundamental dilemma between "information" (such as the need to learn from all populations about their parameter values in the present case) and "control" (such as the objective of sampling only from the best population), cf. Kumar (1985). Another noteworthy example of such problems is in the context of sequential clinical trials, where there are k treatments of unknown efficacy to be chosen sequentially to treat a large class of N patients, cf. Chernoff (1967).

An *adaptive allocation rule* φ is a sequence of random variables $\varphi_1, \ldots, \varphi_N$ with values in the set $\{1, \ldots, k\}$ and such that the event $\{\varphi_i = j\}$, $j = 1, \ldots, k$, belongs to the σ-field $\mathcal{F}_{i-1}$ generated by the previous observations $\varphi_1, x_1, \ldots, \varphi_{i-1}, x_{i-1}$. Letting $\mu(\theta) = \int_{-\infty}^{\infty} x f(x; \theta)\, d\nu(x)$ and $\theta = (\theta_1, \ldots, \theta_k) \in$

Received April 1986; revised January 1987.

[1]This research was supported by the National Science Foundation, the National Institutes of Health and the U. S. Army Research Office.

AMS 1980 *subject classifications*. Primary 62L05; secondary 60G40, 62L12.

Key words and phrases. Sequential experimentation, adaptive control, dynamic allocation, boundary crossings, upper confidence bounds.

Θ^k, it follows that for every $n \leq N$

$$E_\theta S_n = \sum_{i=1}^{n} \sum_{j=1}^{k} E_\theta\left\{E_\theta\left(x_i I_{\{\varphi_i=j\}} | \mathcal{F}_{i-1}\right)\right\} = \sum_{j=1}^{k} \mu(\theta_j) E_\theta T_N(j), \tag{1.1}$$

where

$$T_n(j) = \sum_{i=1}^{n} I_{\{\varphi_i=j\}} \tag{1.2}$$

denotes the number of observations that φ samples from Π_j up to stage n. Hence, the objective of maximizing $E_\theta S_N$ is equivalent to that of minimizing the regret

$$R_N(\theta) = N\mu^*(\theta) - E_\theta S_N = \sum_{j:\, \mu(\theta_j) < \mu^*(\theta)} \left(\mu^*(\theta) - \mu(\theta_j)\right) E_\theta T_N(j), \tag{1.3}$$

where $\mu^*(\theta) = \max_{1 \leq j \leq k} \mu(\theta_j)$. In particular, the usual Bayesian formulation of the multi-armed bandit problem, stated in the form of maximizing $\int E_\theta S_N \, dH(\theta)$, can be restated in the more convenient form of minimizing the Bayes risk $\int R_N(\theta) \, dH(\theta)$, where H is a prior distribution on Θ^k.

In principle, one can use dynamic programming to study the problem of minimizing $\int R_N(\theta) \, dH(\theta)$. For the case where $k = 2$ and Θ has two elements, which we shall denote by a, b with $\mu(a) > \mu(b)$, Feldman (1962) found by this approach that for the prior distribution which assigns probability p to the parameter vector $\theta = (a, b)$ and probability $1 - p$ to the vector (b, a), the allocation rule that chooses Π_1 or Π_2 at stage $i + 1$ according as $p_i \geq \frac{1}{2}$ or $p_i < \frac{1}{2}$ is Bayes, where p_i denotes the posterior probability in favor of the vector (a, b) at the end of stage i ($p_0 = p$). For the case of $k = 2$ Bernoulli populations, Fabius and van Zwet (1970) and Berry (1972) studied the dynamic programming equations analytically and obtained several interesting results about the Bayes rules with respect to general priors. Beyond the two-point priors considered by Feldman, Bayes rules are usually described only implicitly by the dynamic programming equations, whose numerical solution is too complicated for practical implementation when N is large. It is therefore of great interest to develop asymptotic approximations to the Bayes rules for large N and to find simple and easily interpretable adaptive allocation rules that are nearly optimal from both the Bayesian and the frequentist viewpoints.

In this paper we consider the case where the densities $f(y; \theta_j)$, $j = 1, \ldots, k$, belong to the exponential family

$$f(y; \theta) = e^{\theta y - \psi(\theta)}, \tag{1.4}$$

and propose a class of simple adaptive allocation rules which are asymptotically Bayes (as $N \to \infty$) with respect to a wide variety of priors on the natural parameter space Θ and which also have asymptotically optimal frequentist properties. For the exponential family (1.4), $\mu(\theta) = \psi'(\theta)$ is increasing in θ since $\psi''(\theta) = \mathrm{Var}_\theta Y$, and the Kullback–Leibler information number $I(\theta, \lambda) =$

$E_\theta \log[f(Y;\theta)/f(Y;\lambda)]$ is given by

$$(1.5)\quad I(\theta,\lambda) = (\theta-\lambda)\psi'(\theta) - (\psi(\theta)-\psi(\lambda)) = \int_\theta^\lambda (\lambda - t)\psi''(t)\,dt.$$

Based on successive observations $Y_{j,1},\ldots,Y_{j,r}$ from Π_j, we construct an estimator $\hat{\theta}_{j,r}$ of θ_j by the method of maximum likelihood, which leads to the equation

$$(1.6)\quad \mu\big(\hat{\theta}_{j,r}\big) = (Y_{j,1}+\cdots+Y_{j,r})/r.$$

Define an "upper confidence bound" for θ_j of the form

$$(1.7)\quad U_{j,r} = \inf\big\{\theta:\ \theta \geq \hat{\theta}_{j,r} \text{ and } I\big(\hat{\theta}_{j,r},\theta\big) \geq r^{-1}g(r/N)\big\},$$

where the function $g\ (\geq 0)$ satisfies certain assumptions that imply a number of desirable properties for $U_{j,r}$, as will be discussed in detail in Section 2.

If the values $\theta_1,\ldots,\theta_k$ were known, the optimal rule would obviously be to sample from the population with the largest θ_j. In ignorance of $\theta_1,\ldots,\theta_k$, one may try estimating them at stage n with the estimators $\hat{\theta}_{1,T_n(1)},\ldots,\hat{\theta}_{k,T_n(k)}$ and sampling at stage $n+1$ from the population Π_j with the largest $\hat{\theta}_{j,T_n(j)}$. This is the so-called "play-the-leader" rule. The difficulty with this rule is that we may have sampled too little from an apparently inferior population to get a reliable estimate of its parameter and may thereby miss the actually superior population.

Instead of sampling at stage $n+1$ from the population with the largest $\hat{\theta}_{j,T_n(j)}$, we propose herein the following simple modification:

$$(1.8)\quad \text{Sample at stage } n+1 \text{ from the population } \Pi_j \text{ with the largest upper confidence bound } U_{j,T_n(j)},$$

where $U_{j,r}$ is defined in (1.7) and $n \geq k$. (During the first k stages, we sample once from each population.) To explain the heuristic idea behind this approach, we first note that the upper confidence bound $U_{j,r}$ inflates the estimator $\hat{\theta}_{j,r}$ by an amount which decreases with the number r of observations already taken from the population. Thus, $U_{j,r}$ depends not only on the estimator $\hat{\theta}_{j,r}$ but also on the sample size r, and comparing the k populations on the basis of $U_{j,T_n(j)}$ involves not only the parameter estimates but also the sample sizes of all populations. Making use of certain properties of $U_{j,r}$ discussed in Section 2, we show in Sections 3 and 4 that allocation rules of the form (1.8) are asymptotically optimal from both the Bayesian and the frequentist viewpoints as $N \to \infty$. Some simulations results, presented in Section 5, show that these allocation rules also perform well for moderate values of N.

2. A class of upper confidence bounds for the exponential family. Let $Y_1, Y_2, \ldots$ be i.i.d. random variables having the common density $f(x;\theta) = e^{\theta x - \psi(\theta)}$ with respect to some nondegenerate measure ν. Note that the natural parameter space

$$\Theta = \left\{\theta:\ \int e^{\theta x}\,d\nu(x) < \infty\right\}$$

is an interval Since $\mu(\theta) = \psi'(\theta)$ is increasing in θ, $\mu(\Theta)$ is also an interval. Let

$A \subset \Theta$ be an open interval with endpoints $(-\infty \leq)a_1 < a_2(\leq \infty)$ such that

$$\inf_{a_1-r<\theta<a_2+r} \psi''(\theta) > 0, \qquad \sup_{a_1-r<\theta<a_2+r} \psi''(\theta) < \infty \tag{2.1}$$

and ψ'' is uniformly continuous on $(a_1 - r, a_2 + r)$ for some $r > 0$.

In particular, if $Y_1, Y_2, \ldots$ are normally distributed, then Θ is the entire real line and we can take $A = \Theta$.

Let $S_n = Y_1 + \cdots + Y_n$. Based on the observations $Y_1, \ldots, Y_n$, the method of maximum likelihood leads to the equation $\mu(\theta) = S_n/n$, which may not have a solution in Θ. In the sequel, we shall assume that θ is known to lie in the subinterval A of Θ. In this case, the maximum likelihood estimate of θ is given by

$$\begin{aligned} \hat{\theta}_n &= \mu^{-1}(S_n/n), \quad \text{if } \mu(a_1) \leq S_n/n \leq \mu(a_2) \\ &= a_2, \quad \text{if } S_n/n > \mu(a_2) \\ &= a_1, \quad \text{if } S_n/n < \mu(a_1). \end{aligned} \tag{2.2}$$

Let $N > 1$ and let g be a nonnegative function on $(0, \infty)$ satisfying the following assumptions:

$$\sup_{t \geq a} g(t)/t < \infty, \quad \text{for all } a > 0, \tag{2.3}$$

$$g(t) \sim \log t^{-1}, \quad \text{as } t \to 0, \tag{2.4}$$

$$g(t) \geq \log t^{-1} + \xi \log\log t^{-1}, \quad \text{as } t \to 0, \tag{2.5}$$

for some ξ. Based on the n observations $Y_1, \ldots, Y_n$, define the "upper confidence bound"

$$U_n(g, N) = \inf\{\theta \in A: \theta \geq \hat{\theta}_n \text{ and } I(\hat{\theta}_n, \theta) \geq n^{-1} g(n/N)\} \tag{2.6}$$

($\inf \varnothing = \infty$), where $I(\theta, \lambda)$ is the Kullback–Leibler information number given by (1.5). Note in this connection that for fixed $a \in \Theta$ the function $I(a, \theta)$ is increasing in $\theta \geq a$ but decreasing in $\theta \leq a$, with $I(a, a) = 0$. Some basic asymptotic properties of the sequence $\{U_n(g, N)\}$ that will be needed in the sequel are given in

THEOREM 1. *Define $\hat{\theta}_n$ by* (2.2) *and $U_n(g, N)$ by* (2.6), *where g is a nonnegative function on* $(0, \infty)$ *satisfying* (2.3)–(2.5). *Let $\alpha_N < \beta_N$ be positive numbers such that as $N \to \infty$,*

$$\alpha_N \to 0 \quad \textit{and} \quad N^{1/2}\alpha_N \to \infty, \ \beta_N \to \infty \ \textit{and} \ \beta_N = o\big((\log N)^{1/2}\big). \tag{2.7}$$

For $d > 0$ and $\theta \in \Theta$ define

$$\begin{aligned} T(\theta, d) &= \inf\{n: U_n(g, N) \leq \theta + d\}, \\ L(\theta, d) &= \sup\{n: U_n(g, N) \geq \theta + d\}, \\ \#(\theta, d) &= \#\{n: U_n(g, N) \geq \theta + d\}, \end{aligned} \tag{2.8}$$

where $\inf \varnothing = \infty$, $\sup \varnothing = 0$ *and the notation $\#S$ denotes the number of*

elements of a set S.

(i) $T(\theta, d) - 1 \le \#(\theta, d) \le L(\theta, d)$ *and as* $N \to \infty$,

$$P_\theta\{T(\theta, d) \le (1 - \gamma)(\log Nd^2)/I(\theta, \theta + d)\} \to 0, \tag{2.9}$$

$$P_\theta\{L(\theta, d) \ge (1 + \gamma)(\log Nd^2)/I(\theta, \theta + d)\} \to 0, \tag{2.10}$$

for every $0 < \gamma < 1$; *moreover,*

$$E_\theta T(\theta, d) \sim E_\theta \#(\theta, d) \sim E_\theta L(\theta, d) \sim (\log Nd^2)/I(\theta, \theta + d), \tag{2.11}$$

the convergence in (2.9), (2.10) *and* (2.11) *being uniform in* $\alpha_N \le d \le \beta_N$, $\theta \in A$ *and* $\theta + d < a_2 + r$, *where* r *is given in* (2.1).

(ii) *As* $N \to \infty$,

$$P_\theta\{U_n(g, N) \le \theta - d \text{ for some } n\} = O\big((Nd^2)^{-1}(\log Nd^2)^{-\xi - 1/2}\big), \tag{2.12}$$

uniformly in $d \ge \alpha_N$ *and* $a_1 < \theta - d < \theta < a_2$, *where* ξ *is given in* (2.5).

PROOF. Noting that for $x \in \Theta$,

$$U_n(g, N) \le x \Leftrightarrow \hat\theta_n \le x \quad \text{and} \quad I\big(\hat\theta_n, x\big) \ge n^{-1}g(n/N), \tag{2.13}$$

(i) can be proved by a straightforward modification of the proof of Theorem 3 of Lai (1985). To prove (ii), first note that by (2.13),

$$\begin{aligned} &P_\theta\{U_n(g, N) \le \theta - d \text{ for some } n\} \\ &\quad = P_\theta\big\{\hat\theta_n \le \theta - d \text{ and } I\big(\hat\theta_n, \theta - d\big) \ge n^{-1}g(n/N) \text{ for some } n\big\}. \end{aligned} \tag{2.14}$$

Let $\overline{Y}_n = S_n/n$. By Theorem 1(iii) of Lai (1985), as $N \to \infty$,

$$\begin{aligned} &P_\theta\big\{a_1 - r < \mu^{-1}\big(\overline{Y}_n\big) \le \theta - d \\ &\qquad \text{and } I\big(\mu^{-1}\big(\overline{Y}_n\big), \theta - d\big) \ge n^{-1}g(n/N) \text{ for some } n\big\} \\ &\qquad = O\big((Nd^2)^{-1}(\log Nd^2)^{-\xi - 1/2}\big), \end{aligned} \tag{2.15}$$

uniformly in $\theta \in A$ and $\theta - d \in A$ with $d \ge \alpha_N$. Since $\hat\theta_n = (\mu^{-1}(\overline{Y}_n) \wedge a_2) \vee a_1$, (2.12) follows from (2.14), (2.15) and Lemma 1. □

LEMMA 1. (i) *If* $\overline{Y}_n \in \mu(\Theta)$, *then* $I(\mu^{-1}(\overline{Y}_n), \lambda) \ge I(\hat\theta_n, \lambda)$ *for all* $\lambda \in A$.

(ii) *There exists* $\eta > 0$ *such that as* $N \to \infty$,

$$\begin{aligned} &P_\theta\big\{\overline{Y}_n \le \mu(a_1 - r) \text{ and } I\big(\hat\theta_n, \theta - d\big) \ge n^{-1}g(n/N) \text{ for some } n\big\} \\ &\qquad = O\big((Nd^2)^{-1}\exp\big\{-\eta(\log Nd^2)^{1/2}\big\}\big), \end{aligned} \tag{2.16}$$

uniformly in $d \ge \alpha_N$ *and* $a_1 < \theta - d < \theta < a_2$.

PROOF. (i) Consider the case $\hat\theta_n \le \lambda\ (< a_2)$. For fixed λ, the function $I(x, \lambda)$ is decreasing in $x \le \lambda$. Since $\mu^{-1}(\overline{Y}_n) \le \hat\theta_n$ in this case, the desired conclusion follows. The case $\hat\theta_n \ge \lambda\ (> a_1)$ can be treated similarly.

(ii) Noting that $\mu = \psi'$ and that $\hat{\theta}_n = a_1$ if $\bar{Y}_n \le \mu(a_1)$, define

$$T = \inf\{n: S_n \le n\psi'(a_1 - r) \text{ and } I(a_1, \theta - d) \ge n^{-1}g(n/N)\}$$

(inf $\varnothing = \infty$). Choose $c_2 > c_1 > 0$ and $\eta > 0$ such that

$$\eta < 2rc_1c_2^{-1/2}, \qquad c_1 \le \psi''(x)/2 \le c_2 \quad \text{for all } a_1 - r < x < a_2 + r. \tag{2.17}$$

In view of (1.5), we can choose $0 < \rho < 1$ such that for all $0 < t \le \rho$,

$$g(t) \ge \log t^{-1} + \xi \log\log t^{-1} \tag{2.18}$$

and

$$(\log t^{-1})^{-\xi}\exp\{-2rc_1c_2^{-1/2}g^{1/2}(t)\} < \exp\{-\eta(\log t^{-1})^{1/2}\}. \tag{2.19}$$

For $a_1 < \theta - d < \theta < a_2$, we have

$$\begin{aligned} P_\theta(T < \infty) &= \int_{\{T<\infty\}} \exp\{(\theta - a_1)S_T - T(\psi(\theta) - \psi(a_1))\}\, dP_{a_1} \\ &= \int_{\{T<\infty\}} \exp\{(\theta - a_1)(S_T - T\psi'(a_1)) - TI(a_1, \theta)\}\, dP_{a_1} \qquad (2.20) \\ &\le \int_{\{T<\infty\}} \exp\{-T[I(a_1, \theta) + (\theta - a_1)(\psi'(a_1) - \psi'(a_1 - r))]\}\, dP_{a_1}, \end{aligned}$$

since $S_T \le T\psi'(a_1 - r)$ on $\{T < \infty\}$. By (1.5), (2.17) and the mean value theorem,

$$\psi'(a_1) - \psi'(a_1 - r) \ge 2c_1r, \qquad I^{1/2}(a_1, \theta) \le c_2^{1/2}(\theta - a_1), \tag{2.21}$$

$$I(a_1, \theta) = \int_{a_1}^{\theta}(\theta - t)\psi''(t)\, dt \ge I(a_1, \theta - d) + c_1d^2. \tag{2.22}$$

From (2.20)–(2.22), it follows that

$$\begin{aligned} P_\theta(T < \infty) \le \int_{\{T<\infty\}} \exp\{-T[I(a_1, \theta - d) + c_1d^2 \\ + 2rc_1c_2^{-1/2}I^{1/2}(a_1, \theta - d)]\}\, dP_{a_1}. \end{aligned} \tag{2.23}$$

On $\{T \le \rho N\}$, $TI(a_1, \theta - d) \ge g(T/N) \ge \log(N/T) + \xi \log\log(N/T)$ by (2.18), and therefore it follows from (2.23) that

$$\begin{aligned} P_\theta(T < \infty) &\le \int_{\{T>\rho N\}} \exp(-c_1d^2T)\, dP_{a_1} \\ &\quad + N^{-1}\int_{\{T\le\rho N\}} Te^{-c_1d^2T}\{(\log(N/T))^{-\xi} \\ &\qquad \times \exp[-2rc_1c_2^{-1/2}T^{1/2}g^{1/2}(T/N)]\}\, dP_{a_1} \qquad (2.24) \\ &< \exp(-\rho c_1Nd^2) \\ &\quad + (Nd^2)^{-1}\int_{\{T\le\rho N\}} (d^2T)\exp\{-c_1d^2T - \eta(\log(N/T))^{1/2}\}\, dP_{a_1}, \end{aligned}$$

by (2.19). The function $h(x) = -\frac{1}{2}c_1x - \eta\{\log(Nd^2/x)\}^{1/2}$ is decreasing in

$1 \le x \le \rho N d^2$, provided that Nd^2 is sufficiently large. Hence, on $\{1 \le d^2 T \le \rho N d^2\}$,

$$\exp\left\{-\tfrac{1}{2}c_1 d^2 T - \eta(\log(N/T))^{1/2}\right\} \le \exp\left\{-\eta(\log Nd^2)^{1/2}\right\}.$$

Moreover, we have $-\eta(\log(N/T))^{1/2} \le -\eta(\log Nd^2)^{1/2}$ on $\{T < d^{-2}\}$, and $\sup_{x>0} x \exp(-\frac{1}{2}c_1 x) < \infty$. Therefore, the desired conclusion (2.16) follows from (2.24). □

For applications of Theorem 1 to the multi-armed bandit problem described in Section 1, N represents the specified horizon and we therefore have to restrict to $n \le N$ in the confidence sequence $\{U_n(g, N)\}$ and in the definitions (2.8) of $T(\theta, d)$, $L(\theta, d)$ and $\#(\theta, d)$. In this connection, we will use in (2.8) the convention $\inf \varnothing = N + 1$ (instead of $\inf \varnothing = \infty$). Theorem 1 still holds under this convention when n is restricted to $\{1, \ldots, N\}$. By (1.5) and (2.17),

$$(2.25)\quad c_1(\theta - \lambda)^2 \le I(\theta, \lambda) \le c_2(\theta - \lambda)^2, \quad \text{for all } \theta, \lambda \in (a_1 - r, a_2 + r).$$

Hence, Theorem 1 implies that N times the boundary crossing probability (2.12) is of a smaller order of magnitude than $E_\theta T(\theta, d)$ [or $E_\theta \#(\theta, d)$, $E_\theta L(\theta, d)$ by (2.11)] if $\xi > -3/2$, but has a larger order of magnitude than $E_\theta T(\theta, d)$ if $\xi < -3/2$.

Restricting n to $\{1, \ldots, N\}$, we can also restrict the domain of definition of g to $(0, 1]$ and replace the assumption (2.3) by

$$(2.26)\qquad g \text{ is bounded on } [a, 1] \text{ for all } 0 < a < 1.$$

Throughout the sequel we *let $\mathscr{C}$ denote the class of all nonnegative functions g on $(0, 1]$ satisfying* (2.26) *and* (2.4), (2.5) *for some* $\xi > -3/2$. Since $\xi > -3/2$, it follows from Theorem 1(ii) that for $g \in \mathscr{C}$, as $N \to \infty$,

$$(2.27)\quad NP_\theta\{U_n(g, N) \le \theta - d \text{ for some } n \le N\} = o\big((\log Nd^2)/d^2\big),$$

uniformly in $d \ge \alpha_N$ and $a_1 < \theta - d < \theta < a_2$.

We now discuss some background and motivation behind the confidence bounds $U_n(g, N)$ in the following.

EXAMPLE 1. Suppose that $Y_1, Y_2, \ldots$ are i.i.d. normal random variables with mean θ and variance 1. Here $\mu(\theta) = \theta$, $I(\theta, \lambda) = (\theta - \lambda)^2/2$ and $\hat{\theta}_n = S_n/n$. Thus, the confidence bound (2.6) reduces to

$$(2.28)\qquad U_n(g, N) = \hat{\theta}_n + \big(2n^{-1}g(n/N)\big)^{1/2}.$$

For an example of $g \in \mathscr{C}$, consider a nonnegative continuous function on $(0, 1]$ having the asymptotic expansion

$$(2.29)\qquad g(t) = \log t^{-1} - \tfrac{1}{2}\log\log t^{-1} - \tfrac{1}{2}\log 16\pi + o(1), \quad \text{as } t \to 0.$$

The asymptotic expansion (2.29) first arose in the following special bandit problem considered by Chernoff and Ray (1965) and by Chernoff (1967). Suppose that an experimenter can choose at each state n ($\le N$) between sampling from a normal population Π_1 with unknown mean θ and sampling from another normal

population Π_2 with known mean 0. Assuming a normal prior on θ, the Bayes procedure (to maximize the expected sum of N observations) samples from Π_1 until stage

$$T^* = \inf\{n \leq N: \hat{\theta}_n + a_{n,N} \leq 0\}, \tag{2.30}$$

and then takes the remaining $N - T^*$ observations from Π_2, where $a_{n,N}$ are positive constants. Writing

$$t = n/N, \qquad w(t) = (Y_1 + \cdots + Y_n)/N^{1/2}, \qquad \delta = \theta N^{1/2} \tag{2.31}$$

and treating $0 < t \leq 1$ as a continuous variable for large N, we can approximate the Bayes stopping time (2.30) by $N\tau(h)$, where $\tau(h) = \inf\{t \in (0,1]: w(t) + h(t) \leq 0\}$ is the optimal stopping rule in the following continuous-time stopping problem: Assuming a flat prior for the drift coefficient δ of a Wiener process $w(t)$, find the stopping rule $\tau \leq 1$ to maximize $\int_{-\infty}^{\infty} E_\delta(\delta\tau)\, d\delta$. Using an asymptotic analysis of the free boundary problem associated with this continuous-time optimal stopping problem, Chernoff and Ray (1965) found that as $t \downarrow 0$

$$h(t) = \left\{2t\left(\log t^{-1} - \tfrac{1}{2}\log\log t^{-1} - \tfrac{1}{2}\log 16\pi + o(1)\right)\right\}^{1/2}. \tag{2.32}$$

Therefore, letting

$$h(t)/t^{1/2} = (2g(t))^{1/2}, \tag{2.33}$$

$g(t)$ satisfies the asymptotic expansion (2.29). From (2.31) and (2.33), it follows that the $a_{n,N}$ in (2.30) can be approximated by the term $(2n^{-1}g(n/N))^{1/2}$ in (2.28). With this approximation, the Bayes procedure for the one-armed bandit problem can be described in the form:

(2.34) At stage $n+1$ sample from Π_1 or Π_2 according as $U_{T_n(1)}(g, N) > 0$ or $U_{T_n(1)}(g, N) \leq 0$.

Thus, the upper confidence bound (2.6) is an extension of (2.28) in the normal case to the general exponential family, and the allocation rule (1.8) proposed herein is an extension of (2.34) in the one-armed problem to the general case of k populations whose parameters are all unknown.

3. A class of asymptotically optimal adaptive allocation rules. Suppose that the populations Π_j, $j = 1, \ldots, k$, have densities $f(x; \theta_j)$ belonging to the exponential family (1.4) with respect to some nondegenerate measure ν. Let $A \subset \Theta$ be an open interval satisfying (2.1), and assume that $\theta_1, \ldots, \theta_k$ are known to belong to A. Let $g \in \mathscr{C}$. Based on successive observations $Y_{j,1}, \ldots, Y_{j,n}$ from Π_j, define the upper confidence bound $U_{j,n}$ $(= U_{j,n}(g, N))$ for θ_j by (2.6).

During the first k stages, take one observation from each population. For $n \geq k$, sample at state $n+1$ from a population Π_j with the largest upper confidence bound $U_{j,T_n(j)}$, where $T_n(j)$ is the number of observations sampled from Π_j up to stage n. This allocation rule, introduced earlier in Section 1, will be denoted by $\varphi_N(g)$.

Let $\theta = (\theta_1, \ldots, \theta_k)$ and $\theta^* = \max_{1 \le j \le k} \theta_j$. Making use of Theorem 1, we now prove the following theorem which can be applied to evaluate the regret

$$R_N(\theta) = \sum_{j:\, \theta_j < \theta^*} \big(\psi'(\theta^*) - \psi'(\theta_j)\big) E_\theta T_N(j), \tag{3.1}$$

introduced in Section 1, of the allocation rule $\varphi_N(g)$.

THEOREM 2. *Let α_N, β_N be positive numbers satisfying condition* (2.7). *Let $g \in \mathscr{C}$. For the allocation rule $\varphi_N(g)$, we have for every $j = 1, \ldots, k$,*

$$E_\theta T_N(j) \sim \Big(\log\big[N(\theta^* - \theta_j)^2\big]\Big)/I(\theta_j, \theta^*), \tag{3.2}$$

as $N \to \infty$, uniformly in $\theta \in A^k$ such that

$$\beta_N \ge \theta^* - \theta_j \ge \alpha_N. \tag{3.3}$$

PROOF. From the integral representation of $I(\theta, \lambda)$ in (1.5) and the assumption (2.1), it follows that as $\rho \to 0$,

$$I(\theta, \rho\theta + (1-\rho)\lambda)/I(\theta, \lambda) \to 1, \tag{3.4}$$

uniformly in $\theta \in A$ and $\lambda \in A$ with $\theta \ne \lambda$.

For brevity we shall use the abbreviation "unif." after a limiting relation to indicate that the convergence is uniform in $\theta \in A^k$ such that (3.3) holds. Take $\theta \in A^k$ satisfying (3.3), and let $d = \theta^* - \theta_j$. Take $0 < \rho < 1$. Letting $\theta^* = \theta_h$, we note that

$$\begin{aligned} E_\theta T_N(j) \le{}& NP_\theta\{U_{h,n} \le \theta_h - \rho d \text{ for some } n \le N\} \\ &+ E_\theta\big(T_N(j) I_{\{U_{h,n} > \theta_h - \rho d \text{ for all } n \le N\}}\big). \end{aligned} \tag{3.5}$$

Since θ_h and θ_j $(= \theta_h - d)$ belong to the interval A, $\theta_h - \rho d \in A$ and therefore we can apply (2.27) to obtain

$$NP_{\theta_h}\{U_{h,n} \le \theta_h - \rho d \text{ for some } n \le N\} = o\big(d^{-2}\log(Nd^2)\big) \quad \text{unif.} \tag{3.6}$$

Moreover,

$$\begin{aligned} &E_\theta\big(T_N(j) I_{\{U_{h,n} > \theta^* - \rho d \text{ for all } n \le N\}}\big) \\ &\quad \le E_{\theta_j}\big(\#\{n: 1 \le n \le N \text{ and } U_{j,n} \ge \theta^* - \rho d\}\big) \\ &\quad \sim \big(\log Nd^2\big)/I\big(\theta_j, \theta^* - \rho d\big) \quad \text{unif., by (2.11).} \end{aligned} \tag{3.7}$$

From (3.4)–(3.7), we obtain by letting $\rho \to 0$

$$E_\theta T_N(j) \le (1 + o(1))\big(\log Nd^2\big)/I(\theta_j, \theta^*) \quad \text{unif.} \tag{3.8}$$

Again fix $0 < \rho < 1$ and define $\bar\theta = \theta^* + \rho \min(d, r)$, where $r > 0$ is given by (2.1). Note that $a_1 < \bar\theta < a_2 + r$. We now show that as $N \to \infty$,

$$P_\theta\big\{T_N(j) \ge (1-\rho)\big(\log Nd^2\big)/I\big(\theta_j, \bar\theta\big)\big\} \to 1 \quad \text{unif.} \tag{3.9}$$

From (2.10), (2.7) and (2.25), it follows that

$$P_{\theta}(B_N) \to 1 \quad \text{unif., where}$$

$$B_n = \bigcap_{i=1}^{k} \{U_{i,m^*} < \bar{\theta} \text{ for all } N \geq m \geq N/(2k)\}. \tag{3.10}$$

Obviously, there exists $m < N$ such that $\varphi_N(g)$ samples at stage $m+1$ from a population Π_{i_m} with $T_m(i_m) \geq N/(2k)$, and therefore by the definition of the rule $\varphi_N(g)$,

$$U_{j,T_m(j)} \leq U_{i_m,T_m(i_m)} < \bar{\theta}, \quad \text{on } B_N. \tag{3.11}$$

Define $\tau_N = \inf\{n \leq N: U_{j,n} \leq \bar{\theta}\}$. By (3.11),

$$\tau_N \leq T_m(j) \leq T_N(j), \quad \text{on } B_N. \tag{3.12}$$

From (2.9), it follows that

$$P_{\theta_j}\{\tau_N \geq (1-\rho)(\log Nd^2)/I(\theta_j, \bar{\theta})\} \to 1 \quad \text{unif.} \tag{3.13}$$

From (3.10), (3.12) and (3.13), (3.9) follows. In view of (3.4), we obtain by letting $\rho \to 0$ in (3.9)

$$E_{\theta}T_N(j) \geq (1+o(1))(\log Nd^2)/I(\theta_j, \theta^*) \quad \text{unif.} \tag{3.14}$$

From (3.8) and (3.14), (3.2) follows. □

A slight modification of (3.7) in the preceding proof also gives

LEMMA 2. *Fix $\gamma > 0$ and $j \in \{1,\ldots,k\}$. Let $g \in \mathscr{C}$. For the allocation rule $\varphi_N(g)$,*

$$E_{\theta}T_N(j) = O(\log N), \tag{3.15}$$

uniformly in $\theta \in A^k$ such that $\theta^* - \theta_j \geq \gamma$.

The uniformity in θ in the asymptotic relations (3.2) and (3.15) is of particular interest to the asymptotic evaluation of the Bayes risk $\int_{A^k} R_N(\theta)\, dH(\theta)$ of $\varphi_N(g)$. Under certain assumptions on the prior distribution H, we can integrate (3.2) and (3.15) in evaluating the Bayes risk. We first introduce the following notation. For $j = 1,\ldots,k$, let $\theta_j = (\theta_1,\ldots,\theta_{j-1},\theta_{j+1},\ldots,\theta_k)$ (i.e., θ_j consists of all the components of θ except θ_j), and let $\theta_j^* = \max_{i \neq j}\theta_i$. For a prior distribution H of θ, let H_j denote the marginal distribution of the $(k-1)$-dimensional random vector θ_j, and let

$$H^{(j)}(\theta|\theta_j) = P_H\{\theta_j \leq \theta|\theta_j\}$$

denote the conditional distribution function of θ_j given θ_j. We use the notation P_H (and E_H) to denote probability (and expectation) under the distribution H. The asymptotic behavior of the Bayes risk of $\varphi_N(g)$ is given in

THEOREM 3. *Let $A \subset \Theta$ be an open interval satisfying* (2.1), *and let H be a probability distribution on A^k such that for some $\rho > 0$ and every $j = 1,\ldots,k$,*

the following three conditions are satisfied:

$$E_H|\theta_j| < \infty; \tag{3.16}$$

(3.17) *for every fixed* $\theta_j \in A^{k-1}$, $H^{(j)}(\theta|\theta_j)$ *has a positive continuous derivative* $h_j(\theta; \theta_j)$ *for* $\theta \in (\theta_j^* - \rho, \theta_j^* + \rho) \cap A$;

$$\int_{A^{k-1}} \sup_{\theta \in (\theta_j^* - \rho, \theta_j^*] \cap A} h_j(\theta; \theta_j)\, dH_j(\theta_j) < \infty. \tag{3.18}$$

(i) *Let* $g \in \mathscr{C}$. *Then for the allocation rule* $\varphi_N(g)$,

$$\int_{A^k} R_N(\theta)\, dH(\theta) \sim \left\{ \tfrac{1}{2} \sum_{j=1}^{k} \int_{A^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) \right\} (\log N)^2, \quad \text{as } N \to \infty. \tag{3.19}$$

(ii) *Assume, furthermore, that for every compact subset* B *of* A *and for every* j,

$$h_j(\theta; \theta_j)/h_j(\theta_j^*; \theta_j) \to 1, \quad \text{as } \theta \to \theta_j^*, \text{ uniformly in } \theta_j \in B^{k-1}. \tag{3.20}$$

Then, as $N \to \infty$,

$$\inf_{\varphi} \int_{A^k} R_N(\theta)\, dH(\theta) \sim \left\{ \tfrac{1}{2} \sum_{j=1}^{k} \int_{A^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) \right\} (\log N)^2, \tag{3.21}$$

where $\inf_\varepsilon$ *is taken over all adaptive allocation rules* φ.

EXAMPLE 2. We give here a simple example of a prior distribution on θ that satisfies the assumptions of Theorem 3. Suppose that $\theta_1, \ldots, \theta_k$ are i.i.d. with a common continuous positive density function q on the open interval A. Let $Q(t) = \int_{-\infty}^{t} q(\theta)\, d\theta$ denote the distribution function. Then $H(t_1, \ldots, t_k) = Q(t_1) \ldots Q(t_k)$. Moreover, for every j, θ_j^* has distribution function Q^{k-1}, and $h_j(\theta; \theta_j) = q(\theta)$. Assume that $\int |\theta| q(\theta)\, d\theta < \infty$ and that

$$\int_A \left\{ \sup_{\theta \in (\lambda - \rho, \lambda] \cap A} q(\theta) \right\} dQ^{k-1}(\lambda) < \infty, \quad \text{for some } \rho > 0. \tag{3.22}$$

Then conditions (3.16)–(3.18) and (3.20) are all satisfied. Moreover,

$$\sum_{j=1}^{k} \int_{A^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) = k \int_A q(\lambda)\, dQ^{k-1}(\lambda) = k(k-1) \int_A q^2(\lambda) Q^{k-2}(\lambda)\, d\lambda, \tag{3.23}$$

providing a simplification of the formulas (3.19) and (3.21).

Part (ii) of Theorem 3 will be proved in Section 4, where we develop certain lower bounds on the expected sample sizes from inferior populations for general adaptive allocation rules. As an application of Theorem 2, we now give

PROOF OF THEOREM 3(i). Fix $j \in \{1, \ldots, k\}$ and let $b_N = (\log N)^{1/2}$. From (1.5) and (2.1), it follows that as $\theta \to \lambda$,

$$(\psi'(\lambda) - \psi'(\theta))/I(\theta, \lambda) \sim 2(\lambda - \theta)^{-1}, \quad \text{uniformly in } \lambda \in A.$$

Hence, by (3.17), for every fixed $\theta_j \in A^{k-1}$,

$$\begin{aligned}
&\int_{\theta_j^* - b_N^{-1}}^{\theta_j^* - b_N N^{-1/2}} \big(\psi'(\theta_j^*) - \psi'(\theta)\big)\Big\{\log\Big[N(\theta_j^* - \theta)^2\Big]\Big\}/I(\theta, \theta_j^*)\, dH^{(j)}(\theta|\theta_j) \\
&\quad \sim 2\int_{\theta_j^* - b_N^{-1}}^{\theta_j^* - b_N N^{-1/2}} (\theta_j^* - \theta)^{-1}\Big\{\log\Big[N(\theta_j^* - \theta)^2\Big]\Big\} h_j(\theta; \theta_j)\, d\theta \\
&\quad \sim \tfrac{1}{2} h_j(\theta_j^*; \theta_j)(\log N)^2, \quad \text{as } N \to \infty,
\end{aligned}$$

and therefore by Theorem 2 and (3.18),

$$\begin{aligned}
&\int_{b_N N^{-1/2} \le \theta^* - \theta_j \le b_N^{-1}} \big(\psi'(\theta^*) - \psi'(\theta_j)\big) E_\theta T_N(j)\, dH(\theta) \\
&\quad \sim \int_{A^{k-1}} \left\{ \int_{\theta_j^* - b_N^{-1}}^{\theta_j^* - b_N N^{-1/2}} \frac{\big(\psi'(\theta_j^*) - \psi'(\theta)\big)\Big\{\log\Big[N(\theta_j^* - \theta)^2\Big]\Big\}}{I(\theta, \theta_j^*)} \right. \\
&\qquad\qquad \left. \times dH^{(j)}(\theta|\theta_j) \right\} dH_j(\theta_j) \qquad (3.24) \\
&\quad \sim \tfrac{1}{2}(\log N)^2 \int_{A^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j).
\end{aligned}$$

Since $T_N(j) \le N$, a similar argument gives

$$\begin{aligned}
&\int_{0 < \theta^* - \theta_j \le b_N N^{-1/2}} \big(\psi'(\theta^*) - \psi'(\theta_j)\big) E_\theta T_N(j)\, dH(\theta) \\
&\quad \le N \int_{A^{k-1}} \left\{ \int_{\theta_j^* - b_N N^{-1/2}}^{\theta_j^*} \big(\psi'(\theta_j^*) - \psi'(\theta)\big) h_j(\theta; \theta_j)\, d\theta \right\} dH_j(\theta_j) \qquad (3.25) \\
&\quad \sim \tfrac{1}{2} b_N^2 \int_{A^{k-1}} \psi''(\theta_j^*) h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) = O(\log N).
\end{aligned}$$

In view of (1.5) and (2.1), there exists $c > 0$ such that

$$|\psi'(\lambda) - \psi'(\theta)|/I(\theta, \lambda) \le c(\lambda - \theta)^{-1}$$

for all $\theta \in A$ and $\lambda \in A$. Therefore, using Theorem 2, (3.17) and (3.18), we obtain

$$\int_{b_N^{-1} < \theta^* - \theta_j < \rho} \big(\psi'(\theta^*) - \psi'(\theta_j)\big) E_\theta T_N(j)\, dH(\theta)$$

$$\sim \int_{A^{k-1}} \left\{ \int_{\theta_j^* - \rho}^{\theta_j^* - b_N^{-1}} \frac{\big(\psi'(\theta_j^*) - \psi'(\theta)\big)\big\{\log\big[N(\theta_j^* - \theta)^2\big]\big\}}{I(\theta, \theta_j^*)} \right.$$

(3.26)

$$\left. \times dH^{(j)}(\theta|\theta_j) \right\} dH_j(\theta_j)$$

$$\leq c b_N (\log N\rho^2) \int_{A^{k-1}} \left\{ \int_{\theta_j^* - \rho}^{\theta_j^*} h_j(\theta; \theta_j)\, d\theta \right\} dH_j(\theta_j)$$

$$= o\big((\log N)^2\big).$$

Moreover, by (3.16) and Lemma 2,

$$\int_{\theta^* - \theta_j \geq \rho} \big(\psi'(\theta^*) - \psi'(\theta_j)\big) E_\theta T_N(j)\, dH(\theta)$$

(3.27)

$$\leq \Big(\sup_{\theta \in A} \psi''(\theta)\Big) \Big(\sum_{i=1}^{k} E_H|\theta_i| \Big) \sup\big\{E_\theta T_N(j) \colon \theta^* - \theta_j \geq \rho\big\} = O(\log N).$$

From (3.1) and (3.24)–(3.27), (3.19) follows. □

The following variant of the rule $\varphi_N(g)$ can be used in situations where the upper confidence bounds $U_{j,n}$ are computed only with a certain specified degree of numerical accuracy, as will be illustrated in Section 5. The same argument as in Theorems 2 and 3 can be used to prove its asymptotic optimality. This is the content of

THEOREM 4. *With the same notation as in Theorem* 2, *let* ε_N *be a positive constant such that*

$$\varepsilon_N = O(N^{-1/2}), \quad \text{as } N \to \infty. \tag{3.28}$$

Let $\varphi_N^*(g)$ *be an allocation rule which samples at stage* $n+1$ $(n \geq k)$ *from a population* Π_{j^*} *such that*

$$U_{j^*, T_n(j^*)} \geq \max_{1 \leq j \leq k} U_{j, T_n(j)} - \varepsilon_N, \tag{3.29}$$

and which takes one observation from each population during the first k *stages. Then the conclusion* (3.2) *of Theorem* 2 *still holds for the allocation rule* $\varphi_N^*(g)$. *Moreover, if* H *is a probability distribution on* A^k *satisfying the assumptions* (3.16)–(3.18) *of Theorem* 3, *then the conclusion* (3.19) *of Theorem* 3 *still holds for the rule* $\varphi_N^*(g)$.

4. Asymptotic lower bounds on the sample sizes from inferior populations and the proof of Theorem 3(ii). In this section we give the proof of Theorem 3(ii) by developing certain asymptotic lower bounds on the expected sample sizes from the inferior populations of an adaptive allocation rule φ. To develop these asymptotic lower bounds, we introduce modifications of the ideas in Section 2 of Lai and Robbins (1985) who proved

LEMMA 3. *Suppose that the populations* Π_j, $j = 1, \ldots, k$, *have densities* $f(x; \theta_j)$ *belonging to the exponential family* (1.4). *Let* φ *be an allocation rule whose regret satisfies, as* $N \to \infty$, *the condition*

$$R_N(\boldsymbol{\theta}) = o(N^a), \quad \text{for every } a > 0 \text{ and } \boldsymbol{\theta} \in A^k. \tag{4.1}$$

Fix $j \in \{1, \ldots, k\}$. *Let* $\theta^* = \max_{1 \le i \le k} \theta_i$. *Then for every* $\boldsymbol{\theta} \in A^k$ *such that* $\theta_j < \theta^*$, *as* $N \to \infty$,

$$\liminf_{N \to \infty} E_{\boldsymbol{\theta}} T_N(j)/\log N \ge 1/I(\theta_j, \theta^*). \tag{4.2}$$

REMARK. By Theorem 2, the allocation rule $\varphi_N(g)$ satisfies (4.1) and attains the asymptotic lower bound in (4.2). Hence, $\varphi_N(g)$ is asymptotically optimal among all rules satisfying (4.1).

In Lemma 3, the parameter vector $\boldsymbol{\theta}$ is assumed to be fixed while we let $N \to \infty$. To prove Theorem 3(ii) on the asymptotic Bayes character of $\varphi_N(g)$, we need to develop lower bounds that are uniform over certain regions of $\boldsymbol{\theta}$ so that they can be integrated with respect to $\boldsymbol{\theta}$. This is the content of

LEMMA 4. *Fix* $j \in \{1, \ldots, k\}$, $0 < \zeta < 1$ *and* $0 < \gamma < 1$. *Let* $\boldsymbol{\theta}_j$ *denote the* $(k-1)$*-dimensional vector* $(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$ *and let* $\theta_j^* = \max\{\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k\}$. *For* $N \ge 1$, $d > 0$ *and* $\boldsymbol{\theta}_j \in A^{k-1}$ *let*

$$e_{N,d}(\boldsymbol{\theta}_j) = \inf\{E_{\boldsymbol{\theta}}[N - T_N(j)] : \theta_j \in A \text{ and } \theta_j^* + \tfrac{1}{2}\zeta d \le \theta_j \le \theta_j^* + \zeta d\}, \tag{4.3}$$

$$p_{N,d}(\boldsymbol{\theta}_j) = P_{\boldsymbol{\theta}}\{T_N(j) \le (1-\gamma)(\log Nd^2)/I(\theta_j, \theta_j^* + \zeta d)\}, \tag{4.4}$$

$$\text{with } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_k) \text{ and } \theta_j = \theta_j^* - d.$$

Let B *be a Borel subset of* A *such that* $\sup B < a_2$, *and let* G *be a finite measure on* B^{k-1}. *For* $b > 1$, *let* $\Phi_b^j(N, d)$ *denote the class of all allocation rules such that*

$$\int_{B^{k-1}} e_{N,d}(\boldsymbol{\theta}_j)\, dG(\boldsymbol{\theta}_j) \le d^{-2}(\log Nd^2)^b. \tag{4.5}$$

Then as $N \to \infty$ *and* $d \to 0$ *such that* $Nd^2 \to \infty$,

$$\sup_{\varphi \in \Phi_b^j(N,d)} \int_{B^{k-1}} p_{N,d}(\boldsymbol{\theta}_j)\, dG(\boldsymbol{\theta}_j) \to 0. \tag{4.6}$$

PROOF. To fix the ideas, assume that $j = 1$ and consider the case $\boldsymbol{\theta}_1 \in B^{k-1}$ with $\theta_1^* = \theta_2$. Since $E_{\boldsymbol{\theta}}[N - T_N(1)]$ is a continuous function of $\boldsymbol{\theta}$, the inf in (4.3)

is attained, so there exists ε (with $\zeta/2 \le \varepsilon \le \zeta$) for which

$$e_{N,d}(\theta_1) = E_\lambda(N - T_N(1)), \quad \text{where}$$
$$\lambda = (\lambda_1, \ldots, \lambda_k), \quad \text{with } \lambda_1 = \theta_2 + \varepsilon d,\ \lambda_i = \theta_i \text{ for } i \ne 1. \tag{4.7}$$

Note that since $\sup B < a_2$, $\lambda \in A^k$ if d is sufficiently small.

Let $\theta_1 = \theta_2 - d$, $t_N = (1-\gamma)(\log Nd^2)/I(\theta_1, \theta_2 + \zeta d)$. By (2.25),

$$t_N \le (1-\gamma)(\log Nd^2)/(c_1 d^2) \le \tfrac{1}{2}N, \quad \text{as } Nd^2 \to \infty. \tag{4.8}$$

By (4.7) and (4.8), for any allocation rule φ,

$$\begin{aligned} P_\lambda\{T_N(1) \le t_n\} &= P_\lambda\{N - T_N(1) \ge N - t_N\} \\ &\le (N - t_N)^{-1} E_\lambda(N - T_N(1)) \le 2N^{-1} e_{N,d}(\theta_1). \end{aligned} \tag{4.9}$$

Let $Y_1, Y_2, \ldots$ denote successive observations from Π_1, and let $\Lambda_n = Z_1 + \cdots + Z_n$, where

$$Z_i = \log\{f(Y_i; \theta_1)/f(Y_i; \lambda_1)\} = (\theta_1 - \lambda_1)Y_i - (\psi(\theta_1) - \psi(\lambda_1)). \tag{4.10}$$

Since $\theta_i = \lambda_i$ for $2 \le i \le k$, it then follows that

$$\begin{aligned} &P_\lambda\left\{T_N(1) \le t_N,\ \Lambda_{T_N(1)} \le \left(1 - \tfrac{1}{2}\gamma\right)\log Nd^2\right\} \\ &= \int_{\{T_N(1) \le t_N,\ \Lambda_{T_N(1)} \le (1-\frac{1}{2}\gamma)\log Nd^2\}} \exp\left(-\Lambda_{T_N(1)}\right) dP_\theta \\ &\ge (Nd^2)^{-(1-\gamma/2)} P_\theta\left\{T_N(1) \le t_N,\ \Lambda_{T_N(1)} \le \left(1 - \tfrac{1}{2}\gamma\right)\log Nd^2\right\}. \end{aligned} \tag{4.11}$$

By (4.9) and (4.11),

$$P_\theta\left\{T_N(1) \le t_N,\ \Lambda_{T_N(1)} \le \left(1 - \tfrac{1}{2}\gamma\right)\log Nd^2\right\} \le 2d^2(Nd^2)^{-\gamma/2} e_{N,d}(\theta_1). \tag{4.12}$$

Since $E_\theta Y_1 = \psi'(\theta_1)$ and $\sup_{\theta \in A^k} E_\theta[Y_1 - \psi'(\theta_1)]^4 < \infty$ by (2.1) [cf. Lai (1985)], there exists $C > 0$ such that

$$E_\theta\left[\sum_1^n (Y_i - \psi'(\theta_1))\right]^4 \le Cn^2, \quad \text{for all } \theta \in A^k \text{ and } n \ge 1. \tag{4.13}$$

Noting that $\Lambda_n = (\theta_1 - \lambda_1)\Sigma_1^n(Y_i - \psi'(\theta_1)) + nI(\theta_1, \lambda_1)$ by (4.10) and that $t_N I(\theta_1, \lambda_1) \le (1-\gamma)\log Nd^2$, we obtain from (4.13)

$$\begin{aligned} &P_\theta\left\{\max_{n \le t_N} \Lambda_n > \left(1 - \tfrac{1}{2}\gamma\right)\log Nd^2\right\} \\ &\le P_\theta\left\{\max_{n \le t_N} (\theta_1 - \lambda_1)\sum_1^n (Y_i - \psi'(\theta_1)) > \tfrac{1}{2}\gamma \log Nd^2\right\} \\ &\le Ct_N^2\left\{\tfrac{1}{2}\gamma(\log Nd^2)/(\theta_2 + \varepsilon d - \theta_1)\right\}^{-4} \to 0, \end{aligned} \tag{4.14}$$

as $N \to \infty$ and $d \to 0$ such that $Nd^2 \to \infty$, uniformly in $\theta \in A^k$ with $\theta^* = \theta_2 =$

$\theta_1 + d$, in view of (2.25). Noting that

$$\begin{aligned} p_{N,d}(\theta_1) &= P_\theta\{T_N(1) \le t_N\} \\ &\le P_\theta\Big\{T_N(1) \le t_N,\ \Lambda_{T_N(1)} \le \big(1 - \tfrac{1}{2}\gamma\big)\log Nd^2\Big\} \\ &\quad + P_\theta\Big\{\max_{n \le t_N} \Lambda_n > \big(1 - \tfrac{1}{2}\gamma\big)\log Nd^2\Big\}, \end{aligned}$$

the desired conclusion (4.6) follows from (4.5), (4.12) and (4.14). □

PROOF OF THEOREM 3(ii). Take any bounded closed interval $B = [b_1, b_2] \subset A$ and let $0 < \gamma < 1$. In view of Theorem 3(i), we need only show that for all large N,

$$\inf_\varphi \int_{A^k} R_N(\theta)\, dH(\theta) \ge \tfrac{1}{2}(1 - \gamma)^3 \Big\{ \sum_{j=1}^k \int_{B^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) \Big\} (\log N)^2. \tag{4.15}$$

To prove (4.15), we can obviously restrict to allocation rules φ for which

$$\int_{A^k} R_N(\theta)\, dH(\theta) \le C(\log N)^2, \tag{4.16}$$

where $C > \frac{1}{2}\sum_{j=1}^k \int_{A^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j)$.

Fix $j \in \{1, \ldots, k\}$. Let φ be an allocation rule satisfying (4.16). Let $0 < \zeta$ (sufficiently small, as specified later). Note that if $\theta_j > \theta_j^*$, then

$$R_N(\theta) \ge \big(\psi'(\theta_j) - \psi'(\theta_j^*)\big)\big(N - E_\theta T_N(j)\big).$$

Hence, defining $e_{N,d}(\theta_j)$ as in (4.3), we have by (4.16) that for $0 < d < \min\{a_2 - b_2, \rho\}$,

$$\begin{aligned} &C(\log N)^2 \\ &\quad \ge \int_{B^{k-1}} \Big\{ \int_{\theta_j^* + \frac{1}{2}\zeta d}^{\theta_j^* + \zeta d} \big(\psi'(\theta) - \psi'(\theta_j^*)\big)\big(N - E_\theta T_N(j)\big)\, dH^{(j)}(\theta|\theta_j) \Big\}\, dH_j(\theta_j) \\ &\quad \ge \int_{B^{k-1}} e_{N,d}(\theta_j) \Big\{ \int_{\theta_j^* + \frac{1}{2}\zeta d}^{\theta_j^* + \zeta d} \big(\psi'(\theta) - \psi'(\theta_j^*)\big) h_j(\theta; \theta_j)\, d\theta \Big\}\, dH_j(\theta_j) \\ &\quad \sim \tfrac{3}{8}\zeta^2 d^2 \int_{B^{k-1}} e_{N,d}(\theta_j) \psi''(\theta_j^*) h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j), \end{aligned}$$

as $d \to 0$, by (2.1) and (3.20). Therefore, if d is sufficiently small,

$$\begin{aligned} &C(\log N)^2 \ge \tfrac{1}{4}\zeta^2 d^2 \Big(\inf_{\theta \in B} \psi''(\theta)\Big) \int_{B^{k-1}} e_{N,d}(\theta_j)\, dG(\theta_j), \quad \text{where} \\ &\qquad dG(\theta_j) = h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j). \end{aligned} \tag{4.17}$$

Defining $\Phi_b^j(N, d)$ as in Lemma 4, it then follows from (4.17) that there exist positive numbers d_0 and N_0 such that

$$\varphi \in \Phi_\zeta^j(N, d), \quad \text{for all } N \ge N_0 \text{ and } d_0 \ge d \ge N^{-(1-\gamma)/2}. \tag{4.18}$$

Hence, by Lemma 4, as $N \to \infty$ and $d \to 0$ such that $d \geq N^{-(1-\gamma)/2}$,

$$\int_{B^{k-1}} p_{N,d}(\theta_j)\, dG(\theta_j) \to 0, \tag{4.19}$$

the convergence being uniform in all allocation rules that satisfy (4.18), where $p_{N,d}(\theta_j)$ is defined in (4.4).

Since $dG(\theta_j) = h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j)$, it follows from (4.19) that as $N \to \infty$,

$$\begin{aligned} &\int_{N^{-(1-\gamma)/2}}^{(\log N)^{-1}} t^{-1}(\log Nt^2)\left\{\int_{B^{k-1}} p_{N,t}(\theta_j) h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j)\right\} dt \\ &\qquad = o\big((\log N)^2\big). \end{aligned} \tag{4.20}$$

In view of (2.1) and (3.20), we can choose $0 < t_0 < \rho$ and ζ sufficiently small so that

$$\begin{aligned} &h_j(\theta_j^* - t; \theta_j)\big\{\psi'(\theta_j^*) - \psi'(\theta_j^* - t)\big\}/I(\theta_j^* - t, \theta_j^* + \zeta t) \\ &\qquad \geq (1-\gamma)\big\{2t^{-1}h_j(\theta_j^*; \theta_j)\big\}, \quad \text{for all } \theta_j \in B^{k-1} \text{ and } 0 < t \leq t_0. \end{aligned} \tag{4.21}$$

For $\theta \in B^k$ with $\theta_j < \theta_j^*$, writing $\theta_j = \theta_j^* - t$, we have from (4.4) that

$$E_\theta T_N(j) \geq \big\{1 - p_{N,t}(\theta_j)\big\}\big\{(1-\gamma)(\log Nt^2)/I(\theta_j^* - t; \theta_j^* + \zeta t)\big\},$$

and therefore by (4.21) for $0 < t \leq t_0$,

$$\begin{aligned} &h_j(\theta_j^* - t; \theta_j)\big\{\psi'(\theta_j^*) - \psi'(\theta_j^* - t)\big\}E_\theta T_N(j) \\ &\qquad \geq 2(1-\gamma)^2 t^{-1} h_j(\theta_j^*; \theta_j)(\log Nt^2)\big\{1 - p_{N,t}(\theta_j)\big\}. \end{aligned} \tag{4.22}$$

In view of (3.1), the desired conclusion (4.15) follows from

$$\begin{aligned} &\int_{\theta_j < \theta^*} \big(\psi'(\theta^*) - \psi'(\theta_j)\big)E_\theta T_N(j)\, dH(\theta) \\ &\geq \int_{B^{k-1}} \left\{\int_{N^{-(1-\gamma)/2}}^{(\log N)^{-1}} \big(\psi'(\theta_j^*) - \psi'(\theta_j^* - t)\big)E_\theta T_N(j) h_j(\theta_j^* - t; \theta_j)\, dt\right\} dH_j(\theta_j) \\ &\geq 2(1-\gamma)^2 \int_{B^{k-1}} h_j(\theta_j^*; \theta_j)\left\{\int_{N^{-(1-\gamma)/2}}^{(\log N)^{-1}} t^{-1}(\log Nt^2)\right. \\ &\qquad\qquad \left.\times\big[1 - p_{N,t}(\theta_j)\big]\, dt\right\} dH_j(\theta_j) \quad \text{[by (4.22)]} \\ &\sim \tfrac{1}{2}(1-\gamma)^2(1-\gamma^2)(\log N)^2 \int_{B^{k-1}} h_j(\theta_j^*; \theta_j)\, dH_j(\theta_j) \quad \text{[by (4.20)]}, \end{aligned}$$

for every $j = 1, \ldots, k$. □

5. Some numerical results and discussion. In this section we report some simulation results on the asymptotically optimal adaptive allocation rules introduced in Section 3. We shall use the following choice of the function $g \in \mathscr{C}$. Let h be the optimal stopping boundary for the continuous-time Bayes stopping

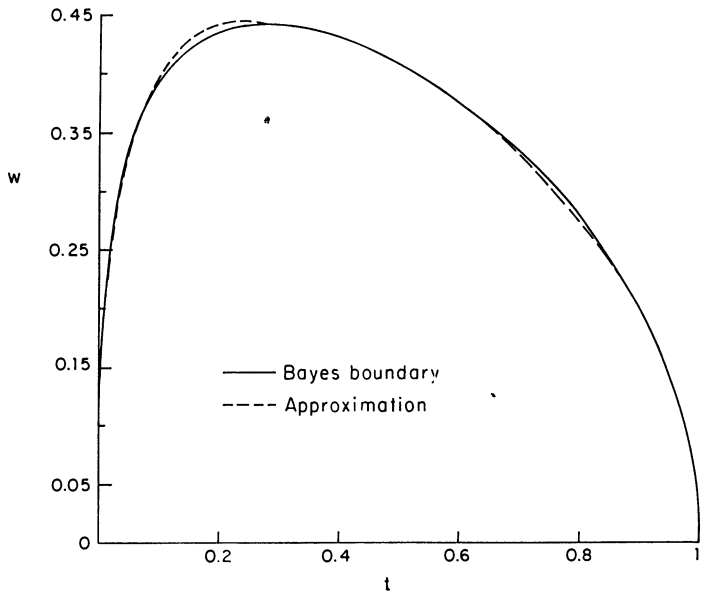

FIG. 1.

problem associated with the Chernoff–Ray one-armed bandit problem described in Example 1. Chernoff and Petkau (1986) have recently tabulated this boundary, which is shown in Figure 1. As shown by Chernoff and Ray (1965), h has the asymptotic expansion (2.32) as $t \downarrow 0$ and

$$h(t) = (t^{-1} - 1)^{1/2}\{0.63883 - 0.40258(t^{-1} - 1) + \cdots\}, \quad \text{as } t \uparrow 1. \tag{5.1}$$

The asymptotic expansions (2.32) and (5.1) together with some simple curve fitting suggest the following approximation to h:

$$\begin{aligned} h_0(t) &= (t^{-1} - 1)^{1/2}\{0.63883 - 0.40258(t^{-1} - 1)\}, \quad \text{if } 0.86 < t \le 1 \\ &= -0.5759t^2 + 0.2987t + 0.4034, \quad \text{if } 0.28 < t \le 0.86 \\ &= -1.58137t + 1.53343t^{1/2} + 0.073271, \quad \text{if } 0.01 < t \le 0.28 \\ &= \{t[2\log t^{-1} - \log\log t^{-1} - \log 16\pi + 0.99232\exp(-0.03812t^{-1/2})]\}^{1/2}, \\ &\qquad \text{if } 0 < t \le 0.01. \end{aligned} \tag{5.2}$$

Figure 1 shows that the Bayes boundary h is closely approximated by h_0. Define

$$g_0(t) = h_0^2(t)/2t. \tag{5.3}$$

Then $g_0 \in \mathscr{C}$.

Table 1 considers the case of $k = 3$ normal populations with means $\theta_1 = 0$, $\theta_2 < 0$, $\theta_3 < 0$, and common variance 1, for horizons $N = 100$ and $N = 2500$. Here Π_2 and Π_3 are the inferior populations, and Table 1 gives the normalized expected sample sizes

$$e_2 = N^{-1}ET_2, \qquad e_3 = N^{-1}ET_3, \tag{5.4}$$

from these inferior populations, for the allocation rule $\varphi_N(g_0)$ introduced in

TABLE 1

Normal three-armed bandits. Normalized expected sample sizes e_2, e_3 from inferior populations Π_2, Π_3 and normalized regret r of the rule $\varphi_N(g_0)$. Also given are the Bayes lower bounds e_2^*, e_3^* derived from the one-armed bandit problem. Each result is based on 1000 simulations, largest mean $= \theta_1 = 0$.

		$N = 100$					$N = 2500$				
		One-armed lower bounds		The rule $\varphi_N(g_0)$			One-armed lower bounds		The rule $\varphi_N(g_0)$		
δ_2 $(= N^{1/2}\theta_2)$	δ_3 $(= N^{1/2}\theta_3)$	e_2^*	e_3^*	e_2	e_3	r	e_2^*	e_3^*	e_2	e_3	r
−0.5	−1	0.37	0.30	0.33	0.27	0.43	0.36	0.27	0.34	0.27	0.44
−1	−2	0.30	0.18	0.33	0.20	0.73	0.27	0.18	0.31	0.21	0.73
−1	−5	0.30	0.07	0.37	0.09	0.81	0.27	0.06	0.37	0.08	0.76
−1	−10	0.30	0.03	0.38	0.04	0.77	0.27	0.03	0.38	0.03	0.67
−2	−5	0.18	0.07	0.26	0.10	1.03	0.18	0.06	0.29	0.08	1.01
−3	−10	0.12	0.03	0.21	0.04	1.04	0.12	0.03	0.21	0.03	0.94
−5	−10	0.07	0.03	0.12	0.04	1.02	0.06	0.03	0.12	0.03	0.94
−10	−15	0.03	0.02	0.04	0.03	0.83	0.03	0.014	0.04	0.018	0.73
−20	−30	0.01	0.01	0.02	0.01	0.73	0.009	0.005	0.012	0.007	0.44
−40	−40	0.01	0.01	0.01	0.01	0.84	0.003	0.003	0.004	0.004	0.32

Section 3. Also given in Table 1 is the normalized regret,

$$r = N^{-1/2}R_N(\theta), \tag{5.5}$$

of the rule $\varphi_N(g_0)$. For fixed $\delta_2 = N^{1/2}\theta_2$ and $\delta_2 = N^{1/2}\theta_3$, note the relative constancy of r, e_2, e_3 as N varies from 100 to 2500, when δ_2 and δ_3 are not too large, in agreement with the Wiener process approximation (2.31).

To develop some benchmark against which we can compare ET_j, $j = 2, 3$, of the rule $\varphi_N(g_0)$, we consider the fictitious case in which the mean θ_1 of Π_1 is known to be 0, the mean θ_j of Π_j is unknown but has a normal prior distribution and the mean of the remaining population is negative and known. Thus, we are in the setting of the one-armed bandit problem involving the normal population Π_j, and as indicated in Example 1, the optimal allocation rule samples from Π_j until stage T_j^* and then takes the remaining $N - T_j^*$ observations from Π_1. In view of the Wiener process approximation (2.31), we have the following approximation to T_j^*:

$$T_j^* \doteq \inf\left\{n \le N: \sum_{i=1}^{n} Y_{j,i} \le -N^{1/2}h(n/N)\right\}. \tag{5.6}$$

Using (5.6), we computed by simulations the normalized expected sample size,

$$e_j^* = N^{-1}ET_j^*, \tag{5.7}$$

for $j = 2, 3$ and the results are given in Table 1. Table 1 shows that e_2 and e_3 of the rule $\varphi_N(g_0)$ for the three-armed bandit problem compares quite well with the "Bayes lower bounds" e_2^* and e_3^* derived from the Chernoff–Ray one-armed bandit problem.

We consider the case of k Bernoulli populations with density $P\{Y_{j,n} = 1\} = p_j = 1 - P\{Y_{j,n} = 0\}$, $j = 1, \ldots, k$. The natural parameter is

$$\theta = \log[p/(1-p)]. \tag{5.8}$$

In terms of p, the Kullback–Leibler information number can be written as

$$I(\tilde{p}, p) = \tilde{p}\log(\tilde{p}/p) + (1-\tilde{p})\log[(1-\tilde{p})/(1-p)]. \tag{5.9}$$

Note that $\psi(\theta) = \log(1-p) = -\log(1+e^{\theta})$, so condition (2.1) does not hold if A is the entire parameter space Θ. We will therefore assume that A corresponds to the interval $\{p: \underline{p} < p < \bar{p}\}$ with $0 < \underline{p} < \bar{p} < 1$, so the maximum likelihood estimate of p_j based on n observations from Π_j is given by

$$\hat{p}_{j,n} = \max\left\{\underline{p}, \min\left(n^{-1}\sum_{i=1}^{n} Y_{j,i}, \bar{p}\right)\right\}. \tag{5.10}$$

We now describe a simple recursive algorithm for finding a population at every stage whose upper confidence bound differs by no more than ε_N from the largest of the k upper confidence bounds at that stage. This algorithm facilitates the implementation of the allocation rule $\varphi_N^*(g)$, introduced in Theorem 4, for which we shall assume that $N > k$ and $g(t) > 0$ if $t < 1$. First, note that the upper confidence bound $U_{j,T_n(j)}$ for population Π_j at stage n, if finite, is the solution of the equation,

$$\begin{aligned} &h_{j,n}(x) = 0, \quad \text{where} \\ &h_{j,n}(x) = I\big(\hat{p}_{j,T_n(j)}, x\big) - \big(T_n(j)\big)^{-1} g\big(T_n(j)/N\big), \\ &\qquad\qquad \text{if } \hat{p}_{j,T_n(j)} \le x \le \bar{p}, \text{ and} \\ &h_{j,n}(x) = h_{j,n}\big(\hat{p}_{j,T_n(j)}\big) \,(<0), \quad \text{if } \underline{p} \le x \le \hat{p}_{j,T_n(j)}. \end{aligned} \tag{5.11}$$

Fix stage n $(\ge k)$ and write $\hat{p}_j = \hat{p}_{j,T_n(j)}$, $U_j = U_{j,T_n(j)}$. Note that since $I(\hat{p}_j, x)$ is increasing in $x \ge \hat{p}_j$,

$$x \gtrless U_j \quad \text{according as } h_{j,n}(x) \gtrless 0. \tag{5.12}$$

Let $c_{k-1} = \bar{p}$, $d_{k-1} = \underline{p}$. Suppose that two numbers $(\bar{p} \ge) c_{n-1} > d_{n-1} (\ge \underline{p})$ have been determined at the end of stage $n-1$ satisfying

$$h_{j,n-1}(d_{n-1}) < 0 \quad \text{for more than one } j; \tag{5.13}$$

$$\text{either } c_{n-1} = \bar{p} \quad \text{or} \quad h_{j,n-1}(c_{n-1}) < 0 \text{ for at most one } j. \tag{5.14}$$

In the case $h_{j,n-1}(c_{n-1}) < 0$ for exactly one j, this j coincides with j^*, where Π_{j^*} is the population from which $\varphi_N^*(g)$ samples at stage n. Note that

$$h_{j,n} = h_{j,n-1}, \quad \text{for all } j \ne j^*. \tag{5.15}$$

In view of (5.12) and (5.15), we use the following bisection-type iterative scheme at stage n to decide from which population $\varphi_N^*(g)$ samples at stage $n+1$.

Let $J(x) = \{j: h_{j,n}(x) < 0\}$, $\hat{p}_{[1]} = \max_{1 \le j \le k} \hat{p}_j$. Define $\hat{p}_{[2]} = \underline{p}$ if $\hat{p}_j = \bar{p}$ for all j, otherwise let $\hat{p}_{[2]} = \max\{\hat{p}_j: j \ne [1] \text{ and } \hat{p}_j < \bar{p}\}$. Since $h_{j,n}(\hat{p}_j) < 0$,

$\#J(\hat{p}_{[2]}) \geq 2$. Initialize the iterative scheme by setting $a_0 = c_{n-1}$, $b_0 = \max(d_{n-1}, \hat{p}_{[1]})$. At $x = a_0$ or b_0, if (i) $\#J(x) = 1$, or (ii) $J^* \triangleq \{j: h_{j,n}(x) = 0\} \neq \varnothing$ and $h_{j,n}(x) > 0$ for $j \notin J^*$, stop the iteration and set $c_n = x$. Otherwise, if (iii) $\max(a_0, b_0) = \bar{p}$ and $J(\bar{p}) \neq \varnothing$, stop the iteration and set $c_n = \bar{p}$. We can set $d_n = \hat{p}_{[2]}$ since $\#J(\hat{p}_{[2]}) \geq 2$, but we let $d_n = d_{n-1}$ instead if $d_{n-1} > \hat{p}_{[2]}$ and $\#J(d_{n-1}) \geq 2$.

Now suppose that (i), (ii), (iii) all fail to hold at a_0 and b_0. Then it follows from (5.13)–(5.15) that $\#J(b_0) \geq 2$ and $\#J(a_0) = 0$, implying that $b_0 < a_0$. Set $x_0 = (a_0 + b_0)/2$. In general, for the ith iterate $x_i = (a_i + b_i)/2$, if $\#J(x_i) \leq 1$, set $a_{i+1} = x_i$, $b_{i+1} = b_i$. On the other hand, if $\#J(x_i) \geq 2$, set $b_{i+1} = x_i$, $a_{i+1} = a_i$. This bisection-type scheme is stopped after the ith iteration, whereupon we set $c_n = a_{i+1}$ and $d_n = b_{i+1}$, if one of the following holds:

(5.16a) $$\#J(x_i) = 1 \quad \left(\text{say, } J(x_i) = \{\tilde{j}\}\right);$$

(5.16b) $$J_i^* \triangleq \{j: h_{j,n}(x_i) = 0\} \neq \varnothing \quad \text{and} \quad h_{j,n}(x_i) > 0 \text{ for } j \notin J_i^*;$$

(5.16c) $$a_{i+1} - b_{i+1} < \varepsilon_N \quad \text{but (5.16a) and (5.16b) both fail.}$$

At stage $n + 1$, sample from $\Pi_{\tilde{j}}$ in case (5.16a), and from Π_j having the largest $\hat{p}_j$ with $j \in J_i^*$ in case (5.16b), but with $j \in J(b_{i+1})$ in case (5.16c), randomizing if there are ties.

The above recursive algorithm can be applied in general and is not restricted only to the Bernoulli case. Instead of direct numerical solution of (5.11) with a prescribed accuracy to compare the k upper confidence bounds, it evaluates the number of elements of the set $J(x_i)$ at each iterate x_i; this is particularly simple since $J(x_i) \subset J(b_i)$. Making use of this algorithm, we have obtained the numerical results in Tables 2 and 3 on the performance of the allocation rule $\varphi_N^*(g_0)$ for Bernoulli k-armed bandits.

TABLE 2

Bernoulli three-armed bandits. Normalized expected sample sizes e_2, e_3 [defined in (5.4)] from inferior populations Π_2, Π_3 and normalized regret $r = 2N^{-1/2}R_N(\theta)$ of the rule $\varphi_N(g_0)$. Largest mean = $p_1 = \frac{1}{2}$. Each result is based on 1000 simulations.

		$N = 100$					$N = 2500$				
δ_2[a]	δ_3[a]	p_2	p_3	e_2	e_3	r	p_2	p_3	e_2	e_3	r
−0.5	−1	0.475	0.450	0.32	0.28	0.44	0.495	0.490	0.35	0.25	0.43
−1	−2	0.450	0.401	0.30	0.21	0.72	0.490	0.480	0.30	0.20	0.70
−1	−5	0.450	0.269	0.35	0.09	0.79	0.490	0.450	0.38	0.08	0.79
−1	−10	0.450	0.119	0.37	0.05	0.73	0.490	0.401	0.37	0.03	0.65
−2	−5	0.401	0.269	0.28	0.10	1.04	0.480	0.450	0.27	0.08	0.95
−3	−10	0.354	0.119	0.22	0.05	1.01	0.470	0.401	0.23	0.03	1.02
−5	−10	0.269	0.119	0.12	0.05	0.96	0.450	0.401	0.13	0.04	1.03
−10	−15	0.119	0.047	0.05	0.04	0.78	0.401	0.354	0.04	0.019	0.67
−20	−30	0.018	0.002	0.04	0.03	0.68	0.310	0.231	0.013	0.007	0.44
−40	−40	0.0003	0.0003	0.03	0.03	0.66	0.168	0.168	0.005	0.005	0.34

[a] See (5.17).

TABLE 3
Bernoulli two-armed bandits with $N = 50$.

(a) The Bayes reward (5.18)

Beta (1,1) prior		Beta (2,6) prior		Beta (4,4) prior		Beta (6,2) prior	
$\varphi_N^*(g_0)$	Bayes rule	$\varphi_N^*(g_0)$	Bayes rule	$\varphi_N^*(g_0)$	Bayes rule	$\varphi_N^*(g_0)$	Bayes rule
0.634	0.641	0.300	0.301	0.558	0.564	0.805	0.807

(b) Normalized reward $N^{-1}E_{\mathbf{p}}S_N$ at $\mathbf{p} = (p_1, p_2)$

p_1	p_2	$\varphi_N^*(g_0)$	Beta (1,1) Bayes rule	Beta (2,6) Bayes rule	Beta (4,4) Bayes rule	Beta (6,2) Bayes rule
0.6	0.5	0.564	0.564	0.560	0.564	0.564
0.9	0.7	0.868	0.864	0.839	0.858	0.871
0.5	0.3	0.453	0.445	0.453	0.455	0.447

Table 2 considers the case of $k = 3$ Bernoulli populations with means $p_1 = \frac{1}{2}$, $p_2 < \frac{1}{2}$, $p_3 < \frac{1}{2}$, for horizons $N = 100$ and $N = 2500$. For comparison with the normal case in Table 1, introduce the natural parameters θ_j as in (5.8) and define

$$\delta_j = \tfrac{1}{2}N^{1/2}(\theta_j - \theta_1), \qquad j = 2, 3, \tag{5.17}$$

the factor $\frac{1}{2}$ being the standard deviation of a Bernoulli random variable with mean $\frac{1}{2}$ $(= p_1)$. We assume that the p_j are known to lie between 0.01 $(= \underline{p})$ and 0.99 $(= \bar{p})$, and use the truncated sample proportions $\hat{p}_{j,n}$ defined in (5.10). Table 2 considers the rule $\varphi_N^*(g_0)$ described above with $\varepsilon_N = 0.05N^{-1/2}$. It shows that the performance of the rule in the present case of Bernoulli populations resembles that in Table 1 for the case of normal populations.

Table 3 considers the horizon $N = 50$ and compares the above rule $\varphi_N^*(g_0)$ (where $\varepsilon_N = 0.05N^{-1/2}$, as before) with certain exact Bayes rules for the case of $k = 2$ Bernoulli populations. These Bayes rules were computed by Wahrenberger, Antle and Klimko (1977), assuming independent and identical Beta (α, β) priors on the two means p_1, p_2, for the values of (α, β) listed in Table 3. Wahrenberger, Antle and Klimko also computed by simulation the normalized reward $N^{-1}E_{\mathbf{p}}S_N$ at certain values of $\mathbf{p} = (p_1, p_2)$ and the Bayes reward,

$$N^{-1}\int_0^1\int_0^1 (E_{\mathbf{p}}S_N)\prod_{i=1}^{2}\left\{p_i^{\alpha-1}(1-p_i)^{\beta-1}/B(\alpha,\beta)\right\}dp_1\,dp_2, \tag{5.18}$$

of these Bayes rules. Their results on these rules are shown in Table 3 for comparison with the rule $\varphi_N^*(g_0)$. Table 3, in which each result is based on 5000 simulation runs, shows that $\varphi_N^*(g_0)$ closely resembles the Bayes rule for each of the priors.

The numerical results above and the asymptotic results of Theorems 2–4 show that the allocation rules based on upper confidence bounds of the k

population parameters have nearly optimal Bayes and frequentist properties. As pointed out in Example 1, these upper confidence bounds first arose in the Chernoff–Ray one-armed bandit problem in which $x_1, \ldots, x_N$ are sampled sequentially from either a normal population with unknown mean or another normal population with known mean, assuming common and known variance for both populations. Making use of the boundary crossing theory developed in Lai (1985), we have extended these upper confidence bounds in Section 2 from the normal case to the general exponential family, while Sections 3 and 4 show that the allocation rule which chooses the population with the largest upper confidence bound is asymptotically optimal as the horizon N approaches ∞, from both the Bayesian and frequentist viewpoints. An important feature of the asymptotic results in Theorems 1 and 2 is their uniformity over a wide range of parameter values, so that they can be integrated with respect to a broad class of prior distributions in evaluating the Bayes risks.

In the past decade, considerable progress was made in a different version of the multi-armed bandit problem. In this version, instead of assuming a finite horizon N, one assumes a discount factor $0 < \beta < 1$, and considers the problem of maximizing the expected value of the infinite discounted sum $\sum_{i=1}^{\infty} \beta^i x_i$. Under the assumption of independent prior distributions Q_j on θ_j, $j = 1 \ldots k$, it has been shown by Gittins (1979) that the optimal allocation rule which maximizes

$$\int \cdots \int E_{\theta}\left(\sum_{i=1}^{\infty} \beta^i x_i \right) dQ_1(\theta_1) \ldots dQ_k(\theta_k) \tag{5.19}$$

is to choose a population at every stage that has the largest "dynamic allocation index." This index (also called the Gittins index) of population Π_j at stage n can be computed by solving an optimal stopping problem that involves only the posterior distribution of θ_j given the observations $Y_{j,1}, \ldots, Y_{j,T_n(j)}$ of Π_j. We have recently shown that as $\beta \to 1$ such that $(1 - \beta)T_n(j) \to 0$, the Gittins index can be approximated by $\psi'(U_{j,T_n(j)})$, where $U_{j,r} = U_{j,r}(g, N)$ is the same as that introduced in Section 3 but with $N = (1 - \beta)^{-1}$. Moreover, the allocation rules $\varphi_N(g)$, $g \in \mathscr{G}$, described in Section 3 provide asymptotically optimal solutions not only of the finite-horizon problem discussed herein, but also of the discounted problem (5.19) as $\beta \to 1$, upon setting $N = (1 - \beta)^{-1}$. The details for the discounted problem are too lengthy to be included here and will be presented elsewhere.

Acknowledgment. The author wishes to express his gratitude to Fridrik Baldursson for helpful discussions and valuable assistance in the numerical work.

REFERENCES

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

CHERNOFF, H. (1967). Sequential models for clinical trials. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 805–812. Univ. California Press.

CHERNOFF, H. and PETKAU, J. (1986). Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Statist. Comput.* **7** 46–59.

CHERNOFF, H. and RAY, S. N. (1965). A Bayes sequential sampling inspection plan. *Ann. Math. Statist.* **36** 1387–1407.

FABIUS, J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41** 1906–1916.

FELDMAN, D. (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33** 847–856.

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177.

KUMAR, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM J. Control Optim.* **23** 329–380.

LAI, T. L. (1985). Boundary crossing problems for sample means. *Ann. Probab.* To appear.

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6** 4–22.

WAHRENBERGER, D. L., ANTLE, C. E. and KLIMKO, L. A. (1977). Bayesian rules for the two-armed bandit problem. *Biometrika* **64** 172–174.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305