ASYMPTOTIC BAYES ANALYSIS FOR THE FINITE HORIZON ONE ARMED BANDIT PROBLEM [1]

Apostolos N. Burnetas[2]

and

Michael N. Katehakis[3]

ABSTRACT

The multi-armed bandit probem is often taken as a basic model for the trade-off between the exploration - utilization required for efficient optimization under uncertainty. In this paper we study the situation in which the unknown performance of a new bandit is to be evaluated and compared with that of a known one over a finite horizon. We assume that the bandits represent random variables with distributions from the one parameter exponential family. When the objective is to maximize the Bayes expected sum of outcomes over a finite horizon, it is shown that optimal policies tend to simple limits when the length of the horizon is large.

1. INTRODUCTION. The multi-armed bandit probem is a basic model for the tradeoffs between the exploration - utilization required for efficient optimization under uncertainty. In this paper we study the situation in which the unknown performance of a new bandit is to be evaluated and compared with that of a known one over a finite horizon. There are two experiments denoted by E_j $(j = 1,2)$. Associated with experiment E_j are i.i.d. random variables which represent the outcomes of the experiment each time it is used. These random variables model, for example, the responses of medical treatments, industrial processes, investment decisions, or even the outcomes of a slot machine (the "bandit"). Associated with each outcome is a reward. We are allowed to use either experiment for N times (finite

[1]Research partially supported by NSF grant DMS- 9703812.

[2]Weatherhead School of Management, Department of Operations, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7235, e-mail: atb4@po.cwru.edu

[3]Dept. of MSIS, Rutgers Business School, Newark and New Brunswick, 111 Washington Street, Newark, NJ 07102. and RUTCOR - Rutgers Center for Operations Research, e-mail: mnk@rci.rutgers.edu

horizon). We wish to maximize the expected value of the sum of the rewards achieved during this finite horizon. Furthermore we assume that the characteristics of experiment E_1 are known in advance, while those of E_2 are not, i.e., experiment E_1 corresponds to a process presently in use, while E_2 corresponds to a new process that is to be evaluated. In this paper we study the case in which the outcomes from E_i (i=1,2) are random variables from the *one parameter exponential family* of distributions. In section 2 we postulate a prior on the unknown parameter of the second experiment, and formulate the problem of maximizing the expected sum of outcomes. We point out that this is equivalent to minimizing a suitably defined regret (expected loss function). In section 3 we summarize a set of results on the the existance of an optimal policy of simple form finite horizon case, obtained in Burnetas and Katehakis (1997a). The main contribution of this paper is to extend the finite horizon results and derive a simple explicit approximation to the optimal policy in the case that the planning horizon is large. This is done in section 4. Section 5 extends the asymptotic approximations to a generalized form of the regret function.

The results of section 4 are related to Lai and Robbins (1985) and Lai (1987) who obtained asymptotic solutions for the more general problem in which one has to choose among k unknown experiments. Our proofs are along different lines, and are based on classical Dynamic Programming arguments, as Bradt, Johnson, and Karlin (1956) did for the binomial case. The results of section 5 are new.

Chernoff and Ray (1965) and Chernoff (1967), obtained asymptotic testing plans for the case of binomial populations using diffusion processes approximations. The approximation technique we use to obtain the asymptotic results is related to that of Schwarz 1962, who derived asymptotic expressions for the hypothesis testing problem, for the case where there is an indifference region separating the two hypotheses. We use a modification of Schwarz's argument to obtain upper and lower bounds for the optimal stopping sets, and then derive asymptotic expressions on these bounds using Laplace's method for the asymptotic expansions of integrals.

The approximation technique we use to obtain the asymptotic results is related to that of Schwarz (1962), who derived asymptotic expressions for the hypothesis testing problem, for the case where there is an indifference region separating the two hypotheses. We use a modification of Schwarz's argument to obtain upper and lower bounds for the optimal stopping sets, and then derive asymptotic expressions on these bounds using Laplace's method for the asymptotic expansions of integrals.

A recent and rather exhaustive survey of the general area is given in Lai (2001); additional recent work in this area is contained in Burnetas and Katehakis (1993), (1996), (1997a) and (1997b), Katehakis and Robbins (1995), and Shimkin and Shwartz (1995), (1995b), . For other related work on the infinite horizon discounted reward version of this problem see Gittins (1979), Varayia Walrand and Buyukkoc (1985), Katehakis and Derman (1987), Katehakis and Veinott (1987), Berry and Fristedt (1985), Agrawal Hedge and Teneketzis (1988), and Glazebrook and Mitchell (2002).

2. THE MODEL. Let E_1 , E_2 be two statistical experiments. With each E_i , i = 1,2 , there are associated: i) a scalar parameter θ_i belonging to some set Θ , and ii) a sequence of random variables X_i , Y_{i1} , Y_{i2} , ... such that Y_{ij} represents the outcome of experiment E_i

the j[th] time it is performed, while X_i is a generic random variable used to denote an outcome from E_i. Given the value of $\theta_i = \theta$, the random variables X_i, Y_{i1}, Y_{i2}, ... are i.i.d., with a probability density function (p.d.f.) $f(x \mid \theta)$ with respect to a non degenerate measure ν. Let $\mu(\theta)$ and $\sigma^2(\theta)$ denote the expected value and variance respectively, of a random variable X distributed according to $f(x \mid \theta)$, i.e. $\mu(\theta) = \mathbf{E}(X \mid \theta)$, $\sigma^2(\theta) = \mathrm{Var}(X \mid \theta)$.

We make the following assumptions.

Assumptions. 1. The p.d.f $f(x \mid \theta)$ belongs to the one–parameter exponential family with a single natural parameter θ, i.e.,

$$f(x \mid \theta) = e^{\theta x - \psi(\theta) + S(x)} \quad . \tag{2.1}$$

2. The parameter space is an interval of the form $\Theta = (\underline{\theta}, \overline{\theta})$, with endpoints that can be infinite, and satisfies the following conditions

$$\zeta_1 = \inf_{\theta \in \Theta} \psi''(\theta) > 0, \quad \zeta_2 = \sup_{\theta \in \Theta} \psi''(\theta) < \infty. \tag{2.2}$$

3. Parameter θ_1 is known in advance, while θ_2 is unknown, and following the Bayesian approach, θ_2 is a random variable with prior distribution: $H_o(\theta)$, $\theta \in \Theta$.

4. We assume that: $\underline{\theta} < \theta_1 < \overline{\theta}$, where $\underline{\theta}$, $\overline{\theta}$ are such that $(\mu(\underline{\theta}), \mu(\overline{\theta})) = \{\mu(\theta) : \theta \in \Theta\}$.

Remark 2.1. **a)** We use the natural parameter representation of the exponential family, c.f., Cox and Hinkley (1974), page 28. It is known that for the one–parameter exponential family $\mu(\theta) = \psi'(\theta)$ and $\sigma^2(\theta) = \psi''(\theta)$, $\mu(\theta)$ is strictly increasing in θ and the set $\{\mu(\theta) : \theta \in \Theta\}$ is an interval of the form $(\mu(\underline{\theta}), \mu(\overline{\theta}))$.

b) Note that if $\theta_1 \leq \underline{\theta}$ $(\theta_1 \geq \overline{\theta})$ then the problem is trivial, because then one should always choose E_2 (E_1).

Let t $(n = N - t)$ denote the number of samples that have already been taken (remain to be taken). At $t = 0$ we have $X_1 \sim f(x \mid \theta_1)$ with respect to $\nu(dx)$, $X_2 \sim f(x \mid \theta_2)$ with respect to $\nu(dx)$, θ_1 known, $\theta_2 \sim H_o(\theta)$.

An observed sample of size k_i from experiment E_i will be denoted by $d_i(k_i) = (y_{i1}, \ldots, y_{i,k_i})$, $i = 1,2$. Let $\underline{k} = (k_1, k_2)$, $\underline{d}(\underline{k}) = (d_1(k_1), d_2(k_2))$.

Since θ_1 is known, the future observations from E_1, Y_{1,k_1+1}, Y_{1,k_1+2}, ..., given $d_1(k_1)$, are i.i.d. random variables with p.d.f. $f(x \mid \theta_1)$, with respect to $\nu(dx)$. Since θ_2 is unknown, the future observations from E_2, Y_{2,k_2+1}, Y_{2,k_2+2}, ... given $\{d_2(k_2)$ and $\theta_2 = \theta\}$, are i.i.d. random variables with p.d.f. $f(x \mid \theta)$, with respect to $\nu(dx)$. Given only $d_2(k_2)$, θ_2 is a random variable with (posterior) distribution $H(\theta \mid d_2(k_2))$, defined as follows

$$dH(\theta \mid d_2(k_2)) = \frac{\tilde{f}(d_2(k_2) \mid \theta)\, dH_o(\theta)}{\tilde{f}(d_2(k_2) \mid H_o)} = \frac{f(y_{2,k_2} \mid \theta)\, dH(\theta \mid d_2(k_2-1))}{\int_{\Theta} f(y_{2,k_2} \mid \theta)\, dH(\theta \mid d_2(k_2-1))}, \tag{2.3}$$

where $d_i(k_i) = (\, d_i(k_i - 1), y_{i,k_i}\,)$, $H(\theta \mid d_2(0)) = H_o(\theta)$, and $\tilde{f}\,(d_2(k_2) \mid \theta)$ (respectively $\tilde{f}\,(d_2(k_2) \mid H_o)$) denotes the joint p.d.f. of the sample $d_2(k_2)$, given $\theta_2 = \theta$ (respectively given the prior H_o) .

Given $d_2(k_2)$, unconditional on the value of θ_2 , the future observations from E_2, Y_{2,k_2+1} , Y_{2,k_2+2} , … , are i.i.d. random variables with distribution determined by the marginal p.d.f (with respect to $\nu(dx)$)

$$f\,(x \mid d_2(k_2)\,) = \int_{\Theta} f(x \mid \theta)\, dH(\theta \mid d_2(k_2)) \,. \tag{2.4}$$

The Bayes estimate of $\mu(\theta_2)$ given the sample $d_2(k_2)$ is equal to

$$\hat{\mu}_2(d_2(k_2)) = \mathbf{E}_{H(\cdot \mid d_2(k_2))}[\, \mu(\theta_2)\,] = \mathbf{E}_{f\,(\cdot \mid d_2(k_2))}[\, Y_{2,k_2+1}\,] \,. \tag{2.5}$$

For notational convenience we use the same symbol f to denote the p.d.f. of an outcome given a specific parameter value, as well as the marginal p.d.f. of an outcome from E_2 given the history of observations $d_2(k_2)$. Although they are different quantities, there is no danger of confusion.

For the one–parameter exponential family case it is well known that the posterior distribution $H(\theta \mid d_2(k_2))$ and the marginal density $f(x| d_2(k_2)\,)$ defined in (2.3) and (2.4) respectively are uniquely determined by the two dimensional sufficient statistic, for the unknown parameter, $(k_2\,, \overline{y}_{2,k_2})$, where $\overline{y}_{2,k} = \frac{1}{k}\sum_{j=1}^{k} y_{2,j}$. Thus, we can assume that in relations (2.3) , (2.4) and (2.5) $d_2(k_2)$ is simply the vector $d_2(k_2) = (k_2\,, \overline{y}_{2,k_2})$.

Given $d_2(k_2 - 1) = (k - 1\,,\ \ y)$ and $Y_{2,k_2} = y_{2,k}$, $d_2(k_2)$ is defined by the following updating scheme

$$d_2(k_2 \mid d_2(k-1), y_{2,k}) = (k,\, \tfrac{k-1}{k} y + \tfrac{1}{k}\, y_{2,k}) = (k, m(k-1, y, y_{2,k}))\,, \tag{2.6}$$

where $m(k,y,x) = (k\,y + x)/\,(k+1)$.

An *N*–stage allocation policy is defined as a rule $\pi = (\pi(0), \pi(1), \ldots, \pi(N-1))$, where

$$\pi(t) = \pi(t \mid d_1(k_1(t,\pi)), d_2(k_2(t,\pi))\,) \tag{2.7}$$

is equal to a_1 or a_2 , according to whether at stage t π dictates to take a sample from E_1 or E_2 respectively, where

$$k_i(t,\pi) = \sum_{j=0}^{t-1} \mathbf{1}_{\{\pi(j)=a_i\}} \,. \tag{2.8}$$

The performance of a policy π is measured by

$$S(t,\pi) = \sum_{j=0}^{t-1} Y_{\pi(j),k_{\pi(j)}(j,\pi)}\,, \tag{2.9}$$

and the expected values

$$\mathbf{E}_\theta S(t,\pi) \;=\; \mathbf{E}\,[S(t,\pi) \;|\; \theta_2{=}\theta] \;=\; \mu(\theta_1)\,\mathbf{E}_\theta\, k_1(t,\pi) \;+\; \mu(\theta)\,\mathbf{E}_\theta\, k_2(t,\pi)\,, \tag{2.10}$$

$$M(t,H_o\,,\pi) \;=\; \mathbf{E}_{H_o}[\,\mathrm{E}_\theta\, S(t,\pi)\,] \;=\; \mathbf{E}_{f(\cdot\,|\,H_o)}\,[S(t,\pi)]\,. \tag{2.11}$$

A policy π^* is optimal for the problem of horizon N and initial prior $H_o(\theta)$ on θ_2, if and only if

$$M(N,H_o\,,\pi^*) \;=\; \max_{\pi} M(N,H_o\,,\pi)\,, \tag{2.12}$$

where the maximum is taken over all sequential policies defined above.

A more general description of the problem is in terms of a loss function $L(\theta,i)$ which represents the expected one step loss incurred when the unknown parameter is equal to θ and a sample from experiment E_i is taken, i.e.,

$$L(\theta,i) \;=\; \mu^*(\theta) - \mathbf{E}_\theta X_i\,, \tag{2.13}$$

where $\mu^*(\theta) \;=\; \max\{\,\mu(\theta_1)\,,\mu(\theta)\,\}$. Then the Bayes risk during the first t observations is

$$R(t,H_o,\pi) \;=\; \mathbf{E}_{H_o}[\,\sum_{j=1}^{t} L(\theta,\pi(j))\,] \;=\; t\,\mathbf{E}_{H_o}[\mu^*(\theta)] \;-\; M(t,H_o\,,\pi)\,. \tag{2.14}$$

Since, $t E_{H_o}[\,\mu^*(\theta)\,]$ in (2.14) is independent of π, maximization of M is equivalent to minimization of R. This leads us to the alternative definition of an optimal policy π^*:

$$R(N,H_o,\pi^*) \;=\; \min_{\pi} R(N,H_o,\pi)\,. \tag{2.15}$$

In section 5 we will consider the following more general form of the loss function

$$L(\theta,i) \;=\; \begin{cases} (\mu(\theta) - \mu(\theta_1))^\beta\,, \text{ if } i=1 \text{ and } \theta \geq \theta_1 + \epsilon, \\ (\mu(\theta_1) - \mu(\theta))^\beta\,, \text{ if } i=2 \text{ and} \theta \leq \theta_1 - \epsilon\,, \\ 0,\ otherwise, \end{cases} \tag{2.16}$$

where $\beta \geq 1$, $\epsilon \geq 0$.

3. OPTIMALITY EQUATIONS – PRELIMINARY RESULTS. In this section we state some preliminary properties and two theorems of Burnetas and Katehakis (1997a) on the structure of an optimal policy for the finite horizon problem. It will be more convenient to discuss the problem in terms of n, the number of samples remaining to be taken until the end of the horizon N.

Let $P(n,k,y)$ be the problem of maximizing the expected sum of observations over a horizon n, when the initial information about θ_2 is summarized by $H(\theta \;|\; (k,y))$, i.e., the posterior

distribution of θ_2 given $d_2(k_2) = (k,y)$. Also let $Q(n,k,y)$ be the problem of minimizing $R(n,H,\pi)$, with the same conventions.

For problems $P(n,k,y)$ and $Q(n,k,y)$ define the optimal value functions

$$V(n,k,y) = \sup_{\pi} M(n, H(\cdot \mid (k,y), \pi), \tag{3.1}$$

$$U(n,k,y) = \inf_{\pi} R(n, H(\cdot \mid (k,y), \pi), \tag{3.2}$$

respectively.

Using standard arguments of Markovian Decision Processes with general state and finite action spaces (cf. Dynkin 1979) one obtains

Proposition 3.1. **a)** The functions $V(n,k,y)$ are the unique solutions of equations (3.3), (3.4) below.

$$V(n,k,y) = \max\{r(k,y;a_1) + V(n-1,k,y),\ r(k,y;\alpha_2) + \mathbf{E}_{f(\cdot \mid (k,y))} V(n-1,k+1,m(k,y,X_2))\},$$

$$n = 1,2,\ldots,N,\ k = 0,1,\ldots,N-n,\ y \in \mathbb{R}, \tag{3.3}$$

$$V(0,k,y) = 0, \tag{3.4}$$

b) The functions $U(n,k,y)$ are the unique solutions of equations (3.5), (3.6) below.

$$U(n,k,y) = \min\{c(k,y;a_1) + U(n-1,k,y),\ c(k,y;a_2) + \mathbf{E}_{f(\cdot \mid (k,y))} U(n-1,k+1,m(k,y,X_2))\},$$

$$n = 1,2,\ldots,N,\ k = 0,1,\ldots,N-n,\ y \in \mathbb{R}, \tag{3.5}$$

$$U(0,k,y) = 0. \tag{3.6}$$

The one step expected reward and cost functions $r(k,y;a_i)$ and $c(k,y;a_i)$, $i = 1,2$, are defined as follows.

$$r(k,y;a_1) = \mathbf{E}_{\theta_1}[X_1] = \mu(\theta_1), \tag{3.7}$$

$$\begin{aligned} r(k,y;a_2) &= \mathbf{E}_{f(\cdot \mid (k,y))} X_2 \\ &= \mathbf{E}_{H(\cdot \mid (k,y))}[\mathbf{E}_{\theta} X_2] = \int_{\Theta} \mu(\theta)\, dH(\theta \mid (k,y)). \end{aligned} \tag{3.8}$$

$$\begin{aligned} c(k,y;a_1) &= \mathbf{E}_{H(\cdot \mid (k,y))}[\mu^*(\theta)] - r(k,y;a_1)] \\ &= \int_{\theta \geq \theta_1} (\mu(\theta) - \mu(\theta_1))\, dH(\theta \mid (k,y)), \end{aligned} \tag{3.9}$$

$$c(k,y;a_2) \;=\; \mathbf{E}_{H(\cdot\,|\,(k,y))}\,[\,\mu^*(\theta)] \;-\; r(k,y;a_2)]$$

$$= \int_{\theta<\theta_1} (\mu(\theta_1) - \mu(\theta))\, dH(\theta \;|\; (k,y))\,. \tag{3.10}$$

Moreover, the supremum and infimum in (3.1) and (3.2) are attained by a policy π^*, and they can be replaced by maximum and minimum respectively.

In the next proposition it is stated that (3.3) and (3.5) are equivalent to the optimality equations of appropriately defined stopping problems, where "stopping" means switching to the known experiment and staying there for the remaining trials. The proof is an extension of that given in (Bradt, Johnson and Karlin (1956)) for the case of binomial populations.

Proposition 3.2. **a)** Eqs. (3.3) are equivalent to the following

$$V(n,k,y) \;=\; \max\{\, n\mu(\theta_1)\,,\, r(k,y;a_2) \;+\; \mathbf{E}_{f(\cdot\,|\,(k,y))} V(n-1,k+1,m(k,y,X_2))\}\,, \tag{3.11}$$

b) Eqs. (3.5) are equivalent to the following

$$U(n,k,y) = \min\{\, nc(k,y;a_1)\,,\, c(k,y;a_2) \;+\; \mathbf{E}_{f(\cdot\,|\,(k,y))} U(n-1,k+1,m(k,y,X_2))\}\,. \tag{3.12}$$

We will use the following quantities in the sequel, where log denotes the base e logarithm.

Definitions 3.1. For $y = \frac{1}{k}\sum_{j=1}^{k} y_{2,j}$, let

$$\ell(\theta,\theta_1 \;|\; y) \;=\; \log \frac{f(y\,|\,\theta)}{f(y\,|\,\theta_1)} \tag{3.13}$$

$$\Lambda(k,y) \;=\; \int_{\Theta} e^{k\,\ell(\theta,\theta_1|y)}\, dH_o(\theta) \tag{3.14}$$

$$d(\theta) \;=\; \theta \;-\; \theta_1\,, \tag{3.15}$$

$$\delta(\theta) \;=\; \mu(\theta) \;-\; \mu(\theta_1)\,, \tag{3.16}$$

$$\omega(\theta) \;=\; \psi(\theta) - \psi(\theta_1)\,. \tag{3.17}$$

Remark 3.1. From (2.1) it is easy to see that

$$\ell(\theta,\theta_1 \;|\; y) = \;\; d(\theta)\, y - \omega(\theta)\,, \tag{3.18}$$

$$k\,\ell(\theta,\theta_1 \;|\; y) + \;\ell(\theta,\theta_1 \;|\; x) \;=\; (k+1)\;\ell(\theta\,,\theta_1 \;|\; m(k,y,x))\,. \tag{3.19}$$

Using Remark 3.1, we can rewrite the optimality equations so that the expectations in the right hand side are taken with respect to the density $f(x\,|\,\theta_1)$ instead of the marginal density $f(x\,|\,(k,y))$. This is done in the next proposition.

Proposition 3.3. **a)** Eqs. (3.11) are equivalent to the following set of equations.

$$v(n,k,y) = \max\{0, q(k,y) + \mathbf{E}_{f(\cdot|\theta_1)} v(n-1, k+1, m(k,y,X_2))\}, \tag{3.20}$$

$$n = 1,2,\ldots,N,\ k = 0,1,\ldots,N-n,\ y \in \mathbb{R},$$

$$v(0,k,y) = 0, \tag{3.21}$$

where

$$v(n,k,y) = (V(n,k,y) - n\mu(\theta_1))\ \Lambda(k,y), \tag{3.22}$$

and

$$q(k,y) = \int_{\Theta} \delta(\theta)\, e^{k\ell(\theta,\theta_1|y)}\, dH_o(\theta). \tag{3.23}$$

b) Eqs. (3.12) are equivalent to the following set of equations.

$$u(n,k,y) = \min\{n\bar{c}(k,y;a_1),\ \bar{c}(k,y;a_2) + \mathbf{E}_{f(\cdot|\theta_1)} u(n-1, k+1, m(k,y,X_2))\}, \tag{3.24}$$

$$n = 1,2,\ldots,N,\ k = 0,1,\ldots,N-n,\ y \in \mathbb{R},$$

$$u(0,k,y) = 0, \tag{3.25}$$

where

$$u(n,k,y) = U(n,k,y)\ \Lambda(k,y), \tag{3.26}$$

$$\bar{c}(k,y;a_1) = \int_{\theta \geq \theta_1} \delta(\theta)\, e^{k\ell(\theta,\theta_1|y)}\, dH_o(\theta), \tag{3.27}$$

$$\bar{c}(k,y;a_2) = -\int_{\theta \leq \theta_1} \delta(\theta)\, e^{k\ell(\theta,\theta_1|y)} dH_o(\theta). \tag{3.28}$$

Theorem 3.1 describes the structure of the optimal policy with respect to stopping and continuation intervals for $y = \frac{1}{k}\sum_{j=1}^{k} y_{2,j}$, while Theorem 3.2 gives a more intuitive characterization in terms of inflation factors added to the Bayes estimate of $\mu(\theta_2)$.

Theorem 3.1. **a)** For each n, k there exists a number $y_n(k)$ with the property

$$\pi^*(n,k,y) = \begin{cases} a_1, & \text{if } y < y_n(k) \\ a_2, & \text{if } y \geq y_n(k) \end{cases} \tag{3.29}_n$$

where $\pi^*(n,k,y)$ is the action indicated by the optimal policy in state (n,k,y).

b) The sequence $y_n(k)$ is nonincreasing in n.

Theorem 3.2. a) For each n, k, y there is a real number $\epsilon(n,k,y)$ with the property

$$\pi^*(n,k,y) = \begin{cases} a_1, & \text{if } \mathbf{E}_{H(\cdot\,|\,(k,y))}[\mu(\theta_2)] + \epsilon(n,k,y) < \mu(\theta_1) \\ a_2, & \text{if } \mathbf{E}_{H(\cdot\,|\,(k,y))}[\mu(\theta_2)] + \epsilon(n,k,y) \geq \mu(\theta_1), \end{cases} \tag{3.30}$$

where $\pi^*(n,k,y)$ is the action indicated by the optimal policy in state (n,k,y), and

$$\epsilon(n,k,y) = -\frac{q(k,y_n(k))}{\Lambda(k,y)}. \tag{3.31}$$

b) The quantities $\epsilon(n,k,y)$ are positive and increasing in n.

Remarks. The threshold $y_n(k)$ represents the amount of immediate reward which we can afford sacrificing in order to obtain information about θ_2, which is valuable for the remaining decisions.

An interpretation of the quantities $\epsilon(n,k,y)$ is that they represent a positive inflation, that is added to the current estimate of the reward of E_2, $\hat{\mu}_2 = \mathbf{E}_{H(\cdot\,|\,(k,y))}[\mu(\theta_2)]$, in order to take into account the uncertainty associated with it.

4. ASYMPTOTICS FOR LARGE N. In this section we obtain properties of the optimal policy that are related to its behavior when the planning horizon is large. Before we proceed with the analysis, we shall make another assumption in addition to those in Section 2. Specifically, we assume that the prior distribution of θ_2 is continuous in $[\underline{\theta}, \overline{\theta}]$, i.e. there is a prior probability density function denoted by $h_o(\theta)$, such that $dH_o(\theta) = h_o(\theta)\,d\theta$, with $\overline{h}_o = \sup_{\Theta} h_o(\theta)$. This assumption helps simplify the derivation of the asymptotic approximations below. However it does not restrict the generality of the results, since the discrete case can be treated in an analogous but simpler way.

The derivations in this section are based on the optimality equations in terms of the regret defined in (3.24). The main results are given in Theorems 4.1 and 4.2 which provide upper and lower bounds for the optimal stopping regions. The proofs of these two theorems are based on a number of intermediate properties, which are given in the appendix in Lemmata A.1–A.7.

For each n define the stopping region $S_n = \{(k,y):\ \pi^*(n,k,y) = 1\}$.

Theorem 4.1. Under the assumptions made, when $n \to \infty$

$$\underline{S}_n \subset S_n \subset \overline{S}_n, \tag{4.1}$$

where

$$\underline{S}_n = \{(k,y):\ n\,\overline{c}(k,y;a_1) < \overline{c}(k,y;a_2)\}, \tag{4.2}$$

$$\overline{S}_n = \{(k,y):\ n\,\overline{c}(k,y;a_1) < 2\sqrt{A(n+k)\,\overline{c}\,(k,y;a_2)}\}. \tag{4.3}$$

Proof. From (3.24) and Lemma A.4 (a) it follows that $u(n,k,y) > 0$. Thus, if $n\,\overline{c}(k,y;a_1) < \overline{c}(k,y;a_2)$, then it is optimal to stop, i.e. $\pi^*(n,k,y) = 1$. This proves the first part of (4.1).

In order to prove the second part, consider the allocation rule $\tau(i)$ denined as follows: a) take a fixed number $i\ (i \le n)$ of samples from E_2, and b) take the remaining $n-i$ samples from E_1 or E_2, according to whether

$$\overline{c}(k+i\,, m(k,y,y_{2,k+1},\ldots,y_{2,k+i})\,;a_1) \;<\; \overline{c}(k+i\,, m(k,y,y_{2,k+1},\ldots,y_{2,k+i});a_2)\,, \tag{4.4}$$

or

$$\overline{c}(k+i\,, m(k,y,y_{2,k+1},\ldots,y_{2,k+i})\,;a_1) \;\ge\; \overline{c}(k+i\,, m(k,y,y_{2,k+1},\ldots,y_{2,k+i});a_2)\,, \tag{4.5}$$

respectively, where $m(k,y,y_{2,k+1},\ldots,y_{2,k+i})$ denotes the new average after the i additional outcomes

$$m(k,y,y_{2,k+1},\ldots,y_{2,k+i}) \;=\; \frac{k\,y}{k+\mathrm{i}} \;+\; \frac{y_{2,\mathrm{k}+1}+\cdots+y_{2,\mathrm{k+i}}}{k+\mathrm{i}}\,. \tag{4.6}$$

Now from Lemma A.4(c), rule $\tau(i)$ has the following risk

$$R^{\tau(i)}(n,k,y) \;=\; i\,\overline{c}(k,y;a_2) \;+\; (n-i)\,\mathbf{E}_{f(\cdot|\theta_1)}[\;\gamma(k+i, m(k,y,Y_{2,k+1},\ldots,Y_{2,k+i}))\;]\,, \tag{4.7}$$

where $\gamma(k,y) \;=\; \min\{\,\overline{c}(k,y;a_1)\,,\,\overline{c}(k,y;a_2)\,\}$.

Note that in (4.7) the risk is the one corresponding to the transformed experiments (see Remark 3.1).

From Lemma A.5, it follows that there exists $A > 0$ such that

$$R^{\tau(i)}(n,k,y) \;<\; i\,\overline{c}(k,y;a_2) \;+\; (n-i)\;\frac{\mathrm{A}}{\mathrm{k+i}} \;=\; \phi(i\,;n,k,y)\;,\; i=0,1,\ldots,n\,. \tag{4.8}$$

If we consider the extension of $\phi(i)$ to the real domain,

$$\phi(i\,;n,k,y) \;=\; i\,\overline{c}(k,y;\,a_2) \;+\; (n-i)\;\frac{\mathrm{A}}{\mathrm{k+i}}\;,\;\; 0 \;\le\; i \;\le\; n\,,\; i\in\mathbb{R}\,, \tag{4.9}$$

then we can differentiate with respect to i

$$\phi'(i) \;=\; \overline{c}(k,y\,;a_2) - \mathrm{A}\;\frac{\mathrm{k+n}}{(\mathrm{k+i})^2}\;, \tag{4.10}$$

$$\phi''(i) \;=\; 2\mathrm{A}\;\frac{\mathrm{k+n}}{(\mathrm{k+i})^3} \;>\; 0\,, \tag{4.11}$$

hence $\phi(i)$ is convex. We also have

$$\phi'(0) \;=\; \overline{c}(k,y\,;\,a_2) - \mathrm{A}\;\frac{\mathrm{k+n}}{\mathrm{k}^2}\;, \tag{4.12}$$

which is negative for n sufficiently large, and

$$\phi'(n) \;=\; \overline{c}(k,y\,;a_2) - \;\tfrac{\mathrm{A}}{\mathrm{k+n}}\;\;, \tag{4.13}$$

which is positive also for n sufficiently large. This means that ϕ attains its minimum at some i^*, $1 \le i^* < n$, for which $\phi'(i^*) = 0$, i.e.

$$i^* \;=\; \sqrt{\tfrac{\mathrm{A}}{\overline{c}(k,y\,;a_2)}\;(k+n)} \;- k\,. \tag{4.14}$$

Let $\lceil i^* \rceil = \min\{\, i \in \mathbb{N} \;:\; i \ge i^* \}$. Then $\lceil i^* \rceil \;<\; i^* + 1$, and, since ϕ is convex

$$\phi(\lceil i^* \rceil) \;<\; \phi(i^* + 1)\,. \tag{4.15}$$

Note that

$$\begin{aligned}
\phi(i^*+1) \;&=\; (\,i^*+1)\,\overline{c}(k,y;a_2) \;+\; (n - i^* - 1)\;\tfrac{\mathrm{A}}{\mathrm{k+i^*+1}} \\
&\le\;\; (\,i^*+1)\,\overline{c}(k,y;a_2) \;+\; (n - i^*)\;\tfrac{\mathrm{A}}{\mathrm{k+i^*}} \\
&=\; 2\sqrt{\mathrm{A}(n+k)\,\overline{c}(k,y\,;\,a_2)} \;-\; (k-1)\,\overline{c}\,(k,y\,;\,a_2) \;-\; \mathrm{A} \\
&<\; 2\sqrt{\mathrm{A}(n+k)\,\overline{c}(k,y\,;a_2)} \;=\; \phi^*(n,k,y)\,.
\end{aligned} \tag{4.16}$$

Combining the above inequalities we have

$$R^{\tau(\lceil i^* \rceil)}(n,k,y) \;<\; \phi^*(n,k,y)\,. \tag{4.17}$$

From this discussion we see that for each (n,k,y) there is an allocation rule, namely $\tau(\,\lceil i^*(n,k,y) \rceil\,)$ as described above, which has expected risk less than ϕ^*. Thus, if $n\overline{c}(k,y\,;\,a_1) \;\ge\; \phi^*$, then it is not optimal to stop, since continuing for i^* more steps gives a better policy. So $n\,\overline{c}(k,vy\,;a_1) \;\ge\; \phi^*$ implies that $\pi^*(n,k,y) \;=\; 2$, or equivalently $\pi^*(n,k,y) \;=\; 1$ implies that $n\,\overline{c}(k,y\,;a_1) \;<\; \phi^*$, which completes the proof of the theorem.

Based on Theorem 4.1 we now derive an asymptotic approximation of the optimal policy $\pi^*(n,k,y)$ as $n \to \infty$.

Let

$$G(k,y,\theta_1) \;=\; \begin{cases} k\,\mathbf{I}\,(\theta^*(y),\theta_1)\,, & \text{if } \mu(\underline{\theta}) < y < \mu(\theta_1) \\ k\;\ell(\underline{\theta},\theta_1)\,, & \text{if } y \le \mu(\underline{\theta})\,, \end{cases} \tag{4.18}$$

where $\mathbf{I}\,(\sigma,\tau) \;=\; \mathbf{E}_\sigma[\,\log\,\tfrac{f\,(X|\,\sigma)}{f\,(X|\,\tau)}\,]$ is the Kullback–Leibler information number .

Theorem 4.2. If $h_o(\theta) > 0$, $\forall\,\theta \in \Theta$, then the optimal policy corresponding to the solution of (3.40) when $n \to \infty$ can be approximated by the following policy.

$$\pi_1(n,k,y) = \begin{cases} a_1, & \text{if } y < \mu(\theta_1) \text{ and } G(k,y,\theta_1) > \log n \\ a_2, & \text{otherwise} \end{cases} \tag{4.19}$$

Proof. We will show that, for large n, the sets $\underline{S}_n$ and $\overline{S}_n$ defined in Theorem 4.1 can both be approximated by the set $K = \{ (k,y) : y < \mu(\theta_1) \text{ and } k\,G(k,y,\theta_1) > \log n \}$. We first consider $\underline{S}_n$, which are described by the following relation

$$\frac{\overline{c}(k,y\,;a_1)}{\overline{c}(k,y\,;a_2)} < \frac{1}{n} . \tag{4.20}$$

From (4.20) we see that, as $n \to \infty$, at least one of the following conditions holds:

$$\overline{c}(k,y\,;a_1) \to 0 , \text{ or } \overline{c}(k,y\,;a_2) \to \infty .$$

From the definition of $\overline{c}(k,y\,;a_1)$ and $\overline{c}(k,y\,;a_2)$ in (3.27), (3.28) it follows that, for any fixed y, in order for either of the above conditions to be true it is necessary that $k \to \infty$. From Lemma A.6 it is easy to see that, when $k \to \infty$, the values of y for which the above ratio tends to 0 are those included in the range $y < \mu(\theta_1)$. We can now use explicitly the results of Lemma A.6 to obtain an asymptotic approximation for the inequality in (4.20). In the case $y < \mu(\theta_1)$, that we are interested in, we have from (A.32) that

$$\overline{c}(k,y;a_1) \sim \frac{2\,h_o(\theta_1)}{(\mu(\theta_1)-y)^2\,k^2} .$$

As for $\overline{c}(k,y;a_2)$, we must consider three cases corresponding to b1, b2 and b3 of Lemma A.6, namely a) $y < \mu(\underline{\theta})$, b) $y = \mu(\underline{\theta})$, and c) $\mu(\underline{\theta}) < y < \mu(\theta_1)$. For each one of these cases (4.20) takes the following forms:

In case (a)

$$\frac{\overline{c}(k,y\,;a_1)}{\overline{c}(k,y\,;a_2)} \sim \frac{2\,h_o(\theta_1)}{(\mu(\theta_1)-y)^2\,(\mu(\theta_1)-\mu(\underline{\theta}))\,h_o(\underline{\theta})\,k\,e^{k\ell(\underline{\theta},\theta_1\,|\,y)}} < \frac{1}{n} . \tag{4.21}$$

Since $y < \mu(\underline{\theta})$, from Lemma A.3, $\ell(\underline{\theta},\theta_1 \mid y) > 0$, therefore, the above approximate expression is decreasing in k, and since $n \to \infty$ the inequality holds for

$$k\,\ell(\underline{\theta},\theta_1|\,y) > \log n - o(\log n) . \tag{4.22}$$

For cases (b) and (c) we can show in the same way that the approximate solution of (5.22) is

$$k\,\ell(\underline{\theta},\theta_1|\,y) > \log n - o(\log n) , \tag{4.23}$$

and

$$k\,\mathbf{I}(\theta^*(y),\theta_1) > \log n - o(\log n) , \tag{4.24}$$

respectively.

We now turn to the inequality which defines the set $\overline{S}_n$ in (4.3). This can be rewritten as

$$\frac{(\overline{c}(k,y\,;a_1))^2}{\overline{c}(k,y\,;a_2)} \;<\; \frac{4A(n+k)}{n^2}\;. \tag{4.25}$$

We can use again the approximations obtained in Lemma A.6, to obtain relations analogous to (4.22), (4.23) and (4.24) for the three cases. For case (a) we now have

$$\frac{(\overline{c}(k,y\,;a_1))^2}{\overline{c}(k,y\,;a_2)} \;\sim\; \frac{4\,h_o^2(\theta_1)}{(\mu(\theta_1)-y)^4\,(\mu(\theta_1)-\mu(\underline{\theta}))\,h_o(\underline{\theta})\,k^3\,e^{k\ell(\underline{\theta},\theta_1\,|\,y)}} \;<\; \frac{4A(n+k)}{n^2}\;. \tag{4.26}$$

Consider (4.26) with the inequality replaced by equality. Assuming that for fixed y the unique solution in k satisfies $k/n \to 0$, the right hand side is of the same order as $1/n$, thus we obtain the following

$$k\,\ell(\underline{\theta},\theta_1\,|\,y) \;=\; \log n \;-\; o(\log n) \tag{4.27}$$

for the asymptotic solution of the equality. This form is in agreement with the assumption $k/n \to 0$. Since the approximate expression in (4.26) is decreasing in k, while the right hand side is increasing, the required inequality will hold for

$$k\,\ell(\underline{\theta},\theta_1\,|\,y) \;>\; \log n \;-\; o(\log n)\,. \tag{4.28}$$

In cases (b) and (c) the corresponding expressions will be

$$k\,\ell(\underline{\theta},\theta_1\,|\,y) \;>\; \log n \;-\; o(\log n)\,, \tag{4.29}$$

and

$$k\,\mathbf{I}\,(\theta^*(y),\theta_1) \;>\; \log n \;-\; o(\log n)\,. \tag{4.30}$$

If we combine (4.22) with (4.28); (4.23) with (4.29); and (4.24) with (4.30), we can see that both $\underline{S}_n$ and $\overline{S}_n$ can be described approximately for large n by the following inequalities

$$k\,\ell(\underline{\theta},\theta_1\,|\,y) \;>\; \log n \;,\; \text{when } y \le \mu(\underline{\theta})\,,\; \text{and} \tag{4.31}$$

$$k\,\mathbf{I}\,(\theta^*(y),\theta_1) \;>\; \log n \;,\; \text{when } \mu(\underline{\theta}) < y < \mu(\theta_1)\,. \tag{4.32}$$

Therefore, based on Theorem 4.1, we can also approximate the stopping set S_n with the same region, and now the asymptotic interpretation of the optimal policy is possible. Namely, in the case $\mu(\underline{\theta}) < y < \mu(\theta_1)$ stopping is required when $k\,\mathbf{I}\,(\theta^*(y),\theta_1) > \log n$, and in the case $y \le \mu(\underline{\theta})$ when $k\,\ell(\underline{\theta},\theta_1\,|\,y) > \log n$.

Remark 4.2. **a)** A consequence of Theorem 4.2 is that, for large n, it is never optimal to stop sampling from E_2 when $y \ge \mu(\theta_1)$, even if the current posterior distribution of θ_2 is unfavorable, i.e. $\mathbf{E}_H[\theta_2] \;<\; \mu(\theta_1)$.

b) The asymptotic policy derived in Theorem 4.2 is independent of the initial prior p.d.f. h_o, when $h_o(\theta) > 0$, $\forall \theta \in \Theta$. If this condition fails, it can still be shown, based on Remark A.1, that a more general form of the asymptotically optimal policy is

$$\pi_1(n,k,y) = \begin{cases} 1\,, & \text{if } \; y < \mu(\xi) \;\text{ and }\; k\,\ell(\tau,\theta_1 \mid y) \;>\; \log n \\ 2\,, & \text{otherwise} \end{cases} \tag{4.33}$$

where

$$\xi \;=\; \inf\{\,\theta \geq \theta_1\,,\; h_o(\theta) > 0\,\}\,, \tag{4.34}$$

and τ is the value of θ which maximizes $\ell(\theta,\theta_1 \mid y)$ in the support of the prior p.d.f.

c) The policy in (4.19) is analogous to that described in Lai and Robbins (1985) and in Lai (1987), in the general case where there are m unknown experiments to be compared. Their asymptotically optimal policy is based on the use of upper confidence bounds (which essentially estimate the unknown parameters) in the following way. If x_j is the average of T_j successive observations from experiment E_j, $j = 1,\ldots,i$, the upper confidence bound is defined as

$$U_j(T_j\,,x_j) \;=\; \inf\{\,\theta > \theta_{x_j}\,,\; \mathbf{I}\,(\theta_{x_j},\theta) \;>\; \tfrac{g(T_j/N)}{T_j}\,\}\,, \tag{4.35}$$

where θ_{x_j} is the maximum likelihood estimate for θ_j given $(T_j\,,x_j)$, and g is a function that satisfies certain assumptions (cf. Lai (1987)), among which is that $g(t) \sim \log t^{-1}$ when $t \to 0$. Then the policy suggests sampling from the experiment with the largest upper confidence bound.

Here, from (4.33) we can see that for every state (n,k,y) there is a number $\theta_1'(n,k,y)$ such that if the known parameter θ_1 of E_1 is less than $\theta_1'(n,k,y)$, then it is optimal to continue, otherwise it is optimal to stop. The value of $\theta_1'(n,k,y)$ can also be determined from (4.33) as follows .

$$\theta_1'(n,k,y) \;=\; \inf\{\,\theta > \theta^*(y)\,,\; \mathbf{I}\,(\theta^*(y),\theta) \;>\; \tfrac{\log n}{k}\,\}\,,\;\text{ if }\; \mu(\underline{\theta}) < y < \mu(\theta_1)\,, \tag{4.36}$$

thus,

$$\theta_1'(n,k,y) \;=\; \inf\{\,\theta > \underline{\theta}\,,\; \ell(\underline{\theta},\theta_1 \mid y) \;>\; \tfrac{\log n}{k}\,\}\,,\;\text{ if }\; y \leq \mu(\underline{\theta})\,. \tag{4.37}$$

Therefore, $\theta_1'(n,k,y)$ plays essentially the same role as the upper confidence bounds, if one considers the fact that $T_j/N \;\to 0$ in (4.35).

5. GENERALIZATION OF THE REGRET. The regret $R(t,H,\pi)$ was defined in (2.16) as the Bayes risk corresponding to the loss function $L(\theta,i)$ defined in (2.15). It was shown that, for this particular choice of loss function, the problems of maximizing the expected sum of outcomes and minimizing the Bayes risk are equivalent. In this section we consider a more general form for $L(\theta,i)$, namely

$$L(\theta,i) \;=\; (\mu^*(\theta) - \mathbf{E}_\theta X_{\mathrm{i}})^{\beta}\,\mathbf{1}_{\{\,|\theta-\theta_1| > \epsilon\,\}}\,, \tag{5.1}$$

or equivalently

$$L(\theta,i) \;=\; \begin{cases} (\mu(\theta)-\mu(\theta_1))^{\beta} \;, \text{ if } i=1 \text{ and } \theta \geq \theta_1+\epsilon \\ (\mu(\theta_1)-\mu(\theta))^{\beta} \;, \text{ if } i=2 \text{ and } \theta \leq \theta_1-\epsilon \\ 0 \;, \qquad\qquad \text{otherwise} \end{cases} \;, \tag{5.2}$$

where $\beta \geq 1$, $\epsilon \geq 0$. This definition of $L(\theta,i)$ includes (2.15) as a special case obtained when $\beta = 1$ and $\epsilon = 0$. It also includes other useful loss functions, such as the quadratic loss $(\beta = 2, \epsilon = 0)$. Furthermore, the case $\epsilon > 0$ corresponds to the existence of an indifference region in a neighborhood of the known value, in which no loss is incurred, i.e. , if $\theta_2 \in (\theta_1 - \epsilon, \theta_1 + \epsilon)$, then both actions are optimal.

Remark 5.1. To avoid trivialities, we assume that $\epsilon < \min\{\theta_1 - \underline{\theta}\,, \overline{\theta} - \theta_1\}$. This ensures that there are possible values of θ_2 at both sides of θ_1, which are distinguishable from θ_1 with respect to the loss function, thus the decision problem is not trivial.

For the loss function defined in (5.1), the second equality in (2.16) is not true in general. Therefore there is no immediate analogue for reward maximization. Nevertheless, we can still formulate optimality equations for the problem of minimization of the regret $R(n,H,\pi)$, as in section 3. For the finite horizon case Theorem 3.1 is still valid. Furthermore, there are analogous expressions for the asymptotic approximations derived in section 4 . In the remainder of this section we highlight the necessary modifications in the formulation, the intermediate properties of the one step regret functions, and the proofs.

For the dynamic programming formulation, we can still define the optimal value function for the regret as in (3.2). Then the optimality equations for $U(n,k,y)$ have exactly the same form as those given in (3.5) and (3.12), the only difference being that the one step cost functions defined in (3.9) and (3.10) now take the form

$$c(k,y;\alpha) \;=\; \mathbf{E}_{\mathrm{H}(\cdot\,|\,(\mathrm{k,y}))}[L(\theta,\alpha)] \;, \tag{5.3}$$

and more specifically

$$c(k,y;a_1) \;=\; \int\limits_{\theta \geq \theta_1+\epsilon} (\mu(\theta)-\mu(\theta_1))^{\beta}\, dH(\theta \;|\; (k,y)) \;, \tag{5.4}$$

$$c(k,y;a_2) \;=\; \int\limits_{\theta \leq \theta_1-\epsilon} (\mu(\theta_1)-\mu(\theta))^{\beta}\, dH(\theta \;|\; (k,y)) \;. \tag{5.5}$$

Proposition 3.3.b still holds, but now with

$$\overline{c}(k,y;a_1) \;=\; \int\limits_{\theta \geq \theta_1+\epsilon} (\delta(\theta))^{\beta}\, e^{k\ell(\theta,\theta_1\,|\,y)}\, dH_o(\theta) \;, \tag{5.6}$$

$$\overline{c}(k,y;a_2) \;=\; \int\limits_{\theta \leq \theta_1-\epsilon} (-\delta(\theta))^{\beta}\, e^{k\ell(\theta,\theta_1\,|\,y)} dH_o(\theta) \;. \tag{5.7}$$

The discussion in section 3 was based on the optimal reward function $v(n,k,y)$ and the optimality equations (3.20) . If we define

$$q(k,y) \;=\; \overline{c}(k,y;a_1) - \overline{c}(k,y;a_2) \tag{5.8}$$

and

$$v(n,k,y) \;=\; n\,\overline{c}(k,y;a_1) - u(n,k,y) \;\;, \tag{5.9}$$

then the quantities $q(k,y)$ and $v(n,k,y)$, although they do not possess immediate interpretation as in section 3, satisfy optimality equations analogous to (3.20). Thus we can establish the structure of the optimal policy for the finite horizon problem analogous to that described in Theorem 4.1.

Now we turn to the asymptotic properties corresponding to those of section 4. Lemmata A7–A8 in the Appendix are the equivalent of A5–A6 for this case.

Let $\xi = \frac{\zeta_1 \epsilon^2}{2}$. The analogue of Theorem 4.1 is presented in the following theorem.

Theorem 5.1. Under the assumptions made, when $n \to \infty$

$$\underline{S}_n \;\subset\; S_n \;\subset\; \overline{S}_n \;\;, \tag{5.10}$$

where

$$\underline{S}_n \;=\; \{\,(k,y):\; n\overline{c}(k,y;a_1) \;<\; \overline{c}(k,y;a_2)\,\}\,, \tag{5.11}$$

$$\overline{S}_n \;=\; \{\,(k,y):\; n\overline{c}(k,y;a_1) \;<\; \tfrac{2\,\overline{c}(k,y;a_2)}{\xi}\;\log\,\tfrac{A\xi n}{\overline{c}(k,y;a_2)}\,\} \tag{5.12}$$

Proof. Eqs. (5.11) can be proved in the same way as (4.2). For (5.12) we can also use the same arguments as in Theorem 4.1, up to inequality (4.8), which here, using Lemma A.8, takes the form

$$R^{\tau(i)}(n,k,y) \;<\; i\,\overline{c}(k,y;a_2) \;+\; n\,A\,e^{-(k+i)\xi} \;=\; \phi(i\,;\,n,k,y) \;\;,\;\; i = 0,1,\ldots,n\,. \tag{5.13}$$

For the extension of ϕ in the real domain we obtain

$$\phi'(i) \;=\; \overline{c}(k,y\,;a_2) - \xi\,n\,A\,e^{-(k+i)\xi} \;\;. \tag{5.14}$$

Thus, ϕ is still convex, $\phi'(0) < 0$, and $\phi'(n) > 0$ for n sufficiently large. Therefore, the minimum is attained at the root of the first derivative,

$$i^* \;=\; \tfrac{1}{\xi}\log\,\tfrac{A\xi n}{\overline{c}(k,y;a_2)} \;-k\,. \tag{5.15}$$

Let $\lceil i^* \rceil = \min\{\, i \in \mathbb{N} \;:\; i \geq i^* \}$. Then $\lceil i^* \rceil \;<\; i^* + 1$, and, since ϕ is convex

$$\phi(\lceil i^* \rceil) \;<\; \phi(i^* + 1)\,. \tag{5.16}$$

But

$$\begin{aligned} \phi(i^* + 1) \;&=\; (\,i^* + 1)\overline{c}(k,y;a_2) \;+\; \;nA\,e^{-(k+i^*+1)\xi} \\ &\leq\; (\,i^* + 1)\overline{c}(k,y;a_2) \;+\; nA\,e^{-(k+i^*)\xi} \end{aligned}$$

$$= \frac{\overline{c}(k,y;a_2)}{\xi} \log \frac{A\xi n}{\overline{c}(k,y;a_2)} - (k-1)\,\overline{c}(k,y\,;a_2) + \frac{\overline{c}(k,y;a_2)}{\xi}$$

$$< 2\,\frac{\overline{c}(k,y;a_2)}{\xi} \log \frac{A\xi n}{\overline{c}(k,y;a_2)} = \phi^*(n,k,y)\,. \tag{5.17}$$

Now the second inclusion relationship in (5.10) follows from (5.17) in the same way that the second inclusion relationship in (4.1) follows from (4.16).

We finally establish the approximation of the optimal policy for large horizon, similarly to Theorem 4.2 , for $\epsilon > 0$. Define the following sets

$$\begin{aligned} K_1(n,k,y) = \{\,(k,y):\ & \mu(\theta_1-\epsilon) \le y < \mu(\theta_1+\epsilon)\,, \\ & \ell(\theta_1-\epsilon,\theta_1 \mid y) > \ell(\theta_1+\epsilon,\theta_1 \mid y) \text{ and} \\ & k\,(\ell(\theta_1-\epsilon,\theta_1 \mid y) - \ell(\theta_1+\epsilon,\theta_1 \mid y)) > \log n\,\}\,, \end{aligned} \tag{5.18}$$

$$\begin{aligned} K_2(n,k,y) = \{\,(k,y):\ & \mu(\underline{\theta}) < y < \mu(\theta_1-\epsilon) \text{ and} \\ & k\,(\mathbf{I}\,(\theta^*(y),\theta_1) - \ell(\theta_1+\epsilon,\theta_1 \mid y)) > \log n\,\}, \end{aligned} \tag{5.19}$$

$$K_3(n,k,y) = \{\,(k,y):\ \ y \le \mu(\underline{\theta}) \text{ and } k\,(\ell(\underline{\theta},\theta_1 \mid y) - \ell(\theta_1+\epsilon,\theta_1 \mid y)) > \log n\,\}\,. \tag{5.20}$$

Theorem 5.2. If $\epsilon > 0$ and $h_o(\theta) > 0$, $\forall \theta \in \Theta$, then the optimal policy as $n \to \infty$ can be approximated by the following policy.

$$\pi_1(n,k,y) = \begin{cases} 1, \text{ if } \ (k,y) \in \bigcup_{i=1}^{3} K_i(n,k,y) \\ 2, \ \text{ otherwise} \end{cases} \tag{5.21}$$

Proof. The approximate characterization of the sets $\underline{S}_n$ of Theorem 5.1 can be derived in the same way as in Theorem 4.2 , now making use of Lemma A.9 for the asymptotic approximation of $\overline{c}(k,y;\alpha)$. As for the approximation of $\overline{S}_n$, we note the following. It must be still true that $\overline{c}(k,y;a_2)$ tends to infinity in order for (5.12) to hold. So $\log \overline{c}(k,y;a_2)$ will be positive for $(k,y) \in \overline{S}_n$, and

$$\phi^*(n,k,y) < 2\,\frac{\overline{c}(k,y;a_2)}{\xi} \log(A\xi n)\,, \tag{5.22}$$

or in set notation

$$\overline{S}_n \subset \{\,(k,y):\ n\overline{c}(k,y;a_1) < \frac{2\,\overline{c}(k,y;a_2)}{\xi} \log(A\xi n)\,\}\,. \tag{5.23}$$

Instead of approximating $\overline{S}_n$, we obtain asymptotic characterizations for the sets on the right hand side of (5.23) . Following the same reasoning as in Theorem 4.2, it can be shown that these sets, as well as $\underline{S}_n$, are approximately described by $K_i(n,k,y)$ as they were defined in (5.18)–(5.20) .

6. CONCLUSIONS AND FURTHER WORK. The asymptotic policy of Theorem 4.2 has interesting properties that are intuitively expected. In each step, if the average of the observed samples taken from the unknown experiment E_2 exceeds the expected value of the

outcome for the known experiment E_1, i.e. $y \geq \mu(\theta_1)$, we continue sampling from E_2. Otherwise the decision is based on the quantity

$$G(k,y) = \begin{cases} k\,\mathbf{I}(\theta^*(y),\theta_1), & \text{if } \mu(\underline{\theta}) < y < \mu(\theta_1) \\ k\,\ell(\underline{\theta},\theta_1), & \text{if } y \leq \mu(\underline{\theta}) \end{cases} \tag{6.1}$$

where k denotes the number of samples taken from E_2, $\theta^*(y)$ is the maximum likelihood estimate of the unknown parameter θ_2 of E_2 based on the average y of the previous k outcomes, and $\mathrm{I}(\theta^*(y),\theta_1)$ is the Kullback–Leibler information number, which represents in some sense the estimated distance between the distributions of the two experiments (Kullback and Leibler (1951)). We continue or stop, according to whether $G(k,y) \leq \log n$ or $G(k,y) > \log n$, respectively. Note that $G(k,y)$ increases when either the number of available samples or the Kullback–Leibler information number increases. So this quantity is a measure of the confidence that the true value of θ_2 is really less than θ_1, when the sample average we have observed is less than $\mu(\theta_1)$.

The following conjectures concerning the asymptotically optimal policy for the case that there are m unknown experiments to be compared, instead of one known and one unknown, (i.e., the multi–armed bandit problem) can be made. A key idea here would be to consider the policy described in Theorem 5.1 as a function of the value θ_1 of the known experiment, as we did in Remark 5.1.b. Using similar sufficient statistics to those used for E_2, we can compute, using (4.36) and (4.37), for each unknown experiment E_i, $i = 1,\ldots,m$, a value θ_{1i} of a hypothetical known experiment $E_{1\mathrm{i}}$, which would make it indifferent to continue sampling from E_i or switch to $E_{1\mathrm{i}}$ for the remaining samples. Then we can compare the "index" values $\theta_{1\mathrm{i}}$, and take the next sample from the experiment with the largest index value. We shall deal with a rigorous statement and justification of these conjectures in a next paper. The idea to replace the unknown parameters with indices equivalent with them in some appropriate sense, appears in the fundamental papers of Lai and Robbins (1985), as we have already discussed, and Gittins (1979), which deals with the discounted infinite horizon version of the multi–armed bandit problem.

APPENDIX

The next lemma, summarizes properties of the Kullback–Leibler information number $\mathbf{I}(\rho,\tau)$

$$\mathrm{I}(\sigma,\tau) = \mathrm{E}_\sigma[\log \tfrac{f(X|\sigma)}{f(X|\tau)}]. \tag{A.1}$$

Lemma A.1. When f belongs to the one parameter exponential family (2.1),

$$\mathrm{I}(\sigma,\tau) = (\sigma-\tau)\,\mu(\sigma) - (\psi(\sigma)-\psi(\tau)), \tag{A.2}$$

$$\mathrm{I}(\sigma,\tau) = \int_\sigma^\tau (\tau-\theta)\,\psi''(\theta)\,d\theta, \tag{A.3}$$

$$\zeta_1 \tfrac{(\tau-\sigma)^2}{2} \leq \mathrm{I}(\sigma,\tau) \leq \zeta_2 \tfrac{(\tau-\sigma)^2}{2}. \tag{A.4}$$

Proof. For (A.1) and (A.2) see Lai (1987). (A.4) is immediate from (2.2) and (A.3).

Lemma A.2 indicates a useful relationship between the log-likelihood ratio $\ell(\theta,\theta_1 \mid x)$ defined in (3.13) and the Kullback Leibler information number.

Lemma A.2. **a)** $\ell(\theta,\theta_1 \mid x)$ is concave in θ.

b) $\forall x \in \mathbb{R}:\ \exists\, \theta^* = \theta^*(x)$, such that $\ell(\theta^*,\theta_1 \mid x) = \max_{\theta \in \Theta} \ell(\theta, \theta_1 \mid x)$, where

$$\theta^*(x) = \begin{cases} \mu^{-1}(x)\ , & \text{if } \mu^{-1}(x) \in \Theta \\ \overline{\theta}\ , & \text{if } \mu^{-1}(x) \notin \Theta \text{ and } \ell(\overline{\theta},\theta_1 \mid x) > \ell(\underline{\theta}, \theta_1 \mid x) \\ \underline{\theta}\ , & \text{if } \mu^{-1}(x) \notin \Theta \text{ and } \ell(\overline{\theta},\theta_1 \mid x) \le \ell(\underline{\theta},\theta_1 \mid x)\ . \end{cases} \tag{A.5}$$

Moreover, if $\mu^{-1}(x) \in \Theta$, then

$$\ell(\theta^*, \theta_1 \mid x) = \mathrm{I}(\theta^*,\theta_1)\,. \tag{A.6}$$

c) If $x < \mu(\underline{\theta})$, then

$$\ell(\underline{\theta}, \theta_1 \mid x) > 0\,. \tag{A.7}$$

Proof. From (3.18)

$$\frac{\partial\, \ell(\theta, \theta_1 \mid x)}{\partial\, \theta} = x - \mu(\theta)\,, \tag{A.8}$$

$$\frac{\partial^2 \ell(\theta, \theta_1 \mid x)}{\partial\, \theta^2} = -\mu'(\theta) = -\psi''(\theta) < 0\,. \tag{A.9}$$

Hence $\ell(\theta,\theta_1 \mid x)$ is concave in θ, and its maximum in $\theta \in \Theta$ is attained either at the point where $\ell_\theta = 0$, i.e., at $\theta^* = \mu^{-1}(x)$, if this point belongs to Θ, or else at one of the end-points. Furthermore,

$$\begin{aligned} \ell(\mu^{-1}(x), \theta_1 \mid x) &= (\mu^{-1}(x) - \theta_1)\, x - (\psi(\mu^{-1}(x)) - \psi(\theta_1)) \\ &= (\mu^{-1}(x) - \theta_1)\, \mu(\mu^{-1}(x)) - (\psi(\mu^{-1}(x)) - \psi(\theta_1)) \\ &= \mathrm{I}(\mu^{-1}(x), \theta_1)\,. \end{aligned} \tag{A.10}$$

This proves (a) and (b).

For (c) we first note that every concave function has at most two roots, lying on opposite sides with respect to its maximizing value. Hence $\forall x \in \mathbb{R}$, the equation $\ell(\theta,\theta_1 \mid x) = 0$, besides θ_1, has at most one more solution $\widetilde{\theta}(x)$, possibly not in Θ, which has the following property

$$\widetilde{\theta}(x) < \mu^{-1}(x) < \theta_1 \quad \text{if } x < \mu(\theta_1)\,, \text{ and}$$

$$\theta_1 < \mu^{-1}(x) < \widetilde{\theta}(x) \quad \text{if } x > \mu(\theta_1)\ .$$

When $x < \mu(\underline{\theta}) < \mu(\theta_1)$ it is true that $\widetilde{\theta}(x) < \mu^{-1}(x) < \underline{\theta} < \theta_1$, thus $\ell(\underline{\theta},\theta_1 \mid x) > 0$.

Lemma A.3. Let $x(n)$ denote the solution of the equation

$$\frac{A}{x^{\alpha} e^{\lambda x}} = \frac{1}{n}\ , \tag{A.11}$$

for $x, n > 0$, and constants $\alpha, A, \lambda > 0$. Then

a) (A.11) holds with " $<$ " for $x > x(n)$.

b) There exists a function $\epsilon(n)$ such that
$\epsilon(n) \sim \alpha \log(\log n)$ as $n \to \infty$ and $\lambda x(n) = \log n - \epsilon(n)$.

Proof. **a)** For $A, \lambda > 0$ the left hand side of (A.11) is increasing in x.

b) For $x = x(n)$ (A.11) can be rewritten

$$\lambda x\,(n) - \log n = \ -\ \alpha \log x(n)\ +\ \log A \tag{A.12}$$

from which it follows that

$$x(n) = \tfrac{\log n - \alpha \log x(n) + \log A}{\lambda}.$$

Substituting this expression for $x(n)$ in the right hand side of (A.12),

$$\lambda x(n) - \log n = \ -\ \alpha \log\left(\tfrac{\log n - \alpha \log x(n) + \log A}{\lambda}\right)\ +\ \log A.$$

Let $\epsilon(n) = \log n - \lambda x(n)$, $B = \alpha \log \lambda + \log A$. Then

$$\epsilon(n) = \alpha \log\left(\tfrac{\log n - \alpha \log x(n) + \log A}{\lambda}\right)\ - \log A$$

$$= \alpha \log\Big(\log n\Big(1 - \tfrac{\alpha \log x(n)}{\log n} + \tfrac{\log A}{\log n}\Big)\Big)\ -\ \log \lambda - \log A,$$

or equivalently

$$\epsilon(n) = \alpha \log\log n\ + \alpha \log\Big(1 - \tfrac{\alpha \log x(n)}{\log n} + \tfrac{\log A}{\log n}\Big)\ -\ B. \tag{A.13}$$

From (A.11) it follows that $x(n) > 0$ and $\lim\limits_{n\to\infty} x(n) = \infty$, thus $\lim\limits_{n\to\infty} \tfrac{\log x(n)}{x(n)} = 0$. Therefore, rewriting (A.12) as

$$\lambda - \tfrac{\log n}{x(n)} = \ -\ \tfrac{\alpha \log x(n)}{x(n)} + \tfrac{\log A}{x(n)}$$

it follows that

$$\lim_{n\to\infty} \tfrac{x(n)}{\log n} = \tfrac{1}{\lambda}\ ,$$

and

$$\lim_{n\to\infty} \frac{\log x(n)}{\log n} = \lim_{n\to\infty} \frac{\log x(n)}{x(n)} \frac{x(n)}{\log n} = 0\,. \tag{A.14}$$

From (A.14) it follows that

$$\lim_{n\to\infty} \log\Big(1 - \frac{\alpha \log x(n)}{\log n} + \frac{\log A}{\log n}\Big) = 0\,,$$

thus, from (A.13), as $n \to \infty, \epsilon(n) \sim \alpha \log\log n$. This completes the proof.

The following lemmata describe useful properties of the transformed one step regret functions.

Lemma A.4. The quantities $\overline{c}(k,y;\alpha)$ defined in (3.27), (3.28) satisfy

a) $\overline{c}(k,y;\alpha) > 0\,,\ \forall\ k,y\,.$ (A.15)

b) $\forall\, k\,,\ \overline{c}(k,y;a_1)$ is increasing in y.
$\forall\, k\,,\ \overline{c}(k,y;a_2)$ is decreasing in y.

c) $\mathrm{E}_{\theta_1}[\,\overline{c}\,(k+1, m(k,y,X)\,;\,\alpha\,)\,] = \overline{c}(k,y\,;\,\alpha)\,.$ (A.16)

Proof. The proof of (a) is immediate by definition. Part (b) can be proved by taking the derivative in y and observing that $\delta(\theta)\, d(\theta) \geq 0$. Part (c) expresses an intuitive martingale property, which can be easily proved as follows. For α=1, we use (3.19) and (3.27) to obtain

$$\begin{aligned}
&\mathrm{E}_{\theta_1}[\,\overline{c}\,(k+1, m(k,y,X)\,;a_1\,)\,] = \int \overline{c}(k+1,m(k,y,x)\,;a_1)\, f(x \mid \theta_1)\, \nu(dx) \\
&= \int_{\mathbb{R}} \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\, e^{(k+1)\,\ell(\theta,\theta_1 | m(k,y,x))} dH_o(\theta) f(x \mid \theta_1)\, \nu(dx) \\
&= \int_{\mathbb{R}} \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\, e^{k\ell(\theta,\theta_1 | y) + \ell(\theta,\theta_1 | x)}\, dH_o(\theta) f(x \mid \theta_1)\, \nu(dx) \\
&= \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\, e^{k\ell(\theta,\theta_1|y)} \int_{\mathbb{R}} f(x \mid \theta)\, \nu(dx)\, dH_o(\theta) \\
&= \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\, e^{k\ell(\theta,\theta_1 | y)} dH_o(\theta) \;=\; \overline{c}(k,y\,;a_1)\,.
\end{aligned} \tag{A.17}$$

The case α=2 can be proved similarly.

Let us define the function

$$\gamma(k,y) \;=\; \min\{\,\overline{c}(k,y;a_1)\,,\,\overline{c}(k,y;a_2)\,\}\,. \tag{A.18}$$

For this quantity the following result holds.

Lemma A.5. $\gamma(k,y) = O(\frac{1}{k})$ uniformly in y.

Proof. It suffices to prove the following intermediate claim.

$$\exists\, A > 0 : \quad \overline{c}(k,\mu(\theta_1)\,;\, a_i) < \frac{A}{k}\ ,\ \forall k = 1,2,\ldots\ ,\ i = 1,2\ . \tag{A.19}$$

Indeed, suppose that (A.19) holds. Then we consider two cases.

Case a. $y \geq \mu(\theta_1)$. From Lemma A.4.b

$$\gamma(k,y) \leq \overline{c}(k,y;a_2) \leq \overline{c}(k,\mu(\theta_1)\,;\,a_2) < \frac{A}{k}\ . \tag{A.20}$$

Case b. $y < \mu(\theta_1)$. In the same way

$$\gamma(k,y) \leq \overline{c}(k,y;a_1) \leq \overline{c}(k,\mu(\theta_1)\,;\,a_1) < \frac{A}{k}\ . \tag{A.21}$$

So $\gamma(k,y) < \frac{A}{k}$, $\forall k,y$, which proves the lemma.

We next prove (A.19).

From (3.27)

$$\overline{c}(k,\mu(\theta_1)\,;\,a_1) = \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\, e^{k\,\ell(\theta,\theta_1|\,\mu(\theta_1))}\, h_o(\theta)\, d\theta\ . \tag{A.22}$$

But $\ell(\theta,\theta_1 \mid \mu(\theta_1)) = (\theta-\theta_1)\mu(\theta_1) - (\psi(\theta)-\psi(\theta_1)) = -\mathrm{I}\,(\theta_1,\theta)$, and from (A.12)

$$-\zeta_2\,\frac{(\theta-\theta_1)^2}{2} \leq \ell(\theta,\theta_1 \mid \mu(\theta_1)) \leq -\zeta_1\,\frac{(\theta-\theta_1)^2}{2}\ . \tag{A.23}$$

From mean value theorems of calculus we obtain

$$\delta(\theta) = \mu(\theta) - \mu(\theta_1) = \psi''(\xi)\,(\theta-\theta_1)\,, \tag{A.24}$$

for some $\xi \in (\theta_1,\theta)$. So for $\theta \geq \theta_1$

$$\delta(\theta) \leq \zeta_2(\theta-\theta_1)\,. \tag{A.25}$$

From (A.23) and (A.25) we obtain

$$\overline{c}(k,\mu(\theta_1)\,;\,a_1) \leq \zeta_2\ \overline{h}_o \int_{\theta_1}^{\overline{\theta}} (\theta-\theta_1)\, e^{-k\,\zeta_1\,\frac{(\theta-\theta_1)^2}{2}}\, d\theta\ . \tag{A.26}$$

Let $A = \frac{\zeta_2\,\overline{h}_o}{\zeta_1}$. Then

$$\overline{c}(k,\mu(\theta_1)\,;\,a_1) \leq \frac{A}{k}\ \left(1 - e^{-k\,\zeta_1\,\frac{(\overline{\theta}-\theta_1)^2}{2}}\right) < \frac{A}{k}\ . \tag{A.27}$$

Following the same reasoning it can be shown that

$$\overline{c}(k,\mu(\theta_1)\,;\,a_2) < \frac{A}{k}\ , \tag{A.28}$$

and (A.19) is proved. This completes the proof of the lemma.

The next two lemmata describe asymptotic properties of $\overline{c}(k,y;\alpha)$.

Lemma A.6. If $h_o(\theta) > 0 \ \forall \theta \in \Theta$ and $y < \mu(\theta_1)$, then the following asymptotic relations hold, as $k \to \infty$.

1. For $a = a_1$,

$$\overline{c}(k,y;a_1) \ \sim \ \frac{h_o(\theta_1)\,\psi''(\theta_1)}{(y-\mu(\theta_1))^2\,k^2} \ . \tag{A.29}$$

2. For $a = a_2$,

a) If $\mu(\underline{\theta}) < y < \mu(\theta_1)$, then

$$\overline{c}(k,y;a_2) \ \sim \ -\,\delta(\theta^*(y))\,h_o(\theta^*(y))\,e^{kI(\theta^*(y),\theta_1)}\sqrt{\frac{2\pi}{\psi''(\theta^*(y))k}} \ . \tag{A.30}$$

b) If $y < \mu(\underline{\theta})$, then

$$\overline{c}(k,y;a_2) \sim \ \frac{\delta(\underline{\theta})\,h_o(\underline{\theta})}{y-\mu(\underline{\theta})} \ \frac{e^{k\,\ell(\underline{\theta},\theta_1\,|\,y)}}{k} \ . \tag{A.31}$$

c) If $y = \mu(\underline{\theta})$, then

$$\overline{c}(k,y;a_2) \ \sim \ -\,\delta(\underline{\theta})\,h_o(\underline{\theta}) \ e^{k\ell(\underline{\theta},\theta_1\,|\,y)} \ \sqrt{\frac{\pi}{\psi''(\underline{\theta})\,k}} \ . \tag{A.32}$$

Proof. The proof is based on the Laplace method for approximating integrals of exponential functions (cf. Erdélyi (1956)). From (3.27) we have

$$\overline{c}(k,y;a_1) \ = \ \int_{\theta_1}^{\overline{\theta}} \delta(\theta)\,e^{k\ell(\theta,\theta_1\,|\,y)}\,h_o(\theta)\,d\theta \ , \tag{A.33}$$

$$\overline{c}(k,y;a_2) \ = \ -\int\limits_{\theta\le\theta_1} \delta(\theta)\,e^{k\ell(\theta,\theta_1\,|\,y)}dH_o(\theta) \ . \tag{A.34}$$

From Lemma A.2 we see that, when $y < \mu(\theta_1)$, $\ell(\theta,\theta_1 \mid y)$ attains its maximum value in $[\,\theta_1,\overline{\theta}]$ for $\theta=\theta_1$, and in $[\,\underline{\theta},\theta_1]$ for $\theta = \tau$, where $\tau = \underline{\theta}$ or $\tau = \theta^*(y)$, according to whether $\mu(\underline{\theta}) < y < \mu(\theta_1)$, or $y = \mu(\underline{\theta})$ respectively. Therefore, when $k \to \infty$, the main contribution to the value of $\overline{c}(k,y;a)$ for $a = 1,2$, will arise from the values of the integrand in a neighborhood of this maximizing value. The main idea of the Laplace method is to introduce a new variable of integration z, such that

$$z^2 \ = \ \ell(\tau,\theta_1 \mid y) - \ell(\theta,\theta_1 \mid y) \ , \tag{A.35}$$

$$z \ < \ 0 \ \ (>0) \ , \ \text{for} \ \ \theta < \tau \ \ (\theta > \tau) \ , \tag{A.36}$$

and to reduce the area of integration in a neighborhood of τ.

We first consider the case a_1. From (A.35)

$$2\,z\,dz \;=\; -\,(y-\mu(\theta))\,d\theta\,, \tag{A.37}$$

$$d\theta \;=\; -\,2\,\frac{z}{y-\mu(\theta)}\,dz\,. \tag{A.38}$$

Here $\tau=\theta_1$, $\ell(\tau,\theta_1 \mid y)=0$, and so for $\eta>0$

$$\overline{c}(k,y;a_1) \;\sim\; \int_{\theta_1}^{\theta_1+\eta} \delta(\theta)\,e^{kl(\theta,\theta_1\mid y)}\,h_o(\theta)\,d\theta \;=\; -\int_0^Z 2\,z\,\frac{\delta(\theta(z))\,h_o(\theta(z))}{y-\mu(\theta(z))}\,e^{-k\,z^2}\,dz\,, \tag{A.39}$$

where

$$Z \;=\sqrt{-\,\ell(\theta_1+\eta\,,\theta_1 \mid y)}\,. \tag{A.40}$$

Since only the values of z close to zero are significant, we can expand the region of integration to infinity

$$\overline{c}(k,y;a_1) \;\sim\; -\int_0^\infty 2\,z\,\frac{\delta(\theta)h_o(\theta(z))}{y-\mu(\theta(z))}\;e^{-k\,z^2}\,dz\;. \tag{A.41}$$

We can further approximate the above expression by substituting $\frac{h_o(\theta)}{y-\mu(\theta)}$ with its value at $z=0$, i.e. at $\theta=\theta_1$. Then we integrate by parts, considering θ a function of the integration variable z.

$$\begin{aligned}
\overline{c}(k,y;a_1) \;&\sim\; -\,\frac{h_o(\theta_1)}{y-\mu(\theta_1)}\;\int_0^\infty 2\,z\,\delta(\theta(z))\;e^{-k\,z^2}\,dz \\
&=\; \frac{h_o(\theta_1)}{(y-\mu(\theta_1))\,k}\;\int_0^\infty \delta(\theta(z))\;de^{-k\,z^2}\,dz \\
&=\; -\,\frac{h_o(\theta_1)}{(y-\mu(\theta_1))\,k}\;\int_0^\infty e^{-kz^2}\;\frac{d\delta(\theta(z))}{dz}\;dz\,.
\end{aligned} \tag{A.42}$$

But

$$\frac{d\delta(\theta(z))}{dz} \;=\; \psi''(\theta(z))\;\frac{d\theta}{dz} \;=\; \psi''(\theta(z))\;\frac{-2\,z}{y-\mu(\theta(z))}\,, \tag{A.43}$$

from (A.38). Substituting again $\frac{\psi''(\theta(z))}{y-\mu(\theta(z))}$ with its value at $z=0$,

$$\overline{c}(k,y;a_1) \;\sim\; -\,\frac{h_o(\theta_1)\,\psi''(\theta_1)}{(y-\mu(\theta_1))^2\,k}\;\int_0^\infty -2z\,e^{-kz^2}\;dz$$

$$= \quad \frac{h_o(\theta_1)\,\psi''(\theta_1)}{(y-\mu(\theta_1))^2\,k^2} \quad , \tag{A.44}$$

and (A.29) is proved.
We now consider the case $a = 2$ and each one of the three subcases.

2.a) $\mu(\theta_1) < y < \mu(\overline{\theta})$. Here $\tau = \theta^*(y)$, while from Lemma A.2, $\ell(\tau, \theta_1 \mid y) = \mathrm{I}(\theta^*(y), \theta_1)$. Following the same method, we obtain a relation analogous to (A.39) for $a = 2$, namely for some $\eta_1, \eta_2 > 0$

$$\overline{c}(k,y;\, a_2) \;\sim\; -\int_{\theta^*-\eta_1}^{\theta^*+\eta_2} \delta(\theta)\, e^{k\,\ell(\theta,\theta_1|y)}\, h_o(\theta)\, d\theta$$

$$= \; e^{k\mathrm{I}(\theta^*(y),\theta_1)} \int_{Z_1}^{Z_2} 2\,z\, \frac{\delta(\theta(z))\, h_o(\theta(z))}{y-\mu(\theta(z))}\; e^{-k\,z^2}\, dz \;, \tag{A.45}$$

where

$$Z_1 \;=\; -\sqrt{\mathrm{I}(\theta^*(y), \theta_1) \;-\ell(\theta^*(y) - \eta_1, \theta_1 \mid y)}\,, \tag{A.46}$$

$$Z_2 \;=\; \sqrt{\mathrm{I}(\theta^*(y), \theta_1) \;-\ell(\theta^*(y) + \eta_2, \theta_1 \mid y)}\,. \tag{A.47}$$

We expand the integration region from $-\infty$ to ∞, and since in this case $z = 0$ corresponds to $\theta = \theta^*(y)$, we substitute $\delta(\theta(z))\, h_o(\theta(z))$ with $\delta(\theta^*(y))\, h_o(\theta^*(y))$.

$$\overline{c}(k,y;\, a_2) \;\sim\; 2\,\delta(\theta^*(y))\, h_o(\theta^*(y))\; e^{k\mathrm{I}(\theta^*(y),\theta_1)} \int_{-\infty}^{\infty} \frac{z}{y-\mu(\theta(z))}\; e^{-k\,z^2}\, dz\,. \tag{A.48}$$

For $z = 0$ it is $y - \mu(\theta(0)) = 0$. Applying de l' Hospital's rule to find the limiting value of $\frac{z}{y-\mu(\theta(z))}$ when $z \to 0$ yields

$$\lim_{z\,\to\,0} \frac{z}{y-\mu(\theta(z))} \;=\lim_{z\,\to\,0} \frac{1}{-\psi''(\theta(z))\,\frac{d\theta}{dz}} \;=\; -\frac{1}{\psi''(\theta^*(y))}\; \lim_{z\,\to\,0} \frac{1}{\frac{-2\,z}{y-\mu(\theta(z))}}\;,$$

thus

$$\Big(\lim_{z\,\to\,0} \frac{z}{y-\mu(\theta(z))}\Big)^2 \;=\; \frac{1}{2\psi''(\theta^*(y))}\;. \tag{A.49}$$

We also note that $\frac{z}{y-\mu(\theta(z))} \;<\; 0$ for all z , which implies that

$$\lim_{z\,\to\,0} \frac{z}{y-\mu(\theta(z))} \;=\; -\sqrt{\frac{1}{2\psi''(\theta^*(y))}}\;, \tag{A.50}$$

and the integral becomes

$$\overline{c}(k,y;a_2) \;\sim\; -\sqrt{\frac{2}{\psi''(\theta^*(y))}}\;\delta(\theta^*(y))\,h_o(\theta^*(y))\;e^{kI(\theta^*(y),\theta_1)}\int_{-\infty}^{\infty} e^{-k\,z^2}\,dz$$

$$= \;-\sqrt{\frac{2}{\psi''(\theta^*(y))}}\;\delta(\theta^*(y))\,h_o(\theta^*(y))\;e^{kI(\theta^*(y),\theta_1)}\;\sqrt{\frac{\pi}{k}}\;. \qquad (A.51)$$

Thus we have established (A.30).

The remaining cases to be proved are 2.b) and 2.c) , which correspond to $y \le \mu(\underline{\theta})$. Now we have that $\tau = \underline{\theta},\ \ell(\tau,\theta_1 \mid y) = \ell(\underline{\theta},\theta_1 \mid y) > 0$ from Lemma A.3.c . Performing the transformations (A.43) , (A.44) and reducing the integration region to a neighborhood of $\underline{\theta}$, i.e. $[\,\underline{\theta},\underline{\theta}+\eta]$ for some $\eta > 0$, we get

$$\overline{c}(k,y;a_2) \;\sim\; \int_{\underline{\theta}}^{\underline{\theta}+\eta} \delta(\theta)\,e^{k\ell(\theta,\theta_1|y)}\,h_o(\theta)\,d\theta$$

$$= \;-\,e^{k\ell(\underline{\theta},\theta_1|y)}\int_0^Z 2z\;\frac{\delta(\theta(z))\,h_o(\theta(z))}{y-\mu(\theta(z))}\;e^{-k\,z^2}\,dz\;, \qquad (A.52)$$

where

$$Z \;=\; \sqrt{\ell(\underline{\theta},\theta_1 \mid y) - \ell(\underline{\theta}+\eta\,,\theta_1 \mid y)}\;. \qquad (A.53)$$

2.b) When $y = \mu(\underline{\theta})$, a relation analogous to (A.59) can be established

$$\lim_{z \to 0^-} \frac{z}{y-\mu(\theta(z))} \;=\; -\sqrt{\frac{1}{2\psi''(\underline{\theta})}}\;. \qquad (A.54)$$

With the same reasoning as in the previous cases

$$\overline{c}(k,y;a_2) \;\sim\; -\sqrt{\frac{2}{\psi''(\underline{\theta})}}\;\delta(\underline{\theta})\,h_o(\underline{\theta})\;e^{k\,\ell(\underline{\theta},\theta_1)}\int_0^{\infty} e^{-k\,z^2}\,dz$$

$$= \;-\sqrt{\frac{2}{\psi''(\underline{\theta})}}\;\delta(\underline{\theta})\,h_o(\underline{\theta})\;e^{k\ell(\underline{\theta},\theta_1)}\;\sqrt{\frac{\pi}{2k}}\;, \qquad (A.55)$$

which proves (A.35).

2.c) Finally when $y < \mu(\underline{\theta})$,

$$\overline{c}(k,y;a_2) \;\sim\; -\;\frac{\delta(\underline{\theta})\,h_o(\underline{\theta})}{y-\mu(\underline{\theta})}\;e^{k\,\ell(\underline{\theta},\theta_1)}\int_0^{\infty}(-2z)\,e^{-k\,z^2}\,dz$$

$$= \;-\;\frac{\delta(\underline{\theta})\,h_o(\underline{\theta})}{y-\mu(\underline{\theta})}\;e^{k\,\ell(\underline{\theta},\theta_1)}\;\frac{1}{k}\int_0^{\infty} de^{-k\,z^2} \;=\; \frac{\delta(\underline{\theta})\,h_o(\underline{\theta})}{y-\mu(\underline{\theta})}\;\frac{e^{k\,\ell(\underline{\theta},\theta_1)}}{k}\;. \qquad (A.56)$$

Remark A.1. In Lemma A.5 we made the assumption that the prior p.d.f h_o is positive on the entire parameter space $\Theta = [\underline{\theta},\overline{\theta}]$. This ensures that the values for which the log–likelihood ratio attains its maximum value in the integration region are independent of $h_o(\theta)$. When this assumption is dropped, the same line of argument remains valid. However the expansion of the integrals becomes more tedious, since one has to consider separately cases such as $h_o(\theta) = 0$, for $\theta \le \theta_1 + \epsilon$, or for $\theta \ge \theta_1 - \epsilon$, or $\theta_1 - \epsilon \le \theta \le \theta_1 + \epsilon$. According to each individual case examined, one must integrate in a neighborhood of a value θ, which is closest to the maximizing value, and has positive prior p.d.f. The corresponding asymptotic expressions cannot be given in advance for the general case, but can be derived following the same general approach.

Lemma A.7. If $h_o(\theta) > 0$, $\forall \theta \in \Theta$ and $y \ge \mu(\theta_1)$, then the following asymptotic relations hold, as $k \to \infty$.

1. For $a = a_1$,

a) If $y = \mu(\theta_1)$, then

$$\overline{c}(k,y,a_1) \sim \frac{h_o(\theta_1)}{k} . \tag{A.57}$$

b) If $\mu(\theta_1) < y < \mu(\overline{\theta})$, then

$$\overline{c}(k,y,a_1) \sim \delta(\theta^*(y))\, h_o(\theta^*(y))\, e^{k\mathrm{I}(\theta^*(y),\theta_1)} \sqrt{\frac{2\pi}{\psi''(\theta^*(y))k}} . \tag{A.58}$$

4. If $y = \mu(\overline{\theta})$, then $\overline{c}(k,y,a_1) \sim \delta(\overline{\theta})\, h_o(\overline{\theta})\; e^{k\ell(\overline{\theta},\theta_1 \mid y)} \sqrt{\frac{\pi}{\psi''(\overline{\theta})\, k}}$. (A.59)

5. If $y > \mu(\overline{\theta})$, then $\overline{c}(k,y,a_1) \sim \frac{\delta(\overline{\theta})\, h_o(\overline{\theta})}{y-\mu(\overline{\theta})} \; \frac{e^{k\ell(\overline{\theta},\theta_1 \mid y)}}{k}$. (A.60)

2. For $a = a_2$,

a) If $\mu(\underline{\theta}) < y < \mu(\theta_1)$, then

$$\overline{c}(k,y;a_2) \sim -\delta(\theta^*(y))\, h_o(\theta^*(y))\, e^{k\mathrm{I}(\theta^*(y),\theta_1)} \sqrt{\frac{2\pi}{\psi''(\theta^*(y))k}} . \tag{A.61}$$

b) If $y < \mu(\underline{\theta})$, then

$$\overline{c}(k,y;a_2) \sim \frac{\delta(\underline{\theta})\, h_o(\underline{\theta})}{y-\mu(\underline{\theta})} \; \frac{e^{k\ell(\underline{\theta},\theta_1 \mid y)}}{k} . \tag{A.62}$$

c) If $y = \mu(\underline{\theta})$, then

$$\overline{c}(k,y;a_2) \sim -\delta(\underline{\theta})\, h_o(\underline{\theta})\; e^{k\ell(\underline{\theta},\theta_1 \mid y)} \sqrt{\frac{\pi}{\psi''(\underline{\theta})\, k}} . \tag{A.63}$$

Proofs of Section 5 .

First note that Lemma A.4 holds unaltered for the more general loss function of Section 5.

Lemma A.8. In the case $\beta \geq 1$ and $\epsilon > 0$, $\gamma(k,y) = \mathrm{O}(\mathrm{e}^{-k\zeta_1 \frac{\epsilon^2}{2}})$, uniformly in y .

Proof. The proof goes along the same lines as in Lemma A.5 , up to relation (A.26) , which takes the form

$$\begin{aligned}
\overline{c}(k,\mu(\theta_1)\,;a_1) \;&\leq\; \zeta_2^{\beta}\;\overline{h}_o \int_{\theta_1+\epsilon}^{\overline{\theta}} (\theta-\theta_1)^{\beta}\, e^{-k\zeta_1 \frac{(\theta-\theta_1)^2}{2}}\, d\theta \\
&\leq\; \zeta_2^{\beta}\;\overline{h}_o\,(\overline{\theta}-\theta_1)^{\beta} \int_{\theta_1+\epsilon}^{\overline{\theta}} e^{-k\zeta_1 \frac{(\theta-\theta_1)^2}{2}}\, d\theta \\
&\leq\; \zeta_2^{\beta}\;\overline{h}_o\,(\overline{\theta}-\theta_1)^{\beta}\; e^{-k\zeta_1 \frac{\epsilon^2}{2}} \int_{\theta_1+\epsilon}^{\overline{\theta}} d\theta \\
&=\; \zeta_2^{\beta}\;\overline{h}_o\,(\overline{\theta}-\theta_1)^{\beta}\;(\overline{\theta}-\theta_1-\epsilon)\, e^{-k\zeta_1 \frac{\epsilon^2}{2}} \;=\; A_1 e^{-k\zeta_1 \frac{\epsilon^2}{2}}\;. \qquad \text{(A.64)}
\end{aligned}$$

Similarly we can show that there exists $A_2 < \infty$ such that

$$\overline{c}(k,\mu(\theta_1)\,;a_1) \;\leq\; A_2 e^{-k\zeta_1 \frac{\epsilon^2}{2}}\;, \qquad \text{(A.65)}$$

and so

$$\gamma(k,y) \;\leq\; A\, e^{-k\zeta_1 \frac{\epsilon^2}{2}}\;, \qquad \text{(A.66)}$$

with $A = \max\{A_1\,, A_2\}$.

For the case $\epsilon > 0$ we can prove the following Lemma, using the same method as in Lemma A.6.

Lemma A.7. If $h_o(\theta) > 0$, $\forall \theta \in \Theta$, then, according to the value of y, $\overline{c}(k,y;\alpha)$ has the following asymptotic forms, as $k \to \infty$.

1. For $\alpha = a_1$, if $y < \mu(\theta_1 + \epsilon)$, then

$$\overline{c}\,(k,y;a_1) \;\sim\; \frac{h_o(\theta_1+\epsilon)\,(\delta(\theta_1+\epsilon))^{\beta}}{\mu(\theta_1+\epsilon)-y} \;\; \frac{e^{k\ell(\theta_1+\epsilon,\theta_1\,|\,y)}}{k}\;. \qquad \text{(A.63)}$$

2. For $\alpha = a_2$,

a) If $y > \mu(\theta_1 - \epsilon)$, then

$$\overline{c}\,(k,y;a_2) \;\sim\; \frac{h_o(\theta_1-\epsilon)\,(-\delta(\theta_1-\epsilon))^{\beta}}{y-\mu(\theta_1-\epsilon)} \quad \frac{e^{k\,\ell(\theta_1-\epsilon,\theta_1\,|\,y)}}{k}\;. \tag{A.64}$$

b) If $y=\mu(\theta_1-\epsilon)$, then

$$\overline{c}\,(k,y;a_2) \;\sim\; h_o(\theta_1-\epsilon)\,(-\delta(\theta_1-\epsilon))^{\beta} e^{k\,\ell(\theta_1-\epsilon,\theta_1\,|\,y)} \sqrt{\frac{\pi}{k\psi''(\theta_1-\epsilon)}}\;. \tag{A.65}$$

c) If $\mu(\underline{\theta})<y<\mu(\theta_1-\epsilon)$, then

$$\overline{c}\,(k,y;a_2) \;\sim\; (-\delta(\theta^*(y)))^{\beta}\, h_o(\theta^*(y))\, e^{k\,\mathrm{I}\,(\theta^*(y),\theta_1)} \sqrt{\frac{2\pi}{\psi''(\theta^*(y))k}}\;. \tag{A.66}$$

d) If $y=\mu(\underline{\theta})$, then

$$\overline{c}\,(k,y;a_2) \sim\; (-\delta(\underline{\theta}))^{\beta}\, h_o(\underline{\theta})\; e^{k\,\ell(\underline{\theta},\theta_1\,|\,y)} \sqrt{\frac{\pi}{\psi''(\underline{\theta})\,k}}\;. \tag{A.67}$$

e) If $y<\mu(\underline{\theta})$, then

$$\overline{c}\,(k,y;a_2) \;\sim\; \frac{(-\delta(\underline{\theta}))^{\beta}\, h_o(\underline{\theta})}{\mu(\underline{\theta})-y} \quad \frac{e^{k\,\ell(\underline{\theta},\theta_1\,|\,y)}}{k}\;. \tag{A.68}$$

ACKNOWLEDGEMENT

We have had the good fortune to know and to have worked with Professor Herbert Robbins and to enjoy his insight, optimism and good humor. He was a fervent source of inspiration and encouragement for this work. We express our deep respect to the memory of Professor Robbins.

REFERENCES

1. Agrawal R., Hedge M., Teneketzis D. (1988). "Asymptotically Efficient Adaptive Allocation Rules for the Multi–armed Bandit Problem with Switching Cost", ***IEEE Trans. Autom. Contr.*** **, AC–33,** 899–906.

2. Bellman R.(1956). "A Problem in the Sequential Design of Experiments", ***Sankhyā,*** **16,** 221–229.

3. Berry D.A., Fridstet B. (1985). ***"Bandit Problems*****:** ***Sequential Allocation of Experiments***" , Chapman and Hall, New York.

4. Bradt, R. N. , Johnson, S. M. and Karlin S. (1956) "On sequential Designs for maximizing the sum of n observations", ***Ann. Math. Stat.,*** **27,** 1060–1074 .

5. Burnetas, A.N. , Katehakis, M.N. (1996). "Optimal Adaptive Policies for Sequential Allocation Problems", Advances in Applied Mathematics, **17**(2), 122–142.

6. Burnetas, A.N. , Katehakis, M.N. (1997a). "On the Finite Horizon One-Armed Bandit Problem", ***Stochatstic Analysis and Applications***, **16**(1), 845-859.

7. Burnetas, A.N. , Katehakis, M.N. (1997b). "Optimal Adaptive Policies for Markov Decision Processes", ***Mathematics of Operations Research,*** 22(1), 222–255.

8. Burnetas, A.N. , Katehakis, M.N. (1993). "On Sequencing Two Types of Tasks on a Single Processor under Incomplete Information". ***Probability in the Engineering and Informational Sciences***, **7**, 85–119.

9. Chernoff, H. and Ray, S. (1965). "A Bayes sequential sampling inspection problem", ***Ann. Math Statist. ,*** **36,** 1387–1407.

10. Chernoff, H. (1967). "Sequential models for clinical trials", ***Proc. Fifth Berkeley Symposium Math. Stat. Probab.,*** **4,** 805–812.

11. Cox, D.R. and Hinkley, D.V. (1974). ***"Theoretical Statistics"*** , Chapman and Hall, New York.

12. Dynkin, E.B. and Yushkevich, A.A. (1979). ***"Controlled Markov Processes"*** , Springer–Verlag, New York.

13. Erdélyi, A. (1956). ***"Asymptotic Expansions" ,*** Dover Publ. Inc., New York.

14. Gittins, J.C. (1979). "Bandit Processes and Dynamic Allocation Indices", ***J. Roy. Statist. Soc.,*** **B, 41,** 148–164.

15. Gittins, J. C. (1989). ***"Multi–armed bandit allocation indices***" , J. Wiley, Chichester, New York.

16. Glazebrook, K. D. and Mitchell.H.M.(2002). "An index policy for a stochastic scheduling model with improving/deteriorating jobs" ***Naval Research Logistics*** (to appear).

17. Katehakis, M. N. and Derman, C. (1987). "Computing Optimal Sequential Allocation Rules In Clinical Trials". In "Adaptive Statistical Procedures and Related Topics" (J. Van Ryzin ed.) ***I.M.S. Lecture Notes-Monograph Series***, **8**, 29–39.

18. Katehakis, M. N. and Veinott, A.F. Jr. (1987). "The Multi-Armed Bandit Problem: Decomposition and Computation". ***Math. Oper. Res.***, **22**(2)**,** 262–268.

19. Katehakis, M. N. and and H. E. Robbins (1995). ``Sequential choice from several populations", Proceedings of National Academy of Sciences U.S.A., **92**, 8584 --8565. .

20. Kullback S., Leibler R.A.(1951)."On Information and Sufficiency". ***Ann. Math. Stat.,*** **22,** 79–86.

21. Lai, T. L. (2001). "Sequential Analysis; Some Classical Problems and New Challen ges", ***Statistica Sinica,*** **11** (2), 303–352 .

22. Lai, T. L. (1987). "Adaptive Treatment Allocation and the Multi–Armed Bandit Problem", ***Annals of Statistics,*** **15**(3), 1091–1114 .

23. Lai, T. L. and Robbins, H. (1985). "Asymptotically Efficient Adaptive Allocation Rules", ***Adv. Appl. Math*** **6 ,** 4–22.

24. Robbins, H. (1952). "Some aspects of the sequential design of experiments", ***Bull. Amer. Math. Monthly,*** **58,** 527–536.

25. Schwarz, G. (1962). "Asymptotic Forms of Bayes Solutions", ***Ann. Math. Stat. , 33,*** 224–236..

26. N. Shimkin and A. Shwartz (1995). "Asymptotically Efficient Adaptive Strategies in Repeated Games, Part I: Certainty Equivalence Strategies", ***Math. Oper. Res.*** **20,** pp. 743–767.

27. N. Shimkin and A. Shwartz (1995b). "Asymptotically Efficient Adaptive Strategies in Repeated Games, Part II: Asymptotic Optimality", ***Math. Oper. Res.*** **21,** pp. 487–512.

28. Varaiya, P., Walrand, J. and Buyukkoc, C. (1985) . "Extensions of the Multiarmed Bandit Problem: The discounted Case" . ***IEEE Trans. Autom. Contr.,*** **AC-30,** 426-439.

29. Whittle , P. (1982). ***"Optimization Over Time",*** Vols. 1,2, J. Wiley, New York.