Stochastic Analysis and Applications

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsaa20

Dynamic allocation policies for the finite horizon one armed bandit problem

Apostolos N. Burnetas [a] & Michael N. Katehakis [b]

[a] Case Western Reserve University,

[b] Rutgers University,

Available online: 03 Apr 2007

DYNAMIC ALLOCATION POLICIES FOR THE FINITE HORIZON ONE ARMED BANDIT PROBLEM

APOSTOLOS N. BURNETAS
Case Western Reserve University

MICHAEL N. KATEHAKIS
Rutgers University

ABSTRACT

The unknown performance of a new experiment is to be evaluated and compared with that of an existing one over a finite horizon. The explicit structure of an optimal sequential allocation policy is obtained under pertinent reward/loss functions, when the experiments are characterized by random variables with distributions from the *one parameter exponential family.*

1. INTRODUCTION. We consider the following version of a classical problem of dynamic allocation of effort among different activities in the presence of partial information about the underlying statistical characteristics, (c.f. [1], [12]). There are two experiments denoted by E_j $(j=1,2)$. Associated with experiment E_j are i.i.d. random variables representing the outcomes (rewards) 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"). The adaptive control problem is to sequentially select the experiment to be used each period, so as to maximize the expected value of the total reward obtained during a finite horizon of length N. Furthermore we assume that the characteristics of experiment E_1 are known in advance, while those of E_2 are not, that is 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 sections 2 and 3 we postulate a prior distribution on the unknown parameter of the second experiment, and formulate the problem of maximizing the expected sum of outcomes over a finite horizon. We point out that this is equivalent to minimizing a suitably defined regret (expected loss function). The main contribution of this paper is the characterization of the structure and properties of optimal dynamic allocation policies given in Theorem 4.1.

The work on the finite horizon problem generalizes results of [3] on the binomial case. For related work see [4]. For the infinite horizon discounted reward version of this problem, see [2], [7], [8], [9], [10], [11], [13], [14].

2. PROBLEM FORMULATION. 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$^{\text{th}}$ 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. $\mu(\theta) = E(X \mid \theta)$, $\sigma^2(\theta) = \text{Var}(X \mid \theta)$.

We make the following assumptions.

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

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

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

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

Assumption 3. Parameter θ_1 is known in advance, while θ_2 is unknown. Following the Bayesian approach, we consider θ_2 as a random variable with prior distribution denoted by $H_0(\theta)$, $\theta \in \Theta$.

Remark 2.1. **a)** It is known (cf. [5]) that, for the one–parameter exponential family, $\mu(\theta)=\psi'(\theta)$ and $\sigma^2(\theta)=\psi''(\theta)$, thus $\mu(\theta)$ is strictly increasing in θ and the set $\{\mu(\theta) : \theta \in \Theta\}$ is an interval of the form $(\mu(\underline{\theta}),\mu(\bar{\theta}))$.

b) Note that if $\theta_1 \leq \underline{\theta}$ $(\theta_1 \geq \bar{\theta})$ then the problem is trivial, because then one should always choose E_2 (E_1) . From now on, we shall assume that $\underline{\theta} < \theta_1 < \bar{\theta}$.

We define the optimization problem as follows. Let t $(n=N-t)$ denote the number of samples that have already been taken (remain to be taken).

At $t=0$ we have $X_1 \sim f(x \mid \theta_1)$ and $X_2 \sim f(x \mid \theta_2)$ with respect to $\nu(dx)$, θ_1 known and $\theta_2 \sim H_0(\theta)$.

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 , X_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 , X_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

$$dH(\theta \mid d_2(k_2)) = \frac{\tilde{f}(d_2(k_2) \mid \theta)\ dH_0(\theta)}{\tilde{f}(d_2(k_2) \mid H_0)}$$

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

where $d_i(k_i)=(d_i(k_i-1), y_{i,k_i})$, $H(\theta \mid d_2(0)) = H_0(\theta)$, and $\tilde{f}(d_2(k_2) \mid \theta)$ (respectively $\tilde{f}(d_2(k_2) \mid H_0))$ denotes the joint p.d.f. of the sample $d_2(k_2)$, given $\theta_2=\theta$ (respectively given the prior H_0) .

Given $d_2(k_2)$, unconditional on the value of θ_2 , the future observations from E_2 , X_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 $\nu(dx)$)

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

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.

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

$$\hat{\mu}_2(d_2(k_2)) = E_{H(\cdot \mid d_2(k_2))}[\mu(\theta_2)] = E_{f(\cdot \mid d_2(k_2))}[X_2] \ . \tag{2.5}$$

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 $f(x \mid d_2(k_2))$ defined in (2.5) and (2.6) respectively are uniquely determined by $(k_2, \bar{y}_{2,k_2})$, where

$$\bar{y}_{2,k} = \frac{1}{k} \sum_{j=1}^{k} y_{2,j} \tag{2.6}$$

is the sample mean; i.e., the pair $(k, \bar{y}_k)$ is a sufficient statistic for the unknown parameter.

Thus, we can assume that in relations (2.5) , (2.6) and (2.7) $d_2(k_2)$ is simply the two dimensional vector $d_2(k_2) = (k_2, \bar{y}_{2,k_2})$. Note that 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

$$d_2(k_2 \mid d_2(k-1), y_{2,k}) = (k, m(k-1, y, y_{2,k})) \ , \tag{2.7}$$

where

$$m(k,y,x) = \frac{ky+x}{k+1} \ . \tag{2.8}$$

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

$$\pi(t) = \pi(t \mid d_1(k_1(t,\pi)), d_2(k_2(t,\pi))) \tag{2.9}$$

is equal to 1 or 2 , according to whether at stage t π dictates to take a sample from E_1 or E_2 respectively, where

$$k_i(t,\pi) = \sum_{j=0}^{t-1} 1_{\{\pi(j)=i\}} \ . \tag{2.10}$$

The performance of a policy π is measured by

$$S(t,\pi) = \sum_{j=0}^{t-1} Y_{\pi(j),k_{\pi(j)}(j,\pi)} , \tag{2.11}$$

and the expected values

$$E_\theta S(t,\pi) = E[S(t,\pi) \mid \theta_2=\theta] = \mu(\theta_1) E_\theta k_1(t,\pi) + \mu(\theta) E_\theta k_2(t,\pi) , \tag{2.12}$$

$$M(t,H_0,\pi) = E_{H_0}[E_\theta S(t,\pi)] = E_{f(\cdot \mid H_0)}[S(t,\pi)] . \tag{2.13}$$

A policy π^* is optimal for the problem of horizon N and initial prior $H_0(\theta)$ on θ_2 , if and only if

$$M(N,H_0,\pi^*) = \max_\pi M(N,H_0,\pi) , \tag{2.14}$$

where the maximum is taken over all sequential policies defined in (2.11) .

An alternative way to describe the problem is in terms of "costs", i.e. suitably defined regrets rather than "rewards", or outcomes. We introduce a loss function $L(\theta,i)$ which represents the one step loss incurred when the unknown parameter is equal to θ and a sample from experiment E_i is taken.

$$L(\theta,i) = \mu^*(\theta) - E_\theta X_i , \tag{2.15}$$

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

$$R(t,H_0,\pi) = E_{H_0}[\sum_{j=1}^{t} L(\theta,\pi(j))] = t E_{H_0}[\mu^*(\theta)] - M(t,H_0,\pi) . \tag{2.16}$$

Since the quantity $t E_{H_0}[\mu^*(\theta)]$ in (2.19) 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^* :

$$R(N,H_0,\pi^*) = \min_\pi R(N,H_0,\pi) . \tag{2.17}$$

3. OPTIMALITY EQUATIONS – PRELIMINARY RESULTS. The main result of this secton is the derivation of a set of Dynamic Programming (D.P.) equations in a form suitable for the study of the problem. This task is accomplished in two steps.

First we obtain the standard D.P. optimality equations for the optimization problem defined in section 2 in terms of maximization of the expected sum of outcomes, as well as of minimization of the Bayes risk and reduce them into those of a stopping problem. This stopping problem reduction is intuitive, because if an optimal policy ever switches from E_2 to E_1 (and θ_1 is known), it means that at the switching time point $\mu(\theta_1)$ appears to be sufficiently larger than $\mu(\theta_2)$, given the up to t information about θ_2 and the number of the remaining to be taken samples, and also no additional information will be gained about the unknown parameter by sampling from E_1 .

Second, using an appropriate change of measure transformation, we bring the D.P. equations to the desired form, used in the subsequent sections for the proofs of the structural and asymptotic results.

In the sequel it is 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 .

We start by defining two sets of optimization problems.

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 $H(\theta \mid (k,y))$, 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. [6]) we obtain

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

$$\begin{aligned} V(n,k,y) = \max\{ & r(k,y;\alpha=1) + V(n-1,k,y) , \\ & r(k,y;\alpha=2) + E_{f(\cdot \mid (k,y))} V(n-1,k+1,m(k,y,X_2))\} , \end{aligned}$$

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

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

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

$$\begin{aligned}U(n,k,y) = \min\{&c(k,y;\alpha=1)+U(n-1,k,y)\ ,\\ &c(k,y;\alpha=2)+E_{f(\cdot\ |\ (k,y))}U(n-1,k+1,m(k,y,X_2))\}\ ,\end{aligned}$$

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

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

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

$$r(k,y;\alpha=1) = E_{\theta_1}[X_1]\ = \mu(\theta_1)\ , \tag{3.7}$$

$$\begin{aligned}r(k,y;\alpha=2) &= E_{f(\cdot\ |\ (k,y))}X_2\\ &=E_{H(\cdot\ |\ (k,y))}[E_\theta X_2] = \int_\Theta \mu(\theta)\,dH(\theta\ |\ (k,y))\ .\end{aligned} \tag{3.8}$$

$$\begin{aligned}c(k,y;\alpha=1) &= E_{H(\cdot\ |\ (k,y))}[\mu^*(\theta)] - r(k,y;\alpha=1)]\\ &=\int_{\theta\geq\theta_1}(\mu(\theta)-\mu(\theta_1))\,dH(\theta\ |\ (k,y))\ ,\end{aligned} \tag{3.9}$$

$$\begin{aligned}c(k,y;\alpha=2) &= E_{H(\cdot\ |\ (k,y))}\,[\mu^*(\theta)] - r(k,y;\alpha=2)]\\ &=\int_{\theta<\theta_1}(\mu(\theta_1)-\mu(\theta))\,dH(\theta\ |\ (k,y))\ .\end{aligned} \tag{3.10}$$

Moreover, the supremum and infimum in (3.1) and (3.2) are attained by 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 for the remaining trials. The proof is an extension of that given in [3] for the case of binomial populations. It is ommitted here for briefness.

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

$$\begin{aligned}V(n,k,y)=\max\{\,&n\mu(\theta_1)\ ,\\ &r(k,y;\alpha=2)+E_{f(\cdot\ |\ (k,y))}V(n-1,k+1,m(k,y,X_2))\}\ ,\end{aligned} \tag{3.11}$$

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

$$U(n,k,y) = \min\{\, nc(k,y;\alpha=1)\ ,\\ c(k,y;\alpha=2) + E_{f(\cdot\ \mid\ (k,y))} U(n-1,k+1,m(k,y,X_2))\}. \tag{3.12}$$

Definitions 3.1 below allow us to use a change of measure transformation in order to obtain the final form of the optimality equations.

Definitions 3.1. Let

$$l(\theta,\theta_1 \mid y) = ln\frac{f(y\mid\theta)}{f(y\mid\theta_1)} \tag{3.13}$$

$$\Lambda(k,y) = \int_{\Theta} e^{kl(\theta,\theta_1\mid y)}dH_0(\theta) \tag{3.14}$$

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

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

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

Remark 3.1. From (2.2) it is easy to see that

$$l(\theta,\theta_1 \mid y) = d(\theta)\,y-\omega(\theta)\ , \tag{3.18}$$

$$kl(\theta,\theta_1 \mid y)+l(\theta,\theta_1 \mid x) = (k+1)\,l(\theta\,,\theta_1 \mid m(k,y,x)) \tag{3.19}$$

In the next proposition we obtain a set of optimality equations which are equivalent to (3.11), (3.12) . The proof is given in the Appendix together with a necessary auxilliary lemma.

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

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

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

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

where

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

and

$$q(k,y) = \int_{\Theta} \delta(\theta)\, e^{kl(\theta,\theta_1 \mid y)}\, dH_0(\theta) \ . \tag{3.23}$$

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

$$u(n,k,y) = \min\ \{n\,\overline{c}(k,y;\alpha=1)\ ,\ \overline{c}(k,y;\alpha=2) + E_{f(\cdot \mid \theta_1)} u(n-1,k+1,m(k,y,X_2))\}\ , \tag{3.24}$$

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

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

where

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

$$\overline{c}(k,y;\alpha=1) = \int_{\theta \geq \theta_1} \delta(\theta)\, e^{kl(\theta,\theta_1 \mid y)}\, dH_0(\theta)\ , \tag{3.27}$$

$$\overline{c}(k,y;\alpha=2) = -\int_{\theta \leq \theta_1} \delta(\theta)\, e^{kl(\theta,\theta_1 \mid y)}\, dH_0(\theta)\ . \tag{3.28}$$

<u>4. STRUCTURE OF OPTIMAL POLICIES.</u> In this section we prove two theorems that describe the structure of the optimal policy for the finite horizon problem formulated in sections 2 and 3 . Theorem 4.1 describes the optimal policy with respect to stopping and continuation intervals for $y = \frac{1}{k} \sum_{j=1}^{k} y_{2,j}$, while Theorem 4.2 gives an alternative intuitive characterization in terms of inflation factors added to the Bayes estimate of $\mu(\theta_2)$. First we prove the following

<u>Lemma 4.1.</u> The quantity $q(k,y)$, defined in (3.23) , is increasing in y .

<u>Proof.</u> From (3.18), $l(\theta,\theta_1 \mid y)$ is decreasing in y for $\theta < \theta_1$, and increasing in y for $\theta > \theta_1$. Also from (3.16), $\delta(\theta) < 0$ (> 0) for $\theta < \theta_1$ ($\theta > \theta_1$) . Thus $\delta(\theta)e^{kl(\theta,\theta_1 \mid y)}$ is increasing in y for every $\theta \in \Theta$, $\theta \neq \theta_1$, and is equal to zero for $\theta = \theta_1$. The lemma follows from this and the definition of $q(k,y)$.

We can now prove the following main result.

<u>Theorem 4.1.</u> **a)** For each n , k there exists a number $y_n(k)$ with the property

$$\pi^*(n,k,y) = \begin{cases} 1\ , & \text{if } y < y_n(k) \\ 2\ , & \text{if } y \geq y_n(k) \end{cases} \tag*{(4.1)$_n$}$$

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 .

Proof. **a)** Define

$$T(n,k,y) = q(k,x) + \int v(n-1,k+1,m(k,y,x)) f(x \mid \theta_1) \nu(dx) \quad . \tag{4.2}$$

We shall prove simultaneously by induction on n that

i) relation $(4.1)_n$ holds,

ii) $T(n,k,y)$ is increasing in y , for all (n,k) . $(4.3)_n$

For $n=1$, $(4.3)_1$ is immediate from Lemma 4.1 . Let

$$y_1(k) = \inf \{ y : \ T(1,k,y) \ \geq \ 0 \} \ , \tag{4.4}$$

where we define $\inf \emptyset = +\infty$.

For $y < y_1(k)$ $T(1,k,y)$ is negative , while for $y \geq y_1(k)$ it is nonnegative. This completes the proof of $(4.1)_1$. Now suppose that $(4.1)_n$ and $(4.3)_n$ hold. For $n+1$ we have

$$v(n+1,k,y) = \max \{ 0 \ , \ T(n+1,k,y) \} \ . \tag{4.5}$$

From $(4.1)_n$ we get

$$v(n,k+1,m(k,y,x)) = \begin{cases} 0 \ , & \text{if } m(k,y,x) < y_n(k+1) \\ T(n,k+1,m(k,y,x)) \ , & \text{if } m(k,y,x) \geq y_n(k+1) \end{cases} \tag{4.6}$$

But the relation $m(k,y,x) \geq y_n(k+1)$ is from (2.10) equivalent to

$$x \ \geq \ x_n(k,y) \ \equiv \ (k+1)\, y_n(k+1) - ky \ , \tag{4.7}$$

hence

$$T(n+1,k,y) = q(k,y) + \int_{x \geq x_n(k,y)} T(n,k+1,m(k,y,x)) f(x \mid \theta_1) \nu(dx) \tag{4.8}$$

In order to prove $(4.3)_{n+1}$, we have from $(4.3)_n$ that $T(n,k+1,m(k,y,x))$ is increasing in m , while $m(k,y,x)$ is increasing in y , and so $T(n,k+1,m(k,y,x))$ is also increasing in y . Also $T(n,k+1,m(k,y,x)) \geq 0$ for $x \geq x_n(k,y)$.

Furthermore $x_n(k,y)$ is decreasing in y , so when y increases the range of integration also increases. $(4.3)_{n+1}$ follows from the above. Relation $(4.1)_{n+1}$ can be established using $(4.3)_{n+1}$ and defining $y_{n+1}(k)$ in the same way as in (4.4) , and this completes the proof of (a) .

b) The proof is an immediate consequence of part (a) and Proposition 3.2 . If $y \geq y_n(k)$, then $\pi^*(n,k,y)=2$, $\pi^*(n+1,k,y)=2$, thus $y \geq y_{n+1}(k)$. Therefore, $y \geq y_n(k)$ implies $y \geq y_{n+1}(k)$, which is equivalent to

$$[y_n(k),\infty) \subseteq [y_{n+1}(k),\infty) \text{ , or}$$

$$y_{n+1}(k) \leq y_n(k) \text{ .} \tag{4.9}$$

<u>**Remark 4.1.**</u> It is clear that $y_n(k)$ is related to the uncertainty due to the ignorance of parameter θ_2 , and represents in some way the amount of immediate reward which we can afford sacrificing in order to obtain information about θ_2 , which is valuable for future decisions. Therefore, the monotonicity of $y_n(k)$ in n is intuitive, since further sampling from E_2 reduces the uncertainty.

Theorem 4.2 below provides an alternative characterization of the structure of the optimal policy.

<u>**Theorem 4.2.**</u> **a)** For each n , k ,y there exists $\epsilon(n,k,y)$ with the property

$$\pi^*(n,k,y) = \begin{cases} 1 , & \text{if } E_{H(\cdot \mid (k,y))}[\mu(\theta_2)] + \epsilon(n,k,y) < \mu(\theta_1) \\ 2 , & \text{if } E_{H(\cdot \mid (k,y))}[\mu(\theta_2)] + \epsilon(n,k,y) \geq \mu(\theta_1) \end{cases} , \tag{4.10}$$

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

$$\epsilon(n,k,y) = -\frac{q(k,y_n(k))}{\Lambda(k,y)} \text{ .} \tag{4.11}$$

b) Quantity $\epsilon(n,k,y)$ is positive and increasing in n for all k,y.

<u>**Proof .**</u> **a)** By Theorem 4.1.b $y_n(k) \leq y_1(k)$. We also have from Theorem 4.1 and Lemma 4.1 that if $\pi^*(n,k,y) = 1$, then $y < y_n(k)$, and so $q(k,y) < q(k,y_n(k))$. But

$$q(k,y) = (r(k,y;\alpha = 2) - \mu(\theta_1))\,\Lambda(k,y))$$

$$= (E_{H(\cdot \mid (k,y))}[\mu(\theta_2)] - \mu(\theta_1))\,\Lambda(k,y)) \text{ ,} \tag{4.12}$$

where $E_{H(\cdot \mid (k,y))}[\mu(\theta_2)]$ denotes the conditional expectation of the reward from E_2, given the information about the previous outcomes.

We thus have

$$\begin{aligned} y < y_n(k) &\Leftrightarrow \quad q(k,y) < q(k,y_n(k)) \\ &\Leftrightarrow \quad E_{H(\cdot \mid (k,y))}[\mu(\theta_2)] - \frac{q(k,y_n(k))}{\Lambda(k,y)} < \mu(\theta_1) , \end{aligned} \tag{4.13}$$

and this completes the proof of (a).

b) Since $y_n(k) \leq y_1(k)$, it is true that $q(k,y_n(k)) < 0$, thus $\epsilon(n,k,y) > 0$. By (4.12), the dependence of $\epsilon(n,k,y)$ on n is due to $q(k,y_n(k))$. Also by Lemma 4.1 $q(k,y_n(k))$ is increasing in $y_n(k)$. Finally, since $y_n(k)$ is nonincreasing in n, $\epsilon(n,k,y)$ is increasing in n.

Remark 4.1. An interpretation of the quantities $\epsilon(n,k,y)$ is that they represent a positive inflation, that we add to the current estimate of the reward of E_2, $\hat{\mu}_2 = E_{H(\cdot \mid (k,y))}[\mu(\theta_2)]$, in order to take into account the uncertainty associated with it. So the properties of $\epsilon(n,k,y)$ stated in part (b) are intuitively expected.

APPENDIX

Lemma A.1. For every function $g(k,y)$ such that $E_{f(\cdot \mid (k,y))}(\mid g(k,X_2) \mid) < \infty$ we have that

$$E_{f(\cdot \mid (k,y))}[\, g(k+1,m(k,y,X_2))\,] =$$

$$\frac{1}{\Lambda(k,y)} E_{f(\cdot \mid \theta_1)}[\, g(k+1,m(k,y,X_2))\, \Lambda(k+1,m(k,y,X_2))]. \tag{A.14}$$

Proof . From (2.6) (3.14) and (3.19)

$$\begin{aligned} f(x \mid (k,y)) &= \frac{1}{\Lambda(k,y)} \int_{\Theta} f(x \mid \theta_1)\, e^{l(\theta,\theta_1 \mid x)} e^{kl(\theta,\theta_1 \mid y)} dH_o(\theta) \\ &= \frac{\Lambda(k+1,m(k,y,x))}{\Lambda(k,y)} f(x \mid \theta_1) \end{aligned} \tag{A.15}$$

Thus,

$$\begin{aligned} E_{f(\cdot \mid (k,y))}[\, g(k+1,m(k,y,X_2))\,] &= \int g(k+1,m(k,y,x))\, f(x \mid (k,y))\, \nu(dx) \\ &= \int g(k+1,m(k,y,x))\, \frac{\Lambda(k+1,m(k,y,x))}{\Lambda(k,y)} f(x \mid \theta_1)\, \nu(dx) \end{aligned}$$

$$=\frac{1}{\Lambda(k,y)}E_{f(\cdot\ |\ \theta_1)}[\ g(k+1,m(k,y,X_2))\,\Lambda(k+1,m(k,y,X_2))]\ ,$$

and the proof is complete.

Proof of Proposition 3.3. **a)** Define

$$W(n,k,y)\ =\ V(n,k,y)-\ n\mu(\theta_1)\ . \tag{A.16}$$

Subtracting $n\mu(\theta_1)$ from both sides of (3.11) , we obtain

$$W(n,k,y) = \max\{0\ ,\ r(k,y;\alpha=2)-\mu(\theta_1)+\ E_{f(\cdot\ |\ (k,y)}[W(n-1,k+1,m(k,y,X_2))]\} \tag{A.17}$$

Using 3.8, and Definitions 3.1

$$r(k,y;\alpha=2) = \frac{1}{\Lambda(k,y)}\int_{\underline{\theta}}^{\overline{\theta}}\ \mu(\theta)\,e^{kl(\theta,\theta_1\ |\ y)}dH_o(\theta)\quad . \tag{A.18}$$

Hence

$$r(k,y;\alpha=2)-\mu(\theta_1)\ = \frac{1}{\Lambda(k,y)}[\int_{\underline{\theta}}^{\overline{\theta}}\ \mu(\theta)\,e^{kl(\theta,\theta_1\ |\ y)}dH_o(\theta)\ -\ \mu(\theta_1)\,\Lambda(k,y)]=\frac{q(k,y)}{\Lambda(k,y)}\ . \tag{A.19}$$

From Lemma A.1

$$E_{f(\cdot\ |\ (k,y)}[W(n-1,k+1,m(k,y,X_2))]=$$

$$\frac{1}{\Lambda(k,y)}E_{f(\cdot\ |\ \theta_1)}[\ W(n-1,k+1,m(k,y,X_2))\ \Lambda(k+1,m(k,y,X_2))]\ , \tag{A.20}$$

Thus, (3.11) is equivalent to

$$W(n,k,y)=$$
$$\max\{0,\ \frac{q(k,y)}{\Lambda(k,y)}+\frac{1}{\Lambda(k,y)}E_{f(\cdot\ |\ \theta_1)}[\ W(n-1,k+1,m(k,y,X_2))\,\Lambda(k+1,m(k,y,X_2))]\}\ . \tag{A.21}$$

To complete the proof we only need to multiply both sides of (A.8) by $\Lambda(k,y)$ (where $\Lambda(k,y)\ >\ 0)$.

The proof of (b) is similar, with the only difference being that we need not perform the first subtractions.

REFERENCES

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

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

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

4. Burnetas, A. N. and Katehakis, M. N. "On Sequencing Two Types of Tasks on a Single Processor under Incomplete Information", *Prob. Eng. Info. Sci.,* 7, 1993, 85-119.

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

6. Dynkin, E.B. and Yushkevich, A.A. *"Controlled Markov Processes"*, Springer-Verlag, New York, 1979.

7. Gittins, J.C. and Jones, D.M. "A Dynamic Allocation Index for the Sequential Design of Experiments", in *"Progress in Statistics"* (J. Gani ed.), North Holland, Amsterdam, 1974, pp. 241-266.

8. Gittins, J.C. "Bandit Processes and Dynamic Allocation Indices", *J. Roy. Statist. Soc. Ser. B,* 41, 1979, 148-164.

9. Gittins, J. C. *"Multi-armed bandit allocation indices"*, J. Wiley, New York, 1989.

10. Katehakis, M. N. and Derman, C. "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, 1986, 29–39.

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

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

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

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