COMPUTING OPTIMAL SEQUENTIAL ALLOCATION RULES IN CLINICAL TRIALS*

Michael N. Katehakis

State University of New York at Stony Brook

and

Cyrus Derman

Columbia University

The problem of assigning one of several treatments in clinical trials is formulated as a discounted bandit problem that was studied by Gittins and Jones. The problem involves comparison of certain state dependent indices. A recent characterization of the index is used to calculate more efficiently the values of these indices.

1. Introduction.

We consider the well known problem of optimal allocation of treatments in clinical trials. A simple version of the problem is as follows. There are several possible treatments for a given disease. When a particular treatment n is used it is either effective with unknown probability θ_n or not effective with probability $1 - \theta_n$. The problem is to find a sequential sampling procedure which maximizes a measure of the expected total number of treatment successes. When the planning horizon is infinite, prior distributions are assigned to the unknown parameters, and one takes the expected total discounted number of successes as the relevant measure of performance of a sequential sampling procedure, the problem can be put into the form of a discounted version of the

* Work supported by USAF Contract AFOSR 840136, NSF Grants DMS-84-05413, ECS-85-07671 and ONR Contract N00014-84-K-0244.

AMS 1980 subject classifications: Primary 62L15; Secondary 60J10.

Key words and phrases: dynamic programming, multi armed bandit problem, sequential allocations, successive approximations.

bandit problem treated successfully by Gittins and Jones (1974), see also Whittle (1980), (1982, p.210). The original formulation of the multiarmed bandit problem and the sequential clinical trials problem is due to Robbins (1952). Gittins and Jones showed that there is an index associated with each state of each bandit such that an optimal procedure always uses the bandit with the largest current index value. Recently, Katehakis and Veinott (1985) have obtained a new characterization of the index which allows the index to be more easily calculated. The purpose of this paper is to illustrate the calculation of the index in the context of the clinical trials problem using this new characterization.

2. Computing dynamic allocation indices.

Suppose N treatments are available for treating patients with a certain disease. Let $Y_n(k) = 1$ $(Y_n(k) = 0)$ denote the outcome that the n-th treatment has been successful (unsuccessful) the k-th time it is used. At times t=1,2,..., based on past observations, one has to decide which treatment to allocate to the next patient. At the start of the experiment we assume that θ_n is a random variable with beta prior density with parameter vector (a_n, b_n); i.e., θ_n has the prior density

$$(1) \quad g_n(\theta) = \Gamma(a_n + b_n)\{\Gamma(a_n)\Gamma(b_n)\}^{-1}\theta^{a_n-1}(1 - \theta)^{b_n-1}, \text{ for every } \theta \in [0,1],$$

where in (1) a_n, b_n are strictly positive constants. Furthermore, we assume that $\theta_1,\ldots,\theta_n$ are independent. If after k trials using treatment n we let $x_n(k) = (\mathbf{s}_n(k), \mathbf{f}_n(k))$, where $\mathbf{s}_n(k)$ $(\mathbf{f}_n(k))$ denotes the number of successes (the number of failures) then, the posterior density of θ_n given $x_n(k)$ is also beta with parameter vector $(a_n + \mathbf{s}_n(k), b_n + \mathbf{f}_n(k))$. Thus, the information obtained during the first k trials from treatment n is summarized by $x_n(k)$. Furthermore,$\{x_n(k), k \geqslant 1\}$ is a Markov chain on
$\mathbf{S} = \{(s,f), s,f = 0,1,2,\ldots\}$ with transition probabilities given by

(2) $$P(x_n(k+1) = (s+1,f) \mid x_n(k) = (s,f))$$

$$= 1 - P(x_n(k+1) = (s,f+1) \mid x_n(k) = (s,f))$$

$$= P(Y_n(k+1) = 1 \mid x_n(k) = (s,f)) = \frac{a_n + s}{a_n + b_n + s + f} .$$

The problem is to determine a policy π which maximizes the expected discounted number of successes; i.e., to maximize $w(\pi,\alpha)$

(3) $$w(\pi,\alpha) = \int \ldots \int E(\Sigma_{t=1}^{\infty} \alpha^{t-1} Y_{\pi(t)})\; g_1(d\theta_1) \ldots g_N(d\theta_N),$$

where $Y_{\pi(t)}$ is $Y_n(k)$ if at time t treatment $\pi(t) = n$ is used for the k-th time and $\alpha \in (0,1)$ is a discount factor. An interpretation of the discount factor α is that $1-\alpha$ is the probability that at any given time the entire experiment will be terminated. Stated otherwise, there are N Markov chains; the problem is to sequentially activate one of them, leaving the others inactive, in order to maximize the expected total discounted reward. In this case the expected reward at any time is the expected posterior probability of success associated with the state of the activated Markov chain; i.e., if the n-th chain is activated for the k-th time when $x_n(k) = (s,f)$, then the corresponding reward is

(4) $$r_n(s,f) = E(Y_n(k+1) \mid x_n(k) = (s,f)) = \frac{a_n + s}{a_n + b_n + s + f} .$$

Within the context of this formulation, Gittins and Jones (1974) showed that this problem can be reduced to N one dimensional problems. Each of the latter problems involves a single Markov chain and its solution is the calculation of a dynamic allocation index $m_n(s,f)$ associated with the current state (s,f) of the Markov chain. Then, at each point of time an optimal policy for the original problem is such that it activates the chain with the largest current index value. Based on an earlier characterization of $(1-\alpha)^{-1} m_n(s,f)$, Gittins and Jones (1979) used an algorithm for computing optimal policies. Recently, Katehakis and Veinott (1985) have obtained a different characterization of the index. This characterization casts the calculation of the index into the form of a familiar replacement problem, e.g., see Derman (1970, p.121). Namely, if C is the class of policies R for controlling $\{x_n(k),\ k \geq 1\}$ by either allowing it to continue or to instantaneously restart it at its initial state $x_n(1) = (s,f)$, then

$$(5) \qquad m_n(s,f) = \sup_R \{E_R(\Sigma_{k=1}^{\infty} r_n(x_n(k)) \mid x_n(1) = (s,f))\}.$$

We next show that (5) can be used to evaluate $m_n(s,f)$ with sufficient accuracy. In the sequel we will be concerned with a single treatment; for notational simplicity we will drop the subscript n. Since computing m(s,f) is essentially the same as computing m(0,0) - it only involves changing the prior vector from (a,b) to (a + s, b + f) - it suffices, without loss of generality, to discuss only the computation of m(0,0). It is well known that solving (5) for the fixed initial state (0,0) involves solving the dynamic programming equations

$$(6) \quad V(s,f) = \max \{\frac{a}{a+b} + \alpha[\frac{a}{a+b} V(1,0) + \frac{b}{a+b} V(0,1)],$$

$$\frac{a+s}{a+b+s+f} + \alpha[\frac{a+s}{a+b+s+f} V(s+1,f) + \frac{b+f}{a+b+s+f} V(s,f+1)]\},$$

for every $(s,f) \in S$.

The fact that equation (6) is for computing m(0,0) is reflected in the appearance of the terms V(1,0) and V(0,1) in the right side of it. Given the solution $\{V(s,f)$, for every $(s,f) \in S\}$ of (6) then m(0,0) = V(0,0).

Equation (6) is of the form $V(s,f) = T_{sf}V$ or equivalently

$$(7) \qquad V = TV,$$

where in (7) V is the vector of values $\{V(s,f)\}$ and T is a contraction operator on a complete metric space. Thus, it has a unique bounded solution.

In computing the solution of (7) we consider the finite subset $S_L = \{(s,f) \in S : s + f \leqslant L\}$ and the two systems of equations

$$u_L(s,f) = T_{sf}u_L, \qquad \text{if } s + f < L, \tag{8a}$$

$$u_L(s,f) = \frac{a + s}{a + b + s + f}\,\frac{1}{1 - \alpha}, \qquad \text{if } s + f = L \tag{8b}$$

$$U_L(s,f) = T_{sf}U_L, \qquad \text{if } s + f < L, \tag{9a}$$

$$U_L(s,f) = \frac{1}{1 - \alpha}, \qquad \text{if } s + f = L. \tag{9b}$$

We will use the following more compact notation for (8) and (9)

$$u_L = T_1u_L, \tag{8c}$$

$$U_L = T_2U_L. \tag{9c}$$

The transformations T, T_1, T_2 are monotone contractions (see Bertsekas (1976)), thus, successive approximations will converge to their unique fixed points for any initial points $V^{(0)}$, $u_L^{(0)}$, $U_L^{(0)}$. That is,

$$\lim_{n\to\infty} V^{(n)} = \lim_{n\to\infty} TV^{(n-1)} = V, \tag{10}$$

$$\lim_{n\to\infty} u_L^{(n)} = \lim_{n\to\infty} T_1u_L^{(n-1)} = u_L, \tag{11}$$

$$\lim_{n\to\infty} U_L^{(n)} = \lim_{n\to\infty} T_2U_L^{(n-1)} = U_L. \tag{12}$$

Moreover, if the points $V^{(0)}$, $u_L^{(0)}$, $U_L^{(0)}$ are chosen propitiously, the convergence in (10), is from below or above as desired and from below (above) in

(11) ((12)).

An algorithm to compute V(0,0) based on (10) involves an infinite number of variables; however, Propositions 1 and 2, below, allow us to use (11) and (12) which involve only a finite number of variables. We first state

PROPOSITION 1. For equations (7), (8) and (9) we have

$$(13) \quad \frac{a + s}{a + b + s + f}(1 - \alpha)^{-1} < V(s,f) < (1 - \alpha)^{-1} \quad \text{for all } (s,f) \in S,$$

and

$$(14) \quad u_L(s,f) < V(s,f) < U_L(s,f), \text{ for all } (s,f) \text{ such that } s + f < L.$$

The proof of Proposition 1 is easy and its details will be omitted. Indeed, the first inequality in (13) follows from the fact that the left hand side is the expected discounted reward achieved by the suboptimal policy that never restarts the process in state (0,0) when the initial state is state (s,f); the second inequality in (13) follows from the fact that the left hand side is the expected discounted reward attained when all one period rewards are replaced by 1 which is an upper bound for them. Inequalities (14) then, follow from (13), equations (10), (11) and (12) and the monotonicity of transformations T, T_1, T_2.

PROPOSITION 2. For any $\varepsilon > 0$ there exist an $L_0 = L(\varepsilon)$ such that

$$(15) \quad U_L(0,0) - u_L(0,0) < \varepsilon, \text{ for all } L > L_0.$$

Proof. Because of (14) it suffices to show that for any positive constants ε_1 and ε_2 there exist $L_1 = L(\varepsilon_1)$ and $L_2 = L(\varepsilon_2)$ such that

$$(16) \quad U_L(0,0) - V(0,0) < \varepsilon_1, \text{ for all } L > L_1,$$

and

$$V(0,0) - u_L(0,0) < \varepsilon_2, \quad \text{for all } L > L_2. \tag{17}$$

We only prove (16) since the proof of (17) is analogous. If we take $U_L^{(0)} = V^{(0)} = (1-\alpha)^{-1}$ in (10) and (12) then, for any L and all $n < L$ we obtain that

$$U_L^{(n)}(0,0) = V^{(n)}(0,0), \tag{18}$$

and the convergence in (10), (12) is from above; thus, using (10) and the fact that $V(s,f) > 0$ we have

$$\begin{aligned} V^{(n)}(0,0) - V(0,0) &< \alpha \sup_{(s,f)} \{V^{(n-1)}(s,f) - V(s,f)\} < \ldots \\ &< \alpha^n \sup_{(s,f)} \{V^{(0)}(s,f) - V(s,f)\} < \alpha^n(1-\alpha)^{-1}. \end{aligned} \tag{19}$$

It follows from (18), (19) that for any L and for all $n < L$

$$U_L^{(n)}(0,0) - V(0,0) < \alpha^n(1-\alpha)^{-1}. \tag{20}$$

Similar arguments using (12) imply that for all $n > 1$

$$U_L^{(n)}(0,0) - U_L(0,0) < \alpha^n(1-\alpha)^{-1}. \tag{21}$$

Thus, using (20) and (21) it is now easy to complete the proof of (16).

REMARK. It was assumed that each clinical trial resulted either in a success or in a failure. The methodology described here extends straightforwardly to the case where the outcome of a trial can be classified into c, $c > 2$, classifications. Then the parameter θ_n, is a vector $(\theta_n^1,\ldots,\theta_n^c)$ where θ_n^i is the probability of the trial resulting in the i-th classification. The beta prior is replaced by a Dirichlet prior and the state space becomes $S = \{(s_1,\dots,s_c),\ s_i = 0,1,\dots\}$, where s_i denotes the number of trials resulting in classification i ($1 \leqslant i \leqslant c$). The reward is a given function of the classification; see, also, Glazebrook (1978).

3. Computations.

For a given (a,b) in order to compute m(0,0) = V(0,0) we use transformations T_1 and T_2 starting from

$$u_L^{(0)}(s,f) = \frac{a+s}{a+b+s+f}\,\frac{1}{1-\alpha}, \quad \text{and} \quad U_L^{(0)}(s,f) = \frac{1}{1-\alpha}.$$

We choose L sufficiently large according to Proposition 1 and iterate until the difference: $U_L^{(n)}(0,0) - u_L^{(n)}(0,0)$ is less than 10^{-4}. We, then, take as our approximation to V(0,0) the midpoint of the final interval.

Since there is always an error in computing the indices, the possibility of not using an optimal policy always exists. In our context, here, this can be overcome by doing enough computations to guarantee that in computing the indices the bounding intervals do not overlap. However in general, Katehakis and Veinott (1985) have shown that if the computed indices are close to the exact indices then the expected discounted return of the policy based on the computed indices will be close to the optimal expected discounted return.

In the following tables the results of some calculations are tabulated. There is a separate table for each value of $\alpha = .5, .75, .9$. An entry in cell (a+s, b+f) is the index for a treatment having prior (a,b) and in state (s,f).

Note that the numbers in Table 2 (for a+s, b+f = 1,2,...,5) are consistent with those published by Gittins and Jones (1979).

Table 1 (α = .5)

b+f a+s	1	2	3	4	5	10	20	30	40	50	100
1	1.118	.751	.560	.444	.367	.194	.099	.066	.049	.039	.019
2	1.411	1.071	.859	.715	.611	.351	.188	.128	.097	.078	.039
3	1.554	1.257	1.051	.902	.789	.482	.269	.186	.142	.115	.058
4	1.639	1.379	1.187	1.040	.925	.592	.342	.240	.185	.150	.077
5	1.697	1.466	1.288	1.147	1.032	.688	.410	.291	.266	.184	.096
10	1.829	1.683	1.558	1.449	1.354	1.017	.677	.507	.405	.337	.183
20	1.908	1.824	1.747	1.675	1.609	1.344	1.008	.807	.672	.575	.335
30	1.937	1.878	1.822	1.769	1.720	1.507	1.207	1.005	.862	.754	.463
40	1.952	1.906	1.863	1.821	1.781	1.605	1.338	1.148	1.004	.892	.573
50	1.961	1.924	1.888	1.854	1.820	1.670	1.433	1.254	1.115	1.003	.668
100	1.980	1.961	1.942	1.923	1.905	1.819	1.668	1.540	1.430	1.335	1.001

Table 2 (α = .75)

b+f a+s	1	2	3	4	5	10	20	30	40	50	100
1	2.484	1.702	1.272	1.007	.829	.428	.212	.139	.103	.082	.040
2	2.986	2.303	1.856	1.548	1.322	.754	.397	.267	.201	.161	.080
3	3.224	2.642	2.221	1.909	1.672	1.018	.563	.386	.293	.236	.119
4	3.367	2.863	2.476	2.174	1.935	1.240	.712	.497	.381	.308	.157
5	3.463	3.019	2.663	2.378	2.143	1.429	.848	.600	.463	.377	.194
10	3.689	3.410	3.164	2.948	2.758	2.076	1.383	1.034	.824	.685	.370
20	3.827	3.666	3.516	3.375	3.245	2.715	2.039	1.631	1.358	1.163	.676
30	3.880	3.766	3.657	3.554	3.456	3.033	2.431	2.026	1.737	1.519	.933
40	3.908	3.819	3.734	3.652	3.574	3.224	2.691	2.308	2.020	1.795	1.153
50	3.925	3.853	3.783	3.715	3.649	3.351	2.877	2.519	2.240	2.016	1.343
100	3.961	3.923	3.886	3.849	3.813	3.643	3.342	3.087	2.867	2.676	2.008

Table 3 (α = .9)

b+f a+s	1	2	3	4	5	10	20	30	40	50	100
1	7.028	5.001	3.796	3.021	2.488	1.268	.607	.390	.285	.224	.107
2	8.000	6.346	5.163	4.341	3.720	2.116	1.097	.729	.543	.431	.210
3	8.542	7.071	6.010	5.184	4.561	2.783	1.523	1.037	.782	.626	.310
4	8.722	7.539	6.578	5.809	5.179	3.332	1.903	1.319	1.005	.810	.408
5	8.905	7.869	6.996	6.276	5.676	3.799	2.247	1.582	1.217	.987	.503
10	9.342	8.694	8.103	7.572	7.100	5.372	3.580	2.672	2.127	1.765	.948
20	9.620	9.243	8.883	8.542	8.223	6.904	5.196	4.158	3.461	2.962	1.716
30	9.729	9.461	9.201	8.950	8.710	7.664	6.156	5.133	4.401	3.849	2.360
40	9.789	9.580	9.375	9.177	8.984	8.120	6.791	5.829	5.101	4.535	2.910
50	9.826	9.655	9.486	9.322	9.161	8.426	7.245	6.348	5.646	5.081	3.385
100	9.907	9.816	9.726	9.637	9.549	9.128	8.381	7.744	7.195	6.718	5.041

REFERENCES

Bertsekas, D.P. (1976). Dynamic Programming and Stochastic Control. Academic Press, New York.

Derman, C. (1970). Finite State Markovian Decision Processes. Academic Press, New York.

Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi and I. Vince (eds.), Progress in Statistics, North Holland, 241-266.

Gittins, J.C. and Jones, D.M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. Biometrika **66** 561-565.

Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. J. Roy. Statist. Soc. Ser. B **41** 148-164.

Glazebrook, K.D. (1978). On the optimal allocation of two or more treatments in a controlled clinical trial. Biometrika **65** 335-340.

Katehakis, M.N. and Veinott, Jr., A.F. (1985). The multi-armed bandit problem: decomposition and computation. Department of Oper. Res., Stanford Univ., Technical Report, 14 pp.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Monthly* **58** 527-586.

Whittle, P. (1980). Multi-armed bandits and the Gittins Index. *J. Roy. Statist. Soc. Ser. B* **42** 143-149.

Whittle, P. (1982). *Optimization over Time*, Vol. 1, John Wiley, New York.