Gittins Procedures for Bandits with Delayed Responses

by

Stephen G. Eick*
University of Minnesota

Technical Report No. 458
December 1985

---

* Partially supported by the University of Minnesota Statistics Alumni Fellowship.

Summary

The existence of dynamic allocation indices is shown for multi-armed delayed response bandits in specified states with geometric discounting. When the information banks for all arms are zero, the arm indicated by the dynamic allocation procedure or Gittins procedure is optimal. A method of calculating dynamic allocation indices is presented and applied to the beta distribution. Optimal strategies for the two-armed delayed response bandit and are completely described when the discount factor $< 1/2$.

1. Introduction and Summary

1.1. The Multi-armed Bandit with Delayed Responses

Imagine a clinical trial in which patients arrive sequentially at times $0,1,2,\cdots$ for treatment of a particular fatal disease. There are k irreversible treatments of unknown efficacy available. The objective is to maximize the expected discounted total patient lifetime. Treatment assignment is sequential and can depend on previous assignments and the censored lifetimes of any surviving patients. When the current patient is to be treated, the treatments used previously are known; it is also known which of the previous patients have died and how long they survived after treatment, and which have survived to the current time. This problem is an example of a multi-armed bandit problem with delayed responses.

Bandit problems have been studied extensively in the statistical literature. Examples include Bradt, Johnson, and Karlin (1956), Bellman (1956), Feldman

(1962), Rodman (1978), Gittins (1979), Whittle (1980), Bather (1981), and Berry and Fristedt (1985). These are all unrealistic when applied to clinical trials because they assume that the results from all previous patients treated are known when the current patient is treated. In a clinical trial it is infeasible to wait for all previous patients to respond before treating the current patient. According to Armitage (1985) and Simon (1977) the response delay in the classical approach is one of several reasons why sequential methods are not widely used.

The two-armed bandit with delayed responses is introduced by Eick (1985a). In the current paper I assume there are k independent stochastic processes $\{X_{ij}|j = 1,2,\cdots\}$, $i = 1,\cdots, k$, representing the k treatments or arms. Conditional on an unknown parameter θ_i, $X_{ij}|\theta_i$, $j = 1,2,\cdots$, are iid geometric with probability mass function $(1-\theta_i)\theta_i^t$, $t = 0,1,2,\cdots$. This assumption is consistent with the patients on arm i being exchangeable and each having constant probability θ_i of surviving to the next time period. This is a discrete-time version of treatments with exponential lifetimes.

I take a Bayesian approach and assume the parameters $\theta_1,\cdots, \theta_k$ are themselves random with priors $\mu_1,\cdots, \mu_k$. I assume $\theta_1,\cdots, \theta_k$ are independent and restrict consideration to priors μ which have a finite life expectancy: $E[X|\mu] = E[\theta/(1-\theta)|\mu] < \infty$. The sufficient statistics for θ_i are S_i, the number of arm i patient time period successes, and F_i, the number of arm i patient failures. I denote the distribution of θ_i conditioned by $S_i = s_i$ and $F_i = f_i$ by $(s_i,f_i)\mu_i$. At time 0, $s_i = f_i = 0$ and $(0,0)\mu_i = \mu_i$.

The bandit state summarizes all relevant information about the allocation

process. The state consists of the tuple $((s_1,f_1)\mu,p_1;\cdots;(s_k,f_k)\mu_k,P_k;\alpha)$, where $((s_i,f_i)\mu_i,p_i)$ is the state of process i, $(s_i,f_i)\mu_i$ is the current distribution of θ_i, and p_i is the number of patients who have been treated with arm i and are currently alive. These patients' lifetimes are censored when the current patient is treated. They form an "information bank": information accrues as they respond in the next time period, either positively or negatively.

The geometric lifetime assumption implies that the arm i bank size is random with a conditional (given θ_i) binomial distribution. Suppose p_i patients are currently in the information bank of arm i. The number of patients in the bank at the next time period, $(P_i|p_i,\theta)$, is either $\text{bin}(p_i+1,\theta)$ or $\text{bin}(p_i,\theta)$ depending on whether or not the current patient receives treatment i. There is a simple relationship between the sufficient statistics and the information bank; S_i is the sum of the bank size over all previous time periods including the current period and F_i is the total number of patients treated with i minus the current bank size. I denote the bandit in state $(\mu_1,p_1;\cdots;\mu_\kappa,p_\kappa;\alpha)$ by the $(\mu_1,p_1;\cdots;\mu_\kappa,p_\kappa;\alpha)$-bandit.

It is natural to assume that $p_i = 0$, $i = 1,\cdots,k$. But since the possibility of positive p_i will have to be reckoned with in future stages, nothing is gained by this assumption. So I allow $p_i \geq 0$. Then any previously treated patients who are still surviving can be included in the information bank. With this convention the treatment assignment at each stage in the trial can be viewed as the initial treatment for the bandit presenting itself at that time. I will characterize the optimal treatment assignment scheme by specifying the initial

treatment for an arbitrary bandit.

In this paper I consider a clinical trial in which n patients are treated where n is random and has a geometric distribution. A new treatment which is clearly better than any currently being considered may be discovered at any time. I assume that the probability of discovery in each time period is constant and equals $1 - \alpha$. When α is near 0 this discovery is regarded as imminent and when α is near 1 discovery in the near future is unlikely. The probability that the trial consists of exactly n patients is proportional to α^{n-1}. The appropriate discount sequence to model this trial is the geometric with factor α: $(1,\alpha,\alpha^2,\cdots)$.

At time 1 the discount sequence for the bandit presenting itself is $(\alpha,\alpha^2,\cdots)$. This latter discount sequence is a multiple of the original; so the allocation process is unchanged.

A strategy or rule τ for the $(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha)$-bandit is a function defined on the states which indicates the arm to use at each stage in the trial. The worth of τ is the expected discounted total patient lifetime when τ is followed:

$$W(\tau) = E_\tau[\sum_1^\infty \alpha^{j-1} Z_j], \qquad (1.1)$$

where Z_j is the lifetime of the patient treated at stage j (i.e. time j-1) when following τ. The objective is the find a strategy that maximizes (1.1).

The value of the $(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha)$-bandit is the supremum over all strategies of their worths:

$$V = V(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha) = \sup_{\tau} W(\tau). \tag{1.2}$$

A strategy is optimal if it achieves the supremum in (1.2). An arm is optimal if it is the first selection of an optimal strategy. The value of selecting arm i initially and then continuing optimally depending on the result is

$$V^{(i)} = \sup\{W(\tau) \mid \tau \text{ indicates arm i initially}\}.$$

The value V satisfies a dynamic programming equation:

$$V(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha) = \bigvee_{i=1}^{k} V^{(i)}(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha),$$

where

$$V^{(i)} = E[X_{i1}|\mu_i] + \alpha E[V((S_1,F_1)\mu_1,P_1;\cdots;(S_k,F_k)\mu_k,P_k;\alpha)]. \tag{1.3}$$

The expectation in the second term on the right-hand side of (1.3) is over the distribution of states at time 1. Arm i is optimal if and only if $V^{(i)} = V$.

1.2. Summary of Results

The results in this paper describe the nature of the optimal strategies. In Section 2.1, I consider the two-armed bandit which is a special case of the k-armed bandit discussed above. When k = 2, the optimal arm is determined by the sign of $\Delta = V^{(1)} - V^{(2)}$. Theorem 2.1 says that $\Delta((s_1,f_1)\mu_1,p_1;(s_2,f_2)\mu_1,p_2;\alpha)$ is monotone in s_1, f_1, s_2, f_2 for $\alpha < 1/2$. This leads to an explicit characterization of the optimal strategies in terms of a manifold in

$(s_1,f_1;s_2,f_2)$-space. In Section 2.2, I specialize further and assume that $\mu_2 = \delta_\lambda$ is known. When either (i) $\alpha < 1/2$ or (ii) $p_1 = 0$, I show in Theorems 2.3 and 2.5 that the arm indicated by the dynamic allocation procedure or Gittins procedure is optimal. There exists a value Λ such that arm 1 is optimal if and only if $\Lambda \geq \kappa$, where $\kappa = \lambda/(1-\lambda)$.

Theorem 3.1 says that for the $(\mu_1,p_1;\cdots;\mu_k,p_k;\alpha)$-bandit, the arm indicated by the Gittins procedure is optimal when $p_1 = \cdots = p_k = 0$. Chapter 4 presents a class of strategies whose worths closely approximate $V^{(i)}$ but are easy to calculate. This leads to a computational method for evaluating dynamic allocation indices which I apply to the beta family.

2. The Two-armed Bandit

Assume $k = 2$ throughout Chapter 2. Define

$$\Delta(\mu_1,p_1;\mu_2,p_2;\alpha) = V^{(1)}(\mu_1,p_1;\mu_2,p_2;\alpha) - V^{(2)}(\mu_1,p_1;\mu_2,p_2;\alpha).$$

Arm 1 is optimal if and only if $\Delta \geq 0$, is nonnegative and arm 2 is optimal if and only if $\Delta \leq 0$.

2.1. The Δ-function for $\alpha < 1/2$

Consider either arm and let X, θ, s, f, and μ without subscripts stand for the corresponding subscripted quantities. For integral nonnegative s and f the conditional distribution of $(\theta|S=s,F=f)$ is defined by

$$d(s,f)\mu = \theta^s(1-\theta)^f d\mu/b(s,f),$$

where

$$b(s,f) = \int_0^1 \theta^s (1-\theta)^f d\mu. \tag{2.1}$$

The mean lifetime is then

$$E[X|(s,f)\mu)] = E\left[\frac{\theta}{1-\theta}\middle|(s,f)\mu\right] = \frac{b(s+1,f-1)}{b(s,f)}. \tag{2.2}$$

In this section s and f are not restricted to the nonnegative integers but are allowed to be arbitrary provided that (2.2) is finite. This defines a family of distributions $(s,f)\mu$ which generalizes the beta distributions. For the beta family, $d(s,f)\mu \propto \theta^{s-1}(1-\theta)^{f-1}d\theta$ and $b(s,f) = be(s,f)$. In this case $E[X|(s,f)\mu] < \infty$ if and only if $s > 0$ and $f > 1$.

It is convenient to think of (s,f) as the prior successes and failures; $(s,f)\mu$ would be the conditional distribution of θ if s successes and f failures were observed when $\theta \sim \mu$.

The next theorem says that when $\alpha < 1/2$, $\Delta((s_1,f_1)\mu_1,p_1;\mu_2,p_2;\alpha)$ is increasing in s_1 and decreasing in f_1 when μ_1 is supported by more than one point. I conjecture that a similar monotonicity result holds for all $0 \leq \alpha < 1$.

Theorem 2.1. Assume $\alpha < 1/2$, μ_1 is not a one-point distribution, and $\mu(\{0,1\}) = 0$. Then for all μ_1,p_1,μ_2,p_2, and α, $\Delta((s_1,f_1)\mu_1,p_1;\mu_2,p_2;\alpha)$ is increasing in s_1 and decreasing in f_1. Furthermore

$$\Delta((s_1,f_1)\mu_1,p_1+1;\mu_2,p_2;\alpha) > \Delta((s_1,f_1+1)\mu_1,p_1;\mu_2,p_2;\alpha). \qquad (2.2)$$

Remark. Because of the symmetry between arm 1 and arm 2, for $\alpha < 1/2$ $\Delta(\mu_1,p_1;(s_2,f_2)\mu_2,p_2;\alpha)$ is increasing in f_2 and decreasing in s_2 and

$$\Delta(\mu_1,p_1;(s_2,f_2+1)\mu_2,p_2;\alpha) > (\mu_1,p_1;(s_2,f_2)\mu_2,p_2+1;\alpha) \qquad (2.3)$$

when μ_2 is not a one-point distribution and $\mu_2(\{0,1\}) = 0$.

Proof of Theorem 2.1. The proof of this result is a direct extension of Theorem 3.1 in Eick (1985b). □

Theorem 2.1 provides an explicit characterization of optimal strategies for the $((s_1,f_1)\mu_1,p_1;(s_2,f_2)\mu_2,p_2;\alpha)$-bandit. For fixed p_1 and p_2 there is a manifold in $(s_1,f_1;s_2,f_2)$-space where $\Delta((s_1,f_1)\mu_1,p_1;(s_2,f_2)\mu_2,p_2;\alpha) = 0$. On one side of the manifold arm 1 is optimal, on the other arm 2 is optimal, and both arms are optimal on the manifold. When arm 1 is optimal and s_1 is increased or f_1 decreased, arm 1 remains optimal; similarly for arm 2. Equations (2.2) and (2.3) relate the manifolds on which Δ vanishes as p_1 and p_2 vary.

In the next section I prove the existence of dynamic allocation indices for particular states in the delayed response bandit state space.

2.2. Dynamic Allocation

Continue to assume $k = 2$, and now assume that $\mu_2 = \delta_\lambda$, so θ_2 is known to equal λ. The state simplifies to $(\mu_1,p_1;\kappa;\alpha)$ where $\kappa = E[X_{2j}|\mu_2] = \lambda/(1-\lambda)$. This is because successes and failures on arm 2 cannot change μ_2. Throughout

this section I suppress the subscript 1, writing $\mu = \mu_1$, $p = p_1$, and $X_j = X_{1j}$.

The function $\Delta(\mu,p;\kappa;\alpha)$ is continuous in κ (see Theorem 4.2, Eick, 1985a). When $\kappa \to 0$, $\Delta \to E[X_j|\mu]/(1-\alpha)$ and when $\kappa \to \infty$, $\Delta \to -\infty$. Thus, the equation $\Delta(\mu,p;\kappa;\alpha) = 0$ has at least one nonzero solution in κ. When it has a unique solution Λ, I define Λ as the dynamic allocation index or DAI (Gittins and Jones, 1974). My notation follows Berry and Fristedt (1985).

Definition 2.2. Suppose for fixed μ and p the equation $\Delta(\mu,p;\kappa;\alpha) = 0$ has a single solution κ^* in κ. Then the dynamic allocation index for arm 1 is $\Lambda(\mu,p;\alpha) = \kappa^*$.

The DAI is the value of κ for which both arm 1 and arm 2 are optimal. For $\kappa \leq \Lambda$ arm 1 is optimal and for $\kappa \geq \Lambda$ arm 2 is optimal. The next theorem says that the DAI exists when $\alpha < 1/2$.

Theorem 2.3. The DAI exists for all μ and p when $\alpha < 1/2$.

Proof. For $\alpha < 1/2$ the geometric discount sequence satisfies the regularity conditions of Theorem 3.1 of Eick (1985b). Then $\Delta(\mu,p;\kappa;\alpha)$ is decreasing in κ and therefore the equation $\Delta(\mu,p;\kappa;\alpha) = 0$ has a unique solution in κ. □

I now show that DAI's exist for all $\alpha < 1$ when $p = 0$. This result depends on the following lemma which extends Theorem 2.1. The lemma says that when $p = 0$, $\Delta((s,f)\mu,0;\kappa;\alpha)$ is increasing in s and decreasing in f and κ.

Lemma 2.4. Suppose μ is supported by more than one point, $\mu(\{0,1\}) = 0$, and $p = 0$. Then for all α, $\Delta((s,f)\mu,0;\kappa;\alpha)$ is increasing in s and decreasing in f

and κ.

Proof. If arm 2 is optimal initially then an optimal strategy indicates arm 2 at all stages. This is so since the state at time 1 after treating the first patient with arm 2 is the same as the original state. So $V^{(2)} = \kappa/(1-\alpha)$. If arm 1 is optimal initially then arm 1 is also optimal at time 1 if the first patient received arm 2: $V^{(2)} = \kappa + \alpha V^{(1)}$. Combining both cases,

$$\Delta((s,f)\mu,0;\kappa;\alpha) = \begin{cases} (1-\alpha)V^{(1)}((s,f)\mu,0;\kappa;\alpha) - \kappa & \text{if arm 1 is optimal,} \\ V^{(1)}((s,f)\mu,0;\kappa;\alpha) - \kappa/(1-\alpha) & \text{if arm 2 is optimal.} \end{cases} \tag{2.3}$$

Eick (1985a, Corollary 6.8) show that $V^{(1)}((s,f)\mu,0;\kappa;\alpha)$ is increasing in s and decreasing in f. Therefore Δ is increasing in s and decreasing in f. Although $V^{(1)}((s,f)\mu,0;\kappa;\alpha)$ is nondecreasing in κ, $V^{(1)}((s,f)\mu,0;\kappa;\alpha) - \kappa/(1-\alpha)$ is decreasing in κ since $\partial V^{(1)}((s,f)\mu,0;\kappa;\alpha)/\partial\kappa < \alpha/(1-\alpha)$. □

The following theorem says that DAI's exist when p = 0.

Theorem 2.5. For all μ and α, the DAI exists for the $(\mu,0;\kappa;\alpha)$-bandit.

Proof. From Lemma 2.4, $\Delta(\mu,0;\kappa;\alpha)$ is decreasing in κ. The hypothesis on μ in Lemma 2.4 are not required to show that Δ is monotone in κ. So the equation $\Delta(\mu,0;\kappa;\alpha) = 0$ has a unique solution in κ. □

3. Gittins Procedures

For classical bandits with k independent arms, Gittins and Jones (1974) show that the following procedure gives rise to an optimal strategy. Evaluate the DAI for each of the k arms. The optimal arm is always one with the largest index. So the arm with the largest index is optimal initially. It remains optimal until its index is no longer largest. At this point an arm whose index was second-best originally becomes optimal. I refer to this as a Gittins procedure. The optimality of Gittins procedures is particularly interesting because it reduces a k-dimensional problem into k one-dimensional problems. Berry and Fristedt (1985), Theorem 6.2.1, show that for classical bandits, Gittins procedures are optimal only if the discounting is geometric.

3.1. Gittins Procedures for Delayed Response Bandits

For delayed response bandits with k independent arms the upcoming Theorem 3.1 shows that when $p_1 = \cdots = p_k = 0$ the arm indicated by the Gittins procedure is optimal. In this case the DAI's exist and so the Gittins procedure is defined from Theorem 2.5.

The statement and proof of this theorem is a modification of that of Whittle (1980).

Theorem 3.1. Suppose $p_1 = \cdots = p_k = 0$. Then the optimal initial selections for the $(\mu_1,0;\cdots;\mu_k,0;\alpha)$-bandit are those i for which

$$\Lambda(\mu_i,0;\alpha) = \bigvee_{j=1}^{k} \Lambda(\mu_j,0;\alpha). \tag{3.1}$$

Furthermore

$$V(\mu_1,0;\cdots;\mu_k,0;\alpha) = \frac{1}{1-\alpha} \lim_{\rho\to\infty} \{\rho-(1-\alpha)^k \int_0^\rho \prod_{j=1}^{k} \frac{\partial}{\partial\kappa} V(\mu_j,0;\kappa;\alpha)d\kappa\}. \quad (3.2)$$

I will temporarily delay the proof of Theorem 3.1. Assume initially that for all (s_i,f_i),

$$E[X_{ij}|(s_i,f_i)\mu_i] \leq M < \infty, \quad (3.3)$$

for $i = 1,\cdots, k$. A necessary and sufficient condition for (3.3) is that the support of μ_i is bounded away from 1.

Let $\frac{\partial}{\partial\kappa}$ be the right derivative operator. The following lemma says that $(1-\alpha)\frac{\partial}{\partial\kappa} V(\mu_i,p_i;\kappa;\alpha)$ exists and is the cumulative distribution function of a probability measure.

Lemma 3.2. Assume $E[X_{ij}|(s_i,f_i)\mu_i] \leq M < \infty$ for $i = 1,\cdots, k$. Then for all $p_1,\cdots, p_k$ the functions

$$\Psi(\mu_i,p_i;\kappa;\alpha) = (1-\alpha)\frac{\partial}{\partial\kappa} V(\mu_i,p_i;\kappa;\alpha), \quad (3.4)$$

$i = 1,\cdots, k$, are the cumulative distribution functions of probability measures with support contained in $[0,M]$.

Proof. Since $V(\mu_i,p_i;\kappa;\alpha)$ is convex in κ (Eick, 1985a, Theorem 4.1), a right-continuous nondecreasing version of $\partial V(\mu_i,p_i;\kappa;\alpha)/\partial\kappa$ exists. For $\kappa \leq 0$, $V(\mu_i,p_i;\kappa;\alpha) = E[X_{i1}|\mu_i]/(1-\alpha)$ and for $\kappa \geq M$, $V(\mu_i,p_i;\kappa;\alpha) = \kappa/(1-\alpha)$. □

Proof of Theorem 3.1. Under assumption (3.3) the theorem is proved by an argument similar to that in Whittle (1980). The general case for arbitrary μ_i (restricted such that $E[X_{ij}|\mu_i] < \infty$) follows by truncation: $d\mu_i^* = 1_{[0,t]}d\mu_i + \mu_i(t,1]\delta_t$, $i = 1,\cdots, k$, and approximation. □

The proof fails if $p_i > 0$ for some i. In this case $\Lambda(\mu_i,p_i;\alpha)$ may not exist if $\alpha \geq 1/2$. However even if $\Lambda(\mu_i,p_i;\alpha)$ does exist, Whittle's argument fails on a more fundamental level. The two-armed classical bandit with immediate responses, arm 2 known with life expectancy κ, and geometric discounting is a stopping problem (Berry and Fristedt, 1979, Theorem 2.1). There exists an optimal strategy which indicates arm 1 at stages 1 through N and arm 2 at all subsequent stages. The stopping time N is random and can be 0 or ∞ with positive probability. A consequence of this is that when arm 2 is optimal an optimal strategy is to indicate arm 2 at all subsequent stages. So $\kappa \geq \Lambda(\mu;\alpha)$ if and only if $V(\mu;\kappa;\alpha) = \kappa/(1-\alpha)$. This characterization is used at a critical step in the proof of Theorem 3.1 to conclude the form of the optimal strategy.

The delayed response $(\mu,p;\kappa;\alpha)$-bandit is a not stopping problem (see Eick, 1985a, Section 3.2) when $p > 0$. The optimal strategy may indicate arm κ while waiting for patients on arm 1 to respond. For delayed response bandits

$$V(\mu,p;\kappa;\alpha) \geq \kappa/(1-\alpha) \tag{3.5}$$

with strict inequality when $p > 0$ for all κ. But when $p = 0$, equality holds in (3.5) if and only if $\kappa \geq \Lambda(\mu,0;\alpha)$. Only in this setting does Whittle's argument provide a proof of the optimality of the Gittins procedure for delayed response

bandits.

I conjecture that DAI's exist for all states for the delayed response bandit and that the Gittins procedures are optimal. In the following Chapter I discuss numerical techniques to calculate DAI's and apply them to the beta family.

4. Computations of Gittins Indices

For delayed response bandits it is difficult to compute $\Lambda(\mu,p;\alpha)$ since it is defined as the zero of $\Delta(\mu,p;\kappa;\alpha) = V^{(1)}(\mu,p;\kappa;\alpha) - V^{(2)}(\mu,p;\kappa;\alpha)$. In this section I present a class of strategies whose worths closely approximate $V^{(i)}$ but are still relatively easy to calculate. I then use them to approximate $\Delta(\mu,p;\kappa;\alpha)$ and calculate $\Lambda(\mu,p;\alpha)$ for the beta distribution. Throughout this section I consider calculations for the $(\mu,p;\kappa;\alpha)$-bandit.

For $1/2 \leq \beta \leq 1$, let $\tau_{i\beta}$ be the strategy which indicates arm i initially and then does as well as possible under the following restrictions:

(1) $\tau_{i\beta}$ indicates arm 1 when the conditional probability that arm 1 is better than arm 2 exceeds β and arm 2 when the converse holds.

(2) When either case above occurs, $\tau_{i\beta}$ indicates that arm at all subsequent stages ignoring any forthcoming information from the information bank.

(3) At time n, $\tau_{i\beta}$ selects one arm to indicate at all subsequent stages.

For $\beta = 1$ and $n = \infty$, τ_{i1} is optimal among those strategies which indicate arm i initially. For $\beta = 1/2$ and $n = 1$, $\tau_{i,1/2}$ is the best one step strategy that indicates arm i initially and subsequently indicates the arm which has the longer expected lifetime at time 1.

The strategy $\tau_{i\beta}$ is truncated when either of (1) or (3) applies. Time n is

the maximum the truncation time. When $n < \infty$ the worth of $\tau_{i\beta}$ can be calculated recursively in at most n+1 steps. The restrictions in (1) and (2) reduce the number of states which must be considered in the recursion which decreases the memory requirements. For $\beta = 1$ and $n = 30$ an evaluation of both $W(\tau_{11})$ and $W(\tau_{21})$ required 30 minutes of CPU time on a VAX 11/750 and a 126325 word array, but for $\beta = .95$ the same calculation required only a 2600 word array. For general n storage requirements are of order $n^4/6$ when $\beta = 1$.

Table 4.1 shows $W(\tau_{1\beta})$, $W(\tau_{2\beta})$, and memory requirements for $\alpha = .8$ and $n = 30$ as β varies from .5 to 1.0 by .05 for the beta distribution. As β increases the memory requirements increase sharply; but even for small β, $W(\tau_{i\beta})$ is a very good approximation to $V^{(i)}$.

Table 4.1.

Worths of $\tau_{1\beta}$, $\tau_{2\beta}$ with $n = 30$ for the $(\mu,0;1;.8)$-bandit

$$d\mu(\theta) \propto \theta^{3/4}(1-\theta)^2 d\theta$$

β	$W(\tau_{1\beta})$	$W(\tau_{2\beta})$	Memory	Approx Bound
0.500	4.7500	5.0000	1	3.7568
0.550	4.7500	5.0000	211	3.7568
0.600	5.7269	5.0000	402	3.7999
0.650	5.7269	5.0000	615	3.3798
0.700	5.7270	5.0000	831	3.3787
0.750	5.7284	5.0000	1071	3.3376
0.800	5.7320	5.0000	1332	3.1515
0.850	5.7325	5.5860	1648	2.8671
0.900	5.3725	5.5860	2052	2.8490
0.950	5.7325	5.5860	2662	0.5324
1.000	5.7325	5.5860	126325	0.0058

$$d\mu(\theta) \propto \theta^1(1-\theta)^2 d\theta$$

β	$W(\tau_{1\beta})$	$W(\tau_{2\beta})$	Memory	Approx Bound
0.500	5.0000	5.0000	1	4.0000
0.550	6.4736	5.0000	138	3.4892
0.600	6.4736	5.0000	379	3.4892
0.650	6.4736	5.0000	598	3.4892
0.700	6.4737	5.0000	819	3.4884
0.750	6.4741	6.1793	1057	3.4524
0.800	6.4752	6.1801	1331	3.3528
0.850	6.4787	6.1830	1642	2.8954
0.900	6.4791	6.1833	2045	0.8268
0.950	6.4791	6.1833	2631	0.7915
1.000	6.4791	6.1833	126325	0.0062

$$d\mu(\theta) \propto \theta^{5/4}(1-\theta)^2 d\theta$$

β	$W(\tau_{1\beta})$	$W(\tau_{2\beta})$	Memory	Approx Bound
0.500	7.1731	5.0000	1	4.3636
0.550	7.1731	5.0000	203	4.3636
0.600	7.2832	5.0000	391	3.7119
0.650	7.2832	5.0000	594	3.7119
0.700	7.2836	6.8269	814	3.7087
0.750	7.2862	6.8290	1047	3.6761
0.800	7.2920	6.8336	1316	3.5278
0.850	7.2954	6.8363	1630	1.3380
0.900	7.2957	6.8366	2029	1.1403
0.950	7.2957	6.8366	2620	1.1097
1.000	7.2957	6.8366	126325	0.0067

The next theorem bounds the error when $W(\tau_{i\beta})$ is used to approximate $V^{(i)}$. The bound, approximated in the last column of Table 4.1, is crude since it comes from comparison to an omniscient strategy which indicates the better arm.

Let γ_i be optimal among those strategies indicating i initially. Then $\gamma_i = \tau_{i1}$ when $n = \infty$ and $W(\gamma_i) = V^{(i)}$. Let $\underset{\sim}{A}$ be the class of states for which

condition (1) of the definition of $\tau_{i\beta}$ applys. Then $((s,f)\mu,p;\kappa;\alpha) \in \underset{\sim}{A}$ if and only if $P\{\theta/(1-\theta) > \kappa | (s,f)\mu\} \geq \beta$ or $\leq 1-\beta$. Let $\underset{\sim}{B}$ be the class of all accessible states at time n. For state $((s,f)\mu,p;\kappa;\alpha)$, define $P_j\{(s,f)\mu,p;\kappa;\alpha) | \gamma_i\}$ as the probability that this is the current state at time j when following γ_i.

Theorem 4.1. For all μ, p, κ, and α,

$$
\begin{aligned}
&V^{(i)}(\mu,p;\kappa;\alpha) - W(\tau_{i\beta}) \\
&\leq (1-\beta) \sum_{j=1}^{n-1} \frac{\alpha^j}{1-\alpha} \sum_{\underset{\sim}{A}} P_j\{((s,f)\mu,p;\kappa;\alpha) | \gamma_i\} E\left[\left|\frac{\theta}{1-\theta} - \kappa\right| | (s,f)\mu\right] \\
&\quad + \frac{\alpha^n}{1-\alpha} \sum_{\underset{\sim}{B}} P_n\{((s,f)\mu,p;\kappa;\alpha) | \gamma_i\} E\left[\left|\frac{\theta}{1-\theta} - \kappa\right| | (s,f)\mu\right].
\end{aligned}
\tag{4.1}
$$

Remark. When $\beta = 1$, (4.1) simplifies:

$$
V^{(i)}(\mu,p;\kappa;\alpha) - W(\tau_{i1}) \leq \frac{\alpha^n}{1-\alpha} E\left[\left|\frac{\theta}{1-\theta} - \kappa\right|\right]. \tag{4.2}
$$

Equation (4.2) is due to Berry and Fristedt (1985) equation (2.6.3).

Proof of Theorem 4.1. Let $\sigma_{i\beta}$ mimic γ_i by indicating the same arm until condition (1), (2), or (3) of the definition of $\tau_{i\beta}$ applies. Since $\tau_{i\beta}$ is the best among the class of strategies satisfying these conditions,

$$
W(\tau_{i\beta}) \geq W(\sigma_{i\beta}).
$$

I complete the proof of Theorem 4.1 by showing that (4.1) holds with $\sigma_{i\beta}$ replacing $\tau_{i\beta}$.

Suppose the current time is $j \leq n-1$. Let Z_{γ_i} and $Z_{\sigma_{i\beta}}$ be the lifetimes of of the current patient when following γ_i and $\sigma_{i\beta}$. Except for states in A, $Z_{\gamma_i} = Z_{\sigma_{i\beta}}$. Suppose the current state $((s,f)\mu,p;\kappa;\alpha) \in \underline{A}$. With conditional probability exceeding β, $\sigma_{i\beta}$ indicates the better arm and with conditional probability at most $1-\beta$, $\sigma_{i\beta}$ indicates the inferior arm. If γ_i always indicates the better arm, or in other words if γ_i is omniscient,

$$\begin{aligned} E[Z_{\gamma_i} - Z_{\sigma_{i\beta}} | (s,f)\mu] &\leq E\left[\tfrac{\theta}{1-\theta} \vee \kappa | (s,f)\mu\right] \\ &\quad - \beta E\left[\tfrac{\theta}{1-\theta} \vee \kappa | (s,f)\mu\right] - (1-\beta)E\left[\tfrac{\theta}{1-\theta} \wedge \kappa | (s,f)\mu\right] \\ &= (1-\beta)E\left[\left|\tfrac{\theta}{1-\theta} - \kappa\right| | (s,f)\mu\right]. \end{aligned} \tag{4.3}$$

A similar calculation for an arbitrary state at time n shows:

$$\begin{aligned} E[Z_{\gamma_i} - Z_{\sigma_{i\beta}} | (s,f)\mu] &\leq E\left[\tfrac{\theta}{1-\theta} \vee \kappa | (s,f)\mu\right] - E\left[\tfrac{\theta}{1-\theta} \wedge \kappa | (s,f)\mu\right]. \\ &= E\left[\left|\tfrac{\theta}{1-\theta} - \kappa\right| | (s,f)\mu\right] \end{aligned} \tag{4.4}$$

Equation (4.1) follows from (4.3) and (4.4) by summing over the appropriate states weighted by the remaining discount sequence. □

For calculations, the bound in Theorem 4.1 is unusable since $P_j\{((s,f)\mu,p;\kappa;\alpha)|\gamma_i\}$ depends on γ_i which is unknown. A convenient approximation replaces $P_j\{((s,f)\mu,p;\kappa;\alpha)|\gamma_i\}$ with $P_j\{((s,f)\mu,p;\kappa;\alpha)|\tau_{i\beta}\}$ which

can be determined recursively when n is finite. This approximation is used in Table 4.1.

For the $(\mu,p;\kappa;\alpha)$-bandit the solution in κ to the equation $W(\tau_{1\beta}) - W(\tau_{2\beta})$ provides an estimate of $\hat{\Lambda}$ of $\Lambda(\mu,p;\alpha)$. The following proposition bounds of the estimation error.

Proposition 4.2. Suppose $|\hat{V}^{(i)} - V^{(i)}| \leq \varepsilon$, $i = 1,2$, and assume $\hat{\Lambda}$ satisfies $|\hat{V}^{(1)}(\mu,0;\hat{\Lambda};\alpha) - \hat{V}^{(2)}(\mu,0,\hat{\Lambda};\alpha)| < \delta$. If $\hat{\Lambda} \geq \Lambda$ then $|\hat{\Lambda}(\mu,0;\alpha) - \Lambda(\mu,0;\alpha)| \leq \delta + 2\varepsilon$.

Proof. The value $V^{(i)}(\mu,p;\kappa;\alpha)$ is convex nondecreasing in κ. Therefore

$$V^{(i)}(\mu,p;\Lambda;\alpha) - V^{(i)}(\mu,p;\hat{\Lambda};\alpha) = d_i(\Lambda-\hat{\Lambda}), \tag{4.5}$$

where

$$\frac{\partial}{\partial\kappa} V^{(i)}(\mu,0;\kappa;\alpha)\Big|_{\kappa=\Lambda} \leq d_i \leq \frac{\partial}{\partial\kappa} V^{(i)}(\mu,0;\kappa;\alpha)\Big|_{\kappa=\hat{\Lambda}},$$

and $\frac{\partial}{\partial\kappa}$ is the right derivative operator. However, for all κ

$$\frac{\partial}{\partial\kappa} V^{(1)}(\mu,0;\kappa;\alpha) \leq \alpha/(1-\alpha) \tag{4.6}$$

and

$$\frac{\partial}{\partial\kappa} V^{(2)}(\mu,u,\kappa;\alpha)\Big|_{\kappa=\hat{\Lambda}} \geq \frac{\partial}{\partial\kappa} V^{(2)}(\mu,0;\kappa;\alpha)\Big|_{\kappa=\Lambda} = 1/(1-\alpha). \tag{4.7}$$

The first inequality in both (4.6) and (4.7) follows since $V^{(i)}$ is convex

nondecreasing in κ and the equality in (4.7) follows since the right-derivative is maximized when arm 2 is indicated at every possible stage. Subtracting (4.5) with i = 2 from (4.5) with i = 1 gives

$$V^{(1)}(\mu,0;\hat{\Lambda};\alpha) - V^{(2)}(\mu,0;\hat{\Lambda};\alpha) = (d_1-d_2)(\Lambda-\hat{\Lambda}).$$

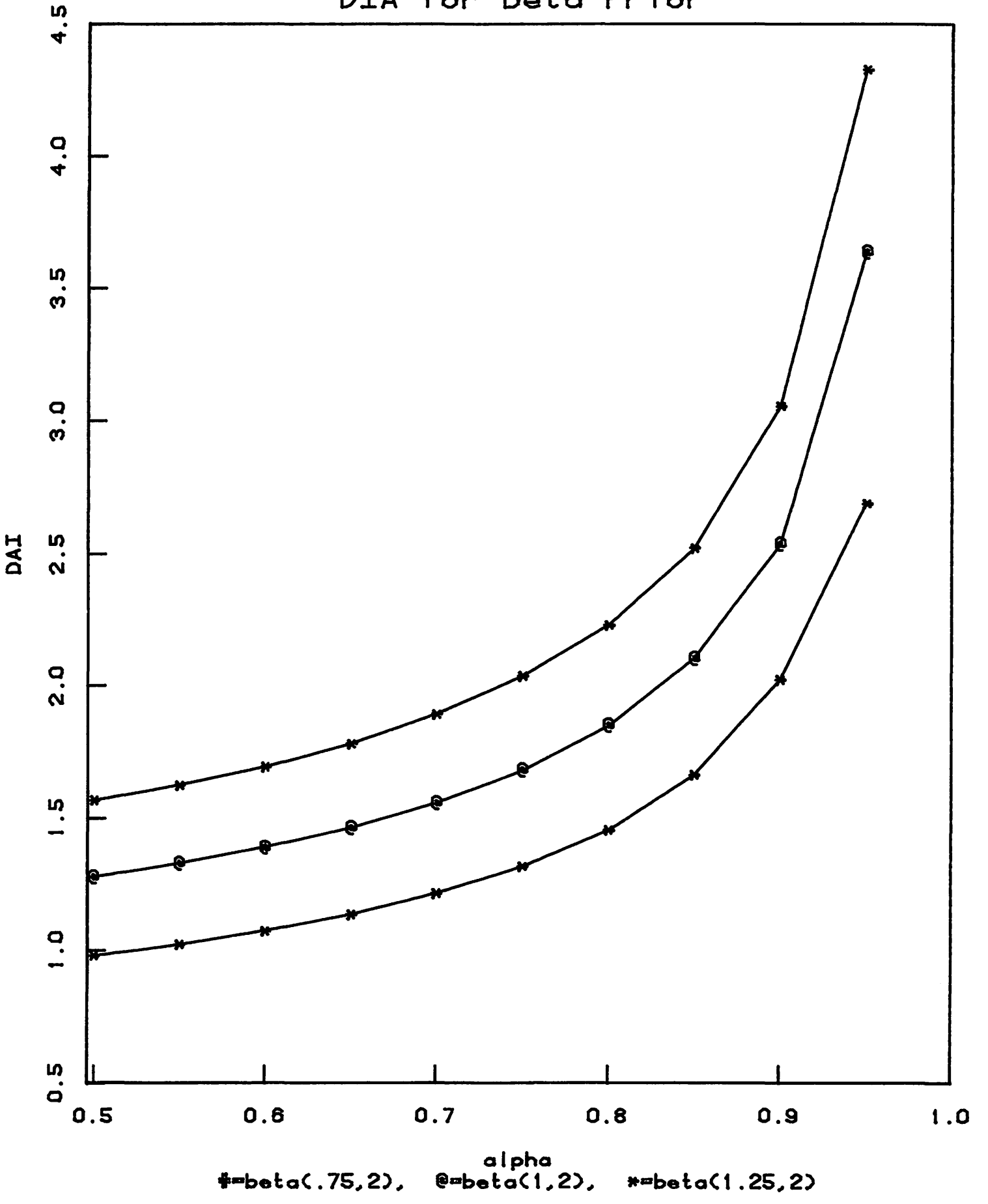
The result follows from the triangle inequality and the bounds (4.6), and (4.7) and the hypothesis on $\hat{V}^{(i)}$ and $\hat{\Lambda}$. □

Figure 4.1 shows $\Lambda((s,f)\mu,0;\alpha)$ for the beta family with density $d(s,f)\mu \propto \theta^{s-1}(1-\theta)^{f-1}d\theta$ for (s,f) = (3/4,2), (1,2), and (5/4,2) as α varies. Computational resource restrictions limited computations to $\alpha \leq .95$.

As α increases $\Lambda((s,f)\mu,0;\alpha)$ increases. This is intuitive since for a larger α there is a greater fraction of the discount sequence at future times and so there is more opportunity to take advantage of a good unknown arm. In the limit as $\alpha \to 1$, $\Lambda(s,f)\mu,0;\alpha) \to \infty$. So for sufficiently large α arm 1 is optimal. For $\alpha = 0$ (not shown on and Figure 4.1), $\Lambda((s,f)\mu,0;0) = E[X|(s,f)\mu] = s/(f-1)$. In this case there is no opportunity to take advantage of anything learned when the initial patient is treated with 1.

Figure 4.1 shows that $\Lambda((s,f)\mu,0;\alpha)$ is an increasing functions of s. This is true in general since $\Delta((s,f)\mu,0;\kappa;\alpha)$ is increasing in s. Similarly, $\Lambda((s,f)\mu,0;\alpha)$ is decreasing in f.

Figure 4.1
DIA for Beta Prior



5. Discussion

In this paper I describe the optimal strategies for the two-armed delayed response bandit with geometric discounting. When both arms are unknown and $\alpha < 1/2$, Theorem 2.1 says that there exists a manifold in $(s_1,f_1;s_2,f_2)$-space which determines the optimal treatment. The class of strategies introduced in Section 4 can be easily modified to approximate this manifold in applications.

When the expected lifetime of arm 2 is known to be κ, I show in Theorem 2.3 and 2.5 that the optimal arm is determined by a DAI when either $\alpha < 1/2$ or $p = 0$. For the k-armed bandit with delayed responses Theorem 3.1 says that the Gittins procedure determines an optimal selection when $p_1 = \cdots = p_k = 0$.

The strategies presented in Section 4 provide a method estimating the DAI's and Thereby implementing the Gittins procedure. Theorem 4.1 and Proposition 4.2 show that the estimates converge as $\beta \to 1$. The numerical results in Table 4.1 indicate that the convergence is very fast leading to significant computational savings.

Acknowledgment

I would like to thank Donald A. Berry for his many helpful comments.

REFERENCES

Armitage, P. (1985). The search for optimality in clinical trials. Int. Statist. Rev. 53:1-13.

Bather, J.A. (1981). Randomized allocation of treatments in sequential experiments. J.R. Royal Statist. Soc. B 43:265-292.

Bellman, R. (1956). A problem in sequential design of experiments. Sankhya A 16:221-229.

Berry, D.A. (1972). A Bernoulli two-armed bandit. Ann. Math. Statist. 43:872-897.

Berry, D.A. and Fristedt, B. (1979). Bernoulli one-armed bandits--arbitrary discount sequences. Ann. Statist. 7:1086-1105.

Berry, D.A. and Fristedt, B. (1985). Bandit Problems Sequential Allocation of Experiments. Chapman-Hall, London.

Blackwell, D. (1965). Discounted dynamic programming. Ann. Math. Statist. 36:226-235.

Bradt, R.N., Johnson, S.M., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. Ann. Math. Statist. 27:1060-1070.

Eick, S.G. (1985a). Two-armed bandits with delayed responses. Univ. of Minnesota Statistics Tech. Rep. No. 456.

Eick, S.G. (1985b). The two-armed bandit with delayed responses. Univ. of Minnesota Statistics Tech. Rep. No. 457.

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

Gittins, J.C., and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. Progress in Statistics (ed. by J. Gani, et al.), pp. 241-266. North-Holland, Amsterdam.

Gittins, J.C. (1979). Bandit processes and dynamic allocation indices (with discussion). J.R. Statist. Soc. B 41:148-177.

Kumar, P.R., and Seidman, T.A. (1981). On the optimal solution to the one-armed bandit adaptive control problem. IEEE Trans. Autom. Control. 26:1176-1184.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bull. Amer. Math. Soc. 58:527-536.

Simon, R. (1977). Adaptive treatment assignment methods and clinical trials. Biometrics 33:743-744.

Whittle, P. (1980). Multi-armed bandits and the Gittins Index. J.R. Soc. B. 42:143-149.