ELSEVIER

Statistics & Probability Letters 60 (2002) 211–217

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Asymptotic properties of bandit processes with geometric responses

Xikui Wang*

Department of Statistics, University of Manitoba, Winnipeg, Man., Canada R3T 2N2

Received October 2001; received in revised form July 2002

Abstract

Asymptotic properties of optimal strategies for two-armed bandit processes with geometrically distributed survival times are derived. These results provide asymptotic boundary conditions and further extend structure properties of optimal strategies for bandit processes with delayed responses.

MSC: 62L05; 62L15

Keywords: Bandit processes; Clinical trial; Dynamic programming; Geometric survival time; Optimal strategy

1. Introduction

Bandit processes with geometrically distributed survival times are studied in Eick (1988) and Wang (2000). Bandit models have been proposed as alternative adaptive designs of clinical trials when the traditional randomized designs become ethically infeasible in desperate medical situations (Pullman and Wang, 2001). The monograph by Berry and Fristedt (1985) is an excellent introduction to the subject of bandit problems.

Assume two treatments x and y for a common disease. Observations of patients' survival times after treatments may be censored. Under treatment y, patients' survival times Y have a known expected value $k>1$. Patients' survival times X on the unknown treatment x are conditionally independent and geometrically distributed with an unknown probability of success $\theta \in (0,1)$. At times $0,1,2,\ldots,$ patients are recruited into the trial sequentially and treated one at a time. Our objective is to sequentially allocate treatments to patients so as to maximize the total expected discounted survival times for all patients.

* Corresponding author. Fax: +1-204-474-7621.

E-mail address: xikui_wang@umanitoba.ca (X. Wang).

PII: S0167-7152(02)00319-X

Under some regularity conditions, Eick (1988) characterizes the optimal strategy by break-even values of parameters. Unfortunately these values are formidable computationally, even in very simple situations. Wang (2000) initiates a study of structural properties of these break-even values, which could potentially lead to efficient computations and simulations.

The purpose of this paper is to explore further structural properties of these break-even values. Main results in Eick (1988) and Wang (2000) are summaried in Section 2. New results and proofs are presented in Section 3.

2. The model and background

Assume that the unknown probability of success $\theta \in (0,1)$ follows a prior distribution μ. Sufficient statistics consist of s and f, which are, respectively, the observed numbers of successes and failures under the unknown treatment x at the current time point. The posterior distribution for θ is then of the form $(s,f)\mu$, with $(0,0)\mu = \mu$. Under the currently updated prior $(s,f)\mu$, the conditional expected survival time under the unknown treatment is denoted as $E(X|(s,f)\mu)$.

A strategy π is a sequence of rules specifying a treatment to be allocated to the patient at time $0,1,2,3,\ldots,$ given current information. If Z_i denotes the ith patient's survival time under a strategy π, then the worth of the strategy π is defined as the expected total discounted survival time

$$W(\pi) = E_\pi\left(\sum_{i=1}^{\infty} \alpha_i Z_i\right),$$

where $D = (\alpha_1, \alpha_2, \ldots)$ is a discount sequence satisfying $\alpha_i \geqslant 0$ and $\sum_{i=1}^{\infty} \alpha_i < \infty$. The goal is to find an optimal strategy π^* such that $W(\pi^*) = V = \max_\pi W(\pi)$. A treatment is optimal for a given patient if it is allocated by an optimal strategy.

Due to the loss-of-memory property of the geometric distribution, the bandit process with delayed responses becomes a discrete time Markov decision process. The state is characterized by $((s,f)\mu, r, D)$, where r is the number of patients previously treated with the unknown treatment x who are still alive. Eick (1988) calls r the size of the information bank. The action space is $\{1,2\}$.

At each state $((s,f)\mu, r, D)$, let $V^{(x)}((s,f)\mu, r, D)$ ($V^{(y)}((s,f)\mu, r, D)$) be the worth of the strategy that allocates initially the unknown treatment x (respectively, the known treatment y) and then follows an optimal strategy. The dynamic programming equation becomes

$$V((s,f)\mu, r, D) = \max\{V^{(x)}((s,f)\mu, r, D), V^{(y)}((s,f)\mu, r, D)\}.$$

Moreover,

$$\Delta((s,f)\mu, r, D) = V^{(x)}((s,f)\mu, r, D) - V^{(y)}((s,f)\mu, r, D)$$

is the advantage of the unknown treatment x over the known treatment y and characterizes the initially optimal selection of treatment: treatment x (treatment y) is optimal at state $((s,f)\mu, r, D)$ if and only if $\Delta((s,f)\mu, r, D) \geqslant (\leqslant) 0$. When $\Delta((s,f)\mu, r, D) = 0$, both treatments x and y are optimal and there is no unique optimal selection.

We assume the following conditions which are also assumed in Eick (1988):

Condition A. $\alpha_i \geqslant \sum_{j=i+1}^{\infty} \alpha_j$ for $i = 1, 2, \ldots$;

Condition B. μ is not concentrated at a single point, and $\mu\{(0,1)\} = 0$.

Under Condition A, Eick (1988) shows that for given μ and r, $\Delta((s,f)\mu, r, D)$ is nondecreasing in s and nonincreasing in f and k. Furthermore, $\Delta((s,f)\mu, r, D)$ is strictly monotone if there is a strict inequality in Condition A for $i = 1$ and Condition B is also true. This implies that the optimal initial selection of treatment is characterized by the equation $\Delta((s,f)\mu, r, D) = 0$.

Unfortunately this equation is formidable to solve in general. To gain insights on structural properties of the solutions to this equation, Wang (2000) demonstrates that for all f, k, r and μ, if $\Delta((s^*, f)\mu, r, D) = 0$ and Condition A is true, then $0 \leqslant s^* \leqslant s_1^*$ where $E(X|(s_1^*, f)\mu) = k$. Moreover, if there is a strict inequality in Condition A for $i = 1$ and Condition B is also true, then s^* is a nondecreasing function of both f and k. Assume that $D = (1, \alpha, \alpha^2, \ldots)$ is geometric and denote $D_n = (1, \alpha, \alpha^2, \ldots, \alpha^{n-1}, 0, 0, \ldots)$. For given f and k, let s_n^* be such that $\Delta((s_n^*, f)\mu, 0, D_n) = 0$. Then under Condition A,

$$0 \leqslant \cdots \leqslant s_n^* \leqslant \cdots \leqslant s_2^* \leqslant s_1^*$$

and the limit $s^* = \lim_{n\to\infty} s_n^*$ exists and satisfies $\Delta((s^*, f)\mu, 0, D) = 0$. Moreover, if $E(X|(s,f)\mu)$ is nonincreasing in f and strictly increasing in s, then $s^* < s_1^*$. If Condition B is also true and $q = P(X = 1|(0,f)\mu) = 1$, then $s^* > 0$.

It is conjectured that similar results hold in general when $r > 0$. These general structural properties have been demonstrated through simulations but theoretical proofs have yet to be found. In this note, we prove some limiting properties for the sequence s_n^*. These properties provide asymptotic boundary conditions for s_n^*.

3. Asymptotic boundary structures for break-even values of s

To explicitly express the dependency of s_n^* on r and f, write s_n^* as $s_n^*(r,f)$ where $\Delta((s_n^*, f)\mu, r, D_n) = 0$. We show that for a given f, $s_n^*(r,f)$ approaches $s_1^*(0,f)$ as the size r of the information bank goes to infinity. On the other hand for given r, $s_n^*(r,f)$ approaches infinity as the number of observed deaths f on the unknown treatment goes to infinity. Throughout this section, we further assume

Condition C. $\lim_{f\to\infty} E(X|(s,f)\mu) = 0$ and $\lim_{f\to\infty} p = 0$ for any given s and μ, where p is the probability of success under the unknown treatment x at the state $((s,f)\mu, r, D)$. Moreover, $E(X|(s,f)\mu)$ is nonincreasing in f and strictly increasing in s, and is continuous in s.

It is worth pointing out that these assumptions are intuitive and nonrestrictive. They are true, for example, when μ is a beta distribution because $(s,f)\mu$ is again a beta distribution. In what follows, denote $q = 1 - p$, $D = (1, \alpha, \alpha^2, \ldots)$ and $D_n = (1, \alpha, \alpha^2, \ldots, \alpha^{n-1}, 0, 0, \ldots)$.

Lemma 1. *Given s, μ and r, the known arm is always optimal when f goes to infinity. That is,* $\lim_{f\to\infty} V((s,f)\mu, r, D_n) = k\sum_{m=0}^{n-1} \alpha^m$ *for any n.*

Proof. We prove by induction. The result is clearly true when $n=1$ because $\Delta((s,f)\mu,r,D_1)=E(X|(s,f)\mu)-k$. Suppose that the result is true when $n=m$. Then for $n=m+1$,

$$\Delta((s,f)\mu,r,D_{m+1})=E(X|(s,f)\mu)-k + \alpha\sum_{i=0}^{r+1}\binom{r+1}{i}V((s+j,f+r+1-j)\mu,j,D_m)p^jq^{r+1-j} - \alpha\sum_{i=0}^{r}\binom{r}{i}V((s+j,f+r-j)\mu,j,D_m)p^jq^{r-j}.$$

Since V is bounded, based on Condition C and the hypothesized result at $n=m$,

$$\lim_{f\to\infty}\Delta((s,f)\mu,r,D_{n+1})=-k+\alpha\lim_{f\to\infty}V((s,f+r+1)\mu,0,D_m) - \alpha\lim_{f\to\infty}V((s,f+r)\mu,0,D_m)=-k<0. \qquad \square$$

Lemma 2.

$$\sum_{i=0}^{r+1}\binom{r+1}{i}F(i)=\sum_{i=0}^{r}\binom{r}{i}[F(i)+F(i+1)]$$

for any function F.

Proof. It is well known that

$$\binom{r+1}{i}=\binom{r}{i}+\binom{r}{i-1}.$$

So

$$\sum_{i=0}^{r+1}\binom{r+1}{i}F(i)=F(r+1)+F(0)+\sum_{i=1}^{r}\binom{r}{i}F(i)+\sum_{i=1}^{r}\binom{r}{i-1}F(i) =\sum_{i=0}^{r}\binom{r}{i}F(i)+\sum_{i=1}^{r+1}\binom{r}{i-1}F(i) =\sum_{i=0}^{r}\binom{r}{i}[F(i)+F(i+1)]. \qquad \square$$

Our first asymptotic boundary property has been observed without proof in Eick (1988).

Theorem 1. $\lim_{r\to\infty}[\Delta((s,f)\mu,r,D_n)-\Delta((s,f)\mu,r,D_1)]=0$ *for any* n.

Proof. From Lemma 2,

$$
\begin{aligned}
&\Delta((s,f)\mu,r,D_n)-\Delta((s,f)\mu,r,D_1)\\
&\quad=\alpha\sum_{i=0}^{r+1}\binom{r+1}{i}V((s+i,f+r+1-i)\mu,i,D_{n-1})p^iq^{r+1-i}\\
&\qquad-\alpha\sum_{i=0}^{r}\binom{r}{i}V((s+i,f+r-i)\mu,i,D_{n-1})p^iq^{r-i}\\
&\quad=\alpha\sum_{i=0}^{r}\binom{r}{i}[V((s+i,f+r+1-i)\mu,i,D_{n-1})q\\
&\qquad+V((s+i+1,f+r-i)\mu,i+1,D_{n-1})p\\
&\qquad-V((s+i,f+r-i)\mu,i,D_{n-1})]p^iq^{r-i}\\
&\quad=\alpha\sum_{i=0}^{r}\binom{r}{i}W(i,r)p^iq^{r-i}.
\end{aligned}
$$

This is the Euler sum of the triangular sequence $W(i,r)$. Clearly, this sequence goes to 0 as $r\to\infty$ for any fixed i because the known arm is always optimal for each of the three V's and $p+q=1$. Therefore the Euler sum converges to 0 as well. That is, $\lim_{r\to\infty}[\Delta((s,f)\mu,r,D_n)-\Delta((s,f)\mu,r,D_1)]=0$. □

Replacing s by $s_n^*(r,f)$ in Theorem 1 implies $\lim_{r\to\infty}\Delta((s_n^*(r,f),f)\mu,r,D_1)=0$ and therefore $\lim_{r\to\infty}E(X|((s_n^*(r,f),f)\mu)=k$. Because of the continuity of $E(X|((s,f)\mu)$ in s, we have

$$E\left(X\,\middle|\,\left(\lim_{r\to\infty}(s_n^*(r,f),f)\mu\right)\right)=k=E(X|((s_1^*(0,f),f)\mu)).$$

Therefore from Condition C,

Corollary 1. $\lim_{r\to\infty}s_i^*(r,f)=s_1^*(0,f)$ *for any* $i=2,3,\ldots$.

From Wang (2000), s^* is a nondecreasing function of f and hence the limit $\lim_{f\to\infty}s_n^*(r,f)$ exists. In fact,

Theorem 2. *For any given* r *and* D_n, $\lim_{f\to\infty}s_n^*(r,f)=\infty$.

Proof. $\Delta((s_n^*(r,f),f)\mu,r,D_n)=0$ implies

$$E(X|(s_n^*(r,f),f)\mu)+\alpha\sum_{i=0}^{r+1}\binom{r+1}{i}V((s_n^*(r,f)+i,f+r+1-i)\mu,i,D_{n-1})p^iq^{r+1-i}$$

$$=k+\alpha\sum_{i=0}^{r}\binom{r}{i}[V((s_n^*(r,f)+i,f+r-i)\mu,i,D_{n-1})p^iq^{r-i}]. \quad (*)$$

If $\lim_{f\to\infty}s_n^*(r,f)=M<\infty$, then $E(X|(s_n^*(r,f),f)\mu)\leqslant E(X|(M,f)\mu)$ and $\lim_{f\to\infty}E(X|(s_n^*(r,f),f)\mu)=0$ from Condition C. Moreover, the same arguments in the proof of Theorem 1 imply

$$\lim_{f\to\infty}\left[\alpha\sum_{i=0}^{r+1}\binom{r+1}{i}V((s+i,f+r+1-i)\mu,i,D_{n-1})p^iq^{r+1-i}\right.$$

$$\left.-\alpha\sum_{i=0}^{r}\binom{r}{i}[V((s+i,f+r-i)\mu,i,D_{n-1})p^iq^{r-i}]\right]=0.$$

Taking the limit as $f\to\infty$ on both sides of equation $(*)$ yields a contradiction, $k=0$. Hence $\lim_{f\to\infty}s_n^*(r,f)=\infty$. □

Finally, we establish the relationship between $s_n^*(r,f)$ and $s_n^*(0,f)$ when $f\to\infty$. On one hand, $\Delta((s_n^*(r,f),f)\mu,r,D_n)=0$. On the other hand,

Theorem 3. $\lim_{f\to\infty}\Delta((s_n^*(0,f),f)\mu,r,D_n)=0$ *for any* r *and* D_n.

Proof. Write $s_n^*(0,f)=s_n^*$. Then $\Delta((s_n^*,f)\mu,0,D_n)=0$ implies

$$E(X|(s_n^*,f)\mu)-k=\alpha V((s_n^*,f)\mu,0,D_{n-1})$$

$$-\alpha V((s_n^*+1,f)\mu,1,D_{n-1})p-\alpha V((s_n^*,f+1)\mu,0,D_{n-1})q.$$

Therefore,

$$\Delta((s_n^*,f)\mu,r,D_n)=E(X|(s_n^*,f)\mu)-k$$

$$+\alpha\sum_{i=0}^{r+1}\binom{r+1}{i}V((s_n^*+j,f+r+1-j)\mu,j,D_{n-1})p^jq^{r+1-j}$$

$$-\alpha\sum_{i=0}^{r}\binom{r}{i}V((s_n^*+j,f+r-j)\mu,j,D_{n-1})p^jq^{r-j}$$

$$=\alpha V((s_n^*,f)\mu,0,D_{n-1})-\alpha V((s_n^*+1,f)\mu,1,D_{n-1})p$$

$$
\begin{aligned}
&- \alpha V((s_n^*, f+1)\mu, 0, D_{n-1})q \\
&+ \alpha \sum_{i=0}^{r+1} \binom{r+1}{i} V((s_n^* + j, f + r + 1 - j)\mu, j, D_{n-1}) p^j q^{r+1-j} \\
&- \alpha \sum_{i=0}^{r} \binom{r}{i} V((s_n^* + j, f + r - j)\mu, j, D_{n-1}) p^j q^{r-j}.
\end{aligned}
$$

Since V is bounded, under Condition C,

$$
\begin{aligned}
&\lim_{f\to\infty} \Delta((s_n^*, f)\mu, r, D_n) \\
&\quad = \alpha \lim_{f\to\infty} [V((s_n^*, f)\mu, 0, D_{n-1}) - V((s_n^*, f+1)\mu, 0, D_{n-1}) \\
&\qquad + V((s_n^*, f+r+1)\mu, 0, D_{n-1}) - V((s_n^*, f+r)\mu, 0, D_{n-1})].
\end{aligned}
$$

From Wang (2000), $s_n^* \leqslant s_{n-1}^*$, s_{n-1}^* is nondecreasing in f, and there is an optimal stopping solution when $r = 0$. Therefore, the known arm is always optimal for each of the four V's in the above expression and $\lim_{f\to\infty} \Delta((s_n^*(0, f), f)\mu, r, D_n) = 0$. □

Corollary 2. *If we assume that $\Delta(\mu, r, D_n)$ is continuous in μ and $\Delta(\mu, r, D_n) = 0$ has a unique root for μ, then $\lim_{f\to\infty}(s_n^*(0, f), f)\mu = \lim_{f\to\infty}(s_n^*(r, f), f)\mu$ given appropriate interpretations of the limits of distributions.*

Proof. The result is clear from

$$
\lim_{f\to\infty} \Delta((s_n^*(0, f), f)\mu, r, D_n) = \Delta\left(\lim_{f\to\infty}(s_n^*(0, f), f)\mu, r, D_n\right) = 0
$$

and $\Delta((s_n^*(r, f), f)\mu, r, D_n) = 0$. □

Acknowledgements

This research is supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Berry, D.A., Fristedt, B., 1985. Bandit Problems—Sequential Allocation of Experiments. Chapman & Hall, London, New York.

Eick, S.G., 1988. The two-armed bandit with delayed responses. Ann. Statist. 16, 254–265.

Pullman, D., Wang, X., 2001. Adaptive designs, informed consent, and the ethics of research. Control. Clin. Trials 22, 203–210.

Wang, X., 2000. A bandit process with delayed responses. Statist. Probab. Lett. 48, 303–307.