A NOTE ON OPTIMAL STRATEGIES OF A GENERALIZED TWO-STAGE BANDIT PROBLEM

Jack C. Yue

Department of Statistics

National Chengchi University

Taipei, Taiwan, R.O.C. 11641

csyue@nccu.edu.tw

Key Words: two-armed Bernoulli bandit; two-stage decision; optimal strategy.

ABSTRACT

In Yue (1), a two-stage approach was used to explore the Bernoulli two-armed bandit problem, where he assumed that one arm has a smaller prior variance than the other arm. In this paper, adapting Yue's assumption, we study the structure of the optimal strategy, which maximizes the expected number of successes. We confirm the conjecture of Pearson (2) that it is never optimal to allocate an equal number of observations to two identical arms in the first stage.

1. INTRODUCTION AND NOTATION

A k-armed bandit is to choose an action (or "pull") among k possible choices at each stage. One receives a payoff from the choice of the j-th arm ($1 \leq j \leq k$), as well as the information about the j-th arm. The objective is to maximize the overall payoff from a number of N pulls. Intuitively one would choose the arm with the largest payoff, but there is also a risk of choosing the inferior arm because of insufficient information. Therefore, one may choose the seemingly inferior arm (and thus a lower payoff) in exchange for more information. The "information vs. immediate payoff" question makes bandit problems difficult to solve. For more details of bandit problems, see Berry and Fristedt (3).

The optimal strategy of a k-armed bandit is often not easy to evaluate. Strategies such as one-step look-ahead (Feldman, (4)) and stay-with-a-winner (Berry, (5)) rules are only optimal in some special cases. Although theoretically the optimal rule can be achieved via elegant approaches such as Gittins index (Gittins and Jones, (6); Gittins, (7)), it is usually solved via a time-consuming method — backward induction — when N is fixed and known. For example, Hardwick and Stout (8) showed that in a Bernoulli k-armed bandit, the time required to evaluate the optimal rule would be $O(N^3), O(N^5)$, and $O(N^6)$ for one-armed, two-armed, and three-armed bandits, respectively.

In this paper, our goal is to explore the structure of an optimal rule. We will focus on rules that we shall avoid in order to reduce the effort of evaluating the optimal rule. In particular, we are interested in proving the conjecture of Pearson (2), that it is never optimal to allocate an equal number of observations to two identical arms in the first stage of a Bernoulli two-stage two-armed bandit. We will also try to discover rules that shall be excluded from consideration because they are dominated by other rules. In this study, we adapt the setting of Yue (1), i.e. a Bernoulli two-stage two-armed bandit with one arm (say, arm 2) with a smaller prior variance than the other arm.

Since the approach we use to explain Pearson's conjecture is related to the one-armed bandit problem, we will first discuss the one-armed case in the next section, and continue with the two-armed case in Section 3. For the rest of this section, we shall introduce the notations used in this study.

Let (θ_1, θ_2) be the probabilities of success of treatments 1 and 2 (i.e. arms 1 and 2), respectively. Let (π_1, π_2) be the independent prior distributions of (θ_1, θ_2). Additionally, for computational simplicity, we also assume conjugate prior for (θ_1, θ_2), i.e., $\pi_1 = Beta(\alpha, \beta)$ and $\pi_2 = Beta(c\alpha, c\beta)$. Here $c \geq 1$ indicates that arm 2 is a better known arm, since

$$Var(\theta_1|\pi_1) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \geq \frac{\alpha\beta}{(\alpha+\beta)^2(c\alpha+c\beta+1)} = Var(\theta_2|\pi_2).$$

Let K_1 and K_2 be the numbers of patients receiving treatments 1 and 2, respectively, in the

first stage. In other words, since the two-stage setting is considered, the better treatment (the one with the larger posterior mean) will be used in the second stage for the remaining $N - K_1 - K_2$ patients. Then the (expected) utility of a strategy τ, with (K_1, K_2) in the first stage, is

$$\begin{aligned} U(\tau) &= U((K_1, K_2), \pi_1, \pi_2, N) \\ &= (K_1 + K_2) \cdot \frac{\alpha}{\alpha + \beta} + (N - K_1 - K_2) \times E(\theta_1 \vee \theta_2 | \pi_1, K_1, \pi_2, K_2) \\ &= (K_1 + K_2) \cdot \frac{\alpha}{\alpha + \beta} + (N - K_1 - K_2) \times \\ &\quad \sum_{m=0}^{K_1} \sum_{j=0}^{K_2} \max\{\frac{\alpha + m}{\alpha + \beta + K_1}, \frac{c\alpha + j}{c\alpha + c\beta + K_2}\} P(s_1(K_1) = m) P(s_2(K_2) = j) \quad (1.1) \end{aligned}$$

where, for $i = 1$ and 2, $s_i(K_i)$ is the number of successes with treatment i among the K_i first-stage observations, $P(s_i)$ is the marginal probability of s_i, and $E(\theta_1 \vee \theta_2 | \pi_1, K_1, \pi_2, K_2)$ is the maximum posterior means for θ_1 and θ_2. Also, the marginal probability of s_2 satisfies $(\, 0 \le x \le K_2)$

$$P(s_2(K_2) = x) = \left(\begin{array}{c} K_2 \\ x \end{array} \right) \frac{\Gamma(c\alpha + c\beta)}{\Gamma(c\alpha)\Gamma(c\beta)} \cdot \frac{\Gamma(c\alpha + x)\Gamma(c\beta + K_2 - x)}{\Gamma(c\alpha + c\beta + K_2)}, \quad (1.2)$$

and the posterior mean of θ_2, with x successes in K_2 patients, is

$$E(\theta_2 | \pi_2, s_2(K_2) = x) = \frac{c\alpha + x}{c\alpha + c\beta + K_2}.$$

The marginal probability function of s_1 and the posterior mean of θ_1 have similar forms. Let (K_1^*, K_2^*) denote the optimal numbers of the first-stage observations assigned to treatments 1 and 2.

Note that, since $\theta_1 \vee \theta_2$ is a convex function of θ_1 and θ_2, adapting the idea of submartingale, we can show that $E(\theta_1 \vee \theta_2 | \pi_1, K_1, \pi_2, K_2)$ is a non-decreasing function of K_1, as well as K_2. In other words, more observations in the first stage (i.e., larger K_1 and K_2) would increase the utility in the second stage. Similarly, since both $E(\theta_1 | \pi_1)$ and $E(\theta_2 | \pi_2)$ are always not larger than $E(\theta_1 \vee \theta_2 | \pi_1, K_1, \pi_2, K_2)$, more observations in the first stage would

reduce the overall utility. This is exactly why the bandit problem is difficult to solve, as mentioned in the first paragraph.

In order to simplify the notation, $U(\tau) = U((K_1, K_2), \pi_1, \pi_2, N)$ will be denoted by $U((K_1, K_2))$, if there is no danger of confusion. Also, although the bandit problem can be applied to areas other than a clinical trial, in this paper we will use the terminology of a clinical trial for the sake of simplification.

2. ONE-ARMED BANDIT CASE

Parallel to Pearson's conjecture in the two-armed case, we can also formulate a similar one-armed case version. Our explanation of Pearson's conjecture is that two identical arms would create redundant information. Because an equal number of first-stage observations from two identical arms might possibly yield identical outcomes (i.e., the same number of successes and thus equal posterior means), choosing the "better" treatment for the second stage would create confusion. Applying this paradigm to a one-armed case, we shall avoid assigning K_1 first-stage patients to treatment 1 if it would produce $E(\theta_1|\pi_1, K_1) = \theta_2$. The following theorem confirms this assertion. Note that, since taking observations from the second treatment(i.e., the known arm) gains no new information, we only need to consider taking first-stage observations from the first treatment in the one-armed bandit case. Therefore, $E(\theta_1 \vee \theta_2|\pi_1, K_1)$ is used to denote $E(\theta_1 \vee \theta_2|\pi_1, K_1, \pi_2, K_2)$ since there is no need to consider sampling the second treatment.

Theorem 1. If there exist integers K and i $(K \geq i \geq 0)$ such that $\frac{\alpha+i}{\alpha+\beta+K} = d = \theta_2$ (i.e., the probability of success of the second treatment is known), then the strategy $(K, 0)$ is never optimal.

Proof: We will proceed with the proof by showing that $U((K-1, 0)) \geq U((K, 0))$. Since $d = \frac{\alpha+i}{\alpha+\beta+K}$ for some integers K and i, it is evident that

$$\frac{\alpha+i-1}{\alpha+\beta+K-1} < \frac{\alpha+i}{\alpha+\beta+K} = d < \frac{\alpha+i}{\alpha+\beta+K-1}.$$

First, let $E(\theta_1 \vee \theta_2|\pi_1, s_1(K) = j)$ denote the maximum posterior mean of θ_1 and θ_2, given

that j out of K first-stage patients from treatment 1 are successes. We shall compare the decision rule with $(K-1,0)$ to that with $(K,0)$. Since

$$E(\theta_1 \vee \theta_2|\pi_1, s_1(K-1)=j) = E(\theta_1 \vee \theta_2|\pi_1, s_1(K)=j) = d$$

for $j \leq i-2$, this implies that there is no difference between taking $K-1$ and K observations if $j \leq i-2$, as in

$$E_{x_K}\left[E(\theta_1 \vee \theta_2|\pi_1, s_1(K-1)=j)\right] = E(\theta_1 \vee \theta_2|\pi_1, s_1(K-1)=j),$$

where x_K is the extra observation from treatment 1, after taking $K-1$ observations from treatment 1. Since we will choose the treatment with the larger probability of success in the second stage, with $j \leq i-2$ and $\frac{\alpha+i-1}{\alpha+\beta+K-1} < \frac{\alpha+i}{\alpha+\beta+K} = d$, taking one more observation from the first treatment won't change our belief in choosing the better treatment in the second stage. This means that the extra observation (i.e., x_K) offers no new information, and is thus wasted.

Likewise, we can also show a similar result if $j \geq i+1$. Therefore, a difference between $E(\theta_1 \vee \theta_2|\pi_1, K-1)$ and $E(\theta_1 \vee \theta_2|\pi_1, K)$ would occur only in the cases where $s_1(K-1)=i-1$ and $s_1(K-1)=i$.

For $s_1(K-1)=i-1$, a success from the extra patient added to treatment 1 would increase the posterior mean by:

$$\frac{\alpha+i}{\alpha+\beta+K} - \frac{\alpha+i-1}{\alpha+\beta+K-1} = \frac{\beta+K-i}{(\alpha+\beta+K-1)(\alpha+\beta+K)}$$

with a probability of $\frac{\alpha+i-1}{\alpha+\beta+K-1}$; a failure, on the other hand, would decrease it by:

$$\frac{\alpha+i-1}{\alpha+\beta+K-1} - \frac{\alpha+i-1}{\alpha+\beta+K} = \frac{\alpha+i-1}{(\alpha+\beta+K-1)(\alpha+\beta+K)}$$

with a probability of $\frac{\beta+K-i}{\alpha+\beta+K-1}$. Through direct calculation it is straightforward to show that the amount of increase is equal to the amount of decrease, and thus these two terms cancel each other. Note that the marginal probability of s_1 can be derived from (1.2).

Similarly, for $s_1(K-1) = i$, adding one more patient to treatment 1 would increase the posterior mean by $\frac{\beta+K-i-1}{(\alpha+\beta+K-1)(\alpha+\beta+K)}$ with a probability of $\frac{\alpha+i}{\alpha+\beta+K-1}$, and decrease it by $\frac{\alpha+i}{(\alpha+\beta+K-1)(\alpha+\beta+K)}$ with a probability of $\frac{\beta+K-i-1}{\alpha+\beta+K-1}$. The amount of increase cancels the decrease as well.

Therefore, we can see that $E(\theta_1 \vee \theta|\pi, K-1) = E(\theta_1 \vee \theta_2|\pi_1, K)$. From equation (1.1), because $E(\theta_1 \vee \theta_2|\pi_1, K-1) \geq d$ is always true, this means that $U((K-1,0)) \geq U((K,0))$.□

We can use the result of Theorem 1 to narrow the range of choices of the optimal rule in the one-armed bandit case. For example, if $\alpha = \beta$ and $d = 1/2$, we shall not take an even number of observations in the first stage; if $\alpha = 2\beta$ and $d = 1/3$, we shall not take a number that is a multiple of three as the number of observations. Then, using the previous results of the one-armed problem, such as when $\theta_1 \sim U(0,1)$, i.e. Pearson (2), then:

$$K_1^* \approx \sqrt{\frac{(N+1)(1-\theta_2)}{\theta_2}} - 1,$$

or when N follows a geometric distribution with a mean of $1/p$ and $\theta_1 \sim U(0,1)$, as in Witmer (9), then:

$$K_1^* \approx \sqrt{(\frac{1}{p}+1)\,(\frac{1-\theta_2}{\theta_2})} - 1,$$

and then we can further refine our search for the optimal strategy.

3. TWO-ARMED BANDIT CASE

Since the two-armed case is more complicated than the one-armed case, we first consider some special cases before discussing Pearson's conjecture. One of the simplest cases is to assume that $\alpha = \beta = c = 1$, i.e., there are two identical arms with uniform $(0,1)$ priors. From (1.2), it is clear that

$$P(s_1(K) = x) = P(s_2(K) = y) = \frac{1}{K+1} \quad \text{for } 0 \leq x, y \leq K,$$

and so

$$U((K,K)) \;=\; K + (N-2K) \times \frac{4K+3}{6(K+1)},$$

$$U((K,K-1)) = K - 1/2 + (N-2K+1) \times \frac{4K+3}{6(K+1)},$$

and

$$U((K+1,K-1)) = K + \frac{N-2K}{K(K+2)} \sum_{i=0}^{K+1} \sum_{j=0}^{K-1} \max\{\frac{i+1}{K+3}, \frac{j+1}{K+1}\}.$$

Comparing these utility functions, it is straightforward that $U((K,K-1)) > U((K,K))$. Also, the difference between $U((K+1,K-1))$ and $U((K,K))$ is

$$\sum_{j=[\frac{K-1}{2}]}^{K-1} [2j+1-K]^+ - \frac{K(K-1)}{6}, \tag{3.1}$$

where $[x]^+$ is the maximum of x and 0. Because (3.1) is always non-negative (Lemma 1), the strategy (K,K) is dominated by $(K+1,K-1)$. Therefore, if $\alpha = \beta = c = 1$, then assigning an equal number of observations to two identical arms is never optimal, i.e., (K,K) is dominated by $(K,K-1)$ and $(K+1,K-1)$.

Lemma 1. $\sum_{j=[\frac{K-1}{2}]}^{K-1} [2j+1-K]^+ \geq \frac{K(K-1)}{6}$.

Theorem 2. It is never optimal to use $K_1 = K_2$ when two arms are identical (i.e., $c = 1$) and $\alpha = \beta = 1$, since it is dominated by $(K,K-1)$ and $(K+1,K-1)$.

Based on the result of Theorem 2, it is natural to guess that (K,K) is dominated by $(K,K-1)$ (or $(K+1,K-1)$) for the general case not requiring $\alpha = \beta = 1$. We shall first use $(K,K-1)$ to check if this is true. Under the two-stage setting, the utility $U((K,K),\pi_1,\pi_2,N)$ is equivalent to the sum of $K \cdot \frac{\alpha}{\alpha+\beta}$ and $U((K,0),\pi_1,\pi_2(K),N-K)$, where $\pi_2(K)$ is the posterior distribution of π_2 after taking K observations from the second treatment. In other words, we can use the result in a one-armed case to show that (K,K) is not optimal. Since for all $s_2(K) = j$ $(0 \leq j \leq K)$ from the second arm, the choice of $K_1 = K$ may possibly let the first arm always produce the same number of successes as the second arm, or $\frac{\alpha+i}{\alpha+\beta+K} = \frac{\alpha+j}{\alpha+\beta+K}$. Therefore, from Theorem 1, the choice of (K,K) is not optimal, and $(K,K-1)$ will dominate (K,K). The following lemma confirms this result, and the proof that $K_1 = K_2$ is not optimal given two identical arms (i.e., $c = 1$) is thus completed.

Lemma 2. If $c = 1$, then $E(\theta_1 \vee \theta_2 | \pi_1, K, \pi_2, K) = E(\theta_1 \vee \theta_2 | \pi_1, K, \pi_2, K-1)$.

Theorem 3. It is never optimal to use $K_1 = K_2$ when two arms are identical.

4. DISCUSSION

In this study, we show that $K_1 = K_2$ is never optimal when $c = 1$, that is to say, allocating an equal number of observations to two identical arms in the first stage is never optimal. In particular, we show that $U((K, K)) \leq U((K, K-1))$ when $c = 1$, under the assumption of Yue (1). In other words, we know that $U((K, K)) \leq U((K, K-1))$ when $c = 1$ and $c \to \infty$. Because $U((K_1, K_2))$ is a continuous function of c (Yue, (1)), we can show that $U((K, K-1)) - U((K, K))$ is a continuous function of c as well. Therefore, it is natural to expect that $U((K, K)) \leq U((K, K-1))$ holds for $c > 1$.

However, when $c > 1$, we can not directly apply the relationship between $E(\theta_1 \vee \theta_2 | \pi_1, K, \pi_2, K)$ and $E(\theta_1 \vee \theta_2 | \pi_1, K, \pi_2, K-1)$ to show that $K_1 = K_2$ is not optimal, since Lemma 2 is not always true. For example, when $\alpha = \beta = 1$ and $c = 2$, $E(\theta_1 \vee \theta_2 | \pi_1, 6, \pi_2, 6) \approx 0.625765$ is greater than $E(\theta_1 \vee \theta_2 | \pi_1, 6, \pi_2, 5) = 0.625$, and is also greater than $E(\theta_1 \vee \theta_2 | \pi_1, 7, \pi_2, 5) = 0.625$. Although we believe that $K_1 = K_2$ is not optimal in the case of $c > 1$, the proof of this is more difficult.

ACKNOWLEDGMENTS

The author is grateful for the helpful comments from the anonymous reviewer. This research was supported in part by grant no. NSC 89-2118-M-004-010 from the National Science Council of Taiwan, R.O.C.

BIBLIOGRAPHY

(1) Yue, J. C. Generalized Two-stage Bandit Problem. Communications in Statistics: Theory and Methods, **1999**, *28(9)*, 2261-2276.

(2) Pearson, L. M. Treatment Allocation for Clinical Trials in Stages. Ph.D. Thesis, University of Minnesota, **1980**.

(3) Berry, D. A.; Fristedt, B. Bernoulli One-armed Bandits – Arbitrary Discount Sequences.

The Annals of Statistics, **1979**, *7(5)*, 1086-1105.

(4) Feldman, D. Contributions to the "Two-armed Bandit" Problem. Annals of Mathematical Statistics, **1962**, *33*, 847-856.

(5) Berry, D. A. A Bernoulli Two-armed Bandit. The Annals of Mathematical Statistics, **1972**, *43(3)*, 871-897.

(6) Gittins, J. C.; Jones, D. M. A Dynamic Allocation Index for the Sequential Design of Experiments. *Progress in Statistics*, J. Gani et al. Eds; **1974**, North Holland, 241-266.

(7) Gittins, J. C. Multiarmed Bandit Allocation Indices. **1989**, John Wiley and Sons.

(8) Hardwick, J.; Stout, Q. F. Optimal Few-stage Designs. Journal of Statistical Planning and Inference, **2001**, *104*, 121-145.

(9) Witmer, J. A. Bayesian Multistage Decision Problems. The Annals of Statistics, **1986**, *14(1)*, 283-297.