GENERALIZED TWO-STAGE BANDIT PROBLEM

Jack C. Yue

Department of Statistics, National Chengchi Univ., Taipei, Taiwan, R.O.C.

Key Words: Two-armed bandit; Two-stage decision; Optimal strategy

ABSTRACT

Two treatments which yield dichotomous outcomes are available for use in a clinical trial with a two-stage setting. Treatments are chosen sequentially in the first stage, and a single treatment (usually the better one) must be used in the second stage. The goal is to find the optimal strategy which maximizes the expected number of successes. The optimal strategy is to choose the unknown treatment when one treatment is known and when the number of patients is known, or unknown but satisfies certain regular conditions. In this paper, we extend the previous study by assuming that both treatments are unknown but that one treatment is better known (i.e. with smaller prior variance), and explore the conditions in which the better known treatment can be omitted in the first stage.

1. INTRODUCTION

Suppose two treatments, namely Treatment 1 and Treatment 2, are available for use in a clinical trial. Patients arrive at the clinic one at a time,

and only one of the treatments is used on each patient. Information of the effectiveness of the treatments accumulates as the trial continues. The overall objective is to treat as many patients as effectively as possible. This kind of problem is an example of a *Two-armed bandit problem.* Cases with the Bernoulli response of patient, i.e. 0 or 1 when the response is a failure or success, is discussed more frequently. For Bernoulli response, the objective can be treated as to maximize the utility, which is the number of successes. The utility function can also include the sampling cost and discount the successes of the patients – both in the early stages. This kind of utility adjustment can be interpreted as the ethical cost since the patients treated in the early stages have a greater chance of receiving the inferior treatment. Note that although we use the terms treatments and patients in this paper, application of the bandit problem is not restricted to clinical trial. It can be used in other fields, such as psychology, economics, and engineering.

Most of the bandit literature takes the Bayesian approach since Bayes' Theorem provides a mathematical formulation that allows for adaptive learning and for choosing either the potentially early payoff or more informed choices later. This "information vs. immediate payoff" question causes two-armed bandit problems to be surprisingly difficult even when the treatments yield Bernoulli outcomes. Except for some special cases, dynamic programming or backward induction is the standard method for constructing the optimal strategies.

Among all other methods, the *myopic rule* and *stay-with-a-winner rule* are used most frequently, although these two methods do not always guarantee the optimal result. Myopic rule is to choose the arm with greater expected immediate gain at every stage, which is similar to the "one-step look-ahead rule" in sequential decision problems. In the two-armed Bernoulli bandit, Feldman (1962) showed that myopic strategies are optimal when the number

of patients is known and the success probability of the two treatments has a two-point prior distribution. However, myopic strategies are not optimal, or even good, generally. The stay-with-a-winner rule is to use the same arm if the last observation from this arm yielded a success at the last stage. The stay-with-a-winner rule is optimal if two arms are independent and if the arm used (yielding a success) at the last stage is optimal. However, nothing can be said if a failure occurs at the last stage, or the arm used at the last stage is not optimal. Fabius and van Zwet (1970), Berry (1972), and Bradt et. al (1956) also considered the stay-with-a-winner rule. For more details of strategies, see Berry and Fristedt (1985).

Other than sequential selection for every observation, separating the medical trial into several stages is a more realistic way of solving the two-armed bandit problem. Since the data can be collected at intervals throughout the trial, there is no need to know the result of previous patients before giving the next patient treatment, and the calculations can thus be simplified. Canner (1970) discussed the Bernoulli two-armed bandit in a two-stage setting with a sampling cost for the first stage patient, but the same number of patients are assigned to both treatments. Witmer (1986) also discussed the Bernoulli two-armed bandit in a multi-stage setting but with one treatment known. Clayton and Witmer (1988) considered the Bernoulli one-armed bandit in a two-stage setting, and the successes in the first stage are discounted by a factor. Gittins and Wang (1992) showed that if two arms have the same prior mean and only one arm is to be used for all patients, the arm with greater prior variance shall yield greater outcomes. In this study, we will focus on the Bernoulli two-armed bandit in a two-stage setting. But, instead of assuming that one arm is known, we assume that one arm is better known (i.e. has smaller prior variance) than the other arm. We will use the Bayesian approach to explore the influence of

variance on the treatment allocation in the first stage.

2. ASSUMPTIONS

Suppose (θ_1, θ_2) are the success probability of arms 1 and 2, respectively. Let (π_1, π_2) be the prior distribution of (θ_1, θ_2) and in this paper, our focus is mainly on the case in which

$$\pi_1 = Beta(\alpha, \beta) \qquad \text{and} \qquad \pi_2 = Beta(c\alpha, c\beta), \quad \text{where } c \geq 1.$$

That is, arm 2 is better known and so the prior variance of θ_2 is smaller, or

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

while both arms have the same expected immediate payoff, or

$$E(\theta_1|\pi_1) = \frac{\alpha}{\alpha+\beta} = E(\theta_2|\pi_2).$$

Let N be the number of patients and K be the number of patients treated in the first stage. Also, let K_1 and K_2 $(K_1 + K_2 = K)$ be the number of patients receiving treatment 1 and treatment 2, respectively, in the first stage. The treatment allocation of the K first stage patients is predetermined and is nonsequential if N is known. After collecting the information of effectiveness of the treatments from the first stage, the better treatment will be assigned to the remaining $N - K$ patients in the second stage. Our goal is to find the optimal strategy which maximizes the number of successes among these N patients. When the optimal selection of the first stage is not unique, we will choose the one with the smallest K. If several optimal strategies have the same number of K, we will choose the one with smallest K_2. Also, $K^* = (K_1^*, K_2^*)$ is denoted as the optimal numbers of treatment 1 and 2 used in the first stage.

When arm 2 is well-known, i.e. a special case where $c = \infty$ in our setting, Berry and Pearson (1984) showed that $K_2^* = 0$ is the optimal strategy when

N is known, and when $\theta_1 \sim U(0,1) = Beta(1,1)$,

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

Witmer (1986) extended this result to the case where N follows a geometric distribution with mean $1/p$, and if $\theta_1 \sim U(0,1)$,

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

The other extreme is that arms 1 and 2 are equally known, or $c = 1$ in our setting, and Pearson (1980) showed by examples in which $K_1 = K_2$ is never optimal.

3. KNOWN TRIAL LENGTH

When N is fixed and known, the order of using arm 1 and arm 2 in the first stage will not affect the utility. Then, given the strategy τ of using (K_1, K_2) in the first stage, the utility of τ is

$$\begin{aligned} U(\tau) &= U((K_1,K_2),\pi_1,\pi_2(c),N)) \\ &= (K_1+K_2)\cdot\frac{\alpha}{\alpha+\beta} + (N-K_1-K_2)\times \\ &\quad \sum_{i=0}^{K_1}\sum_{j=0}^{K_2} \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{c\alpha+j}{c\alpha+c\beta+K_2}\} P(s_1=i)P(s_2=j) \quad (1) \end{aligned}$$

where s_i is the number of successes using treatment i in K_i first-stage observations and $P(s_i)$ is the marginal probability of s_i for $i = 1, 2$. Also, let $U((K_1,K_2),\pi_1,d,N)$ denote the utility of the strategy in the one-armed bandit problem, where d is the (known) probability of success from arm 2.

Since the one-armed Bernoulli bandit can be treated as a special case in our study, it is natural to expect that $K_2^* \to 0$ as $c \to \infty$. When c is large (or N is small compared to c), taking observations from arm 2 at the first stage can not provide enough evidence to change our belief on θ_2 and there is no

new information gained. Therefore, we will use only arm 1 in the first stage, i.e. $K_2^* = 0$, if c is sufficiently large.

Theorem 3.1 *Given α, β, and N, we will not use the second arm if c is sufficiently large.*

Proof: We will show that the utility of the strategy $(K_1, 0)$ is larger than that of (K_1, K_2) for any $K_1 > 0$ and $K_2 > 0$.

$$
\begin{aligned}
& U((K_1,0),\pi_1,\pi_2(c),N) - U((K_1,K_2),\pi_1,\pi_2(c),N) \\
= \;& K_1 \cdot \frac{\alpha}{\alpha+\beta} + (N-K_1)\cdot \sum_{i=0}^{K_1} \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{\alpha}{\alpha+\beta}\} P(s_1=j) \\
& -(K_1+K_2)\cdot \frac{\alpha}{\alpha+\beta} \\
& -(N-K_1-K_2)\cdot \sum_{i=0}^{K_1}\sum_{j=0}^{K_2} \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{c\alpha+j}{c\alpha+c\beta+K_2}\} P(s_1=i)P(s_2=j) \\
= \;& K_2 \cdot \left[\sum_{i=0}^{K_1} \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{\alpha}{\alpha+\beta}\} P(s_1=i) - \frac{\alpha}{\alpha+\beta} \right] \\
& -(N-K_1-K_2)\times \sum_{i=0}^{K_1} P(s_1=i) \times \\
& \left[\sum_{j=0}^{K_2} \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{c\alpha+j}{c\alpha+c\beta+K_2}\} P(s_2=j) - \max\{\frac{\alpha+i}{\alpha+\beta+K_1}, \frac{\alpha}{\alpha+\beta}\} \right] \\
\equiv \;& \Delta_1 - \Delta_2.
\end{aligned}
$$

Note that

$$\Delta_1 \geq K_2 \cdot \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+K_1)}{\Gamma(\alpha+\beta+K_1)} \cdot \frac{\beta K_1}{(\alpha+\beta)(\alpha+\beta+K_1)} > 0 \quad (2)$$

and

$$\Delta_2 \leq (N-K_1-K_2)\cdot \frac{\beta N}{(\alpha+\beta)(c\alpha+c\beta+N)}, \quad (3)$$

and $\Delta_2 \to 0$ as $c \to \infty$.

Similar argument holds for the case that $K_1 = 0$, and we can show that $U((0,K_2),\pi_1,\pi_2(c),N)) \leq U((K_2,0),\pi_1,\pi_2(c),N))$ if c is sufficiently large.

Thus, we will use only arm 1 in the first stage if c is sufficiently large.□

Given α, β, and N, we can use the ratio of Δ_1 and Δ_2, to find the lower bound $c_0 = c_0(\alpha, \beta, N)$ such that $K_2^* = 0$ when $c \geq c_0$. Similarly, the result in Theorem 3.1 can be extended to the case in which the ratio of prior variances of arm 1 to arm 2 is sufficiently large, since taking observations from arm 2 provides no new information. The proof is similar to that of Theorem 3.1 and is thus omitted.

Corollary 3.2 *Suppose that* $\theta_1 \sim Beta(\alpha_1, \beta_1)$, $\theta_2 \sim Beta(\alpha_2, \beta_2)$, *and* N *is known. Then* $K_2^* = 0$ *if the ratio of prior variances, i.e.* $Var(\theta_1)/Var(\theta_2)$, *is sufficiently large.*

When N, K_1, K_2, α, and β are fixed, the utility of every strategy is a non-increasing function of c. This can be seen from the $\max\{\cdot\}$ term in (1). Also, because the marginal probability function of s_2, i.e.

$$P(s_2 = x) = \begin{pmatrix} K_2 \\ x \end{pmatrix} \frac{\Gamma(c\alpha + c\beta)}{\Gamma(c\alpha)\Gamma(c\beta)} \cdot \frac{\Gamma(c\alpha + x)\Gamma(c\beta + K_2 - x)}{\Gamma(c\alpha + c\beta + K_2)} \text{ for } 0 \leq x \leq K_2,$$

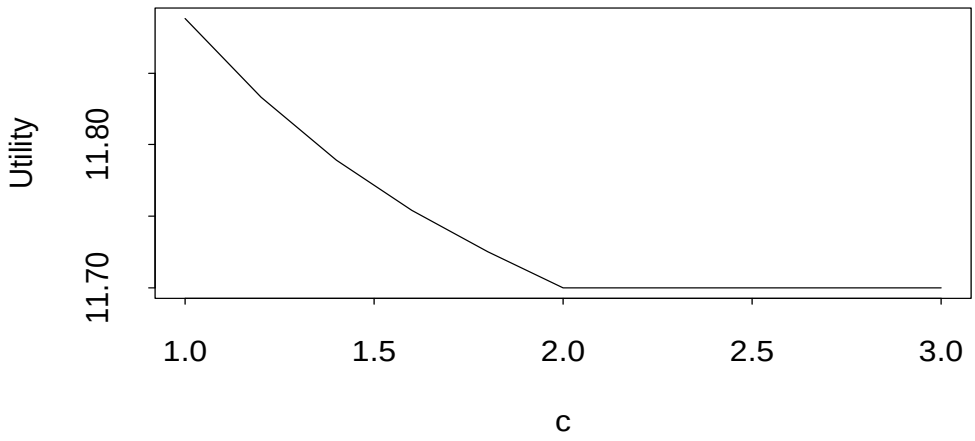
is a continuous function of c, the utility is a continuous function of c as well. Similarly, the utility of the optimal strategy is also a nonincreasing function of c.

Theorem 3.3 *Given* α, β, K_1, K_2, *and* N, $U((K_1, K_2), \pi_1, \pi_2(c), N)$ *is a continuous and non-increasing function of* c.

Corollary 3.4 *Given* α, β *and* N, *the utility of the optimal strategy is a continuous and non-increasing function of* c.

Figure 3.1 is the utility of the optimal rule, given $\alpha = \beta = 1$ and $N = 20$. We can see that the utility is a nonincreasing function of c, as shown in Corollary 3.4. Since $K_2 = 0$ if $c \geq 2$, the utility of the optimal rule is fixed when $c \geq 2$. Similar pattern can be found in other cases of α, β, and N.

Figure 3.1 The utility of the optimal rule, given $\alpha = \beta = 1, N = 20$.



Although Theorem 3.1 guarantees that arm 2 should not be used in the first stage of the optimal strategy if c is large enough, the lowest bound of c is very difficult to find even in the case where $\alpha = \beta$. From Theorem 3.3, we would expect that the lowest bound of c is also a monotonic function of $\alpha = \beta$. However, this is not true, as shown in the following table (although the utility is a decreasing function of $\alpha = \beta$).

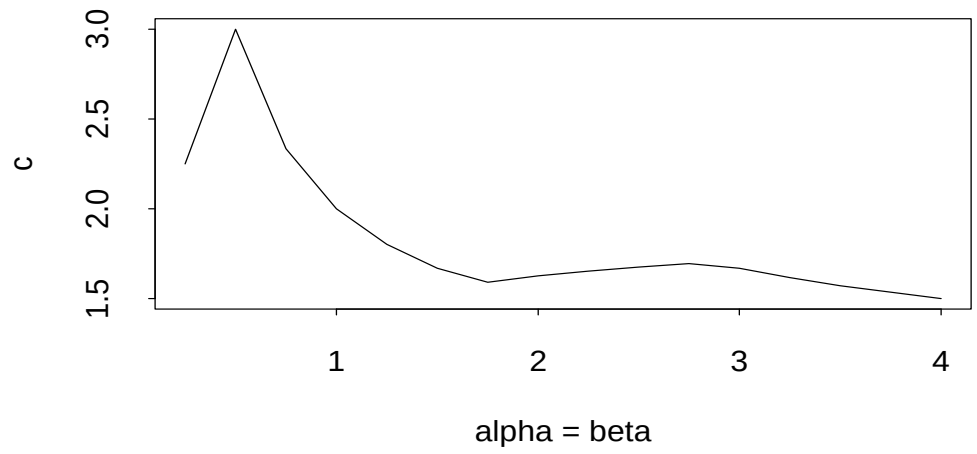
Table 3.1 The Smallest c for which $K_2^* = 0$, given $N = 10$

$\alpha = \beta$	c	K_1	$U((K_1^*, K_2^*))$
1	**1.25**	1	5.75
2	**1.50**	1	5.45
3	1.33	3	5.33
4	1.25	3	5.27
5	1.2	3	5.22

The cases where $N = 20$ and $N = 30$ also show that the lowest bound of c is not necessarily a monotonic function of $\alpha = \beta$. Further, the lowest bound

of c is not an unimodal function of $\alpha = \beta$ either, as shown in Figure 3.2.

Figure 3.2 The smallest c for $K_2 = 0$ vs. $\alpha = \beta$, given $N = 20$



Similar to Theorem 3.3, the utility of the optimal strategy shall be a monotonic function of α and β, if c is fixed. Since increasing α and β at the same rate would have the same effect as if increasing c, the following result thus holds. The proof is similar to that in Theorem 3.3 and is omitted.

Corollary 3.5 *Given $\alpha, \beta(= a\alpha, a > 0), c$ and N, the utility of the optimal strategy is a continuous and non-increasing function of a.*

For the rest of this section, we will focus on the conditions in which arm 2 is not used in the first stage. From Theorem 3.1, it is obvious that arm 2 need not be used if c is large or N is small (and so K^* is small as well). The following are some examples, and the proof can be shown by listing all possible strategies.

Lemma 3.6 *If $N = 2$ or 3, then $(K_1^*, K_2^*) = (1, 0)$.*

Lemma 3.7 *If* $K^* = 1$, *then* $(K_1^*, K_2^*) = (1, 0)$.

When c is small and N is large, $K_2 = 0$ is not necessary to be the optimal selection. For example, $(K_1^*, K_2^*) = (2, 1)$ when $c = 1.2, N = 20$, and $\alpha = \beta = 1$. However, if we can only choose one treatment in the first stage, arm 1 will probably be the choice. This is because both treatments have the same prior mean (same expected immediate payoff) and θ_1 has larger prior variance (larger future outcomes). Note that since only one treatment can be used in the first stage, there is no new information gained on the other treatment. Thus, the decision of choosing the better treatment in the second stage is to compare the posterior mean of the treatment used in the first stage and the prior mean of the unused treatment. Because both arms have the same prior mean, it is equivalent to showing that (in the one-armed bandit setting)

$$U((K, 0), \pi_1, d, N) \geq U((K, 0), \pi_2(c), d, N) \tag{4}$$

with $d = \alpha/(\alpha + \beta)$, if we want to show that arm 1 will be used in the first stage.

We need the following lemmas to prove (4):

Lemma 3.8 *Given* α, β, K, *and* $c > 1$,
(i) $P(s_1 = 0) \geq P(s_2 = 0)$ *and* $P(s_1 = K) \geq P(s_2 = K)$, *where the equality holds only at* $K = 1$, *and*
(ii)$P(s_2 = x)/P(s_1 = x)$ *is an increasing function of* x *if* $x \leq [\frac{\alpha(K-1)}{\alpha+\beta}]$ *and a decreasing function of* x *if* $x \geq [\frac{\alpha(K-1)}{\alpha+\beta}]$.

Proof: The proof can be shown by plugging into the probability functions of s_1 and s_2.□

Based on Lemma 3.8, we can show the following result:

Lemma 3.9 *For every* α, β, K, *and* $c > 1$, *there exists* $x_0 = x_0(\alpha, \beta, K, c)$, *such that*

$$P(s_1 \leq x) \geq P(s_2 \leq x), \text{ if } x \leq x_0$$

and

$$P(s_1 \geq x) \geq P(s_2 \geq x), \text{ if } x \geq x_0.$$

Lemma 3.10 *For every α, β, K, and $c > 1$,*

$$\frac{\alpha + x}{\alpha + \beta + K} \leq \frac{c\alpha + x}{c\alpha + c\beta + K}, \quad \text{if } x \leq [\frac{\alpha}{\alpha + \beta} K]$$

and

$$\frac{\alpha + x}{\alpha + \beta + K} \geq \frac{c\alpha + x}{c\alpha + c\beta + K}, \quad \text{if } x \geq [\frac{\alpha}{\alpha + \beta} K].$$

Theorem 3.11 *For $c \geq 1$, $K \geq 0$, and $0 \leq d \leq 1$,*

$$U((K,0), \pi_1, d, N) \geq U((K,0), \pi_2(c), d, N).$$

Proof: Define

$$g(c,d) = \sum_{i=1}^{K} \max\{\frac{c\alpha + i}{c\alpha + c\beta + K}, d\} P(s_2 = i). \tag{5}$$

Thus, it is equivalent to showing that $g(1,d) \geq g(c,d)$ for $c \geq 1$. It is trivial that

$$g(c,d) = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } 0 \leq d \leq \frac{c\alpha}{c\alpha+c\beta+K} \\ d & \text{if } \frac{c\alpha+K}{c\alpha+c\beta+K} \leq d \leq 1 \end{cases}$$

and by Lemma 3.8,

$$g(1,d) > g((c,d) \text{ if } \begin{cases} \frac{\alpha}{\alpha+\beta+K} < d < \frac{c\alpha}{c\alpha+c\beta+K} \\ \frac{c\alpha+K}{c\alpha+c\beta+K} < d < \frac{\alpha+K}{\alpha+\beta+K}. \end{cases}$$

From Lemma 3.9, without loss of generality, suppose that $x_0 \leq [\frac{\alpha}{\alpha+\beta} K]$. Then by Lemma 3.10, $g(1,d) \geq g(c,d)$ if $0 \leq d \leq x_0/K$ and $[\frac{\alpha}{\alpha+\beta} K]/K \leq d \leq 1$. Because $g(c,d)$ is a continuous function of d, it follows that $g(1,d) \geq g(c,d)$ for $0 \leq d \leq 1$. The case where $x_0 \geq [\frac{\alpha}{\alpha+\beta} K]$ is similar.□

Using $\alpha/(\alpha+\beta)$ to replace d in Theorem 3.11, we have

Corollary 3.12 *For $c \geq 1$ and $K \geq 0$,*

$$U((K,0), \pi_1, \pi_2(c), N) \geq U((0,K), \pi_1, \pi_2(c), N).$$

The result in Corollary 3.12 indicates that we shall choose the arm with larger variance (i.e. the less certain arm) in the first stage, if two arms have the same expected immediate payoff and only one arm is allowed to be used. This is similar to that of Wang and Gittins (1992), but in a two-stage setting. This result can be directly extended to the case where the less certain arm has a larger expected immediate payoff, i.e. to the case where $E(\theta_1) \geq E(\theta_2)$ and $Var(\theta_1) \geq Var(\theta_2)$.

4. UNKNOWN TRIAL LENGTH

In this section we consider the case in which N is unknown, and in particular, we will focus on the case that N follows a geometric distribution. In the known trial length case, taking observations from the (second) arm with small prior variance gives little information and thus $K_2^* = 0$. This should also be true in the case where $E(\theta_1) \geq E(\theta_2)$, $Var(\theta_1) \geq Var(\theta_2)$, and N follows a geometric distribution, since arm 2 does not provide a larger expected immediate payoff or larger expected outcomes in the future. The proof is similar to that in Theorem 3.1 and is omitted.

Theorem 4.1 *Let* $\theta_1 \sim Beta(\alpha_1, \beta_1)$, $\theta_2 \sim Beta(\alpha_2, \beta_2)$, *and* $N \sim Geom(p)$. *Then* $K_2^* = 0$ *if* $E(\theta_1) \geq E(\theta_2)$ *and* $Var(\theta_1)/Var(\theta_2)$ *is sufficiently large.*

The result in Theorem 4.1 is not necessarily true if $E(\theta_1) < E(\theta_2)$ since choosing arm 2 would give a larger expected immediate payoff. However, because arm 2 yields a larger expected immediate payoff, we should use arm 2 before arm 1 in the first stage if $K_2 \neq 0$. Let τ_i denote the treatment used on

the ith patient for strategy τ, i.e. $\tau_i = 1$ or 2 for all i, then

Theorem 4.2 *Given α, β and $N \sim Geom(p)$, there is an optimal strategy of the form: $\tau_1 = \cdots = \tau_{K_2} = 2$, $\tau_{K_2+1} = \cdots = \tau_{K_1+K_2} = 1$.*

Then based on this result, we can also show that:

Corollary 4.3 *Let $\theta_1 \sim Beta(\alpha_1, \beta_1)$, $\theta_2 \sim Beta(\alpha_2, \beta_2)$, and $N \sim Geom(p)$. Then $K_2^* = 0$ if $Var(\theta_1)/Var(\theta_2)$ is sufficiently large.*

5. DISCUSSIONS

In this paper, we showed that the one-armed bandit is a special case in our study. When we know the performance of a treatment sufficiently well, there is no need to experiment on this treatment in the first stage, i.e. $K_2^* = 0$ if c is sufficiently large and N is bounded, or is unbounded and from a geometric distribution. Given $K_2^* = 0$, most of the results in the one-armed bandit, such as the formula of K_1^* in Berry and Pearson (1984) and Witmer (1986) stated in Section 2, are also true in our study.

When $K_2 = 0$ is not the optimal strategy, we need to decide what values of K_1 and K_2 would yield the maximal utility. Following the same interpretation of Corollary 3.12, we would expect that the strategies with $K_2 > K_1$ are dominated by other strategies with $K_1 > K_2$, and in our calculation we found that $U((K_2, K_1)) \geq U((K_1, K_2))$ is always true given $K_2 > K_1$. The proof of this result would be more complicated than that in Theorem 3.11, since the posterior mean of either arm in the $\max\{\cdot\}$ of (1) is changing at the same time. Then, based on the result, we only need to consider the strategies with the form $K_1 > K_2$, and this can halve the work in finding the optimal strategy.

We also found that $K_1 = K_2$ is not optimal in our calculation as well.

This is similar to the conjecture by Berry (1972), where Berry conjectured that $K_1 = K_2$ is never optimal in the Bernoulli two-armed bandit. (This conjecture has not yet been proven.) We found that $U((K_1, K_1))$ is smaller than $U((K_1, K_1 - 1))$ if $K_1 \geq 2$, or smaller than $U((K_1 + 1, K_1))$ if $K_1 \leq 1$. This calculation result might be helpful in proving that $K_1 = K_2$ is never optimal.

BIBLIOGRAPHY

Berry, D. A. (1972) A Bernoulli Two-armed Bandit. *The Annals of Mathematical Statistics*, Vol. 43, No. 3, 871-897.

Berry, D. A. and Fristedt, B. (1979) Bernoulli One-armed Bandits – Arbitrary Discount Sequences. *The Annals of Statistics*, Vol. 7, No. 5, 1086-1105.

Berry, D. A. and Pearson, L. (1985) Optimal Designs for Two-stage Clinical Trials with Dichotomous Responses. *Statistics in Medicine*, Vol. 4, 497-508.

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

Canner, P. L. (1970) Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association*, 65, 293-306.

Clayton, M. K. and Witmer, J. A. (1988) Two-stage Bandits. *The Annals of Statistics*, Vol. 16, No. 2, 887-894.

Fabius, J. and van Zwet, W. R. (1970) Some Remarks on the Two-armed bandit. *Annals of Mathematical Statistics*, Vol. 41, 1906-1916.

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

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

Gittins, J. C. and Wang, Y. (1992) The Learning Component of Dynamic Allocation Indices. *The Annals of Statistics*, 20, 1625-1636.

Joshi, V. M. (1975) A Conjecture of Berry Regarding a Bernoulli Two-armed Bandit. *The Annals of Statistics*, Vol. 3, NO. 1 189-202.

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

Simons, G. (1986) Bayes Rules for a Clinical Trials Model with Dichotomous Responses. *The Annals of Statistics*, Vol. 14, No. 3, 954-970.

Witmer, J. A. (1983) *Bayesian Multistage Decision Problems.* Ph.D. Thesis, University of Minnesota, USA.

Witmer, J. A. (1986) Bayesian Multistage Decision Problems. *The Annals of Statistics*, Vol. 14, No. 1, 283-297.