Finite-Time Lower Bounds for the Two-Armed Bandit Problem*

Sanjeev R. Kulkarni
Department of Electrical Engineering
Princeton University, Princeton, N.J., 08544
kulkarni@ee.princeton.edu

Gábor Lugosi
Department of Economics,
Pompeu Fabra University,
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain.
lugosi@upf.es

Submitted: Oct.24, 1997

Revised: May 28, 1999

Abstract

We obtain minimax lower bounds on the regret for the classical two-armed bandit problem. We provide a finite-sample minimax version of the well-known $\log n$ asymptotic lower bound of Lai and Robbins. The finite-time lower lower bound allows us to derive conditions for the amount of time necessary to make any significant gain over a random guessing strategy. These bounds depend on the class of possible distributions of the rewards associated with the arms. For example, in contrast to the $\log n$ asymptotic results on the regret, we show that the minimax regret is achieved by mere random guessing under fairly mild conditions on the set of allowable configurations of the two arms. That is, we show that for *every* allocation rule and for *every* n, there is a configuration such that the regret at time n is at least $1-\epsilon$ times the regret of random guessing, where ϵ is any small positive constant.

Key Words: two-armed, bandit, learning, estimation, minimax, regret, lower bound

*This work was supported in part by the National Science Foundation under NYI grant IRI-9457645. G. Lugosi acknowledges the support of DIGES Grant PB96-0300.

1 Introduction

Bandit problems have received considerable interest (e.g., see [5, 6]) originating from the work in [8, 9]. In the classical two-armed bandit problem, there are two unknown distributions P_1 and P_2 associated with arm 1 and arm 2, respectively. At each time we are allowed to select an arm from which to receive a reward drawn independently according to the distribution for that arm. Our goal is to maximize the expected sum of the rewards. Let m_1 and m_2 denote the expected values corresponding to P_1 and P_2, respectively. If we knew which one of m_1 and m_2 is larger, we could keep selecting the arm with larger mean, and after time n, our expected reward would be $n \max(m_1, m_2)$. Since the distributions P_1 and P_2 are unknown, the expected reward will always be smaller than this optimal value. The difference between $n \max(m_1, m_2)$ and the expected reward is called the regret. Note that if, in each step, we select an arm independently with equal probabilities, the regret is $n\Delta/2$, where $\Delta = |m_1 - m_2|$. The results of Lai and Robbins [7] and subsequent extensions by others (e.g., [1, 2, 3, 4]) showed that in a fairly strong asymptotic sense the optimum achievable regret is $\Delta \log n/\tilde{I}$, where $\tilde{I}$ is the Kullback-Leibler divergence between P_1 and P_2.

In this paper, we consider the problem from a non-asymptotic minimax perspective. In the minimax setup, one assumes that the possible configurations (P_1, P_2) belong to a fixed family of pairs of distributions. Then one is interested in the smallest possible regret achievable by any strategy of selecting the arms. Here we obtain lower bounds on the regret of such optimal strategies. Our starting point is a class of *two* configurations, and we show that even if the player knows this class, he/she cannot perform better than the lower bounds obtained here.

First we offer a finite-sample minimax version of the Lai-Robbins lower bound (see Theorem 1 below). This result can be used to provide bounds on the sample size necessary to guarantee a desired performance. Also, in sharp contrast to the well-known $\log n$ asymptotic results on the regret, we show that for small sample sizes the wrong arm will be pulled a constant fraction of times. In particular, the minimax regret is about $n\Delta/2$ under fairly mild conditions on the set of allowable configurations of the two arms. We show that if the set of allowable configurations is sufficiently "large," then for any n, for any small ϵ, and for any strategy of selecting arms, there is a configuration such that the regret is larger than $(1-\epsilon)n\Delta/2$. In other words, regardless of how large n is, up to time n, the "bad" arm will be played almost half of the time for some configuration. That is, in the minimax sense, no arm-selection strategy can perform better than completely random selections.

2 Formulation and Lower Bounds

Let a configuration $\Theta = (\theta_1, \theta_2)$ be a pair of parameters determining the distributions of the two arms. That is, if arm i is pulled, a reward is payed independently according to the probability density f_{θ_i} $(i = 1, 2)$. All densities are understood with respect to a common dominating σ-finite measure λ on the real line. Denote the respective means by

$$m_i = \int x f_{\theta_i}(x) d\lambda(x), \qquad i = 1, 2,$$

and let $\Delta = |m_1 - m_2|$. Assume without loss of generality that $m_1 > m_2$, that is, arm 1 is optimal. We denote the measure and expectation with respect to the distribution associated with θ by P_θ and E_θ, respectively.

Introduce the alternative configuration $\Theta' = (\theta_1', \theta_2)$, for some θ_1' such that $m_1' = \int x f_{\theta_1'}(x) d\lambda(x) = m_2 - \Delta$. Thus, the distribution of the reward after pulling arm 2 is unchanged, but arm 2 is optimal in configuration Θ'.

Our bounds involve the *information divergence* (or Kullback-Leibler number) between the densities f_{θ_1} and $f_{\theta_1'}$ given by

$$I = I(\theta_1', \theta_1) = \int f_{\theta_1'}(x) \log\left(\frac{f_{\theta_1'}(x)}{f_{\theta_1}(x)}\right) d\lambda(x) = E_{\theta'}\left\{\log\left(\frac{f_{\theta_1'}(x)}{f_{\theta_1}(x)}\right)\right\},$$

as well as a variance-like quantity, denoted V, related to the information divergence I by

$$V = \int f_{\theta_1'}(x) \log^2\left(\frac{f_{\theta_1'}(x)}{f_{\theta_1}(x)}\right) d\lambda(x) - I^2 = E_{\theta'}\left\{\log^2\left(\frac{f_{\theta_1'}(x)}{f_{\theta_1}(x)}\right)\right\} - I^2.$$

(All logarithms are of the natural base.)

An adaptive allocation rule $\Phi = (\phi_1, \phi_2, \ldots)$ is a sequence of random variables taking values in $\{1, 2\}$ such that ϕ_i is measurable with respect to the σ-field $\mathcal{F}_{i-1}$ generated by the previous values $\phi_1, X_1, \ldots, \phi_{i-1}, X_{i-1}$, where $X_1, X_2, \ldots$ are the random variables denoting the sequence of rewards obtained. That is, based on the previous rewards $(X_1, \ldots, X_{i-1})$ and the previous selections $(\phi_1, \ldots, \phi_{i-1})$, ϕ_i denotes whether arm 1 or arm 2 is to be pulled at time i. Under a particular adaptive allocation rule and configuration Θ, our reward up to and including time n

is

$$S_n = \sum_{i=1}^{n} X_i.$$

Since $E[X_i|\mathcal{F}_{i-1}] = m_{\phi_i}$, the expected reward is

$$E[S_n] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} E\left(E[X_i|\mathcal{F}_{i-1}]\right) = E\left[\sum_{i=1}^{n} m_{\phi_i}\right].$$

and the regret at time n is

$$R_n(\Theta) = n\max(m_1, m_2) - E[S_n].$$

In other words, $R_n(\Theta)$ is Δ times the expected number of times the arm with worse expected payoff is pulled.

Our goal is to obtain lower bounds on the minimum value of

$$\max(R_n(\Theta), R_n(\Theta')),$$

over all possible adaptive allocation rules. For any integer $n > 0$ and $\alpha > 0$, introduce

$$\Lambda_{\alpha,n} = e^{-\alpha I}\left(1 - \max_{1\leq i\leq n} p_{\alpha,i}\right),$$

where

$$p_{\alpha,n} = P_{\theta'}\left\{\sum_{i=1}^{n} \log\left(\frac{f_{\theta_1}(x_i)}{f_{\theta_1'}(x_i)}\right) \leq -\alpha I\right\}.$$

Theorem 1 *For any n, $a_n \in (0,1)$, $c_n \in (0,n)$, $\alpha > c_n$, and for any adaptive allocation rule,*

$$\max(R_n(\Theta), R_n(\Theta')) \geq \Delta \min\left(c_n(1-a_n), (n-c_n)a_n\Lambda_{\alpha,c_n}\right).$$

To interpret the theorem, we need lower bounds for $\Lambda_{\alpha,n}$, that is, upper bounds for $p_{\alpha,n}$. Note that $p_{\alpha,n}$ is the probability that the sum of n i.i.d. random variables (with negative mean $-I$) is less than the mean of the sum minus $(\alpha - n)I$. Thus, it follows by Chebyshev's inequality that

$$p_{\alpha,i} \leq \frac{iV}{(\alpha-i)^2 I^2},$$

and therefore, since the right-hand side is a monotonically increasing function of i, we have

$$\Lambda_{\alpha,n} \geq e^{-\alpha I}\left(1 - \frac{nV}{(\alpha-n)^2 I^2}\right). \tag{1}$$

We will see that (1) is a satisfactory bound if the ratio V/I is not too large. This is indeed the case for many interesting cases. The next two examples serve as illustration.

Example. Let f_{θ_1} be the normal density with mean m_1 and variance σ^2, and let $f_{\theta_1'}$ be the normal density with mean m_1' and variance σ^2. Then straightforward calculation shows that $V = I$ for all values of m_1, m_1', and σ.

Example. Let θ_1 correspond to the Bernoulli distribution $P_{\theta_1}(\{0\}) = p$, $P_{\theta_1}(\{1\}) = 1-p$, and let θ_1' be defined by $P_{\theta_1'}(\{0\}) = 1-p$, $P_{\theta_1'}(\{1\}) = p$ and assume that $p > 1/2$. Then using the inequality $\log x \le x - 1$,

$$\frac{V}{I} = \frac{(1-(2p-1)^2)\log^2(p/(1-p)}{(2p-1)\log(p/(1-p)} \le 4p \le 4.$$

Note that in this example we can take f_{θ_2} to be any density with mean $m_2 = 1/2$, for example, we may let $P_{\theta_2}(\{1/2\}) = 1$.

In specific situations, one may get much sharper estimates. For example, if both f_{θ_1} and $f_{\theta_1'}$ are Gaussian with variance σ^2, then $\log(f_{\theta_1}(X)/f_{\theta_1'}(X))$ also has a Gaussian distribution, so one may get sharper estimates for $p_{\alpha,n}$ by using standard bounds for the tail of a Gaussian distribution, but we do not detail these, rather straightforward, bounds here.

Corollary 1 *Fix any $\epsilon \in (0,1)$. If n is so large that*

$$n^{\epsilon^2} \ge \max\left(4\frac{(1-\epsilon)^2}{\epsilon}\frac{\log n}{I}, e^{2V/(I(1-\epsilon))}\right),$$

then

$$\max(R_n(\Theta), R_n(\Theta')) \ge \Delta\left\lfloor(1-\epsilon)^2\frac{\log n}{I}\right\rfloor.$$

Proof. In Theorem 1 take $a_n = \epsilon$, $c_n = \lfloor(1-\epsilon)\log n/I\rfloor$, and $\alpha = (1+\epsilon)c_n$. Then (1) and a straightforward calculation shows that

$$c_n(1-a_n) \le (n-c_n)a_n\Lambda_{\alpha,c_n}$$

whenever the condition for n is satisfied, and therefore $\max(R_n(\Theta), R_n(\Theta')) \ge c_n(1-a_n)$. $\square$

Remark. The classical lower bound of Lai and Robbins [7] states that for any allocation rule, the regret $R_n(\Theta)$ is asymptotically not smaller than $(\Delta/\tilde{I})\log n$, where $\tilde{I}$ is the Kullback-Leibler divergence between P_1 and P_2. The main novelty of Corollary 1 is in its nonasymptotic nature. Note however, that the new bound is for $\max(R_n(\Theta), R_n(\Theta'))$ instead of just $R_n(\Theta)$, and the divergence I appearing in the bound is different from $\tilde{I}$. However, it is easy to see that by adding extra regularity conditions similar to those of Lai and Robbins, the bound of Corollary 1 may easily be converted into a nonasymptotic analogue of the Lai-Robbins lower bound.

For smaller values of n, Theorem 1 may be used to derive much larger lower bounds:

Corollary 2 *Let $a \in (0,1)$. If $n^{1-a} \geq 4e$ and $n^a I + \sqrt{n^a V/2} \leq 1$ then*

$$\max(R_n(\Theta), R_n(\Theta')) \geq \Delta \frac{\lfloor n^a \rfloor}{2}.$$

Proof. Take $c_n = \lfloor n^a \rfloor$, $a_n = 1/2$, and $\alpha = n + \sqrt{nV/(2I^2)}$ in Theorem 1. □

Taking $c_n = n/2$ and $a_n = 1/2$ in Theorem 1 we obtain the lower bound $(n\Delta/4)\Lambda_{\alpha,n/2}$. The following theorem improves on this:

Theorem 2 *For any n, $\alpha > n$, and for any adaptive allocation rule,*

$$\max(R_n(\Theta), R_n(\Theta')) \geq \frac{n\Delta}{2} \Lambda_{\alpha,n},$$

The next corollary points out that for sample sizes of the order of $1/I$, the number of times the bad armed is pulled is linear. This means that for sample sizes smaller than this, it is basically impossible to learn which is the best arm.

Corollary 3 *Let c be a positive constant. For any $n \leq c/I$ and for any adaptive allocation rule,*

$$\max(R_n(\Theta), R_n(\Theta')) \geq \frac{\Delta n}{4} e^{-c-\sqrt{cV/(2I)}}.$$

Proof. Note that if we take $\alpha = n + \sqrt{nV/(2I^2)}$,

$$1 - \frac{nV}{(\alpha - n)^2 I^2} = \frac{1}{2},$$

and therefore, the corollary follows by applying Theorem 2 with (1) for $n \leq c/I$. □

Corollary 4 *Let $\epsilon > 0$ be arbitrary. If there exists an $\alpha > n$ such that $\alpha \leq -\log(1-\epsilon)/2I$ and $(\alpha - n)^2/n \geq 2V/(\epsilon I^2)$, then*

$$\max(R_n(\Theta), R_n(\Theta')) \geq \frac{\Delta n}{2}(1 - \epsilon).$$

Proof. Straightforward calculation shows that if the conditions are satisfied then $e^{-\alpha I} \geq \sqrt{1-\epsilon}$ and $1 - nV/\left((n-\alpha)^2 I^2\right) \geq \sqrt{1-\epsilon}$, so the statement follows by Theorem 2 and the bounds of (1). □

Corollary 4 may be applied with arbitrary ϵ in many cases when, in the class of allowable configurations, there are pairs (θ_1, θ_1') with arbitrarily small information divergence. The following two special cases illustrate such situations.

Corollary 5 *Suppose P_2 is an arbitrary distribution with mean zero (which can even be known). Suppose P_1 is Gaussian with mean either Δ or $-\Delta$ with arbitrary variance. Then for* every *adaptive allocation rule, for every $\epsilon > 0$, and for every n, there is a configuration such that the regret at time n is at least* $(1-\epsilon)n\Delta/2$.

Corollary 6 *Let S be the set of all configurations $\Theta = (\theta_1, \theta_2)$ such that both P_{θ_1} and P_{θ_2} are Bernoulli distributions (i.e., of the form $P_\theta(\{0\}) = p$, $P_\theta(\{1\}) = 1-p$ for some p). Then for every adaptive allocation rule Φ, for every $\epsilon > 0$, and for every n,*

$$\sup_{\Theta \in S} R_n(\Theta) \geq \frac{n\Delta}{2}(1-\epsilon),$$

where Δ is the difference between the means corresponding to the two arms.

3 A Change-of-Measure Lemma

As before, let the vector $\mathbf{X} = (X_1, \ldots, X_n)$ of random variables denote the rewards up to time n if a particular adaptive allocation rule is used. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{R}^n$ denote a fixed realization of $\mathbf{X}$.

$T_{\mathbf{x}}(1)$ and $T_{\mathbf{x}}(2)$ denote the number of times arm 1, and arm 2 are pulled up to time n.

The key part of the proofs of the results in the previous section is the following measure-transformation lemma, which is based on ideas of Lai and Robbins [7].

Lemma 1 *For any integer $k \in [0, n]$, and $\alpha > 0$,*

$$P_\Theta\{T_{\mathbf{x}}(1) = k\} \geq e^{-\alpha I}(1 - p_{\alpha,k}) P_{\Theta'}\{T_{\mathbf{x}}(1) = k\}.$$

Proof. Let $J \subset \{1, \ldots, n\}$ be a set of indices. On J, introduce the likelihood ratio

$$L_J(\mathrm{x}) = \sum_{j \in J} \log \left(\frac{f_{\theta_1}(x_j)}{f_{\theta_1'}(x_j)} \right).$$

The first step of the proof is trivial:

$$P_\Theta\{T_{\mathrm{x}}(1) = k\} \geq P_\Theta\{T_{\mathrm{x}}(1) = k,\ L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I\}, \tag{2}$$

where $B(\mathrm{x})$ is the set of indices indicating the times when arm 1 is pulled by the allocation rule based on the sequence of observations x.

For each index set J, define $A_J \subset \mathcal{R}^n$ by

$$A_J = \left\{ \mathrm{x} : B(\mathrm{x}) = J, L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I \right\}.$$

Thus,

$$P_\Theta\{T_{\mathrm{x}}(1) = k, L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I\} = P_\Theta \left\{ \bigcup_{J:|J|=k} A_J \right\} = \sum_{J:|J|=k} P_\Theta\{A_J\}.$$

Now

$$\begin{aligned} P_\Theta\{A_J\} &= \int_{A_J} \left(\prod_{j \in J} f_{\theta_1}(x_j) \right) \left(\prod_{j \notin J} f_{\theta_2}(x_j) \right) d\lambda(x_1) \cdots d\lambda(x_n) \\ &= \int_{A_J} \left(\prod_{j \in J} \frac{f_{\theta_1}(x_j)}{f_{\theta_1'}(x_j)} \right) \left(\prod_{j \in J} f_{\theta_1'}(x_j) \right) \left(\prod_{j \notin J} f_{\theta_2}(x_j) \right) d\lambda(x_1) \cdots d\lambda(x_n). \end{aligned}$$

But for each $\mathrm{x} \in A_J$, we have $L_J(\mathrm{x}) > -\alpha I$, so

$$\left(\prod_{j \in J} \frac{f_{\theta_1}(x_j)}{f_{\theta_1'}(x_j)} \right) > e^{-\alpha I},$$

and therefore

$$P_\Theta\{A_J\} \geq e^{-\alpha I} P_{\Theta'}\{A_J\}.$$

It follows that

$$\begin{aligned} & P_\Theta\{T_{\mathrm{x}}(1) = k,\ L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I\} \\ & \geq\ e^{-\alpha I}\, P_{\Theta'}\{T_{\mathrm{x}}(1) = k,\ L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I\} \\ & =\ e^{-\alpha I}\, P_{\Theta'}\{T_{\mathrm{x}}(1) = k\}\, P_{\Theta'}\{L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I | T_{\mathrm{x}}(1) = k\}. \end{aligned} \tag{3}$$

But by the definition of $p_{\alpha,k}$, we have

$$P_{\Theta'}\{L_{B(\mathrm{x})}(\mathrm{x}) > -\alpha I | T_{\mathrm{x}}(1) = k\} \geq 1 - p_{\alpha,k}. \tag{4}$$

Summarizing (2), (3), and (4), the proof of the lemma is complete. □

4 Proofs of Theorems 1 and 2

Proof of Theorem 1. There are two cases. If $P_{\Theta'}\{T_x(1) \leq c_n\} < a_n$, then by Markov's inequality, we have

$$E_{\Theta'}[T_x(1)] \geq c_n(1 - a_n).$$

If, on the other hand, $P_{\Theta'}\{T_x(1) \leq c_n\} \geq a_n$, then

$$\begin{aligned} P_\Theta\{T_x(1) \leq c_n\} &= \sum_{k=1}^{c_n} P_\Theta\{T_x(1) = k\} \\ &\geq \sum_{k=1}^{c_n} P_{\Theta'}\{T_x(1) = k\} e^{-\alpha I}(1 - p_{\alpha,k}) \quad \text{(by Lemma 1, whenever } \alpha > c_n\text{)} \\ &\geq \Lambda_{\alpha,c_n} P_{\Theta'}\{T_x(1) \leq c_n\} \\ &\geq a_n \Lambda_{\alpha,c_n}. \end{aligned}$$

Thus,

$$P_\Theta\{T_x(2) > n - c_n\} = P_\Theta\{T_x(1) \leq c_n\} \geq a_n \Lambda_{\alpha,c_n},$$

and by Markov's inequality,

$$E_\Theta[T_x(2)] \geq (n - c_n) a_n \Lambda_{\alpha,c_n}.$$

Therefore,

$$\begin{aligned} \max(R_n(\Theta), R_n(\Theta')) &= \Delta \max\left(E_\Theta[T_x(2)], E_{\Theta'}[T_x(1)]\right) \\ &\geq \Delta \min\left(c_n(1 - a_n), (n - c_n) a_n \Lambda_{\alpha,c_n}\right), \end{aligned}$$

and the theorem is proved. □

Proof of Theorem 2. After time n, the regret under configuration Θ is

$$R_n(\Theta) = \Delta E_\Theta[T_{\mathbf{x}}(2)].$$

For any $\alpha > n$, we have

$$\begin{aligned} &\max(R_n(\Theta), R_n(\Theta')) \\ &\quad \geq \frac{R_n(\Theta) + R_n(\Theta')}{2} \\ &\quad = \Delta \frac{E_\Theta[T_{\mathbf{x}}(2)] + E_{\Theta'}[T_{\mathbf{x}}(1)]}{2} \\ &\quad = \frac{\Delta}{2} \sum_{i=1}^{n} \left(P_\Theta\{T_{\mathbf{x}}(2) \geq i\} + P_{\Theta'}\{T_{\mathbf{x}}(1) \geq i\}\right) \\ &\quad \geq \frac{\Delta}{2} \sum_{i=1}^{n} \left(P_\Theta\{T_{\mathbf{x}}(2) \geq i\} + P_\Theta\{T_{\mathbf{x}}(1) \geq i\}\right) e^{-\alpha I}(1 - p_{\alpha,k}) \quad \text{(by Lemma 1)} \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\Delta}{2}\Lambda_{\alpha,n}\sum_{i=1}^{n}\left(P_{\Theta}\{T_{\mathbf{x}}(2)\geq i\}+P_{\Theta}\{T_{\mathbf{x}}(1)\geq i\}\right)\\
&= \frac{\Delta}{2}\Lambda_{\alpha,n}\left(E_{\Theta}[T_{\mathbf{x}}(2)]+E_{\Theta}[T_{\mathbf{x}}(1)]\right)\\
&= \frac{n\Delta}{2}\Lambda_{\alpha,n},
\end{aligned}
$$

and the proof is complete.

References

[1] R. Agrawal, M. Hedge, D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," *IEEE Trans. Automatic Control*, Vol. AC-33, No. 10, pp. 899-906, 1988.

[2] R. Agrawal, D. Teneketzis, V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: finite parameter space," *IEEE Trans. Automatic Control*, Vol. AC-34, No. 3, pp. 258-266, 1989.

[3] V. Anantharam, P. Varaiya, J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays — part I: i.i.d. rewards," *IEEE Trans. Automatic Control*, Vol. AC-32, No. 11, pp. 968-976, 1987.

[4] V. Anantharam, P. Varaiya, J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays — part II: Markovian rewards," *IEEE Trans. Automatic Control*, Vol. AC-32, No. 11, pp. 977-982, 1987.

[5] D.A. Berry and B. Fristedt, *Bandit Problems*, Chapman and Hall, 1985.

[6] J.C. Gittins, *Multi-armed Bandit Allocation Indices*, Wiley & Sons, 1989.

[7] T.L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, Vol. 6, pp. 4-22, 1985.

[8] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, Vol. 58, pp. 527-535, 1952.

[9] W.R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, Vol. 25, pp. 275-294, 1933.