AN ASYMPTOTIC MINIMAX THEOREM FOR THE TWO ARMED BANDIT PROBLEM

By Walter Vogel[1]

University of Chicago and Universität Tübingen[2]

1. Introduction. Let Ex I and Ex II be two experiments, the outcomes of which are described by the two random variables X and Y. Let $P(X = 1) = p = 1 - P(X = 0)$, $P(Y = 1) = q = 1 - P(Y = 0)$ and $0 < p, \; q < 1$. An experimenter has to do n experiments, one after another, and at every step he may choose between Ex I or Ex II. He does not know the values of p and q and he wants to maximize the sum of all outcomes. Therefore he will choose a strategy, i.e. a procedure which tells him which experiment to use at the kth step as a function of his previous choices and the previous outcomes of the experiments. The question how to find a suitable strategy is known as the problem of the two armed bandit. For approaches other than the one used in this paper see [1], [2], [4], [5] and [6].

We will measure the value of a strategy by a loss function. Let Π_k be the unconditional probability of choosing Ex I at the kth step. The expected value of the performed experiment at the kth step will be

$$\Pi_k p + (1 - \Pi_k)q = p - (p - q)(1 - \Pi_k) = q - (q - p)\Pi_k \, .$$

We define as loss at the kth step:

$$\max\,(p, q) - (\Pi_k p + (1 - \Pi_k)q) = (p - q)(1 - \Pi_k)$$

if $p \geqq q$ or $(q - p)\Pi_k$ if $p \leqq q$. The loss $L(p, q)$ for the whole game is then

$$(p - q) \sum_{k=1}^{n} (1 - \Pi_k) \quad \text{or} \quad (q - p) \sum_{k=1}^{n} \Pi_k \, .$$

In $L(p, q)$ and $\Pi_k(p, q)$ the first argument is always related to Ex I. Let $\sigma = \max\,(p, q)$ and $\tau = \min\,(p, q)$; then

$$L(\sigma, \tau) = (\sigma - \tau) \sum_{k=1}^{n} (1 - \Pi_k(\sigma, \tau))$$

and

$$L(\tau, \sigma) = (\sigma - \tau) \sum_{k=1}^{n} \Pi_k(\tau, \sigma).$$

As we do not suppose any previous knowledge about p and q, it seems natural

Received April 2, 1959; revised January 18, 1960.

[1] Research carried out in part at the Statistical Research Center, University of Chicago, under the sponsorship of the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] Present address: Mathematisches Institut der Universität Tübingen, Germany.

to use only strategies which are symmetric in Ex I and Ex II, i.e. for which $L(\sigma, \tau) = L(\tau, \sigma)$. Every strategy s can be made symmetric. Define s' just as s, but with Ex I and Ex II interchanged, and then choose s or s' with probabilities $\frac{1}{2}, \frac{1}{2}$.

We give an example: Let s be: use Ex I all the time. The loss is $L(p, q) = 0$ for $p \geqq q$ and $L(p, q) = n(q - p)$ for $p \leqq q$. The corresponding symmetric strategy is: choose Ex I all the time or Ex II all the time with probabilities $\frac{1}{2}$, $\frac{1}{2}$; the loss is then $L(p, q) = n| p - q |/2$.

For a symmetric strategy $L(\sigma, \tau) = L(\tau, \sigma)$, and we can therefore write

$$L(p, q) = \frac{\sigma - \tau}{2} \sum_{k=1}^{n} (1 - \Pi_k(\sigma, \tau) + \Pi_k(\tau, \sigma)). \tag{1.1}$$

This is also the loss for an arbitrary strategy if the possibilities $p = \sigma$, $q = \tau$ and $p = \tau$, $q = \sigma$ have *a priori* probabilities $\frac{1}{2}$, $\frac{1}{2}$. We shall always use (1.1) as the loss-function.

Besides the strategy of the experimenter there is a strategy of "nature" which consists in choosing a pair p, q. The use of (1.1) as loss function may be interpreted in two ways.

First interpretation: Nature's strategy is to choose a pair σ, τ and then to play either Ex I with σ and Ex II with τ or vice versa. The experimenter is free to use any strategy.

Second interpretation: Nature is free to use any strategy but the experimenter is restricted to symmetric strategies.

Let s be a strategy of the experimenter and let t be a strategy of nature. We will write $L(s, t)$ in place of $L(p, q)$ in order to exhibit the dependence of L on both strategies.

2. Statement of the theorems.

In this paper we are interested in sequences of strategies. Both, the experimenter and nature have to choose strategies for every n. Let $S = \{s_n\}$ be a sequence of strategies of the experimenter and let $T = \{t_n\}$ be a sequence of strategies of nature. This defines a sequence

$$\{L_n = L(s_n, t_n)\}$$

of loss functions. We use the "order of infinity" of this sequence to construct a "weak" loss-function $l(S, T)$. For λ large enough we have $L_n = o(n^\lambda)$. Now let

$$l(S, T) = \inf \{\lambda \mid L_n = o(n^\lambda)\}. \tag{2.1}$$

The following theorems will be proved.

THEOREM 1. $\min_S \max_T l(S, T) = \max_T \min_S l(S, T) = \frac{1}{2}$, i.e. *there are sequences S_0 and T_0 such that for every sequence S and for every sequence T*

$$l(S_0, T) \leqq l(S_0, T_0) = \tfrac{1}{2} \leqq l(S, T_0).$$

An example of a sequence T_0 is given by

$$(2.2) \qquad | p_n - q_n | = mn^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}); \qquad p_n , q_n \rightarrow p$$

where $m > 0$ and $0 < p < 1$.

An example of a sequence S_0 is given by the strategies defined in [7], These strategies are given by numbers α (α serves to construct a sequential plan), and in order to get a sequence S_0 we must choose

$$(2.3) \qquad \alpha_n = a \cdot n^{\frac{1}{2}} + o(n^{\frac{1}{2}}), \qquad a > 0.$$

We will reserve the letters S_0 and T_0 for sequences of strategies for which (2.3) and (2.2) hold.

Of course the strategies S_0 will not be the only minimax strategies for the weak loss-function $l(S, T)$. There may be a large class of such strategies. To pick an especially good one out of this class one has to use a stronger criterion than l.

THEOREM 2. *Let $T_0 = \{t_n^{(0)}\}$ be a sequence as defined by* (2.2). *Then we have for every sequence S*

$$(2.4) \qquad \liminf_{n\to\infty} (L(s_n , t_n^{(0)})/n^{\frac{1}{2}}) \geqq C_1 > 0.$$

C_1 depends on the values of m and p in (2.2)

COROLLARY TO THEOREM 2. *Let $\{t_n^{(0)}\}$ be defined by $| p_n - q_n | = 0.849\ n^{-\frac{1}{2}}$, $p_n , q_n \rightarrow \frac{1}{2}$. Then we have for every S*

$$(2.5) \qquad \liminf_{n\to\infty} (L(s_n , t_n^{(0)})/n^{\frac{1}{2}}) \geqq 0.1876.$$

THEOREM 3. *Let $\{s_n^{(0)}\}$ be the sequence of minimax strategies as defined in* [7] (see also (4.2) in this paper). *For these strategies,* (2.3) *is valid with $a = 0.292$ and we have for every sequence T*

$$(2.6) \qquad \limsup_{n\to\infty} (L(s_n^{(0)}, t_n)/n^{\frac{1}{2}}) \leqq 0.376.$$

THEOREM 4. *Let $S_0 = \{s_n^{(0)}\}$ be a sequence of strategies as defined by* (2.3). *Then we have for every sequence T*

$$(2.7) \qquad \limsup_{n\to\infty} (L(s_n^{(0)}, t_n)/n^{\frac{1}{2}}) \leqq C_2 < \infty.$$

C_2 depends on the value of a in (2.3).

REMARK: If the random variables X and Y of Ex I and Ex II are normally distributed with common known variance, we can prove virtually the same theorems.

Theorem 1 follows immediately from Theorems 2 and 3 or 4. The proofs of Theorems 3 and 4 depend on results of [7]. The minimum over all possible C_2 of theorem 4 is 0.96; therefore Theorem 3 is not merely a corollary to Theorem 4. The proof of Theorem 2 is independent of [7]. The main idea of the proof of Theorem 2 is as follows: Construct a new game which obviously gives a smaller loss than the old one and which is simple enough to allow the computation of a best strategy. Then the new game can be used to obtain a lower bound for the loss of the old game.

The author is very grateful to the referee for his valuable suggestions.

3. Proof of Theorem 2. We need a lemma.

Lemma. *Let $Z_{i,n}(i = 1, 2, \cdots n)$ be n identically distributed random variables with*

$$E(Z_{i,n}) = m_n = mn^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}), \qquad \operatorname{Var}(Z_{i,n}) = \sigma_n^2 \to \sigma^2,$$

$$\sigma^2 > 0, \qquad E(| Z_{i,n} - E(Z_{i,n}) |^3) = b_n \to b.$$

Let further $U_k^{(n)} = \sum_{i=1}^k Z_{i,n}$. Then (for $n \to \infty$) we have

$$\frac{1}{n} \sum_{k=1}^{n} P(U_k^{(n)} < 0) \to \int_0^1 \Phi\left(-\frac{m}{\sigma} (x)^{\frac{1}{2}}\right) dx$$

where Φ denotes the standardised normal cumulative distribution function.

Proof: Let

$$A = \left| \frac{1}{n} \sum_{k=1}^{n} P(U_k^{(n)} < 0) - \int_0^1 \Phi\left(-\frac{m}{\sigma} (x)^{\frac{1}{2}}\right) dx \right| \leqq \sum_{k=1}^{n} A_{k,n}$$

where

$$n \cdot A_{k,n} = \left| P(U_k^{(n)} < 0) - n \int_{(k-1)/n}^{k/n} \Phi\left(-\frac{m}{\sigma} (x)^{\frac{1}{2}}\right) dx \right| \leqq 2.$$

As a first step we prove that $n \cdot A_{k.n} \to 0$ uniformly for all $k \geqq n^{\frac{1}{2}}$. By a theorem due to Berry and Esseen ([3], p. 201) we have

$$\left| P(U_k^{(n)} < 0) - \Phi\left(-\frac{m_n}{\sigma_n} (k)^{\frac{1}{2}}\right) \right| \leqq c \frac{b_n}{\sigma_n^3} k^{-\frac{1}{2}} \leqq c \frac{b_n}{\sigma_n^3} n^{-\frac{1}{4}} \to 0. \tag{3.1}$$

By the mean value theorem there is a $\xi_{k,n}$ with $(k - 1)/n \leqq \xi_{k,n} \leqq k/n$ so that $n\int_{(k-1)/n}^{k/n} \Phi(- (m/\sigma)(x)^{\frac{1}{2}})dx = \Phi(- (m/\sigma)(\xi_{k,n})^{\frac{1}{2}})$. Therefore

$$\left| \phi\left(-\frac{m_n}{\sigma_n} (k)^{\frac{1}{2}}\right) - n \int_{(k-1)/n}^{k/n} \Phi\left(-\frac{m}{\sigma} (x)^{\frac{1}{2}}\right) dx \right| = (2\pi)^{-\frac{1}{2}} \left| \int_{\underline{a}}^{\bar{a}} e^{-t^2/2}\, dt \right| = B.$$

Here $\underline{a} = -(m/\sigma)(k/n)^{\frac{1}{2}} + o(n^{-\frac{1}{2}})$ and $\bar{a} = -(m/\sigma)\xi^{\frac{1}{2}}_{k,n}$. Now $| \bar{a} - \underline{a} | \to 0$ uniformly in k and therefore $B \to 0$ uniformly in k. This together with (3.1) shows that $n \cdot A_{k,n} \to 0$ uniformly in k (for $k > n^{\frac{1}{2}}$). It follows that

$$\textstyle\sum_{k=[(n)^{\frac{1}{2}}]+1}^{n} A_{k,n} \to 0$$

and we have

$$0 \leqq A \leqq \sum_{k=1}^{[(n)^{\frac{1}{2}}]} A_{k,n} + \sum_{k=[(n)^{\frac{1}{2}}]+1}^{n} A_{k,n} \leqq 2n^{-\frac{1}{2}} + \sum_{k=[(n)^{\frac{1}{2}}]+1}^{n} A_{k,n} \to 0.$$

This proves the lemma.

We define a new game with a loss-function Λ so that

$$\inf_s L(s, t) \geqq \inf_s \Lambda(s, t) \tag{3.2}$$

for every fixed n and t. The new game will give the experimenter greater strategic possibilities.

We can play the old game in the following manner: At every step an umpire performs Ex I and Ex II. The experimenter simply states whether he wants Ex I or Ex II this time and then the umpire tells him the outcome of that experiment which he had chosen and this outcome counts for the loss-function.

The new game will be played as follows: The umpire performs the two experiments. The experimenter states which experiment he wants. Then the umpire tells the experimenter the outcomes of Ex I and of Ex II. For the loss-function, only the outcome of the chosen experiment counts, but of course this time the information which the experimenter gets at every step is greater. As the experimenter is free to use this additional information or not he has all the strategic possibilities he had before as well as some new ones. The set of values Λ on the right side of inequality (3.2) contains the set of values of L on the left side. Therefore the inequality is justified.

We will now show, that there is a uniformly best strategy in the new game, namely to play whichever experiment is ahead so far.

In the old game the choice at a certain step will not only influence the gain at this step but also the information available at the following steps. In the new game the information can not be influenced by a strategy, so that the best thing to do is to minimize the loss step by step. All we have to do is to get a strategy which (at the $k + 1$th step) minimizes $1 - \Pi_{k+1}(\sigma, \tau) + \Pi_{k+1}(\tau, \sigma)$ or maximizes $\Pi_{k+1}(\sigma, \tau) - \Pi_{k+1}(\tau, \sigma)$ where $\sigma \geqq \tau$ (see (1.1)).

The $k + 1$th step of a strategy is given by a function

$$f = f(X_1, \cdots Y_k ; Y_1, \cdots Y_k)$$

with $0 \leqq f \leqq 1$. f gives the probability of choosing Ex I as a function of the previous history and we have $\Pi_{k+1} = E(f)$. In the new game $X_1 \cdots X_k$ and $Y_1 \cdots Y_k$ are independent binomial random variables with parameters p and q. Since $(\sum_{i=1}^k X_i = \mu, \sum_{i=1}^k Y_i = \nu)$ is sufficient for (p, q) we may restrict ourselves to functions of μ and ν alone. Therefore

$$\Pi_{k+1}(\sigma, \tau) = E(f \mid \sigma, \tau) = \sum_{\mu=1}^{k} \sum_{\nu=1}^{k} f(\mu, \nu) P\left(\sum_{i=1}^{k} X_i = \mu \mid \sigma\right) P\left(\sum_{i=1}^{k} Y_i = \nu \mid \tau\right)$$

with $P(\sum_{i=1}^k X_i = \mu \mid \sigma) = \binom{k}{\mu} \sigma^\mu (1 - \sigma)^{k-\mu}$ and similarly for y, ν, τ. It follows that

$$\Pi_{k+1}(\sigma, \tau) - \Pi_{k+1}(\tau, \sigma) = \sum_{\mu=1}^{k} \sum_{\nu=1}^{k} f(\mu, \nu) \binom{k}{\mu} \binom{k}{\nu} \sigma^\mu (1 - \sigma)^{k-\mu} \tau^\nu (1 - \tau)^{k-\nu} D_{\mu\nu}$$

where $D_{\mu\nu} = 1 - (\tau(1 - \sigma)/\sigma(1 - \tau))^{\mu-\nu} \geqq 0$ if $\mu > \nu$ and $D_{\mu\nu} \leqq 0$ if $\mu < \nu$.

This shows that we maximize $\Pi_{k+1}(\sigma, \tau) - \Pi_{k+1}(\tau, \sigma)$ by choosing $f(\mu, \nu) = 1$ for $\mu > \nu$ and $f(\mu, \nu) = 0$ for $\mu < \nu$. For $\mu = \nu$ we choose $f(\mu, \nu) = \frac{1}{2}$ in order to have a symmetric strategy. It follows that the best symmetric strategy (regardless of the strategy of nature) is to play Ex I if $\mu > \nu$ and to play Ex I with probability $\frac{1}{2}$ when $\mu = \nu$.

Let $X_i - Y_i = Z_i$ and $\sum_{i=1}^{k} Z_i = U_k$. If we use the best symmetric strategy we have

$$\Pi_{k+1} = P(U_k > 0) + \tfrac{1}{2}P(U_k = 0)$$

The formula is correct even for $k = 0$ if we put $U_0 \equiv 0$. The loss-function Λ is then (see (1.1))

$$\Lambda = \frac{\sigma - \tau}{2} \sum_{k=0}^{n-1} (1 - P(U_k > 0 \mid \sigma, \tau) - \tfrac{1}{2}P(U_k = 0 \mid \sigma, \tau) \\ + P(U_k > 0 \mid \tau, \sigma) + \tfrac{1}{2}P(U_k = 0 \mid \tau, \sigma))$$

Using $P(U_k = 0 \mid \sigma, \tau) = P(U_k = 0 \mid \tau, \sigma)$ and $P(U_k > 0 \mid \tau, \sigma) = P(U_k < 0 \mid \sigma, \tau)$ we get

$$\Lambda = (\sigma - \tau) \sum_{k=0}^{n-1} (P(U_k < 0 \mid \sigma, \tau) + \tfrac{1}{2}P(U_k = 0 \mid \sigma, \tau)). \tag{3.3}$$

Let nature use T_0, i.e. let $\sigma_n - \tau_n = m \cdot n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}})$ and $\sigma_{n i}\tau_n \rightarrow p$. U_k will depend now on n and we write $U_k^{(n)}$. In (3.3) we drop the terms $P(U_k = 0 \mid \sigma, \tau)$ and we write $\geqq$ in place of $=$. It then follows, that,

$$n^{-\frac{1}{2}}\Lambda_n \geqq m \frac{1}{n} \sum_{k=0}^{n-1} P(U_k^{(n)} < 0 \mid \sigma, \tau) + o(1).$$

For $U_k^{(n)}$ the assumptions of the lemma hold and therefore

$$\liminf_{n \to \infty} n^{-\frac{1}{2}}\Lambda n \geqq m \int_0^1 \Phi\left(-\frac{m}{\delta}(x)^{\frac{1}{2}}\right) dx,$$

where $\delta^2 = 2p(1 - p)$. As we have used the best possible strategy (say $\bar{s}$) we conclude that

$$n^{-\frac{1}{2}}L(s, t_n^{(0)}) \geqq \inf_s n^{-\frac{1}{2}}\Lambda(s, t_n^{(0)}) = n^{-\frac{1}{2}}(\bar{s}, t_n^{(0)}).$$

But then we have

$$\liminf_{n \to \infty} (L(s_n, t_n^{(0)})/n^{\frac{1}{2}}) \geqq C_1 > 0$$

with $C_1 = m\int_0^1 \Phi(- mx^{\frac{1}{2}}/\delta)\, dx$.

Proof of the Corollary to Theorem 2: C_1 attains its maximum for $\delta^2 = \frac{1}{2}$ (i.e. for $p = \frac{1}{2}$) and $m = 0.849$. The value of the maximum is 0.187. This follows by numerical computations.

4. Proof of Theorem 3. We give some results from [7] which will be needed in the sequel. The strategy treated in [7] runs as follows: At the first $2K$ steps (K is a random variable) Ex I and Ex II will be used alternately so that we have a sequence of random variables $X_1, Y_1, X_2, Y_2, \cdots X_k, Y_k$. Let $U_k = \sum_{i=1}^{k} (X_i - Y_i)$. As long as $-\alpha < U_k < +\alpha$ we make another pair of experiments and so get X_{k+1}, Y_{k+1}. When $U_k \geqq +\alpha$ we continue for the remaining $n - 2k$ steps with Ex I and when $U_k \leqq -\alpha$ we do so with Ex II. In every case, we stop after n experiments. A sequence of strategies for the experimenter consists of choosing a number α for each n, and from now on s_n always means this kind of strategy.

It was shown in [7] that

$$(4.1) \qquad L_n \leqq n(\sigma - \tau)/(u^\alpha + 1) + \alpha(u^\alpha - 1)^2/(u^\alpha + 1)^2 = M$$

where $\sigma = \max(p, q)$ and $\tau = \min(p, q)$ and $u = (\sigma(1 - \tau)/\tau(1 - \sigma))$. There are strategies $s_n^{(0)}$ and $t_n^{(0)}$ such that, for all s_n and t_n,

$$(4.2) \qquad M(s_n^{(0)}, t_n) \leqq M(s_n^{(0)}, t_n^{(0)}) \leqq M(s_n, t_n^{(0)}),$$

i.e. there are minimax strategies for the approximate loss-function M. For the strategies $s_n^{(0)}$ we have

$$(4.3) \qquad \alpha_n = 0.292\, n^{\frac{1}{2}} + o(n^{\frac{1}{2}}),$$

and for the strategies $t_n^{(0)}$ we have

$$(4.4) \qquad \sigma_n - \tau_n = 1.89\, n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}); \qquad \sigma_n ; \tau_n \longrightarrow \tfrac{1}{2},$$

and

$$(4.5) \qquad u_n^{\alpha_n} \longrightarrow 9.06.$$

It is now easy to prove Theorem 3. We use (4.1) and (4.2) and get

$$n^{-\frac{1}{2}} L(s_n^{(0)}, t_n) \leqq n^{-\frac{1}{2}} M(s_n^{(0)}, t_n) \leqq n^{-\frac{1}{2}} M(s_n^{(0)}, t_n^{(0)}).$$

The last expression converges (using (4.3), (4.4) and (4.5)) to 0.367. Therefore

$$\limsup_{n \to \infty} \frac{L(s_n^{(0)}, t_n)}{n^{\frac{1}{2}}} \leqq 0.367.$$

5. Proof of Theorem 4. Now let $s_n^{(0)}$ be a strategy as defined in (2.3) i.e. let $\alpha_n = an^{\frac{1}{2}} + o(n^{\frac{1}{2}})$ and $a > 0$. Then

$$(5.1) \quad n^{-\frac{1}{2}} L(s_n^{(0)}, t) \leqq n^{-\frac{1}{2}} M(s_n^{(0)}, t)$$
$$= n^{\frac{1}{2}}(\sigma - \tau)/(u^{\alpha_n} + 1) + (a + o(1))(u^{\alpha_n} - 1)^2/(u^{\alpha_n} + 1)^2.$$

The last term is smaller or equal to $a + o(1)$. To get an upper bound for the term $n^{\frac{1}{2}}(\sigma - \tau)/(u^{\alpha_n} + 1)$ we first show

$$\ln u \geqq (\sigma - \tau) \ln 5.$$

From $\sigma = \tau$ follows $u = 1$. Therefore the inequality holds for $\sigma = \tau$. Now let $\sigma > \tau$. Because $\tau(1 - \sigma) < \frac{1}{4}$ we have

$$(\sigma - \tau)^{-1} \ln u = (\sigma - \tau)^{-1} \ln (1 + ((\sigma - \tau)/\tau(1 - \sigma)))$$

$$> (\sigma - \tau)^{-1} \ln (1 + 4(\sigma - \tau)) \geqq \operatorname*{mim}_{0<x\leqq 1} x^{-1} \ln (1 + 4x) = \ln 5.$$

Therefore

$$n^{\frac{1}{2}}(\sigma - \tau)/(u^{\alpha_n} + 1) \leqq n^{\frac{1}{2}}(\sigma - \tau)/5^{n^{\frac{1}{2}}(\sigma-\tau)(a+o(1))}$$

$$\leqq \max_{x\geqq 0} x/5^{x(a+0(1))} = (e\cdot a\cdot \ln 5)^{-1} + o(1)$$

This proves that

$$\limsup_{n\to\infty} n^{-\frac{1}{2}} M(s_n^0, t_n) \leqq C_2$$

with $C_2 = a + (ae \ln 5)^{-1}$. Taking account of (5.1) it also proves (2.7).

REFERENCES

[1] R. Bellman, "A problem in the sequential design of experiments," *Sankhyā*, Vol. 16 (1956), pp. 221–229.

[2] R. N. Bradt, S. M. Johnson, and S. Karlin, "On sequential designs for maximizing the sum of n observations," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1060–1074.

[3] B. V. Gnedenko and A. N. Kolmogorov, "*Limit distributions for sums of independent random variables*," translated by K. L. Chung, Cambridge, Addison-Wesley, 1954.

[4] J. R. Isbell, "On a problem of Robbins," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 606–610.

[5] H. E. Robbins, "Some aspects of the sequential design of experiments," *Bull. Am. Math. Soc.*, Vol. 55 (1952), pp. 527–535.

[6] H. Robbins, "A sequential decision problem with a finite memory," *Proc. Nat. Acad. Sci.*, Vol. 42 (1956), pp. 920–923.

[7] Walter Vogel, "A sequential design for the two armed bandit," Vol. 31 (1960), pp. 430–443.