17 Simulation Results for a New Two-armed Bandit Heuristic

Ronald L. Rivest and Yiqun Yin

17.1 Introduction

Bandit problems were first introduced by Robbins [10] in 1952. The name derives from an imagined slot machine with $k \geq 2$ arms. When an arm is pulled, the player wins a random reward according to an unknown probability distribution π_j. The player's problem is to choose a sequence of pulls on the k arms, depending on the results of previous trials, so as to maximize the long-run total reward. In general, we consider the problem of sampling $x_1, x_2, \ldots$ sequentially from k statistical populations (arms, medical treatments, etc.) specified by density functions $f(x, \theta_j)$ with respect to some measure v, where $f(\cdot, \cdot)$ is known and the θ_j's are unknown parameters belonging to some set Θ. We assume that the average reward

$$\mu(\theta) = \int_{-\infty}^{\infty} x f(x, \theta)\, dv(x) \tag{1}$$

is well defined for all $\theta \in \Theta$. The goal is to maximize, in some sense, the expected value of the sum

$$S_n = x_1 + x_2 + \cdots + x_n \tag{2}$$

as $n \to \infty$. There have been several different approaches to this problem based on different formulations of optimality.

In 1985 Lai and Robbins [6] constructed a class of asymptotically efficient strategies (also called "adaptive allocation rules"), and many works in recent years are based on their results. An adaptive allocation rule ϕ for a k-armed bandit problem is a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values in the set $\{1, 2, \ldots, k\}$. We will give a brief survey of their algorithms in section 17.2.

Another approach is to consider for large fixed n (finite horizon) the Bayes problem of maximizing

$$\int_{\Theta} E_\theta S_n\, dH(\theta), \tag{3}$$

where $H(\theta)$ is a prior distribution on Θ^k, and where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ gives the parameters defining the probability distribution for the rewards of each arm. Berry and Fristedt [1] studied the dynamic programming equations for the Bayes optimal solution analytically and obtained several

interesting results about the Bayes optimal rules with respect to general priors. However, Bayes rules are usually described only implicitly by the dynamic programming equations, and they are usually difficult to compute numerically.

Besides the bandit problem we discussed above, there is also a class of "discounted multiarmed bandit problems," in which a discount factor of β, for some $0 < \beta < 1$, is introduced. Here we consider the problem of maximizing the expected value of the series

$$\sum_{i=1}^{\infty} \beta^{i-1} x_i. \tag{4}$$

Major advances in this problem were made by Gittins and Jones (see [3] for a survey of their work); their strategies are usually called "Gittins index rules." These rules have been shown to be optimal for the discounted problem.

In another point of view, the classical bandit problem can also be viewed as a learning process, in which we make decisions according to what we have learned in the past. Narendra and Thathachar [9] used "learning automata" as a framework for attacking this problem. Their basic idea is to update the probabilities of pulling each arm at every stage, based on the previous results. Most of these schemes were shown to be ε-optimal, which means that for every ε there exists a learning automaton that can achieve an asymptotic average reward rate that is within ε of optimality. Because learning automata have limited number of states, one cannot expect optimal performance from them.

17.2 Asymptotically Efficient Adaptive Allocation Rules

Let

$$\mu^*(\theta) = \max_{1 \le j \le k} \mu(\theta_j) = \mu(\theta^*) \tag{5}$$

for some $\theta^* \in \{\theta_1, \theta_2, \ldots, \theta_k\}$. Robbins [10] formulated a notion of asymptotic optimality as obtaining

$$\lim_{n \to \infty} n^{-1} E_\theta S_n = \mu^*(\theta) \quad \text{for all} \quad \theta \in \Theta^k. \tag{6}$$

For the case $k = 2$, he also introduced a class of simple allocation rules that attains (6). A natural question is how to make the rule so that $n^{-1}E_\theta S_n$ approaches $\mu^*(\theta)$ as quickly as possible.

Lai and Robbins [6] introduced the concept of "regret" as

$$R_n(\theta) = n\mu^*(\theta) - E_\theta S_n = \sum_{j:\mu(\theta_j)<\mu^*(\theta)} (\mu^*(\theta) - \mu(\theta_j))E_\theta T_n(j) \tag{7}$$

where $T_n(j)$ is the total number of observations from π_j up to stage n. Therefore, maximizing $E_\theta S_n$ is equivalent to minimizing the regret $R_n(\theta)$. Their main theoretical result is that for every reasonably good allocation rule (one that satisfies $R_n(\theta) = o(n^a)$ for every $a > 0$ for every fixed θ), we also have

$$\liminf_{n\to\infty} \frac{R_n(\theta)}{\ln n} \geq \sum_{j:\mu(\theta_j)<\mu^*(\theta)} \frac{(\mu^*(\theta) - \mu(\theta_j))}{I(\theta_j, \theta^*)} \tag{8}$$

for all $\theta \in \Theta^k$, where

$$I(\theta, \lambda) = \int_{-\infty}^{\infty} \{\ln[f(x,\theta)/f(x,\lambda)]\} f(x,\theta)\, dv(x) \tag{9}$$

is the Kullback-Leibler information number which gives a measure of the difference between two density functions. Moreover, a class of allocation rules that asymptotically attains this theoretical lower bound is constructed.

For $j = 1, 2, \ldots, k$, let $Y_{j1}, Y_{j2}, \ldots, Y_{jT_n(j)}$ denote the successive observations from π_j. Define

$$\hat{\mu}_n(j) = \frac{Y_{j1} + Y_{j2} + \cdots + Y_{jT_n(j)}}{T_n(j)}$$

as the estimated sample mean, and define a certain upper confidence bound for the mean of each population π_j as

$$U_n(j) = g_{n,T_n(j)}(Y_{j1}, \ldots, Y_{jT_n(j)}). \tag{10}$$

Define $j_n \in \{1, 2, \ldots, k\}$ such that

$$\hat{\mu}_n(j_n) = \max\{\hat{\mu}_n(j) : T_n(j) \geq \delta n\}. \tag{11}$$

At stage $n + 1$, then, where $j = (n + 1) \bmod k$, we select arm j only if

$$\hat{\mu}_n(j_n) \leq U_n(j); \tag{12}$$

otherwise we select arm j_n. Lai and Robbins proved that this rule satisfies the equation

$$E_\theta(T_n(j)) \sim \frac{\ln n}{I(\theta_j, \theta^*)} \tag{13}$$

for every j such that $\mu(\theta_j) < \mu(\theta^*)$.

For normal, Bernoulli, Poisson, and double exponential populations, they expressed the upper confidence bound as

$$U_n(j) = \inf\{\lambda \geq \hat{\mu}_n(j) : I(\hat{\mu}_n(j), \lambda) \geq a_{ni}\}, \tag{14}$$

where a_{ni} ($n = 1, 2, \ldots, i = 1, 2, \ldots, n$) are positive constants satsifying certain conditions. For example, in the case of a two-armed Bernoulli bandit a_{ni} can be chosen as $(\ln n)/i$.

17.3 A New Heuristic Algorithm

In this section we propose a simple heuristic algorithm for the bandit problem, which seems to have (empirically) better performance than the algorithm of Lai and Robbins [6] discussed in the previous section.

As before, let $Y_{j1}, Y_{j2}, \ldots, Y_{jT_n(j)}$ denote the successive observations from π_j up to stage n. We define $\hat{\mu}_n(j)$ as the estimated sample mean (as above), and define $\hat{\sigma}_n(j)$ as the estimated standard deviation of the sample mean, for $j = 1, 2, \ldots, k$. The new allocation rule is the following:

At stage $n + 1$, we associate a random variable $Z_n(j)$ with each arm j, where $Z_n(j)$ has a normal distribution with mean $\hat{\mu}_n(j)$ and standard deviation $\hat{\sigma}_n(j)$. We then sample from population j_n, where

$$Z_n(j_n) = \max\{Z_n(1), Z_n(2), \ldots, Z_n\{k\}\}.$$

We call this new heuristic the "Z-heuristic." The $Z_n(j)$ variables are intended to reflect the learner's uncertainty about the true values for $\theta(j)$.

We now apply both the Lai and Robbins algorithm and our new heuristic to construct allocation rules for normal and Bernoulli populations. The simulation results are given in the next section.

17.4 Experimental Results

We considered the bandit problem for two kinds of arms: Bernoulli variables and normal variables. In both case the parameter θ was equal to the

expected reward. For the Bernoulli arms with parameter θ a reward of value 1 was received with probability θ, and a reward of value 0 was received with probability $(1 - \theta)$. For the normal arms the mean reward was θ, and the variance was equal to 1.

We implemented the Lai and Robbins algorithm and the Z-heuristic and ran these algorithms for $n = 10^7$ trials for $k = 2$ and four sets of probabilities:

$\theta = (0.1, 0.9)$,

$\theta = (0.46, 0.54)$,

$\theta = (0.496, 0.504)$, and

$\theta = (0.4996, 0.5004)$.

We also calculated for these experiments the theoretical bound from equation (13). Because this formula behaved poorly for small values of n, we also calculated the value of the heuristic formula

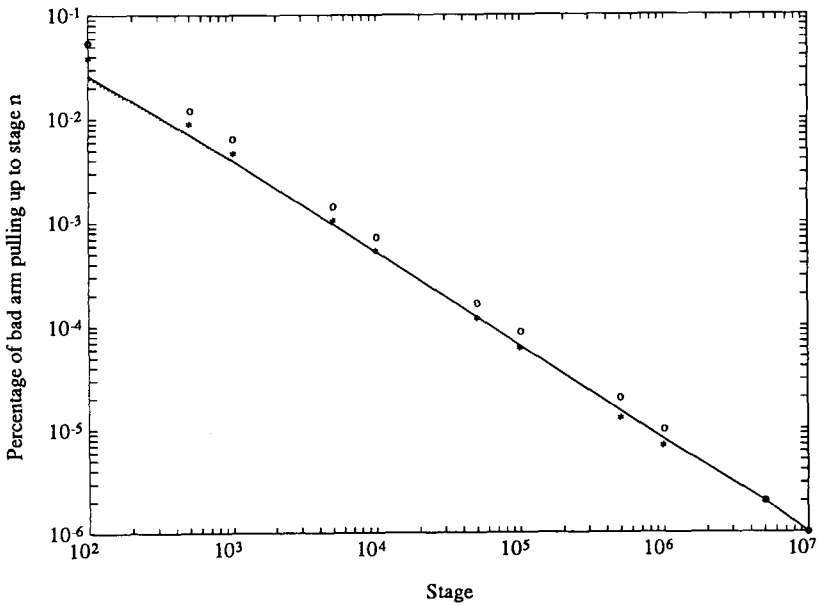

Figure 17.1
Two Bernoulli arms: $\theta = (0.1, 0.9)$

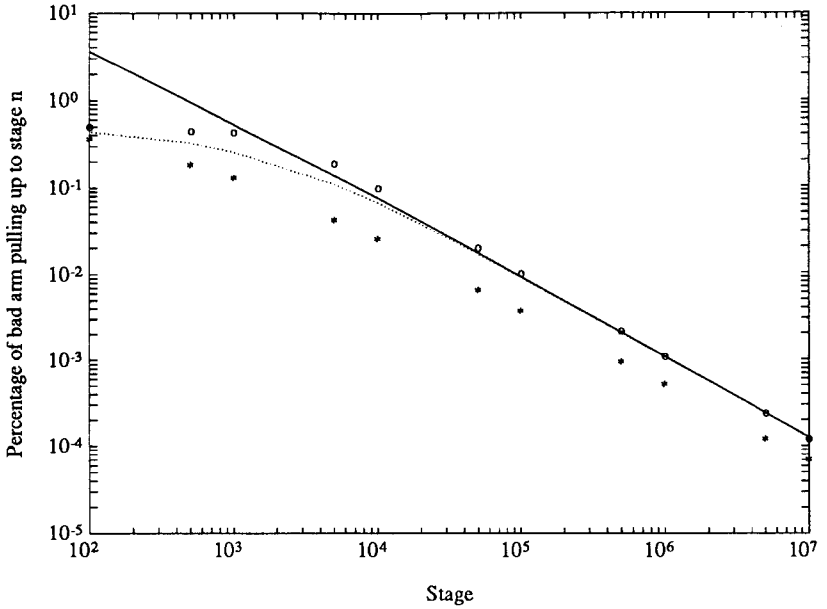

Figure 17.2
Two Bernoulli arms: $\theta = (0.46, 0.54)$.

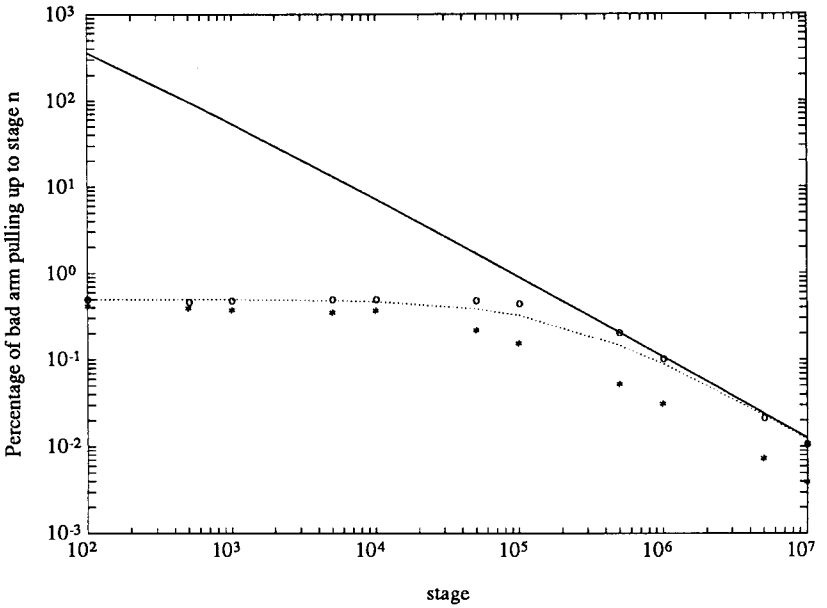

Figure 17.3
Two Bernoulli arms: $\theta = (0.496, 0.504)$

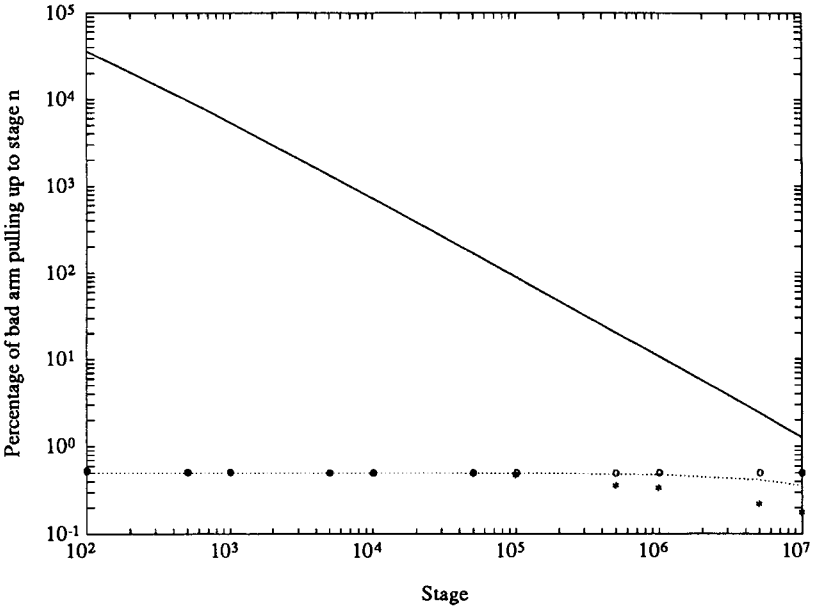

Figure 17.4
Two Bernoulli arms: $\theta = (0.4996, 0.5004)$

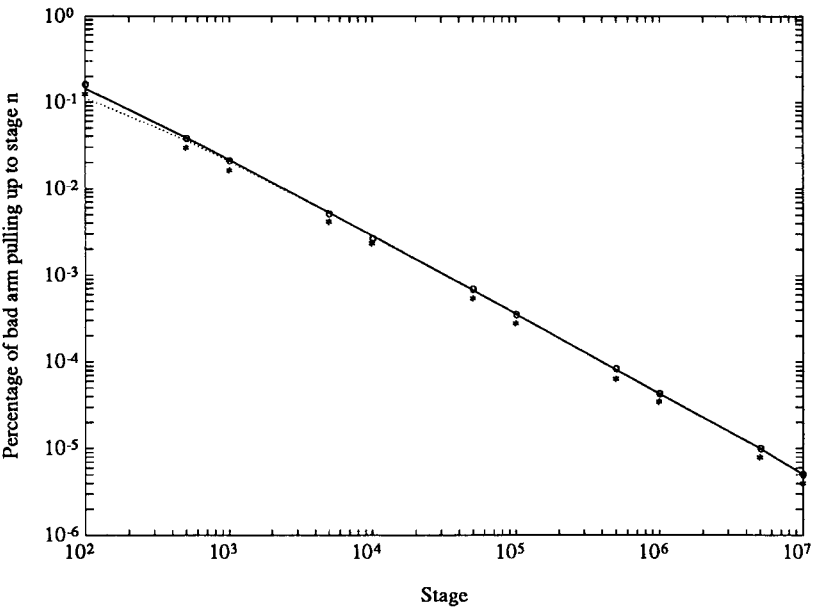

Figure 17.5
Two normal arms: $\theta = (0.1, 0.9)$

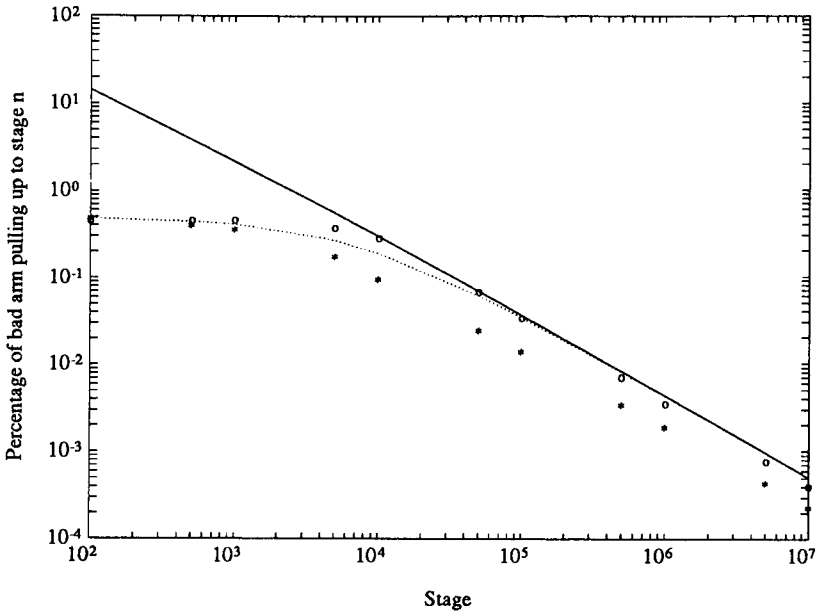

Figure 17.6
Two normal arms: $\theta = (0.46, 0.54)$

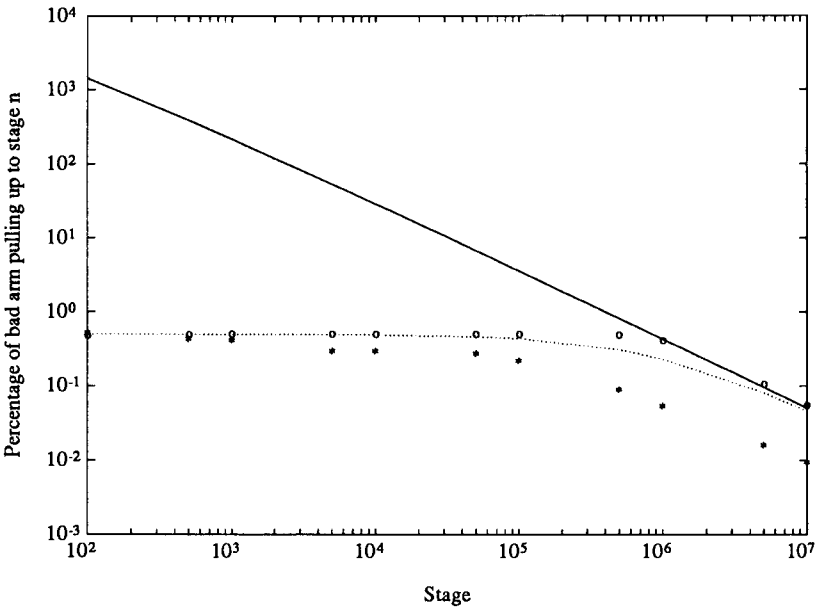

Figure 17.7
Two normal arms: $\theta = (0.496, 0.504)$

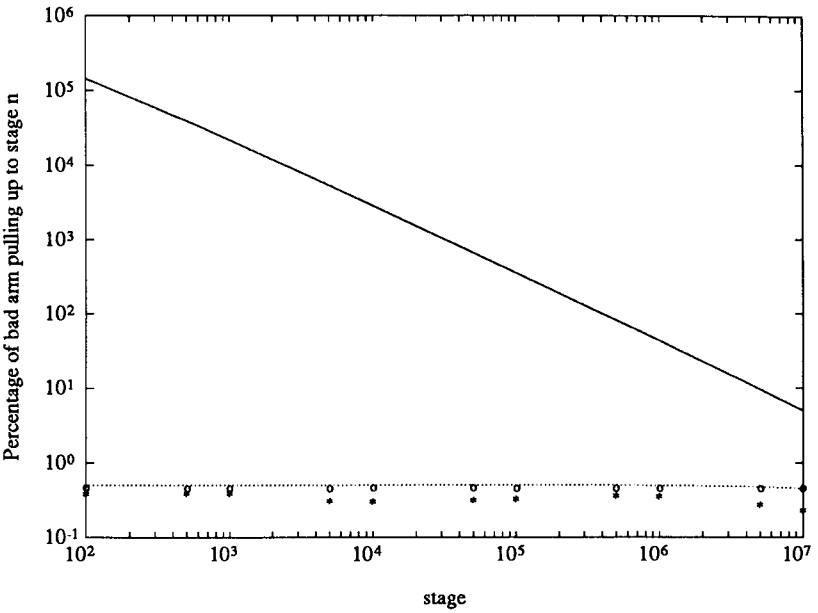

Figure 17.8
Two normal arms: $\theta = (0.4996, 0.5004)$

$$E_\theta(T_n(j)) \sim \frac{\ln n}{I(\theta_1, \theta_2) + \dfrac{2 \ln n}{n}} \tag{15}$$

(which is approximately $n/2$ for small n, but approaches the Lai and Robbins bound asymptotically).

The results are plotted in figures 17.1–17.8. In each case, the Lai and Robbins bound is plotted as a heavy line, and formula (15) is plotted as a dotted line. The performance of the Lai and Robbins algorithm is plotted as o's, and the Z-heuristic is plotted as $*$'s.

We see that the Z-heuristic performs much better than the Lai and Robbins algorithm for the experiments we tried. We conjecture that the Z-heuristic is asymptotically optimal, and we are working to prove this conjecture.

Notes

Supported by NSF grant CCR-8914428, ARO grant DAAL03-86-K-0171, and the Siemens Corporation.

References

[1] Berry, D. A., and Fristedt, B. *Bandit Problems-Sequential Allocation of Experiments.* Chapman and Hall, New York, 1985.

[2] Chang, Fu, and Tze Leung Lai. Optimal stopping and dynamic allocation. *Advances in Applied Probability* 19:829–853, 1987.

[3] Gittins, J. C. *Multi-armed Bandit Allocation Indices.* John Wiley and Sons, New York, 1989.

[4] Lai, T. L. Asymptotic solutions of bandit problems. In W. Fleming and P. L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications.* Springer-Verlag, Berlin, 1988.

[5] Lai, T. L., and Herbert Robbins. Optimal sequential sampling from two populations. *Proceedings National Academy Science USA* 81:1284–1286, 1984.

[6] Lai, T. L., and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6:4–22, 1985.

[7] Lai, Tze Leung. Adaptive treatment allocation and the multi-armed bandit problem. *Annals of Statistics* 15 (3): 1091–1114, 1987.

[8] Lai, Tze Leung, and Zhiliang Ying. Open bandit processes and optimal scheduling of queueing networks. *Advances in Applied Probability* 20:447–472, 1988.

[9] Narendra, Kumpati S., and Mandayam A. L. Thathachar. *Learning Automata—An Introduction.* Prentice-Hall, New York, 1989.

[10] Robbins, H.,Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society* 55:527–535, 1952.

[11] Tsitsiklis, John N. A lemma on the multiarmed bandit problem. *IEEE Transactions on Automatic Control*, AC-31 (6), 1986.

[12] Whittle, Peter. *Optimization Over Time*, Vol. 2. John Wiley and Sons, New York, 1983.