Bandit Problems With Infinitely Many Arms

Author(s): Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, Larry A. Shepp

Source: *The Annals of Statistics*, Vol. 25, No. 5 (Oct., 1997), pp. 2103-2116

Published by: Institute of Mathematical Statistics

Stable URL: http://www.jstor.org/stable/2959020

Accessed: 23/11/2010 14:21

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=ims.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *The Annals of Statistics.*

http://www.jstor.org

The Annals of Statistics
1997, Vol. 25, No. 5, 2103–2116

BANDIT PROBLEMS WITH INFINITELY MANY ARMS[1]

BY DONALD A. BERRY, ROBERT W. CHEN, ALAN ZAME,
DAVID C. HEATH AND LARRY A. SHEPP

*Duke University, University of Miami, University of Miami,
Cornell University and AT&T Bell Laboratories*

We consider a bandit problem consisting of a sequence of n choices from an infinite number of Bernoulli arms, with $n \to \infty$. The objective is to minimize the long-run failure rate. The Bernoulli parameters are independent observations from a distribution F. We first assume F to be the uniform distribution on $(0, 1)$ and consider various extensions. In the uniform case we show that the best lower bound for the expected failure proportion is between $\sqrt{2}/\sqrt{n}$ and $2/\sqrt{n}$ and we exhibit classes of strategies that achieve the latter.

1. Introduction.

A bandit problem consists of a sequence of choices from among a set of stochastic processes, or arms. We consider discrete time and Bernoulli arms, with Arm i having success probability p_i, $i = 1, \dots, A$. The parameters p_i are unknown. Regard them to be independent random variables with p_i having (known) distribution F_i. The (unconditional) probability of success on the initial selection of Arm i is the mean of p_i: $E(p_i) = \int pF_i(dp)$. The decision maker chooses an arm for observation at each of a number n of decision epochs; n is called the horizon. Information accrues about an arm that is selected for observation: namely, the distribution of p_i is updated using Bayes' theorem based on the resulting observation. Choices are sequential in the sense that they can depend on which arms were chosen previously and on the resulting observations. A strategy (or decision procedure) specifies which arm to select at any time for every possible history of previous selections and observations.

The general setting of bandit problems is described in Berry and Fristedt (1985). Other pertinent references include Whittle (1982, 1983) and Gittins (1989). A common objective in bandit problems is to maximize the expected value of some function of the sequence of observations. A strategy is optimal if this expected value is maximal when following that strategy. An example objective function is the sum of the n observations. The decision problem that corresponds to this objective in the Bernoulli case is finding a strategy that maximizes the expected number of successes.

If $n = 1$ (and $A < \infty$), then the decision problem is trivial: choose the arm with the largest mean, and the expected maximum is $\max\{Ep_1, \dots, Ep_A\}$. The decision problem is trivial for any n if all the F_i are one-point distribu-

Received October 1995; revised February 1997.

[1] Research supported in part by Grant DMS-94-04200.

AMS 1991 *subject classifications.* 62C25, 62L05, 60F99.

Key words and phrases. Bandit problems, sequential experimentation, dynamic allocation of Bernoulli processes, staying with a winner, switching with a loser.

tions, or if one F_i is wholly to the right of the others (i.e., if there exist an i and a u such that $F_i(u) = 0$ and $F_j(u) = 1$ for all $j \neq i$). However, when $A > 1$, $n > 1$ and none of the F_i is wholly to the right of the others, then the problem is nontrivial. Generally speaking, when n is large the decision maker is willing to sacrifice immediate gain (eschewing an arm with large mean if necessary) while testing arms and searching for one with an even larger mean that will yield a long-term benefit. But when n is small, information has less value for helping to glean future profits; such information might reasonably be eschewed in favor of selecting an arm that has a large mean. Therefore, during the course of a trial and as the horizon nears, arms having greater mean become more appealing even if they have little potential for providing information.

We consider the horizon n to be infinite. Typically, when $n = \infty$ the expected number of successes is infinite for a wide class of strategies—perhaps even for every strategy—and so the decision problem is trivial. There are at least two modified objectives that give rise to interesting decision problems. One is to discount future observations [Berry and Fristedt (1985), Chapter 3]. A special case is geometric discounting [Gittins (1989); Banks and Sundaram (1992)] in which an observation at time t has utility β^{t-1}, where $0 \leq \beta < 1$, and the goal is to maximize the expected discounted number of successes. Banks and Sundaram (1992) show that in this discounted case, optimal strategies have the characteristic that when an arm is indicated for selection there is a positive probability that the same arm will be selected forever into the future, and that this is so whether A is finite or $A = \infty$.

The other variation is that of Robbins (1952), who considered maximizing the long-run success proportion. Robbins considered $A = 2$ arms and showed that basing selections on the immediately preceding observation (staying with a winner and switching on a loser) dominates strategies that do not use the history of observations. He also exhibited selection strategies that are asymptotically optimal in the sense that their long-run success proportion is $\max\{p_1, p_2\}$.

In this paper, we adopt Robbins's objective, maximizing the long-run success proportion. We assume throughout that $A = \infty$ and that $F_i = F$ for all i. The latter assumption means that all arms are exchangeable before making any observations. In view of exchangeability we can restrict consideration to strategies that call for using Arm 1 first and for using the smallest numbered arm whenever a new arm (one never used before) is selected. A *nonrecalling strategy* is one that always uses a new arm when switching from the current arm. So a nonrecalling strategy indicates Arm 1 for a period of time and then Arm 2 for a period of time and then Arm 3 for a period of time and so on. Also, if the period of time on Arm i is empty, then the same is true for that of Arm j for $j > i$.

We assume that F is not a one-point distribution. Therefore, once Arm 1 has been selected and has produced an observation, it is no longer exchangeable with the other arms: the distribution of p_1 becomes $pF(dp)/\int pF(dp)$ after a success and $(1-p)F(dp)/\int(1-p)F(dp)$ after a failure, while the distribution of the other p_i continues to be $F(dp)$. It seems reasonable to expect (and it is true) that Arm 1 should be used again after an immediate success and never again after an immediate failure. But suppose Arm 1 yields s immediate successes and then a failure, is it better to switch to Arm 2 or to stay with Arm 1? The answer depends on s and on the horizon n, and it is not easy to find in general. It is even more difficult to say which arm should be used when sample information including both successes and failures becomes available about several of the p_i. We address such questions in this paper for the case in which $n \to \infty$.

Infinite horizon problems are forgiving of finite (and some infinite) exploratory excursions. So when there is a strategy that maximizes the expected long-run proportion of successes, there are an infinite number that do as well. In the current problem there even exist strategies that maximize the expected long-run proportion of successes uniformly over the space of distributions F. For example, Herschkorn, Pekoz and Ross (1995) show that the nonrecalling strategy that indicates Arm i until it gives i failures in a row yields a long-run proportion of successes that is the essential supremum of $p \sim F$. While such robustness is appealing, such a strategy performs relatively poorly for given F and any fixed value of $n < \infty$. Suppose F is uniform on $(0, 1)$. Then the limiting proportion of successes for this strategy is 1. However, the strategy of Herschkorn, Pekoz and Ross can spend inordinate amounts of time waiting for a long run of failures before dispensing with arms whose performances are clearly mediocre. For example, when $n = 500$ this strategy's success proportion is only 0.79. We give strategies (in Section 4) that also have a limiting success proportion of 1 but that achieve success rates as high as 0.92 when $n = 500$. The focus of the current paper is the order of magnitude (depending on n) of a strategy's failure proportion; all strategies that we consider are optimal in the sense that their limiting expected failure proportions equal 0.

Bandit problems have applications in clinical trials, in on-line industrial experimentation and in many other settings. The version in which there are an unlimited number of exchangeable arms has applications as described by Banks and Sundaram (1992): (1) labor markets in which a worker has a many opportunities for jobs [A worker who accepts a job receives salary (on-line payoff) and she also gets information about her ability to do the job and therefore to earn good money in the long term. See also Jovanovic (1979), Wilde (1979) and Viscusi (1979). In a related vein, there may be a large number of cities in which a salesperson can ply his trade. Each day he chooses a city in which to operate and either has a successful sales day or not. Success brings immediate payoff but it also suggests that the city is a fruitful location to visit in the future.]; (2) a voter model of repeated elections in which a single voter elects a representative from a large set of candidates to represent her during the current period; (3) worker selection in which a firm chooses one at a time from a pool of workers and gets to observe the performance of the workers chosen for as long a period as desired; (4) general search for nondurable experience goods in which the consumer has numerous brands from which to choose; (5) dating (or marriage) in which an individual strives to maximize the proportion of successful dates—however defined—and can choose from among a large pool of candidates, with the possibility of repeating dates in an exploratory way to assess that candidate's success proportion. Two other possible applications are as follows: (6) mining for valuable resources such as gold or drilling for oil when there are many areas available for exploration (the miner or mining company can move to another location or continue in the same location, depending on results); (7) drug testing when there is an abundant supply of molecules available for consideration [in cancer research, for example, drugs can be tested on one patient at a time, with a success being defined as a patient's tumor responding (shrinking)].

The assumptions made in this paper may not be appropriate in any particular setting of the applications mentioned above. When they are not appropriate then the results of this paper do not apply, at least not perfectly. In the following example settings the assumptions are inappropriate or questionable. In drug testing the available molecules may be neither exchangeable nor independent. In labor markets observations of performance may not be exchangeable within a particular worker, as when a worker becomes better at a job because of experience gained while working at the job. We assume a 0–1 payoff; in drilling for oil, for example, successful wells have different values. Also, drilling in different locations may entail different costs and we do not consider sampling costs. Finally, the objective in a practical setting may be other than maximizing the long-run proportion of successes. Firms tend to discount successes that may be obtained in the distant future. Also, when a firm's distribution F is uniform on $(0, 1)$, maximizing the long-run success proportion means that the firm should fire many employees after but a single mistake (see Section 2). Employers associate negative utility with firing employees. There may be few actual settings in which firms would use a strategy that is optimal based solely on maximizing long-run success proportion. (One may be when the employees are air traffic controllers.) In medical trials it may be ethically questionable to gather information about an untested drug when a well-studied drug has been found to be successful on a substantial proportion of patients. Despite these caveats, there are a variety of instances within each of the applications described in which the assumptions made and results obtained in this paper apply reasonably well.

In Section 2, we assume that F is beta(1, 1), the uniform distribution on $(0, 1)$. Section 3 extends Section 2 to the case in which F is a uniform distribution on a subset of $(0, 1)$. This extension is important for the following reason. When p_i is uniform on $(0, 1)$ [or on $(a, 1)$], the expected number of successes before the first failure on arm i is infinite. Therefore, one may be tempted to discard an arm as soon as it gives a failure, and indeed such a strategy is not bad (see Section 2). However, when p_i is uniform on $[a, b]$ for $b < 1$, some proportion of failures (at least $1-b$) is inevitable and therefore optimal strategies are qualitatively different. In Section 4 we consider the case in which F is an arbitrary distribution. In this paper we do not consider the interesting and difficult hierarchical setting in which F is itself unknown and observations on Arm i give information about F and therefore also about Arm j for $j \neq i$.

2. Uniform distribution on (0, 1). In this section, we consider the case in which the distribution F is uniform on the interval $(0, 1)$, that is, beta(1, 1). We show that discarding an arm whenever it yields a failure and never using it again is an excellent strategy. In particular, its asymptotic failure proportion has order of magnitude $1/\ln(n)$. But we show that this strategy can be improved. Namely, we find that $\sqrt{2/n}$ [which is less than $1/\ln(n)$] is an asymptotic lower bound over all strategies for the expected failure proportion and we exhibit several strategies with this asymptotic failure proportion.

First some notation and definitions. For each positive integer k, a strategy is called a *k-failure strategy* if it calls for using the same arm until that arm produces k failures, and when this happens, it calls for switching to a new arm (never returning to arms that have yielded failures). With the possible exception of the arm being used at the end of the experiment (i.e., after a total of n observations), every arm that has been used at least once yields exactly k failures.

For each constant α in $[0, 1)$, a strategy is called an *α-rate strategy* if it stays on the same arm until that arm has produced a failure rate greater than α, and when this happens discarding it and switching to a new arm. Again, discarded arms are never used again.

A 1-failure strategy and a 0-rate strategy are the same. This strategy is a modification of Robbins's stay-with-a-winner/switch-on-a-loser strategy to the infinite-arm setting. The failure proportion of this strategy in n trials is asymptotically $1/\ln(n)$. To see this consider the number S of immediate successes with any particular arm having parameter $p \sim F$. For $s = 1, 2, \dots, n$, the probability of at least s immediate successes on this arm is $P(S \geq s) = \int p^s F(dp) = 1/(s+1)$. Hence

$$E(S) = \sum_{s=1}^{n} \frac{1}{s+1} \approx \ln(n).$$

Thus, when following the 1-failure strategy, the expected number of failures in the first n trials, which is within one of the expected number of arms used, is approximately $n/\ln(n)$. Hence the expected proportion of failures in n trials is asymptotically $1/\ln(n)$.

Consider the k-failure strategy. Let $N(n, k)$ be the expected number of trials until the kth failure or until the horizon is reached, whichever comes first. For $n \leq k$, $N(n, k) = n$. The more interesting case is $n > k$:

$$\begin{aligned} N(n, k) &= \int_0^1 \sum_{j=0}^{k-1} j\binom{n}{j} u^{n-j}(1-u)^j \, du + \int_0^1 \sum_{j=k}^{n} j\binom{j-1}{k-1} u^{j-k}(1-u)^k \, du \\ &= \sum_{j=0}^{k-1} \frac{n}{n+1} + \sum_{j=k}^{n} \frac{k}{j+1} = k\left[1 + \sum_{j=k}^{n-1} \frac{1}{j+1}\right]. \end{aligned}$$

This applies for fixed k and asymptotically as $n \to \infty$.

The next result implies that asymptotically the best strategy among k-failure strategies has $k = 1$.

THEOREM 1. *As $n \to \infty$ the expected failure proportion for k-failure strategies is increasing in k.*

PROOF. Asymptotically, the expected failure proportion when following a k-failure strategy is

$$\frac{k}{N(n,k)} = \frac{1}{[1 + \sum_{j=k}^{n-1} 1/(j+1)]}.$$

Since $\sum_{j=k}^{n-1} 1/(j+1)$ is decreasing in k, when n is large, $k/N(n,k)$ is increasing in k. So, asymptotically, the expected failure proportion is increasingly in k. □

Theorem 1 implies that the 1-failure strategy has the smallest asymptotic expected failure rate among k-failure strategies. The next theorem implies that, asymptotically, the advantage of $k = 1$ is not great: all k-failure rate strategies have the same asymptotic failure proportion.

THEOREM 2. *For any fixed k the expected failure proportion of the k-failure strategy is asymptotically* $1/\ln(n)$.

PROOF. For $n = 1, 2, \dots$, let $\phi(n,k)$ be the expected number of failures in n trials produced by the k-failure strategy. Fix $k \in \{1, 2, \dots\}$. For $n \le k$, $\phi(n,k) = n/2$. For $n \ge k$, $\phi(n,k)$ can be found recursively as follows, where $\phi(0,k) = 0$:

$$\begin{aligned}
\phi(n,k) &= \int_0^1 \sum_{j=1}^{k-1} j\binom{n}{j} u^{n-j}(1-u)^j\,du \\
&\quad + \int_0^1 \sum_{j=k}^{n} [k + \phi(n-j,k)]\binom{j-1}{k-1} u^{j-k}(1-u)^k\,du \\
&= \sum_{j=1}^{k-1} j\binom{n}{j}\frac{(n-j)!\,j!}{(n+1)!} + \sum_{j=k}^{n} k\binom{j-1}{k-1}\frac{(j-k)!\,k!}{(j+1)!} \\
&\quad + \sum_{j=k}^{n} \phi(n-j,k)\binom{j-1}{k-1}\frac{(j-k)!\,k!}{(j+1)!} \\
&= k - \frac{1}{n+1}\binom{k+1}{2} + k\sum_{j=k}^{n}\frac{\phi(n-j,k)}{j(j+1)}.
\end{aligned}$$

Let $G_k(t)$ be the generating function of $\{\phi(n,k)\}_{n\ge1}$; that is,

$$G_k(t) = \sum_{n=0}^{\infty} \phi(n,k)t^n.$$

Then

$$G_k(t) = \sum_{n=0}^{k} \frac{n}{2} t^n + \sum_{n=k+1}^{\infty} \left\{ k - \frac{1}{n+1} \binom{k+1}{2} \right\} t^n + k \sum_{n=k+1}^{\infty} \sum_{j=k}^{n} \frac{\phi(n-j, k)}{j(j+1)} t^n$$
$$= \sum_{n=1}^{k} \frac{n}{2} t^n + \frac{k}{1-t} t^{k+1} - \binom{k+1}{2} \frac{1}{t} \int_0^t \frac{1}{1-u} u^{k+1}\, du$$
$$+ k \sum_{j=k}^{\infty} \sum_{n=j}^{\infty} \phi(n-j, k) \left\{ \frac{1}{j} t^j - \frac{1}{j+1} t^j \right\} t^{n-j}.$$

Hence

$$G_k(t) \left\{ 1 - k \int_0^t \frac{1}{1-u} u^{k-1}\, du + \frac{k}{t} \int_0^t \frac{1}{1-u} u^k\, du \right\}$$
$$= \left\{ \sum_{n=1}^{k} \frac{n}{2} t^n + \frac{k}{1-t} t^{k+1} - \binom{k+1}{2} \frac{1}{t} \int_0^t \frac{1}{1-u} u^{k+1}\, du \right\}.$$

Alternatively,

$$G_k(t) \left\{ 1 - t^{k-1} + \frac{k}{t} (1-t) \int_0^t \frac{1}{1-u} u^{k-1}\, du \right\}$$
$$= \left\{ \sum_{n=1}^{k} \frac{n}{2} t^n + \frac{k}{1-t} t^{k+1} - \binom{k+1}{2} \frac{1}{t} \int_0^t \frac{1}{1-u} u^{k+1}\, du \right\}.$$

Therefore,

$$\frac{(1-t)}{t} G_k(t) \left\{ \sum_{n=0}^{k-1} t^n + k \int_0^t \frac{1}{1-u} u^{k-1}\, du \right\}$$
$$= \frac{1}{t(1-t)} \left\{ (1-t) \sum_{n=1}^{k} \frac{n}{2} t^{n+1} + k t^{k+2} - \binom{k+1}{2} (1-t) \int_0^t \frac{1}{1-u} u^{k+1}\, du \right\}.$$

That is,

$$G_k(t) = \frac{1}{(1-t)^2} \left\{ \frac{H_1(t) + H_2(t)}{J_1(t) + J_2(t)} \right\}.$$

Here,

$$H_1(t) = (1-t) \left\{ \sum_{n=1}^{k} \frac{n}{2} t^{n+1} - \binom{k+1}{2} \int_0^t \frac{1}{1-u} u^{k+1}\, du \right\},$$
$$H_2(t) = k t^{k+2}, \qquad J_1(t) = \sum_{n=0}^{k-1} t^n, \qquad J_2(t) = k \int_0^t \frac{1}{1-u} u^{k-1}\, du.$$

It is easy to see that as $t \to 1-$, $H_1(t)/H_2(t) \to 0$ and $J_1(t)/J_2(t) \to 0$. Therefore,

$$\frac{H_1(t) + H_2(t)}{J_1(t) + J_2(t)} \approx \frac{H_2(t)}{J_2(t)} \approx \frac{1}{\ln(1/(1-t))} \quad \text{as } t \to 1-.$$

Since, for fixed k, $\phi(n,k)$ is increasing in n, it follows by Theorem 5 of Feller [(1971), page 447] that, for any fixed k, $\phi(n,k)/n \approx 1/\ln(n)$ as $n \to \infty$. □

The 1-failure strategy indicates a switch to a new arm whenever the current arm produces a failure. Discarding an arm that has given a very large number of successes and a single failure seems wasteful. One might expect an α-rate strategy to do better for some α. However, the following argument shows that, for any $\alpha > 0$, the α-rate strategy has a positive expected failure proportion asymptotically.

Since F is uniform on $(0,1)$, for any $\alpha > 0$ there is a positive probability that any particular arm has parameter p such that

$$1 - \tfrac{3}{4}\alpha < p < 1 - \tfrac{1}{4}\alpha.$$

By the strong law of large numbers, the failure rate produced by this arm is between $\alpha/2$ and α with positive probability. Hence the proportion of failures when following the α-rate strategy is at least $\alpha/2$ with positive probability. Therefore, the expected failure proportion of the α-rate strategy is greater than a positive constant as $n \to \infty$ for any $\alpha > 0$. Therefore, for any $\alpha > 0$ and sufficiently large n, the α-rate strategy is inferior to the 1-failure strategy.

Is it possible to do better than the asymptotic expected failure proportion of the 1-failure strategy, $1/\ln(n)$? We seek the best lower bound for the expected failure proportion. To this end, define an *m-run strategy* as one that follows the 1-failure strategy until either the current arm has produced a success run of length m or Arm m is used. If the former obtains, then the current arm is used for the remaining trials. If the latter obtains, then the arm with highest proportion of successes among the m arms used so far is used for the remaining trials. So an m-run strategy uses at most m arms; if it uses m arms, then the best performing arm is recalled and used for the duration.

The next two theorems show that $\sqrt{2/n}$ is a lower bound for the expected failure proportion over all strategies and the failure rate of the $\sqrt{n}$-run strategy has the same order of magnitude as this lower bound.

THEOREM 3. *$\sqrt{2/n}$ is a lower bound for the expected failure proportion over all strategies.*

PROOF. Let C be the number of arms actually used when following a particular strategy. Given $C = c$, and since each arm is used at least once, the conditional expected number of failures is greater than or equal to

$$\frac{c}{c+1} + \frac{c-1}{c+1} + \cdots + \frac{1}{c+1} + \frac{n-c}{c+1}.$$

(Imagine that each arm is used once and then the best performing arm of these c arms is used for the remaining $n - c$ trials.) Hence, by Jensen's inequality, the expected number of failures is greater than or equal to

$$\frac{E(C+1)}{2} + \frac{n+1}{E(C+1)} - \frac{3}{2}.$$

Since $a/2 + x/a \geq \sqrt{2x}$, the expected failure proportion is greater than or equal to

$$\frac{\sqrt{2(n+1)}}{n} - \frac{3}{2n} \sim \sqrt{\frac{2}{n}}$$

for any particular strategy. Therefore, $\sqrt{2/n}$ is a lower bound for the expected failure proportion over all strategies. □

THEOREM 4. *The expected failure proportion for the $\sqrt{n}$-run strategy is less than or equal to $2/\sqrt{n}$.*

PROOF. Let C be the number of arms used in the $\sqrt{n}$-run strategy and let T be the corresponding number of failures produced. It is easy to see that

$$E(T) \leq \sqrt{n} + \frac{n}{\sqrt{n}+1} \leq 2\sqrt{n}.$$

Therefore, the expected failure proportion for the $\sqrt{n}$-run strategy is $E(T)/n \leq 2/\sqrt{n}$. □

The $\sqrt{n}$-run strategy is not the only strategy that has expected failure proportion less than or equal to $2/\sqrt{n}$. The following is another.

A strategy is called an *m-learning strategy* if it follows the 1-failure strategy for the first m trials (with the arm selected at trial m used until it yields a failure), and then for the remaining trials it calls for using the arm that performed best during the first m trials. The next theorem shows that a $\ln(n)\sqrt{n}$-learning strategy has expected failure proportion less than or equal to $2/\sqrt{n}$.

THEOREM 5. *The $\ln(n)\sqrt{n}$-learning strategy has expected failure proportion less than or equal to $2/\sqrt{n}$.*

PROOF. Since the expected number of trials to the first failure of each new arm is asymptotically equal to $\ln(n)$, the expected number of arms used during the learning period will be $\ln(n)\sqrt{n}/\ln(n)$. Also, the expected number of failures and the expected number of arms used during this learning period is $\sqrt{n}$. The (expected) probability of failure on the best of these $\sqrt{n}$ arms is $1/(\sqrt{n}+1)$. So the expected number of failures is less than or equal to

$$\sqrt{n} + (n - \ln(n)\sqrt{n})\frac{1}{\sqrt{n}+1} \leq \sqrt{n} + \sqrt{n} = 2\sqrt{n}.$$

Therefore, the expected failure proportion of the $\ln(n)\sqrt{n}$-learning strategy is asymptotically less than or equal to $2/\sqrt{n}$ and the proof of Theorem 5 now is complete. □

The $\sqrt{n}$-run strategy of Theorem 4 and the $\ln(n)\sqrt{n}$-learning strategy of Theorem 5 are recalling strategies. Is there a nonrecalling strategy that is asymptotically as good as these two recalling strategies? The next theorem gives an affirmative answer.

A strategy is a *nonrecalling m-run strategy* if it uses the 1-failure strategy until an arm produces a success run of length m at which time this arm is used for all remaining trials; if no arm produces a success run of length m, the 1-failure strategy is used for all n trials. The following theorem says that the nonrecalling $\sqrt{n}$-run strategy has expected failure proportion asymptotically less than or equal to $2/\sqrt{n}$.

THEOREM 6. *The expected failure proportion of the nonrecalling $\sqrt{n}$-run strategy is less than or equal to $2/\sqrt{n}$ asymptotically.*

PROOF. Let B be the number of arms tried until finding one that produces a success run of length $\sqrt{n}$. If an arm has produced a success run of length $\sqrt{n}$, then this arm is expected to produce no more than $n(1/\sqrt{n})$ failures. Hence the expected number of failures produced by the nonrecalling $\sqrt{n}$-run strategy will be less than or equal to $\sqrt{n} + E(B)$, where

$$E(B) = \sum_{j=1}^{n} jP(B = j) = \sum_{j=1}^{n} P(B \geq j) \leq 1 + \sum_{j=1}^{n} P(B > j).$$

$P(B > j)$ is the probability that none of the first j arms has produced a success run of length $\sqrt{n}$, which is the jth power of the probability that any particular arm has not produced a success run of length $\sqrt{n}$. That is,

$$\begin{aligned} P(B > j) &= \left\{\int_0^1 [(1-u) + u(1-u) + \cdots + u^{\sqrt{n}-1}(1-u)]\,du\right\}^j \\ &= \left\{\int_0^1 (1 - u^{\sqrt{n}})\,du\right\}^j = \left(1 - \frac{1}{\sqrt{n}+1}\right)^j = \left(\frac{\sqrt{n}}{\sqrt{n}+1}\right)^j. \end{aligned}$$

Therefore,

$$\sum_{j=1}^{n} P(B > j) = \sum_{j=1}^{n}\left(\frac{\sqrt{n}}{\sqrt{n}+1}\right)^j \leq \sum_{j=1}^{\infty}\left(\frac{\sqrt{n}}{\sqrt{n}+1}\right)^j = \sqrt{n}.$$

Therefore, the expected number of failures produced by the nonrecalling $\sqrt{n}$-run strategy is less than $2\sqrt{n} + 1$ and the expected failure proportion is less than or equal to $2/\sqrt{n}$ asymptotically. □

Based on Theorems 3, 4, 5 and 6, we have the following corollary.

COROLLARY 1. *The best lower bound for the expected failure proportion over all strategies is between $\sqrt{2/n}$ and $2/\sqrt{n}$.*

REMARK. We suspect that the best lower bound for the expected failure proportion over all strategies is $2/\sqrt{n}$. However, we do not have a proof.

3. Uniform distribution on $[a, b]$. In this section we investigate the situation in which the prior distribution F is uniform over a subinterval $[a, b]$ of the unit interval.

Using the same argument as for Theorem 3, we can show that the expected failure proportion is greater than or equal to $(1-b)+(b-a)(\sqrt{2/n}-3/2n)$. Therefore, $(1-b)+(b-a)\sqrt{2/n}$ is a lower bound for the expected failure proportion over all strategies for the assumed F.

THEOREM 7. *If the distribution F is uniform over subinterval $[a,b]$ of the unit interval, then $(1-b)+(b-a)\sqrt{2/n}$ is a lower bound for the expected failure proportion over all strategies.*

With little modification, the arguments for Theorems 4–6 apply to the current situation. Hence we have the following theorems.

THEOREM 8. *The expected failure proportion of the $\sqrt{n(b-a)}$-run strategy is less than or equal to $(1-b)+2\sqrt{(b-a)/n}$ asymptotically.*

THEOREM 9. *The expected failure proportion of the $\sqrt{n(b-a)}\ln(n(b-a))$-learning strategy is less than or equal to $(1-b)+2\sqrt{(b-a)/n}$ asymptotically.*

THEOREM 10. *The expected failure proportion of the nonrecalling $\sqrt{n(b-a)}$-run strategy is less than or equal to $(1-b)+2\sqrt{(b-a)/n}$.*

REMARK. We suspect that the best lower bound for the expected failure proportion is $(1-b)+2\sqrt{(b-a)/n}$. However, we do not have a proof.

4. Arbitrary prior distributions.

In this section, we briefly discuss the case in which F is an arbitrary distribution on the interval $[0,1]$. We assume that F is continuous, $F(0)=0$ and $F(1)=1$.

Suppose that the number of arms used over the course of the n trials is C. Then, given $C=c$, there will be at least

$$\begin{aligned} G(c) &= c\int_0^1 (1-\alpha)\,dF(\alpha)+(n-c)\int_0^1(1-\alpha)\,dF^c(\alpha)\\ &= c\int_0^1 F(\alpha)\,d\alpha+(n-c)\int_0^1 F^c(\alpha)\,d\alpha \end{aligned}$$

(on the conditional space) expected failures. (Imagine that each arm is used only once; then an oracle tells us which of these c arms is the best arm and we use this best arm for the remaining $n-c$ trials.) Since

$$G''(c) = -2\int_0^1 F^c(\alpha)\ln F(\alpha)\,d\alpha+(n-c)\int_0^1 F^c(\alpha)[\ln F(\alpha)]^2\,d\alpha > 0$$

if $c<n$, $G(c)$ is a convex function. By Jensen's inequality,

$$E\{G(C)\}\geq G(E(C)) = E(C)\int_0^1 F(\alpha)\,d\alpha+(n-E(C))\int_0^1 F^{E(C)}(\alpha)\,d\alpha.$$

Since $G'(1) < 0$, $G(1) = G(n) = n\int_0^1 F(\alpha)\,d\alpha$, and $G''(c) > 0$ if $c < n$, there exists a positive integer c_n such that

$$1 < c_n < n \quad \text{and} \quad G(c_n) = \min_{1<c<n} G(c).$$

Therefore, we have the following theorem.

THEOREM 11.
$$\frac{G(c_n)}{n} = \frac{1}{n}\left\{c_n \int_0^1 F(\alpha)\,d\alpha + (n - c_n)\int_0^1 F^{c_n}(\alpha)\,d\alpha\right\}$$
is a lower bound for the expected failure proportion over all strategies.

For $k = 0, 1, 2, \dots, n$, let

$$H(k) = k + (n-k)\int_0^1 F^k(\alpha)\,d\alpha.$$

We have that $H(0) = H(n) = n$,

$$H'(0) = n\int_0^1 \ln(F(\alpha))\,d\alpha < 0,$$

$$H'(n) = 1 - \int_0^1 F^n(\alpha)\,d\alpha > 0,$$

$H''(k) > 0$ for $0 \le k < n$. Therefore there exists a positive integer k_n such that $H(k_n) = \min_{0\le k\le n} H(k)$.

Let $h(0) = 0$, and for each $i = 1, 2, \dots, n$ let

$$e(i) = \sum_{j=0}^{n-h(i-1)} \int_0^1 (1-\alpha)\alpha^j\,dF(\alpha)$$

and $h(i) = h(i-1) + e(i)$. Define

$$M(i) = i + (n - h(i))\int_0^1 F^i(\alpha)\,d\alpha$$

for all $i = 0, 1, 2, \dots$ and $h(i) \le n$. Since h is increasing and $e(i)$ is decreasing, $M(i)$ is a convex function. Since $M(0) = n$ and $M(i^*) = n$ if $h(i^*) = n$, there exists a positive integer i_n such that $M(i_n) = \min_{0\le i,\, h(i)\le n} M(i)$.

When n is large, i_n is small, and $e(1) \approx e(2) \approx \cdots \approx e(i_n)$. Asymptotically we can write

$$M(i) = i + (n - ie(1))\int_0^1 F^i(\alpha)\,d\alpha.$$

Then we can find an i_n^* such that $M(i_n^*) = \min_{0\le i\le n/e(1)} M(i)$.

Let B_u be the number of arms tried until finding one that produces a success run of length u. If an arm has produced a success run of length u, then the expected number of failures on this arm is no more than

$$m\left\{1 - \int_0^1 \alpha^{u+1}\,dF(\alpha) \Big/ \int_0^1 \alpha^u\,dF(\alpha)\right\},$$

where m is the number of remaining trials. Hence the expected number of failures produced by the nonrecalling u-run strategy is less than or equal to

$$E(B_u) + \{n - E(B_u)E(W_u)\}\left\{1 - \int_0^1 \alpha^{u+1}\, dF(\alpha) \Big/ \int_0^1 \alpha^u\, dF(\alpha)\right\},$$

where W_u is the number of trials when following a 1-failure strategy that does not produce a success run of length u; $E(B_u)$ and $E(W_u)$ are easy to compute. Now we define

$$N(u) = E(B_u) + \{n - E(B_u)E(W_u)\}\left\{1 - \int_0^1 \alpha^{u+1}\, dF(\alpha) \Big/ \int_0^1 \alpha^u\, dF(\alpha)\right\}$$

for each $u = 1, 2, \dots$ such that $E(B_u)E(W_u) < n$. By a similar argument, we can show that, for each n, there exists a u_n such that $N(u_n) = \min_{0<u<n} N(u)$.

With slight modifications of the proofs of Theorem 4 and 5, we have the following theorems.

THEOREM 12. *The expected failure proportion for the k_n-run strategy is less than or equal to $H(k_n)/n$ asymptotically.*

THEOREM 13. *The $e(1)i_n^*$-learning strategy has expected failure proportion less than or equal to $M(i_n^*)/n$ asymptotically.*

By a slight modification of Theorem 6, we have the following theorem.

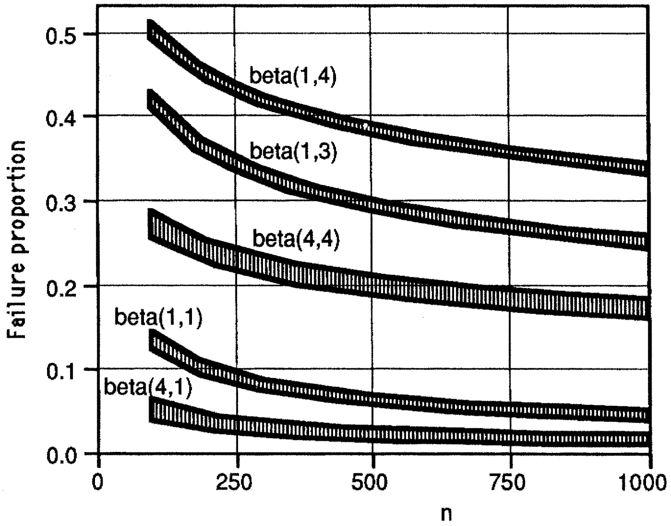

FIG. 1. *Lower and upper bounds (and possible range, shown shaded) of failure proportion depending on n (shown for n between* 100 *and* 1000*) for each of five distributions F, as labeled. The lower bound is $G(c_n)/n$ and the upper bound is $M(i_n^*)/n$.*

THEOREM 14. *The expected failure proportion of the nonrecalling u_n-run strategy is less than or equal $N(u_n)/n$ asymptotically.*

It is difficult to find the minimum values of G, H, M and N analytically. However, numerical investigations are straightforward. All three strategies considered in this section perform well in the sense that the asymptotic expected failure proportions are close to the lower bound. Figure 1 shows graphs of the lower bound $G(c_n)/n$ and the upper bound $M(i_n^*)/n$ when F is one of the five indicated beta distributions and for n varying from 100 to 1000. In all cases considered in Figure 1, $M(i_n^*)/n$ is the smallest of the three upper bounds presented in this section.

REFERENCES

BANKS, J. S. and SUNDARAM, R. K. (1992). Denumerable-armed bandits. *Econometrica* **60** 1071–1096.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocations of Experiments.* Chapman and Hall, London.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd ed. Wiley, New York.

GITTINS, J. C. (1989). *Multi-armed Bandit Allocation Indices*. Wiley, New York.

HERSCHKORN, S. J., PEKOZ, E. and ROSS, S. M. (1995). Policies without memory for the infinite-armed Bernoulli bandit under the average-reward criterion. *Probab. Engrg. Inform. Sci.* **10** 21–28.

JOVANOVIC, B. (1979). Job-search and the theory of turnover. *Journal of Political Economy* **87** 972–990.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–536.

WHITTLE, P. (1982, 1983). *Optimization over Time* **1**, **2**. Wiley, New York.

WILDE, L. (1979). An information-theoretic approach to job quits. In *Studies in the Economics of Search* (S. Lippman and J. McCall, eds.) 35–52. North-Holland, Amsterdam.

VISCUSI, W. (1979). Job-hazards and worker quit rates: an analysis of adaptive worker behavior. *Internat. Econ. Rev.* **20** 29–58.

DONALD A. BERRY
INSTITUTE OF STATISTICS
AND DECISION SCIENCES
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708-0251
E-MAIL: db@isds.duke.edu

ROBERT W. CHEN
ALAN ZAME
DEPARTMENT OF MATHEMATICS
AND COMPUTER SCIENCE
UNIVERSITY OF MIAMI
CORAL GABLES, FLORIDA 33124

DAVID C. HEATH
SCHOOL OF OR&IE
RHODES HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801
E-MAIL: davidh@orie.cornell.edu

LARRY A. SHEPP
DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
NEW BRUNSWICK, NEW JERSEY
E-MAIL: las@research.att.com