ELSEVIER

Journal of Economic Dynamics & Control
27 (2002) 87–108

JOURNAL OF
Economic
Dynamics
& Control

www.elsevier.com/locate/econbase

# Optimal learning and experimentation in bandit problems

Monica Brezzi[a], Tze Leung Lai[b,*,1]

[a] *Dipartimento per le politiche di sviluppo e coesione, Ministero del Tesoro Via Nerva 1-00187 Rome, Italy*

[b] *Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305-4065, USA*

Received 10 January 2000; accepted 2 April 2001

**Abstract**

This paper studies how and how much active experimentation is used in discounted or finite-horizon optimization problems with an agent who chooses actions sequentially from a finite set of actions, with rewards depending on unknown parameters associated with the actions. Closed-form approximations are developed for the optimal rules in these 'multi-armed bandit' problems. Some refinements and modifications of the basic structure of these approximations also provide a nearly optimal solution to the long-standing problem of incorporating switching costs into multi-armed bandits.
© 2002 Elsevier Science B.V. All rights reserved.

*JEL classification:* C44; C63; D83

*Keywords*: Optimal stopping; Corrected binomial algorithm; Multi-armed bandits; Switching costs; Incomplete learning

## 1. Introduction

In many situations, rational economic agents face the dilemma between the objective of reward maximization and the need for experimentation with

* Corresponding author. Tel.: +1-650-723-2622; fax: +1-650-725-8977.
*E-mail address:* lait@stat.stanford.edu (T.L. Lai).
[1] Research supported by the National Science Foundation and the Center for Advanced Study in the Behavioral Sciences.

0165-1889/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0165-1889(01)00028-8

potentially suboptimal actions to learn about their expected rewards. Prototypical examples are multi-armed bandit problems, in which an agent chooses actions sequentially from a finite set $\{a_1,\ldots,a_k\}$ such that the reward $R(a_j)$ of action $a_j$ has a probability distribution depending on an unknown parameter $\theta_j$ which has a prior distribution $\Pi^{(j)}$. The agent's objective is to maximize the total discounted reward

$$\int\cdots\int \mathrm{E}_{\theta_1,\ldots,\theta_k}\left\{\sum_{t=0}^{\infty}\beta^t R(X_{t+1})\right\}\mathrm{d}\Pi^{(1)}(\theta_1)\ldots\mathrm{d}\Pi^{(k)}(\theta_k), \tag{1}$$

where $0<\beta<1$ is a discount factor and $X_t$ denotes the action chosen by the agent at time $t$. The optimal solution to this problem, commonly called the 'discounted multi-armed bandit problem', was shown by Gittins and Jones (1974) and Gittins (1979) to be the 'index rule' that chooses at each stage the action with the largest 'dynamic allocation index' (DAI). In Section 2 a precise definition of the DAI of action $a_j$ at stage $t$ is given, and it is a complicated function of the posterior distribution of $\theta_j$ given the rewards, up to stage $t$, at the times when action $a_j$ is used. We develop in Section 2 a simple and easily interpretable approximation of the DAI. It is based on numerical solution of an optimal stopping problem for a limiting diffusion. A computational method to solve this optimal stopping problem, which has been studied analytically via free boundary problems for the heat equation and integral representations by Chang and Lai (1987) and Brezzi and Lai (1999), is also given.

In the finite-horizon version of bandit problems, the agent's objective is to maximize the total reward

$$\int\cdots\int \mathrm{E}_{\theta_1,\ldots,\theta_k}\left\{\sum_{t=0}^{N-1}R(X_{t+1})\right\}\mathrm{d}\Pi(\theta_1,\ldots,\theta_k), \tag{2}$$

where $\Pi$ is a prior distribution of the vector $(\theta_1,\ldots,\theta_k)$. Even when the $\theta_i$ are independent under $\Pi$ (so that $\Pi$ is a product of marginal distributions as in (1)), the optimal rule that maximizes (2) does not reduce to an index rule. In principle, one can use dynamic programming to maximize (2). For the case $k=2$ and $(\theta_1,\theta_2)\in\{(\alpha,\gamma),(\gamma,\alpha)\}$, where $\alpha$ and $\gamma$ are known numbers with $\mu(\alpha)>\mu(\gamma)$, Feldman (1962) found by this approach that the optimal rule chooses $a_1$ or $a_2$ at stage $n+1$ if $\pi_n^{(1)}\geq 1/2$ or $\pi_n^{(1)}<1/2$, where $\pi_n^{(1)}$ is the posterior probability in favor of $(\alpha,\gamma)$ at the end of stage $n$. In the case of $k=2$ Bernoulli populations with independent Beta priors for their parameters, Fabius and van Zwet (1970) and Berry (1972) studied the dynamic programming equations analytically and obtained several qualitative results concerning the optimal rule. Beyond the two-point priors considered by Feldman, optimal rules in the finite-horizon multi-armed bandit problem are defined only implicitly by the dynamic programming equations, whose numerical solution becomes formidable for large horizon $N$. Refining the

earlier work of Lai and Robbins (1985), Lai (1987) showed that although index rules do not provide exact solutions to the optimization problem (2), they are asymptotically optimal as $N \to \infty$, and have nearly optimal performance from both the Bayesian and frequentist viewpoints for moderate and small values of $N$. Section 3 gives a brief review of these nearly optimal index rules in the finite-horizon case, which are easily implementable and whose indices can be interpreted as certain upper confidence bounds for the expected rewards of $a_1, \ldots, a_k$. Making use of these simple approximations to the optimal policy, we analyze the value of experimentation in Section 3, where our results show that unless the horizon $N$ or the discount factor $\beta$ is large enough, experimentation does not have much value since the optimal rule that involves active learning by experimentation has little improvement over the myopic rule that chooses the action with the largest posterior mean reward. On the other hand, Section 3 also shows that for large horizon $N$ or for $\beta$ close to 1, the inefficiency of the myopic rule due to inadequate learning is much more pronounced.

The theory of multi-armed bandits has been applied to pricing under demand uncertainty (cf. Rothschild, 1974), decision making in labor markets (cf. Jovanovich, 1979; Mortensen, 1985), general search problems (cf. Gittins, 1989; Banks and Sundaram, (1992)), and resource allocation among competing projects (cf. Gittins, 1989). Banks and Sundaram (1994) have pointed out the need to incorporate switching costs into bandit problems, since 'it is difficult to imagine a relevant economic decision problem in which the decision-maker may costlessly move between alternatives'. They show that unfortunately it is not possible to define indices which have the property that the resulting index strategy is optimal when there are switching costs. In Section 4 we assume a cost of switching from one arm to another in multi-armed bandits. Although the index rules in Sections 2 and 3 are no longer applicable, they provide important insights into how and how much active experimentation is used by optimal rules to generate information about the unknown parameters of the different actions. Making use of these insights, we develop in Section 4 nearly optimal and easily implementable procedures for bandit problems with switching costs. Some concluding remarks are given in Section 5.

## 2. Gittins indices in discounted multi-armed bandits

As pointed out in Section 1, the optimal solution to the discounted $k$-armed bandit problem (1) is the 'index rule' that chooses at each stage the action with the largest dynamic allocation index (also called the 'Gittins index'). Specifically, at stage $t$, let $n_t(j) = \sum_{i=1}^{t} \mathbf{1}_{\{X_i = a_j\}}$ denote the total number of times that action $a_j$ has been used so far, and let $\Pi^{(j)}_{n_t(j)}$ denote the posterior

distribution of $\theta_j$ based on $Y_{j,1},\ldots,Y_{j,n_t(j)}$, where the $Y_{j,i}$ denote the rewards at the successive times when $a_j$ is used. The index rule chooses at stage $t$ the action $a_j$ with the largest $v(\Pi^{(j)}_{n_t(j)})$, where $v(\cdot)$ is the Gittins index, defined by (3) below, associated with the posterior distribution. Here and in the sequel, $\mathbf{1}_A$ denotes the indicator variable of an event $A$ (i.e., $\mathbf{1}_A = 1$ or 0 depending on whether $A$ occurs or not).

Let $R(a_j)$ have distribution function $F_{\theta_j}$ (depending on the unknown parameter $\theta_j$) so that $Y_{j,1}, Y_{j,2},\ldots$ are independent random variables with common distribution function $F_{\theta_j}$. Let $\Pi^{(j)}$ be a prior distribution on $\theta_j$. The Gittins index $v(\Pi^{(j)})$ associated with $\Pi^{(j)}$ is defined as

$$v(\Pi^{(j)}) = \sup_{\tau}\left\{\frac{\int \mathrm{E}_{\theta_j}\left(\sum_{i=0}^{\tau-1}\beta^i Y_{j,i+1}\right)\,\mathrm{d}\Pi^{(j)}(\theta_j)}{\int \mathrm{E}_{\theta_j}\left(\sum_{i=0}^{\tau-1}\beta^i\right)\,\mathrm{d}\Pi^{(j)}(\theta_j)}\right\}, \tag{3}$$

where the supremum is over all stopping times $\tau \geq 1$ defined on $\{Y_{j,1}, Y_{j,2},\ldots\}$ (cf. Gittins, 1979). As is well known, the conditional distribution of $(\theta_j, Y_{j,n+1}, Y_{j,n+2},\ldots)$ given $(Y_{j,1},\ldots,Y_{j,n})$ can be described by that $Y_{j,n+1}, Y_{j,n+2},\ldots$ which are independent having common distribution function $F_{\theta_j}$ and that $\theta_j$ has distribution $\Pi^{(j)}_n$, which is the posterior distribution of $\theta_j$ given $(Y_{j,1},\ldots,Y_{j,n})$. Letting $\mu(\theta_j) = \mathrm{E}_{\theta_j}(R(a_j))$, it then follows that for $m > n$,

$$\mathrm{E}[Y_{j,m}|Y_{j,1},\ldots,Y_{j,n}] = \int \mu(\theta_j)\,\mathrm{d}\Pi^{(j)}_n(\theta_j) = \mathrm{E}[\mu(\theta_j)|Y_{j,1},\ldots,Y_{j,n}]. \tag{4}$$

The Gittins index (3) of $\Pi^{(j)}$ can be equivalently defined as the infimum of the set of solutions $M$ of the equation

$$\sup_{\tau}\int E_{\theta_j}\left\{\sum_{n=0}^{\tau-1}\beta^n\int \mu(\theta_j)\,\mathrm{d}\Pi^{(j)}_n(\theta_j) + M\sum_{n=\tau}^{\infty}\beta^n\right\}\,\mathrm{d}\Pi^{(j)}(\theta_j) = M\sum_{n=0}^{\infty}\beta^n, \tag{5}$$

where we set $\Pi^{(j)}_0 = \Pi^{(j)}$ (cf. Whittle, 1980).

Chapter 7 of Gittins (1989) describes computational methods to calculate Gittins indices for normal, Bernoulli and exponential $F_\theta$, with the prior distribution of $\theta$ belonging to a conjugate family. These methods involve approximating the infinite horizon in the optimal stopping problem (5) by a finite horizon $N$ and using backward induction. When $\beta$ is near 1, a good approximation requires a very large $N$ and becomes computationally prohibitive. In this case, we can get around the computational difficulties by using a diffusion approximation, which involves the Gittins index for a Wiener process and is described in the next three subsections.

### 2.1. Gittins index for a Wiener process

Let $w(t), t \geq 0$, be a Wiener process with drift coefficient $\theta$ which has a normal distribution with mean $u_0$ and variance $v_0$. The posterior distribution of $\theta$ given $\{w(s), s \leq t\}$ is normal with variance $v_t$ satisfying $v_t^{-1} = v_0^{-1} + t$ and mean $u_t = v_t\{w(t) + u_0/v_0\}$. In analogy with (5), where we set $\beta = e^{-c}$, define the Gittins index $M_c(u_0, v_0)$ as the infimum of the set of solutions $M$ of the equation

$$\sup_{T \geq 0} E_{u_0,v_0} \left\{ \int_0^T u_t e^{-ct}\, dt + M \int_T^\infty e^{-ct}\, dt \right\} = M \int_0^\infty e^{-ct}\, dt, \tag{6}$$

where the supremum is taken over all stopping times $T \geq 0$. Under the change of variables

$$v = (v_0^{-1} + t)^{-1}, \quad Y(v) = u_t - u_0, \quad s = v/c, \quad Z(s) = Y(cs)/\sqrt{c},$$

$$z_0 = (u_0 - M)/\sqrt{c}, \quad s_0 = v_0/c, \tag{7}$$

$\{Z(s), 0 < s \leq s_0\}$ is a Brownian motion in the $-s$ scale and (6) can be rewritten as

$$z_0 e^{-1/s_0} = \inf_S E[Z(S)e^{-1/S} | Z(s_0) = z_0], \tag{8}$$

where $\inf_S$ is over all stopping times (in the $-s$ scale) of the Brownian motion with initial value $Z(s_0) = z_0$. The optimal stopping rule $S^*$ that attains the infimum in the right-hand side of (8) has a continuation region $\mathscr{C}$ of the form

$$\mathscr{C} = \{(z, s): z > -b(s)\}, \tag{9}$$

in which $b(\cdot)$ is a nonnegative function such that

$$\begin{aligned} b(s) &= (2^{-1/2} + o(1))s \qquad \text{as } s \to 0, \\ &= \{2s[\log s - \tfrac{1}{2}\log\log s - \tfrac{1}{2}\log 16\pi + o(1)]\}^{1/2} \quad \text{as } s \to \infty, \end{aligned} \tag{10}$$

see Chang and Lai (1987). Since (8) is equivalent to $z_0 = -b(s_0)$ (i.e., $(z_0, s_0)$ belongs to the boundary of $\mathscr{C}$), the Gittins index can be represented via (7) as

$$M_c(u_0, v_0) = u_0 + \sqrt{c}\, b(v/c). \tag{11}$$

We can therefore determine the values of the Gittins indices $M_c(\cdot, \cdot)$ by computing the optimal stopping boundary $-b(\cdot)$.

### 2.2. Numerical computation of the optimal stopping boundary

To compute the optimal stopping boundary for the Brownian motion $Z(s)$ (in the $-s$ scale) corresponding to the loss function $L(s, z) = ze^{-1/s}$, we use

the corrected binomial method due to Chernoff and Petkau (1986) together with a representation of the optimal value function given in Brezzi and Lai (1999) to initialize the algorithm. Letting $\lambda(s,z) = \inf_S \mathrm{E}[L(S, Z(S))|Z(s) = z]$ be the optimal value function, the procedure is described below.

The basic idea of the Chernoff–Petkau method is to approximate Brownian motion (in the $-s$ scale) by a symmetric Bernoulli random walk with time increment $-\delta$ and space increment $\sqrt{\delta}Y_i$, where the $Y_i$ are independent Bernoulli random variables with $P(Y_i = 1) = 1/2 = P(Y_i = -1)$ so that $\lambda$ can be approximated by the recursion

$$\lambda(s_i, z) = \min\{L(s_i, z), [\lambda(s_{i-1}, z + \sqrt{\delta}) + \lambda(s_{i-1}, z - \sqrt{\delta})]/2\} \tag{12}$$

with $s_i = i\delta$ (so that $s_i - \delta = s_{i-1}$) and $z \in \mathbf{Z}_\delta := \{\sqrt{\delta}n: n \text{ is an integer}\}$. The subtle point in the present problem is that because $L$ converges to 0 exponentially fast as $s \to 0$, initializing the recursion (12) at $s_0 (= 0)$ with the obvious boundary condition $\lambda(0, z) = 0$ leads to numerical difficulties due to finite-precision arithmetic. We can get around these difficulties by initializing at some $s_{i_0} > 0$ and using the following representation of $\lambda$ derived in Brezzi and Lai (1999):

$$\begin{aligned} \lambda(s,z) &= \int_0^s \mathrm{E}\{g(t, z + \sqrt{s-t}Z)\mathbf{1}_{\{z+\sqrt{s-t}Z \le -b(t)\}}\}\,\mathrm{d}t \quad \text{if } z > -b(s), \\ &= L(s,z) \quad \text{if } z \le -b(s), \end{aligned} \tag{13}$$

where $g(s,z) = ((1/2)\partial^2/\partial z^2 - \partial/\partial s)L(s,z) = -s^{-2}z\mathrm{e}^{-1/s}$ and $Z$ is a standard normal random variable. Although representation (13) involves the optimal stopping boundary $-b(\cdot)$ which is to be determined, we can use the approximation $b(t) \doteq t/\sqrt{2}$ given by (10) for $t \le s$ when $s = s_{i_0} (= i_0\delta)$ is small. The integrand in (13) can be expressed in terms of the standard normal density and distribution functions and the integral can be evaluated by numerical integration. In our implementation, we take $\delta = 3 \times 10^{-5}$ and $s_{i_0} = 5 \times 10^{-3}$.

Each point $z \in \mathbf{Z}_\delta$ can be determined to be a stopping or continuation point at time $s_i$ depending on whether $\lambda(s_i, z) = L(s_i, z)$ or $\lambda(s_i, z) < L(s_i, z)$. We use the following continuity correction, proposed by Chernoff and Petkau (1986), to compute the optimal stopping boundary for the Brownian motion. Let $b_\delta(s_i) = \max\{z \in \mathbf{Z}_\delta: \lambda(s_i, z) = L(s_i, z)\}$, $b_{\delta,0}(s_i) = b_\delta(s_i) + \sqrt{\delta}, b_{\delta,1}(s_i) = b_\delta(s_i) + 2\sqrt{\delta}$, and define

$$D_j(x_i) = L(s_i, b_{\delta,j}(s_i)) - \lambda(s_i, b_{\delta,j}(s_i)) \quad \text{for } j = 0, 1.$$

The continuity correction involves $\sqrt{\delta}$, $D_0(s_i)$ and $D_1(s_i)$, and subtracting it from $b_{\delta,0}(s_i)$ yields

$$b(s_i) = b_{\delta,0}(s_i) - \sqrt{\delta}|D_1(s_i)/\{2D_1(s_i) - 4D_0(s_i)\}|. \tag{14}$$

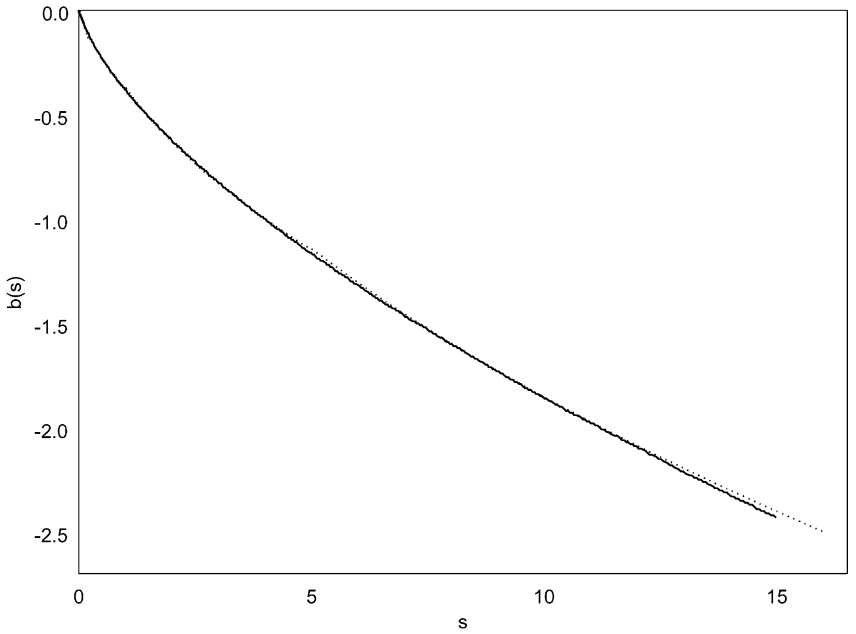


Fig. 1. Optimal stopping boundary (solid curve) and its approximation (dotted curve).

### 2.3. Closed-form approximations to Gittins indices

Fig. 1 plots the optimal stopping boundary computed by the above method. The plot and the asymptotic behavior (10) suggest the closed-form approximation $b(s) \doteq \sqrt{s}\psi(s)$, where

$$\psi(s) = \begin{cases} \sqrt{s/2} & \text{if } s \le 0.2, \\ 0.49 - 0.11s^{-1/2} & \text{if } 0.2 < s \le 1, \\ 0.63 - 0.26s^{-1/2} & \text{if } 1 < s \le 5, \\ 0.77 - 0.58s^{-1/2} & \text{if } 5 < s \le 15, \\ \{2\log s - \log\log s - \log 16\pi\}^{1/2} & \text{if } s > 15. \end{cases}$$

The approximation $\sqrt{s}\psi(s)$ is also plotted in Fig. 1 (dotted curve) and is in good agreement with the solid curve representing $b(s)$. Putting this approximation in (11) yields the following approximation to the Gittins index for the Wiener process $w(t)$:

$$M_c(u_0, v_0) \doteq u_0 + \sqrt{v_0}\psi(v_0/c). \tag{15}$$

We now use (15) to provide closed-form approximations to the Gittins index $\nu(\Pi_n^{(j)})$, defined in (3), of the posterior distribution $\Pi_n^{(j)}$ of $\theta_j$ given

$Y_{j,1},\ldots,Y_{j,n}$. Let $\mu_{j,n}$ and $v_{j,n}$ denote the mean and variance, respectively, of $\Pi_n^{(j)}$. Under mild regularity conditions, $\Pi_n^{(j)}$ is asymptotically normal $N(\mu_{j,n}, v_{j,n})$ as $n \to \infty$, with probability 1 (cf. LeCam, 1953). Let $\sigma^2(\theta_j)$ be the variance of $F_{\theta_j}$, which may depend on the unknown parameter $\theta_j$, and let $\beta = e^{-c}$. Then the functional central limit theorem can be used to show that as $n \to \infty$ and $\beta \to 1$ (or equivalently $c \to 0$), $v(\Pi_n^{(j)}) - \mu_{j,n}$ is asymptotically equivalent to $\sqrt{c}\sigma(\mu_{j,n})M_c(0, v_{j,n}/c\sigma^2(\mu_{j,n}))$, where $M_c(u_0, v_0)$ is the Gittins index of the prior normal distribution $N(u_0, v_0)$ for the drift coefficient of a Wiener process (as defined in Section 2.1). The approximation (15) for $M_c(\cdot,\cdot)$ therefore yields

$$v(\Pi_n^{(j)}) \doteq \mu_{j,n} + \sqrt{v_{j,n}}\psi(v_{j,n}/c\sigma^2(\mu_{j,n})) \tag{16}$$

for large $n$ and small $c$.

*Example 1.* Let $Y_1, Y_2, \ldots$ be independent random variables from a Bernoulli distribution whose parameter $p$ has a Beta prior distribution. Then the posterior distribution of $p$ is also a Beta distribution. Table 1 gives the absolute value of the difference between the Gittins index of a Beta$(a,b)$ distribution, computed by Gittins (1989) in his Table 8, and the approximation (16), for

Table 1
Difference (absolute value) between the Gittins index and its approximation for a Beta $(a,b)$ prior distribution when $\beta = 0.8$

| $a$ | 12 | 14 | 16 | 18 | 20 | Index | |
|---|---|---|---|---|---|---|---|
| $b$ | | | | | | min | max |
| 2 | 0.0111 | 0.0105 | 0.0100 | 0.0093 | 0.0087 | 0.8756 | 0.9183 |
| 4 | 0.0072 | 0.0072 | 0.0074 | 0.0069 | 0.0064 | 0.7730 | 0.8463 |
| 6 | 0.0060 | 0.0059 | 0.0061 | 0.0060 | 0.0059 | 0.6901 | 0.7836 |
| 8 | 0.0057 | 0.0059 | 0.0056 | 0.0054 | 0.0053 | 0.6226 | 0.7291 |
| 10 | 0.0059 | 0.0054 | 0.0054 | 0.0053 | 0.0051 | 0.5666 | 0.6814 |
| 12 | 0.0056 | 0.0054 | 0.0053 | 0.0052 | 0.0049 | 0.5195 | 0.6394 |
| 14 | 0.0054 | 0.0049 | 0.0047 | 0.0049 | 0.0047 | 0.4797 | 0.6021 |
| 16 | 0.0052 | 0.0047 | 0.0050 | 0.0047 | 0.0043 | 0.4455 | 0.5689 |
| 18 | 0.0049 | 0.0049 | 0.0048 | 0.0045 | 0.0041 | 0.4158 | 0.5390 |
| 20 | 0.0046 | 0.0045 | 0.0043 | 0.0041 | 0.0044 | 0.3897 | 0.5120 |
| 22 | 0.0044 | 0.0041 | 0.0045 | 0.0042 | 0.0039 | 0.3666 | 0.4877 |
| 24 | 0.0041 | 0.0043 | 0.0040 | 0.0037 | 0.0040 | 0.3460 | 0.4656 |
| 26 | 0.0043 | 0.0040 | 0.0036 | 0.0038 | 0.0041 | 0.3276 | 0.4433 |
| 28 | 0.0039 | 0.0041 | 0.0037 | 0.0040 | 0.0036 | 0.3111 | 0.4268 |
| 30 | 0.0041 | 0.0037 | 0.0039 | 0.0036 | 0.0037 | 0.2961 | 0.4097 |
| 32 | 0.0037 | 0.0038 | 0.0035 | 0.0036 | 0.0033 | 0.2825 | 0.3938 |
| 34 | 0.0037 | 0.0034 | 0.0036 | 0.0038 | 0.0033 | 0.2701 | 0.3792 |
| 36 | 0.0033 | 0.0035 | 0.0036 | 0.0033 | 0.0034 | 0.2587 | 0.3656 |
| 38 | 0.0034 | 0.0036 | 0.0032 | 0.0033 | 0.0035 | 0.2482 | 0.3529 |
| 40 | 0.0034 | 0.0032 | 0.0033 | 0.0034 | 0.0029 | 0.2386 | 0.3411 |

$\beta = 0.8$. To compare the magnitude of each difference with the actual index, we also summarize the range of Gittins indices for each row of Table 1. As shown in the tables of Gittins (1989), the Gittins index of Beta$(a, b)$ is an increasing function of $a$ for each fixed value of $b$. In order to not present an overly large table with similar results, we only consider in Table 1 even values of $a$ and $b$ on page 226 of Gittins (1989), c whose Table 8 consists of four pages. The approximation differs from the index by less than 1.5%. Moreover, with the exception of $b = 2$, for which the mean $a/(a + b)$ is close to the Gittins index when $a$ lies in the range considered in Table 1, incorporating the second summand in (16) reduces the approximation error of the simple approximation using only the first summand (which is equal to $a/(a+b)$ in the present example) by a factor between 51% and 75%. Similar results, not reported in Table 1, also hold for other values of $a$ and $b$ in Table 8 of Gittins (1989) and for his other tables dealing with $\beta = 0.9$, 0.95 and 0.99. This shows that (16) provides good closed-form approximations to Gittins indices for $\beta \geq 0.8$ and $\min(a, b) \geq 4$. There is also closer agreement between the Gittins index and the approximation (16) as $\beta$ increases through these values.

## 3. The value of experimentation

Brezzi and Lai (2000) give the following upper and lower bounds for $v(\Pi_n^{(j)})$ involving only $\mu_{j,n}$ and the posterior variance $v_{j,n}$:

$$\mu_{j,n} \leq v(\Pi_n^{(j)}) \leq \mu_{j,n} + \beta\sqrt{v_{j,n}}\,/(1-\beta) \tag{17}$$

and use these bounds to give a simple proof of the incomplete learning theorem for $k$-armed bandits with $k \geq 2$: With positive probability, the optimal rule chooses the optimal action only a finite number of times and it can estimate consistently only one of the $\theta_j$. This generalizes the results of Rothschild (1974), McLennan (1984) and Banks and Sundaram (1992). Since the optimal rule is an index rule that chooses at stage $t$ the action $a_j$ with the largest $v(\Pi_{N_t(j)}^{(j)})$, (16) and (17) show the extent of experimentation in the optimal rule. When $\beta$ is small, (17) shows that $v(\Pi_n^{(j)})$ differs little from the posterior mean $\mu_{j,n}$, so the index rule has very little active experimentation. On the other hand, as $\beta(=\mathrm{e}^{-c}) \to 1, \psi(v_{j,n}/c\sigma^2(\mu_{j,n})) \to \infty$ and (16) shows that the difference between $\mu_{j,n}$ and $v(\Pi_n^{(j)})$ becomes infinite, suggesting continued experimentation with $a_j$ to reduce the variance of the posterior distribution of $\theta_j$.

The expected total discounted reward (1) involves an infinite series of rewards and is not easy to compute directly by Monte Carlo simulations for the performance analysis of different allocation strategies. We have to first

replace the infinite series $\sum_{t=0}^{\infty}$ by a finite sum $\sum_{t=0}^{N-1}$ with suitably large $N$ that depends on $\beta$ and the distribution of $R(a_j)$, $1 \le j \le k$. For simplicity, we shall consider in the following simulation study a fixed horizon $N$ in the undiscounted case $\beta=1$. This corresponds to the finite-horizon bandit problem that maximizes the expected total reward (2). Although the optimal procedure that maximizes (2) is no longer an index rule, there are simple index rules that are asymptotically optimal as $N \to \infty$, as shown by Lai (1987).

The starting point in Lai's approximation is to consider the normal case. Suppose that an experimenter can choose at each stage $n(\le N)$ between sampling from a normal population with known variance 1 but unknown mean $\mu$ and sampling from another normal population with known mean 0. Assuming a normal prior distribution $\mathrm{N}(\mu_0, v)$ on $\mu$, the optimal rule that maximizes the expected sum of $N$ observations samples from the first population (with unknown mean) until stage $T^* = \inf\{n \le N: \hat{\mu}_n + a_{n,N} \le 0\}$ and then takes the remaining observations from the second population (with known mean 0), where $\hat{\mu}_n$ is the posterior mean based on observations $Y_1, \ldots, Y_n$ from the first population and $a_{n,N}$ are positive constants that can be determined by backward induction. Writing $t=n/N$, $w(t)=(Y_1+\cdots Y_n)/\sqrt{N}$, and treating $0 < t \le 1$ as a continuous variable, Lai (1987) approximates $a_{n,N}$ by $\sqrt{v_n}\, h(n/N)$, where $v_n$ is the posterior variance of $\mu$ and

$$h(t) = \begin{cases} \{2\log t^{-1} - \log\log t^{-1} - \log 16\pi \\ \quad +0.99\exp(-0.038t^{-1/2})\}^{1/2} & \text{if } 0 < t \le 0.01, \\ -1.58\sqrt{t} + 1.53 + 0.07t^{-1/2} & \text{if } 0.01 < t \le 0.28, \\ -0.576t^{3/2} + 0.299t^{1/2} + 0.403t^{-1/2} & \text{if } 0.28 < t \le 0.86, \\ t^{-1}(1-t)^{1/2}\{0.639 - 0.403(t^{-1}-1)\} & \text{if } 0.86 < t \le 1. \end{cases}$$

The function $h$ is obtained by first evaluating numerically the boundary of the corresponding optimal stopping problem and then developing a simple closed-form approximation to the boundary. As shown by Lai (1987, p. 1108), there is close agreement between $h$ and the boundary for the continuous-time finite-horizon optimal stopping problem, similar to Fig. 1 for the infinite-horizon discounted case.

Without assuming a prior distribution on the unknown parameters, suppose $Y_{j,1}, Y_{j,2}, \ldots,$ are independent random variables from a one-parameter exponential family with density function $f_{\theta_j}(y) = \exp\{\theta_j y - g(\theta_j)\}$ with respect to some dominating measure. Then $\mu(\theta) = \mathrm{E}_\theta Y_1 = g'(\theta)$ is increasing in $\theta$ since $\mathrm{Var}(Y_{j,1}) = g''(\theta_j)$, and the Kullback–Leibler information number is

$$I(\theta, \lambda) = \mathrm{E}_\theta[\log(f_\theta(Y)/f_\lambda(Y))] = (\theta - \lambda)\mu(\theta) - (g(\theta) - g(\lambda)). \qquad (18)$$

Assuming that all the $\theta_j$ lie in some open interval $\Gamma$ such that $\inf_{\theta\in\Gamma} g''(\theta) > 0$ and $\sup_{\theta\in\Gamma} g''(\theta) < \infty$ and letting $\hat{\theta}_{j,n}$ be the maximum likelihood estimate of $\theta_j$ based on $Y_{j,1}, \ldots, Y_{j,n}$, Lai (1987) considered an upper confidence bound

for $\mu(\theta_j)$ of the form $\mu(\theta^*_{j,n})$, where

$$\theta^*_{j,n} = \inf\{\theta \in \Gamma\colon \theta \geq \hat{\theta}_{j,n},\ 2nI(\hat{\theta}_{j,n}, \theta) \geq h^2(n/N)\}. \tag{19}$$

Note that $nI(\hat{\theta}_{j,n}, \theta_0)$ is the generalized likelihood ratio statistic for testing $\theta = \theta_0$, so the above upper confidence bound adopts the usual construction of confidence limits by inverting a generalized likelihood ratio test.

Define the regret of an allocation rule at $(\theta_1, \ldots, \theta_k)$ by

$$r_N(\theta_1, \ldots, \theta_k) = N \max_{1 \leq i \leq k} \mu(\theta_i) - \mathrm{E}_{\theta_1, \ldots, \theta_k} \left\{ \sum_{t=0}^{N-1} R(X_{t+1}) \right\}. \tag{20}$$

Note that the problem of maximizing the expected value of $\sum_{t=0}^{N-1} R(X_{t+1})$ is equivalent to that of minimizing the regret. Lai and Robbins (1985) derived the following asymptotic lower bound for the regret $r_N(\theta_1, \ldots, \theta_k)$ of uniformly good rules:

$$r_N(\theta_1, \ldots, \theta_k) \geq (\log N) \sum_{j:\theta_j < \theta^*} \left\{ \frac{\mu(\theta^*) - \mu(\theta_j)}{I(\theta_j, \theta^*)} + o(1) \right\}, \tag{21}$$

where $\theta^* = \max_{1 \leq i \leq k} \theta_i$, and a rule is said to be 'uniformly good' if $r_N(\theta_1, \ldots, \theta_k) = O(\log N)$ for any fixed $(\theta_1, \ldots, \theta_k) \in \Gamma^k$. Lai (1987) showed that the preceding upper confidence bound rule is uniformly good and attains the lower bound (21) not only at fixed $(\theta_1, \ldots, \theta_k)$ as $N \to \infty$ (so that the rule is asymptotically optimal from the frequentist viewpoint), but also uniformly over a wide range of parameter configurations, which can be integrated to show that the rule is asymptotically Bayes with respect to a large class of prior distributions $\Pi$ for $(\theta_1, \ldots, \theta_k)$.

This asymptotic theory for the finite-horizon undiscounted case is closely related to the asymptotic theory, as the discount factor $\beta$ approaches 1, for the discounted multi-armed bandit problem, in which the discounted regret of an allocation rule at $(\theta_1, \ldots, \theta_k)$ is defined by

$$\tilde{r}_\beta(\theta_1, \ldots, \theta_k) = \mathrm{E}_{\theta_1, \ldots, \theta_k} \left[ \sum_{t=0}^{\infty} \beta^t \left\{ \max_{1 \leq i \leq k} \mu(\theta_i) - R(X_{t+1}) \right\} \right]. \tag{22}$$

Making use of (21) with $N \sim (1-\beta)^{-1}$, Chang and Lai (1987) showed that as $\beta \to 1$,

$$\tilde{r}_\beta(\theta_1, \ldots, \theta_k) \geq |\log(1-\beta)| \sum_{j:\theta_j < \theta^*} \left\{ \frac{\mu(\theta^*) - \mu(\theta_j)}{I(\theta_j, \theta^*)} + o(1) \right\} \tag{23}$$

for every rule that satisfies $\tilde{r}_\beta(\theta'_1,\ldots,\theta'_k)=O(|\log(1-\beta)|)$ for all $(\theta'_1,\ldots,\theta'_k)\in \Gamma^k$. They also showed that Gittins' index rule and its approximation that replaces the Gittins indices by the simpler upper confidence bounds (19) with $h(n/N)$ replaced by $\psi(\{(1-\beta)n\}^{-1})$ attain the asymptotic lower bound (23) and are also asymptotically Bayes with respect to a large class of priors (not necessarily assuming independence among $\theta_1,\ldots,\theta_k$).

While the regret of the preceding upper confidence bound rule is of logarithmic order as $N\to\infty$ in the finite-horizon case or as $\beta\to 1$ in the discounted case, the regret of the myopic rule that chooses the action with the largest posterior mean reward (or the largest sample average reward without assuming a prior distribution on the unknown parameters) has regret of order $N$ (in the finite-horizon case) or order $(1-\beta)^{-1}$ (in the discounted case); see Kumar (1985). Therefore the upper confidence bound rule is considerably more efficient than the myopic rule for large horizons $N$ or for discount factors approaching 1, showing the importance of active experimentation in these cases. On the other hand, it is difficult to improve on the myopic rule when $N$ is moderate or small, or when there is substantial discounting, since the long-term benefit of active experimentation to improve future performance cannot be realized when there is not much time left before the final action or when future values become insignificant after discounting.

*Example 2.* To illustrate this point, consider the case of Bernoulli bandits with $k=2$ (arms) and $a=b=1$ for the parameters of the prior Beta distribution. The Bayesian myopic (BM) rule chooses the arm with the larger posterior mean at each stage, using randomization in the case of ties. The frequentist myopic (M) rule does not assume the Beta$(1,1)$ (or uniform) prior distribution for $\theta_1$ or $\theta_2$, and replaces the posterior mean in the BM rule by the sample mean. The upper confidence bound (UCB) rule described above uses the upper confidence bound (19) in lieu of the posterior mean. Table 2 gives the regret (20) of each rule at different values of $(\theta_1,\theta_2)$ for $N=20,100,300,3000$. Also given in Table 2 is the Bayes regret

$$\int_0^1\int_0^1 r_N(\theta_1,\theta_2)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 = N\int_0^1\int_0^1 \max(\theta_1,\theta_2)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 - R_{\Pi}$$
$$= 2N/3 - R_{\Pi} \tag{24}$$

for each rule, where $R_{\Pi}$ is the Bayes reward defined by (2) with uniform $\Pi$. Each result in the table is based on 1000 simulations. Table 2 shows that all three rules M, BM and UCB are nearly Bayes for $N=20$, as they have small Bayes regret. While the regret function and the Bayes regret increase slowly with $N$ for the UCB rule, they grow much faster with $N$ for the myopic rules M and BM. For $N=3000$, the UCB rule that incorporates an appropriate amount of active experimentation in a simple way shows great

Table 2
Regret for the myopic (M,BM) and upper confidence bound (UCB) rules in Bernoulli two-armed bandits

| $(\theta_1, \theta_2)$ | $N = 20$ | | | $N = 100$ | | | $N = 300$ | | | $N = 3000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | BM | UCB | M | BM | UCB | M | BM | UCB | M | BM | UCB |
| (0.1, 0.7) | 1.16 | 0.84 | 0.87 | 1.68 | 0.86 | 1.00 | 3.77 | 0.93 | 1.61 | 65.6 | 1.11 | 2.89 |
| (0.2, 0.8) | 1.41 | 0.96 | 0.83 | 1.44 | 1.46 | 1.20 | 5.60 | 1.92 | 1.68 | 122.1 | 3.62 | 2.57 |
| (0.25, 0.75) | 1.51 | 1.11 | 1.03 | 2.40 | 1.76 | 1.57 | 5.76 | 3.17 | 2.29 | 34.1 | 9.10 | 4.33 |
| (0.3, 0.5) | 1.15 | 1.21 | 1.10 | 4.74 | 4.21 | 1.57 | 13.32 | 10.14 | 4.37 | 124.6 | 95.7 | 7.95 |
| (0.4, 0.5) | 0.84 | 0.78 | 0.75 | 3.51 | 3.74 | 3.07 | 10.32 | 10.08 | 5.99 | 102.4 | 108.4 | 12.91 |
| (0.5, 0.65) | 1.01 | 1.09 | 0.94 | 4.33 | 4.49 | 3.22 | 11.76 | 12.57 | 5.74 | 117.4 | 120.1 | 9.73 |
| Bayes | 1.00 | 0.85 | 0.70 | 3.83 | 2.65 | 2.00 | 12.8 | 10.56 | 5.88 | 78.11 | 35.49 | 9.74 |

improvement over the myopic rules, from both the frequentist and Bayesian viewpoints.

## 4. Multi-armed bandit problems with switching costs

When switching costs are present, even the discounted multi-armed bandit problem does not have an optimal solution in the form of an index rule, as shown by Banks and Sundaram (1994). At any stage one has a greater propensity to stick to the current arm instead of switching to the arm with the largest index and incurring a switching cost. Although the optimal solution becomes much more complicated when there are switching costs, the basic ideas in Sections 2 and 3 can be extended to multi-armed bandits with switching costs.

To reduce switching costs, Agrawal et al. (1988) modified the construction by Lai and Robbins (1985) of rules that attain the asymptotic lower bound (21) for the regret at every fixed $(\theta_1, \ldots, \theta_k)$ so that the total switching cost up to time $t$ is of smaller order than the regret (i.e., is $o(\log t)$). Specifically, they divide time into 'frames' numbered $0, 1, 2, \ldots$ and further subdivide each frame $f$ into blocks of equal length $\max\{f, 1\}$ such that $m_f - m_{f-1} = \lceil (2^{f^2} - 2^{(f-1)^2})/f \rceil kf$ for $f \geq 1$, where $m_f$ denotes the time instant at the end of frame $f$, with $m_0 = k$. Thus the pair $(f, i)$ denotes block $i$ in frame $f$. The time instant $t$ when $(f, i)$ begins is a comparison instant at which upper confidence bounds $U_{n_t(j)}(j)$ for $\mu(\theta_j)$ are computed for $j = 1, \ldots, p$, and the action $a_{j^*}$ with the largest $U_{n_t(j)}(j)$ is chosen for the entire block $(f, i)$.

The upper confidence bounds $U_{n_t(j)}(j)$ used by Agrawal et al. are the same as those in Lai and Robbins (1985) and do not involve the horizon $N$ or the discount factor $\beta$. By incorporating $N$ into the construction of upper confidence bounds, Lai (1987) improved the finite-sample performance

of the corresponding index-type rule in finite-horizon bandit problems without switching costs. In this connection, recall the role of $N$ in $h(n/N)$ or of $c(=-\log\beta)$ in $\psi(v_{j,n}/c\sigma^2(\mu_{j,n}))$ in determining the amount of active experimentation in (19) or (16). Moreover, the choice of blocks by Agrawal et al. (1988) does not involve $N$ (or $\beta$). We can improve its performance by suitably incorporating this basic parameter into the definition of the blocks, as in the following construction of nearly optimal allocation rules in the presence of switching costs.

### 4.1. Normal two-armed bandits

To begin with, consider the finite-horizon bandit problem with $k=2$ normal arms, assuming common known variance 1 for each arm. For notational simplicity let $n_t(1)=m_t$, $n_t(2)=n_t$, $Y_j=Y_{j,1}$, $Z_j=Y_{j,2}$, $\bar{Y}_j=(Y_1+\cdots+Y_j)/j$, $\bar{Z}_j=(Z_1+\cdots+Z_j)/j$. The generalized likelihood ratio (GLR) statistic $\ell_t$ for testing $\mathrm{H}_0\colon \mathrm{E}Y_1=\mathrm{E}Z_1$ based on $Y_1,\ldots,Y_{m_t}$, $Z_1,\ldots,Z_{n_t}$ is given by

$$\ell_t=\frac{m_t n_t}{2(m_t+n_t)}(\bar{Y}_{m_t}-\bar{Z}_{n_t})^2. \tag{25}$$

Note that $\ell_t$ has the same form as the GLR statistic $n\bar{X}_n^2/2$ for testing $\mathrm{H}_0'\colon \mu=0$ based on i.i.d. normal $X_1,\ldots,X_n$ with mean $\mu$ and variance 1, if we replace $n$ by $m_t n_t/(m_t+n_t)$ and $\bar{X}_n$ by $\bar{Y}_{m_t}-\bar{Z}_{n_t}$. As noted by Lai (1987), the upper confidence bound (19) in the UCB rule can be constructed by inverting a GLR test, and for the finite-horizon problem of choosing between a normal population $\Pi_1$ with unknown mean $\mu$ and another normal population $\Pi_2$ with known mean 0, a nearly optimal rule samples from the population with unknown mean until stage

$$\begin{aligned} T&=\inf\{n\le N\colon 2nI(\bar{X}_n,0)\ge h^2(n/N)\}\\ &=\inf\{n\le N\colon n\bar{X}_n^2\ge h^2(n/N)\}, \end{aligned} \tag{26}$$

and then samples the remaining $N-T$ observations from $\Pi_1$ or $\Pi_2$ depending on whether $\bar{X}_T>0$ or $\bar{X}_T<0$.

Note that $n\bar{X}_n=w(n)$, where $w(\cdot)$ is a Wiener process with drift coefficient $\mu$. Letting $\mu=\mathrm{E}Y_1-\mathrm{E}Z_1$, Robbins and Siegmund (1974) have shown that the random sequences $\{mn(\bar{Y}_m-\bar{Z}_n)/(m+n)\}$ and $\{w(mn/(m+n))\}$ have the same joint distribution for any sequence of integer pairs $(m,n)$ which is nondecreasing in each coordinate. This suggests that in analogy with (26), after stage

$$\tau=\inf\left\{t\colon m_t+n_t\le N,\ \frac{m_t n_t}{m_t+n_t}(\bar{Y}_{m_t}-\bar{Z}_{n_t})^2\ge h^2\left(\frac{m_t n_t}{(m_t+n_t)N}\right)\right\}, \tag{27}$$

we can stop sampling from $Y$ or $Z$ depending on whether $\bar{Y}_{m_\tau} < \bar{Z}_{n_\tau}$ or $\bar{Z}_{n_\tau} < \bar{Y}_{m_\tau}$. Prior to stage $\tau$, we can use an adaptive sampling rule that carries out active experimentation with an apparently inferior population in blocks of consecutive time periods to reduce switching costs. This is the basic idea underlying the following sampling scheme.

Take an even integer $b$ (depending on the horizon $N$) and partition time into blocks so that the length of the $j$th block is $b^j - b^{j-1}$. In the first block, sample $b/2$ observations first from $Y$ and then from $Z$ with probability $1/2$, and sample $b/2$ observations first from $Z$ and then from $Y$ with probability $1/2$. For the $j$th block ($j \geq 2$), we define the leading population as that having the maximum of the two sample means at the end of the $(j-1)$st block. Sample the first $(b^j - b^{j-1})/2$ observations of the $j$th block from the leading population. Then switch to sampling from the other population until stage

$$\tau_j = \inf\{t\colon m_t + n_t \leq b^j,\ m_t n_t(\bar{Y}_{m_t} - \bar{Z}_{n_t})^2/(m_t + n_t) \geq h^2(m_t n_t/N(m_t + n_t))\}. \tag{28}$$

If the set in (28) is non-empty, stop experimentation and sample the remaining $N - \tau_j$ observations from $Y$ (or $Z$) if $\bar{Y}_{m_{\tau_j}} > (\text{or} <)\ \bar{Z}_{n_{\tau_j}}$. In particular, if $\tau_j$ occurs at the time of switching with the leading population still having the larger sample mean, then no switching actually occurs as the apparently inferior population is eliminated from further sampling. If the set in (28) is empty, let $\tau_j = b^j$ and note by induction that in this case we have sampled $b^j/2$ observations from each population at the end of the $j$th block. If $N = b^J$ for some integer $J$, the preceding definition applies to all $J$ blocks. If $b^{J-1} < N < b^J$, we modify the definition of the $J$th block by proceeding as before until the $N$th (instead of the $b^J$th) observation. We shall call this rule the *block experimentation* (BE) rule. It experiments with an apparently inferior population within blocks of consecutive times to reduce switching costs. The amount of experimentation is similar to that of the UCB rule, as illustrated in the following.

*Example 3.* The regret (20) of the BE rule that uses $b = 10$ to form the blocks is compared with that of the UCB rule and the frequentist myopic (M) rule described in Section 3, where the Bernoulli populations in Example 2 are replaced by normal populations. Note that the first block of the BE rule consists of the first 10 stages, the second block consists of stages 11 through 100, etc. Let $\mu = \mathrm{E}Y_1 - \mathrm{E}Z_1$. Table 3 gives the results for various values of $\mu$ and for $N = 100$ or 1000. They show that the BE rule has a somewhat larger regret than the UCB rule but a substantially smaller regret than the myopic rule when $N = 1000$, although the regret of the myopic rule is only slightly larger than that of the BE rule when $N = 100$. The expected number

Table 3
Regret and expected number of switches for the myopic (M), upper confidence bound (UCB) and block experimentation (BE) rules in normal two-armed bandits

| $\mu$ | $N=100$ | | | | | | | | $N=1000$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Switch # | | | | Regret | | | | Switch # | | | Regret | | |
| | M | UCB | BE | $BE^{(2)}$ | M | UCB | BE | $BE^{(2)}$ | M | UCB | BE | M | UCB | BE |
| 1 | 2.31 | 5.01 | 3.90 | 5.82 | 12.35 | 4.12 | 8.72 | 7.30 | 2.89 | 8.60 | 3.90 | 95.5 | 6.6 | 10.2 |
| 0.8 | 2.38 | 5.94 | 3.78 | 6.22 | 10.97 | 4.85 | 9.17 | 7.59 | 2.41 | 11.1 | 3.83 | 82.3 | 7.7 | 11.5 |
| 0.6 | 2.55 | 6.98 | 3.69 | 6.58 | 11.46 | 5.06 | 9.44 | 8.21 | 2.43 | 14.1 | 3.74 | 118.2 | 8.8 | 14.3 |
| 0.4 | 2.65 | 9.33 | 3.36 | 7.22 | 11.35 | 6.08 | 10.07 | 7.85 | 2.71 | 17.8 | 4.03 | 104.3 | 10.5 | 16.1 |
| 0.2 | 2.82 | 10.55 | 3.23 | 7.87 | 8.33 | 5.75 | 7.08 | 6.49 | 2.79 | 28.6 | 4.42 | 72.5 | 19.4 | 28.4 |
| 0.1 | 2.85 | 11.79 | 3.12 | 7.87 | 4.96 | 3.93 | 4.25 | 3.55 | 3.08 | 34.9 | 4.32 | 41.8 | 22.5 | 27.0 |

of switches of the UCB rule, however, is considerably larger than that of the BE rule or the myopic rule.

Unlike the rigid choice of frames and blocks in the rule of Agrawal et al. (1988), the choice of $b$ in the BE rule can depend on $N$ and the switching cost. In particular, it will be shown in Theorem 1 below that as $N \to \infty$, by choosing $b \sim (\log N)^{\varepsilon}$ with $1/2 < \varepsilon < 1$, the expected number of switches converges to 3.5 for fixed $\mu \neq 0$, while the regret of the BE rule is asymptotically equivalent to that of the UCB rule. On the other hand, for moderate values of $N$ and relatively small switching costs, it may be desirable to choose $b$ as small as 2. In particular, for the case $N=100$ in Table 3, the rule $BE^{(2)}$ uses $b=2$. Its regret is closer to that of the UCB rule than that of BE (which uses $b=10$) while the expected number of switches increases substantially.

### 4.2. Extension to the exponential family and general $k$

The preceding block experimentation and sequential GLR testing ideas can be readily generalized to $k$ populations $\pi_1,\ldots,\pi_k$ such that $\pi_i$ has density function $f_{\theta_i}(y)=\exp\{\theta_i y - g(\theta_i)\}$ with respect to some common dominating measure for $i=1,\ldots,k$. Let $b$ be a positive integer divisible by $k$ and partition time into frames such that the length of the $j$th frame is $b$ for $j=1$ and is $b^j - b^{j-1}$ for $j \geq 2$. The $j$th frame is further subdivided into $k$ blocks of equal length so that $(j,i)$ refers to the $i$th block in frame $j$. Let $(\sigma(1),\ldots,\sigma(k))$ be a random permutation of $(1,\ldots,k)$ (i.e., all $k!$ permutations are equally likely). The block $(1,i)$ in the first frame is devoted to sampling from $\pi_{\sigma(i)}$. For the $j$th frame ($j \geq 2$), denote the population with the $i$th largest sample mean among all populations not yet eliminated at the end of the $(j-1)$st frame by $\pi_{\sigma_j(i)}$. Let $I_j$ denote the number of such populations and let $\hat{i}=\sigma_j(i)$. Let $\pi_{i^*}$ denote the population with the largest sample mean among all populations not yet eliminated at the end of the block $(j,i-1)$, where the end of the block

$(j,0)$ means the end of the frame $j-1$. Let $Y_{i,1}, Y_{i,2}, \ldots$ denote successive observations from $\pi_i$ and $\bar{Y}_{i,t}$ be the sample mean based on $Y_{i,1}, \ldots, Y_{i,t}$. For the block $(j,i)$, which will be denoted by $B_{j,i}$ (with $1 \le i \le I_j$), we sample from $\pi_{\hat{i}}$ until stage

$$\tau = \inf \left\{ t \in B_{j,i} \colon \ell(\bar{Y}_{i^*,n_t(i^*)}, \bar{Y}_{i,n_t(\hat{i})}; n_t(i^*), n_t(\hat{i})) \right.$$
$$\left. \ge \frac{1}{2} h^2 \left( \frac{n_t(i^*) n_t(\hat{i})}{N[n_t(i^*) + n_t(\hat{i})]} \right) \right\}, \tag{29}$$

where $\tau$ is defined as the largest number in $B_{j,i}$ if the set in (29) is empty, and $\ell(\bar{Y}_{k,m}, \bar{Y}_{i,n}; m, n)$ is the GLR statistic for testing $H_0$: $\mathrm{E}Y_k = \mathrm{E}Y_i$ based on $Y_{k,1}, \ldots, Y_{k,m}, Y_{i,1}, \ldots, Y_{i,n}$ and is given by (30) below. If the set in (29) is non-empty, eliminate $\pi_{\hat{i}}$ (or $\pi_{i^*}$) from further sampling if $\bar{Y}_{i,n_\tau(\hat{i})} <$ (or $>$) $\bar{Y}_{i^*,n_\tau(i^*)}$, and the remaining observations in the block $(j,i)$ are sampled from $\pi_{i^*}$ (or $\pi_{\hat{i}}$). Note that (29) reduces to (28) in the normal case, for which the GLR statistics are given by (25). For $I_j < i \le k$, the block $(j,i)$ is devoted to sampling from the population with the largest sample mean among all populations not yet eliminated at the end of block $(j, I_j)$. We call this rule the BE rule for the $k$-armed bandit problem.

If $N = b^J$ for some integer $J$, the preceding definition of the BE rule applies to all $J$ frames. If $b^{J-1} < N < b^J$, we modify the definition of the $J$th frame by proceeding as before until the $N$th observation. The following theorem, whose proof is given in the appendix, shows that the BE rule has an asymptotically optimal regret and also gives the asymptotic behavior of its expected number of switches. As in Lai (1987), our analysis of boundary crossing probabilities of GLR statistics in the proof of the theorem requires the regularity condition that $\theta_1, \ldots, \theta_k$ all belong to an open interval $\Gamma = (\gamma, \gamma^*)$ such that $-\infty \le \gamma < \gamma^* \le \infty$, $\inf_{\gamma-\delta<\theta<\gamma^*+\delta} g''(\theta) > 0$, $\sup_{\gamma-\delta<\theta<\gamma^*+\delta} g''(\theta) < \infty$ and $g''$ is uniformly continuous on $(\gamma-\delta, \gamma^*+\delta)$ for some $\delta > 0$. Note that this condition is satisfied in the normal case with $\gamma = -\infty$ and $\gamma^* = \infty$. The maximum likelihood estimate of $\mu(\theta_i)$ based on $Y_{i,1}, \ldots, Y_{i,m}$ is $\mu(\gamma) \vee (\bar{Y}_{i,m} \wedge \mu(\gamma^*))$, where $\vee$ and $\wedge$ denote maximum and minimum, respectively, and the GLR statistic for testing $H_0 : \mu(\theta_k) = \mu(\theta_i)$ based on $Y_{k,1}, \ldots, Y_{k,m}, Y_{i,1}, \ldots, Y_{i,n}$ is

$$\ell(\bar{Y}_{k,m}, \bar{Y}_{i,n}; m, n) = mL(\bar{Y}_{k,m}) + nL(\bar{Y}_{i,n})$$
$$- (m+n)L((m\bar{Y}_{k,m} + n\bar{Y}_{i,n})/(m+n)) \tag{30}$$

with $L(y) = \tilde{L}(\mu(\gamma) \vee (y \wedge \mu(\gamma^*)))$ and $\tilde{L}(z) = z\mu^{-1}(z) - g(\mu^{-1}(z))$, noting that the function $\mu = g'$ is continuous and increasing and therefore has an inverse.

Moreover, for this asymptotic theory, we can replace, as in Lai (1987), the specific form of $h$ in Section 3 by more general positive functions on $(0,1]$ such that for some $\xi > -3/2$,

$$h^2(t) \sim 2\log t^{-1} \quad \text{and} \quad h^2(t)/2 \geq \log t^{-1} + \xi \log\log t^{-1} \quad \text{as } t \to 0. \tag{31}$$

*Theorem 1. In the* BE *rule above, suppose* $b \sim (\log N)^{\varepsilon}$ *for some* $1/2 < \varepsilon < 1$ *and* $h:(0,1] \to (0,\infty)$ *satisfies* (31) *for some* $\xi > -3/2$. *Let* $I(\theta,\lambda)$ *denote the Kullback–Leibler information number defined in* (18). *Define the regret* $r_N(\theta_1,\ldots,\theta_k)$ *of the BE rule by* (20), *and let* $s_N(\theta_1,\ldots,\theta_k)$ *denote its expected number of switches up to stage* $N$. *Let* $\theta^* = \max_{1\leq i\leq k}\theta_i$.

(i) *At every fixed* $(\theta_1,\ldots,\theta_k) \in \Gamma^k$, *as* $N \to \infty$,

$$r_N(\theta_1,\ldots,\theta_k) \sim (\log N) \sum_{j:\theta_j<\theta^*} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j,\theta^*),$$

$$s_N(\theta_1,\ldots,\theta_k) \to 2k - k^{-1} \text{ if } \theta^* = \theta_i \text{ for only one } i.$$

(ii) *Let* $\#_N(j)$ *denote the expected number of observations from* $\pi_j$ *and* $S_N(j)$ *denote the expected number of switches to and from* $\pi_j$ *up to stage* $N$. *Then as* $\theta^* - \theta_j \to 0$ *but* $N(\theta^* - \theta_j)^2 \to \infty$,

$$\#_N(j) \sim (\log[N(\theta^* - \theta_j)^2])/I(\theta_j,\theta^*),$$

$$S_N(j) = O(\max\{1, |\log(\theta^* - \theta_j)|/\log\log N\}).$$

Parts (i) and (ii) of Theorem 1 can be used to show that the Bayes regret $\int\ldots\int r_N(\theta_1,\ldots,\theta_k)\,\mathrm{d}\Pi(\theta_1,\ldots,\theta_k)$ is asymptotically minimal over a large class of prior distributions $\Pi$, as in Lai (1987).

### 4.3. Discounted bandit problems with switching costs

We can easily modify the BE rule for the infinite-horizon discounted bandit problem with regret $\tilde{r}_\beta(\theta_1,\ldots,\theta_k)$ defined by (22). Simply replace $N$ in (29) by $(1-\beta)^{-1}$ and remove the upper bound $J$ on the number of frames. This modified rule will be denoted by $\mathrm{BE}_\beta$. The analogue of $s_N(\theta_1,\ldots,\theta_k)$ in the discounted problem is

$$\tilde{s}_\beta(\theta_1,\ldots,\theta_k) = E_{\theta_1,\ldots,\theta_k}\left[\sum_{t=1}^{\infty} \beta^t \mathbf{1}_{\{\text{a switch occurs at time t}\}}\right].$$

*Theorem 2. Suppose* $h:(0,\infty)\to(0,\infty)$ *satisfies* (31) *for some* $\xi > -3/2$ *and* $b \sim |\log(1-\beta)|^{\varepsilon}$ *for some* $1/2 < \varepsilon < 1$ *in the* $\mathrm{BE}_\beta$ *rule. Then at every fixed* $(\theta_1,\ldots,\theta_k)\in\Gamma^k$, *as* $\beta\to 1$,

$$\tilde{r}_\beta(\theta_1,\ldots,\theta_k) \sim |\log(1-\beta)| \sum_{j:\theta_j<\theta^*} (\mu(\theta^*)-\mu(\theta_j))/I(\theta_j,\theta^*),$$

$$\tilde{s}_\beta(\theta_1,\ldots,\theta_k) \to 2k-k^{-1} \quad \textit{if } \theta^*=\theta_i \textit{ for only one } i.$$

*Moreover, the conclusion of Theorem* 1(ii) *also holds for the rule* $\mathrm{BE}_\beta$.

## 5. Conclusion

In Section 2, we provide closed-form approximations to Gittins indices so that the optimal index rule can be easily implemented for discounted bandit problems. Although index rules are no longer optimal for finite-horizon (instead of discounted infinite-horizon) multi-armed bandit problems, they are asymptotically optimal from both Bayesian and frequentist viewpoints for large horizons. They also perform well for small or moderate values of the horizon $N$, for which even the myopic rule that does not incorporate active experimentation is shown to perform well in Section 3. When switching costs are present, even the discounted multi-armed bandit problem does not have an optimal solution in the form of an index rule, as shown by Banks and Sundaram (1994). Nevertheless, Section 4 has shown how index rules can be modified by not switching within prespecified blocks of time to come up with asymptotically optimal rules in the discounted or finite-horizon multi-armed bandit problem with switching costs.

The incomplete learning theorem for discounted multi-armed bandits established by Rothschild (1974), Banks and Sundaram (1992) and Brezzi and Lai (1999b) shows that in feedback control of a system with unknown parameters, the control objective may preclude full knowledge of the parameter values in the long run. However, one still needs sufficient information about the unknown parameters to come up with an appropriate action at every stage. A good control rule therefore introduces adjustments into the myopic rule so that some active experimentation is used to generate information about the unknown parameters. For discounted or finite-horizon multi-armed bandits, we have shown how such adjustments can be implemented by using an index which replaces the sample estimates of the parameters by suitable upper confidence bounds. In view of the duality between confidence intervals and hypothesis testing, we can also perform these adjustments by testing the hypothesis whether an apparently superior action is indeed superior. In particular, we have used this hypothesis testing approach in Section 4 to address the long-standing problem of switching costs in multi-armed bandits.

## Appendix.

*Proof of Theorem 1.* Consider the special case $k=2$, as the general case can be treated by a similar argument. Without loss of generality, we shall assume that $\theta_1 > \theta_2$. Let $d = \theta_1 - \theta_2$. In view of (31), we can make use of Lemma 2.6 of Zhang (1992) on boundary crossing probabilities of GLR statistics (with a modification of the statement to accommodate unequal sample sizes from the two populations but with essentially the same proof) to show that as $Nd^2 \to \infty$ with $O < d = o((\log N)^{1/2})$,

$$\mathrm{P}\{\ell(\bar{Y}_{1,n_t(1)}, \bar{Y}_{2,n_t(2)}; n_t(1), n_t(2)) \geq \frac{1}{2}h^2\,(n_t(1)n_t(2)/N[n_t(1)+n_t(2)]) \text{ and } \bar{Y}_{1,n_t(1)} < \bar{Y}_{2,n_t(2)} \text{ for some } t \leq N\} = O((Nd^2)^{-1}(\log Nd^2)^{-\xi-1/2}). \tag{A.1}$$

Consider the event $F = \{\text{Population 1 is not eliminated at any stage t} \leq N\}$, and let $F^c$ denote its complement. Since $F^c$ is a subset of the event in (A.1), $\mathrm{P}(F^c) = O((Nd^2)^{-1})$ by (A.1). Note that

$$\begin{aligned}\mathrm{E}\{n_N(2)\mathbf{1}_F\} \leq \#_N(2) &\leq N\mathrm{P}(F^c) + \mathrm{E}\{n_N(2)\mathbf{1}_F\}\\ &= \mathrm{E}\{n_N(2)\mathbf{1}_F\} + O(d^{-2}(\log Nd^2)^{-\xi-1/2}).\end{aligned} \tag{A.2}$$

Since $\xi > -3/2$, it follows from (A.2) that $\#_N(2) - E\{n_N(2)\mathbf{1}_F\} = o(d^{-2}\log Nd^2)$.

First consider the case of fixed $d > 0$, as in part (i) of the theorem. By standard large deviation bounds for sample means from an exponential family, there exists $\rho > 0$ such that

$$\mathrm{P}\{\bar{Y}_{1,m} < \bar{Y}_{2,n}\} = O(\mathrm{e}^{-\rho m} + \mathrm{e}^{-\rho n}) \quad \text{as } m \to \infty \quad \text{and} \quad n \to \infty. \tag{A.3}$$

For the BE rule, $n_b(1) = n_b(2) = b/2 \sim (\log N)^e/2$. On $F \cap \{\bar{Y}_{1,n_b(1)} > \bar{Y}_{2,n_b(2)}\}$, the BE rule samples from $\pi_1$ for stages $n_b(1)+1, \ldots, n_b(1)+(b^2-b)/2$ and then switches to sampling from $\pi_2$ until the end of the second frame or the elimination of $\pi_2$, whichever occurs sooner. Let $\mu_i = \mu(\theta_i)$. Since $\tilde{L}'(y) = \mu^{-1}(y)$, it follows that as $m/n \to \infty$,

$$\tilde{L}\left(\frac{m\mu_1 + n\mu_2}{m+n}\right) = \tilde{L}(\mu_1) + \frac{n}{m+n}(\mu_2 - \mu_1)\theta_1 + O\left(\left(\frac{n}{m}\right)^2\right).$$

Since $I(\theta_2, \theta_1) = (\theta_2 - \theta_1)\mu_2 - (g(\theta_2) - g(\theta_1))$ and $\tilde{L}(\mu_i) = \theta_i\mu_i - g(\theta_i)$, it then follows that

$$\begin{aligned}&m\tilde{L}(\mu_1) + n\tilde{L}(\mu_2) - (m+n)\tilde{L}((m\mu_1 + n\mu_2)/(m+n))\\ &\quad = n\{\tilde{L}(\mu_2) - \tilde{L}(\mu_1) - (\mu_2 - \mu_1)\theta_1 + O(n/m)\}\\ &\quad = n\{I(\theta_2, \theta_1) + O(n/m)\}\end{aligned} \tag{A.4}$$

as $m/n \to \infty$. Noting that $mn/(m+n) \sim n$ as $m/n \to \infty$ and using (31) and (A.4) together with the law of large numbers, it can be shown that with probability 1, $\pi_2$ is eliminated at some stage $t$ in frame 2 with $n_t(2) \sim (\log N)/I(\theta_2, \theta_1)$. Uniform integrability arguments can then be used to show that

$$\mathrm{E}\{n_N(2)\mathbf{1}_{F\cap\{\bar{Y}_{1,n_b(1)} > \bar{Y}_{2,n_b(2)}\}}\} \sim (\log N)/I(\theta_2, \theta_1). \tag{A.5}$$

Making use of exponential bounds for the large deviation probabilities of sample means and noting that $n_N(2) < b^2$ on the event that $\pi_2$ is eliminated in the second frame, we obtain that

$$\mathrm{E}\{n_N(2)\mathbf{1}_{F\cap\{\bar{Y}_{1,n_b(1)} \le \bar{Y}_{2,n_b(2)}\}}\} \le b^2 \mathrm{e}^{-\rho b} + O(N\mathrm{e}^{-\rho b^2}), \tag{A.6}$$

where the second summand on the right hand side simply bounds $n_N(2)$ by $N$. Since $b^2 \sim (\log N)^{2\varepsilon}$ and $2\varepsilon > 1$, it then follows that

$$\mathrm{E}\{n_N(2)\mathbf{1}_F\} = \mathrm{E}\{n_N(2)\mathbf{1}_{F\cap\{\bar{Y}_{1,n_b(1)} > \bar{Y}_{2,n_b(2)}\}}\} + o(1) \sim (\log N)/I(\theta_2, \theta_1).$$

The preceding proof also shows that on $F \cap \{\bar{Y}_{1,n_b(1)} > \bar{Y}_{2,n_b(2)}\}$, there are two switches in the second frame, to and from $\pi_2$, with probability 1 as $N \to \infty$. There is also one switch in the middle of the first frame, and with probability 1/2 one more switch at the end of the first frame (when $\pi_1$ is chosen at the beginning). Uniform integrability arguments and bounds of the type in (A.6) then show that $s_N(\theta_1, \theta_2) \to 3.5$ in this case.

We next consider the case $d \to 0$ but $Nd^2 \to \infty$, as in part (ii) of the theorem. In this case, (A.2) still holds and $\pi_2$ is eliminated at some stage $t$ belonging to some frame $j$ with $n_t(2) \sim (\log Nd^2)/I(\theta_2, \theta_1)$, with probability approaching 1 as $Nd^2 \to \infty$. At the end of frame $j-1$, the BE rule has sampled $b^{j-1}/2$ observations from each population on the event $F$. Therefore, similar uniform integrability arguments can be used to show that $\mathrm{E}\{n_N(2)\mathbf{1}_F\} \sim (\log Nd^2)/I(\theta_2, \theta_1)$. Moreover, the expected number of switches is $O(J_d)$, where $b^{J_d} > (1+o(1))(\log Nd^2)/I(\theta_2, \theta_1) > b^{J_d-1}/2$. Since $I(\theta_2, \theta_1) \sim d^2 g''(\theta_1)/2$, we have

$$J_d \log(\log N)^{\varepsilon} = \log(\log N + \log d^2) - \log d^2 + O(\log\log N),$$

yielding $\varepsilon J_d \sim (-\log d^2)/(\log\log N) + O(1)$.

*Proof of Theorem 2.* Take any $a > 0$ and choose a positive integer $N(a) \sim a(1-\beta)^{-1}$. We can derive the desired conclusions as $\beta \to 1$ by applying Theorem 1 to the horizon $N(a)$ with $a$ arbitrarily large.

## References

Agrawal, R., Hegde, M.V., Teneketzis, D., 1988. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching costs. IEEE Transactions on Automatic Control 33, 899–906.

Banks, J.S., Sundaram, R.K., 1992. Denumerable-narmed bandits. Econometrica 60, 1071–1096.
Banks, J.S., Sundaram, R.K., 1994. Switching costs and the Gittins index. Econometrica 62, 687–694.
Berry, D.A., 1972. A Bernoulli two-armed bandit. Annals of Mathematical Statistics 43, 871–897.
Brezzi, M., Lai, T.L., 1999. Optimal stopping for Brownian motion in bandit problems and sequential analysis, Working Paper, Department of Statistics, Stanford University.
Brezzi, M., Lai, T.L., 2000. Incomplete learning from endogenous data in dynamic allocation. Econometrica 68, 1511–1516.
Chang, F., Lai, T.L., 1987. Optimal stopping and dynamic allocation. Advances of Applied Probability 19, 829–853.
Chernoff, H., Petkau, A.J., 1986. Numerical solutions for Bayes sequential decision problems. SIAM Journal of Scientific and Statistical Computing 7, 46–59.
Fabius, J., van Zwet, W.R., 1970. Some remarks on the two-armed bandit. Annals of Mathematical Statistics 41, 1906–1916.
Feldman, D., 1962. Contributions to the two-armed bandit problem. Annals of Mathematical Statistics 33, 847–856.
Gittins, J.C., 1979. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society, Series B 41, 148–177.
Gittins, J.C., 1989. Multi-Armed Bandit Allocation Indices. Wiley, New York.
Gittins, J.C., Jones, D.M., 1974. A dynamic allocation index for the sequential design of experiments. In: Gani, J., Sarkadi, K., Vineze, I. (Eds.), Progress in Statistics. North-Holland, Amsterdam, pp. 241–266.
Jovanovich, B., 1979. Job-search and the theory of turnover. Journal of Political Economy 87, 972–990.
Kumar, P.R., 1985. A survey of some results in stochastic adaptive control. SIAM Journal of Control and Optimization 23, 329–380.
Lai, T.L., 1987. Adaptive treatment allocation and the multi-armed bandit problem. Annals of Statistics 15, 1091–1114.
Lai, T.L., Robbins, H., 1985. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics 6, 4–22.
LeCam, L., 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. University of California Publications in Statistics 1, 277–330.
McLennan, A., 1984. Price disperson and incomplete learning in the long run. Journal of Economic Dynamics and Control 7, 331–347.
Mortensen, D., 1985. Job-search and labor market analysis. In: Ashenfelter, O., Layard, R. (Eds.),, Handbook of Labor Economics, Vol. 2. North Holland, Amsterdam, pp. 849–919.
Robbins, H., Siegmund, D., 1974. Sequential tests involving two populations. Journal of the American Statistical Association 69, 132–139.
Rothschild, M., 1974. A two-armed bandit theory of market pricing. Journal of Economic Theory 9, 185–202.
Whittle, P., 1980. Multi-armed bandits and the Gittins index. Journal of the Royal Statistical Society, Series B 42, 143–149.
Zhang, L., 1992. Asymptotically optimal sequential tests of linear hypotheses in multiparameter exponential families. Ph.D. Dissertation, Department of Statistics, Stanford University.