Econometrica, Vol. 68, No. 6 (November, 2000), 1511–1516

NOTES AND COMMENTS

INCOMPLETE LEARNING FROM ENDOGENOUS DATA IN DYNAMIC ALLOCATION

BY MONICA BREZZI AND TZE LEUNG LAI[1]

1. INTRODUCTION

THIS PAPER STUDIES THE PROBLEM of learning from endogenous data by an economic agent who 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 θ_j that has a prior distribution $\Pi^{(j)}$. The agent's objective is to maximize the total discounted reward

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

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). The theory of multi-armed bandits has been applied to decision making in labor markets (cf. Jovanovic (1979), Mortensen (1985)), general search problems involving nondurable goods (cf. Banks and Sundaram (1992)) and pricing under demand uncertainty (cf. Rothschild (1974)).

The DAI (also called the "Gittins index") of action a_j at stage t is a complicated function of the posterior distribution of θ_j given the rewards, up to stage t, at the times when action a_j is used. In Section 2 we derive simple bounds for the DAI involving only the mean and the standard deviation of the conditional expected reward of a_j given θ_j, with respect to the posterior distribution of the latter. Making use of these bounds, we give in Section 3 a simple proof of the incompleteness of optimal learning from endogenous data in the discounted multi-armed bandit problem. Specifically, for $k \geq 2$, we show that with positive probability the index rule uses the optimal action a_{j^*} only finitely often and that it can estimate consistently only one of the θ_j. This generalizes Rothschild's (1974) result for Bernoulli two-armed bandits, and also the result of Banks and Sundaram (1992) who show that there is positive probability of incomplete learning in multi-armed bandits with general distributions of rewards if the priors have finite support.[2] A comprehensive theory about the limits of beliefs and actions as $t \to \infty$ under the optimal rule is given in Section 3. This theory solves completely, in the context of discounted multi-armed bandits, the fundamental problem concerning the extent of experimentation and the long-run beliefs and actions of optimizing agents who learn by

[1]The first author gratefully acknowledges financial support provided by Universita di Padova and Consiglio Nazionale delle Ricerche, Italy. The research of the second author is partially supported by the National Science Foundation.

[2]Banks and Sundaram proved this result when the number of arms is denumerable (possibly infinite); we explain how to adapt our result to their setting at the end of the paper.

doing (cf. Blume and Easley (1984), McLennan (1984), Easley and Kiefer (1988), Aghion, Bolton, Harris, and Julien (1991), Banks and Sundaram (1992), El-Gamal and Sundaram (1993)).

2. THE GITTINS INDEX

Let $R(a_j)$ have distribution function F_{θ_j} (depending on the unknown parameter θ_j) and let $Y_{j,1}, Y_{j,2}, \ldots$ be independent random variables with common distribution function F_{θ_j}, representing the rewards at successive times when action a_j is taken. Let $\Pi^{(j)}$ be a prior distribution on θ_j. The Gittins index $\nu(\Pi^{(j)})$ associated with $\Pi^{(j)}$ is defined as

$$\nu(\Pi^{(j)}) = \sup_{\tau} \left\{ \frac{\int E_{\theta_j}\left(\sum_{t=0}^{\tau-1} \beta^t Y_{j,t+1}\right) d\Pi^{(j)}(\theta_j)}{\int E_{\theta_j}\left(\sum_{t=0}^{\tau-1} \beta^t\right) d\Pi^{(j)}(\theta_j)} \right\}, \tag{2.1}$$

where the supremum is over all stopping times $\tau \geq 1$ defined on $\{Y_{j,1}, Y_{j,2}, \ldots\}$ (cf. Gittins (1979)). Note that 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 $Y_{j,n+1}, Y_{j,n+2}, \ldots$ are independent having common distribution function F_{θ_j} and that θ_j has distribution $\Pi_n^{(j)}$, which is the posterior distribution of θ_j given $(Y_{j,1}, \ldots, Y_{j,n})$. Letting $\mu(\theta_j) = E_{\theta_j}(R(a_j))$, it then follows that for $m > n$,

$$E[Y_{j,m} \mid Y_{j,1}, \ldots, Y_{j,n}] = \int \mu(\theta_j) d\Pi_n^{(j)}(\theta_j) = E[\mu(\theta_j) \mid Y_{j,1}, \ldots, Y_{j,n}]. \tag{2.2}$$

The Gittins index (2.1) of $\Pi^{(j)}$ can be equivalently defined as the infimum of the set of solutions M of the equation

$$\sup_{\tau} \int E_{\theta_j}\left\{ \sum_{n=0}^{\tau-1} \beta^n \int \mu(\theta_j) d\Pi_n^{(j)}(\theta_j) + M \sum_{n=\tau}^{\infty} \beta^n \right\} d\Pi^{(j)}(\theta_j) = M \sum_{n=0}^{\infty} \beta^n, \tag{2.3}$$

where we set $\Pi_0^{(j)} = \Pi^{(j)}$ (cf. Whittle (1980)).

To compute $\nu(\Pi^{(j)})$, one has to solve the optimal stopping problem in the right-hand side of (2.1), or a family of optimal stopping problems (indexed by M) in the left-hand side of (2.3) so that the index can be obtained as the value M that equates both sides of (2.3). The following theorem gives simple bounds for $\nu(\Pi^{(j)})$ involving the mean and standard deviation of $\mu(\theta_j)$ under the distribution $\Pi^{(j)}$ for θ_j. For notational simplicity we shall fix j and omit the subscript (or superscript) in θ_j, $Y_{j,t}$ (or $\Pi^{(j)}$) in the remainder of this section. To distinguish from the case where θ is treated as an unknown parameter, we shall write Θ (instead of θ) when it is treated as a random variable (having distribution Π). *The symbol E will denote expectation with respect to the probability measure under which Θ has distribution Π and $Y_1, Y_2, \ldots$, are independent random variables with common distribution function F_θ conditional on $\Theta = \theta$. Thus E_Θ denotes conditional expectation given the random variable Θ.* This notation will be used throughout the sequel.

THEOREM 1: *Suppose that $\int \mu^2(\theta) d\Pi(\theta) < \infty$. Let μ_Π denote the mean and σ_Π^2 denote the variance of $\mu(\Theta)$ (under the distribution Π for Θ). Then*

$$\mu_\Pi \leq \nu(\Pi) \leq \mu_\Pi + \sigma_\Pi \frac{\beta}{1-\beta}.$$

PROOF: First note that

$$E_\theta\left(\sum_{t=0}^{\tau-1} \beta^t Y_{t+1} \right) = E_\theta\left(\sum_{t=1}^{\tau} \beta^{t-1} Y_t \right) = \sum_{t=1}^{\infty} \beta^{t-1} E_\theta(Y_t \mathbf{1}_{\{\tau \geq t\}}). \tag{2.4}$$

To justify interchanging the order of expectation and summation in the second equality above, write $Y_t = Y_t^+ - Y_t^-$ and apply the monotone convergence theorem (cf. Williams (1991)), noting that $\int E_\theta Y_t^- d\Pi(\theta) < \infty$. Since $\{\tau \geq t\}$ is the complement of $\{\tau \leq t-1\}$, which depends on $Y_1, \ldots, Y_{t-1}$ because τ is a stopping time, it follows from independence between Y_t and $(Y_1, \ldots, Y_{t-1})$ that $E_\theta(Y_t \mathbf{1}_{\{\tau \geq t\}}) = \mu(\theta) E_\theta \mathbf{1}_{\{\tau \geq t\}}$. Therefore (2.4) implies that

$$E_\theta\left(\sum_{t=0}^{\tau-1} \beta^t Y_{t+1}\right) = \mu(\theta) E_\theta\left(\sum_{t=1}^{\tau} \beta^{t-1}\right) = \mu(\theta)(1 - E_\theta \beta^\tau)/(1-\beta).$$

Hence the Gittins index (2.1) can be expressed in the form

$$\nu(\Pi) = \sup_\tau \left\{ \frac{E[\,\mu(\Theta)(1 - E_\Theta \beta^\tau)]}{1 - E\beta^\tau} \right\}, \tag{2.5}$$

where $\sup_\tau$ is over all stopping times $\tau \geq 1$. In particular, for the stopping time $\tau \equiv 1$, (2.5) yields $\nu(\Pi) \geq E\mu(\Theta) = \mu_\Pi$.

It remains to prove the upper bound of Theorem 2.1. Note that

$$\begin{aligned} E[\,\mu(\Theta)(1 - E_\Theta \beta^\tau)] &= (E\mu(\Theta))(1 - E\beta^\tau) - E[\,\mu(\Theta)(E_\Theta \beta^\tau - E\beta^\tau)] \\ &= (E\mu(\Theta))(1 - E\beta^\tau) - \operatorname{cov}(E_\Theta \beta^\tau, \mu(\Theta)), \end{aligned} \tag{2.6}$$

since $E(E_\Theta(\beta^\tau)) = E[E(\beta^\tau \mid \Theta)] = E\beta^\tau$. Moreover, letting $\sigma^2(\tau) = \operatorname{var}(E_\Theta \beta^\tau)$, we have $|\operatorname{cov}(E_\Theta \beta^\tau, \mu(\Theta))| \leq \sigma(\tau)\sigma_\Pi$, and therefore (2.5) and (2.6) imply that

$$\nu(\Pi) \leq E\mu(\Theta) + \sigma_\Pi \sup_\tau \frac{\sigma(\tau)}{1 - E\beta^\tau}. \tag{2.7}$$

Clearly $\sigma^2(\tau) \leq E[E_\Theta(\beta^\tau)]^2 \leq \beta^2$ while $1 - E\beta^\tau \geq 1 - \beta$ for $\tau \geq 1$. Hence it follows from (2.7) that $\nu(\Pi) \leq \mu_\Pi + \sigma_\Pi \beta/(1-\beta)$. *Q.E.D.*

Chapter 7 of Gittins (1989) describes computational methods to calculate Gittins indices for normal, Bernoulli, and exponential F_θ, with the prior distribution of θ belonging to a conjugate family. These methods involve approximating the infinite horizon in the optimal stopping problem by a finite horizon N and using backward induction. When β is near 1, a good approximation requires a very large horizon N, which becomes computationally prohibitive. In this case Brezzi and Lai (1999) developed a much simpler approximation to the Gittins index, by first approximating the posterior distribution with a normal distribution having the same mean and variance and then solving the corresponding optimal stopping problem for Brownian motion.

3. LIMITING BELIEFS AND ACTIONS OF THE OPTIMAL ALLOCATION RULE

We now make use of Theorem 1 to show that with probability 1 the optimal allocation rule, which is Gittin's index rule, samples only finitely often from all except one of the k populations when $\mu(\theta_1), \ldots, \mu(\theta_k)$ are distinct. Moreover, with positive probability the population from which the optimal policy samples exclusively after some finite time is not the best population. Such incomplete learning by the optimal policy can be attributed to the discount factor, which downweights the need for acquiring information to benefit long-run future performance.

The following notation and assumptions will be used in Theorem 2 below. Suppose that the unknown parameter θ_j takes the values in the same parameter space Γ for $j = 1, \ldots, k$, and that if $\theta_j = \theta$ then $R(a_j)$ has distribution P_θ, depending only on θ and

not on j. Let

$$l_\Gamma = \inf_{\theta \in \Gamma} \mu(\theta), \quad \hat{\mu}_{j,n} = \mu_{\Pi_n^{(j)}}, \quad \text{and} \quad \sigma_{j,n} = \sigma_{\Pi_n^{(j)}},$$

where μ_Π and σ_Π^2 denote the mean and the variance of $\mu(\Theta)$ under the distribution Π for Θ, as in Theorem 1. The Gittins index of an action a_j at stage t will be denoted by $\nu(\Pi_{N_t(j)}^{(j)})$ as in Section 2, since it is a function of the posterior distribution $\Pi_{N_t(j)}^{(j)}$ based on the $N_t(j)$ rewards associated with a_j. Let $N_\infty(j) = \lim_{t \to \infty} N_t(j)$. Assume the following conditions for every $\theta \in \Gamma$ and $j = 1, \ldots, k$:

(C1) $P_\theta\{\lim_{n \to \infty} \hat{\mu}_{j,n} = \mu(\theta)\} = 1$, $P_\theta\{\lim_{n \to \infty} \sigma_{j,n} = 0\} = 1$.

(C2) $P_\theta\{\hat{\mu}_{j,n} > l_\Gamma\} = 1$ *for all* n.

(C3) *For any* $\lambda > l_\Gamma$, *there exists* $m = m(\theta, \lambda)$ *such that* $P_\theta\{\hat{\mu}_{j,m} < \lambda\} > 0$.

THEOREM 2: *Assume* (C1). *For any* $(\theta_1, \ldots, \theta_k) \in \Gamma^k$ *such that* $\mu(\theta_i) \neq \mu(\theta_j)$ *if* $i \neq j$,
(i) $P_{\theta_1, \ldots, \theta_k}\{N_\infty(j) < \infty$ *for all except one* $j\} = 1$,
(ii) $P_{\theta_1, \ldots, \theta_k}\{N_\infty(j^*) < \infty\} > 0$ *if* $\min_{1 \le i \le k} \mu(\theta_i) > l_\Gamma$ *and* (C2) *and* (C3) *also hold, where* $\mu(\theta_{j^*}) = \max_{1 \le i \le k} \mu(\theta_i)$.

PROOF: In what follows "a.s." refers to almost surely under the measure $P_{\theta_1, \ldots, \theta_k}$. From (C1) and Theorem 1, it follows that $\nu(\Pi_{N_t(j)}^{(j)}) \to \mu(\theta_j)$ a.s. on the event $\{N_\infty(j) = \infty\}$. Since the $\mu(\theta_j)$ are distinct, this implies that $\nu(\Pi_{N_t(j)}^{(j)})$ and $\nu(\Pi_{N_t(i)}^{(i)})$ converge a.s. to two distinct numbers on the event $\{N_\infty(j) = \infty = N_\infty(i)\}$ for $i \neq j$. For definiteness suppose $\mu(\theta_i) > \mu(\theta_j)$. Because the index rule does not sample from $\Pi^{(j)}$ whenever $\nu(\Pi_{N_t(j)}^{(j)}) < \nu(\Pi_{N_t(i)}^{(i)})$, this implies that the event $\{N_\infty(j) = \infty = N_\infty(i)\}$ cannot occur with positive probability, proving part (i) of the theorem.

To prove part (ii), assume without loss of generality that $j^* = 1$. Take $\tilde{\eta} \in (l_\Gamma, \mu(\theta_2))$. By (C1), there exists n_0 such that

$$p := P_{\theta_2}\{\hat{\mu}_{2,n} \ge \tilde{\eta} \text{ for all } n \ge n_0\} > 0. \tag{3.1}$$

By (C2), $P_{\theta_2}\{\min_{1 \le n \le n_0} \hat{\mu}_{2,n} > l_\Gamma\} = 1$, and therefore there exists $\eta \in (l_\Gamma, \tilde{\eta}]$ such that

$$P_{\theta_2}\{\hat{\mu}_{2,n} \ge \eta \text{ for all } n \le n_0\} \ge 1 - p/2. \tag{3.2}$$

Combining (3.1) and (3.2) yields

$$P_{\theta_2}\{\hat{\mu}_{2,n} \ge \eta \text{ for all } n \le 1\} \ge 1 - (1 - p) - p/2 = p/2.$$

This and the lower bound in Theorem 1 gives

$$P_{\theta_2}\{\nu(\Pi_n^{(2)}) \ge \eta \text{ for all } n \ge 1\} \ge p/2. \tag{3.3}$$

By (C3) and the assumption on $\sigma_{1,n}$ in (C1), there exists m such that

$$q := P_{\theta_1}\left\{\hat{\mu}_{1,m} + \beta(1 - \beta)^{-1} \sigma_{1,m} < \eta\right\} > 0. \tag{3.4}$$

Using independence and the upper bound in Theorem 1, we obtain from (3.3) and (3.4) that

$$P_{\theta_1, \ldots, \theta_k}\{\nu(\Pi_m^{(1)}) < \eta \le \nu(\Pi_n^{(2)}) \text{ for all } n \ge 1\} \ge qp/2.$$

Since the index rule chooses at every stage the action with the largest index, $N_\infty(1) \le m$ on the event $\{\nu(\Pi_m^{(1)}) < \eta \le \nu(\Pi_n^{(2)})$ for all $n \ge 1\}$, proving part (ii) of the theorem. *Q.E.D.*

Theorem 2 implies that under the measure $P_{\theta_1,\ldots,\theta_k}$, the posterior distribution of $(\mu(\theta_1),\ldots,\mu(\theta_k))$ does not shrink to the true value except for one component and that this component may vary among different realizations of the bandit process (i.e. it is a random variable instead of the desired j^*th component). If we consider the Bayesian (mixture) measure $P(A) = \int P_{\theta_1,\ldots,\theta_k}(A) d\Pi^{(1)}(\theta_1)\ldots d\Pi^{(k)}(\theta_k)$, the martingale convergence theorem implies that almost surely $[P]$, the vector of posterior means at stage n converges as $n \to \infty$ to a random vector, which differs from $(\mu(\theta_1),\ldots,\mu(\theta_k))$ in view of Theorem 2(i).

EXAMPLE 1: Suppose that the $Y_{j,t}$ are independent normal random variables with means θ_j and common variance 1 and that $\mu(\theta) = \theta$. Suppose that the θ_j are independent normal random variables with common mean μ_0 and variance v. Let $\overline{Y}_{j,n} = n^{-1}\Sigma_{t=1}^n Y_{j,t}$. Then

$$\hat{\mu}_{j,n} = \frac{\mu_0 v^{-1} + n\overline{Y}_{j,n}}{n + v^{-1}}, \quad \sigma_{j,n}^2 = (v^{-1} + n)^{-1}.$$

It is easy to see that conditions (C1)–(C3) all hold with $l_\Gamma = -\infty$ and $m = 1$.

EXAMPLE 2: Suppose that $Y_{j,t}$ are independent Bernoulli random variables such that $P_{\theta_j}\{Y_{j,t} = 1\} = \theta_j = 1 - P_{\theta_j}\{Y_{j,t} = 0\}$. Suppose that $\mu(\theta) = \theta$ and that the θ_j have independent prior distributions Π_j such that $\Pi_j(A) > 0$ for any open subset A of the unit interval (0,1). Then conditions (C1)–(C3) hold with $l_\Gamma = 0$, noting that $P_{\theta_j}(S_{j,m} = 0) = (1 - \theta_j)^m$, where $S_{j,m} = \Sigma_{t=1}^m Y_{j,t}$. In particular, if Π_j is a Beta(a,b) distribution, then $\hat{\mu}_{j,n} = (a + S_{j,n})/(a + b + n)$ and $\sigma_{j,n}^2 = (a + S_{j,n})(b + n - S_{j,n})/\{(a + b + n)^2(a + b + 2)\}$ clearly satisfy conditions (C1)–(C3).

Our proof of the incomplete learning theorem for discounted multi-armed bandits (Theorem 2) is considerably more direct and transparent than previous proofs in somewhat different contexts. Rothschild's (1974) seminal work considers the case $k = 2$ and Bernoulli distributions for the rewards, with success probabilities θ_1, θ_2. His approach is to express the dynamic programming equation defining the optimal rule as a functional equation involving the numbers of observed successes and failures from both arms, and to prove certain topological properties associated with the functional equation, leading to the construction of a set with positive probability under P_{θ_1,θ_2} on which the better arm is played only finitely many times. Banks and Sundaram (1992) assume general distributions for the rewards and independent priors for their parameters, as we do in Theorem 2, but that the priors have finite support so that the Gittins index of arm i is a function of the finite-dimensional vector of posterior probabilities for the possible values of θ_i. Their approach is to make use of the separating hyperplane theorem and the martingale convergence theorem to show the existence of a set with positive probability, under the mixture measure $P(= \int P_{\theta_1,\ldots,\theta_k} d\Pi^{(1)}(\theta_1)\ldots d\Pi^{(k)}(\theta_k))$, on which an inferior arm can be played forever once it is chosen. They have actually proved this result for the more general setting in which there are denumerably many (possible infinite) arms with the same parameter $\theta_i (1 \le i \le k)$ and with $\Pi^{(1)} = \ldots = \Pi^{(k)}$. They have also shown that

for such "stationary denumerable-armed bandits" the index rule is well defined and optimal. Clearly Theorem 2 and its proof can be easily extended to this setting if we redefine $N_t(j)$ as the total number of times that an action belonging to A_j has been used up to stage j, where A_j is the set of all actions with a common reward distribution that corresponds to the same parameter θ_j. Thus $N_t(j) = \sum_{n=1}^{t} I_{\{X_n \in A_j\}}$. In fact, with this new definition of $N_t(j)$, there is no loss of generality in the assumption $\mu(\theta_i) \neq \mu(\theta_j)$ for $i \neq j$ in Theorem 2.

Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, U.S.A.

Manuscript received December, 1998; final revision received August, 1999.

REFERENCES

AGHION, P., P. BOLTON, C. HARRIS, AND B. JULIEN (1991): "Optimal Learning by Experimentation," *Review of Economic Studies*, 58, 621–654.

BANKS, J. S., AND R. K. SUNDARAM (1992): "Denumerable-Armed Bandits," *Econometrica*, 60, 1071–1096.

BLUME, L. E., AND D. EASLEY (1984): "Rational Expectations Equilibrium: An Alternative Approach," *Journal of Economic Theory*, 34, 116–129.

BREZZI, M., AND T. L. LAI (1999): "Optimal Learning and Experimentation in Bandit Problems," forthcoming in *Journal of Economic Dynamics and Control*.

EASLEY, D., AND N. KIEFER (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 56, 1045–1064.

EL-GAMAL, M. A., AND R. K. SUNDARAM (1993): "Bayesian Economists... Bayesian Agents," *Journal of Economic Dynamics and Control*, 17, 355–383.

GITTINS, J. C. (1979): "Bandit Processes and Dynamic Allocation Indices," *Journal of Royal Statistical Society, Series B*, 41, 148–177.

——— (1989): *Multi-Armed Bandit Allocation Indices*. New York: Wiley.

GITTINS, J. C., AND D. M. JONES (1974): "A Dynamic Allocation Index for the Sequential Design of Experiments," in *Progress in Statistics*, ed. by J. Gani et al. Amsterdam: North Holland, 241–266.

JOVANOVIC, B. (1979): "Job-Search and the Theory of Turnover," *Journal of Political Economy*, 87, 972–990.

MCLENNAN, A. (1984): "Price Dispersion 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 *Handbook of Labor Economics*, Vol. 2, ed. by O. Ashenfelter and R. Layard. New York: North-Holland, 849–919.

ROTHSCHILD, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

WILLIAMS, D. (1991): *Probability with Martingales*. Cambridge: Cambridge University Press.

WHITTLE, P. (1980): "Multi-Armed Bandits and the Gittins Index," *Journal of the Royal Statistical Society, Series B*, 42, 143–149.