BANDIT PROBLEMS

By

Dirk Bergemann and Juuso Välimäki

January 2006

COWLES FOUNDATION DISCUSSION PAPER NO. 1551

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Bandit Problems*

Dirk Bergemann† Juuso Välimäki‡

January 2006

Abstract

We survey the literature on multi-armed bandit models and their applications in economics. The multi-armed bandit problem is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. This classic problem has received much attention in economics as it concisely models the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff).

JEL CLASSIFICATION: C72, C73, D43, D83.

KEYWORDS: One-Armed Bandit, Multi-Armed Bandit, Bayesian Learning, Experimentation, Index Policy, Matching, Experience Goods,

*The authors gratefully acknolwedge financial support through the National Science Foundation Grants CNS 0428422 and SES 0518929 and the Yrjö Jahnsson's Foundation. The survey was prepared as an entry for the New Palgrave Dictionary of Economics, 2nd edition.

†Department of Economics, Yale University, New Haven, CT 06520-8268, U.S.A., dirk.bergemann@yale.edu.

‡Department of Economics, Helsinki School of Economics and University of Southampton, 00100 Helsinki, Finland, juuso.valimaki@hse.fi

Introduction The multi-armed bandit problem, originally described by Robbins (1952), is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. In the multi-arm bandit problem, the gambler has to decide which arm of K different slot machines to play in a sequence of trials so as to maximize his reward. This classical problem has received much attention because of the simple model it provides of the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff). Each choice of an arm results in an immediate random payoff, but the process determining these payoffs evolves during the play of the bandit. The distinguishing feature of bandit problems is that the distribution of returns from one arm only changes when that arm is chosen. Hence the rewards from an arm do not depend on the rewards obtained from other arms. This feature also implies that the distributions of returns do not depend explicitly on calendar time.

Practical examples of the bandit problem include clinical trials where different treatments need to be experimented with while minimizing patient losses, or adaptive routing efforts for minimizing delays in a network. In an economics environment, experimental consumption is an example of intertemporal allocation problems where the trade-off between current payoff and value of information plays a key role. Alternatively, the use of arms may change their physical properties as in learning by doing where experience with the arm increases its future payoffs.

Basic Model It is easiest to formulate the bandit problem as an infinite horizon Markov decision problem in discrete time with time index $t = 0, 1, ...$ At each t, the decision maker chooses amongst K arms and we denote this

choice by $a_t \in \{1, ..., K\}$. If $a_t = k$, a random payoff x_t^k is realized and we denote the associated random variable by X_t^k. The state variable of the Markovian decision problem is given by s_t. We can then write the distribution of x_t^k as $F^k(\cdot; s_t)$. The state transition function ϕ depends on the choice of the arm and the realized payoff:

$$s_{t+1} = \phi\left(x_t^k; s_t\right)$$

Let S_t denote the set of all possible states in period t. A feasible Markov policy $a = \{a_t\}_{t=0}^{\infty}$ selects an available alternative for each conceivable state s_t, i.e.

$$a_t : S_t \to \{1, ..., K\}$$

The following two assumptions must be met for the problem to qualify as a bandit problem.

1. Payoffs are evaluated according to the discounted expected payoff criterion where the discount factor δ satisfies $0 \leq \delta < 1$.

2. The payoff from each k depends only on outcomes of periods with $a_t = k$. In other words, we can decompose the state variable s_t into K components $\left(s_t^1, ..., s_t^K\right)$ such that for all k :

$$\begin{aligned} s_{t+1}^k &= s_t^k \qquad \text{if} \quad a_t \neq k, \\ s_{t+1}^k &= \phi(s_t^k, x_t) \quad \text{if} \quad a_t = k, \end{aligned}$$

and

$$F^k(\cdot, s_t) = F^k\left(\cdot; s_t^k\right).$$

Notice that when the second assumption holds, the alternatives must be statistically independent.

It is easy to see that many situations of economic interest are special cases of the above formulation. First, it could be that $F^k\left(\cdot;\theta^k\right)$ is a fixed distribution with an unknown parameter θ^k. The state variable is then the posterior probability distribution on θ^k. Alternatively, $F^k\left(\cdot;s^k\right)$ could denote the random yield per period from a resource k after extracting s^k units.

The value function $V\left(s_0\right)$ of the bandit problem can be written as follows. Let $X^k\left(s_t^k\right)$ denote the random variable with distribution $F^k\left(\cdot;s_t^k\right)$. Then the problem of finding an optimal allocation policy is the solution to the following intertemporal optimization problem:

$$V\left(s_0\right)=\sup_a\left\{\mathbb{E}\sum_{t=0}^{\infty}\delta^t X^{a_t}\left(s_t^{a_t}\right)\right\}.$$

The celebrated index theorem due to Gittins and Jones (1974) transforms the problem of finding the optimal policy into a collection of k stopping problems. For each alternative k, we calculate the following index $\gamma^k\left(s_t^k\right)$, which only depends on the state variable of alternative k:

$$m^k\left(s_t^k\right)=\sup_\tau\left\{\frac{\mathbb{E}\sum_{u=t}^{\tau}\delta^t X^k\left(s_u^k\right)}{\mathbb{E}\sum_{u=t}^{\tau}\delta^t}\right\}, \tag{1}$$

where τ is a stopping time with respect to $\{s_t^k\}$. The idea is to find for each k the stopping time τ that results in the highest discounted expected return per discounted expected number of periods in operation. The Gittins index theorem then states that the optimal way of choosing arms in a bandit problem is to select in each period the arm with the highest Gittins index, $m^k\left(s_t^k\right)$, as defined by (1).

Theorem 1 (Gittins-Jones (1974))

The optimal policy satisfies $a_t=k$ for some k such that

$$m^k\left(s_t^k\right)\geq m^j\left(s_t^j\right)\ \textit{for all}\ j\in\{1,...,K\}.$$

To get the economic intuition behind this theorem, consider the following variation on the original problem. This reasoning follows the lines suggested in Weber (1992). The arms are owned and operated by separate risk neutral agents. The owner can rent a single arm at a time to the operators and there is a competitive market of potential operators. As time is discounted, it is clearly optimal to obtain high rental incomes in early periods of the model. The rental market is operated as a descending price auction where the fee for operating an arbitrary arm is lowered until an operator accepts the price. At the accepted price, the operator is allowed to operate the arm as long as it is profitable. Since the market for operators is competitive, the price is such that under an optimal stopping rule, the operator breaks even. Hence the highest acceptable price for arm k is the Gittins index $m^k\left(s_t^k\right)$, and the operator operates the arm until its Gittins index falls below the price, i.e. its original Gittins Index. Once an arm is abandoned, the process of lowering the price offer is restarted. Since the operators get zero surplus and they are operating under optimal rules, this method of allocating arms results in the maximal surplus to the owner and thus to the largest sum of expected discounted payoffs.

The optimality of the index policy reduces the dimensionality of the optimization problem. It says that the original $K-$dimensional problem can be split into K independent components, and then be knitted together after the solutions of the indices for the individual problems have been computed, as in formula (1). In particular, in each period of time, at most one index has to be re-evaluated, the other indices remain frozen.

The multi-armed bandit problem and many variations are presented in detail in Gittins (1989) and Berry and Fristedt (1985). An alternative proof of the main theorem, based on dynamic programming can be found in Whittle

(1982). The basic idea is to find for every arm a retirement value M_t^k, and then to choose in every period the arm with the highest retirement value. Formally, for every arm k and retirement value M, we can compute the optimal retirement policy given by:

$$V^k\left(s_t^k, M\right) \triangleq \max\left\{\mathbb{E}\left[X^k\left(s_u^k\right) + \delta V^k\left(s_t^{k+1}, M\right), M\right]\right\} \tag{2}$$

The auxiliary decision problem given by (2) compares in every period the trade-off between continuation with the reward process generated by arm k or stopping with a fixed retirement value M. The index of arm k in the state s_t^k is the highest retirement value at which the decision is just indifferent between continuing with arm k or retiring with $M = M\left(s_t^k\right)$:

$$M^k\left(s_t^k\right) = V^k\left(s_t^k, M^k\left(s_t^k\right)\right).$$

The resulting index $M^k\left(s_t^k\right)$ is equal to the discounted sum of flow index $m^k\left(s_t^k\right)$, or $M^k\left(s_t^k\right) = m^k\left(s_t^k\right)/\left(1-\delta\right)$.

Extensions Even though it is easy to write down the formula for the Gittins index and to give it an economic interpretation, it is normally impossible to obtain analytical solutions for the problem. One of the few settings where such solutions are possible is the continuous time bandit model where the drift of a Brownian motion process is initially unknown and learned through observations of the process. Karatzas (1984) provides an analysis of this case when the volatility parameter of the process is known.

From an analytical standpoint, the key property of bandit problems is that they allow for an optimal policy that is defined in terms of indices that are calculated for the individual arms. It turns out that this property does not generalize easily beyond the bandit problem setting. One instance where

such a generalization is possible is the branching bandit problem where new arms are born to replace the arm that was chosen in the previous period (see Whittle (1981)).

An index characterization of the optimal allocation policy can still be obtained without the Markovian structure. Varaiya, Walrand, and Buyukkoc (1985) give a general characterization in discrete time, and Karoui and Karatzas (1997) provide a similar result in a continuous time setting. In either case, the essential idea is that the evolution of each arm only depends on the (possibly entire) history and running time of the arm under consideration, but not on the realization nor the running time of the other arms. Banks and Sundaram (1992) show that the index characterization remains valid under some weak additional condition even if the number of indices is countable, but not necessarily finite.

On the other hand, it is well known that an index characterization is not possible when the decision maker must or can select more than a single arm at each t. Banks and Sundaram (1994) also show further that an index characterization is not possible when an extra cost must be paid to switch between arms in consecutive periods. Bergemann and Välimäki (2001) consider a stationary setting in which there is an infinite supply of ex ante identical arms available. Within the stationary setting, they show that an optimal policy follows the index characterization even when many arms can be selected at the same time or when a switching cost has to be paid to move from one arm to another.

Market Learning In economics, Bandit problems have first been used to model search processes. The first paper that used a one-armed bandit problem in economics is Rothschild (1974) in which a single firm is facing

a market with unknown demand. The true market demand is given by a specific probability distribution over consumer valuations. However the firm initially has a prior probability over several possible market demands. The problem for the firm is find an optimal sequence of prices to learn more about the true demand while maximizing its expected discounted profits. In particular, Rothschild shows that ex ante optimal pricing rules may well end up using prices that are ex post suboptimal (i.e. suboptimal if the true distribution were to be known). If several firms were to experiment independently in the same market, they might offer different prices in the long run. Optimal experimentation may therefore lead to price dispersion in the long run as shown formally in McLennan (1984).

In an extension of Rothschild, Keller and Rady (1999) consider the problem of the monopolist facing an unknown demand that is subject to random changes over time. In a continuous time model, they identify conditions on the probability of regime switch and discount rate under which either very low or very high intensity of experimentation is optimal. With a low intensity policy, the tracking of the actual demand is poor and the decision maker eventual becomes trapped, in contrast with a high intensity policy demand is tracked almost perfectly. Rustichini and Wolinsky (1995) examine the possibility of mis-pricing in a two-armed bandit problem when the frequency of change is small. Nonetheless, they show that it is possible that learning will cease even though the state of demand continues to change.

The choice between various research projects often takes the form of a bandit problem. In Weitzman (1979) each arm represents a distinct research project with a random prize associated with it. The issue is to characterize the optimal sequencing over time in which the projects should be undertaken. It shows that as novel projects provide an option value to the research, the op-

timal sequence is not necessarily the sequence of decreasing expected rewards (even when there is discounting). Roberts and Weitzman (1981) consider a richer model of choice between R&D processes.

Many Agent Experimentation The multi-armed bandit models have recently been used as a canonical model of experimentation in teams. In Bolton and Harris (1999) and Keller, Rady, and Cripps (2005) a set of players choose independently between the different arms. The reward distributions are fixed, but characterized by parameters that are initially unknown to the players. The model is one of common values in the sense that all players receive independent draws from the same distribution when choosing the same arm. It is assumed that outcomes in all periods are publicly observable and as a result a free riding problem is created. Information is a public good and each individual player would prefer to choose the current payoff maximizing arm and let other players perform costly experimentation with currently inferior arms. These papers characterize equilibrium experimentation under different assumptions on the reward distributions. In Bolton and Harris (1999), the model of uncertainty is a continuous time model with unknown drift and know variance, whereas in Keller, Rady, and Cripps (2005) the underlying uncertainty is modelled by an unknown Poisson parameter.

Experimentation and Matching The bandit framework have been successfully applied to learning in matching markets such as labor and consumer good markets. An early example of this is given in the job market matching model of Jovanovic (1979) who applies a bandit problem to a competitive labor markets. Suppose that a worker must choose employment in one of K firms and her (random) productivity in firm k is parametrized by a real

variable θ^k. The bandit problem is then a natural framework for the study of learning about the match specific productivities. For each k, s_0^k is then simply the prior on θ^k and s_t^k is the posterior distribution given s_0^k and x_s^k for $s < t$. Over time, a worker's productivity in a specific job becomes know more precisely. In the event of a poor match, separation occurs in equilibrium and job turnover arises as natural by-product of the learning process. On the other hand, over time the likelihood of separation eventually decreases, as conditional on being still on the job, the likelihood of a good match increases. The model generates hence a number of interesting empirical implications which have since been investigated extensively. Miller (1984) enriches the above setting by allowing for a priori different occupations, and hence the sequence in which a worker is matched over time to different occupations is determined as part of the equilibrium.

Experimentation and Pricing In a related recent literature, bandit problems have been taken as a starting point for the analysis of division of surplus in an uncertain environment. In the context of a differentiated product market and a labor market respectively, Bergemann and Välimäki (1996) and Felli and Harris (1996) consider a model with a single operator and a separate owner for each arm. The owners compete for the operator's services by offering rental prices. These models are interested in the efficiency and the division of the surplus resulting from the equilibrium of the model. In both models, arms are operated according to the Gittins index rule and the resulting division of surplus leaves the owners of the arms as well as the operator with positive surpluses. In Bergemann and Välimäki (1996), the model is set in discrete time and a general model of uncertainty is considered. They interpret the experiment as the problem of choosing among two

competing experience goods, in which both seller and buyer are uncertain about the quality of the match between the product and the preferences of the buyer. In contrast, Felli and Harris (1996) consider a continuous model with uncertainty represent by a Brownian motion and interpret the model in the context of a labor market. Both models show that even though the models allow for a genuine sharing of the surplus, allocation decisions are surplus maximizing in all Markovian equilibria and each competing seller receives his marginal contribution to the social surplus in the unique cautious Markovian equilibrium. Bergemann and Välimäki (2006) generalizes the above efficiency and equilibrium characterization from two sellers to an arbitrary finite number of sellers in a deterministic setting. Their proof uses some of the techniques first introduced in Karoui and Karatzas (1997). On the other hand, if the market consists of many buyers and each one of them is facing the same experimentation problem, then the issue of free-riding arises again. Bergemann and Välimäki (2002) analyzes a continuous time model as in Bolton and Harris (1999) but with strategic sellers. Surprisingly, the inefficiency observed in the earlier paper is now reversed and the market equilibrium displays too much information. As information is a public good, the seller has to compensate an individual buyer only for the impact his purchasing decision has on his own continuation value but not on the change in continuation value of the remaining buyers. As experimentation leads in expectation to more differentiation, hence less price competition, the sellers prefer more differentiation, hence more experimentation to less. As each seller only has to compensate the individual buyers, but not all buyers, the social price of the experiment is above the equilibrium price, leading to excess experimentation in equilibrium.

Experimentation in Finance Recently, the paradigm of the bandit model has also been applied in corporate finance and asset pricing. Bergemann and Hege (1998) and Bergemann and Hege (2005) model a new venture or innovation as a Poisson bandit model with variable learning intensity. The investor controls the flow of funding allocated to the new project and hence the rate at which information about the new project arrives. The optimal funding decision is subject to a moral hazard problem in which the entrepreneur controls the unobservable decision to allocate the funds to the project. Hong and Rady (2002) introduce experimentation in an asset pricing model with uncertain liquidity supply. In contrast to the standard noise trader model, the strategic seller can learn about liquidity from past prices and trading volume. This learning implies that strategic trades and market statistics such as informational efficiency are path-dependent on past market outcomes.

References

Banks, J., and R. Sundaram (1992): "Denumerable-Armed Bandits," *Econometrica*, 60, 1071–1096.

——— (1994): "Switching Costs and the Gittins Index," *Econometrica*, 62, 687–694.

Bergemann, D., and U. Hege (1998): "Dynamic Venture Capital Financing, Learning and Moral Hazard," *Journal of Banking and Finance*, 22, 703–735.

——— (2005): "The Financing of Innovation: Learning and Stopping," *RAND Journal of Economics*, 36, forthcoming.

Bergemann, D., and J. Välimäki (1996): "Learning and Strategic Pricing," *Econometrica*, 64, 1125–49.

——— (2001): "Stationary Multi Choice Bandit Problems," *Journal of Economic Dynamics and Control*, 25, 1585–1594.

——— (2002): "Information Acquisition and Efficient Mechanism Design," *Econometrica*, 70, 1007–1033.

——— (2006): "Dynamic Price Competition," *Journal of Economic Theory*, forthcoming.

Berry, D., and B. Fristedt (1985): *Bandit Problems*. Chapman and Hall, London.

Bolton, P., and C. Harris (1999): "Strategic Experimentation," *Econometrica*, 67, 349–374.

Felli, L., and C. Harris (1996): "Job Matching, Learning and Firm-Specific Human Capital," *Journal of Political Economy*, 104, 838–868.

Gittins, J. (1989): *Allocation Indices for Multi-Armed Bandits*. London, Wiley.

Gittins, J., and D. Jones (1974): "A Dynamic Allocation Index for the Sequential Allocation of Experiments," in *Progress in Statistics*, ed. by J. Gani, pp. 241–266. North-Holland, Amsterdam.

Hong, H., and S. Rady (2002): "Strategic Trading and Learning About Liquidity," *Journal of Financial Markets*, 5, 419–450.

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

KARATZAS, I. (1984): "Gittins Indices in the Dynamic Allocation Problem for Diffusion Processes," *Annals of Probability*, 12, 173–192.

KAROUI, N. E., AND I. KARATZAS (1997): "Synchronization and Optimality for Multi-Armed Bandit Problems in Continuous Time," *Computational and Applied Mathematics*, 16, 117–152.

KELLER, G., AND S. RADY (1999): "Optimal Experimentation in a Changing Environment," *Review of Economic Studies*, 66, 475–507.

KELLER, G., S. RADY, AND M. CRIPPS (2005): "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73, 39–68.

MCLENNAN, A. (1984): "Price Dispersion and Incomplete Learning in the Long-Run," *Journal of Economic Dynamics and Control*, 7, 331–347.

MILLER, R. (1984): "Job Matching and Occupational Choice," *Journal of Political Economy*, 92, 1086–1120.

ROBBINS, H. (1952): "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 55, 527–535.

ROBERTS, K., AND M. WEITZMAN (1981): "Funding Criteria for Research, Development and Exploration of Projects," *Econometrica*, 49, 1261–1288.

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

RUSTICHINI, A., AND A. WOLINSKY (1995): "Learning About Variable Demand in the Long Run," *Journal of Economic Dynamics and Control*, 19, 1283–1292.

VARAIYA, P., J. WALRAND, AND C. BUYUKKOC (1985): "Extensions of the Multiarmed Bandit Problem: The Discounted Case," *IEEE Transactions on Automatic Control*, AC-30, 426–439.

WEBER, R. (1992): "On the Gittins Index for Multi-Armed Bandits," *Annals of Applied Probability*, 2, 1024–1033.

WEITZMAN, M. (1979): "Optimal Search for the Best Alternative," *Econometrica*, 47, 641–654.

WHITTLE, P. (1981): "Arm-Acquiring Bandits," *Annals of Probability*, 9, 284–292.

——— (1982): *Optimization Over Time*, vol. 1. Wiley, Chichester.