JOURNAL OF ECONOMIC THEORY 9, 185–202 (1974)

A Two-Armed Bandit Theory of Market Pricing*

MICHAEL ROTHSCHILD

Princeton University, Princeton, New Jersey 08540

Received September 10, 1973

1. INTRODUCTION

Economics lacks a good theory of how stores should set their prices when they do not know the demand functions of their customers. Traditional theory either assumes that firms know their demand curves, or that they can, if necessary, find them out easily and costlessly from market experience. The market will inform the perfectly competitive firm of its demand function with ruthless efficiency. A firm which charges more than the market price will, as a matter of definition, lose all its customers. If it charges less than the market price, it will have all the customers it can handle. It is not difficult to explain how a perfectly competitive firm discovers what the market price is and all it needs to know of its demand function.

Conventional theories of monopoly and imperfect competition attribute to firms' complete knowledge of all relevant parameters of their demand relationships. Recently developed theories of firm behavior under uncertainty (for example, Baron [2] or Leland [10]) are not exceptions. The firm is typically supposed to face a stochastic demand function of the general form $D = D(\psi, \theta)$, where θ is a random variable and $D(\cdot, \cdot)$ a known function. The firm is assumed to maximize expected profits (or the expected utility of profits). The theory does not explain how the firm comes to know $D(\cdot, \cdot)$ nor is it concerned with explaining how the firm might come to know its demand function more accurately and thus reduce the uncertainty it faces.[1]

* This paper reports on work done jointly with Menahem Yaari. The basic idea of the model was worked out together and he is responsible for much that appears here. I am grateful to Kenneth Arrow, Peter Diamond, Joseph Gastwirth, Rose Preiss, and Eytan Sheshinski for helpful discussion and the National Science Foundation and the Industrial Relations Section of Princeton University for research support, but exampt them from responsibility. The description of the model in Section 2 follows closely the sketch in the Appendix to my paper [14].

[1] As it might if the stochastic demand function were of the form $D(p, \theta) = d(p) \cdot \theta$ (where $d(\cdot)$ is an ordinary demand function and θ a positive random variable) and the model allowed the firm to reduce the variance of θ (at a cost).

Theories recently developed to explain the behavior of economic agents in situations of disequilibrium and uncertainty also avoid the problem of what firms should do when they do not know their demand functions. These theories either tend to assume that firms are less intelligent or avaricious than they might be [7, 19, 20], or that they have much more information than they could reasonably have [6, 13].[2]

We attempt in this paper to fill this gap by constructing a highly stylized model of the behavior of rational and optimizing sellers in a market where they are initially ignorant of the demand curves they face. The basic idea of our model is very simple. A firm which does not know the consequences of charging a particular price has an obvious way of finding out. It may charge that price and observe the result. However, such experimental determination of its demand curve is costly. Customers turned away by a price which is objectively too high may not return, and sales made at prices which are too low represent losses which cannot be recouped. Formulating the firm's optimal strategy consists of finding some way of weighing the value of new information gained from charging a particular price against the cost of not charging the price which present information indicates is most profitable.

There are at least two reasons, aside from pure intellectual curiosity, for being interested in such a model. The first is simply that situations in which businesses are somewhat ignorant of the environment in which they operate are common and important. The airline industry would probably behave differently if the elasticity of demand for air travel were known. It is hard to imagine a serious study of the women's apparel industry which assumed that manufacturers knew what the demand for their products would be. Our attempts to understand such industries will be hampered unless we have satisfactory models of what their participants should, and possibly do, do.[3]

The second reason is that the major conlusion of our analysis is that, although it is possible to do so, it does not pay firms to acquire perfect information about their customers' demand functions. As a consequence some sellers will persist in making incorrect inferences about the nature of the market situation in which they are involved, and will continue to charge prices other than those they would charge if the demand function were known. Only some stores will make these errors. Others will charge the correct prices. Thus, our attempt to provide a reasonable theory of

[2] For an elaboration of these remarks, see [14].

[3] We ignore the question of how many firms consciously follow the strategies our model suggests they should. It may be that our work will have a normative significance for businessmen faced with the sorts of situations we describe. If this is the case, then the model will come to have more descriptive validity than it does at present.

how stores behave when they do not know their demand functions yields a solution to another theoretical problem: How to explain a diversity of prices within a single market. Following the fundamental work of Stigler [16], many authors have shown how consumers should behave in a market where the same good is being sold at a variety of prices.[4] None have presented convincing explanations for this diversity of price. Stigler attributes it to continual random shocks and limitations on the computational ability of market participants. He does not expect variety of prices to persists in stationary circumstances. The authors who followed Stigler have made even less of an effort to explain the price diversity which is such a crucial part of their models.

Since price variability is such a pervasive phenomenon, it seems unsatisfactory to regard it as simply an artifact of disequilibrium. Furthermore, the failure to explain price diversity means that Stigler's conclusions and those of his followers are based on analysis of one side of the market alone. Since the behavior of price setters is unexplained, it cannot be affected by consumer behavior.

As I have argued elsewhere [14], this sort of partial–partial equilibrium analysis can lead to paradoxes. Until we have a complete model of a market with persisting price variability, we should treat propositions about the causes and effects of price variability which are based on analysis of but one side of the market with great caution. We do not provide such a model here, but the results of this paper provide a foundation on which such a model can be built.

In the next section we present our model and show how it is identical to a problem familiar to students of probability theory—the two-armed bandit problem. In the third section we prove our main result: Optimal behavior can lead to a distribution of store prices. The final section discusses generalizations and implications. We note here that, although we describe our model in the language of retail stores, our analysis applies with only slight changes in vocabulary to labor markets.

2. The Model

Consider a store which is trying to price a particular item in its inventory. Potential customers arrive at a constant rate—which we may take to be one per period by an appropriate choice of units—unaffected by the

[4] See, for example, the papers in [12] and Gastwirth [8], Kohn and Shavell [9], Rothschild [15], and Telser [17].

price the store is charging.[5] The assumption that price has no influence on customer flow is an obviously limiting assumption and is in general not correct.[6] Still it is not an unreasonable assumption for some markets and some stores. Consider, for example, markets where repeat business cannot be expected either because customers buy the item in question so rarely that their trips to the store are separated by long periods of time in which the price policy of particular stores are largely forgotten, or because the customers are transients unlikely to return to the store under any circumstances. Examples are refrigerator dealers and cosmetic counters in airports. Another class of enterprises to which our model might apply is the store which sells many items. Although customer flow does depend on its overall price reputation, the pricing of a particular item will—within a range of possible prices—not affect this. Examples are supermarkets or department stores.[7] Customers are ignorant of the price of the item in question.[8] They enter the store, ask what the price is,[9] and then either buy or leave the store. From the point of view of the store, each potential customer is a binomial random variable which will buy one unit[10] with probability Π and not buy with probability $1 - \Pi$. The probability is a function of the price quoted to the customer. If the function $\Pi(p)$ were known, then the firm would in each period simply choose p to maximize expected profits, $\Pi(p)(p - c)$, where c is the constant cost of the item to the store owner.[11] However, if the firm does not know the function $\Pi(p)$, which in this model is all that not knowing the demand function of his customers[12] could mean, then what he should do is not so straightforward.

[5] Strict constancy of arrival rates is not essential to our analysis. Our results would go through if costomer arrivals were stochastic but independent of the store's price policy.

[6] For a model which describes store policy when customers' future behavior *is* a function of price, see Rothschild [13].

[7] Other examples might be real estate salesmen or used car salesmen. Problems are raised by the nonuniqueness of the items sold.

[8] This is a good assumption if the customers are truly transients—as in the airport drugstore example—or if the store sells too many items to advertise them effectively. Despite the wide circulation of their catalog, it is hard to imagine that it could become a matter of common knowledge that Sears sold blue jeans at $6.98.

[9] Explicit contact with a sales person is not required. Customers who visit supermarkets can be assumed to ask what prices are by observing them on shelves.

[10] We show below that it is trivial to drop the assumption that the customer buys but a single unit.

[11] For the case of a store which sells more than a single item, we ignore cross elasticities of demand.

[12] Again the model could be trivially generalized to take account of the store owner charging different prices to different customers. Thus the model does allow for haggling and bargaining. Unless we assume our store owner has some strategy for discerning

We can simplify the store owner's problem considerably by assuming that prices must be chosen from a finite set, indexed by i. If the store charges price p_i, then the true probability of a sale is Π_i. Again, if the store knew the parameters Π_i, it would simply choose i to maximize $\Pi_i q_i$, where $q_i = p_i - c$, the profit from making a sale at price p_i, but since the store does not, by assumption, know the values of the Π_i, its problem is more complex. However, the store can, if it chooses to, learn the value of any particular parameter Π_i. For Π_i is simply the probability of success in a single binomial trial. If the trial is repeated infinitely often, an observer will, according to the strong law of large numbers, be able to estimate Π_i exactly (that is, with probability one). The store could come to learn all the probabilities Π_i simply by choosing a strategy which involved charging each of the prices an infinite number of times—as, for instance, by playing them in turn. We shall show below that such asymptotically perfect information strategies are not optimal.

It should be clear from this discussion that each time the store charges a particular price p_i it can anticipate getting two things: a profit of q_i if a sale is made, and, whether the customer buys or not, some information about the demand function of his customers (in the form of a sharper estimate of the parameter Π_i). An optimal policy for a store involves putting a value on the potential gain in information from charging p_i. We show below that this can be done by setting up a dynamic programing equation in which the state variables describe the store owner's present beliefs about the demand functions of his customers.

Before we describe and analyze this dynamic programing problem, we shall point out that the problem faced by the store owner is one of a class of problems studied by statisticians and probabilists under the general heading of two-armed bandit problems. Consider a gambler condemned to, or bent on, putting a quarter into one of several (by convention the number is usually two) slot machines (one-armed bandits) from now until the end of time. The ith machine gives a payoff of q_i with probability Π_i, and nothing with probability $1 - \Pi_i$.[13] He wants

exactly what each customer's reservation price is, then he must guess. He may decide that different classes of customers have different reservation prices and decide that it makes sense to quote different prices to them. However, all that is necessary for the basic structure of our model to apply is that the store owner recognize that he can learn something about the behavior of future potential customers from the behavior—buying or not buying at a given price—of present customers.

[13] We see here how the generalization to the case where customers may purchase varying numbers of units is carried out. This is equivalent to a two-armed bandit problem in which the arms are multinomial rather than binomial random variables. It will be apparent that this generalization raises notational rather than substantive problems.

to choose a strategy which will maximize his expected discounted earnings or (to put the problem more realistically) minimize his expected discounted losses. If he knew how the two machines were set, he would achieve his goal simply by selecting the most favorable machine. However, it is assumed that, although he may have prior beliefs, the man does not know which machine offers the most favorable odds. The gambler's dilemma is formally identical to that of the store owner trying to decide what price to charge. (Prices are slot machines, payoffs are profits from sales, etc.) Looking at our model this way has two advantages: First, it allows us to take advantage of the insights of those who have studied two-armed bandit problems—in particular, Bellman [3].[14] Second, it permits us to use a much more vivid and intuitive vocabulary in the analysis below. We can use this language to state the major conclusion of our analysis.

If the discount rate is positive, almost all those who follow optimal strategies will after an initial period of sampling settle on one slot machine and play it in preference to all others. However, the machine chosen will not necessarily be the correct one. With positive probability a person persuing an optimal strategy will play the most favorable machine but a finite number of times while he plays a less attractive machine infinitely often.

The proof of this proposition, given in the next section, is somewhat involved. We give here a heuristic argument, which shows why it should be true. Consider a special case of the two-armed bandit problem. Suppose the probability of a payoff on one of the machines, say the second one, is known with certainty. The state of the player's information is described by his estimate of the probability of a payoff on the first machine. As we shall show, his choice of machine at any point in time is entirely determined by this estimate. When the first machine is played, the gambler receives, in addition to a random payoff, information which allows him to revise his estimate of the probability of a success on a first machine. When he plays the second machine, he receives only a payoff (if he is lucky) or nothing (if he is not). The outcome cannot affect his estimate of

[14] A brief note on the literature: The basic model seems to have been devised by Thompson [18] in 1933. Various extensions and elaborations have appeared since then. On the whole the literature appears quite fragmentary. Models and results vary from author to author. The model closest to ours was developed by Richard Bellman [3]. For recent surveys and expositions of two-armed bandit problems, see Berry [4], DeGroot [5], and Obregon [11]. Although the literature refers to *two*-armed bandit problems, the generalization of many results to an arbitrary finite number of arms is striaghtforward. In particular, Theorem I, below, goes through with minor notational changes.

the probability of success on the second machine since it is assumed this is known with subjective certainty. It also cannot affect his estimate of the probability of a payoff on the first machine. Thus, if the state of the player's information was such that optimal strategy dictated that the second machine be played, it is unchanged after the second machine has been played again. If the machine whose payoff probability is known is ever played, it will be played forever more. Thus, the optimal policy is in the form of a stopping rule. It is an answer to the question: When is information on the unknown arm so disappointing that play on it should be suspended? This question has an answer. There is a finite sequence of results on the unknown arm so discouraging that the rational gambler will abandon experimentation and play the known arm only. Even if the unknown arm is objectively better than the known arm, the gambler may observe such a sequence and thus play the unknown, but superior, device a finite number of times and the known, but inferior, device an infinite number of times. The proof of the next section uses a continuity argument to generalize the argument of the above paragraph (due, in essence, to Bellman [3]) to the case where the payoff probabilities of both devices are unknown.

3. Analysis of the Model

A. *Preliminaries*

We consider two machines (referred to also as devices and arms). Each trial on the ith machine yields payoff q_i with probability Π_i and nothing with probability $(1 - \Pi_i)$. The player does not know the parameters Π_i with certainty. He decides which machine to play at each stage after consulting his prior beliefs about the parameters Π_i and examining the record of successes and failures on the arms so far.

As is well known, T_i, the number of trials, and N_i, the number of successes, are a set of sufficient statistics for the parameter Π_i. It will occasionally be convenient to use a different set of sufficient statistics to represent the information in the sample. We call these statistics μ_i and ρ_i and define them by

$$\rho_i = \frac{1}{1 + T_i}, \tag{1}$$

$$\mu_i = \frac{N_i}{1 + T_i}. \tag{2}$$

These definitions imply particularly simple rules for updating (μ_i, ρ_i)

with experience. If arm i is played, ρ_i becomes $\rho_i/(1 + \rho_i)$. If there is a success on arm i, μ_i becomes

$$s(\mu_i) = \frac{(\mu_i + \rho_i)}{(1 + \rho_i)}, \tag{3}$$

while if there is failure μ_i becomes

$$f(\mu_i) = \frac{\mu_i}{(1 + \rho_i)}. \tag{4}$$

Although μ_i is not equal to the sample mean ($\bar{\mu}_i = N_i/T_i$), it approaches it as the number of trials increases; that is,

$$\lim_{\rho_i \to 0} \mu_i = \lim_{T_i \to \infty} \frac{N_i}{T_i + 1} = \lim_{T_i \to \infty} \frac{N_i}{T_i} = \lim_{T_i \to \infty} \bar{\mu}_i. \tag{5}$$

The information from the sample is contained in $(\mu_1, \mu_2, \rho_1, \rho_2)$, which we shall often write as (μ, ρ). With these conventions, sample information is confined to a fourfold copy of the closed unit interval [0, 1]. We shall call this subset of $\mathbf{R}^4$, Δ. If a success on arm 1 is observed, (μ, ρ) is updated to

$$s_1(\mu_1, \mu_2, \rho_1, \rho_2) = (s(\mu_1), \mu_2, \rho_1/(1 + \rho_1), \rho_2);$$

$s_2(\mu, \rho)$, $f_1(\mu, \rho)$, and $f_2(\mu, \rho)$ are defined in the obvious manner.

The player's prior beliefs about the parameters are summarized by a prior density function $g(\pi_1, \pi_2)$. If the player believes the probabilities of success on the two arms are independent, then the density factors and $g(\pi_1, \pi_2) = g_1(\pi_1)\, g_2(\pi_2)$. Most previous work on two-armed bandit problems has been restricted to the case of independent prior beliefs. We shall not adopt this restriction. Instead we work only with priors which are continuous and consider all nonextreme combinations of Π_1, Π_2 to be possible. That is,

$$g(\pi_1, \pi_2) > 0 \qquad \text{for all} \quad (\pi_1, \pi_2) \in (0, 1) \times (0, 1). \tag{6}$$

This assumption is not terribly restrictive. However, recalling the economic interpretation of the model, it does not allow store owners to know for certain that lowering the price will increase the probability of a sale. We require the store owners regard it as possible, albeit unlikely, that customers will respond perversely to price increases. We could relax (6) or replace it by more palatable assumptions, but it seems both complicated

and uninteresting to try to find weaker sufficient conditions (on store owners' priors) for our result.[15]

If a player has experience (μ, ρ), he will use Bayes' rule to update his prior beliefs from $g(\pi_1, \pi_2)$ to $h(\pi_1, \pi_2, \mu, \rho)$, the probability density proportional to

$$\pi_1^{\mu_1/\rho_1}(1-\pi_1)^{[1-(\mu_1+\rho_1)]/\rho_1}\,\pi_2^{\mu_2/\rho_2}(1-\pi_2)^{[1-(\mu_2+\rho_2)]/\rho_2}\,g(\pi_1, \pi_2).$$

Thus, if we define

$$\lambda_i(\mu, \rho) = \int_0^1 \int_0^1 \pi_i h(\pi_1, \pi_2, \mu, \rho)\, d\pi_1\, d\pi_2, \tag{7}$$

$\lambda_i(\mu, \rho)$ is the posterior mean of the players' belief about the value Π_i given the sample information (μ, ρ) and the prior density g. That is, the player, if he were risk neutral, would regard a bet, which paid \$1 for a success on arm i and 0 for a failure, as worth $\$\lambda_i(\mu, \rho)$.

Note that, although this interpretation of $\lambda_i(\mu, \rho)$ only makes sense when

$$\frac{\mu_i}{\rho_i} = N_i \qquad \text{and} \qquad \frac{1-(\mu_i+\rho_i)}{\rho_i} = T_i - N_i$$

are integers, the function $\lambda_i(\mu, \rho)$ is defined and is continuous for all (μ, ρ) such that $\rho_i > 0$, $i = 1, 2$. It may be extended continuously to the boundary of Δ since

$$\lim_{\rho_i \to 0} \lambda_i(\mu, \rho) = \mu_i. \tag{8}$$

This follows from the fact that $\lambda_i(\mu, \rho)$ is the mean of a posterior distribution based on $(1-\rho_i)/\rho_i$ observations. As the number of observations becomes large (as $\rho_i \to 0$), the posterior distribution approaches a normal distribution with mean equal to the sample mean $\tilde{\mu}_i$. Thus (8) follows from (5).

B. *The Dynamic Programming Equations and Their Properties*

In this section we show that for the player maximizing the expected discounted value of his return over an infinite horizon, there is a continuous

[15] Even in the economic interpretation of the model assumption (6) is not as unreasonable as it might appear. Elsewhere [15] I have shown that when demands observed by stores are caused by customers following optimal search rules, these demand functions will not necessarily be downward sloping unless all customers know the distribution of prices on the market with certainty.

real-valued function $V(\cdot\ \cdot)$ with domain Δ such that $V(\mu, \rho)$ is equal to the maximum expected discounted profits which the player can make if the present state of information is described by (μ, ρ). (Of course, $V(\mu, \rho)$ depends on g, but we shall ignore this dependence henceforth.) We show that $V(\mu, \rho)$ satisfies the basic functional equation

$$V(\mu, \rho) = \max_i W_i(\mu, \rho), \tag{9}$$

where

$$W_i(\mu, \rho) = \lambda_i(\mu, \rho)\, q_i + \delta[\lambda_i(\mu, \rho)\, V(s_i(\mu, \rho)) + (1 - \lambda_i(\mu, \rho))\, V(f_i(\mu, \rho))], \tag{10}$$

and δ is a positive discount factor less than one.

We prove the existence of V and W_i satisfying Eqs. (9) and (10) in the standard way—by induction. The only reason that we bother with this familiar excercise is that the continuity of V and W_i is crucial to our approach and we know no other way to establish this.

Define the functions $V^t(\mu, \rho)$, $W_i^t(\mu, \rho)$ by

$$V^0(\mu, \rho) = 0, \tag{11}$$

$$V^t(\mu, \rho) = \max_i W_i^t(\mu, \rho), \tag{12}$$

where

$$W_i^t(\mu, \rho) = \lambda_i(\mu, \rho)\, q_i + \delta[\lambda_i(\mu, \rho)\, V^{t-1}(s_i(\mu, \rho)) + (1 - \lambda_i(\mu, \rho))\, V^{t-1}(f_i(\mu, \rho))]. \tag{13}$$

Then $V^t(\mu, \rho)$ may be interpreted as the maximum value of the expected discounted winnings of a player who will gamble for only t more periods.

LEMMA 1. *The functions $V^t(\mu, \rho)$, $W_i^t(\mu, \rho)$ are continuous.*

Proof. We use an inductive argument. The continuity of $\lambda_i(\mu, \rho)$ and (11) establish the proposition for $t = 1$. Suppose $V^{t-1}(\mu, \rho)$ and $W_i^{t-1}(\mu, \rho)$ are continuous. Then $V^t(\mu, \rho) = \max_i(W_i^{t-1}(\mu, \rho))$ is continuous, and $W_i^t(\mu, \rho)$ is, by the induction hypothesis, a sum of continuous functions.

A similar inductive argument will show that $V^t(\mu, \rho)$, and $W_i^t(\mu, \rho)$ are monotone increasing in t. Furthermore, if M is any number greater than $\max_i(q_i)$, then $V_i^t(\mu, \rho) \leqslant M \sum_{\tau=0}^{t} \delta^\tau \leqslant M \sum_{\tau=0}^{\infty} \delta^\tau = M/(1 - \delta)$. The sequences $V^t(\mu, \rho)$, and $W_i^t(\mu, \rho)$ are bounded and converge.

Let $V(\mu, \rho) = \lim_{t\to\infty} V^t(\mu, \rho)$ and $W_i(\mu, \rho) = \lim_{t\to\infty} W_i^t(\mu, \rho)$. Clearly the functions $V(\mu, \rho)$ and $W^i(\mu, \rho)$ satisfy (9) and (10).

LEMMA 2. *$V^t(\mu, \rho)$ and $W_i^t(\mu, \rho)$ converge uniformly to $V(\mu, \rho)$ and $W_i(\mu, \rho)$, respectively.*

Proof. Let $\overline{V}^t(\mu, \rho)$ be the present discounted value of the expected sum of payments from the first t periods when following a policy implied by (9) and (10). Clearly $V^t(\mu, \rho) \geqslant \overline{V}^t(\mu, \rho)$, while

$$V(\mu, \rho) \leqslant \overline{V}^t(\mu, \rho) + \delta^t \sum_{\tau=0}^{\infty} \delta^\tau M \leqslant V^t(\mu, \rho) + \delta^t \frac{M}{1 - \delta},$$

so that $| V(\mu, \rho) - V^t(\mu, \rho)| \leqslant \delta^t(M/(1 - \delta))$. This inequality is independent of (μ, ρ), so that uniform convergence is established.

The uniform convergence of the $W_i^t(\mu, \rho)$ can be demonstrated in a similar manner. An immediate consequence is the following lemma.

LEMMA 3. *$V(\mu, \rho)$ and $W_i(\mu, \rho)$ are continuous on Δ.*

The principle of optimality states that an optimal policy for the two-armed bandit problem is to play arm i whenever

$$(\mu, \rho) \in A_i = \{(\mu, \rho) \in \Delta \mid j \neq i \text{ implies } W_i(\mu, \rho) > W_j(\mu, \rho)\}.$$

(It does not matter what arm is played if (μ, ρ) belongs to neither A_1 nor A_2.) We have the following important consequence of the continuity of $W_i(\mu, \rho)$.

LEMMA 4. *A_i is an open set.*

Proof. We shall show that for every point $(\bar{\mu}, \bar{\rho}) \in A_i$ there is an open set $U \subset A_i$ such that $(\bar{\mu}, \bar{\rho}) \in U$. By hypothesis there are numbers α, β such that, if $j \neq i$, $W_i(\bar{\mu}, \bar{\rho}) > \alpha > \beta > W_j(\bar{\mu}, \bar{\rho})$. Let

$$\overline{U} = \{(\mu, \rho) \in \Delta \mid W_i(\mu, \rho) > \alpha\} \quad \text{and} \quad \underline{U} = \{(\mu, \rho) \in \Delta \mid W_j(\mu, \rho) < \beta\}.$$

Then $\overline{U}$ and $\underline{U}$ are open sets (because $W_i(\mu, \rho)$ and $W_j(\mu, \rho)$ are continuous), and $(\bar{\mu}, \bar{\rho}) \in U = \overline{U} \cap \underline{U}$, an open set contained in A_i.

C. *The Basic Results*

We are now in a position to state our main result.

THEOREM I. *If the true parameters satisfy*

$$0 < \Pi_1 < \Pi_2 < 1, \tag{14}$$

then a player who follows an optimal strategy will with positive probability play machine 1 *infinitely often and machine* 2 *only a finite number of times.*

The proof uses Lemmas 3 and 4 to extend the heuristic argument given above. That argument was based on the player having perfect information about one of the parameters Π_i. In this context perfect information about Π_i corresponds to an infinite amount of experience with arm i or to $\rho_i = 0$. The continuity of $V(\mu, \rho)$ and $W(\mu, \rho)$ and the openness of A_i state precisely that a lot of information is very much like perfect information.

Proof. To begin we must specify how we are assigning probabilities to outcomes of plays. For $i = 1, 2$ let Ω_i be the space of all outcomes of an infinite sequence of trials on arm i; thus elements of Ω_i are infinite sequences of 0's and 1's. Let $\mathscr{B}_i$ be the σ-algebra generated by finite subsets of these sequences and P_{Π_i} the measure induced on Ω_i by the probabilities of success on successive trials being independent and equal to Π_i. Then, the space of outcomes is the probability space $(\Omega, \mathscr{B}, P)$, where $\Omega = \Omega_1 \times \Omega_2$, $\mathscr{B} = \mathscr{B}_1 \times \mathscr{B}_2$, and the measure P is defined by $P(B_1 \times B_2) = P_{\Pi_1}(B_1)\, P_{\Pi_2}(B_2)$ for any $(B_1 \times B_2) \in \mathscr{B}$.

There are several optimal strategies, since if $W_1(\mu, \rho) = W_2(\mu, \rho)$ it does not matter which arm the player chooses when his information is (μ, ρ). For the remainder of the proof we assume a particular optimal strategy has been chosen although we do not specify what it is beyond requiring that arm i be chosen whenever $(\mu, \rho) \in A_i$. Given this optimal strategy and an $\omega \in \Omega$, one may in principle calculate the number of times each arm is played. Let $T_i(\omega)$ be the number of times arm i is played when the sequences ω are observed, and define

$$F_i(t) = \{\omega \in \Omega \mid \text{arm } i \text{ is played on trial } i\}$$
$$F_i = \{\omega \in \Omega \mid T_i(\omega) = \infty\}.$$

Since the $F_i(t)$ are measurable, $F_i = \limsup_t F_i(t)$ is measurable as is its complement $E_i = \{\omega \in \Omega \mid T_i(\omega) < \infty\}$. Thus, P assigns probability to these sets. To prove the theorem we must show that $P(F_1 \cap E_2) > 0$.

To do this we first show that if a player is sufficiently sure that playing arm 2 is hopeless then he will play arm 1 if he thinks it sufficiently likely that playing arm 1 will have positive results. This is the content of the following, somewhat technical lemma.

LEMMA 5. *For every δ_1, δ_2 such that $\Pi_1 > \delta_1 > 0$ and $1 > \delta_2 > 0$, there is an $\epsilon > 0$ such that $W_1(\mu, \rho) > W_2(\mu, \rho)$ whenever $\mu_2 + \rho < \epsilon$ and either $\mu_1 \geqslant \Pi_1 - \delta_1$ or $\rho_1 \geqslant \delta_2$.*

Proof. Consider the compact set

$$K = \{(\mu_1, \mu_2, \rho_1, \rho_2) \in A \mid \mu_1 \geqslant \Pi_1 - \delta_1 \text{ or } \rho_1 \geqslant \delta_2 \text{ and } \mu_2 = \rho_2 = 0\}.$$

Then $K \subset A_1$ for

$$W_1(\mu, 0, \rho, 0) \geqslant \lambda_1(\mu, 0, \rho, 0)\, q_1 > 0,$$

while if

$$W_2(\mu, 0, \rho, 0) \geqslant W_1(\mu, 0, \rho, 0)$$

then

$$\begin{aligned} V(\mu, 0, \rho, 0) &= W_2(\mu, 0, \rho, 0) = 0 + \delta(0 \cdot V(\mu, 0, \rho, 0)) + 1V(\mu, 0, \rho, 0) \\ &= \delta V(\mu, 0, \rho, 0), \end{aligned}$$

implying

$$V(\mu, 0, \rho, 0) = 0 < W_1(\mu, 0, \rho, 0).$$

Since A_1 is open, centered at every point (μ, ρ), K is an open ball $B(x, r)$ of radius r [16] (which depends on (μ, ρ)) contained in A_1. These balls cover K, and thus there is a finite index set I such that

$$\bigcup_{i \in I} B(x_i, r_i) \supset K.$$

Let

$$\epsilon = \min_{i \in I} r_i .$$

Then any point $(\mu_1, \mu_2, \rho_1, \rho_2)$, such that $\mu_2 + \rho_2 < \epsilon$ and either $\mu_1 \geqslant \Pi_1 - \delta$ or $\rho_1 \geqslant \delta_2$, belongs to one of the $B(x_i, r_i)$ and thus to A_i. This completes the proof.

This lemma states that poor experience on arm 2, if sufficiently bad, will lead the player to play arm 1. To complete the proof we only must show that it is possible that experience on arm 1 is not so erratic that he will never play arm 2 again. Recall that, while $(\mu_1, \mu_2, \rho_1, \rho_2)$ describes a player's experience on both arms, the parameters μ_1 and ρ_1 depend only on his experience on arm 1. Let $\mu_1(\omega_1(t))$ be the value of μ_1 after observing the first t elements of $\omega_1 \in \Omega_1$, and define $\rho_1(\omega_1(t))$ similarly.

Consider

$$\begin{aligned} G(\delta_1, \delta_2) = \{\omega_1 \in \Omega_1 \mid \mu_1(\omega_1(t)) &\geqslant \Pi_1 - \delta_1, \\ &\text{whenever } \rho_1(\omega_1(t)) = (1+t)^{-1} \leqslant \delta_2\}. \end{aligned}$$

Suppose $\omega_1 \in G(\delta_1, \delta_2)$ for some $\delta_1 > 0$, $\delta_2 > 0$ and that, for the $\epsilon(\delta_1, \delta_2)$ whose existence is guarenteed by Lemma 5,

$$\begin{aligned} &\omega_2 \in H(\epsilon(\delta_1, \delta_2)) \\ &\quad = \{\omega_2 \in \Omega_2 \mid \rho_2(\omega_2(t)) + \mu_2(\omega_2(t)) < \epsilon(\delta_1, \delta_2) \text{ for some } t\}. \end{aligned}$$

[16] We adopt the max norm: $\| x_1, x_2, x_3, x_4 \| = \max_i | x_i |$.

Then, by construction, $\omega = (\omega_1, \omega_2) \in (F_1 \cap E_2)$. To complete the proof we must show that there exist δ_1, δ_2 such that

$$P(G(\delta_1, \delta_2) \times H(\epsilon(\delta_1, \delta_2)) = P_{\Pi_1}(G(\delta_1, \delta_2)\, P_{\Pi_2} H(\epsilon(\delta_1, \delta_2)) > 0.$$

Fix $\delta_1 > 0$. Let

$$Q = \{\omega_1 \in \Omega_1 \mid \lim_{t\to\infty} \mu_1(\omega_1(t)) > \Pi_1 - \delta\}.$$

Then by the law of large numbers $P_{\Pi_1}(Q) = 1$. For $u > 0$, define

$$Q_u = \{\omega_1 \in \Omega_1 \mid \mu_1(\omega_1(u) \leqslant \Pi_1 - \delta_1, \mu_1(\omega_1(t)) > \Pi_1 - \delta_1 \text{ for all } t > u\}.$$

If $u \neq v$, $Q_u \cap Q_v = \varnothing$ and, with the obvious definition of Q_0,

$$Q = \bigcup_{u=0}^{\infty} Q_u,$$

so that

$$\sum_{u=0}^{\infty} P_{\Pi_1}(Q_u) = P_{\Pi_1}(Q) = 1.$$

Thus $P_{\Pi_1}(Q_t) > 0$ for some t. But

$$Q_t \subset G(\delta_1, (2 + t)^{-1}),$$

and thus

$$P_{\Pi_1}(G(\delta_1, (2 + t)^{-1})) > 0.$$

Furthermore, it is clear that $P_{\Pi_2}(H(\epsilon)) > 0$ for all $\epsilon > 0$, since if

$$J = \{\omega_2 \in \Omega_2 \mid \omega_2(t) = 0 \text{ for all } t < \epsilon^{-1} + 2\}$$

then $J \subset H(\epsilon)$ and $P_{\Pi_2}(J) > 0$. This completes the proof.

The construction makes no use of the fact that $\Pi_2 > \Pi_1$. A similar argument can be made to show that there is a set U of positive probability in Ω such that $\omega \in U$ implies arm 1 is played only a finite number of times.

A natural question to ask is whether gamblers pursuing optimal policies will play more than one arm infinitely often or whether each gambler will eventually settle on one particular arm and play it to the exclusion of all others. It seems reasonable that, if two arms are not objectively identical, then if each is played often enough the gambler will come to know which arm is really better. If they are the same, no amount of

experience could force him to choose between the two. This is the content and the basis of the proof of the following theorem.

THEOREM II. *If*

$$\Pi_i q_i \neq \Pi_j q_j \qquad \text{for} \quad i \neq j, \tag{15}$$

then optimal strategy will, with probability one, lead to but one arm being played infinitely often.

Proof. Suppose that $\Pi_1 q_1 > \Pi_2 q_2$. Then $(\Pi_1, 0, \Pi_2, 0) \in A_1$ and there exist $\epsilon > 0$ and $\delta > 0$ such that $\mu_i \in J_i(\epsilon) = (\Pi_i - \epsilon, \Pi_i + \epsilon)$ and $\rho_i < \delta$ for $i = 1, 2$ implies $(\mu_1, \rho_1, \mu_2, \rho_1) \in A_1$. Let

$$Q_i = \{\omega_i \in \Omega_i \mid \lim_{t \to \infty} \mu_i(\omega_i(t)) \in J_i(\epsilon)\}.$$

Then $P_{\Pi_i}(Q_i) = 1$. If

$$Q_i(t) = \{\omega_i \in \Omega \mid \mu_i(\omega_i(t)) \notin J_i(\epsilon), \mu_i(\omega_i(u)) \in J_i(\epsilon) \text{ for all } u > t\},$$

then $Q_i(t) \cap Q_i(u) = \varnothing$ if $u \neq t$ and

$$\bigcup_{t=0}^{\infty} Q_i(t) = Q_i .$$

Let $R_{tu} = Q_1(t) \times Q_2(u)$. Then

$$\sum_{t=0}^{\infty} \sum_{u=0}^{\infty} P(R_{tu}) = \sum_{t=0}^{\infty} P_{\Pi_1}(Q_1(t)) \sum_{u=0}^{\infty} P_{\Pi_2}(Q_2(u)) = 1.$$

It will suffice to show that if $\omega \in R_{tu}$ then both arms cannot be played infinitely often. Suppose $\omega \in R_{tu}$ and $(T+1)^{-1} < \delta$. Then if $M \geqslant \max(t, u, T)$ the optimal strategy can lead to each arm being played at most M times. For after $M+1$ plays on each arm, $(\mu, \rho) \in A_1$ on all succeeding plays. This completes the proof.

4. CONCLUSION

Returning to the economic interpretation of the model, these theorems state that, if a store owner operates in a market in which he does not know the demand functions of his customers, then if he chooses his prices optimally he will eventually end up charging one single price and sticking to it; however, this price may be any one of several prices. Nothing guarantees that he will choose the right price (that is the price he would

have chosen had he known his customers' demand function). If there are many stores operating in the same market, and if they all follow optimal strategies, then it is almost certain that a distribution of prices will result, even though the stores have identical market opportunities and costs of production.

Of course these results were obtained only for a simple and rather special example. I assumed that customers would buy at most a single unit of the commodity, that there were only two prices that stores could charge, that stores were not aware of each others' existence, and that they did not compete with each other. These results continue to be valid when some of these simplifications are abandoned, but it is not clear what happens when others are dispensed with. The above proofs would go through if customers were to buy several units of the commodity; the store is only interested in estimating the expected profit which results from charging a particular price. This can always be done perfectly—but possibly at great cost—by charging that price infinitely often. Similarly, the proofs did not depend on the player having only two choices; they would go through with only minor changes if he were allowed any finite number of choices. However prices are often considered to be naturally continuous variables, and it is not clear that Theorems I and II hold when they are. This seems an open and difficult question.

In the same manner the model assumes stores behave as if they were the only stores in the market. One could well ask whether they would be content charging the prices that they think are best while observing that other stores—presumably rational—are charging different prices. I do not think this is a particularly compelling point. Unless store A has access to store B's books, the mere fact that store B is charging a price different from A's and not going bankrupt is not conclusive evidence that A is doing the wrong thing. Who is to say A's experience is not a better guide to the true state of affairs than B's? Of course, if A had access to B's books, then he would eventually know who was doing the more sensible thing. But competitors do not usually exchange this information, so this does not seem a particularly important objection.

It is a more serious problem that the model does not allow stores to compete with one another in the most natural way—by attracting potential customers at the expense of other stores. The model assumed that the flow of customers was independent of the stores' actions. This is a crucial assumption, and I do not know how to abandon it. Some rudimentary and transitory brand loyalty could be introduced by assuming, along the lines of [13], that new customers search at random while old customers (who eventually die) search in a way which is conditioned by their experience, but this is far from a satisfactory answer to this objection.

In defense of the results, it can be said that the model did not incorporate two things which would tend to create and perpetuate price diversity. We assumed that stores' information processing facilities were both limitless and costless to operate; we did not allow stores to entertain the possibility that they might be operating in a changing environment. If either of these assumptions is abandoned, it is clear that a store which is initially unsure of the demand curve it faces will remain forever less than perfectly informed about it; the price it charges at any time will be determined by beliefs which experience is too weak to contradict and by the accidents of its recent history as well as by "objective" market conditions.

The primary import of these results is that, if reasonable assumptions are made about the difficulties of obtaining information, then the assumption that economic agents optimize serves to constrain their actions much less than when these difficulties are ignored. For example, if one is to explain how a distribution of prices can persist as an equilibrium on some market, one must explain why or how rational stores should choose to charge different prices when they face what is objectively the same situation. One answer could be that this distribution of prices was a sort of Nash equilibrium, that each price charged maximized the store's profits given the prices at all otker stores. On the basis of some preliminary work I have done on this problem with Arrow [1], I am inclined to believe that this can only happen in rare circumstances. A more natural explanation for price diversity would be one based on the results of this paper.

REFERENCES

1. K. J. ARROW AND M. ROTHSCHILD, Equilibrium price distributions with limited information, paper presented at the Dec. 1973 meetings of the Econometric society.
2. D. P. BARON, Price uncertainty, utility, and industry equilibrium in pure competition, *Int. Econ. Rev.* **11** (October 1970), 463–480.
3. R. BELLMAN, A problem in the sequential design of experiments, *Sankhȳa* **16** (1956), 221–229.
4. D. A. BERRY, A Bernouili, two-armed bandit, *Ann. Math. Stat.* **43** (1972), 871–897.
5. M. H. DEGROOT, "Optimal Statistical Decisions," McGraw–Hili, New York, 1970.
6. P. A. DIAMOND, A model of price adjustment, *J. Econ. Theory* **3** (1971), 156–168.
7. F. M. FISHER, Quasi-competitive price adjustment by individual firms: A preliminary paper, *J. Econ. Theory* **2** (1970), 195–206.
8. J. L. GASTWIRTH, On probabilistic models of consumer search for information, *Quart. J. Econ.*, to appear
9. M. KOHN AND S. SHAVELL, The theory of search, *J. Econ. Theory* **9** (1974).
10. H. LELAND, The theory of the firm facing uncertain demand, *Amer. Econ. Rev.* **72** (1972), 278–291.
11. I. OBREGON, The *N*-armed bandit problem and other topics in sequential decision processes, Department of Operations Research, Massachusetts Institute of Technology, Cambridge, MA, 1968.

12. E. S. PHELPS *et al.*, "Microeconomic Foundations of Employment and Inflation Theory," Norton, New York, 1970.
13. M. ROTHSCHILD, Prices, information and market structure, unpublished (1970).
14. M. ROTHSCHILD, Models of market organization with imperfect information: A survey, *J. Polit. Econ.* **81**, (November 1973), 1283–1308.
15. M. ROTHSCHILD, Searching for the lowest price when the distribution of prices is unknown, *J. Polit. Econ.* **82**, (July 1974), 689–712.
16. G. J. STIGLER, The economics of information, *J. Polit. Econ.* **69** (1961), 213–225.
17. L. TELSER, Searching for the lowest price, *Amer. Econ. Rev. Proc.* **63** (1973), 41–49.
18. W. R. THOMPSON, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika* **25** (1933), 285–394.
19. S. WINTER, JR., Satisficing, selection, and the innovating remnant, *Quart. J. Econ.* **85** (1971), 237–261.
20. S. WINTER, JR., An SSRI model of markup pricing, unpublished (1971).