Econometrica, Vol. 64, No. 5 (September, 1996), 1125–1149

LEARNING AND STRATEGIC PRICING

BY DIRK BERGEMANN AND JUUSO VÄLIMÄKI[1]

We consider the situation where a single consumer buys a stream of goods from different sellers over time. The true value of each seller's product to the buyer is initially unknown. Additional information can be gained only by experimentation. For exogeneously given prices the buyer's problem is a multi-armed bandit problem. The innovation in this paper is to endogenize the cost of experimentation to the consumer by allowing for price competition between the sellers. The role of prices is then to allocate intertemporally the costs and benefits of learning between buyer and sellers. We examine how strategic aspects of the oligopoly model interact with the learning process.

All Markov perfect equilibria (MPE) are efficient. We identify an equilibrium which besides its unique robustness properties has a strikingly simple, seemingly myopic pricing rule. Prices below marginal cost emerge naturally to sustain experimentation. Intertemporal exchange of the gains of learning is necessary to support efficient experimentation. We analyze the asymptotic behavior of the equilibria.

KEYWORDS: Learning, experimentation, dynamic oligopoly, Markov perfect equilibrium, infinite stochastic game, multi-armed bandit.

1. INTRODUCTION

MUCH OF THE EXISTING LITERATURE on dynamic choice under uncertainty has focused on the case where a single decision-maker chooses sequentially among a fixed set of alternatives. In many economic situations, the alternatives are supplied by a separate economic agent (or a group of economic agents) and the decision theoretic analysis then provides only a description of the demand side of the market.

We develop a simple dynamic equilibrium model of price formation under learning and uncertainty. In an infinite horizon model with price competition, a buyer chooses sequentially between products whose qualities are initially unknown to all parties in the model; the buyer does not know the underlying characteristics of the products while the producers are uncertain about the tastes of the buyer. Each purchase yields additional information about the true product quality to all parties in the model.

In the decision theoretic situation where prices are exogenously fixed, it is well known that the optimal purchasing strategy by the buyer may involve experimentation. That is, in some periods the buyer is willing to sacrifice some

[1]We would like to express our gratitude to George J. Mailath for his encouraging support since the very beginning of this project. We thank Dieter Balkenborg, Pierpaolo Battigalli, Bruno Biais, In-Koo Cho, Matthias Kahl, Richard Kihlstrom, Roger Lagunoff, Stephen Morris, Andrew Postlewaite, Rafael Rob, and Jean Tirole for many helpful comments. We are grateful to an editor and two anonymous referees for very helpful comments and suggestions. The first author would like to thank Avner Shaked for his hospitality during a stay at the SFB 303, University of Bonn. Financial support by the German Science Foundation and Yrjö Jahnsson Foundation is gratefully acknowledged.

of her *current* payoff in order to gain additional information which is valuable for *future* decisions. This temporal separation of costs and benefits causes no ex ante efficiency losses in the single player case since the costs of experimentation have to be born by the same agent who enjoys any gains from successful experiments. However, if prices are set by profit maximizing producers, a nontrivial problem of intertemporal allocation arises as current and future prices determine the costs and benefits of experimentation to all the parties in the model.

To illustrate the point, consider a factory manager (the buyer) choosing between two alternative technologies supplied by outside contractors (the sellers). In each period, she signs a one period lease with one of the contractors. The output from each technology is a random variable depending on both the true productivity of the technology in the factory (the value of the match between the factory and the technology) and some outside random effects. If output is publicly observable, then all parties receive a common noisy signal about the true productivity of the technology chosen in each period. Beliefs are updated in a Bayesian fashion and posterior beliefs determine the relative competitive positions of the two contractors. We want to compare the incentives of the factory manager in the equilibrium model, where the contractors react optimally to changes in beliefs, to the decision theoretic model, where prices are exogenously fixed. In particular, the incentives to undertake experimentation may be fundamentally different under the two scenarios.

Suppose that the beliefs about product qualities are such that the optimal strategy in the decision theoretic case suggests experimentation (i.e., choosing the product with lower expected quality) in the current period. Is the buyer still willing to pay for an experiment in the equilibrium model if the outcomes of experiments are publicly observed? A bad outcome in the experiment results in more pessimistic beliefs on the quality of the product purchased. If the buyer decides to switch suppliers, the bad outcome results in a higher price to be paid in the next period since the relative competitive position of the nonselling firm has improved. On the other hand, if the outcome of the experiment is good, the buyer becomes more optimistic about the product quality and as a consequence, her willingness to pay for the product is increased. Since the outcome is publicly observable, the seller observes an increase in her monopoly power relative to the competitors and has an incentive to raise the price in future periods to appropriate the maximal amount of consumer surplus from the buyer. Since neither a good nor a bad outcome leads to an increase in consumer surplus, the willingness of the buyer to experiment is reduced. If it is in the firms' interest to sustain experimentation, the consumer will have to be compensated for experimentation costs in the current period.

For the case of two sellers and one buyer, we characterize the set of Markov perfect equilibria. The central result of the paper states that in spite of future rent seeking by the firms, all Markov perfect equilibria in the model are efficient, i.e. the discounted expected sum of consumer surplus and the two

firms' profits is maximized along any equilibrium path.[2] In particular, an efficient amount of experimentation is undertaken on any Markov perfect equilibrium path. Using this fact, we can deduce the sequencing of consumer purchases immediately, since the efficient paths coincide with the solution paths of the buyer's decision problem when prices are fixed to be identically zero. A solution for this decision problem is available in the statistical literature on multi-armed bandits. The remaining task is thus to calculate the prices that support efficient experimentation in equilibrium and determine the division of surplus between the buyer and the sellers along the efficient path.

In analogy with the one-shot pricing game with heterogeneous product quality, we need a refinement similar in nature to trembling hand perfection to select a unique equilibrium. In this equilibrium, which we call the cautious equilibrium, current prices provide the buyer with insurance against future rent seeking resulting from successful experiments. Note, however, that a bad outcome for the currently selling firm is a good outcome for the nonselling firm. As a consequence, prices do not provide insurance against bad outcomes in the experimentation. The buyer is left worse off since the nonselling firm acts as an outside option for the buyer in the determination of selling prices. Rent seeking by the nonselling firm in these contingencies allows the selling firm to charge higher prices. The equilibrium pricing rule is quite simple. In each period, the selling price is equal to the difference in expected qualities and is hence similar to the equilibrium price in the myopic Bertrand game. The identity of the seller does not, however, coincide with the myopic game since the efficient path involves experimentation at some nodes.

Recent papers by Smith (1992) and Bolton and Harris (1993) also introduce dynamic learning models with many agents. Smith considers a sequence of sellers entering the market. Each seller can individually observe the fraction of incumbents charging a low price, before solving his own bandit problem. He shows that the most profitable pricing option is eventually chosen by the market with probability one, as opposed to the case of an individual seller, who might charge the less profitable price forever even under optimal learning, as Rothschild (1974) showed. Bolton and Harris introduce strategic interaction in a learning model in which N players face simultaneously the same experimentation problem. Although the alternatives are still exogenously given, the informational externality, which arises through the public good aspect of experimentation, transforms the bandit problem into a game of strategic experimentation. The idea of an informational externality arising in a sequential learning model is already central to Rob (1991), who studies a dynamic model of entry when the size of the market is uncertain. In our work, public observability of the signals creates no informational externalities since the product qualities are assumed to

[2] With multiple buyers and publicly observed signals, the externalities involved in the experimentation process cannot be fully internalized by the firms and hence the equilibrium path will differ from the efficient path, in general. The restriction to two sellers is made purely for expositional convenience. We discuss various extensions to the model in Section 3.

be statistically independent. With multiple buyers and publicly observed signals, free riding on other buyers' experimentation becomes a problem. The firms are, however, able to internalize this externality to a large extent and this may reverse the typical results on underinvestment in information as discussed in Section 3.

While we consider the situation where a single consumer makes purchases over time in an oligopolistic market, several other economic applications could be analyzed within our framework. The bandit framework is often used as a matching model in the analysis of the labor market as in Jovanovic (1979) and Miller (1984), which ignore however the aspect of strategic interaction between employee and employer.[3] The choice and financing of new and uncertain technologies and R&D projects also fits exactly into the framework we develop in this paper.

A brief outline of the paper follows. The duopoly model is introduced in Section 2. In Section 3 we investigate the efficiency of the Markov perfect equilibria for the infinite game. In Section 4 we single out a particular equilibrium (we call it *cautious equilibrium*), which besides its unique robustness properties, has very appealing economic features. Subsequently we analyze the asymptotic properties of the equilibria and characterize the entire set of Markov perfect equilibria by the lower and upper bounds of the payoffs. We conclude in Section 5 with a discussion of some variants of the basic model.

2. THE MODEL

In this section, we describe the players, the learning environment, and the strategies. Then we compare the strategic pricing model briefly with the multi-armed bandit model. The comparison will prove useful for the welfare analysis of the equilibrium model in Section 3.

Price competition between two firms, indexed by $i = 1, 2$, takes place in discrete time with an infinite horizon, $t = 0, 1, 2, \ldots$. The firms announce in each period their prices, p_t^i, simultaneously. The goods produced by the two firms differ only with respect to their (expected) quality. Firms have the same unit costs normalized to zero. The buyer has unit demand in each period. At time t, the buyer's expected valuation of a purchase is a linear function of the expected quality and the price:

$$E_t X_t^i - p_t^i = x_t^i - p_t^i,$$

where the random realization of the quality of product i in period t is denoted by X_t^i.[4] Each X_t^i is a nonnegative real valued random variable with finite expectations on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$. The expected value of the

[3]Recently, Felli and Harris (1994) introduced a matching model in continuous time where wages are renegotiated at every instant of time. The equilibrium in their basic model is the continuous time equivalent of the cautious equilibrium in our model.

[4]Any quasilinear utility function $U(X_t^i) - p_t^i$ could be used alternately.

quality realization, X_t^i, conditional on the history until period t, is given by $x_t^i = E_t X_t^i$. All parties have common priors about the reward processes $X^i = \{X_t^i\}_{t=0}^{\infty}$ at the beginning of the game. Moreover, the sample realization X_t^i is publicly observable.

We concentrate our attention for simplicity on *sampling processes*. A sampling process is a sequence $X^i = \{X_t^i\}_{t=0}^{\infty}$ of independent, identically distributed random variables $X_0^i, X_1^i, \ldots$, drawn from a distribution with an unknown (vector-valued) parameter θ^i belonging to a family of distributions $\mathscr{D}$. The associated density functions are denoted by $f^i(\cdot|\theta^i)$. The *prior density* for the parameter $\theta^i \in \mathbb{R}^n$ is given by $\pi_0^i(\cdot)$. The posterior beliefs are represented by $\pi_t = (\pi_t^1, \pi_t^2)$. After observing the random variable X_t^i in period t, π_t^i is converted by Bayes rule into π_{t+1}^i:

$$\pi_{t+1}^i(\theta^i|X_t^i) = \frac{\pi_t^i(\theta^i)\cdot f^i(X_t^i|\theta^i)}{\int \pi_t^i(\phi)\cdot f^i(X_t^i|\phi)\,d\phi}. \tag{2.1}$$

Starting with prior beliefs and applying (2.1) recursively, we obtain a sequence of beliefs $\{\pi_t\}_{t=0}^{\infty}$.[5]

The consumer and the firms discount the future with the same discount factor β, with $0 \le \beta < 1$. Past quality realizations together with past prices and past consumer decisions constitute the history of the game. We denote with H_t the set of all possible histories up to, but not including period t. An element $h_t \in H_t$ includes all past prices, $p_s = (p_s^1, p_s^2)$, $0 \le s < t$, the consumers decision variable, $d_s = (d_s^1, d_s^2)$, where

$$d_s^i = \begin{cases} 1 & \text{if the consumer } \textit{accepts} \text{ the offer of firm } i \text{ in period } s, \\ 0 & \text{if the consumer } \textit{rejects} \text{ the offer of firm } i \text{ in period } s, \end{cases}$$

and the random realizations X_s^i of the purchased product i, $0 \le s < t$. Hence h_t is

$$h_t = (p_0, d_0, X_0^s, \ldots, p_{t-1}, d_{t-1}, X_{t-1}^s),$$

where the upper index $s = 1,2$ indicates the identity of the selling firm.

A *pricing strategy* of seller i at any time t is a function from the history into a distribution on the real numbers,

$$p_t^i\colon H_t \to \Delta(\mathbb{R}).$$

The buyer makes her purchase decision knowing the past play and the prices currently offered. Her *acceptance strategy* is a function from the history and the current prices into her decision space $\Delta(\{0,1\}\times\{0,1\}\setminus(1,1))$:

$$d_t\colon H_t \times \mathbb{R} \times \mathbb{R} \to \Delta(\{0,1\}\times\{0,1\}\setminus(1,1)).$$

[5]While our exposition will be restricted to sampling processes, all our results remain true for a general non-i.i.d. filtration set-up. Similarly all our results extend naturally from a duopoly to a general N-seller oligopoly model.

Notice that unit demand imposes the constraint $d_t^1 + d_t^2 \leq 1$, which allows the buyer not to purchase at all, if she prefers to do so. We denote by $\boldsymbol{d}_s = \{d_t\}_{t=s}^{\infty}$ the sequence of decision functions starting in period s. Similarly $\boldsymbol{p}_s^i = \{p_t^i\}_{t=s}^{\infty}$ is the sequence of future pricing strategies of firm i starting in period s.

The discounted expected profit for firm i under a given strategy triple $(\boldsymbol{d}_s, \boldsymbol{p}_s^1, \boldsymbol{p}_s^2)$ at time s is

$$E_s\left[\sum_{t=s}^{\infty} \beta^{t-s} d_t^i p_t^i\right], \tag{2.2}$$

and the expected present value for the consumer in period s is

$$E_s\left[\sum_{t=s}^{\infty} \beta^{t-s}\left[d_t^1(X_t^1 - p_t^1) + d_t^2(X_t^2 - p_t^2)\right]\right]. \tag{2.3}$$

Each player acts so as to maximize the expected discounted return given the beliefs over the return processes and the strategies of the other players. To facilitate the equilibrium analysis in the next section, we compare our model with the multi-armed bandit problem.

An n-armed bandit consists of n statistically independent alternatives (arms) which may be chosen in any order and one at a time. We concentrate our attention without loss of generality to $n = 2$. The maximization problem of the decision maker is to find an allocation strategy $\boldsymbol{d}^*$ which solves

$$\max_{d} E_t\left[\sum_{t=0}^{\infty} \beta^t d_t^1 X_t^1 + \sum_{t=0}^{\infty} \beta^t d_t^2 X_t^2\right], \quad \text{subject to} \tag{2.4}$$

$$d_t^1 + d_t^2 \leq 1.$$

The solution to (2.4) is the celebrated index policy. Gittins and Jones (1974) showed that it is possible to assign to each alternative i an *index function* $M^i(\pi_t^i)$ which depends only on the state π_t^i of project i. The optimal policy based on the index function is simple: Compute at any given time the indices of the different alternatives and select a project with maximal index. The index of project i is defined in terms of the following optimization problem involving only project i. Suppose the decision maker is facing in each period only the choice between continuing with the random sequence $\boldsymbol{X}^i$ or stopping the sequence to obtain a terminal reward z. The value $G^i(\pi_t^i, z)$ of this problem is defined by the dynamic programming equation

$$G^i(\pi_t^i, z) = \max\left\{z, E_{\pi^i}\left[X_t^i + \beta E G^i(\pi_{t+1}^i, z)\right]\right\}. \tag{2.5}$$

The *dynamic allocation* or *Gittins index* $M^i(\pi_t^i)$ is given through the equation (2.5).

DEFINITION 1: The *dynamic allocation index of alternative* i is defined as

$$M^i(\pi_t^i) = \sup\{z \in \mathbb{R} | G^i(\pi_t^i, z) > z\},$$
$$= \inf\{z \in \mathbb{R} | G^i(\pi_t^i, z) = z\}.$$

In words, the index $M^i(\pi_t^i)$ of alternative i in state π_t^i is the supremum over all terminal rewards, such that the decision-maker still prefers to continue with the random stream; or alternatively, it is the infimum over all terminal rewards such that the decision maker is indifferent between continuing with the random sequence and retiring with the stopping reward $M^i(\pi_t^i)$. The *index process* $M_t^i = \{M^i(\pi_t^i), t \in N\}$ reduces the n-dimensional problem to a comparison of n 1-dimensional problems.[6]

The buyer's decision problem in our model differs from the multi-armed bandit problem in two important aspects. First, in the strategic model the return stream of the buyer is affected by the pricing policies of the sellers, where each pricing policy is in turn the solution to the firm's profit maximization problem (2.2). The second difference is central to the intertemporal aspect of the game. In the multi-armed bandit problem, the value of the random sequence i depends only on the information acquired along the sequence i, but not on the history of the other random sequences. In the duopoly game, however, we naturally expect any pricing strategy of seller i to react not only to its own quality realizations, but also to those of the competing alternative. Hence, the *current* expected reward, $x_t^1 - p_t^1$ or $x_t^2 - p_t^2$, and the expectation over *future* rewards of the competing alternatives are naturally dependent.

3. EQUILIBRIUM PRICE COMPETITION AND EFFICIENCY

The conceptual distinction between our strategic model and the decision theoretic learning model is that in our model the alternatives are owned by separate economic agents. The pricing of alternative i is now a strategic decision made by seller i in each period. By introducing the separation in ownership we can examine how the costs and benefits of learning are allocated intertemporally between the buyer and sellers in equilibrium. First, we investigate the efficiency of the learning process in the presence of strategic interaction. In the next section, we analyze the dynamic allocation of the payoffs needed to sustain the equilibrium learning process.

We are interested in Markov perfect equilibria of the game for which π_t is the state variable.[7] By requiring that players base their decisions in equilibrium only on payoff relevant variables, current prices, and current information (summarized in the densities $\pi_t = (\pi_t^1, \pi_t^2)$), we focus the equilibrium analysis

[6] Whittle (1982) and Gittins (1989) are excellent references for more details on the multi-armed bandit theory.

[7] See Maskin and Tirole (1994) for a detailed account of the Markov perfect equilibrium concept.

on the strategic effects of the learning process. The Markov property will also limit the set of equilibria significantly and we will discuss the main differences between Markovian and non-Markovian equilibria briefly in the next section.

An equilibrium in the game is defined as follows.

DEFINITION 2: A *subgame perfect equilibrium* (*SPE*) is a triple of decision rules $\{\boldsymbol{d}, \boldsymbol{p}^1, \boldsymbol{p}^2\}$ which form a Nash equilibrium in every subgame.[8]

Player i's strategy is Markov if it depends only on the payoff relevant history.

DEFINITION 3: A *Markov perfect equilibrium* (*MPE*) is an SPE, where

$$p_t^i(h_t) = p^i(\pi), \quad \text{and}$$

$$d_t^i(\pi_t, p_t^1, p_t^2) = d^i(\pi, p^1, p^2) \quad \text{for } i = 1, 2.$$

The definition explicitly states that the Markov strategies should be time-invariant or stationary in the sense that whenever $\pi_t = \pi_{t+s}$, then $p^i(\pi_t) = p^i(\pi_{t+s})$ and also $d_t^i(\pi_t, p_t^1, p_t^2) = d_{t+s}^i(\pi_{t+s}, p_{t+s}^1, p_{t+s}^2)$ have to hold.[9]

For the characterization of the Markov equilibria we cast the players' decision problems in a stochastic dynamic programming framework. We define the value function of each player i as

$$V^i(\pi_t) \equiv V^i(\pi_t | \cdot, \cdot),$$

where we take the strategies of the other players as given and π_t as the state variable of the system. The buyer has to choose simultaneously between the current returns, $X_t^i - p^i(\pi_t)$ or $X_t^j - p^j(\pi_t)$ and their associated learning opportunities as indicated by (π_t, X_t^i) or (π_t, X_t^j). We shall write $V^B(\pi_t, X_t^i)$ rather than $V^B(\pi_{t+1})$ to indicate which sample, X_t^i or X_t^j, conditions the transition from π_t to π_{t+1}. The Bellman equation for the buyer is given by

$$V^B(\pi_t) = \max E_t\{X_t^1 - p_t^1 + \beta V^B(\pi_t, X_t^1), \tag{3.1}$$
$$X_t^2 - p_t^2 + \beta V^B(\pi_t, X_t^2), 0 + \beta V^B(\pi_t)\}.$$

The consumer can always refuse to make a purchase and receive a reservation value of zero. If the best decision for the buyer should be to accept neither offer, then under the Markov assumption, this will remain her best possible

[8]There are two alternative ways to look at this game. The first is an incomplete information game where nature moves at the first node to select the types of sellers. Information is partially revealed at subsequent nodes. In this representation the game has no proper subgames. An alternative representation, adopted for this paper, is a complete information game with a unique starting node given by the priors and a perfectly observed move by nature in each period determining the transition on the state variables of all players. In this representation, each choice by the buyer starts a new subgame.

[9]We index π with t henceforth only as a matter of recording time.

decision forever. Consider next the dynamic programming problem for the firms. Each seller, when choosing his price, has to consider the benefits of realizing a sale today, or foregoing that possibility today and instead betting on future sales in a possibly changed environment. The value function for the first seller is

$$V^1(\pi_t) = \max_{p^1} E_t\{d_t^1[p^1(\pi_t) + \beta V^1(\pi_t, X_t^1)] + d_t^2 \beta V^1(\pi_t, X_t^2) + (1-d_t^1)(1-d_t^2)\beta V^1(\pi_t)\}, \tag{3.2}$$

and for the second seller it is symmetrically

$$V^2(\pi_t) = \max_{p^2} E_t\{d_t^2[p^2(\pi_t) + \beta V^2(\pi_t, X_t^2)] + d_t^1 \beta V^2(\pi_t, X_t^1) + (1-d_t^1)(1-d_t^2)\beta V^2(\pi_t)\}. \tag{3.3}$$

If the buyer decides not to accept any offer in π_t, then we have of course $\pi_{t+1} = \pi_t$.

We concentrate our attention for the moment on MPE in pure strategies. It will be shown in Proposition 1, that this involves no loss of generality.

LEMMA 1: *In any pure MPE the buyer makes a purchase in every period and*

$$V^B(\pi_t) = E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] \geq 0,$$

where $s = 1, 2$ is the accepted seller.

PROOF: Notice first that $V^B(\pi_t) \geq 0$, $\forall \pi_t$, by the no purchase option and $V^i(\pi_t) \geq 0$, $\forall \pi_t$, since the seller can always ask for positive prices, which can at most be refused. Suppose that no purchase is made in period t. By Markov assumption no sales are made in future periods either and hence all agents' value is zero. Since $E_t X_t^i > 0$ by nonnegativity of X^i, one of the firms can offer a strictly positive price $0 < p_t^i < E_t X_t^i$ that yields a strictly positive value to the firm as well as the buyer. Hence a sale has to be made in all periods in equilibrium. *Q.E.D.*

For pure strategy equilibria, price competition implies that the consumer is indifferent between the choices offered by the two firms (or between one firm and the no-purchase option) at all points in time. At equilibrium prices, the selling firm, s, must (weakly) prefer to make the sale, whereas the nonselling firm, n, must (weakly) prefer to concede in the current period. The following (in-) equalities for the buyer (B), the selling firm (S^s), and the nonselling firm (S^n) are equilibrium conditions and central to the following analysis. We state them independently in the following lemma.

LEMMA 2: *Assume $E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] > 0$. The strategy triple $\{\boldsymbol{d}, \boldsymbol{p}^1, \boldsymbol{p}^2\}$ is a pure MPE if and only if the conditions (B), (S^s), and (S^n) are met*

for all t and π_t:

$$(B) \qquad E_t\left[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)\right] = E_t\left[X_t^n - p^n(\pi_t) + \beta V^B(\pi_t, X_t^n)\right],$$

$$(S^s) \qquad p^s(\pi_t) + \beta E_t V^s(\pi_t, X_t^s) \geq \beta E_t V^s(\pi_t, X_t^n),$$

$$(S^n) \qquad \beta E_t V^n(\pi_t, X_t^s) \geq p^n(\pi_t) + \beta E_t V^n(\pi_t, X_t^n),$$

where $n \neq s$.

PROOF: ($\Rightarrow$) Consider (B) first. Suppose not. Then the two terms in the Bellman equation (3.1) of the consumer differ by a strictly positive number, which implies that the firm offering the higher value to the consumer could raise her price by $\epsilon > 0$ and, by $E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] > 0$, this would not affect the consumer's decision, which contradicts the equilibrium assumption. Consider now (S^s) and (S^n). If (S^s) does not hold, then by (B) we know that a deviation to a higher price induces the buyer to switch sellers and consequently is a profitable deviation. Similarly, if (S^n) does not hold, a downward deviation by ε is profitable for the nonselling firm by (B).

($\Leftarrow$) Recall that the value function of each player is defined given the opponents' strategies. Conditions (B), (S^s), and (S^n) are then the appropriate conditions to make sure that no one-shot deviations are profitable for any of the players. Since the payoffs are bounded from below, we refer to the principle of optimality to conclude that strategies satisfying conditions (B), (S^s), and (S^n) are optimal given the other players' strategies and hence form an equilibrium. *Q.E.D.*

REMARK: Lemma 2 is true for all pure MPE for which $E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] > 0$ holds along the equilibrium path for all π_t. If the assumption doesn't hold, an equilibrium, and some π_t, could conceivably exist such that $E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] = 0$. The equilibrium price $p^n(\pi_t)$ would then not necessarily satisfy (S^n) and this could possibly break the equality (B). It is easy to verify that whenever such an equilibrium exists, an outcome equivalent MPE with price $\bar{p}^n(\pi_t) < p^n(\pi_t)$ also exists, which contrary to the latter satisfies both conditions (B) and (S^n). We will see in Proposition 3 that all pure MPE for which (B), (S^s), and (S^n) are satisfied have $E_t[X_t^s - p^s(\pi_t) + \beta V^B(\pi_t, X_t^s)] > 0$, which allows the conclusion that all pure MPE are characterized by (B), (S^s), and (S^n), and that the qualification made for the moment is only a temporary one.

Price competition between the sellers in each period makes the stage game similar to a static Bertrand pricing game in which firms have different costs. To illustrate this point, we take the continuation values of each subgame for the moment as given. Upon entering the competition each firm has to consider the benefits as well as the costs of making a sale today. The benefits for firm i of selling today come from the realized current price and the future sales following

(π_t, X_t^i), but by doing so firm i foregoes all possible sales along the continuation game of (π_t, X_t^j):

$$\overbrace{p^i(\pi_t) + \beta E_t V^i(\pi_t, X_t^i)}^{\text{benefits of making a sale}} - \overbrace{\beta E_t V^i(\pi_t, X_t^j)}^{\text{costs of making a sale}}.$$

If we take the difference in the future payoffs of the two paths, $\beta E_t V^i(\pi_t, X_t^j)$ and $\beta E_t V^i(\pi_t, X_t^i)$, to be the net costs $c^i(\pi_t)$ of making a sale today, which are of course endogenous in equilibrium:

$$(3.4) \qquad c^i(\pi_t) \equiv \beta E_t V^i(\pi_t, X_t^j) - \beta E_t V^i(\pi_t, X_t^i),$$

then we can read conditions (S^s) and (S^n) simply as

$$(S^{s\prime}) \qquad p^s(\pi_t) \geq c^s(\pi_t),$$

$$(S^{n\prime}) \qquad p^n(\pi_t) \leq c^n(\pi_t).$$

The equilibrium price for the selling firm, s, must exceed the costs of making a sale, whereas the price of the conceding seller, n, must not exceed the costs of making a sale, for otherwise he could lower his price slightly and attract the consumer. We may note at this point that the dynamic duopoly model inherits the multiplicity of equilibria present in the static Bertrand game with different costs.[10] Since the buyer is indifferent between the sellers in equilibrium, choosing n instead of s is without costs for her. By quoting a high enough price, seller n makes sure that a deviation by the buyer is not to his disadvantage. In other words, he is *cautious* enough to ask for a price which he does not regret should he be chosen against all expectations. For future reference we shall call the equilibrium in which the conceding seller bids exactly his intertemporal costs of competition, *cautious equilibrium.*

DEFINITION 4: An *MPE is cautious* if seller n is indifferent between selling and conceding to the competitor:

$$(3.5) \qquad p^n(\pi_t) + \beta E_t V^n(\pi_t, X_t^n) = \beta E_t V^n(\pi_t, X_t^s).$$

Before we attack the question of how efficient the learning process is under competition, we clarify a more technical issue concerning the *similarity* of the mixed strategy equilibria with the pure strategy equilibria.

PROPOSITION 1: *Every Markov perfect equilibrium is outcome equivalent to some Markov perfect equilibrium in pure strategies.*

PROOF: We notice first that there is no MPE in which all sellers use mixed strategies in any one period simultaneously. For given continuation payoffs the

[10] The reader may recall that in a static Bertrand game where firms have different costs $c_1 < c_2$, the "standard" equilibrium is $p_1 = p_2 = c_2$ and the low cost firm is making the sale. However any price combination $p_1 = p_2 \in [c_1, c_2)$ can also be sustained as an equilibrium if the consumer chooses the low cost producers with probability one.

price setting game in any period is just like a static Bertrand duopoly game with different costs for each seller and unit demand. By a standard, but tedious, argument which we omit here, one can show that both sellers do not simultaneously engage in mixed strategies. We now show that only the seller who is not chosen by the consumer in equilibrium can ever use mixed strategies. Consider a pure strategy $p^i(\pi_t)$ by seller i when seller j is employing a mixed strategy. Take the price $\hat{p}^j(\pi_t)$ at which the buyer is just willing to buy from j:

$$E_t\left[X_t^i - p^i(\pi_t) + \beta V^B(\pi_t, X_t^i)\right] = E_t\left[X_t^j - \hat{p}^j(\pi_t) + \beta V^B(\pi_t, X_t^j)\right].$$

Should the inequality

$$\hat{p}^j(\pi_t) + \beta E_t V^j(\pi, X_t^j) > \beta E_t V^j(\pi_t, X_t^i)$$

hold, then seller j would never use a price lower than $\hat{p}^j(\pi_t)$ in his mixed price strategy. But at any price higher than $\hat{p}^j(\pi_t)$, he would be rejected by the buyer; thus seller j offers a unique price $\hat{p}^j(\pi_t)$. If seller j is then using a mixed strategy with $\hat{p}^j(\pi_t)$ in its support, it must satisfy

$$\hat{p}^j(\pi_t) + \beta E_t V^j(\pi, X_t^j) \leq \beta E_t V^j(\pi_t, X_t^i).$$

The prices $p^j(\pi_t)$ which are in the support of the mixed strategy, cannot be lower than $\hat{p}^j(\pi_t)$, because otherwise j would be chosen by the buyer, although he prefers not to. Thus the support can only contain $\hat{p}^j(\pi_t)$ and higher prices. But at higher prices than $\hat{p}^j(\pi_t)$, j will never be chosen by the buyer. Thus this equilibrium is outcome equivalent to the one in which the seller, who plays the mixed strategy, simply charges the lower bound in the support of his mixed strategies, namely $\hat{p}^j(\pi_t)$. Finally, if the buyer is randomizing, then both sellers have to be indifferent between selling and nonselling; otherwise one of them would deviate. And if both sellers are indifferent, then choosing one of them with probability one is again a pure strategy equilibrium. *Q.E.D.*

It will prove instructive to express the value function of each player for given (equilibrium) policies explicitly by the entire sequence of payoffs. The buyer's alternating between the sellers as she acquires experience can be represented by two sequences of switching times, which are in fact stopping times: $\{\sigma_n\}_{n=1}^{\infty}$ and $\{\tau_n\}_{n=1}^{\infty}$. The switching times are random times which depend on the sample path and the prices offered along the sample path. We define τ_n as a (stochastic) time at which the consumer stops buying from the first seller and switches to the second, and σ_n as a (stochastic) time at which the reverse switching behavior occurs. In other words, σ_n is a time period in which the buyer begins her nth round of purchases from the first seller and τ_n is the period in which the buyer begins her nth round of purchases from the second seller. The value function of

the buyer then admits the following representation

$$V^B(\pi_t) = E_t\left\{\sum_{n=1}^{\infty}\sum_{s=\sigma_n}^{\tau_n-1}\beta^{s-t}\left[X_s^1 - p^1(\pi_s)\right] + \sum_{n=1}^{\infty}\sum_{s=\tau_n}^{\sigma_{n+1}-1}\beta^{s-t}\left[X_s^2 - p^2(\pi_s)\right]\right\}. \tag{3.6}$$

The value of the buyer is the discounted expected sum of the per period net gains, $X_s^1 - p^1(\pi_s)$ or $X_s^2 - p^2(\pi_s)$, along all sample paths.[11] For seller i it is the expected discounted sum of all realized sales. The value of the game for the first seller is then

$$V^1(\pi_t) = E_t\left\{\sum_{n=1}^{\infty}\sum_{s=\sigma_n}^{\tau_n-1}\beta^{s-t}p^1(\pi_s)\right\},$$

and for the second seller

$$V^2(\pi_t) = E_t\left\{\sum_{n=1}^{\infty}\sum_{s=\tau_n}^{\sigma_{n+1}-1}\beta^{s-t}p^2(\pi_s)\right\}.$$

The social value $W(\pi_t)$ of the game in any state π_t is simply $W(\pi_t) \equiv V^B(\pi_t) + V^1(\pi_t) + V^2(\pi_t)$ and can be explicitly expressed through

$$W(\pi_t) = E_t\left\{\sum_{n=1}^{\infty}\sum_{s=\sigma_n}^{\tau_n-1}\beta^{s-t}X_s^1 + \sum_{n=1}^{\infty}\sum_{s=\tau_n}^{\sigma_{n+1}-1}\beta^{s-t}X_s^2\right\}. \tag{3.7}$$

We define an efficient equilibrium.

DEFINITION 5: An equilibrium is *efficient* if it maximizes the social value $W(\pi_t)$ for all π_t.

The notion of efficiency should, of course, be understood as a notion of (informationally) constrained or ex ante Pareto efficiency.

The problem of maximizing the social value of the game as depicted in (3.7) is in fact identical to the multi-armed bandit problem given in (2.4). An equilibrium is then efficient if and only if the stopping times $\{\sigma_n\}_{n=1}^{\infty}$ and $\{\tau_n\}_{n=1}^{\infty}$ coincide with the stopping times prescribed by the dynamic allocation index policy. With this identification in place no ambiguity should arise when we refer to the *efficient path* or the *efficient* (*inefficient*) or *superior* (*inferior*) *alternative* $i(j)$, by which we simply mean that $M^i(\pi_t) > M^j(\pi_t)$.[12]

[11] By Lemma 1, there is no time period where she does not buy at all, so that the purchasing behavior is completely described by $\{\sigma_n\}_{n=1}^{\infty}$ and $\{\tau_n\}_{n=1}^{\infty}$.

[12] By the index theorem efficiency follows already by $M^i(\pi_t^i) > M^j(\pi_t^j)$, i.e. the index $M^i(\pi_t^i)$ is independent of π_t^j. To save on notation we shall neglect this distinction.

The explicit representation of the payoff sequence of the buyer in (3.6) suggests also that the buyer should be willing to support the efficient allocation path through her choices if she is guaranteed to share sufficiently, *today* and in the *future*, in the gains from learning, which are limited only by the prices she has to pay. Ultimately, the question of efficient experimentation then hinges on the price path induced through the duopoly game.

We may distinguish two different situations. When long-run efficiency as indicated by the allocation index and high current return coincide in alternative i, i.e. $M^i(\pi_t) \geq M^j(\pi_t)$ and $E_t X_t^i \geq E_t X_t^j$, then firm i should be able to attract the buyer and yet obtain a relatively high price for its product since the competing product is inferior on both accounts. The inferior firm's best strategy is to wait, since a price low enough to attract the consumer today would not be justified on the current expectation for future profits.

The intertemporal incentives are more complicated when long-run efficiency and current high returns do not coincide, i.e. $M^i(\pi_t) \geq M^j(\pi_t)$ but $E_t X_t^i < E_t X_t^j$. Suppose for simplicity of the argument that the quality of firm j's product is known with certainty. We may ask how long firm i is willing to make sales. In this simplified case, we know by the Markovian assumption that once the buyer selects firm j, she will buy from firm j forever and consequently firm i's profit is zero from then on. As long as the total surplus along paths beginning with a sale by firm i exceed the total value of paths switching immediately, firm i can offer low enough prices today to attract the consumer while making a positive expected profit in future periods. But this is exactly the condition stated earlier in the form of the dynamic allocation indices: $M^i(\pi_t) \geq M^j(\pi_t)$. The following theorem shows that this intuition extends to the case of two uncertain products. In the following section, we determine equilibrium prices needed to support the appropriate intertemporal division of gains from trade.

THEOREM 1 (Efficiency): *All Markov perfect equilibria are efficient: If seller i is chosen in period t, then*

$$M^i(\pi_t) \geq M^j(\pi_t).$$

PROOF: [13] Suppose that firm i is chosen in period t. By Lemma 2 the following (in-)equalities have to hold:

$$(B) \qquad E_t\left[X_t^i - p^i(\pi_t) + \beta V^B(\pi_t, X_t^i)\right] = E_t\left[X_t^j - p^j(\pi_t) + \beta V^B(\pi_t, X_t^j)\right],$$

$$(S^1) \qquad p^i(\pi_t) + \beta E_t V^i(\pi_t, X_t^i) \geq \beta E_t V^i(\pi_t, X_t^j), \quad \text{and}$$

$$(S^2) \qquad \beta E_t V^j(\pi_t, X_t^i) \geq p^j(\pi_t) + \beta E_t V^j(\pi_t, X_t^j).$$

[13] We thank the editor for suggesting a different proof strategy, which led to a shorter and more transparent argument.

By summing both sides of the three (in-)equalities and recalling the definition of $W(\pi)$, we get

$$E_t X_t^i + \beta E_t W(\pi_t, X_t^i) \geq E_t X_t^j + \beta E_t W(\pi_t, X_t^j).$$

Hence the social value of the game, $W(\pi_t)$, calculated along the equilibrium path, satisfies the following functional equation:

$$W(\pi_t) = \max\{E_t X_t^i + \beta E_t W(\pi_t, X_t^i), E_t X_t^j + \beta W(\pi_t, X_t^j)\}. \tag{3.8}$$

But (3.8) characterizes the value function of the planner's problem as well. An easy application of the contraction mapping theorem establishes the uniqueness of solutions to (3.8). Consequently, firm i is selected in equilibrium only if firm i is selected along some optimal path in the planner's problem. By the Gittins index theorem, this is equivalent to $M^i(\pi_t) \geq M^j(\pi_t)$. *Q.E.D.*

The message of Theorem 1 is unambiguous. The fact that all MPE are efficient demonstrates that no firm has an interest in stopping the efficient learning process, since the costs involved in doing so are too high at each stage. In particular, the conceding seller, rather than forcing a sale through a very low price, prefers to postpone any sales in the expectation of a more favorable competitive context in the future.

The efficiency result of Theorem 1 continues to be valid in more general settings. If the buyer is not restricted to a single experiment, but can allocate up to N experiments among the sellers in each period, the resulting equilibria are still efficient. The necessary modification to establish efficiency in this framework is to allow firms the use of nonlinear pricing schemes.

The sellers are offering nonlinear pricing schedules to the buyer:

$$\boldsymbol{p}_t^j = \{p_t^{j1}, p_{2t}^j, \ldots, p_t^{jN}\},$$

where p_t^{jk} denotes the unit price of firm j's product if the buyer purchases k units. The purchasing decision of the buyer is then represented by two numbers:

$$\boldsymbol{d}_t = \{n_t^1, n_t^2\},$$

where n_t^j denotes the number of units demanded from firm j.

The optimization problem of the consumer in the value function form is then given by

$$V(\pi_t) = \max_{0 \leq n_t^1 + n_t^2 \leq N} E_t\Big\{n_t^1\big(X_t^1 - p_t^{1n_t^1}\big) + n_t^2\big(X_t^2 - p_t^{2n_t^2}\big) + \beta V(\pi_t, n_t^1, n_t^2)\Big\}.$$

The sellers' problems are described similarly. As in the case of a single unit, the consumer will always buy N units and will never use her no-purchase option. Analogues to Lemmas 1 and 2 continue to hold in this setting and using the

same equilibrium concepts and welfare criteria as above we are able to prove the following theorem.

THEOREM 2: *All MPE in the multi-unit game are efficient.*

Furthermore, all the results pertaining to the price path of the cautious equilibrium discussed in Section 4 continue to hold in the multi-unit case. We also point out that the efficiency result would continue to hold if the quality realizations of the products were mutually dependent.[14]

Extending these results to a game with multiple buyers proves to be substantially more difficult. Since the experiments are publicly observed, each purchase creates an informational externality on the other buyers. In the case of fixed prices this leads to well-known free-rider problems as in Bolton and Harris (1993). In our model, the firms are able to internalize some of these effects since a successful experiment by one consumer leads to an improved competitive position with respect to all consumers in the next period. It turns out, however, that we cannot expect efficient experimentation in general, even with nonlinear prices. A consumer has to be compensated for her experimentation costs only, while the gains of experimentation are collected from all consumers through higher prices. As a consequence, experimentation tends to be *too* cheap from the firm's point of view in the multiple buyer case and ex ante optimal experimentation is not achieved.[15] This has to be contrasted to the single buyer *with* multiple unit demand, who perfectly internalizes the price increase on *all* units in future periods.

4. CHARACTERIZATION OF THE MARKOV PERFECT EQUILIBRIA

The equilibrium choice path of the buyer has been established by the efficiency property of the MPE and we focus now on the equilibrium price path. Prices determine the intertemporal allocation of gains from experimentation between buyer and sellers. We focus on the *cautious equilibrium* where the pricing path provides the intertemporal incentives to experiment in a surprisingly simple and intuitive way. Finally, we characterize the *entire set* of MPE by giving upper and lower bounds on the payoffs for the players in Proposition 3.

4.1. *The Cautious Equilibrium*

We recall that the equilibrium condition as given in Definition 4 made the conceding seller (S^n) indifferent between realizing a sale or foregoing the sale in

[14]All model extensions as mentioned above have, however, the drawback that efficient policies cannot be characterized by index policies anymore, since they are either not known or simply don't exist as in the case of mutually dependent alternatives.

[15]A well known fact on ex ante efficient experimentation in multi-armed bandit models states that ex post efficiency fails with positive probability. Since experimentation is relatively cheap in the multiple buyer case, an interesting conjecture to be checked in future research is that equilibrium in the pricing game is closer to the ex post efficient path than the ex ante efficient path.

the current period,

$$p^n(\pi_t) + \beta E_t V^n(\pi_t, X_t^n) = \beta E_t V^n(\pi_t, X_t^s). \tag{4.1}$$

In other words, in the cautious equilibrium the conceding seller always sets his price equal to his net costs of competing $c^n(\pi_t) = \beta E_t V^n(\pi_t, X_t^s) - \beta E_t V^n(\pi_t, X_t^n)$:

$$p^n(\pi_t) = c^n(\pi_t).$$

The intertemporal allocation of the costs and benefits of the learning process in the *cautious* MPE are described completely in the following theorem.

THEOREM 3: *The cautious equilibrium is unique, efficient, and*

$$p^s(\pi_t) = E_t X_t^s - E_t X_t^n = x_t^s - x_t^n, \tag{4.2}$$

$$p^n(\pi_t) = \beta E_t V^n(\pi_t, X_t^s) - \beta E_t V^n(\pi_t, X_t^n). \tag{4.3}$$

The pricing rule of the conceding seller is a submartingale:

$$p^n(\pi_t) \leq \beta E_t\{p^n(\pi_t, X_t^s)\}. \tag{4.4}$$

PROOF: The nonselling firm in period t, say j, is indifferent between selling and not selling by the definition of the cautious equilibrium:

$$p^j(\pi_t) + \beta E_t V^j(\pi_t, X_t^j) = \beta E_t V^j(\pi_t, X_t^i). \tag{4.5}$$

By the consumer's indifference,

$$E_t\left[X_t^i - p^i(\pi_t) + \beta V^B(\pi_t, X_t^i)\right] = E_t\left[X_t^j - p^j(\pi_t) + \beta V^B(\pi_t, X_t^j)\right], \tag{4.6}$$

we can express $V^B(\pi_t)$ for a given equilibrium either as

$$V^B(\pi_t) = E_t\left[X_t^i - p^i(\pi_t) + \beta V^B(\pi_t, X_t^i)\right], \tag{4.7}$$

or as

$$V^B(\pi_t) = E_t\left[X_t^j - p^j(\pi_t) + \beta V^B(\pi_t, X_t^j)\right]. \tag{4.8}$$

We can extend (4.7) and (4.8) in this way for any number of periods. Since the equality (4.6) has to hold in each period, we are free to choose which alternative, i or j, to use in any period in the particular extensions. For now we extend (4.7) and (4.8) by the continuation game in which j is accepted forever. Extending (4.7) we get

$$V^B(\pi_t) = E_t\left\{\sum_{s=t}^{\infty} \beta^{s-t}\left[X_s^j - p^j(\pi_s)\right]\right\}, \tag{4.9}$$

and extending (4.8) we have

$$V^B(\pi_t) = E_t\left\{X_t^i - p^i(\pi_t) + \sum_{s=t+1}^{\infty} \beta^{s-t}\left[X_s^j - p^j(\hat{\pi}_s)\right]\right\}. \tag{4.10}$$

We decorated the state variable $\hat{\pi}_s$ in (4.10) to distinguish it from the state variables π_s in (4.9), since their experimenting paths are different. We solve equation (4.6) for the price of the superior seller i in period t, when the inferior seller j adheres to his pricing policy (4.5). By extending (4.5) in the same manner as (4.9) or (4.10) we obtain

$$E_t\left\{\sum_{s=t}^{\infty}\beta^{s-t}p^j(\pi_s)\right\}=E_t\left\{\sum_{s=t+1}^{\infty}\beta^{s-t}p^j(\hat{\pi}_s)\right\}.$$

We can finally use equality (4.5) to simplify equality (4.6) and obtain

$$p^i(\pi_t)=E_tX_t^i-E_tX_t^j=x_t^i-x_t^j,$$

describing the price of the successful seller in period t. To obtain an expression of $p^j(\pi_t)$ for the nonselling firm, we start again with equation (4.5) and extend both sides by one period. Since the Gittins index of seller j might rise above the one of seller i in $t+1$ after an observation of X_t^j in t, equality (4.5) might turn into an inequality in $t+1$, conditional on X_t^j:

$$p^j(\pi_t,X_t^j)+\beta E_tV^j(\pi_t,X_t^j,X_{t+1}^j)\geq\beta E_tV^j(\pi_t,X_t^j,X_{t+1}^i). \tag{4.11}$$

As a consequence we obtain from (4.5):

$$p^j(\pi_t)+\beta^2E_tV^j(\pi_t,X_t^j,X_{t+1}^i)\leq\beta E_tp^j(\pi_t,X_t^i)+\beta^2E_tV^j(\pi_t,X_t^i,X_{t+1}^j),$$

resulting in

$$p^j(\pi_t)\leq\beta E_tp^j(\pi_t,X_t^i),$$

which concludes the proof. *Q.E.D.*

A few aspects of the pricing strategies in the cautious equilibrium deserve discussion. Experimentation is efficient and all sales are made at $p^s(\pi_t)=E_tX_t^s-E_tX_t^n$. Notice that efficiency, $M^s(\pi_t)\geq M^n(\pi_t)$, does not imply $E_tX_t^s\geq E_tX_t^n$. It may be that $M^s(\pi_t)\geq M^n(\pi_t)$, although $E_tX_t^s<E_tX_t^n$, in which case $p^s(\pi_t)=E_tX_t^s-E_tX_t^n<0$. Negative (or below cost) prices then appear in equilibrium as a natural instrument to support the learning process of the consumer. Negative prices are associated with states π_t where $M^s(\pi_t)\geq M^n(\pi_t)$ *but* $E_tX_t^s<E_tX_t^n$. In these states seller s is willing to offer negative prices because the superiority of his dynamic allocation index indicates that he will be able to recover the initial losses through higher future prices. On the other hand, the consumer expects negative prices as an advance payment because the gains from learning will eventually be diluted through higher prices.

The net value of the sale to the buyer is the current expected quality of the dynamically inefficient firm: $E_tX_t^s-p^s(\pi_t)=E_tX_t^n$. As long as the buyer does not switch from seller i to seller j, her net value of a purchase remains thus constant at $E_tX_t^j$ and the price of the successful seller i forms a martingale.

Marginal gains and costs of experimentation are reflected in the price set by seller i in response to new information. The buyer is thus insured during any single *round* of experimentation with a particular seller and realizes intertemporal costs or benefits only when she is switching from one seller to the other.

The conceding seller hence represents, in equilibrium, the insurance or outside option to the buyer. If the buyer continues to experiment with i although the estimate $E_t X_t^i$ decreases, she weakens her future insurance position vis-à-vis seller j. Or, put differently, the future switching costs from i to j are increasing when the expected quality of seller i decreases. Consequently, if seller i still intends to make the sale, i has to compensate the buyer for the induced future risk of higher switching costs. Conversely, if the buyer expects to switch from i to j because the value of learning is high, but not because the current return of j dominates i, then i provides the buyer with a comparatively high insurance level in the future. In consequence seller i can ask today for a price higher than that justified by current quality differences. The discrepancy between the intertemporal pricing rule and the static pricing rule can be systematically linked to this insurance effect.

In a myopic environment the optimal pricing rule $p_m^i(\pi_t)$ of the successful seller i is given by

$$p_m^i(\pi_t) - p^j(\pi_t) = x_t^i - x_t^j. \tag{4.12}$$

The price $p_m^i(\pi_t)$ at which i can make a sale is such that the price difference between i and j equates the estimated quality difference between the competing products. We contrast this with the incentives provided by the intertemporal pricing rules. Recall from the definition of the cautious equilibrium that

$$p^j(\pi_t) = \beta E_t\left[V^j(\pi_t, X_t^i) - V^j(\pi_t, X_t^j)\right], \tag{4.13}$$

and from Theorem 4.1 that

$$p^i(\pi_t) = x_t^i - x_t^j. \tag{4.14}$$

Comparing equations (4.12) and (4.14) we observe immediately that the price differential in the cautious equilibrium is smaller than in the myopic case if $p^j(\pi_t) > 0$. By (4.13) we can relate this condition to the future competitive position of the currently conceding seller j. When $p^j(\pi_t) > 0$, then the expected continuation payoff for firm j is higher if the buyer experiments today with firm i's product rather than with j's product:

$$p^j(\pi_t) > 0 \Leftrightarrow E_t V^j(\pi_t, X_t^i) > E_t V^j(\pi_t, X_t^j).$$

Experimentation with seller i must hence provide better prospects for an improvement in j's competitive position than a direct experiment with j himself. The improvement can only come through a decrease in the expected quality of firm i's product along the path of the play. But the decrease in $E_t X_t^i$ along the equilibrium path implies that the outside option seller i provides to the buyer, conditional on switching from seller i to seller j, is expected to decrease. Since

the buyer realizes the negative future impact of experimentation with seller i today, she is not willing to pay the myopic price $p_m^i(\pi_t)$ and seller i has to settle for less than the myopic price:

$$p^i(\pi_t) < p_m^i(\pi_t).$$

The potential of future rent-seeking by seller j, after current experimentation with i, is consequently expressed in the fact that only a strictly positive price determined by (4.13) makes seller j indifferent between a sale today and the improvement in the competitive position induced by experimenting with seller i. A similar argument can be given for $p^j(\pi_t) < 0$, in which case seller i can extract a higher than myopic price since he provides the buyer with a relatively stable insurance value and consequently:

$$p^i(\pi_t) > p_m^i(\pi_t).$$

With more accurate estimates of the product qualities, the changes in the competitive environment become smaller over time. In turn, we would expect the deviation from the myopic pricing policy to become less significant as more information accumulates. In the next subsection, we describe the asymptotic properties of the cautious equilibrium and show that in general, the long-run prices are different from the myopic Bertrand prices under perfect information.

4.2. *The Asymptotic Behavior of the Cautious Equilibrium*

As time goes by, the buyer will learn more about the true value of the purchased products. By Lemma 1 a purchase is made in every period and at least one seller will be chosen infinitely often. The value of learning, $(1-\beta)M^i(\pi_t) - x_t^i$, as distinct from the value of the current return, diminishes as the dynamic allocation index (and the posterior mean) converges to the true mean of the reward process. Consequently the value of sampling decreases and the ranking of the alternatives with respect to their current pay-off tends to coincide with the ranking of the indices. Below cost prices associated with states π_t, where the learning effects dominate the current return effect, should therefore gradually disappear. We may ask whether the pricing and acceptance policies will approach those of the static Bertrand competition in the limit.

PROPOSITION 2 (Asymptotic Behavior): *The asymptotic behavior of the cautious MPE is given by*

(i) $\lim_{t\to\infty} p^s(\pi_t) = x^s - \lim_{t\to\infty} E_t X_t^n \geq 0$,

(ii) $\lim_{t\to\infty} V^n(\pi_t) = 0$.

PROOF: (i) By Definition 1 and the Martingale convergence theorem,

$$\lim_{t\to\infty} M^\infty(\pi_t) = x^\infty/(1-\beta) \quad \text{a.s.}$$

where M^∞ and x^∞ are the dynamic allocation index and the true mean of any alternative which is chosen infinitely often. Since

$$M^s(\pi_t) \geq M^n(\pi_t) \geq E_t X_t^n/(1-\beta)$$

has to hold by the efficiency of the MPE, we have

$$\lim_{t\to\infty} p^s(\pi_t) = \lim_{t\to\infty} [E_t X_t^s - E_t X_t^n] = x^s - \lim_{t\to\infty} E_t X_t^n \geq 0.$$

Since the claim applies to the prices of all sellers who are chosen infinitely often, we don't exclude the case where switching between the sellers occurs infinitely often. $\mathrm{Lim}_{t\to\infty} E_t X_t^n$ may not converge to x^n, since if a seller j is abandoned after a finite time, convergence to x^n is by no means guaranteed.

(ii) Clearly $\liminf_{t\to\infty} V^{n_t}(\pi_t) \geq 0$ by the no sale option and we want to show that in fact $\lim_{t\to\infty} V^{n_t}(\pi_t) = 0$, where n_t is the conceding seller in period t, and s_u is the successful seller in period u. We proceed by contradiction. Assume that $\limsup_{t\to\infty} V^{n_t}(\pi_t) > 0$, which implies that there exists $\epsilon > 0$ such that

$$\lim_{t\to\infty} \Pr[s_u = n_t, u > t | \pi_t] \geq \epsilon,$$

where $\Pr[\cdot]$ is the probability assessment of the players along the path of the play. Standard convergence arguments imply that

$$\lim_{t\to\infty} \Pr\left[s_u = n_t, u > t | x^i \neq x^j\right] = 0,$$

and consequently if

$$\lim_{t\to\infty} \Pr[\pi_u, s_u = n_t, u > t | \pi_t] \geq \epsilon,$$

then

$$\lim_{t\to\infty} \Pr\left[x^i = x^j | \pi_t\right] = 1,$$

in which case it has to be that

$$\lim_{t\to\infty} p^{s_u}(\pi_u, s_u = n_t) = \lim_{t\to\infty} [E_t X_t^s - E_t X_t^n] = 0,$$

which implies that $\limsup_{t\to\infty} V^{n_t}(\pi_t) = 0$, concluding the proof. *Q.E.D.*

The equilibrium converges exactly to the myopic Bertrand equilibrium if both sellers are chosen infinitely often and all learning possibilities are exhausted, in which case even $\lim_{t\to\infty} E_t X_t^n = x^n$ holds. The asymptotic behavior is somewhat different when only one seller is chosen infinitely often in equilibrium. The price $p^n(\pi_t)$ is determined by Theorem 4.1 as

$$p^n(\pi_t) = \beta E_t V^n(\pi_t, X_t^s) - \beta E_t V^n(\pi_t, X_t^n).$$

By Proposition 2 we also have $\lim_{t\to\infty} V^n(\pi_t) = \lim_{t\to\infty} \beta E_t V^n(\pi_t, X_t^s) = 0$, so that

$$\lim_{t\to\infty} p^n(\pi_t) = -\beta E_t V^n(\pi_t, X_t^n).$$

Now, if a single experiment with n at π_t could change the ranking of the indices, then we have $\beta E_t V^n(\pi_t, X_t^n) > 0$ and therefore $p^n(\pi_t) < 0$. The fact that $V^n(\pi_t)$ converges nevertheless to zero is only confirming that in equilibrium it is optimal not to explore the remaining learning opportunities. This has to be contrasted to statistical decision problems which generally converge in the limit to the short-run, myopic decision problems as in Aghion et al. (1991, Proposition 2.2–2.4). Here it is the strategy of the conceding seller which indicates that if some uncertainty remains unresolved in the game then the convergence will not be complete.

4.3. *The Set of MPE*

We come to the characterization of the entire set of MPE. By Theorem 1, all MPE are efficient and hence have the same social value $W^*(\pi_t)$. The multiplicity of equilibria then pertains only to different allocations of the surplus, $W^*(\pi_t)$, among the players. The equilibrium which maximizes the buyer's payoff is therefore simultaneously minimizing the sellers' payoff. Conversely, the equilibrium which is maximizing the seller's payoff is simultaneously minimizing the buyer's payoff. The characterization of these two extremal equilibria is then sufficient to describe the lower and upper bounds on the payoffs of the players.

PROPOSITION 3 (Characterization of MPE): *The set of Markov perfect equilibria is characterized by the lower and upper bounds on the payoffs. The lower bounds are given by*:

$$(b) \qquad V^B(\pi_t) = E_t\left\{ \sum_{n=1}^{\infty} \sum_{s=\sigma_n}^{\tau_n - 1} \beta^{s-t} x^2_{\sigma_n} + \sum_{n=1}^{\infty} \sum_{s=\tau_n}^{\sigma_{n+1}-1} \beta^{s-t} x^1_{\tau_n} \right\},$$

$$(s^1) \qquad V^1(\pi_t) = 0,$$

$$(s^2) \qquad V^2(\pi_t) = 0.$$

The upper bounds are given by:

$$(B) \qquad V^B(\pi_t) = W^*(\pi_t),$$

$$(S^1) \qquad V^1(\pi_t) = E_t\left\{ \sum_{n=1}^{\infty} \sum_{s=\sigma_n}^{\tau_n - 1} \beta^{s-t} \left[x^1_s - x^2_{\sigma_n} \right] \right\},$$

$$(S^2) \qquad V^2(\pi_t) = E_t\left\{ \sum_{n=1}^{\infty} \sum_{s=\tau_n}^{\sigma_{n+1}-1} \beta^{s-t} \left[x^1_s - x^2_{\tau_n} \right] \right\}.$$

We omit the proof, which can be found in Bergemann and Välimäki (1995), since the construction of the extremal equilibria, while entirely straightforward, is long and tedious.

The equilibrium which generates payoffs $\{(B),(s^1),(s^2)\}$ reduces the payoffs of the sellers permanently to zero, which is their individual participation constraint. The buyer receives the entire value of the game. The second equilibrium, which generates $\{(b),(S^1),(S^2)\}$ is in fact the *cautious equilibrium* of Theorem 4.1. Let us here just recall the payoff structure of this equilibrium. The consumer is buying the product of the superior seller i at the price $p^i(\pi_t) = E_t X_t^i - E_t X_t^j = x_t^i - x_t^j$. Let $i=1$ and $j=2$. Starting at τ_n, when the buyer switches from j to i and until σ_n when she switches back to j, the estimate of $E_t X_t^j$ does not change since no new information on j's product becomes available. We can consequently write $E_t X_t^j = E_{\tau_n} X_{\tau_n}^j = x_{\tau_n}^j$ for t between $\tau_n \le t \le \sigma_n - 1$. During this time span the buyer is experimenting with i, and only the estimate of X_t^i, $E_t X_t^i$ is changing over time. For the same time interval, $\tau_n \le t \le \sigma_n - 1$, the buyer's periodic return is constant and given by $E_t X_t^i - p^i(\pi_t) = E_t X_t^i - (x_t^i - x_{\tau_n}^j) = x_{\tau_n}^j$, which is represented in (b).

The symmetry in the extremal equilibria is apparent. The equilibrium which maximizes the buyer's payoff makes the successful seller s always indifferent between selling and not selling: the equilibrium condition (S^s) holds as an equality. In the equilibrium which minimizes the buyer's payoff it is, on the contrary, always the conceding seller n who is indifferent between selling and not selling and in turn (S^n) holds as an equality. Since the "successful" prices, $p^s(\pi_t)$, and the "threat" prices, $p^n(\pi_t)$, that each seller is employing are strategically almost independent devices, it is not difficult to show that the entire convex hull spanned by the payoffs of the extremal equilibria constitute the set of MPE payoffs.

Proposition 3 tells us that the main difference between the set of Markovian and the set of non-Markovian equilibria lies in the possibility of collusion among the sellers. This may imply efficiency losses as the following example demonstrates. Consider the following collusive equilibrium, in which the sellers alternate in selling at prices $p^s = E_t X_t^s$ and $p^n \le E_t X_t^n$, and use the trigger strategy to convert to the equilibrium $\{(B),(s^1),(s^2)\}$ should a price deviation by one of the sellers occur. The buyer's payoff is now reduced forever to zero and the allocation path is clearly inefficient, since the alternating is independent of the actual learning experience.

5. CONCLUSION

We presented a simple dynamic equilibrium pricing model under uncertainty where the players take into account the costs and benefits of learning. All MPE are efficient. The cautious MPE, which was the focus of our analysis implements the efficient learning solution by a simple and intuitive equilibrium pricing policy of the firms. The restriction to Markovian equilibria allowed us to focus on the interaction of pricing and learning policies.

Since there was only one large consumer and the qualities of the firms were statistically independent, experimentation did not give rise to any externalities. The extension to statistically dependent alternatives is straightforward and would yield exactly the same conclusions in terms of efficiency and equilibrium

prices as the statistically independent case. The reader may verify that the independence assumption was only used for the characterization of the efficient policy in terms of the dynamic allocation index. This result underlines the basic mechanism at work in the dynamic pricing model. Strategic competition can sustain efficient learning outcomes if the exchange of the costs and benefits is frictionless both intertemporally *and* interpersonally. The necessity of intertemporal exchange was a major theme throughout the paper. The extension of our model to a multiple buyer market illustrates the necessity of frictionless interpersonal exchange to sustain efficient experimentation. While our results for multiple buyers are only preliminary, they suggest some interesting possibilities. In the simplest case of one known and one unknown product with many buyers, market experimentation will continue beyond the social optimum.

As in the case with a single buyer, the seller of the unknown product has to offer negative prices to compensate for the current quality differential. The essential difference arises as the seller with the unknown product needs to compensate *his* buyers only for their own future expected losses. Since experimentation is public, he will eventually appropriate the benefits of a positive sample path from all consumers. Conversely, the firm with the known product would need to offer to each and every consumer a price low enough in order to completely prevent experimentation with the unknown product. This policy is very costly and, rather than trying to exclude his competitor entirely, he prefers to let experimentation continue beyond the social optimum to further weaken his opponent's position. In Bergemann and Välimäki (1996), we show that the underinvestment in learning result as described in Bolton and Harris (1993) can be reversed and the equilibrium may involve socially excessive experimentation.

Dept. of Economics, Yale University, P.O. Box 208268, New Haven, CT 06520-8268, U.S.A., and Institut d'Analisi Economica, CSIC, Barcelona, Spain

and

Dept. of Economics, Northwestern University, 2003 Sheridan Rd., Evanston, IL 60208, U.S.A.

Manuscript received March, 1994; final revision received November, 1995.

REFERENCES

AGHION, P., P. BOLTON, C. HARRIS, AND B. JULLIEN (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.

BERGEMANN, D., AND J. VÄLIMÄKI (1995): "Learning and Strategic Pricing: Further Results," Mimeo, Northwestern University and Yale University.

——— (1996): "Market Experimentation and Pricing," Mimeo, Northwestern University and Yale University.

BOLTON, P., AND C. HARRIS (1993): "Strategic Experimentation," Discussion Paper TE/93/261, LSE, London.

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

FELLI, L., AND C. HARRIS (1994): "Job Matching, Learning and the Distribution of Surplus," Mimeo, LSE.

GITTINS, J. C. (1989): *Multi-Armed Bandit Allocation Indices*. Chichester: 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.

KIHLSTROM, R., L. MIRMAN, AND A. POSTLEWAITE (1984): "Experimental Consumption and the 'Rothschild Effect,'" in *Bayesian Models of Economic Theory*, ed. by M. Boyer and R. Kihlstrom. Amsterdam: Elsevier.

MASKIN, E., AND J. TIROLE (1988): "A Theory of Dynamic Oligolpoly, II," *Econometrica*, 56, 571–599.

——— (1994): "Markov Perfect Equilibrium," mimeo.

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

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

ROB, R. (1991): "Learning and Capacity Expansion under Demand Uncertainty," *Review of Economic Studies*, 58, 655–675.

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

SMITH, L. (1992): "Error Persistence, and Experimental versus Observational Learning," Mimeo, MIT.

WHITTLE, P. (1982): *Optimization over Time*, Vols. 1&2. Chicester: Wiley.