Dynamic Learning in Strategic Pricing Games

Matthew Stern*and John R. Birge†

University of Chicago

Booth School of Business

Chicago, IL 60637 USA

October 25, 2018

Abstract

In monopoly pricing situations, firms should optimally vary prices to learn demand. The variation must be sufficiently high to ensure complete learning. In competitive situations, however, varying prices provides information to competitors and may reduce the value of learning. Such situations may arise in the pricing of new products such as pharmaceuticals and digital goods. This paper shows that firms in competition can learn efficiently in certain equilibrium actions which involve adding noise to myopic estimation and best-response strategies. The paper then discusses how this may not be the case when actions reveal information quickly to competitors. The paper provides a setting where this effect can be strong enough to stop learning so that firms optimally

*stern@chicagobooth.edu

†john.birge@chicagoobooth.edu. This work was supported by the University of Chicago Booth School of Business. The authors thank seminar participants at City University of Hong Kong, Columbia University, Georgetown University, Johns Hopkins University, Massachusetts Institute of Technology, McGill University, Singapore Management University, Stanford University, Syracuse University, Tsinghua University, University of British Columbia, University of Houston, University of Michigan, University of Minnesota, ECSO 2017 (Rome, Italy), University of Wisconsin-Madison, and WHU Otto Beisheim School of Management for valuable comments and suggestions.

reduce any variation in prices and choose not to learn demand. The result can be that the selling firms achieve a collaborative outcome instead of a competitive equilibrium. The result has implications for policies that restrict price changes or require disclosures.

Keywords: Revenue management, sequential estimation, dynamic pricing, learning, competition

Contents

1 Introduction

Developing effective pricing strategies in new or uncertain markets involves balancing the tradeoff between learning about consumer demand and earning profits in each period. For markets with a small number of firms, this tradeoff becomes increasingly difficult to manage as firms must not only account for the unknown demand from consumers, but also for the response from their competitors. As the market matures and firms learn from the experience of selling their products, each firm becomes more proficient at setting prices; however, the resulting increased level of competition may cause profits to decline for the industry as a whole. This effect brings into question whether it is in the best interest of firms to actively learn about their business environment, or rely only on their current information to compete for profits. In other words, do competing firms always want to learn about demand?

Our primary research focus concerns the impact of competition on the design and implementation of dynamic pricing strategies. Dynamic pricing strategies are a useful tool for firms that face a changing marketplace for their product. As market conditions evolve over time, firms can employ such strategies by adjusting their prices periodically to match their environment. For example, consider the market for a single, new product where the demand for that product is initially unknown. From the firm's perspective, the marketplace is changing as data from the sales of the product is collected, providing more information about the underlying demand. In each period, the firm must weigh the expected revenue that their pricing decision will yield in the current period against the information value that charging varied or experimental prices may provide for future pricing decisions. Balancing this learning and earning tradeoff for a monopolist firm has been the focus of recent research in revenue management and has produced effective pricing policies under various forms of uncertainty. Our goal is to assess the performance of these dynamic strategies when multiple firms are selling products in an uncertain marketplace and examine the effects of competition on the learning and earning tradeoff.

Designing strategies that account for both demand learning and the impact of competition leads to the following questions. If firms adopt dynamic pricing strategies taken from the

monopolist literature for demand learning, will they eventually learn the underlying demand conditions for their products? At same time, will revealing the market conditions increase the degree of competition and "learn" away profits for all of the firms involved? To address these questions, we introduce two models of competition under demand uncertainty. The first tests the performance of various monopolist strategies and heuristics in these competitive environments, while the second explores potential equilibrium strategies for these markets.

The first model allows each competing firm to choose from a family of pricing strategies from the monopolist literature on dynamic pricing. When these strategies are applied simultaneously, we show that the prices of the individual firms can approach an overall Nash equilibrium of the game with complete information if the firms sufficiently vary their prices. On the other hand, if firms do not vary prices sufficiently, the firms can incompletely learn the demand with prices that converge to a different set of prices from the complete-information Nash equilibrium. These prices can be beneficial or harmful to individual firms relative to the complete-information Nash equilibrium, and, in some case, can actually be better than the complete-information solution for all firms (an aspect which we explore more fully in the second part of the paper).

The model assumes that each firm estimates its underlying demand curves using least squares estimation. We find that these strategies do not necessarily produce desirable outcomes. Firms that face uncertainty from both their consumers and their competitors need to account for how their pricing decisions will influence their competitors' future prices. Firms eventually learn their underlying demand curves, but drive the market into a state of increased competition. Instead, firms would prefer to remain ignorant of the true demand in order to charge collusive prices in the market.

Our second model shows that willful ignorance is not merely a byproduct of the dynamic pricing strategies we investigated, but rather is a rational outcome for firms competing in a perfect Bayes equilibrium (PBE). In particular, we develop several simple demand environments, where the equilibrium strategies for the competing firms actively avoid learning the true value of demand and attain a collusive outcome even in finite time horizons.

2 Literature Review

This paper relates to an extensive literature on dynamic learning and pricing in the operations research literature. These papers build on the classical work of Lai and Robbins [1982] that sequential optimization and estimation of a linear model can lead to incomplete learning without sufficient variation in the choice of the controls. The main focus in the subsequent papers is then to generate conditions under which the model can be learned completely and the controls can yield the same rewards asymptotically as with full information. Papers in this stream include Besbes and Zeevi [2009], Keskin and Zeevi [2014], den Boer and Zwart [2013], Cheung et al. [2015], Harrison et al. [2012], and Broder and Rusmevichientong [2012].

In contrast to most of the work in this area, our paper considers a competitive environment in which individual agents have private information. Some papers that explore this area include Cooper et al. [2015] and Bertsimas and Perakis [2006]. Cooper et al. [2015] considers a case where one agent is unaware of the competition and shows that this can possibly attain better outcomes from all firms than if they knew they were in competition. The Bertsimas and Perakis paper considers a similar situation to that of our model but considers only restrictive strategies and more of a full information situation. In contrast, our work considers asymmetric information but full knowledge of the competitive landscape.

This paper also relates to the extensive economics literature on collusive pricing and equilibrium price experimentation as well as the recent research on dynamic pricing and learning in stationary demand environments. In particular, the second part of the paper explores the maintenance of collusion as, for example, in Stigler [1964] and Maskin and Tirole [1988], but instead of uncertainty or non-differentiability leading to collusive arrangements, collusion in our case will follow from uncertainty and the threat of providing information to competitors.

Our modeling framework fits into the literature on prisoner's dilemma situations and the maintenance of cooperative behavior. Such papers include Kreps et al. [1982], Green and Porter [1984], Abreu et al. [1990], Sannikov and Skrzypacz [2007], and Rahman [2014]. The mechanism to sustain collusion, however, in this paper is again different in its reliance

on the potential information leak from individual agent actions, leading to disincentives to experiment and incomplete but ultimately beneficial incomplete learning.

Pricing when demand is uncertain requires balancing the trade-off between learning about demand and earning revenues in the short run. In order to accurately estimate the parameters of a demand model, firms must vary the prices of their products across the selling season; however, this practice typically requires firms to deviate from the myopic price that earns the most expected revenue given their current information and forecasts. Early works by Rothschild [1974] and McLennan [1984] showed that following a myopic strategy can lead to incomplete learning in the long run. In the case of a monopolist firm, such strategies will inevitably underperform more forward looking policies that explore the potential pricing alternatives. By introducing a performance metric known as *regret*, researchers in revenue management have developed semi-myopic polices to avoid incomplete learning by strategically experimenting with prices. However, learning the demand curve in the long run may not be a desirable strategy for firms in a competitive marketplace. For instance, firms would prefer to have mistaken beliefs if the resulting market prices outperform the full-information competitive equilibrium. Adhering to a myopic strategy would therefore be rational in competitive environments provided that firms can recognize situations where their mistaken beliefs generate collusive prices. Our work extends the notion of regret to competitive environments, develops pricing strategies that avoid incomplete learning, and analyzes the connection between demand uncertainty and collusive pricing.

We should also note that the type of behavior in this paper may be observed in new markets, such as those for new classes of pharmaceuticals. In these cases, firms may not know the effect of price changes and tend not to vary prices until many entries have occurred. Recent experience for new Hepatitis C treatments and new cholesterol-lowering agents (PCSK9 inhibitors) appears to follow this pattern.

3 Setup and Notation

We consider N firms, each about to introduce a single, new product to the marketplace. Firms are aware of their competitors' products and that the prices chosen by their competitors will influence the demand for their own product. The underlying market conditions, however, are initially unknown. The demand for each distinct product is related to the others, but it is uncertain whether these products will behave as substitutes or complements, nor is it clear what magnitude a change each product's own price will have on its demand. Firms learn about the market conditions by pricing their product and privately observing their resulting sales.

We model their interactions as a T-period, dynamic game of incomplete information. Let $I := \{1, \ldots, N\}$ be the set of firms. The market conditions are characterized by a state of the world θ, which remains constant throughout the selling season, and is drawn randomly from a compact set Θ before selling begins. At the start of each period t, firms use their available data to estimate the state of the world and forecast the prices that their competitors will choose. Firms then price their products publicly and simultaneously with firm i choosing its price p_{it} from a compact subset of the positive real line $P \subset \Re^+$. After prices are chosen, each firm receives a private signal $d_{it} \in \Re$ drawn randomly from consumers. For our purposes, each firm's private signal represents that firm's observed demand for its product in period t. In other words, prices are observed by both consumers and competing firms, whereas the demand for each product is observed privately by each firm.

We are primarily focused on the role of dynamic pricing for the purposes of learning demand while earning profits. To that end, we assume that the firms have no capacity limitations for their products and ignore the effects of marginal and fixed costs. Observed sales therefore represent samples of the uncensored demand distribution, and firms can maximize their profits by choosing prices that increase revenues. The distribution of consumer demand for each product depends on the vector of prices chosen in that period, $p_t \in P^N$, and the unknown state of the world θ. Each firm's profit in a given period t is defined as the product of their chosen price and the consumer demand $u_i(p_t, d_{it}) = p_{it} d_{it}$.

4 Model 1: Efficient learning with least-squares estimates and restricted strategies

Model I evaluates the performance of myopic and semi-myopic heuristics from the revenue management literature on dynamic pricing. Drawing from the setup in Keskin and Zeevi [2014], each firm correctly assumes that their demand is a linear function of the prices charged in each period,

$$d_{it} = \theta_i^\top \begin{bmatrix} 1 \\ p_{1t} \\ \vdots \\ p_{Nt} \end{bmatrix} + \epsilon_{it}, \quad i = 1, \ldots, N.$$

The unknown market conditions $\theta = (\theta_1, \ldots, \theta_N)^\top$ are initially drawn from a compact set $\Theta \subset \Re^{N(N+1)}$ and the demand shocks, ϵ_{it} are drawn from a static, mean-zero distribution independently for each of the N products and serially independent across time. For convenience, let $x_t = (1,\ p_t^\top)^\top$, where $p_t = (p_{1t}, \ldots, p_{Nt})^\top$ is the vector of prices chosen by each firm in period t. It will also be useful to distinguish between the various demand parameters within each firm's demand model. Thus, with a slight overuse of notation, let $\theta_{i0} = \alpha_i$, $\theta_{ii} = \beta_i$ and $\theta_{ij} = \gamma_{ij}$ for $i \neq j$. The demand curve for firm i can now be expressed as

$$d_{it} = \alpha_i + \beta_i p_{it} + \sum_{j \neq i} \gamma_{ij} p_{jt} + \epsilon_{it}, \quad i = 1, \ldots, N.$$

Though equivalent, both the vector and component form for the demand curves will be useful for our analysis as will be evident in context.

At the start of each period t, firms use their available data to estimate the state of the world, to forecast the prices that their competitors will choose, and to determine their own pricing action. We describe each agent's strategy as comprising these elements of estimation, forecast, and action, all based on public information revealed in the past actions of all agents and each individual agent's private information on past payoffs.

In considering demand estimates, we note that observed sales represent samples of the uncensored demand distribution since we have ignored capacity considerations. As a simple

estimation strategy, we consider that each agent uses ordinary least squares regression. At the start of period t, firm i's least squares estimator of θ_i is

$$\hat{\theta}_{it} = \left(\sum_{s=1}^{t-1} x_s x_s^\top \right)^{-1} \left(\sum_{s=1}^{t-1} d_{is} x_s \right).$$

Since each firm's realized demand history $\{d_{is}\}_{i=\pm 1, s=1,\ldots,t}$ is private information, only firm i has knowledge of the least-squares estimate $\hat{\theta}_{it}$. However, the least-squares estimates for the N firms are interconnected as they share the same empirical Fisher information matrix given by

$$\mathcal{J}_t \;=\; \sum_{s=1}^{t} x_s x_s^\top.$$

This matrix plays a crucial role in our analysis, as its size (as measured by its smallest eigenvalue $\lambda_{\min}(\mathcal{J}_t) \geq 0$) indicates how close each firm's estimates are to the true value of the parameters. Combining the definition of the least-squares estimates with the true demand model yields the following expression for the estimation error:

$$\hat{\theta}_{it} - \theta_i = \mathcal{J}_t^{-1} \sum_{s=1}^{t} \epsilon_{is} x_t.$$

As the estimation error is inversely proportional to the empirical Fisher information, increasing the size of this matrix will enable each firm to learn their underlying demand; however, joint control of the Fisher information is non-trivial, since it is generated through the combined pricing actions of all of the firms in the market.

The distribution of the demand noise ϵ_{it} plays a important role in the estimation error as well. To ensure that the demand noise does not dominate the estimation error, we assume that it follows a light-tailed distribution, as in Keskin and Zeevi [2014]; that is, there exists a positive constant z_0 such that $\mathbb{E}[e^{z\epsilon_{it}}] < \infty$ for all $|z| < z_0$. The least-squares estimates can be improved further by using the knowledge that the true parameters θ belong to the compact set Θ. Let $\vartheta_{it} := \text{argmin}_{\vartheta \in \Theta_i}\{||\vartheta - \hat{\theta}_{it}||\}$ be the L^2-projection of the least-squares estimates onto firm i's subset of the demand parameter set $\Theta_i \subset \Theta$.

Along with their demand estimates, each firm develops a forecast for their competitors' future prices p_t^e at start of period t. We allow a range of assumptions that include the following that have appeared in the literature:

1. *Cournot adjustment*: $p_t^e = p_{t-1}$;
2. *Time average over H-horizon:* $p_t^e = \frac{1}{H}\sum_{\tau=t-H}^{t-1} p_\tau$ for $H < T$;
3. *Exponential smoothing:* $p_t^e = \sum_{\tau=1}^{t-1} \lambda^{t-\tau} p_\tau$ for $\lambda < 1$.

Each firm selects a forecasting strategy before selling begins; indeed, rival firms need not be aware of their competitors' choices. Given these estimates and forecasts, we define an estimated expected demand for each agent i's demand given their action p_{it} as $\hat{d}_{it}(p_t^e, p_{it}, \vartheta_{it})$.

A *general pricing strategy* for firm i is a sequence of functions $\sigma_i := (\sigma_{i1}, \sigma_{i2}, \ldots \sigma_{iT})$ where $\sigma_{it} : \Re^{(N+1)(t-1)} \to \mathcal{P}(P)$ is a measurable mapping from firm i's observable history $\mathcal{H}_{it} = (d_{i1}, p_1, \ldots, d_{i(t-1)}, p_{(t-1)})$, to the space of probability measures on the closed interval of prices $P \subset \Re^+$. However, by introducing the demand estimates and price forecasts, we consider a restricted set of *admissible pricing strategies* $\sigma_{it} : \Re^{2N+1} \to \mathcal{P}(P)$ such that firm i's price in period t is distributed as

$$p_{it} \sim \sigma_{it}(\vartheta_{it}, p_t^e),$$

where $p_t^e = \sum_{\tau=1}^{t-1} \lambda_\tau^t p_\tau$ with $\lambda_\tau^t \geq 0$ and $\sum_{\tau=1}^{t-1} \lambda_\tau^t \leq 1$ (i.e., p_t^e is in the convex hull of observed prices $p_1, \ldots, p_{t-1}$) with the following additional restriction: for some $0 < \bar{T} < \infty$ and all $t \geq \bar{T}$ and some $0 < \delta < 1$,

$$\sum_{i=1}^{t-1} \lambda_s^t \frac{\log s}{\sqrt{s}} \leq \frac{\delta^2}{\Gamma}\frac{\log t}{\sqrt{t}}, \lambda_1^t \leq (1-\delta)\frac{\log t}{\sqrt{t}}, \tag{4.1}$$

where $1 > \delta^2 > \Gamma := 2\max_{\theta\in\Theta}\left[\left(\frac{\sum_{j\neq i}\gamma_{ij}}{-2\beta_i}\right)^2\right]$ and the feasibility of this relationship holds by assumption. We note that the three forecasting policies given above, as well as many more-complex prediction functions, satisfy these assumptions.

Our analysis revolves around strategies that focus on best-response (as, for example, also assumed in Simon [2007]). Consider the firm's expected single-stage payoffs given their

estimates and forecasts. In this case, firms can maximize their single-stage profits by choosing prices that maximize expected revenues (since we assume zero marginal cost). Each firm's profit in a given period t is their best-response to the forecast prices of competitors defined as the product of their chosen price and the consumer demand $u_i(p_t, d_{it})$.

The single-period best-response function for firm i given an estimate ϑ_{it} and price forecast p_t^e is

$$\varphi(\vartheta_{it}, p_t^e) = \arg\max_{p_{it}} u_i(p_{it}, \hat{d}_{it}(p_t^e, p_{it}, \vartheta_{it})) = \frac{\hat{\alpha}_i}{-2\hat{\beta}_i} + \sum_{j \neq i} \frac{\hat{\gamma}_{ij}}{-2\hat{\beta}_i} p_{jt}^e.$$

If the true demand parameters θ were known, then solving for the best-response prices for each firm simultaneously yields the unique single-period Nash equilibrium prices:

$$p_i^{NE} = \varphi(\theta_i, p^{NE}).$$

Since the best-response in a monopolistic setting can yield incomplete learning, we add a perturbation to the best-response in the form of a noise term. We call this set of policies *best response with random dithering* (following Lobo and Boyd [2003]), given as follows:

$$\sigma_{it}(\vartheta_i, p_t^e) = \varphi(\vartheta_{it}, p_t^e) + \nu_{it},$$

such that the ν_{it} are mean-zero random variables, that are conditionally independent over i and t given the history at time t. Therefore, at the start of period t and given history $\mathcal{H}_{it}$, firm i's price is a random variable with distribution $\sigma_{it}(\vartheta_i, p_t^e)$ and mean $\varphi(\vartheta_{it}, p_t^e)$. A simple way to generate admissible random dithering policies is to choose distributions for the random noise terms before the selling season begins and then truncate these distributions in each period so that prices remain in P.

Note that in order for best response with random dithering polices to be *admissible* strategies, the prices that each firm charges in each period must lie in a closed interval P. An immediate consequence of this property is that the ν_{it} noise terms must be bounded in each period and that the best response function $\varphi(\vartheta_i, p^e) \in P$ for all $\vartheta_i \in \Theta_i$ and all $p^e \in P^N$. To that end, for a random dithering strategy to be an *admissible random dithering strategy*, we assume that each firm's best-response price lies within an open interval $\text{int}(P)$ whenever

its competitors' prices also lie within that interval. This assumption allows for the added noise terms ν_{it} to have positive variance in each period and has the following implications for the parameter space Θ.

Proposition 4.1. If $\varphi(\vartheta_i, p_t^e)) \in \text{int}(P)$ for all $\vartheta_i \in \Theta_i$ and all $p_t^e \in P^N$, then:

i. $\beta_i < 0$ for all $i \in 1 \dots N$ and all $\theta \in \Theta$;

ii. $|\sum_{j \neq i} \gamma_{ij}| \leq -2\beta_i$ for all $i \in 1 \dots N$ and all $\theta \in \Theta$.

Proof. First note that if $\beta_i \geq 0$, then demand for product i increases with the price of product i and by definition the best response would be unbounded $\varphi(\vartheta_i, p_t^e) = \infty$. Let $P = [l, u]$ where $u > l > 0$. Then the second property comes from enforcing that $\varphi(\vartheta_i, ue) \leq u$ and $\varphi(\vartheta_i, le) \geq l$, where e is the N-vector of ones. Combining these inequalities yields $\varphi(\vartheta_i, ue) - \varphi(\vartheta_i, le) \leq u - l$ or $(\sum_{j \neq i} \hat{\gamma}_{ij} / -2\hat{\beta}_i)(u - l) \leq u - l$. Since both $-2\hat{\beta}_i$ and $u - l$ are positive, this implies that $\sum_{j \neq i} \hat{\gamma}_{ij} \leq -2\hat{\beta}_i$. Similarly, the inequality $-(\sum_{j \neq i} \hat{\gamma}_{ij}) \leq -2\hat{\beta}_i$ follows through the combination of $\varphi(\vartheta_i, ue) \geq l$ and $\varphi(\vartheta_i, le) \leq u$. □

Additionally, it is necessary to strengthen the second property of Proposition 4.1 and assume that the parameter space has the property that $|\sum_{j \neq i} \gamma_{ij}| \leq -\beta_i$ for all $i \in 1 \dots N$ and all $\theta \in \Theta$. In the following, we expand our definition of *admissible strategies* to include this restriction on the parameter space. This assumption is required due to a byproduct of our proof technique which bounds the forecasting errors inherent to our allowed options for competitor forecasts.

We restrict the analysis below to this limited set of admissible policies since finding equilibria in fully general settings is not realistic. It requires, for example, extraordinary rationality (see Blume and Easley [1995]) for each agent to fully consider all other agents' information states, updating capabilities, and strategy choices. In Model 2 (Section 5), however, we analyze a restricted action and parameter space setting where we can more fully describe equilibria in the repeated game.

4.1 Measuring Regret

In order to analyze the pricing strategies outlined in the previous section, we introduce a performance metric called *regret*, which compares the revenues generated by each firm to the revenues generated by the competitive Nash equilibrium. The T-period regret for firm i is given by

$$\Delta_i^\sigma(\theta, T) = \sum_{t=1}^{T} \mathbb{E}_\theta^\sigma[p_i^{NE}(\alpha_i + \beta_i p_i^{NE} + \sum_{j \neq i} \gamma_{ij} p_j^{NE}) - p_{it}(\alpha_i + \beta_i p_{it} + \sum_{j \neq i} \gamma_{ij} p_{jt})],$$

where the expectation $\mathbb{E}_\theta^\sigma$ is taken with respect to the probability of the price sequences $\{p_t\}_{t=1,\ldots,T}$ induced by the firms' chosen strategies σ and the true value of demand parameters θ. We note that this is distinct from the definition of regret often used in the machine learning literature (see Zinkevich et al. [2007]). In that context, regret is measured as the difference relative to the best possible static competitor *actions*; that is, rather than comparing realized profits to the complete-information Nash equilibrium, firms would instead compare their outcomes to the profits they would have accrued if their competitors chose to charge a constant price in each period. We consider our admissible policy responses as more fully capturing rational behavior. As all firms are pricing to learn demand and earn revenues, charging a constant price would prohibit learning not only for those firms that choose to adopt that strategy, but also for the marketplace as a whole.

Choosing to benchmark against the competitive Nash outcome is also useful because it has a direct analogue to the monopolist learning and earning literature. In particular, it separates into two quantities of interest that we describe as the *regret due to learning* and the *regret due to influence* as the two terms in the following representation of the regret:

$$\Delta_i^\sigma(\theta, T) \;\; = \;\; -\beta_i \sum_{t=1}^{T} \mathbb{E}_\theta^\sigma \left[(p_i^{NE} - p_{it})^2\right] - 2\beta_i \sum_{t=1}^{T} \mathbb{E}_\theta^\sigma \left[p_{it}\left(\varphi(\theta_i, p^{NE}) - \varphi(\theta_i, p_t)\right)\right].$$

The first term in the sum measures the expected squared distance between firm i's Nash equilibrium price and the prices chosen in each period according to its strategy σ_i. We identify this term as measuring firm i's regret due to learning, as it tracks the firm's knowledge of the underlying demand parameters and the system's progress towards the complete-information Nash equilibrium. The expected squared distance is precisely the regret metric

that a monopolist would face in this market when balancing the tradeoffs between learning and earning. The worst-case regret of a monopolist is

$$\Delta^{\sigma}(T) \leq \sup\{-|K| \sum_{t=1}^{T} \mathbb{E}_{\theta}^{\sigma} \left[||p^{CE} - p_t||^2\right] : \theta \in \Theta\},$$

where p^{CE} is the vector of cooperative Nash equilibrium prices and K is a known constant. The remaining term, called the *regret due to influence*, represents the regret that firm i incurs as a result of the mistaken beliefs about its competitor. Notice that when the products in the market are substitutes (complements), this term is negative when competing firms charge above (below) their Nash equilibrium prices. Our definition of regret allows for such negative values as in learning scenarios in which a firm i can exploit the ignorance of its competitors to earn additional revenues. Hence, firms may consider allowing an increase in the regret due to learning in an effort to influence their competitor to charge a more favorable price.

4.2 Conditions for Efficient Learning

This section analyzes the use of the admissible strategies in achieving an efficient learning outcome, as measured by the worst-case regret due to learning. In particular, we show first that the information grows at the rate of the variance of the pricing strategies σ, as controlled by the noise terms in random dithering policies. These results depend on the information metric $\mathcal{J}_t$ and its minimum eigenvalue, denoted $\lambda_{\min}(\mathcal{J}_t)$. To bound this minimum eigenvalue, we use the following matrix version of the Freedman bound. Let $\mathbb{E}_s$ and Var_s denote the conditional expectation and variance of an adapted sequence of random matrices given a history of realizations up to time s.

Theorem 4.2 (Matrix Freedman, Tropp [2011])**.** Consider a finite adapted sequence $\{Y_s\}$ of random, self-adjoint matrices with dimension d. Assume that

$$\mathbb{E}_{s-1} Y_s = 0 \text{ and } \lambda_{\max}(Y_s) \leq R \text{ almost surely.}$$

Define the finite series

$$Z := \sum_s Y_s \text{ and } W := \sum_s \mathbb{E}_{s-1}(Y_s^2).$$

Then, for $\delta \geq 0$ and $v > 0$,

$$\mathbb{P}\left\{\lambda_{\max}(Z) \geq \delta \text{ and } \lambda_{\max}(W) \leq v^2\right\} \leq d \cdot \exp\left(\frac{-\delta^2/2}{v^2 + R\delta/3}\right).$$

The Freedman bound can be applied directly to our model as follows:

Corollary 4.3. Let $\mathcal{J}_t = \sum_{s=1}^{t} x_s x_s^\top$ be the Fisher information matrix at time $t \leq T$ generated by admissible strategies σ. Assume that $\exists v > 0$ such that

$$||\sum_{s=1}^{t} \text{Var}_{s-1}[x_s x_s^\top]|| < v, \quad \text{almost surely,}$$

then

$$\mathbb{P}\left\{\lambda_{\min}(\mathcal{J}_t) \leq -\delta + \sum_{s=1}^{t} \lambda_{\min}\left(\mathbb{E}_{s-1}[x_s x_s^\top]\right)\right\} \leq N \cdot \exp\left(\frac{-\delta^2/2}{v^2 + R\delta/3}\right).$$

Proof. This follows through a direct application of the Freedman bound and Weyl's inequalities. Since the sequence of public price vectors x_s is a finite sequence of adapted random vectors, the random matrices $Y_s := \mathbb{E}_{s-1}[x_s x_s^\top] - x_s x_s^\top$ form an adapted sequence of random, self-adjoint matrices each with conditional mean zero.

By the definition of our admissible strategies, the prices charged in each period lie in an interval $P = [l, u]$ with $0 < l < u$, so $||x_t||^2 \leq Nu^2 := R$. The fact that prices are uniformly bounded implies that the maximum eigenvalues of the Y_s are also uniformly bounded by the following argument. Consider the spectral norm of the Y_s,

$$||Y_s|| = ||\mathbb{E}_{s-1}[x_s x_s^\top] - x_s x_s^\top|| = \max\{||\mathbb{E}_{s-1}[x_s x_s^\top]||, ||x_s x_s^\top||\} \leq R.$$

The above expression for spectral norm follows from the fact that both $\mathbb{E}_{s-1}[x_s x_s^\top]$ and $x_s x_s^\top$ are positive semi-definite matrices and the bound follows from the following application of Jensen's inequality:

$$||\mathbb{E}_{s-1}[x_s x_s^\top]|| \leq \mathbb{E}_{s-1}||x_s x_s^\top|| \leq \mathbb{E}_{s-1}||x_s||^2 \leq R.$$

Since the spectral norm of a Hermitian matrix is $||X|| = \max\{\lambda_{\max}(X), -\lambda_{\min}(X)\}$, we have that

$$\lambda_{\max}(Y_s) \leq ||Y_s|| \leq R.$$

Applying the Matrix Freedman bound yields:

$$\begin{aligned} n\exp\left(\frac{-\delta^2/2}{\sigma^2+R\delta/3}\right) &\geq \mathbb{P}\left\{\lambda_{\max}\left(-\mathcal{J}_t+\sum_{s=1}^{t}\mathbb{E}_{s-1}[x_s x_s^\top]\right)\geq\delta\right\} \\ &\geq \mathbb{P}\left\{\lambda_{\max}(-\mathcal{J}_t)\geq\delta-\sum_{s=1}^{t}\lambda_{\min}(\mathbb{E}_{s-1}[x_s x_s^\top])\right\} \\ &= \mathbb{P}\left\{\lambda_{\min}(\mathcal{J}_t)\leq-\delta+\sum_{s=1}^{t}\lambda_{\min}(\mathbb{E}_{s-1}[x_s x_s^\top])\right\}. \end{aligned}$$

Notice that the matrix $W=\sum_{s=1}^{t}\mathrm{Var}_{s-1}[x_s x_s^\top]$ in the Freedman bound is omitted. This is due to the fact that the spectral norm of a matrix is greater than or equal its maximum eigenvalue and the condition $\lambda_{\max}(W)\leq v^2$ holds almost surely by assumption. □

We can now use the corollary above to bound the minimum eigenvalues of the Fisher information matrix in our setting.

Lemma 4.4. If firms choose admissible strategies and there exist constants $c_i, C_i > 0$ such that $\frac{c_i}{\sqrt{t}} \leq \mathrm{Var}_{t-1}[\sigma_{it}] \leq \frac{C_i}{\sqrt{t}}$ almost surely for $i = 1,\ldots,N$ and $t < T$, then there exist constants $\kappa_0, \kappa_1 > 0$ such that

$$\mathbb{P}(\lambda_{\min}(\mathcal{J}_t) < \kappa_0\sqrt{t}) \leq \frac{\kappa_1}{\sqrt{t}}.$$

Proof. The proof will proceed as follows. First, we will provide a lower bound for the minimum eigenvalue for the sum of the conditional expectations of $x_s x_s^\top$ for $s = 1,\ldots,t$. Second, we will provide an upper bound for the sum of the conditional variances of $x_s x_s^\top$ for $s = 1,\ldots,t$. These bounds then allow us to apply the Matrix Freedman corollary and complete the proof. We express the price vectors in each period as the sum:

$$x_s = z_s + \nu_s,$$

where $z_s := \mathbb{E}_{s-1}[x_s]$ and $\nu_s \in \Re^{N+1}$ is a random vector with zero mean and variance equal to $\mathrm{Var}_{s-1}[x_s]$.

Claim 1. $\lambda_{\min}\left(\sum_{s=1}^{t}\mathbb{E}_{s-1}[x_s x_s^\top]\right) \geq \delta_0\sqrt{t}$ for some $\delta_0 > 0$.

Proof of Claim 1

Separate the conditional expectations into z_s and ν_s components,

$$\begin{aligned}\sum_{s=1}^{t}\mathbb{E}_{s-1}[x_s x_s^\top] &= \sum_{s=1}^{t} z_s z_s^\top + \sum_{s=1}^{t}\mathbb{E}_{s-1}[\nu_s\nu_s^\top] \\ &= \sum_{s=1}^{t}(z_s-\bar{z})(z_s-\bar{z})^\top + \sum_{s=1}^{t}\bar{z}\bar{z}^\top + \sum_{s=1}^{t}\mathbb{E}_{s-1}[\nu_s\nu_s^\top] \\ &\succeq \sum_{s=1}^{t}\bar{z}\bar{z}^\top + \sum_{s=1}^{t}\mathbb{E}_{s-1}[\nu_s\nu_s^\top],\end{aligned}$$

where $\bar{z} := \frac{1}{t}\sum_{s=1}^{t} z_s$. Let $y = (y_1, \ldots, y_{N+1}) \in \Re^{N+1}$ be an arbitrary unit vector and y_* and $\bar{z}_*$ be vectors consisting of the last N components of $y, \bar{z}$, respectively. Let $c = \min_{i=1,\ldots N}\{c_i\}$, then

$$\begin{aligned}y^\top\left(\sum_{s=1}^{t}\bar{z}\bar{z}^\top + \sum_{s=1}^{t}\mathbb{E}_{s-1}[\nu_s\nu_s^\top]\right)y &= \sum_{s=1}^{t}(y_1 + y_*^\top\bar{z}_*)^2 + \sum_{i=1}^{N} y_{i+1}^2\mathrm{Var}_{s-1}[\sigma_{is}] \\ &\geq \sum_{s=1}^{t}(y_1 - ||y_*||\cdot||\bar{z}_*||)^2 + ||y_*||^2\frac{c}{\sqrt{t}} \\ &\geq (y_1 - ||y_*||\cdot||\bar{z}_*||)^2 t + ||y_*||^2 c\sqrt{t}.\end{aligned}$$

Since y is an arbitrary unit vector, the Rayleigh-Ritz theorem implies that the minimum eigenvalue of $\sum_{s=1}^{t}\mathbb{E}_{s-1}[x_s x_s^\top]$ grows by least $\delta_0\sqrt{t}$ for some constant $\delta_0 > 0$. †

Claim 2. $||\sum_{s=1}^{t}\mathrm{Var}_{s-1}[x_s x_s^\top]|| \leq v_0\sqrt{t}$ almost surely for some $v_0 > 0$.

Proof of Claim 2

Let $C = \max_{i=1,\ldots,N}\{C_i\}$ and $R = Nu^2$ where $P = [l, u]$ is the admissible price interval with

$0 < l < u$. By the definition of the conditional variance for random matrices:

$$\begin{aligned}
\mathrm{Var}_{s-1}\left[x_s x_s^\top\right] &= \mathrm{Cov}_{s-1}\left[z\nu_s^\top, x_s x_s^\top\right] + \mathrm{Cov}_{s-1}\left[\nu_s z^\top, x_s x_s^\top\right] + \mathrm{Cov}_{s-1}\left[\nu_s\nu_s^\top, x_s x_s^\top\right] \\
&= \mathbb{E}_{s-1}\left[(\nu_s^\top\nu_s) z z^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top z) z\nu_s^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top\nu_s) z\nu_s^\top\right] \\
&\quad + \mathbb{E}_{s-1}\left[(\nu_s^\top z)\nu_s z^\top\right] + \mathbb{E}_{s-1}\left[(z^\top z)\nu_s\nu_s^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top z)\nu_s\nu_s^\top\right] \\
&\quad + \mathbb{E}_{s-1}\left[(\nu_s^\top\nu_s)\nu_s z^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top z)\nu_s\nu_s^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top\nu_s)\nu_s\nu_s^\top\right] \\
&\quad - \mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right]^2 \\
&= \mathbb{E}_{s-1}\left[||\nu_s||^2 x_s x_s^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top z) x_s x_s^\top\right] + \mathbb{E}_{s-1}\left[(\nu_s^\top z)\nu_s\nu_s^\top\right] \\
&\quad + (z^\top z)\mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right] - \mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right]^2.
\end{aligned}$$

Apply the spectral norm to both sides of the above equation and use Jensen's inequality on each term to generate an upper bound,

$$\begin{aligned}
||\mathrm{Var}_{s-1}\left[x_s x_s^\top\right]|| &\leq \mathbb{E}_{s-1}\left[||\nu_s||^2 ||x_s x_s^\top||\right] + \mathbb{E}_{s-1}\left[|\nu_s^\top z|\ ||x_s x_s^\top||\right] \\
&\quad + \mathbb{E}_{s-1}\left[|\nu_s^\top z|\ ||\nu_s\nu_s^\top||\right] + \ z^\top z\mathbb{E}_{s-1}\left[||\nu_s\nu_s^\top||\right] \\
&\leq R\left(\mathbb{E}_{s-1}\left[||\nu_s||^2\right] + 2\mathbb{E}_{s-1}\left[|\nu_s^\top z|\right] + ||\mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right]||\right) \\
&\leq R\left(\mathbb{E}_{s-1}\left[||\nu_s||^2\right] + 2|z|^\top\mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right]^{1/2} + ||\mathbb{E}_{s-1}\left[\nu_s\nu_s^\top\right]||\right) \\
&\leq \frac{N\cdot R\cdot C}{\sqrt{s}} + O(\frac{1}{\sqrt{s}}).
\end{aligned}$$

The resulting matrix variance statistic is then

$$\sum_{s=1}^{t} ||\mathrm{Var}_{s-1}[x_s x_s^\top]|| \leq \sum_{s=1}^{t} \frac{NR}{\sqrt{s}} + O(\frac{1}{\sqrt{s}}) \leq v_0\sqrt{t},$$

for some $v_0 > 0$. †

Let $\delta = \frac{\delta_0}{2}\sqrt{t}$ and $v = v_0\sqrt{t}$ and apply the Matrix Freedman corollary,

$$\begin{aligned}
N\cdot\exp\left(\frac{-\delta^2/2}{v^2 + R\delta/3}\right) &\geq \mathbb{P}\left\{\lambda_{\min}(\mathcal{J}_t) \leq -\delta + \sum_{s=1}^{t}\lambda_{\min}\left(\mathbb{E}_{s-1}[x_s x_s^\top]\right)\right\} \\
Ne^{-\kappa_1\sqrt{t}} &\geq \mathbb{P}\left\{\lambda_{\min}(\mathcal{J}_t) \leq \kappa_0\sqrt{t}\right\}.
\end{aligned}$$

□

We now invoke the following result from Keskin and Zeevi [2014].

Lemma 4.5 (Keskin and Zeevi [2014], Lemma 3)**.** There exist finite positive constants ρ and k such that,

$$\mathbb{P}\left\{||\hat{\theta}_{it} - \theta_i|| > \delta, \lambda_{\min}(\mathcal{J}_t) \geq m\right\} \leq kt \exp\left(-\rho(\delta \wedge \delta^2)m\right),$$

for all $\delta, m > 0$ and all $t \geq 3$.

The proof approach in the result above combined with our earlier results then yield the following result for the overall estimation error.

Proposition 4.6. If firms choose admissible strategies and there exist constants $c_i, C_i > 0$ such that $\frac{c_i}{\sqrt{t}} \leq \mathrm{Var}_{t-1}[\sigma_{it}] \leq \frac{C_i}{\sqrt{t}}$ almost surely for $i = 1, \ldots, N$ and $t < T$, then firm i's expected estimation error in period t is:

$$\mathbb{E}_\theta^\sigma\left[||\theta_i - \vartheta_{it}||^2\right] = O\left(\frac{\log(t)}{\sqrt{t}}\right).$$

Proof. We break up the expectation into cases where the minimum eigenvalue of $\mathcal{J}_t$ is large with respect to the current time period and when it is small.

$$\begin{aligned}
\mathbb{E}_\theta^\sigma\left[||\theta_i - \vartheta_{it}||^2\right] &= \int_0^\infty \mathbb{P}_\theta^\sigma\left(||\theta_i - \vartheta_{it}||^2 > x, \lambda_{\min}(\mathcal{J}_t) \geq \kappa_0\sqrt{t}\right) dx \\
&\quad + \int_0^\infty \mathbb{P}_\theta^\sigma\left(||\theta_i - \vartheta_{it}||^2 > x, \lambda_{\min}(\mathcal{J}_t) < \kappa_0\sqrt{t}\right) dx \\
&\leq \int_0^\infty \mathbb{P}_\theta^\sigma\left(||\theta_i - \hat{\theta}_{it}||^2 > x, \lambda_{\min}(\mathcal{J}_t) \geq \kappa_0\sqrt{t}\right) dx \\
&\quad + \int_0^{K_1} \mathbb{P}_\theta^\sigma\left(||\theta_i - \vartheta_{it}||^2 > x, \lambda_{\min}(\mathcal{J}_t) < \kappa_0\sqrt{t}\right) dx \\
&\leq \int_0^\infty \mathbb{P}_\theta^\sigma\left(||\theta_i - \hat{\theta}_{it}||^2 > x, \lambda_{\min}(\mathcal{J}_t) \geq \kappa_0\sqrt{t}\right) dx \\
&\quad + K_1 \mathbb{P}_\theta^\sigma\left(\lambda_{\min}(\mathcal{J}_t) < \kappa_0\sqrt{t}\right),
\end{aligned}$$

where $K_1 = \max_{\theta,\theta' \in \Theta} ||\theta - \theta'||^2$. The first inequality is due to the fact that the estimation errors of the projected least squares estimate ϑ_{it} are bounded by K_1 and are weakly smaller than the estimation errors of $\hat{\theta}_{it}$.

In the proof of Keskin and Zeevi [2014], Theorem 2, the authors prove that Lemma 4.5 implies that

$$\int_0^\infty \mathbb{P}_\theta^\sigma \left(||\theta_i - \hat{\theta}_{it}||^2 > x, \lambda_{\min}(t) \geq \kappa_0 \sqrt{t} \right) dx = O(\frac{\log t}{\sqrt{t}}).$$

Since the steps are identical to their analysis, they are omitted for brevity. We can then apply Lemma 4.4 to obtain the desired conclusion:

$$\mathbb{E}_\theta^\sigma \left[||\theta_i - \vartheta_{it}||^2 \right] \leq \frac{A_0 \log t}{\sqrt{t}} + \frac{A_1}{\sqrt{t}},$$

for some constants $A_0, A_1 > 0$. Therefore, the expected estimation error in each period is order $O(\frac{\log t}{\sqrt{t}})$. □

4.3 Low Regret under Random Dithering Policies

Using the information result of the previous section, we can devise a best-response with random dithering strategy that achieves low regret due to learning.

Theorem 4.7. If firms adopt best-response with admissible random dithering strategies and there exist constants $c_i, C_i > 0$ such that $\frac{c_i}{\sqrt{t}} \leq \text{Var}_{t-1}[\sigma_{it}] \leq \frac{C_i}{\sqrt{t}}$ almost surely for $i = 1, \ldots, N$ and $t < T$, then the worst-case regret due to learning is $O(\sqrt{T} \log T)$ for all firms.

Proof. Consider the maximum t period contribution to regret due to learning for any firm. Let $i^* = \arg\max_{i \in I} \mathbb{E}_\theta^\sigma \left[(p_i^{NE} - p_{it})^2 \right]$. Next, separate the regret due to learning into estimation error and forecast error and dithering as follows:

$$\begin{aligned}
\mathbb{E}_\theta^\sigma \left[(p_{i^*}^{NE} - p_{i^*t})^2 \right] &= \mathbb{E}_\theta^\sigma \left[(\varphi(\theta_{i^*}, p^{NE}) - \varphi(\vartheta_{i^*t}, p_t^e) - \nu_{i^*t})^2 \right] \\
&\leq 2\mathbb{E}_\theta^\sigma \left[\left(\varphi(\theta_{i^*}, p^{NE}) - \varphi(\vartheta_{i^*t}, p^{NE}) \right)^2 \right] \\
&\quad + 2\mathbb{E}_\theta^\sigma \left[\left(\varphi(\vartheta_{i^*t}, p^{NE}) - \varphi(\vartheta_{i^*t}, p_t^e) - \nu_{i^*t} \right)^2 \right] \\
&= 2\mathbb{E}_\theta^\sigma \left[\left(\varphi(\theta_{i^*}, p^{NE}) - \varphi(\vartheta_{i^*t}, p^{NE}) \right)^2 \right] \\
&\quad + 2\mathbb{E}_\theta^\sigma \left[\left(\varphi(\vartheta_{i^*t}, p^{NE}) - \varphi(\vartheta_{i^*t}, p_t^e) \right)^2 \right] + 2\text{Var}_\theta^\sigma(\nu_{i^*t}).
\end{aligned}$$

The first term on the right hand side representing the estimation error component can be bounded using the mean value theorem:

$$|\left(\varphi(\theta_{i^*}, p^{NE}) - \varphi(\vartheta_{i^*t}, p^{NE})\right)| \leq \max_{\substack{\omega \in \Theta, j \in I, \\ k \in \{1,\ldots,N+1\}}} \sqrt{(N+1)} \left|\frac{\partial \varphi(\omega_j, p^{NE})}{\partial \omega_{jk}}\right| ||\theta_{i^*} - \vartheta_{i^*t}||.$$

Therefore,

$$\begin{aligned}\mathbb{E}_\theta^\sigma\left[(p_{i^*}^{NE} - p_{i^*t})^2\right] &\leq 2K_0 \mathbb{E}_\theta^\sigma\left[||\theta_{i^*} - \vartheta_{i^*t}||^2\right] \\ &\quad + 2\mathbb{E}_\theta^\sigma\left[\left(\varphi(\vartheta_{i^*t}, p^{NE}) - \varphi(\vartheta_{i^*t}, p_t^e)\right)^2\right] + 2\operatorname{Var}(\nu_{i^*t}),\end{aligned}$$

where K_0 is defined by the maximization in the previous equation.

Let $\vartheta_{i^*t} = (a_{i^*t}, b_{i^*t}, c_{i^*jt})^\top$ and substitute the definition of $\varphi(\vartheta_{i^*t}, p_t^e)$ into the above equation:

$$\begin{aligned}\mathbb{E}_\theta^\sigma\left[(p_{i^*}^{NE} - p_{i^*t})^2\right] &\leq 2K_0 \mathbb{E}_\theta^\sigma\left[||\theta_{i^*} - \theta_{i^*t}||^2\right] \\ &\quad + 2\mathbb{E}_\theta^\sigma\left[\left(\sum_{j \neq i^*} \frac{c_{ijt}}{-2b_{it}}(p_j^{NE} - p_{jt}^e)\right)^2\right] + 2\operatorname{Var}(\nu_{i^*t}).\end{aligned}$$

Choosing the firm with the largest gap between the Nash equilibrium price and the forecasted price,

$$\begin{aligned}\mathbb{E}_\theta^\sigma\left[(p_{i^*}^{NE} - p_{i^*t})^2\right] &\leq 2\max_{\vartheta \in \Theta}\left\{\left(\frac{\sum_{j \neq i^*} c_{i^*jt}}{-2b_{i^*t}}\right)^2\right\} \max_{j \in I}\left\{\mathbb{E}_\theta^\sigma\left[\left(p_j^{NE} - p_{jt}^e\right)^2\right]\right\} \\ &\quad + 2K_0 \mathbb{E}_\theta^\sigma\left[||\theta_{i^*} - \theta_{i^*t}||^2\right] + 2\operatorname{Var}(\nu_{i^*t}).\end{aligned}$$

Note that $\Gamma = 2\max_{\theta \in \Theta}\left[\left(\frac{\sum_j \gamma_j}{-2\beta}\right)^2\right] < 1$ as used in the admissible pricing condition (4.1) by both Proposition 4.1 and the stronger assumption that $|\sum_{j \neq i} \gamma_{ij}| \leq -\beta_i$ for all $i \in 1 \ldots N$ and all $\theta \in \Theta$. Therefore, when firms use an admissible forecast scheme the regret due to learning is

$$\begin{aligned}\mathbb{E}_\theta^\sigma\left[(p_{i^*}^{NE} - p_{i^*t})^2\right] &\leq \Gamma \max_{j \in I}\left\{\mathbb{E}_\theta^\sigma\left[\left(p_j^{NE} - p_{jt}^e\right)^2\right]\right\} \\ &\quad + 2K_0 \mathbb{E}_\theta^\sigma\left[||\theta_{i^*} - \theta_{i^*t}||^2\right] + 2\operatorname{Var}(\nu_{i^*t}).\end{aligned}$$

Applying Proposition 4.6, and the assumptions on the variance of the added noise,

$$\mathbb{E}_\theta^\sigma\left[(p_{i^*}^{NE} - p_{i^*t})^2\right] \leq \Gamma \max_{j \in I}\left\{\mathbb{E}_\theta^\sigma\left[\left(p_j^{NE} - p_{jt}^e\right)^2\right]\right\} + K(1-\delta)\frac{\log t}{\sqrt{t}},$$

where δ is a parameter that satisfies the admissible pricing properties in (4.1) and $K > 0$ is determined by δ, the bounds C_{i^*}, K_0, and the upper bound on the estimation error. Next, representing the price forecasts in the above equation as a weighted sum of past prices,

$$\mathbb{E}_\theta^\sigma \left[(p_{i^*}^{NE} - p_{i^*t})^2\right] \leq \Gamma \max_{j \in I} \left\{ \mathbb{E}_\theta^\sigma \left[\left(p_j^{NE} - \sum_{\tau=1}^{t-1} \lambda_\tau^t p_{j\tau} \right)^2 \right] \right\} + K(1-\delta)\frac{\log t}{\sqrt{t}}.$$

Applying Jensen's inequality to the sum,

$$\begin{aligned} \mathbb{E}_\theta^\sigma \left[(p_{i^*}^{NE} - p_{i^*t})^2\right] &\leq \Gamma \max_{j \in I} \left\{ \mathbb{E}_\theta^\sigma \left[\sum_{\tau=1}^{t-1} \lambda_\tau^t \left(p_j^{NE} - p_{j\tau} \right)^2 \right] \right\} + K(1-\delta)\frac{\log t}{\sqrt{t}} \\ &\leq \Gamma \sum_{\tau=1}^{t-1} \lambda_\tau^t \max_{j \in I} \left\{ \mathbb{E}_\theta^\sigma \left[\left(p_j^{NE} - p_{j\tau} \right)^2 \right] \right\} + K(1-\delta)\frac{\log t}{\sqrt{t}}, \end{aligned}$$

which forms the autoregressive sequence:

$$y_t \leq \Gamma \sum_{\tau=1}^{t-1} \lambda_\tau^t y_\tau + K(1-\delta)\frac{\log t}{\sqrt{t}}, \tag{4.2}$$

where $y_t = \max_{i \in I} \mathbb{E}_\theta^\sigma \left[(p_i^{NE} - p_{it})^2\right]$. The proof proceeds to establish an order bound on $\sum_{t=1}^{\infty} y_t$ by bounding y_t for $t \geq \bar{T} + 1$, where the admissible pricing condition (4.1) holds for all $t \geq \bar{T}$.

With the admissible policy assumptions, we can show that for all $t \geq \bar{T} + 1$,

$$y_t \leq K\frac{\log t}{\sqrt{t}}, \tag{4.3}$$

where $y_1 < K < \infty$.

This can be assumed for all $2 \leq t \leq \bar{T} + 1$ for some choice of K (by increasing K in (4.2) if necessary); so, we follow by induction with the hypothesis for all $s \leq t - 1$ with $t > \bar{T} + 1$ and wish to show it holds for t. From (4.2) and the assumption of (4.3) for $s \leq t - 1$,

$$\begin{aligned} y_t &\leq \Gamma \left(\lambda_1^t y_1 + \sum_{s=2}^{t-1} \lambda_s^t (K\frac{\log s}{\sqrt{s}}) \right) + K(1-\delta)\frac{\log t}{\sqrt{t}} && (4.4) \\ &\leq \Gamma \left((1-\delta)\frac{\log t}{\sqrt{t}} K + \frac{\delta^2}{\Gamma}\frac{\log t}{\sqrt{t}} K \right) + K(1-\delta)\frac{\log t}{\sqrt{t}} && (4.5) \\ &\leq \delta K\frac{\log t}{\sqrt{t}} + K(1-\delta)\frac{\log t}{\sqrt{t}} = K\frac{\log t}{\sqrt{t}}, && (4.6) \end{aligned}$$

which yields (4.3) for $s = t$ to complete the induction.

Summing across time, we attain our desired result through the integral bound:

$$\begin{aligned} \max_{i \in I} \mathbb{E}_\theta^\sigma \left[(p_i^{NE} - p_{it})^2 \right] &= O(\frac{\log t}{\sqrt{t}}) \quad \forall \theta \in \Theta \\ &\Downarrow \\ \max_{\theta \in \Theta, i \in I} \sum_{t=1}^{T} \mathbb{E}_\theta^\sigma \left[(p_i^{NE} - p_{it})^2 \right] &= O(\sqrt{T} \log T). \end{aligned}$$

□

4.4 Managerial Implications

These results imply that an equilibrium among firms using a policy of best-response against forecast plus noise can achieve complete learning and attain the full-information Nash equilibrium revenues if the noise terms satisfy the conditions above. In numerical results, however, if firms do not add sufficient noise then incomplete learning may occur and different effects can occur, including the possibility that all firms earn greater revenues than under the full-information Nash equilibrium. An example appears in Figures 1 and 2. The red and blue curves in the figures correspond to the regions that are favorable to each player relative to the Nash equilibrium which occurs at the lower intersection of the two curves. The upper section corresponds to higher revenues for player 1 and the right section corresponds to higher revenues for product 2. The crosses correspond to the repeated actions of best responses with different initial observations or priors and varying amounts of noise. The yellow trajectory provides each player with a random and unbiased prior, the green trajectory has priors that give low-biased initial prices, and the purple trajectory has a prior that gives high-biased initial prices. Figure 1 shows a trajectory of 400 price pairs while Figure 2 shows only the last 50 price pairs to show where the trajectories are converging. As shown in Figure 2, the yellow prices are concentrated around the full-information Nash equilibrium while the green prices are favorable for Product 1 and unfavorable for Product 2 relative to the Nash equilibrium revenues. The purple prices are concentrated in an area that corresponds to higher revenues for both Products 1 and 2 relative to the Nash equilibrium revenues. In this

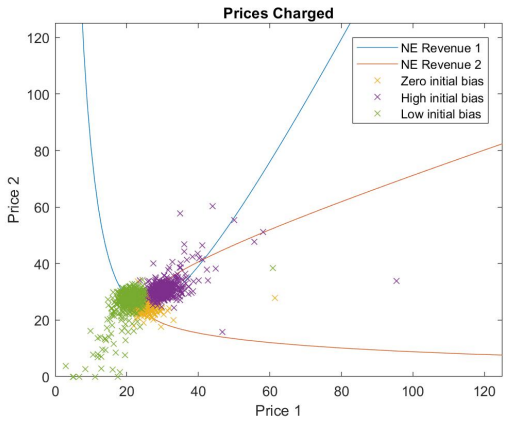

Figure 1: Incomplete learning price trajectories.

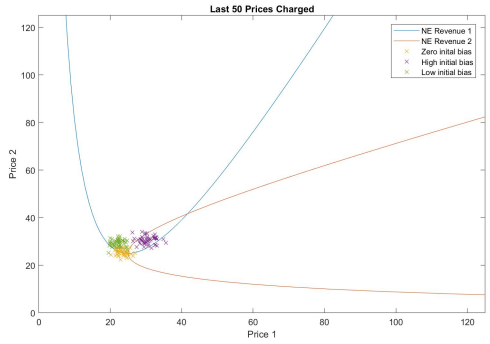

Figure 2: Incomplete learning price trajectories.

case, incomplete learning leads to better outcomes for both products than if both players had full information.

This type of behavior was also observed in the situation of unknown competition explored in Cooper et al. [2015]. The phenomenon suggests that it may be beneficial for firms not to experiment on prices if the information from such revelation can lead to competitors' learning (perhaps free-riding) and reduced revenues. To make the analysis of such a situation tractable, the next section introduces a model with simplified action and belief states but fully general policy structure in the form of a repeated prisoner's dilemma game which demonstrates that strategic lack of learning can occur in this setting, even with a finite time

horizon.

5 Model 2: Incomplete learning as an equilibrium strategy

This section addresses the issue of incomplete learning under competition by directly investigating equilibrium pricing policies in dynamic games of incomplete information. The model shows that incomplete learning of demand information is not merely a byproduct of the dynamic pricing strategies investigated in the first model, but is rather a rational outcome for firms competing in a Markov perfect equilibrium (MPE). In particular, we develop several simple demand environments, where the equilibrium strategies for the competing firms actively avoid learning the true value of demand and attain a collusive outcome even in finite time horizons. Hence, firms prefer to remain willfully ignorant of the marketplace for their products.

By simplifying the information assumptions of our first model, we present a dynamic game where the equilibria can be characterized using simple Markov strategies. Consider two firms that choose between two potential prices, the single period Nash equilibrium price and the cooperative equilibrium price. Initially, they do not know which price represents cooperation and which represents competition. In this respect, the competitive single period Nash equilibrium price and the cooperative equilibrium price form a prisoner's dilemma. Assuming a common Bayesian prior over the two possibilities, each firm decides whether or not to stick with the historically charged price or to experiment. Our results show that there exist conditions, parameterized by the value of the game's payoffs and prior beliefs, such that the firms will deliberately avoid experimenting with the prices in an effort to remain uncertain and to keep their opponent uncertain.

The significance of the results here are that firms can rationally choose not to experiment and learn the environment because their competitor also benefits from the information and can use that knowledge to the detriment of the experimenting agent. This threat of

information leakage can dominate for any finite time horizon, leading to maintenance of a cooperative equilibrium and continued uncertainty over the environment. This contrasts with the traditional literature in which opportunities for collusion are reduced in conditions with uncertainty (since deviations from cooperation are less likely to be detected and punished).

5.1 Model

Consider a stochastic game where two firms compete in a common market, with each selling a distinct product over a fixed T period selling season. The sequence of actions and outcomes in each period follow the same steps as in our first model. Firms begin each period with a belief about an uncertain demand environment and privately choose a price for their product from a fixed set of feasible prices. The firms then announce their chosen prices publicly and simultaneously, and in turn are given private realizations of demand from their customers. The distinguishing feature of this model, compared to our previous OLS approach, is that the set of possible prices for each firm is limited to a finite set of actions, instead of a closed interval, and the parameters for the underlying demand are restricted to a set of discrete demand "scenarios", rather than a compact set Θ.

Before this selling season begins, a state of the world is drawn from one of two possible market scenarios. For each state of the world, the profits to each firm are conditionally deterministic; that is, if a firm knew the true underlying market scenario, then that firm would know exactly what revenues would result from each possible pair of prices. In the first period, firms are unaware of the true market scenario and share a common prior π, which equals the probability that the underlying state is scenario 1. Specifically, each scenario corresponds to a different Prisoner's dilemma,

Scenario 1:

	a_1	a_2
a_1	X, X	T_1, S_1
a_2	S_1, T_1	R_1, R_1

Scenario 2:

	a_1	a_2
a_1	X, X	S_2, T_2
a_2	T_2, S_2	P_2, P_2

where $S_1 < X < R_1 < T_1$ and $S_2 < P_2 < X < T_2$.

At the start of each period, firms simultaneously choose between two prices: price a_1 and price a_2. When both firms elect to charge price a_1, the outcome is known to be a fixed value, X. This joint action reveals no new information and the game continues to the next period. However, if either of the firms choose to charge price a_2, then the true demand scenario is revealed to both firms after the revenues for that period are collected. For this reason, we refer to the pure-strategy of charging a_1 as *staying put* and the pure-strategy of charging a_2 as *experimenting*. The firms, therefore, have joint control of the information state of the game.

To solve for the Markov perfect equilibria of this game of uncertain information, we begin by considering the single period game, $T = 1$. Let p be the probability that firm 1, (the ROW player in the Prisoner's dilemma), chooses price a_1 and let q be the probability that firm -1, (the COL player) chooses price a_2. Then for a given q the best response for firm 1 is

$$\mathcal{BR}(q) = \arg\max_{p\in[0,1]} \begin{pmatrix} pq \\ p(1-q) \\ (1-p)q \\ (1-p)(1-q) \end{pmatrix}^{\top} \left[\pi \begin{pmatrix} X \\ T_1 \\ S_1 \\ R_1 \end{pmatrix} + (1-\pi) \begin{pmatrix} X \\ S_2 \\ T_2 \\ P_2 \end{pmatrix} \right],$$

with the left vector representing the probability of each joint action and the vector sum on the right representing the expected value of each action for firm 1. Through the following transformations, the best-response function is fully represented by a 2-dimensional state of game coordinates $x_\pi \in \Re$ and $y_\pi \in \Re$.

$$
\begin{aligned}
\mathcal{BR}(q) = \quad & \arg\max_{p\in[0,1]} \quad p \begin{pmatrix} q \\ (1-q) \\ -q \\ -(1-q) \end{pmatrix}^{\top} \left[\pi \begin{pmatrix} X \\ T_1 \\ S_1 \\ R_1 \end{pmatrix} + (1-\pi) \begin{pmatrix} X \\ S_2 \\ T_2 \\ P_2 \end{pmatrix} \right] + \alpha(\pi, q) \\
= \quad & \arg\max_{p\in[0,1]} p \quad \begin{pmatrix} q \\ (1-q) \end{pmatrix}^{\top} \left[\pi \begin{pmatrix} X - S_1 \\ T_1 - R_1 \end{pmatrix} + (1-\pi) \begin{pmatrix} X - T_2 \\ S_2 - P_2 \end{pmatrix} \right] + \alpha(\pi, q) \\
= \quad & \arg\max_{p\in[0,1]} p \quad \begin{pmatrix} q \\ (1-q) \end{pmatrix}^{\top} \begin{pmatrix} x_\pi \\ y_\pi \end{pmatrix} + \alpha(\pi, q),
\end{aligned}
$$

where

$$
\begin{aligned}
x_\pi &:= \pi(X - S_1) - (1-\pi)(T_2 - X) \\
y_\pi &:= \pi(T_1 - R_1) - (1-\pi)(P_2 - S_2) \\
\alpha(\pi, q) &:= q\left(\pi S_1 + (1-\pi)T_2\right) + (1-q)\left(\pi R_1 + (1-\pi)P_2\right),
\end{aligned}
$$

with x_π representing the expected gain to firm 1 (given π) of price a_1 when the opponent chooses to stay put ($q = 1$) and y_π representing the expected gain of staying put when the opponent experiments ($q = 0$).

Written in a simplified form,

$$
\mathcal{BR}(q) = \arg\max_p \alpha(\pi, q) + \beta(\pi, q)p \quad = \quad \begin{cases} 0, & \beta(q) \leq 0, \\ 1, & \beta(q) > 0, \end{cases}
$$

where $\beta(\pi, q)$ represents the expected payoff of charging a_1 and possibly remaining ignorant of demand, given the strategy of the opponent is q. Specifically, $\beta(\pi, q) = qx_\pi + (1-q)y_\pi$ is the q weighted convex combination of x_π and y_π. Firm 1's decision is then straightforward. If the expected payoff of charging a_1 is positive, then firm 1 charges a_1. If it is negative, then firm 1 experiments and chooses a_2. If it is zero, then firm 1 is indifferent between experimenting and staying put and could chose either action or adopt a mixed strategy.

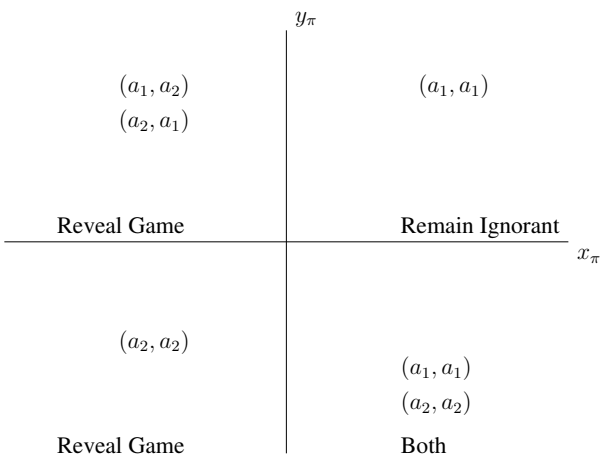

Figure 3: Pure-Strategy Nash EQ in Single-Period Game

The two-dimensional graph in Figure 3 illustrates the regions of the (x_π, y_π) graph and the pure-strategy equilibrium actions in each quadrant.

Each quadrant of the (x_π, y_π) represents a region where for a given belief π, the prescribed pure strategy actions are single-period Nash equilibria. For instance, if for a given π the values of x_π and y_π were both positive, then the joint action $(p^{NE}, q^{NE}) = (a_1, a_1)$ is a pure-strategy Nash equilibrium. Note that in quadrant II and quadrant IV of Figure 3, mixed strategy equilibria exist where both firms 1 and 2 are indifferent between learning the game and remaining uncertain.

Next, we extend the single-period stage game to a T-period dynamic game. We use a general setting of Markov perfect equilibrium as our solution concept. A Markov perfect equilibrium is a subgame perfect equilibrium where the firms are restricted to Markov strategies $\sigma(s) : S \to \mathcal{P}([0,1])^T$ where $S = \{0, \pi, 1\}$ denotes the set of three possible belief states for the firms. When $s = 1$, both firms are aware that scenario 1 represents the underlying payoff structure of the game and likewise, when $s = 0$, the game is known to be scenario 2. Let $\sigma_1(s) = (p_1(s), \ldots, p_T(s)) \in [0,1]^T$ and $\sigma_{-1} = (q_1(s), \ldots, q_T(s)) \in [0,1]^T$ denote the Markov strategies of firm 1 and -1, respectively, and the set Σ denote the set of all Markov strategies. The term p_t denotes the probability that firm 1 stays put in period t; q_t is analogously defined for firm -1. Given these strategies, the expected payoff to firm i for the

subgame beginning in period t is denoted $u_i(\sigma, \sigma_{-i}, t)$. A pair of strategies form a Markov perfect equilibrium when

$$u_i(\sigma_i, \sigma_{-i}, t) \geq \max_{\bar{\sigma}_i \in \Sigma} u_i(\bar{\sigma}_i, \sigma_{-i}, t), \quad i \in \{1, -1\}, \forall t \in \{1, \ldots, T\}.$$

The space of Markov strategies Σ can be reduced significantly, as there are unique equilibria for the states $s = 1$ and $s = 0$. If either of these states are reached in period t, by either firm charging a_2 in period $t-1$, then the only subgame perfect equilibrium strategy is for both firms to charge a_1 if $s = 1$ or a_2 if $s = 0$ for all periods t through T. This is due to the standard backwards induction argument for a finite horizon, Prisoner's dilemma game. Both firms know that their opponent has a dominant strategy, which is to choose these prices in the final period, and each has no incentive to cooperate in the period prior. However, if the game is uncertain in period t, we now show that the deterrent to price exploration enforces a more collusive outcome, due to the uncertainty between two Prisoner's dilemma games.

The expected payoffs of each firm for the T-period games are determined using backwards induction. Let $V_\pi^\sigma[t]$ denote the expected payoff of firm i from periods t through T given that the game has not yet been revealed (i.e., $s = \pi$) by period t. Let $v_0 = \pi X + (1-\pi)P_2$ denote the single-period payoff that firm i expects, given his beliefs, the revealed game to yield in the next period. Then the expected payoff for firm i given strategies σ are

$$\begin{aligned} V_\pi^\sigma[t] &= SG_\pi(0) + \mathbb{E}[u_i(\sigma, t+1)] \\ &= SG_\pi(0) + p_t q_t V_\pi^\sigma[t+1] + (1 - p_t q_t) v_0(T-t) \\ &= SG_\pi(0) + p_t q_t \left(V_\pi^\sigma[t+1] - v_0(T-t)\right) + v_0(T-t), \end{aligned}$$

where the value $SG_\pi(0)$ is equal to the expected payoff to firm i of the single-period stage game that was analyzed previously. Using the (x_π, y_π) graphical analysis to characterize equilibria strategies, let $SG(\Delta)$ represent the expected payoff to firm i of an equilibrium strategy to an augmented single-period game with $(x', y') = (x_\pi + \Delta, y_\pi)$ as the game coordinates. That is, plot the point (x', y') on the graph in Figure 3 and select a pure-strategy Nash equilibrium (p^{NE}, q^{NE}) from the appropriate quadrant. The value $SG(\Delta)$ is then the

expected value of this augmented game for firm i:

$$\begin{aligned} SG_\pi(\Delta) &:= \left\{u(p^{NE}, q^{NE}, T) : (x', y') = (x_\pi + \Delta, y_\pi)\right\} \\ &= \alpha(\pi, q^{NE}) + \left(q^{NE}x' + (1 - q^{NE})y'\right) p^{NE} \\ &= \alpha(\pi, q^{NE}) + \left(q^{NE}(x_\pi + \Delta) + (1 - q^{NE})y_\pi\right) p^{NE}. \end{aligned}$$

Since there exist multiple equilibria for certain quadrants in the (x_π, y_π) graph, the value of $SG_\pi(\Delta)$ is not well-defined without an explicit equilibria selection criterion. However, this formulation allows us to identify the Markov perfect equilibria to the dynamic game as follows. Consider the expected payoff equation for firm i introduced earlier:

$$\begin{aligned} V_\pi^\sigma[t] &= SG_\pi(0) + p_t q_t \left(V_\pi^\sigma[t+1] - v_0(T-t)\right) + v_0(T-t) \\ &= SG_\pi(0) + p_t q_t \Delta_{t+1} + v_0(T-t) \\ &= SG_\pi(\Delta_{t+1}) + v_0(T-t). \end{aligned}$$

The term $\Delta_{t+1} := V_\pi^\sigma[t+1] - v_0(T-t)$ represents the expected future gains when both firms stay put in period t less the gains from playing a complete information game for the next $T-t$ periods. Hence, the multi-period game is solved by reducing the decision in each period t for each firm to a single-period game. The future value of either knowing the scenario information or remaining uncertain is incorporated appropriately into this augmented game. Using this technique, we identify the following payoff and belief conditions where the strategy of $(p_t, q_t) = (1, 1)$ forms a Markov perfect equilibrium, and both firms choose to stay put for the entire $T-$period game. Inefficient learning is then an equilibrium outcome for competing firms under these conditions.

Consider the case where $x_\pi > 0$; that is, the expected payoff of charging a_1 is positive when firm i knows that their opponent will also charge a_1 in a single-period game. In this case, firms would choose to remain ignorant of the underlying game in the terminal period T, provided that firms reached that period without changing their prices. Iterating backwards, we show that the added value of remaining ignorant persists in each period and as a result $\Delta_t > \Delta_{t+1}$ for all t and firms choose (a_1, a_1) in all periods. This process is illustrated in

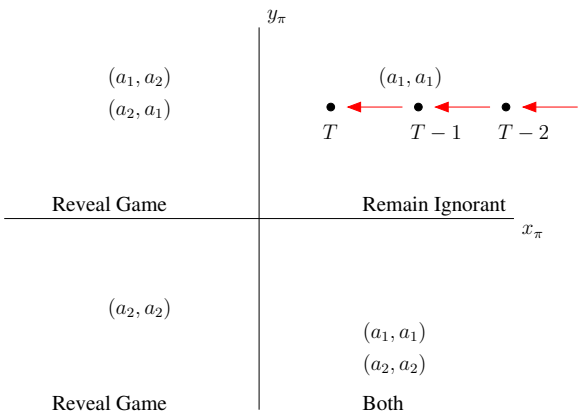

Figure 4: Inefficient Learning MPE in the Dynamic Game

Figure 4. Essentially, the analysis shows that if firms are willing to remain ignorant in the terminal period, they will also be willing to remain ignorant over an arbitrarily long, finite horizon.

Proposition 5.1 (Equilibrium with Incomplete Learning)**.** If the set of payoff parameters and beliefs are such that $x_\pi > 0$, then the strategy $(p_t, q_t) = (1, 1)$ for all $t \in \{1, \ldots, T\}$ is a Markov perfect equilibrium.

Proof. First recall from the analysis in Figure 3 that when $x_\pi > 0$ the subgame beginning in period T has a Nash equilibrium $(p_T, q_T) = (1, 1)$. Hence, if the firms were to choose the action (a_1, a_1) for all periods $t < T$ and remain ignorant of the underlying game payoffs, they would continue to choose (a_1, a_1) in the terminal period. Continuing with this graphical approach, we next show that $\Delta_t \geq 0$ for $t < T$ and thus $x' \geq x_\pi$ for these periods as well. The backwards induction argument proceeds as follows. First, consider the terminal case where $\Delta_T = \alpha(\pi, 1) + x_\pi - v_0$. Substituting the definitions of α and x_π,

$$\begin{aligned} \Delta_T &= \alpha(\pi, 1) + x_\pi - v_0 \\ &= (1 - \pi)(X - P_2) > 0. \end{aligned}$$

Next, assume for a given $t < T$ that $(p_\tau, q_\tau) = (1, 1)$ and $\Delta_\tau \geq 0$ for $\tau > t$. Since $\Delta_{t+1} \geq 0$ the game coordinates $(x_\pi + \Delta_{t+1}, y_\pi)$ lie in quadrant I or quadrant IV of the single-period

equilibrium graph. Therefore, the action $(p_t, q_t) = (1, 1)$ is an equilibrium action for period t. It remains to show that $\Delta_t \geq 0$ as follows.

$$\begin{aligned}\Delta_t &= V_\pi^\sigma[t] - v_0(T-t+1) \\ &= (\alpha(\pi,1) + x_\pi)(T-t+1) - v_0(T-t+1) \\ &= \Delta_T(T-t+1) > 0.\end{aligned}$$

□

5.2 Competition with Demand and Opponent Uncertainty

In the model of the previous section, firms receive the same private information in each round and are effectively aware of their competitor's information state. In this section, we introduce uncertainty into the payoff signals and, hence, distinct private information states. We show that, although firms can become aware of the game environment, they may still have an incentive to sustain the cooperative equilibrium.

Consider two games with the following payoff structure:

Scenario 1:

	a_1	a_2
a_1	X_1^t, X_{-1}^t	T_1, S_1
a_2	S_1, T_1	R_1, R_1

Scenario 2:

	a_1	a_2
a_1	X_1^t, X_{-1}^t	S_2, T_2
a_2	T_2, S_2	P_2, P_2

where

$$\begin{array}{ccccccccc} S_2 & < & P_2 & < & \left[\begin{array}{l}\text{supp}\, X_1^t \cup X_{-1}^t \\ \text{supp}\, X_1^t \cup X_{-1}^t\end{array}\right] & < & T_2 & & \\ & & S_1 & < & \left[\begin{array}{l}\text{supp}\, X_1^t \cup X_{-1}^t \\ \text{supp}\, X_1^t \cup X_{-1}^t\end{array}\right] & < & R_1 & < & T_1, \end{array}$$

for all periods $t \leq T$. The key distinction between these scenarios and the game considered in the previous section is that the rewards for playing action (a_1, a_1) are independent Bernoulli random variables. If the scenario is equal to 1 then,

$$X_i^t \sim \begin{cases} X_0, & \text{w.p. } \gamma \\ X_0 \pm \delta_1, & \text{w.p. } \frac{1}{2}(1-\gamma_t) \end{cases} \qquad i = \pm 1, t = 1, \ldots, T.$$

If the scenario is equal to 2 then,

$$X_i^t \sim \begin{cases} X_0, & \text{w.p. } \gamma \\ X_0 \pm \delta_2, & \text{w.p. } \frac{1}{2}(1-\gamma_t) \end{cases} \qquad i = \pm 1, t = 1, \ldots, T,$$

for $\delta_1 \neq \delta_2$. There is a $(1-\gamma_t)$ probability in each period that firm i will learn the underlying game scenario, even if both firms choose the typically uninformative action of pricing (a_1, a_1). Furthermore, firm i may learn the game, with a realization of $X_0 \pm \delta_1$ or $X_0 \pm \delta_2$, while its opponent realizes X_0 and remains ignorant. These scenarios generate a partially observable stochastic game, where firms face uncertainty over the game outcomes and uncertainty about their opponent's information state. Each firm i maintains a private state $s_{it} \in \{1, \ldots, 5\}$ for $t \in \{1, \ldots, T\}$. These states are defined as follows:

$$s_{it} = \begin{cases} 1 & \text{Game Revealed - Scenario 1} \\ 2 & \text{Game Revealed - Scenario 2} \\ 3 & \text{Game Not Revealed - all } X_0 \\ 4 & \text{Game Not Revealed - occurrence of } X_0 \pm \delta_1 \\ 5 & \text{Game Not Revealed - occurrence of } X_0 \pm \delta_2. \end{cases}$$

States 1 and 2 indicate that at some time $\tau < t$, one of the firms charged a_2 and thus both firms are aware of the underlying game scenario and are aware that their opponent also has this knowledge. The remaining states all correspond to the setting where both firms charged price a_1 for the past $t-1$ periods. Therefore, neither firm knows whether their opponent has seen a revealing outcome, $(X_0 \pm \delta_1)$ or $(X_0 \pm \delta_2)$, in any of the past periods. State 3 corresponds to the situation where firm i is unaware of the game scenario, and states 4 and 5 correspond to firm i's knowledge of scenario 1 or scenario 2 respectively.

Denote the strategy of firm i as $\sigma_{it}(s_i)$. To solve for a Markov perfect equilibrium, we assume that firm i knows its current state in each period and best-responds to a known strategy matrix $\sigma_{-it}(s)$ for $s \in \{1, \ldots, 5\}$ of its opponent. The main result of this section is that for a range of values for the common prior $\pi_0 = Pr(\text{Scenario } 1 | s_{i1} = 3)$, the horizon

length T, and the probability of uninformative outcomes γ_t, there exists Markov perfect equilibria where firms choose (a_1, a_1) in all periods. In other words, the theorem states that cooperation persists when the game has not been revealed, even when one firm or both firms has discovered the environment. Each firm's uncertainty over the private state of their opponent is enough to enforce cooperation throughout the game.

First, we update some of the notation used in the previous model to include the addition of opponent uncertainty. Let s_{1t} denote the vector $p_t \in [0,1]^5$ with each component representing the probability that firm 1 will choose price a_1 in period t and state s. Likewise, represent s_{-1t} as the vector $q_t \in [0,1]^5$. There are two belief sequences for firm 1. The belief matrix, $\mu^t_{s\bar{s}} = Pr(s_{-1t} = \bar{s} | s_{1t} = s)$, maintains firm 1's belief in period t of its opponents state. Additionally, let $\pi_t = Pr(\text{Scenario } 1 | s_{it} = 3)$ denote firm 1's belief that the underlying environment is scenario 1, provided this information is still unknown. Applying the same game coordinate system that was used in the previous model, define

$$\begin{aligned} x_{\pi_t} &:= \pi_t(X_0 - S_1) - (1-\pi_t)(T_2 - X_0) \\ y_{\pi_t} &:= \pi_t(T_1 - R_1) - (1-\pi_t)(P_2 - S_2). \end{aligned}$$

Incorporating these new factors into the analysis used in the previous section yields the following theorem.

Theorem 5.2. There exists a Markov perfect equilibrium strategy with $p_t = q_t = (1,0,1,1,1)^\top$ for $t \in \{1, \ldots, T-1\}$, and $p_T = q_T = (1,0,1,1,0)^\top$ provided that

i. $x_{\pi_0} > 0$;

ii. $\prod_{t=1}^{T-1} \gamma_t > \left(\frac{T_2 - X_0}{T_2 - P_2}\right)$.

Proof. Note that the only equilibrium action in state 1 is for both firms to choose (a_1, a_1) and similarly firms choose (a_2, a_2) in state 2. Each of these states generate finite horizon, complete-information games with dominant Nash equilibrium strategies. As a result, $p_{1t} = q_{1t} = 1$ and $p_{2t} = q_{2t} = 0$ are the only potential equilibrium actions for these revealed states.

The action (a_1, a_1) is also a dominant strategy for state 4. To see this, assume that firm 1 is in state 4 at period t and therefore knows that the underlying game is scenario 1, but

is unaware of its opponent's state. Given knowledge of the underlying game, firm 1 has no myopic incentive to charge a_2 since $S_1 < R_1 < x_0$. Firm 2 also has no information-state incentive to charge a_2 as the value of the realized game (state 1) is the same as the value of obscured game (state 4) provided $q_{1t} = q_{3t} = q_{4t} = 1$. Firm 1's best response is therefore to charge a_1 for the remaining horizon.

Assuming that firm -1 adopts q_t as its strategy, the best-response strategy for firm 1 while in states 3 and 5 is determined using the analysis of the previous section. In the terminal period T, the best-response for firm 1 in state 3 is

$$\begin{aligned} \mathcal{BR}_3(q_T) &= \arg\max_p \alpha(\pi_T, \mu_{3\cdot} q_T) + \beta(\pi_T, \mu_{3\cdot} q_T) p \\ &= \arg\max_p \left[(\mu_{33}^T + \mu_{34}^T) x_{\pi_T}\right] p. \end{aligned}$$

By assumption $x_{\pi_0} > 0$ and $q_{3t} = q_{4t} = q_{5t} = 1$ for $t < T$. Applying Bayes rule to the prior belief on scenarios, $\pi_t = Pr(Scenario = 1 | s_{1t} = 3)$ yields

$$\pi_{t+1} = \pi_t \frac{\mu_{4\cdot}^T q_t}{\mu_{3\cdot}^T q_t} = \pi_t,$$

and thus $\pi_T = \pi_0$. As a result, $(\mu_{33}^T + \mu_{34}^T) x_{\pi_0} > 0$ and the best-response for firm 1 in state 3 and period T is to charge the uninformative price a_1. If firm 1 were instead in state 5 in the terminal period T, the best-response to q_T is

$$\begin{aligned} \mathcal{BR}_5(q_T) &= \arg\max_p \alpha(0, \mu_{5\cdot} q_T) + \beta(0, \mu_{5\cdot} q_T) p \\ &= \arg\max_p \left[(\mu_{53}^T + \mu_{54}^T) x_0\right] p. \end{aligned}$$

By definition, $x_0 := -(T_2 - X_0) < 0$, and therefore firm 1 would choose price a_2 in the terminal period. To simplify the notation for the backwards induction, let

$$\begin{aligned} v_0 &:= \pi_0 V_{1,T} + (1 - \pi_0) V_{2,T} \\ \Delta_{3,t+1} &:= V_{3,t+1} - v_0 \\ \Delta_{5,t+1} &:= V_{5,t+1} - V_{2,t+1}. \end{aligned}$$

The terms $V_{s,t}$ represent the expected payoffs to firm 1 of the subgame starting in period t with information state s, provided that firm -1 chooses strategy q_t. Note that $\Delta_{3,T} =$

$(1-\pi_0)(X_0-P_2)>0$ is the single-period expected gain from being ignorant of both the underlying game and the opponent's information-state, and $\Delta_{5,T}=\mu_{53}^T(T_2-P_2)$ is the single-period expected gain from being ignorant of the opponent's information-state.

Iterating backwards, assume that both $\Delta_{3,t+1}$ and $\Delta_{5,t+1}$ are positive and consider the value function for state 3,

$$\begin{aligned} V_{3,t} &= \max_p \alpha(\pi_0,1)+\beta(\pi_0,1)p \\ &\quad + p\left(\gamma_t V_{3,t+1}+(1-\gamma_t)\left[\pi_0 V_{4,t+1}+(1-\pi_0)V_{5,t+1}\right]\right) \\ &\quad + (1-p)\left(\pi_0 V_{1,t+1}+(1-\pi_0)V_{2,t+1}\right). \end{aligned}$$

In words, the value of state 3 in period t is equal to the sum of myopic stage-game value $\alpha(\pi_0,1)+\beta(\pi_0,1)p$, the expected value of the subgame in period $t+1$ if firm 1 were to charge a_1, and the expected value of the subgame if firm 1 were to charge a_2. Simplifying the above equation,

$$V_{3,t} = \max_p \left(x_{\pi_0}+\gamma_t\Delta_{3,t+1}+(1-\gamma_t)(1-\pi_0)\Delta_{5,t+1}\right)p+\alpha(\pi_0,1)+v_0(T-t).$$

From condition (1) of the theorem, $x_{\pi_0}>0$ and by the inductive hypothesis $\Delta_{3,t+1}>0$ and $\Delta_{5,t+1}>0$. Therefore, the strategy $p_{3,t}=1$ is the best response to q_t. Next, consider the value function for state 5.

$$\begin{aligned} V_{5,t} &= \max_p \alpha(0,1)+\beta(0,1)p+pV_{5,t+1}+(1-p)V_{2,t+1} \\ &= \max_p \left(\beta(0,1)+\Delta_{5,t+1}\right)p+\alpha(0,1)+V_{2t+1} \\ &= \max_p \left(x_0+\Delta_{5,t+1}\right)p+\alpha(0,1)+V_{2t+1}. \end{aligned}$$

By the analysis above, $p_{5,t}=1$ is the best response strategy to q_t provided that $\Delta_{5,t+1}>-x_0$. Note that this condition holds in the terminal period by condition (2) of the theorem.

$$\begin{aligned} \Delta_{5,T} &= \mu_{53}^T(T_2-P_2) \\ &= (T_2-P_2)\prod_{t=1}^{T-1}\gamma_t \\ &> T_2-x_0 \\ &> -x_0. \end{aligned}$$

It remains to show that $\Delta_{3,t} > 0$ and that $\Delta_{5,t} > -x_0$. Assume that this condition holds for $\tau \geq t$. Then,

$$\begin{aligned}
\Delta_{5,t} &= V_{5,t} - V_{2,t} \\
&= \Delta_{5,t+1} + x_0 + \alpha(0,1) - P_2 \\
&= \Delta_{5,t+1} + (X_0 - T_2) + T_2 - P_2 \\
&= \Delta_{5,t+1} + (X_0 - P_2) \\
&> \Delta_{5,t+1} \\
&> -x_0,
\end{aligned}$$

and

$$\begin{aligned}
\Delta_{3,t} &= V_{3,t} - (\pi_0 V_{1,t} + (1-\pi_0)V_{2,t}) \\
&= \gamma_t \Delta_{3,t+1} + (1-\gamma_t)(1-\pi_0)\Delta_{5,t+1} + x_{\pi_0} + \alpha(\pi_0, 1) - v_0 \\
&> x_{\pi_0} + \alpha(\pi_0, 1) - v_0 \\
&= X_0 - v_0 \\
&= X_0 - (\pi_0 X_0 + (1-\pi_0)P_2) > 0.
\end{aligned}$$

□

Overall, this demonstrates that for a finite range, an equilibrium strategy has a positive probability of not learning and leading to a collusive outcome for any finite time horizon.

6 Conclusion

This paper presents dynamic models of price competition with unknown demand. In the first part, we presented a model in which the players assume linear demand functions and can observe others' actions but not their payoffs. We showed that, in the class of strategies characterized by best responses plus noise, an equilibrium exists in which each player adjusts their noise to be sufficient for learning and decreasing to ensure convergence. We also gave an

example to show that it is possible, however, for other strategies to dominate this outcome for all players. The lack of information, however, makes it difficult for the players to coordinate on this outcome.

The second model considers relaxation of the policies to obtain an equilibrium in which learning does not occur and players jointly attain a cooperative outcome. To make this analyzable, the second model restricts the uncertainty to finite unknown states of the world and actions in a dynamic Prisoner's dilemma situation. When players' actions can reveal the state to their competitor and the payoffs overlap in a way that does not reveal the state with other actions, an equilibrium can result in which the players maintain non-revealing actions and achieve the cooperative outcome for any time horizon. This result can even occur after one agent becomes informed.

These results imply that firms in competition can achieve learning as efficiently as a monopolist firm, but this requires either restriction in strategies or information structure. With sufficiently diffuse information but rapid learning by all participants from variations in actions, equilibria can exist in which learning does not occur. This result suggests that limitations on actions (such as laws, for example, those for insurance products, that restrict price adjustments or impose price caps) and privacy restrictions can lead to inefficient outcomes (or to losses in consumer welfare).

This work suggests several potential follow-on studies. Further theoretical studies could consider general conditions for the non-learning phenomenon (or for its non-existence) in a broader game context. Mechanisms to reduce the possibilities of inefficient outcomes, such as the presence of a monitor or intermediary, could also be considered. For empirical extensions, as mentioned, pharmaceutical examples could be studied for the observed phenomenon. In addition, the effects of price restrictions as in insurance markets might also provide useful empirical research.

References

D. Abreu, D. Pearce, and E. Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society*, pages 1041–1063, 1990.

D. Bertsimas and G. Perakis. *Dynamic pricing: A learning approach*. Springer, 2006.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

L.E. Blume and D. Easley. What has the rational learning literature taught us? In A. Kirman and M. Salmon, editors, *Learning and rationality in economics*. B. Blackwell, 1995.

J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Wang Chi Cheung, David Simchi-Levi, and He Wang. Dynamic pricing and demand learning with limited price experimentation. *Available at SSRN 2457296*, 2015.

W.L. Cooper, T. Homem-de Mello, and A.J. Kleywegt. Learning and pricing with models that do not explicitly incorporate competition. *Operations Research*, 63(1):86–103, 2015.

A.V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783, 2013.

E.J. Green and R.H. Porter. Noncooperative collusion under imperfect price information. *Econometrica: Journal of the Econometric Society*, pages 87–100, 1984.

J.M. Harrison, N.B. Keskin, and A. Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

N. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62:1142–1167, 2014.

David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

T.L. Lai and H. Robbins. Iterated least squares in multiperiod control. *Advances in Applied Mathematics*, 3(1):50–73, 1982.

M.S. Lobo and S. Boyd. Pricing and learning with uncertain demand. In *INFORMS Revenue Management Conference*, 2003.

Eric Maskin and Jean Tirole. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica*, 56(3):571–599, 1988.

A. McLennan. Price dispersion and incomplete learning in the long run. *Journal of Economic dynamics and control*, 7(3):331–347, 1984.

D. Rahman. The dilemma of the cypress and the oak tree. Technical report, Discussion paper, 2014.

M. Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

Yuliy Sannikov and Andrzej Skrzypacz. Impossibility of collusion under imprerfect monitoring with flexible production. *American Economic Review*, pages 1794–1823, 2007.

C. Simon. *Dynamic pricing with demand learning under competition.* PhD thesis, Massachusetts Institute of Technology, 2007.

G.J. Stigler. A theory of oligopoly. *The Journal of Political Economy*, pages 44–61, 1964.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, 2011.

M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pages 1729–1736, 2007.