# Algorithmic pricing with strategic consumers

ANDREAS BECH

USC

January 29, 2026

ABSTRACT. Abstract

KEYWORDS. Pricing, monopolist, reinforcement learning.

CONTACT. EMAILS. Thanks to.....

## 1. INTRODUCTION

Machine learning/AI is widespread in business operations, pricing, recommendation, fraud detection, etc. In this paper I on focus algorithmic pricing using so-called learning algorithms.

The problem is that firms are using algorithmic pricing but rational consumers might anticipate prices and adapt to algorithms. Little is known about how common algorithms, which are single-player algorithms, optimized on an exogenous environment, perform with strategic adversaries. An interaction between algorithm and best responding consumer is potentially complex and untractable.

## 2. LITERATURE

This work builds on the extensive literature regarding dynamic pricing with learning. This field generally studies how a firm can learn the optimal price when the demand curve is initially unknown. A comprehensive survey of this literature is provided by Den Boer (2015), who categorizes various approaches to the exploration-exploitation trade-off. While earlier works focused on parametric demand, recent advances have leveraged machine learning and bandit algorithms in more complex environments. For instance, Misra, Schwartz, and Abernethy (2019) develop mechanisms for dynamic online pricing that account for unknown demand parameters. However, the standard assumption in this literature is that consumers are non-strategic, meaning their purchasing decisions in period $t$ depend only on the price in period $t$, independent of the firm's history or future expectations.

Second, this paper is motivated by the recent surge of interest in algorithmic collusion. This body of work examines how independent pricing algorithms, specifically those based on Q-learning, can learn to sustain supra-competitive prices without explicit communication. Calvano et al. (2020) show that reinforcement learning algorithms consistently converge to collusive strategies in repeated oligopoly games. Similarly, Klein (2021) and Hansen, Misra, and Pai (2021) provide evidence that such algorithmic cooperation is robust to various environmental factors. While these papers focus on the interaction between two algorithms, this work examines the interaction between a single algorithm and a strategic human agent, asking if the algorithm can be similarly ”manipulated” or exploited by a rational adversary.

Third, we incorporate the classical literature on strategic consumers in operations management and economics. Su (2007) provides a seminal analysis of dynamic pricing when consumers are strategic (patient) and differ in their valuations and patience levels. This literature establishes that when consumers anticipate future price drops, firms must adjust their pricing paths (often leading to the well-known Coase conjecture dynamics).

Our contribution lies at the intersection of these fields. While Su (2007) assumes the firm is rational and knows the demand curve, and the bandit literature (Den Boer, 2015) assumes the firm learns but consumers are myopic, we examine the gap where the firm is learning (using standard algorithms like UCB1 or Q-learning) and consumers are strategic. We investigate whether the ”regret” bounds guaranteed in the first stream of literature hold up when the stationarity assumptions are broken by the strategic behavior modeled in the third stream.

## 3. MODEL

The game proceeds over $t = 1, 2, \ldots, T$, where a monopolists sets a price $p_t$ each period in order to maximize average profit. In each period a new consumer is drawn according to $v_t \sim F$ and lives for at most two periods. If the consumer buys in the first period they exit, otherwise they are alive in the second period.

The consumer is assumed to discount the future by $\beta$, and given a belief about the future price, $p^e_{t+1}$, the consumer buys in the first period if

$$v_t - p_t \geq \max\{0, \beta(v_t - p^e_{t+1})\}$$

$F$ is assumed to be unknown to the monopolist such that it uses a pricing policy $\Psi$ that learns from past experience: $p_t = \Psi\left(\{p_\tau, \pi_\tau \mid \tau = 1, \ldots, t-1\}\right)$. The consumer forms beliefs about the future price based on past prices.

### 3.1. Regret

Let the prices the firm can choose be from a grid of size $K$. The regret of a policy, $\Psi$, through time $t$ is

$$\begin{aligned}
\text{Regret}\left(\Psi, t\right) &= \mathbb{E}\left[\sum_{\tau=1}^{\text{t}} \pi^* - \pi_{p_\tau}\right] \\
&= \sum_{\tau=1}^{t} \left(\pi^* - p_\tau(1 - F(p_\tau)\right) \\
&= \pi^* t - \sum_{k=1}^{K} p_k(1 - F(p_k))\mathbb{E}\left[n_{kt}\right]
\end{aligned}$$

It is the ex-post regret from not choosing the constant, optimal price $p^*$

### 3.2. Multi-Armed Bandits and the UCB1 Algorithm

The so-called multi-armed bandits problem is a sequential decision-making problem where an agents has to choose one of multiple options (arms) in each period. Each arm yields a reward with an unknown distribution. The goal to maximize cumulative reward. When choosing among the arms the agent faces the famous exploration vs. exploitation tradeoff. One of the most common algorithms for balancing this tradeoff is the **UCB1** algorithm: At time $t$ select arm (price) $k$ that maximizes:

$$\bar{\pi}_k + \sqrt{\frac{2 \ln t}{n_{kt}}}$$

where $\bar{\pi}_k$: average profit for arm $k$, $t$: total plays, $n_{kt}$: plays for arm $k$ at time $t$.

### 3.3. Q-learning

Another framework for sequential learning is reinforcement learning. Reinforcement learning assumes the agent faces an unknown Markov decision process where rewards in each period depends on the state, $s$, and the action, a. One of the simpler, and very popular, algorithms in this framework is Q-learning. Q-learning aims to learn a value function $Q(s, a)$ through interacting with the Markov process. It is model-free, meaning that it does not know the transition probabilities.

The goal is again to maximize the expected cumulative reward and this is done by estimating the expected value of taking action $a$ in state $s$, $Q(s, a) = \mathbb{E}[R_t + \gamma \max_{a'} Q(s', a')]$. This is done by learning from experience using the update rule

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left( r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)$$

, where $\alpha$ is the learning rate, $\gamma$ is the discount factor, $r$ is the reward, and $s'$ is next state.

## 4. BENCHMARK REGRET BOUND

### 4.1. Single period consumers

When consumers only live for one period (or equivalently $\beta = 0$), the strategic link between periods is severed. The consumer's purchasing decision simplifies to the static condition: buy if $v_t \geq p_t$. Consequently, the consumer does not react to the firm's pricing history or anticipate future prices.

In this case, the demand and profit associated with any chosen price $p_k$ are independent and identically distributed (i.i.d.), governed solely by the distribution $F$. This perfectly maps the pricing problem to the standard stochastic Multi-Armed Bandit framework described in Section 3.2, where the environment is exogenous and stationary. The learning algorithm only needs to balance exploration and exploitation to find the optimal monopoly price $p^*$. The main result from the bandit literature (e.g., using the UCB1 algorithm) is that the cumulative Regret grows logarithmically:

$$\text{Regret}\,(\Psi, T) = \mathcal{O}(\log T).$$

This sublinear growth implies that the average regret per period converges to zero as $T \to \infty$, meaning the algorithm successfully learns the optimal price. See Auer et al. (2002) for details.

### 4.2. Two period game

In a simplified version of the game where $T = 2$ and only a single consumer, the solution can be formulated in a game theory framework. Assume that it is a two period game with no commitment and that $F$ is Uniform(0, 1). Both consumer and monopolist discounts the future with discount factor $\beta$. This model captures a key strategic tension: the monopolist wants to sell at a high price today, but the consumer knows that if they wait, the monopolist will likely lower the price tomorrow to capture remaining demand.

The equilibrium prices are (see Appendix):

$$p_1 = \frac{(2-\beta)^2}{2(4-3\beta)} < 1/2, \quad p_2 = \frac{(2-\beta)}{2(4-3\beta)} < 1/2$$

Figure 1 (a) below shows the prices as a function of the discount factor. Prices are always below $1/2$ and approach $1/2$ as the discount factor goes to $1$

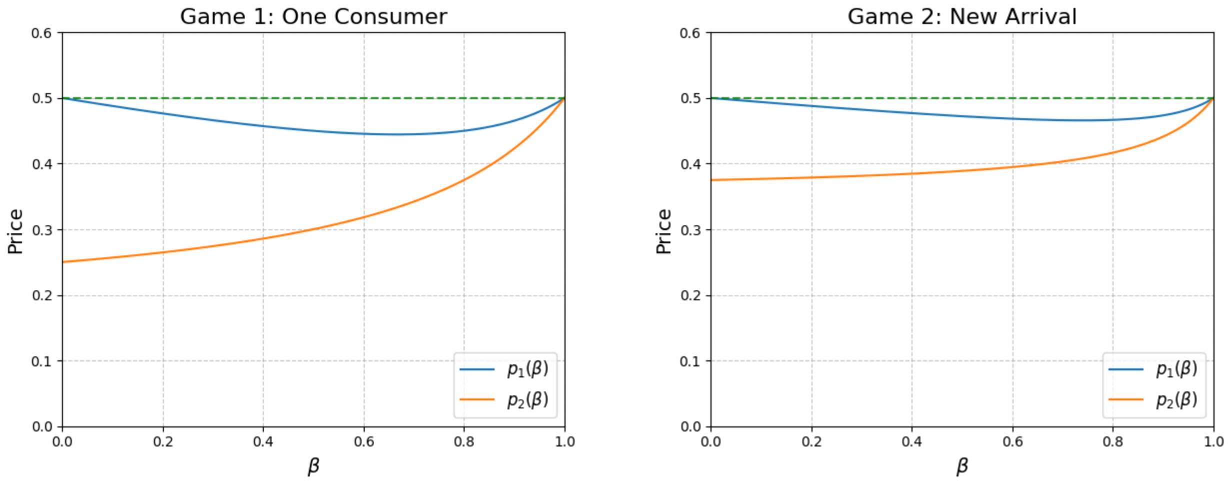
### 4.3. Two period game with two consumers

A slightly extended version of the previous game is when $T = 2$ but a second consumer enters in the second period. If the first consumer waits the monopolist has to set same price, $p_2$, for both old and new consumer. This modification mimics a stage game of the original model, where the monopolist now has a reason to keep the price high in the second period.

The equilibrium prices in this game are (see Appendix):

$$p_1^* = \frac{\beta^2 - 8\beta + 8}{2(8-7\beta)}, \quad p_2^* = \frac{6-5\beta}{2(8-7\beta)}$$

and are plotted in Figure 1 (b).



(a) One Consumer

(b) New Arrival

Figure 1: Equilibrium prices as a function of $\beta$

## 5. SIMULATIONS

### 5.1. UCB1

$F$ is assumed to be Uniform(0,1). Monopolist sets price using $p_t = \textbf{UCB1}\left(\{p_\tau, \pi_\tau \mid \tau = 1, \ldots, t-1\}\right)$. Consumer at time $t$ expects future price $p_{t+1}^e = \bar{p}_t$.

### 5.2. Thompson

Thompson sampling is alternative MAB algorithm.

### 5.3. Q-learning

The state is whether previous consumer is still alive, along with the previous price (20 states)

## 6. EXTENSIONS

## 7. DISCUSSION

- Find a proper nested stage game
- Prove sublinear regret (or lack thereof) for stylized problem

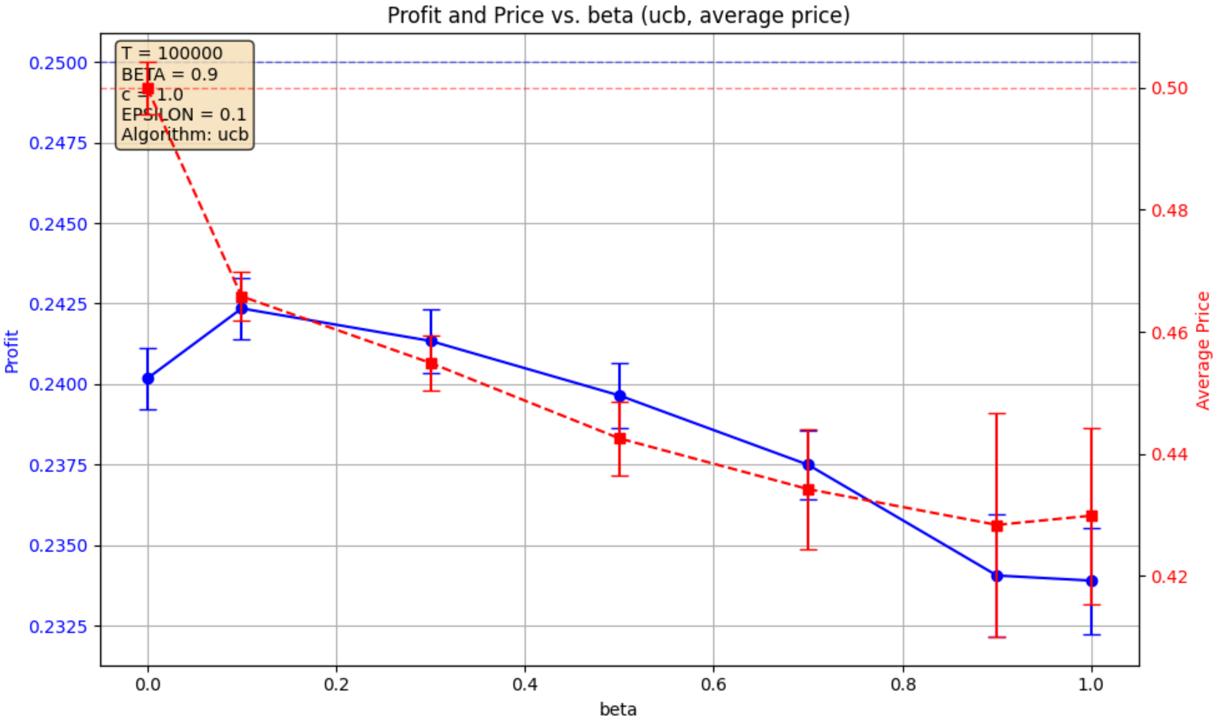


Figure 2: Simulation UCB1

## References

## 8. APPENDIX

### 8.1. Two period game

The game can be solved using a cutoff in the valuation $\hat{v}$, which we call the marginal consumer. Consumers with $v < \hat{v}$ buy the item in the first period at price $p_1$. In the second period, if the consumer waits, the monopolist infers the demand curve $v \sim U(0, \hat{v})$ and sets price $p_2(\hat{v}) = \hat{v}/2$.

The indifference condition for the consumer is $\hat{v} - p_1 = \beta(\hat{v} - p_2)$. Substitute $p_2(\hat{v})$ to get $p_1$ as a function of $\hat{v}$ (choosing $p_1$ is equivalent to chooses $\hat{v}$):

$$p_1(\hat{v}) = \hat{v}\left(1 - \frac{\beta}{2}\right)$$

The monopolist chooses $\hat{v}$ in order to maximize total discounted profit

$$\Pi(\hat{v}) = \underbrace{p_1(\hat{v})(1 - \hat{v})}_{\text{period 1 profit}} + \beta \underbrace{\left[p_2(\hat{v}) \cdot (\hat{v} - p_2(\hat{v}))\right]}_{\text{period 2 profit}}$$

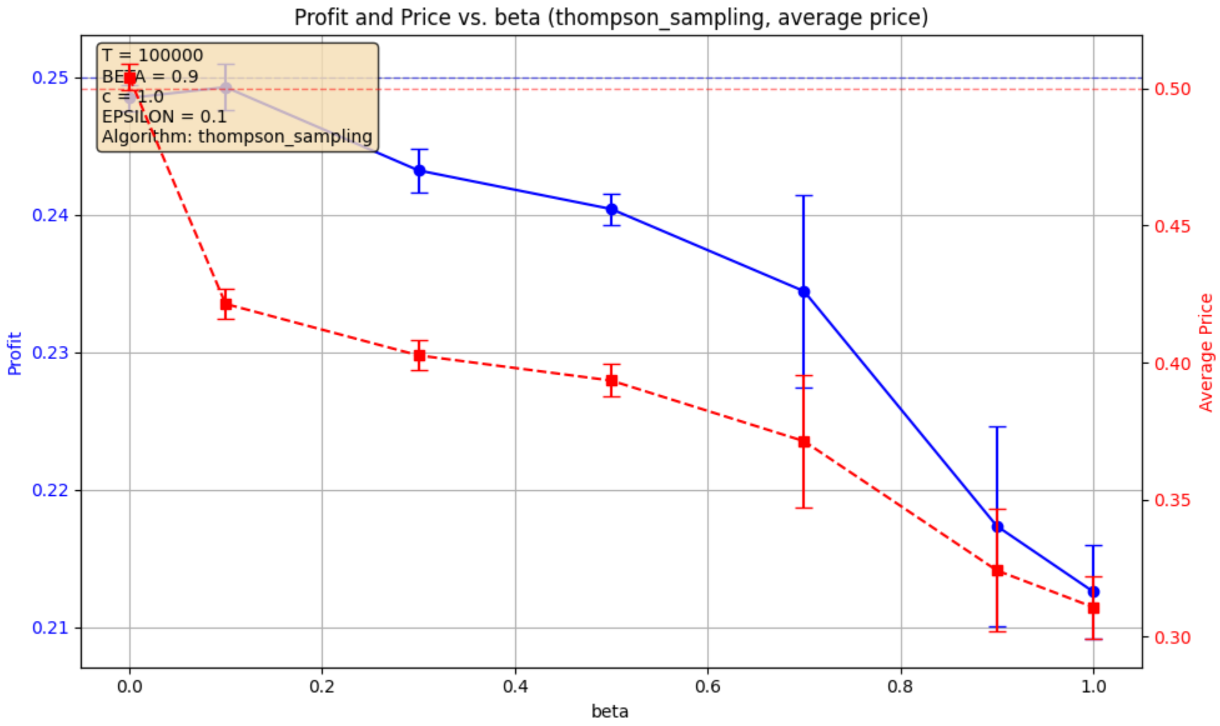


Figure 3: Simulation Thompson

Substitute and simplify to get

$$\Pi(\hat{v}) = \left(1 - \frac{\beta}{2}\right)\hat{v} - \left(1 - \frac{3\beta}{4}\right)\hat{v}^2$$

and from FOC

$$\hat{v}^* = \frac{1 - \beta/2}{2(1 - 3\beta/4)} = \frac{2 - \beta}{4 - 3\beta}$$

Finally the prices, from plugging back in, are:

$$p_1^* = \frac{(2-\beta)^2}{2(4-3\beta)}, \quad p_2^* = \frac{2-\beta}{2(4-3\beta)}$$

### 8.2. Two period game with two consumers

In period 2, a new consumer enters with valuation $v \sim U[0,1]$. The residual demand from the first consumer consists of those with $v \in [0, \hat{v})$. The monopolist sets a single price $p_2$. The total demand in period 2 is the sum of the residual demand and the new entrant demand: $Q_2(p_2) = (\hat{v} - p_2) + (1 - p_2) = 1 + \hat{v} - 2p_2$, and thus sets price $p_2(\hat{v}) = \frac{1+\hat{v}}{4}$ and derives profit

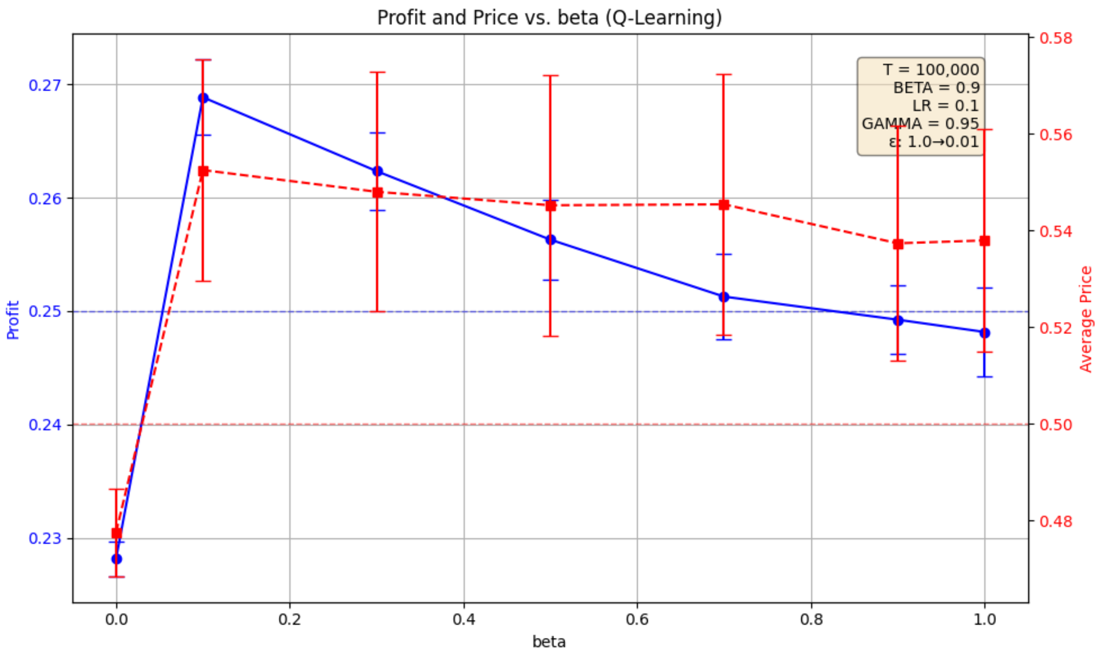


Figure 4: Simulation Q-learning

$\pi_2(\hat{v}) = (1+\hat{v})^2/8$.

The marginal consumer satisfies $\hat{v} - p_1 = \beta(\hat{v} - p_2)$. Substituting gives $p_1$ as a function of $\hat{v}$:

$$p_1(\hat{v}) = \frac{\beta}{4} + \hat{v}\left(1 - \frac{3\beta}{4}\right)$$

The monopolist chooses $\hat{v}$ in order to maximize total discounted profit

$$\Pi(\hat{v}) = \underbrace{p_1(\hat{v})(1-\hat{v})}_{\text{Period 1}} + \beta\,\underbrace{\pi_2(p_2^*)}_{\text{Period 2}}$$

$$\left[\frac{\beta}{4} + \hat{v}\left(1 - \frac{3\beta}{4}\right)\right](1-\hat{v}) + \beta\frac{(1+\hat{v})^2}{8}$$

and from the FOC

$$\hat{v}^* = \frac{1 - \frac{3\beta}{4}}{2\left(1 - \frac{7\beta}{8}\right)} = \frac{4-3\beta}{8-7\beta}$$

giving prices

$$p_1^* = \frac{\beta^2 - 8\beta + 8}{2(8-7\beta)}, \quad p_2^* = \frac{6-5\beta}{2(8-7\beta)}$$