Dynamic Pricing with Bayesian Updates from Online Reviews

Jose R. Correa*
Universidad de Chile
correa@uchile.cl

Mathieu Mari*
University of Warsaw
mathieu.mari@ens.fr

Andrew Xia*
MIT
axia@mit.edu

Abstract

When launching a new product, firms face uncertainty about market reception. Online reviews provide valuable information not only to consumers but also to firms, allowing them to adjust the product characteristics, including its selling price. In this paper we consider a pricing model with online reviews in which the quality of the product is uncertain and both the seller and the buyers Bayesianly update their belief to make purchasing or pricing decisions. Quite naturally, we connect the problem to a bandits problem. In this problem a player is unsure if a slot machine is *good* or *bad*. If the machine is good then, in each round, the player wins one dollar with probability p, whereas if it is bad the winning probability is q, where $q < p$. At each point in time, the only decision of the player is whether to play one more round at a cost of c (where $p > c > q$) or to stop playing forever, giving rise to an optional stopping problem. Initially the player has a prior x about the quality of the machine, i.e., she believes that with probability x the machine is good. We show a connection between this problem and the celebrated Catalan numbers, allowing us to efficiently compute the overall future discounted reward of the player. With this tool we can determine the optimal dynamic pricing strategy for the firm in the original pricing problem. We also discuss how this differs from the optimal static pricing, particularly in terms of the probability of effectively learning the quality of the product.

1 Introduction

Online decision making plays a crucial role in a variety of decision making settings, particularly related to the internet. Two landmark examples that have been widely studied are *dynamic pricing* and *online reviews*. Online review systems are powerful platforms for users to be informed about the product and for the firm to understand how a given market is receiving the product. The study of these systems has been vast for the last two decades [5, 8]. More recently, modeling simple like/dislike reviews as bandits problems has become standard [2, 3, 10, 12]. Dynamic pricing on the other hand, is an active area of research in economics, computer science and operations research, and has become a common practice in several industries such as transportation and retailing.

Very recently, there has been a growing interest in combining the two areas as a way to design more effective pricing mechanisms that gather information from current reviews to update prices and make the product more attractive [4, 9, 13]. Crapis et al consider social learning from likes/dislikes in a market with non-Bayesian agents, and the resulting pricing decision of a monopolist [4]. Shin et al consider a setting in which the volume of sales is large and optimize for revenue via fluid dynamics and ODEs [13]. Online reviews update the market's belief about the quality of the product and

*Equal Contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

thus influence product pricing. However the complexity of many models limit their practicality and already the Bayesian updating of beliefs becomes intractable.

In this paper, we continue on this line of research and consider a straightforward model that precisely determines the optimal pricing strategies using information from online reviews. We show that online reviews not only influence how successful a product is but also help to find more effective dynamic pricing strategies. These dynamic pricing strategies ultimately lead to more efficient allocations. Indeed, dynamic pricing with online reviews gives a foundation for the common practice of temporarily pricing a product below its production cost, leading to short term revenue losses. These losses come with a potential boost in future purchases, even at a higher price, and ultimately may lead to more sales.

The Pricing Problem. Consider a seller marketing a new product. As it is common in the literature (see e.g. [9]) we assume that the product may be either *good* or *bad*. A *good* product will be liked by a user with probability p (so that roughly a fraction p of the market is satisfied with the product), while a *bad* product will be liked by a user with probability $q < p$. Neither the seller nor the buyers are informed about the quality of the product but only receive a public signal x representing the prior probability of the product being good. The market is composed by an infinite stream of users arriving at times $t = 0, 1, 2, \ldots$, which are offered the product at a certain price π. Upon receiving the offer, a user (which we assume risk-neutral) evaluates whether his expected utility for buying the product. Initially, since the first buyer's prior is x, his expected utility can be evaluated as $xp + (1-x)q$, and thus, if this quantity exceeds the price he decides to buy. After buying, the user experiences the product and may like it or not, in both cases he submits an online review in the like/dislike format. These reviews allow following users to update their priors using Bayes' rule. More precisely, given a prior x, if the product receives one like we update the prior to $W(x)$ as follows:

$$W(x) := \mathbb{P}_x(\text{good} \,|\text{like}) = \frac{\mathbb{P}_x(\text{like}|\text{good})\mathbb{P}_x(\text{good})}{\mathbb{P}_x(\text{like})} = \frac{x \cdot p}{x \cdot p + (1-x) \cdot q}. \tag{1}$$

Similarly, given a dislike, we update the prior as follows:

$$L(x) := \mathbb{P}_x(\text{good} \,|\text{dislike}) = \frac{\mathbb{P}_x(\text{dislike} \,|\text{good})\mathbb{P}_x(\text{good})}{\mathbb{P}_x(\text{dislike})} = \frac{x \cdot (1-p)}{x \cdot (1-p) + (1-x) \cdot (1-q)}. \tag{2}$$

As we will see later, an interesting feature of our model is that the updated prior after a sequence of likes and dislikes only depends on the number of such likes and dislikes and not on the full sequence. This is in sharp contrast with most models of online reviews and it allows both the seller and the users to update their beliefs based solely on these figures.

A basic problem faced by the seller is thus to find an optimal pricing strategy. One alternative for the seller is to adopt a *static* price π. In this situation, users will buy the product so long as the current prior x satisfies $xp+(1-x)q \geq \pi$, yielding positive expected value, and whenever $xp+(1-x)q < \pi$ the process will stop forever. Thus, sales will continue until the prior reaches $x_{\min} = \frac{\pi - q}{p - q}$. Let c denote the per-unit cost to produce the product. Thus, given the prior x and the price π, the seller, who discounts the future at rate δ, can compute her expected revenue by computing the probability of stopping at time t, $\mathbb{P}_{x,\pi}(\tau = t)$, for all times t, including ∞.[2] This results in an expected revenue[3] of $\frac{\pi - c}{1-\delta}\left(\mathbb{P}_{x,\pi}(\tau = \infty) + \sum_{t=0}^{\infty}(1-\delta^t)\mathbb{P}_{x,\pi}(\tau = t)\right)$. Finally the seller will optimize this function over the values of $\pi \in (c, xp + (1-x)q]$.[4]

On the other hand, the seller may opt for a *dynamic* pricing approach. In this setting the price may be adjusted according to the current reviews the product has received. To maximize revenue the seller will just make the user indifferent whenever she decides to continue selling the product, offering the product at time t with prior x at exactly $\pi_t = xp + (1-x)q$. The decision of when to stop however becomes more involved. Although at times $xp + (1-x)q < c$ may hold, such that the seller will incur in a loss, it may still be worth to continue selling. The reason for this relies on the information gain provided by one more sale and the impact this information has on future purchases. In this paper we explicitly compute the prior value x^* that determines the stopping time of the seller under a

[2] As we will show, with positive probability the process will continue forever.

[3] We denote by $\mathbb{P}_{x,\pi}(\cdot)$ to the probability of an event given prior x and price π.

[4] Clearly π has to be at least c for the revenue to be positive. Also if $\pi > xp + (1-x)q$ then no user will ever buy and then the revenue is zero.

dynamic pricing strategy. We show a surprising connection between this value and an unexplored generalization of the celebrated Catalan numbers in combinatorics.

Our Results. In this paper we study the dynamic and static pricing with online reviews in detail. We start by formulating the underlying dynamic process as a simple multi-armed bandit problem and design a dynamic programming strategy for it. Then, we take a combinatorial approach using the Gittins index [7], and we explicitly determine the optimal underlying stopping time. For the latter, we uncover a connection with Catalan numbers that allow us to explicitly determine the stopping rule. This rule takes the form of a threshold prior such that the process continues so long as the current prior is at least this threshold. In the case of static pricing the threshold is simply given by $x_{\min} = \frac{\pi - q}{p - q}$. In the case of dynamic pricing the situation is more involved. The intuition is that even at the prior $\frac{c-q}{p-q}$, the seller may still be incentivized to sell in order to gain more information about the product. Thus, the threshold x^*,[5] which we compute explicitly, such that we are actually indifferent to continue, is smaller than $\frac{c-q}{p-q}$.

With this tool at hand, we go back to the pricing problem and pin down the optimal static and dynamic pricing strategies. Finally we study, in for each pricing strategy, the probability of achieving full efficiency. Note that this fully efficient situation occurs when the product is good and is sold forever[6], i.e., the product is good and the market learns this fact. To this end we exploit that the stochastic process governing the prior updates is a martingale and we can thus use the optional stopping theorem.

The Bandits Problem. We can model our dynamic pricing problem using a bandit framework. Imagine there is a single slot machine (bandit) which could either be a *good* or *bad* machine. This machine costs c to play and will yield a return of 1 (a *win*) with some probability and 0 (a *loss*) with some other probability.[7] If the machine is good it has a fixed, known probability of p of returning 1, while if it is bad it has known probability of q of returning 1. We have a prior x that the machine is good. Thus, given a prior x, the expected earning of a single pull is $xp + (1 - x)q - c$. Finally, we discount the future at rate δ, so that the value of earnings in time t is discounted by a factor of δ^t. As we play, we update our prior using Bayes rule and the problem we consider is that of determining the prior, x^*, under which we should stop playing.

The correspondence between the dynamic pricing problem and the bandits problem is straightforward. The slot machine corresponds to the product, which may be good or bad, and the prior for this is given by x. The cost c to play corresponds to the cost of the product. The slot machine returning 1 or 0 corresponds to a like or dislike in the online review, and the probabilities p and q have the same meaning in the problems. It should be noted that in the dynamic pricing problem the optimal price always corresponds to the expected utility of the next buyer $xp + (1 - x)q$ so long as the total discounted future revenue is positive. Thus the total expected reward in the bandits problem and the the revenue in the dynamic pricing problem coincide.

The only subtle apparent difference in the problems comes from the available information. In bandits problems we typically assume that we have the whole history of pulls whereas in the dynamic pricing problem we only want to assume that the users get to see the number of likes and dislikes the product has received so far. However, as we show next this is not an issue since the updated prior after a number of pulls only depends on the number of wins and losses and not on the sequence itself. Indeed, given a prior x, denote by $W(x)$ and $L(x)$ the updated priors after a win and a loss, respectively defined with equations (1) and (2). With these expression we can establish the important observation that the value of the prior after a given sequence of losses and wins is independent from the order, it only depends on the number of wins and losses. Furthermore, even the probability of the sequence only depends on the number of wins and losses. Denote $x_{w,\ell}$ as the value after seeing w wins and ℓ losses.

Lemma 1. *Given a prior x, the updated prior after a sequence of ℓ losses and w wins does not depend on the order and this value is $x_{w,\ell} = \frac{xp^w(1-p)^\ell}{xp^w(1-p)^\ell + (1-x)q^w(1-q)^\ell}$. Furthermore, the probability of each such sequence only depends on w and ℓ.*

Roadmap. We first describe two approaches to the bandit problem, with a dynamic programming approach in section 2 and a combinatorial approach providing a formal method to compute x^* and

[5] This x^* is a function of p, q, c, δ

[6] This is the best possible situation for the whole market, considering both the seller and the buyers.

[7] To avoid trivial cases we assume that $0 \leq q \leq c \leq p \leq 1$.

the expected long term reward in section 3. Finally, equipped with these tools we go back to the pricing problem and study their relative efficiency in section 4.

2 A Dynamic Programming Approach

In this section, we describe how to determine x^*, the lowest prior such that the player will continue playing the slot machine, via dynamic programming.

Let $V(x)$ denote the global expected reward from a prior x, having the ability to play on infinitely if so desired. We compute the Gittins index for our slot machine $V(x)$ as the discounted global reward, such that as long as $V(x)$ is positive we continue playing [7]. We compute $V(x)$ as a function of the expected local reward and the global reward from updated priors $W(x)$ and $L(x)$. Thus, we can formulate $V(x)$ recursively as follows:

$$V(x) = \max\Big(0, R(x) + \delta \cdot \mathbb{P}_x(\text{win}) \cdot V(W(x)) + \delta \cdot \mathbb{P}_x(\text{loss}) \cdot V(L(x))\Big) \tag{3}$$

where the local reward $R(x) = xp + (1-x)q - c$, the probability of winning is $\mathbb{P}_x(\text{win}) = xp + (1-x)q$, the probability of losing is $\mathbb{P}_x(\text{loss}) = x(1-p) + (1-x)(1-q)$, and $W(x)$ and $L(x)$ are the Bayesian updates given a win and a loss respectively. We define x^* as the largest prior x such that $V(x) = 0$.

In this section, we propose a fast dynamic programming approach to compute the threshold prior x^* and the global expected reward from any initial prior x. Let us denote the set of possible updated priors by $\mathcal{P}(x) := \{x_{w,\ell} \mid w, \ell \geq 0\}$. We say that a set $S \subseteq [0,1]$ is *discrete* if all elements of S has a neighbourhood that contains no other elements of S.

Claim 2. *The space of possible updated priors $\mathcal{P}(x)$ is discrete if and only if $\frac{\log(\frac{p}{q})}{\log(\frac{1-q}{1-p})} \in \mathbb{Q}$.*

Corollary 3. *If $\frac{\log(\frac{1-q}{1-p})}{\log(\frac{q}{p})} = \frac{a}{b}$ where a and b are two integers such that $gcd(a,b) = 1$, then there exists an infinite increasing sequence $(x_i)_{i\in\mathbb{Z}}$ such that for all $i \in \mathbb{Z}$ we have $W(x_i) = x_{i+a}$ and $L(x_i) = x_{i-b}$.*

In this setting, we can re-write equation (3) as

$$V(x_i) = \max\Big(0, R(x_i) + \delta \cdot \mathbb{P}_{x_i}(\text{win}) \cdot V(x_{i+a}) + \delta \cdot \mathbb{P}_{x_i}(\text{loss}) \cdot V(x_{i-b})\Big) \tag{4}$$

The algorithm. To apply a dynamic programming approach, let us fix an $\epsilon > 0$ that corresponds to the threshold between the precision of the solution returned and the running time: the smaller ϵ, the greater the precision and the running time. We note that this approach is not mathematically rigorous since its validity requires smoothness conditions on the function $V(\cdot)$ that do not follow from (3). In the next section we will derive a formal approach while now we continue with this approach that is computationally tractable. First, assuming that V is differentiable at $x = 1$ we get from equation (3) that $V(1) = \frac{p-c}{1-\delta}$ and $V'(1) = \frac{-(p-q)}{1-\delta}$. See Claim 8 in the Appendix for a proof.

For all indexes i such that $x_i > 1-\epsilon$, we make the estimation that $V(x_i) \approx V(1) + (1-x_i)V'(1) = \frac{p-c}{1-\delta} - (1-x_i)\frac{p-q}{1-\delta}$. Let i_{start} be the greatest index i such that $x_{i_{start}} < 1-\epsilon$. We have the following estimate: $i_{start} = O_x(a \cdot \log_{p/q}(1/\epsilon))$.

Then, for all $i \leq i_{start}$, we recursively compute $V(x_i)$, in decreasing order, with:

$$V(x_i) := \frac{V(x_{i+b}) - R(x_{i+b}) - \delta \cdot \mathbb{P}_{x_{i+b}}(\text{win}) V(x_{i+b+a})}{\delta \cdot \mathbb{P}_{x_{i+b}}(\text{loss})}$$

until we have $V(x_i) \leq 0$. We call this index i^* and then set $V(x_i) := 0$ for all $i \leq i^*$. The prior x_{i^*} gives an estimation of x^*. This method is fast but it is difficult to prove an upper bound on the precision of the results obtained. In the next section, we give a combinatorial method to compute x^* and the global expected reward, that enables us to provide a strong guarantee on the solution.

3 A Combinatorial Approach

In this section we intend to give a exact formulation of the value $V(x)$. To compute $V(x)$ we will make use of the concepts of *Catalan's triangle* and *Catalan's trapezoid*. Conceptually, *Catalan's*

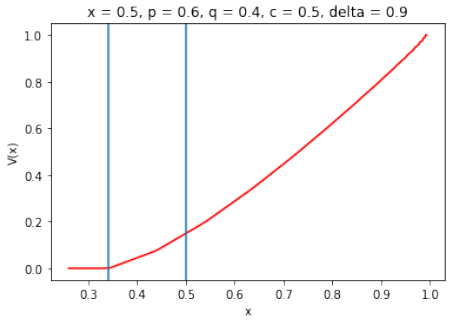

Figure 1: An example of computing $V(x)$ through our dynamic programming approach. We see that $x^* = 0.33$, which is lower than when $x = 0.5 = \frac{c-q}{p-q}$ causes the local reward to be 0.

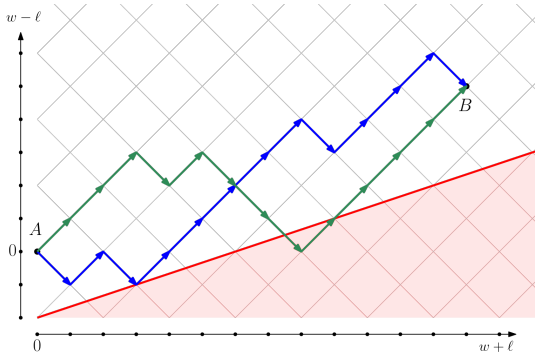

Figure 2: The number of paths from A to B along the grid, that do not enter the light red area (like the blue path but not like the green one) is the Catalan's quadrilateral number $C_3^{1,2}(9,4) = 570$ (for comparison, the unconditional number of paths from A to B is $\binom{9+4}{4} = 715$). More generally $C_m^{a,b}(w,\ell)$ is the number of paths starting on point $A = (0,0)$ that consists of moving w times "up" i.e. $(i,j) \to (i+1, j+1)$ and w times "down" i.e. $(i,j) \to (i+1, j-1)$, and that do not enter the open half-plane $\{(x,y) \mid (a-b)x - (a+b)y < -2m\}$. The slope of the boundary depends on parameters a, b. The classic Catalan's trapezoid ($a = b = 1$) arises when the boundary is horizontal.

triangle is a number triangle whose entries $C(w,l)$ corresponds to the number of strings with w "wins" and l "losses" such that no initial segment of the string has more losses than wins. *Catalan numbers* belongs to Catalan's triangle: $C(n,n) = \frac{1}{n+1}\binom{2n}{n}$.

Catalan's trapezoid is an extension of Catalan's triangle, in which $C_m(w,l)$ counts the number of strings with w wins and ℓ losses such that every initial segment has at least m more wins than losses[8]. In particular the Catalan's triangle corresponds to the special case where $m = 0$. We have the following closed form for the Catalan's trapezoid [11]: $C_m(w,\ell) = \binom{w+\ell}{\ell}$ if $0 \le \ell \le m$, $C_m(w,\ell) = \binom{w+\ell}{\ell} - \binom{w+\ell}{\ell-m-1}$ if $m < \ell \le w+m$ and $C_m(w,\ell) = 0$ otherwise.

Definition 4 (Catalan's quadrilateral)**.** *Given integers w, ℓ and parameters a, b and m we denote $C_m^{a,b}(w,\ell)$ the number of strings consisting of w W-s and ℓ L-s such that in every initial segments of the string that consist of w' W-s and ℓ' L-s, the value $a \cdot w' - b \cdot \ell'$ is always at least $-m$.*

Let us say that these numbers form *Catalan's quadrilateral* since the Catalan's trapezoid corresponds to $a = b = 1$. Figure 2 provides a geometrical interpretation of these numbers.

We have the following induction to compute these numbers: $C_m^{a,b}(w,\ell) = 0$ whenever $a \cdot w - b \cdot \ell < -m$. Otherwise $C_m^{a,b}(w,\ell) = 1$ when $w = 0$ or $\ell = 0$. And generally: $C_m^{a,b}(w,\ell) = C_m^{a,b}(w-1,\ell) + C_m^{a,b}(w,\ell-1)$.

[8] We use here a slightly different definition than in [11]. We have $C'_m(w,\ell) = C_{m-1}(w,\ell)$ where $C'_m(w,\ell)$ denote the original Catalan Trapezoid numbers.

Thus, computing $C_m^{a,b}(w,\ell)$ can be done in time $O(w \cdot \ell)$ with a simple dynamic program implementing the above inductive formulation. No non-recurrence-based formula exists for Catalan's quadrilaterals, though computation of partial values in the quadrilateral are derived in [6]. The purpose of defining these particular numbers lies in the following lemma that enables us to provide an expression of the global expected reward.

Given a initial prior x, we let X_t be the (random) updated prior after t pulls. In particular, $X_0 = x$. Recall that $x_{w,\ell}$ denotes the updated prior after a sequence of w wins and ℓ losses from an initial prior x, and $R(x) = xp + (1-x)q - c$ is the local reward when the prior is x. The global expected prior from a prior x is given by $V(x) = \sum_{w,\ell \geq 0, t=w+\ell} \delta^t \cdot \mathbb{P}_x(X_t = x_{w,\ell}) \cdot R(x_{w,\ell})$.

We now use the Catalan's quadrilateral to provide an expression of $\mathbb{P}_x(X_t = x_{w,\ell})$.

Lemma 5. *Let w, ℓ two integers, and $t = w + \ell$. Given any prior x, we have $\mathbb{P}_x(X_t = x_{w,\ell}) = C_m^{a,b}(w,\ell) \cdot p_x(w,\ell)$ where $p_x(w,\ell) := xp^w(1-p)^\ell + (1-x)q^w(1-q)^\ell$ is the probability of having a given ordered sequence of w wins and ℓ losses; $a = \log(p/q)$; $b = \log(\frac{1-q}{1-p})$; and $m = \log_{\frac{1-q}{1-p}}\left(\frac{x(1-x^*)}{x^*(1-x)}\right)$.*

To compute the value of $V(x)$, first we need to compute the value of the threshold x^*.

3.1 Computing x^*

By definition, x^* is the prior for which stopping or continuing to play gives the same global expected reward. Assuming that the initial prior is $x = x^*$, we have $m = 0$ and we obtain the following equation after simplification:

$$0 = V(x^*) = \sum_{w,\ell \geq 0} \delta^{w+\ell} C_0^{a,b}(w,\ell) \cdot \left(x^* p^w (p-c)(1-p)^\ell + (1-x^*)q^w(q-c)(1-q)^\ell\right)$$

where $a = \log(p/q)$ and $b = \log(\frac{1-q}{1-p})$. If we set $\Phi(p) := \sum_{w,\ell \geq 0} \delta^{w+\ell} \cdot C_0^{a,b}(w,\ell) \cdot (p-c)p^w(1-p)^\ell$, the above equation becomes $0 = x^*\Phi(p) + (1-x^*)\Phi(q)$. Thus we can express x^* as $x^* = \frac{\Phi(q)}{\Phi(q)-\Phi(p)}$.

To get a precise estimate of the value $\Phi(p)$, we only need to focus on sequences of wins and losses that do not exceed a certain length. More precisely, fix any $\epsilon > 0$. Since $\Phi(p)$ is defined as a series of positive terms, we know that there exists an integer t_ϵ such that

$$\sum_{w,\ell \geq 0, w+\ell \geq t_\epsilon} \delta^{w+\ell} \cdot C_0^{a,b}(w,\ell) \cdot (p-c)p^w(1-p)^\ell \leq \epsilon$$

and since this series is upper bounded by a convergent geometric series, we have the following estimate $t_\epsilon = O(\log 1/\epsilon)$. Thus, we can compute an ϵ-estimate $\widehat{x^*}$ of x^*, i.e. $|\widehat{x^*} - x^*| < \epsilon$, in time $O(\log(1/\epsilon)^2)$. When the ratio a/b is a rational number, the set of possible updated priors from x is discrete, so that choosing an ϵ sufficiently small enables to compute an exact value for x^*. Eventually, since the rational number are dense in $[0,1]$, we can slightly change p and q so that a/b is rational.

Once we have a precise estimation of x^*, we can proceed similarly to compute an arbitrarily closed estimation of $V(x)$ for any prior x.

In the *symmetric* case, when the values of p and q are such that $q = 1-p$, we can even get a closed expression for x^* and $V(x)$. Indeed, we have $a = b = 1$ and we can use the closed formula of the coefficients of the Catalan Trapezoid.

4 Pricing Strategies

In this section, we revisit our original seller's problem. When the seller has the option to price the product dynamically, the problem reduces to our slot machine scenario. Thus she maximizes her revenue by setting the dynamic price as $\pi = xp + (1-x)q$, where x is the current prior. In this case the buyer has negligible utility but is still willing to purchase the product. The seller is willing to set such price up until x reaches x^*, upon which the seller stops selling the product. In particular, she

may sell the product at certain times $\pi < c$ in order to gain more information about the product. Of course, any price $\pi < xp + (1-x)q$ also works although the surplus will then be split between the seller and the users; the important part, in term of social welfare, is that the process continues so long as the current prior is at least x^*. This dynamic pricing scenario maximizes the overall welfare of the system. We now discuss how to compute the optimal static price and then discuss exhibit some comparisons between these two pricing strategies.

4.1 Optimal static price

Given production cost c, the seller's goal is to determine the price π at which she wants to sell it, assuming that the price will be the same at all times. Of course, π must be at least equal to c and and should depend on p, q and the original prior x. On the other side, the buyer has access to the public prior x and buys if $xp + (1-x)q \geq \pi$. When the prior x is less than $x_{\min} := \frac{\pi - q}{p - q}$, the buyer stops buying. Let τ denote the random variable that corresponds to the first time t when x goes below $x_{\min}$.

First, we note that if the machine is used t times, then the profit achieved is $\sum_{i=0}^{t} \delta^i(\pi - c) = (\pi - c)\frac{1-\delta^t}{1-\delta}$. The expected revenue of the seller is then

$$Rev(\pi) := \mathbb{E}\left((\pi - c)\frac{1-\delta^\tau}{1-\delta}\right) = \frac{\pi - c}{1-\delta}\left(\mathbb{P}_{x,\pi}(\tau = \infty) + \sum_{t \geq 0} \mathbb{P}_{x,\pi}(\tau = t) \cdot (1 - \delta^t)\right).$$

We first prove that with constant probability the player will play on forever and give an expression of how often that happens. Given an initial prior x, and given a time $t \geq 0$ we define the (random) variable X_t that is the updated prior after t pulls. $(X_t)_{t \geq 0}$ is a martingale with $X_0 = x$. Let us fix $\epsilon > 0$. The random time τ at which X_t reaches $x_{\min}$ or $1 - \epsilon$ is a stopping time. Since τ has finite expectation, by the optional stopping theorem, the expected value of X_τ is equal to the initial prior, i.e., $\mathbb{E}(X_\tau) = x$. Then we get

$$x = \mathbb{P}_{x,\pi}(X_\tau < x_{\min})\mathbb{E}(X_\tau | X_\tau < x_{\min}) + \mathbb{P}_{x,\pi}(X_\tau > 1 - \epsilon)\mathbb{E}(X_\tau | X_\tau > 1 - \epsilon)$$

We know that $L(x_{\min}) \leq \mathbb{E}(X_\tau | X_\tau < x_{\min}) < x_{\min}$ and $1 - \epsilon < \mathbb{E}(X_\tau | X_\tau > 1 - \epsilon) \leq W(1 - \epsilon)$.

Thus, when ϵ goes to zero, we obtain $\mathbb{P}_{x,\pi}(X_\tau = \infty) = \frac{x - \mathbb{E}(X_\tau | X_\tau < x_{\min})}{1 - \mathbb{E}(X_\tau | X_\tau < x_{\min})}$.

Lemma 6. *We have the following estimation of playing forever:* $\frac{x - x_{\min}}{1 - x_{\min}} < \mathbb{P}_{x,\pi}(X_\tau = \infty) \leq \frac{x - L(x_{\min})}{1 - L(x_{\min})}$.

In the symmetric case when $q = 1 - p$, we can have a better estimation. Indeed, if $\tau = t$ then necessarily, $X_{t-1} = x_{\min}$ and we got a loss at time $t - 1$. Thus, $\mathbb{E}(X_\tau | X_\tau < x_{\min}) = L(x_{\min})$ so that $\mathbb{P}_{x,\pi}(X_\tau = \infty) = \frac{x - L(x_{\min})}{1 - L(x_{\min})}$.

Now, the probability of stopping at a finite time t can be expressed via the Catalan's quadrilateral, using Lemma 5. This is justified by the following Claim. Recall that X_t denotes the prior at time t and $x_{w,\ell}$ denotes the value of the updated prior from an initial prior after a sequence of w wins and ℓ losses.

Claim 7. *Given an initial prior x, and a cost c, the probability that the buyer stops buying at time $t \geq 1$ is $\mathbb{P}_{x,\pi}(\tau = t) = \mathbb{P}_x(X_t = x_{w,\ell}) = C_m^{a,b}(w, \ell) \cdot p_x(w, \ell)$ where $\ell = \lfloor \frac{m + at}{a + b} \rfloor$ and $w = t - \lfloor \frac{m + at}{a + b} \rfloor$; $a = \log(p/q)$ and $b = \log(\frac{1-q}{1-p})$; and $m = \log_{\frac{1-q}{1-p}}\left(\frac{x(1 - x_{\min})}{x_{\min}(1 - x)}\right)$; and $p_x(w, \ell) := xp^w(1-p)^\ell + (1-x)q^w(1-q)^\ell$.*

We now give an approximate expression of the expected revenue for the seller for a fixed price π. Let us fix an $\epsilon > 0$. There is a time $t_\epsilon = O(\log 1/\epsilon)$ for which $\mathbb{P}_{x,\pi}(\tau \geq t_\epsilon) \leq \mathbb{P}_{x,\pi}(\tau = \infty) + \epsilon \cdot \frac{1-\delta}{\pi - c}$. To get an estimate, we consider that if the buyers have not stopped buying at time t_ϵ, then they will continue to buy the product forever. Thus using the value for Catalan's quadrilateral, we can compute the following approximation of the expected revenue in time $O(t_\epsilon^2) = O((\log(1/\epsilon))^2)$.

$$Rev_\epsilon(\pi) := \frac{\pi - c}{1-\delta}\left(\mathbb{P}_{x,\pi}(\tau \geq t_\epsilon) + \sum_{t \leq t_\epsilon} \mathbb{P}_{x,\pi}(\tau = t) \cdot (1 - \delta^t)\right)$$

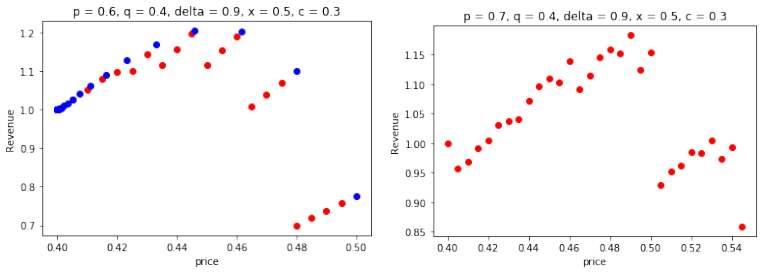

Figure 3: In the symmetric setting (left), the blue points represent the revenue on the efficient frontier, which is the maximum possible price per discrete x_{stop}. The red points are not optimal because such prices yield the same number of possible net dislikes before the buyers stop buying. In the general setting (right), we see that revenue as a function of price is not as well-defined.

and we have a good estimate: $|Rev_\epsilon(\pi) - Rev(\pi)| \leq \epsilon$. To compute the optimal price, i.e. the price that maximizes the revenue for the seller, one can compute $Rev_\epsilon(\pi)$ for different prices π and pick the best one. The set of values to try can be restricted to a discrete set of possible updated priors $\mathcal{P}(x)$. In the symmetric setting, we observe that for a set price π, buyers can tolerate up to a net of m losses, where m is a function of x, p, q, c, δ, π, before no longer buying. Thus, for a fixed m, we can maximize revenue $Rev_\epsilon(\pi)$ by setting the price π such that the buyer is indifferent to purchase after m net losses ($\pi = x_{0,m}p - (1 - x_{0,m})q$). This efficient frontier appears to be concave, as shown in figure 3, so finding the optimal price is simply a binary search procedure along the efficient frontier points. In the asymmetric setting, the function is not as well-defined.

4.2 Comparison of the two pricing strategies.

In this section we investigate the probability that the market learns the true value of the product, i.e. the probability of stopping in finite time when the product is bad and the probability of playing forever when the production is good. With dynamic pricing, learning occurs with larger probability. The main conclusion of this section is to express this additional gain as a function of the primitives of the model, proving a simple quantification of the potential gains of dynamic pricing over static pricing,

A first observation (see Lemma 9) is that in both models the market discovers that the machine is bad when the machine is actually bad. On the other hand, when the product is good, learning may fail to occur. To quantify this efficiency loss we define x_{stop} as the threshold prior from which the users stop buying. In the dynamic pricing model we have $x_{\text{stop}} = x^*$ and in the static price model, we have $x_{\text{stop}} = x_{\min} = \frac{\pi - q}{p - q}$. In particular we know that $x^* < c < x_{\min}$. Our main result (Lemma 10) establishes that when the product is good the probability of learning it is at least $\frac{x - x_{\text{stop}}}{x(1 - x_{\text{stop}})}$. Moreover, the latter holds with equality in the symmetric case.

With these results we can bound the ratio of not learning under the considered pricing strategies. This happens exactly when the machine is good but the market does not discover it and stops buying in finite time. Let FN_{static} and FN_{dynamic} be the probabilities of stopping when the product is good in the static and in the dynamic prices scenarios, respectively. In the case when $p = 1 - q$ we have $\frac{FN_{\text{static}}}{FN_{\text{dynamic}}} = \frac{1/x^* - 1}{1/x_{\min} - 1}$.

Acknowledgments and Disclosure of Funding

Work done while Andrew Xia and Mathieu Mari were visiting Universidad de Chile. Jose Correa was partially supported by ANID Chile through grant CMM-AFB 170001. Mathieu Mari was supported by École Normale Supérieure (Paris). Andrew Xia was funded by a Fulbright Research Scholarship.

References

[1] D. Austen-Smith, C. Martinelli. Optimal Exploration. *GMU Working Paper in Economics* No. 18-25

[2] O. Besbes, M. Scarsini. On information distortions in online ratings. *Operations Research* 66(3):597-892, 2018.

[3] T. Bonald, A. Proutiere. Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards. NIPS 2013.

[4] D. Crapis, B. Ifrach, C. Maglaras, M. Scarsini. Monopoly Pricing in the Presence of Social Learning. *Management Science* 63(11):3586–3608, 2017.

[5] K. Dave, S. Lawrence, D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. WWW 2003.

[6] Y. Fukukawa, Counting generalized Dyck paths. Manuscript 2013.

[7] J. Gittins, D. Jones, A Dynamic Allocation Index for the Discounted Multiarmed Bandit Problem, *Biometrika* 66(3):561–565, 1979.

[8] M. Hu, B. Liu, Mining and summarizing customer reviews. KDD 2004.

[9] B. Ifrach, C. Maglaras, M. Scarsini, A. Zseleva. Bayesian Social Learning from Consumer Reviews. *Operations Research* 67(5):1209-1221, 2019.

[10] F. Monachou, I. Ashlagi. Discrimination in Online Markets: Effects of Social Bias on Learning from Reviews and Policy Design. NeurIPS 2019.

[11] S. Reuveni. Catalan's Trapezoids. *Probability in the Engineering and Informational Sciences*, 28(3):353-361, 2014.

[12] V. Shah, R. Johari, and J. Blanchet. Semi-parametric dynamic contextual pricing. NeurIPS 2019.

[13] D. Shin, S. Vacarri, A. Zeevi. Dynamic Pricing with Online Reviews. Manuscript 2019.

Appendix

Proofs

Proof. (Lemma 1) Let us start by computing $L(W(x))$, the value of the updated prior after seeing a win then a loss:

$$\begin{aligned} L(W(x)) &= \frac{(1-p)W(x)}{(1-p)W(x)+(1-q)(1-W(x))} \\ &= \frac{(1-p)\frac{p\cdot x}{p\cdot x+q\cdot(1-x)}}{(1-p)\frac{p\cdot x}{p\cdot x+q\cdot(1-x)}+(1-q)(1-\frac{p\cdot x}{p\cdot x+q\cdot(1-x)})} \\ &= \frac{p(1-p)\cdot x}{p(1-p)\cdot x+q(1-q)\cdot(1-x)} \end{aligned}$$

Clearly, this expression when swapping p (resp. q) with $1-p$ (resp. $1-q$) is unchanged, thus $W(L(x)) = L(W(x))$. The proof of the formula for $x_{w,\ell}$, with w wins and ℓ losses, uses induction and works similarly.

To see the second part of the statement simply note that

$$\begin{aligned} \mathbb{P}_{W(x)}(\text{loss})\cdot\mathbb{P}_x(\text{win}) &= (W(x)(1-p)+(1-W(x))(1-q))\cdot(xp+(1-x)q) \\ &= \left(\frac{xp}{xp+(1-x)q}(1-p)+(1-\frac{xp}{xp+(1-x)q})(1-q)\right)\cdot(xp+(1-x)q) \\ &= xp(1-p)+(1-x)q(1-q). \end{aligned}$$

And we easily get the same expression for $\mathbb{P}_{L(x)}(\text{win})\cdot\mathbb{P}_x(\text{loss})$. The result follows by induction. □

Proof. (Claim 2) For our purpose we prove only the sufficiency and leave the proof of necessity to the interested reader.

Assume that $\frac{\log(\frac{p}{q})}{\log(\frac{1-q}{1-p})} = \frac{a}{b}$ where a and b are two integers such that $pcd(a,b) = 1$. This is equivalent to $\left(\frac{1-q}{1-p}\right)^a = \left(\frac{p}{q}\right)^b$. We show that there exists an infinite increasing sequence $(x_i)_{i\in\mathbb{Z}}$ such that $x_0 = x$ and for all $i \in \mathbb{Z}$ we have $W(x_i) = x_{i+a}$ and $L(x_i) = x_{i-b}$.

We first show that the updated prior after a sequence of b wins and a losses is unchanged, i.e. $x_{b,a} = x_{0,0}$. Indeed, by Lemma 1, we have $x_{b,a} = \frac{xp^b(1-p)^a}{xp^b(1-p)^a+(1-x)q^b(1-q)^a} = \frac{x}{x+(1-x)(\frac{q}{p})^b(\frac{1-q}{1-p})^a} = x$.

Then, for any $i \in \mathbb{Z}$, we can set $x_i := x_{w,\ell}$ where w, ℓ is any pair of integers such that $w \cdot a - \ell \cdot b = i$. The sequence (x_i) is strictly increasing because $aw - b\ell > aw' - b\ell'$ if and only if $(\frac{q}{p})^w(\frac{1-q}{1-p})^\ell > (\frac{q}{p})^{w'}(\frac{1-q}{1-p})^{\ell'}$, i.e. if and only if $x_{w,\ell} > x_{w',\ell'}$. □

Claim 8. *Assuming that solution V of equation (3) is differentiable in $x = 1$, we have $V(1) = \frac{p-c}{1-\delta}$ and $V'(1) = \frac{-(p-q)}{1-\delta}$.*

Proof. If $x = 1$, then the machine must be good, therefore the expected global reward is

$$V(1) = \sum_{t\geq 0}(p-c)\delta^t = \frac{p-c}{1-\delta}.$$

Now fix a small $\epsilon > 0$. Assume that the function $V(x)$ admits a derivative in $x = 1$, we can write $V(1-\epsilon) = V(1) + \epsilon V'(1) + o(\epsilon)$. With equation 3 we have for $x = 1-\epsilon$ close to 1:

$$\begin{aligned}
V(1-\epsilon) &= R(1-\epsilon) + \delta \cdot \mathbb{P}_x(\text{win}) \cdot V(W(1-\epsilon)) + \delta \cdot \mathbb{P}_x(loose) \cdot V(L(1-\epsilon)) \\
&= p - c - \epsilon(p-q) + \delta(p - \epsilon(p-q))(V(1 - \frac{q}{p}\epsilon + o(\epsilon))) \\
&\qquad + \delta((1-p) + \epsilon(p-q))(V(1 - \frac{1-q}{1-p}\epsilon + o(\epsilon))) \\
&= p - c - \epsilon(p-q) + \delta(p - \epsilon(p-q))(V(1) + \frac{q}{p}\epsilon V'(1) + o(\epsilon)) \\
&\qquad + \delta((1-p) + \epsilon(p-q))(V(1) + \frac{1-q}{1-p}\epsilon V'(1) + o(\epsilon)) \\
&= p - c + \delta V(1) + \epsilon\left(-(p-q) + \delta V'(1)\right) + o(\epsilon) = V(1) + \epsilon V'(1) + o(\epsilon)
\end{aligned}$$

Thus, $V'(1) = \frac{-(p-q)}{1-\delta}$. □

Proof. (Claim 7) Suppose we stop exactly at time t after a sequence of w wins and ℓ losses. First $w + \ell = t$. Moreover, $aw - b\ell < -m$ and $aw - b(\ell-1) \geq -m$. This implies that $\ell = \lfloor \frac{m+at}{a+b} \rfloor$ and $w = t - \ell$. Then we apply Lemma 5. □

Proof. (Lemma 5) By Lemma 1, the updated prior after a sequence of w wins and ℓ losses only depends on w and ℓ so as the probability of such each sequence. Therefore, $\mathbb{P}_x(X_t = x_{w,\ell})$ is the product of the probability of one sequence and the number of such sequences.

We first show that the probability of having a given ordered sequence of w wins and ℓ losses is $p_x(w,\ell) := xp^w(1-p)^\ell + (1-x)q^w(1-q)^\ell$. We proceed by induction on the length $w + \ell$ of the sequence $w + \ell$. The base case follows from Lemma 1. Using the induction hypothesis, we obtain

$$\begin{aligned}
p_x(w+1,\ell) &= p_{W(x)}(w,\ell) \cdot \mathbb{P}_x(\text{win}) \\
&= (W(x)p^w(1-p)^\ell + (1-W(x))q^w(1-q)^\ell) \cdot (xp + (1-x)q) \\
&= \left(\frac{xp}{xp+(1-x)q}p^w(1-p)^\ell + (1 - \frac{xp}{xp+(1-x)q})q^w(1-q)^\ell\right)(xp + (1-x)q) \\
&= xp^{w+1}(1-p)^\ell + (1-x)q^{w+1}(1-q)^\ell
\end{aligned}$$

The calculation for $p_x(w,\ell+1) = xp^w(1-p)^{\ell+1} + (1-x)q^w(1-q)^{\ell+1}$ works similarly.

Now for any integers w, ℓ, it is easy to see that $x_{w,\ell} < x^*$ if and only if $a \cdot w - b \cdot \ell < -m$ where $a = \log(p/q)$, $b = \log(\frac{1-q}{1-p})$ and $m = \log_{\frac{1-q}{1-p}}\left(\frac{x(1-x^*)}{x^*(1-x)}\right)$. Thus, the number of sequences of w wins and ℓ losses such that the prior at any time is at least x^* equals the Catalan's quadrilateral $C_m^{a,b}(w,\ell)$. □

Comparison of the pricing strategies

Lemma 9. *Assuming that the product is bad, we stop in finite time almost surely.*

Proof. Let $D_t = aw_t - b\ell_t$ denote the random variable that corresponds to the weighted difference between the number of wins and loses after t pulls, where $a = \log(p/q)$ and $b = \log((1-q)/(1-p))$. We define the stopping time τ as the first time t when $D_t < -m$ where m depends on the original prior x and the threshold prior x_{stop}.

Given that the machine is bad, we have: $\mathbb{E}(D_t - D_{t-1}) = a \cdot \mathbb{P}(\text{win}|\text{bad}) - b \cdot \mathbb{P}(\text{loss}|\text{bad}) = aq - b(1-q) =: \mu < 0$ for any $0 < q < p$. Then, $\mathbb{E}(D_t) = \mu \cdot t \to_t -\infty$. We deduce, for t sufficiently large, using Bienaymé-Tchebychev inequality that

$$\mathbb{P}(\tau \geq t) \leq \mathbb{P}\left(|D_t - \mathbb{E}(D_t)| \geq -\mu \cdot t - m\right) \leq O(1/t).$$

□

Note that since $x^* < x_{\min}$, then we will learn that the product is bad in the static price scenario earlier than in the dynamic pricing model. Conversely, we learn that the product is good with higher probability in the dynamic pricing model.

Lemma 10. *Assume that the product is good. In both pricing scenarios, the probability of learning that the product is good is at least* $\frac{x - x_{stop}}{x(1-x_{stop})}$.

Proof. The probability of playing forever is

$$\mathbb{P}(\tau = \infty) = \mathbb{P}(\tau = \infty \mid \text{good}) \cdot \mathbb{P}(\text{good}) + \mathbb{P}(\tau = \infty \mid \text{bad}) \cdot \mathbb{P}(\text{bad}) = \mathbb{P}(\tau = \infty \mid \text{good}) \cdot x,$$

since $\mathbb{P}(\tau = \infty \mid \text{bad}) = 0$ by Lemma 9. We conclude using Lemma 6. □

In particular, in the symmetric case the probability of learning that the product is good is exactly $\frac{x - x_{\text{stop}}}{x(1-x_{\text{stop}})}$. In this case we can even give an alternative expression for this probability.

Lemma 11. *Assume that $q = 1 - p$. Then when the machine is good, the probability of playing forever is equal to* $1 - \left(\frac{1-\sqrt{1-4p(1-p)}}{2p}\right)^m$ *where m is the smallest integer such that after m losses the prior goes below x_{stop}.*

Proof. Let p_m denote this probability. We have $p_m = 0$ if $m \leq 0$; $p_m = p \cdot p_{m+1} + (1-p) \cdot p_{m-1}$ otherwise, and $\lim_m p_m = 1$. The roots of the polynomial $pX^2 - X + (1-p)$ are 1 and $\frac{1-\sqrt{1-4p(1-p)}}{2p} < 1$. Thus, we deduce easily the expected expression. □

We now intend to compare the two pricing strategies in the symmetric case. We already know that we stop earlier in the static price scenario. Conversely, when the product is actually good the probability of not learning it is greater for the static price strategy. We now give the ratio between these two *false negative* probabilities.

Lemma 12. *Let FN_{static} and $FN_{dynamic}$ be the probabilities of stopping when the product is good, respectively in the static and in the dynamic prices scenarios. We have $FN_{static} \geq FN_{dynamic} > 0$ and in the case when $p = 1 - q$:*

$$\frac{FN_{static}}{FN_{dynamic}} = \frac{x_{\min}}{x^*} \cdot \frac{1 - x^*}{1 - x_{\min}} = \frac{1/x^* - 1}{1/x_{\min} - 1}.$$

Proof. By Lemma 10, the probabilities of having a false negative is $\frac{x_{\text{stop}}}{1-x_{\text{stop}}} \cdot \frac{1-x}{x}$ and the formula follows. □