Multi–Armed Bandit for Pricing

Francesco Trovò FRANCESCO1.TROVO@POLIMI.IT
Stefano Paladino STEFANO.PALADINO@POLIMI.IT
Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Nicola Gatti NICOLA.GATTI@POLIMI.IT
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milano, Italy

Abstract

This paper is about the study of Multi–Armed Bandit (MAB) approaches for pricing applications, where a seller needs to identify the selling price for a particular kind of item that maximizes her/his profit without knowing the buyer demand. We propose modifications to the popular Upper Confidence Bound (UCB) bandit algorithm exploiting two peculiarities of pricing applications: 1) as the selling price increases it is rational to assume that the probability for the item to be sold decreases; 2) since usually people compare prices from different sellers and track price changes over time before buying (specially for online purchases), the number of times that a certain kind of item is purchased is only a small fraction of the number of times that its price is visualized by potential buyers. Leveraging on these assumptions, we consider refinements of the concentration inequality used in the UCB1 algorithm, that result to be significantly tighter than the original one, specially in the early learning stages when only a few samples are available. We provide empirical evidence on the effectiveness of the proposed variations in speeding up the learning process of UCB1 in pricing applications.

Keywords: Multi–Armed Bandit, Upper Confidence Bound Approach, Pricing.

1. Introduction

The problem of designing policies for pricing has been pursued by resorting to different perspectives, for instance, approaches have been developed by considering techniques coming from the economic field (Bertsimas and Perakis, 2006), operations research one (Gallego and Van Ryzin, 1994) and machine learning one (den Boer, 2015). The solutions proposed in the literature consider different sets of assumptions holding on the nature of the considered product (e.g., of the same nature or of different class, perishable or non–perishable, with finite or infinite availability), the buyers' behavior (e.g., fixed and known or generic, varying with time or stationary) and the optimality condition they considered (e.g., social welfare or revenue).

Here we want to address the basic problem of a seller, with an unlimited amount of non–perishable goods, who wants to assign a price to a good without knowing the value that buyers assign to it. This is an interesting problem from the seller point of view, since the profit depends by the pricing policy and the best pricing policy is determined by the buyers' preferences, which are usually unknown *a priori*. In this work, two realistic assumptions have been considered: the conversion probability, i.e., the probability that a buyer purchases the good, is decreasing as the revenue margin is increasing and it is lower than a certain value that is *a priori* known by the seller, for positive item prices. The former assumption comes from the fact that buyers are more willing to

buy a product at lower price, while the latter assumption is based on the fact that sellers usually have *a priori* information, coming from past transactions (e.g., from different products), on the amount of buyers which are not willing to buy any item, but are interested in checking the price anyway, which leads to a low probability of purchasing a good. The described setting could apply to both a local hardware store, as well as to a wider online selling website (e.g., Amazon or Kelkoo). For instance, this setting could be directly applied to the pricing of flight tickets made by an Online Travel Agency (OTA) in a day, where the cost of the ticket provided by the airline is fixed and the availability of the tickets is larger w.r.t. the buyers' demand.

In this scenario, we do not assume to have any further information on the distribution of the buyers' demand, thus we resort to the widely known framework of Multi–Armed Bandit (MAB), where we are able to propose a single price to a buyer (pull a single arm) and we are able to observe the outcome due to this choice (buy/refuse to buy). In this setting, the goal is to design a pricing policy that minimizes the regret, i.e., the loss the seller incurs in the choice of a suboptimal arm, and, at the same time, that is able to converge to the optimal price choice. While most of the algorithms developed in the MAB setting do not consider correlations existing among arms, here the knowledge of the outcome of an arm is able to provide information regarding other arms. For instance, if a buyer accepts to buy a product at a specific price, it is reasonable to expect that she/he would have bought it also at a lower price, while if she/he refuses the good at a specific price the same would have been also if the seller had proposed a higher price.

MAB approaches have been already considered for dynamic pricing (den Boer, 2015) exploiting both Bayesian (Leloup and Deveaux, 2001; Cope, 2007) and non–Bayesian (Kleinberg and Leighton, 2003; Tehrani et al., 2012) formulations. As observed in several works, dynamic pricing can be modeled as a *partial monitoring* game (Cesa-Bianchi et al., 2006; Bartók et al., 2014), for which several studies on asymptotic regret bounds have been produced in the least decade both in stochastic (Bartók et al., 2011, 2012) and adversarial (Piccolboni and Schindelhauer, 2001; Cesa-Bianchi et al., 2006; Foster and Rakhlin, 2012) settings.

In this paper, we consider dynamic pricing in the stochastic settings and we propose two variations to the popular UCB1 algorithm (Auer et al., 2002) that, exploiting the correlation among the outcomes of different arms (prices) as well as the *a priori* information on the maximum conversion probability, produce tighter Upper Confidence Bounds (UCBs) over the expected utility of the arms. We focused on UCB1 due to its strong regret bounds together with its computational efficiency as needed for e–commerce applications where thousands of pricing operations per second may be requested. Although the proposed variations do not improve the UCB1 regret bounds, they provide significant advantages in many realistic pricing situations as will be shown by empirical results.

2. Problem Formulation

We consider a scenario where an unlimited non–perishable amount of goods is available to a monopolistic seller, who wants to propose at a chosen price the product she/he is selling to a buyer. Each buyer is modeled as a deterministic agent which will buy the item only if the proposed price is lower or equal than a threshold $s \in \mathbb{R}^+$. Since buyers have generally different thresholds s, we considered s as drawn from a random variable S with probability density function (pdf) $\mathcal{S}$ over $\Omega \subseteq \mathbb{R}^+$. Here we consider a stationary setting where the threshold distribution $\mathcal{S}$ does not change over time.

Given an arm $a_i \in \mathbb{R}^+$, that corresponds to a specific price, we define a variable $X_i \sim Be(\mu_i)$ which represents the *outcome* (buy/refuse to buy) of the transaction. The expected *conversion prob-*

ability μ_i corresponding to price a_i is defined as the probability that a user will purchase the product:

$$\mu_i = \Pr(s \geq a_i) = 1 - \int_0^{a_i} \mathcal{S}(x)dx. \tag{1}$$

It is clear that for the non–negativity of the distribution function $\mathcal{S}$ and for the property of the integral we have that $a_i < a_j \Rightarrow \mu_i > \mu_j$, i.e., the expected conversion probability is decreasing w.r.t. the price. We would like to point out that by adopting this model of the user we automatically obtain the first assumption that will be exploited in the following section.

The second assumption is that the seller knows that only a certain proportion of the buyers $\mu_{\max} \in [0, 1]$ will really consider the possibility of purchasing the good, while the remaining proportion $1 - \mu_{\max}$ will not buy at any price. This can be introduced in the user model by considering S with pdf equal to $\mathcal{S}(a) = (1 - \mu_{\max}) \cdot \delta(0) + \mu_{\max} \cdot \mathcal{C}(a), a \in \Omega$, where $\delta(0)$ is a Dirac delta distribution centered in 0 and $\mathcal{C}$ is a pdf defined over Ω. In this context, the seller faces a problem that can be modeled by considering a MAB setting. In fact, at each time t (or equivalently as soon as the t–th client appears), we are allowed to collect the reward (either 0 or a_i) from a previously selected arm a_i, since we are able to propose a single price to a buyer. We consider a finite set of ordered arms (prices) $A = \{a_1, \ldots, a_K\}$, where $|A| = K < \infty$ and $a_i \geq a_j$ iff $i \leq j$.[1] Thus, the *reward* of arm a_i at time t can be modeled as a random variable $V_i = a_i X_i$, $\forall i \in \{1, \ldots, K\}$. By considering the definition of outcome X_i, we have that $\nu_i := \mathbb{E}[V_i] = a_i \mathbb{E}[X_i] = a_i \mu_i, \forall i \in \{1, \ldots, K\}$.

A (pricing) *policy* over a MAB setting is an algorithm $\mathfrak{U}(h_t)$ that chooses the next arm to play $a_{\bar{i}}$ at time t according to history h_t, defined as the sequence of past plays and obtained rewards. Remind that at each time t we observe a single realization of the reward $v_{\bar{i},t}$ coming from the arm $\mathfrak{U}(h_t) = a_{\bar{i}}$. After t time points, we are provided with a set of realizations $D_{it} = \{x_{it'} | \mathfrak{U}(h_{t'}) = a_i, t' < t\}$ for each arm $a_i \in A$, where $x_{it'}$ is the realization of the outcome X_i at time t'. The objective of a policy is to maximize the expected cumulative reward or equivalently to minimize the loss w.r.t. the optimal decision (in terms of revenue). This loss is usually addressed as *total regret*, whose definition over a finite time horizon N is as follows:

$$R_N = \nu^* N - \sum_{i=1}^{K} \nu_i \mathbb{E}[T_i(N)] \tag{2}$$

where $\mathbb{E}[\cdot]$ is the expectation w.r.t. the stochastic components of the algorithm, $\nu^* := \max_{i \in \{1,\ldots,K\}} \nu_i$ and $T_i(t) := |D_{it}|$ is the number of times the algorithm selects the i–th arm in the first t time points.

3. Proposed Methodology

In what follows we propose a framework, composed of two different parts, which is able to incorporate into UCB–like algorithms information about the correlation existing among arms and *a priori* knowledge about the maximum conversion probability of the arms. The main idea for exploiting correlation is to use the decreasing structure of the expected conversion probabilities $\{\mu_i\}_{i=1}^K$ to tighten their UCBs. Moreover, to incorporate in our framework the *a priori* information about the maximum conversion probability we consider a form of the Chernoff bound which is tighter than the one proposed by Hoeffding when the conversion probability is particularly low ($\mu_i < \mu_{\max} \ll 1$). In the methodological section, we consider the two aspects as separated entities, but their joint use, which fully exploits the characteristics of the studied pricing problem, will be shown in the experimental section.

1. We denote with $|A|$ the cardinality of a set A.

3.1 Arms Correlation

Due to the specific structure of the considered problem, we include the information about the ordering of the expected values $\{\mu_i\}_{i=1}^K$ along the arms. If we consider arm a_i, we have that the realizations of all the arms $a_j < a_i$ provide information for the computation of the UCB on the expected value μ_i. In fact, since $\mu_i \leq \mu_j$ we can use the samples in D_{jt} as optimistic estimates of what would have happened to realizations coming from arm a_i at time instant t.

Let us define the estimator for the average value of X_{ij}, that is a random variable resulting from a convex combination of the random variables $\{X_j, \dots, X_i\}$:

$$\hat{X}_{ji,t} = \frac{\sum_{x \in D_{ji,t}} x}{T_{ji}(t)}, \tag{3}$$

where $D_{ji,t} := \bigcup_{k=j}^{i} D_{kt}$ is the union of the sets $\{D_{jt}, \dots, D_{it}\}$ and $T_{ji}(t) := \sum_{k=j}^{i} T_k(t)$ is the cardinality of the set $D_{ji,t}$. The following theorem holds:

Theorem 1 *Given a set of sequence of realizations $\{D_{1t}, \dots, D_{Kt}\}$ drawn from $X_i \sim P(\Theta)$ with $i \in \{1, \dots, K\}$ having $\mu_i := \mathbb{E}[X_i]$ and $\mu_i \leq \mu_j$ iff $j \leq i$, where P is a probability distribution over a finite set $\Theta = [a, b]$, the following holds for any $j \in \{1, \dots, i\}$:*

$$\mathbb{P}(\mu_i \geq \hat{X}_{it} + \varepsilon_i) \leq e^{-\frac{2T_{ji}(t)\left(\hat{X}_{it} - \hat{X}_{ji,t} + \varepsilon_i\right)^2}{(b-a)^2}}, \tag{4}$$

where $\hat{X}_{it} = \frac{\sum_{x \in D_{it}} x}{T_i(t)}$.

Proof By considering a generic variable X_i, we have that the following holds $\forall j \in \{1, \dots, i\}$:

$$\mathbb{P}(\mu_i \geq \hat{X}_{it} + \varepsilon_i) \underbrace{\leq}_{\mu_{ji} \geq \mu_i} \mathbb{P}(\mu_{ji} \geq \hat{X}_{it} + \varepsilon_i) = \mathbb{P}(\mu_{ji} \geq \hat{X}_{ji,t} - \hat{X}_{ji,t} + \hat{X}_{it} + \varepsilon_i) \leq e^{-\frac{2T_{ji}(t)(\hat{X}_{it} - \hat{X}_{ji,t} + \varepsilon_i)^2}{(b-a)^2}},$$

where the last inequality follows from the Hoeffding bound and μ_{ji} is the mean of a random variable X_{ji} with values over Θ, i.e., $\mathbb{E}[X_{ji}] =: \mu_{ji} \geq \mu_i$. ■

In the setting considered in this work, $X_i \sim Be(\mu_i)$, thus we have $P = Be(\mu_i)$ and $\Theta = \{0, 1\}$, thus the inequality provided in Theorem 1 reduces to:

$$\mathbb{P}(\mu_i \geq \hat{X}_{it} + \varepsilon_i) \leq e^{-2T_{ji}(t)\left(\hat{X}_{it} - \hat{X}_{ji,t} + \varepsilon_i\right)^2}. \tag{5}$$

If we need an UCB over μ_i holding with confidence of at least $1 - p$ with $p \in [0, 1]$ we have:

$$\mathbb{P}(\mu_i \geq \hat{X}_{it} + \varepsilon_i) \leq e^{-2T_{ji}(t)\left(\hat{X}_{it} - \hat{X}_{ji,t} + \varepsilon_i\right)^2} = p, \qquad \varepsilon_i = \hat{X}_{ji,t} - \hat{X}_{it} + \sqrt{-\frac{\log(p)}{2T_{ji}(t)}}.$$

By considering $\bar{j} = \arg\min_{j \in \{1, \dots, i\}} e^{-2T_{ji}(t)\left(\hat{X}_{it} - \hat{X}_{ji,t} + \varepsilon_i\right)^2}$, we will have the tighter bound among those provided by Theorem 1. This constitutes a potential improvement over the traditional bound obtained by considering realizations coming from a single arm. In fact, this novel UCB considers $T_{ji}(t) \geq T_i(t)$ samples and it is tighter or, in the worst case, equal to the Hoeffding's one. For instance, when the empirical means for the arms are monotonically decreasing (i.e., $\hat{X}_{it} < \hat{X}_{jt},\ \forall i > j$) the use of realizations coming from other arms provides little contribution to tighten the bound, but in the case of an inversion ($\exists i > j | \hat{X}_{it} > \hat{X}_{jt}$) we have that the bound in Theorem 1 can significantly improve over the original one. Such a situation is exemplified in Figure 1, where, in contrast with the assumption about the monotonicity of the conversion probabilities, the empirical conversion $\hat{X}_{1t}$ of arm $a_1 = 1$ is lower than the one $\hat{X}_{2t}$ of arm $a_2 = 2$. This happens because arm a_2 has been selected much less than arm a_1 and so its empirical mean is more uncertain. In this

Algorithm 1 UCB1-O

Initialization

Play each arm $a_i \in A$ once.

Loop

At time instant t

for $i = 1, \ldots, K$ **do**

Compute:

$$u_i = \min_{1 \le j \le i} \left\{ \hat{X}_{ji,t} + \sqrt{\frac{2\log(t)}{|T_{ij}(t)|}} \right\}$$

end for

Play $\bar{i}$-th arm s.t.: $\bar{i} = \arg\max_i a_i u_i$

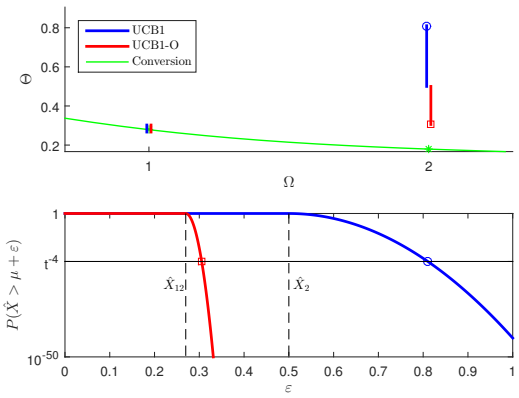

Figure 1: Example of inversion of estimates for decreasing expected values and corresponding UCBs.

case, the samples of arm a_1 decrease the UCB for arm a_2 from the value denoted by the blue circle to the value denoted by the red square (top). This reduction of the UCB for arm a_2 does not imply a reduction in the confidence level, since the two values have been obtained from different bounds, but with the same confidence level $(1 - t^{-4})$, as shown in Figure 1 (bottom).

The proposed algorithm (UCB1-O) is presented in Algorithm 1. At first, the algorithm selects in turns each arm at least once, to have at least an outcome realization coming from each arm. Once the initial round is completed, it considers, for each arm a_i:

$$u_i = \varepsilon_i + \hat{X}_{it} = \hat{X}_{\bar{j}i,t} + \sqrt{\frac{2\log(t)}{T_{\bar{j}i}(t)}}, \tag{6}$$

where $\bar{j}$ is selected in order to minimize the bound in Theorem 1 and the confidence level is selected, as done for the UCB1 algorithm (Auer et al. (2002)), by setting $p = t^{-4}$. At last, the algorithm selects for the next round the arm $a_{\bar{i}}$ providing the maximum upper bound over the expected reward.

3.2 *A Priori* Information

To include the information about the prior over the conversion probability, i.e., that $\mu_i \le \mu_{\max}$, $\forall i \in \{1, \ldots, K\}$, we considered the fact that the Hoeffding bound (Hoeffding (1963)) does not constitute a tight bound in the case we are considering random variables with bounded expected value. More specifically, one of the approximations considered in the bound for the generic outcome X_i is:

$$e^{-T_i(t)D(\mu_i+\varepsilon||\mu_i)} \le e^{-2T_i(t)\varepsilon^2}, \tag{7}$$

where $D(\mu_i + \varepsilon||\mu_i)$ is the Kullback-Leibler (KL) divergence between two Bernoulli variables with mean $\mu_i + \varepsilon$ and μ_i, respectively. As shown in Figure 2, the bound based on the KL divergence (solid lines) and the one on Hoeffding's inequality (dash–dotted line) diverge as μ decreases. To reduce the gap, we consider the following results known in literature as one of the formulation of the Chernoff bound (Chernoff (1981)):

Theorem 2 *Given a set of realizations D_{it} extracted from $X_i \sim Be(\mu_i)$, we have:*

$$\mathbb{P}(\mu_i \ge \hat{X}_{it} + \varepsilon_i) \le e^{-\frac{T_i(t)\varepsilon^2}{\mu_i}}. \tag{8}$$

Algorithm 2 UCB1-P

Initialization

Input: $\mu_{\max}$

Play each arm $i \in \mathcal{K}$ once.

Loop

At time instant t

for $i = 1, \ldots, K$ **do**

Compute:

$$u_i = \hat{X}_{it} + \sqrt{\frac{4\mu_{\max} \log(t)}{T_i(t)}}$$

end for

Play $\bar{i}$-th arm s.t.: $\bar{i} = \arg\max_i a_i u_i$

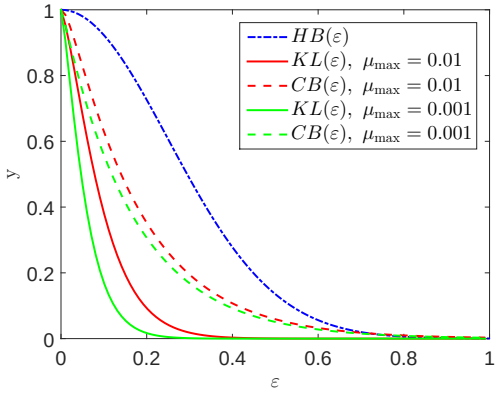

Figure 2: Example of bounds $y = e^{-x(\varepsilon)}$ obtained with different $x(\varepsilon)$: Hoeffding Bound ($HB(\varepsilon)$), Kullback Leiber divergence ($KL(\varepsilon)$) and Chernoff Bound ($CB(\varepsilon)$).

In the application scenario we are considering, it is interesting to derive UCBs from it, since it provides a better approximation of the KL divergence in the case $\mu_i \leq \mu_{\max} \ll 1$, as shown in Figure 2 it provides tighter estimates (dashed lines) of the one derived from KL divergence. If we want an UCB with confidence $1 - p$ with $p \in [0, 1]$ we have from Theorem 2:

$$\mathbb{P}(\hat{X}_{it} \geq \mu_i + \varepsilon) \leq e^{-\frac{T_i(t)\varepsilon^2}{\mu_i}} \leq e^{-\frac{T_i(t)\varepsilon^2}{\mu_{\max}}} = p, \qquad \varepsilon = \sqrt{\frac{4\mu_{\max} \log p}{T_i(t)}},$$

where the last inequality derives from the trivial fact that $\mu_{\max} \geq \mu_i, \forall i \in \{1, \ldots, K\}$ and ε is computed by considering the positive root of the second order equality associated to the first inequality. It is possible to compute, by comparing the two bounds provided by Hoeffding and Chernoff, a sufficient condition that identifies when the former is more convenient than the latter. In fact, it is easy to infer that when $\mu_{\max} > \frac{1}{2}$ the bound in Equation (8) is wider than the one in the RHS of Equation (7), thus, if we cannot guarantee low conversion probabilities, it is better to resort to the traditional Hoeffding bound. Further insights will be given in the experimental section. The proposed algorithm (UCB1-P) is presented in Algorithm 2. In UCB1-P we set, similarly to what has been considered in the UCB1 algorithm, $p = t^{-4}$ and we choose the next arm to be pulled by selecting the one having the maximum expected revenue.

4. Experimental Section

To evaluate the performance of the proposed framework, we run several simulations on a variety of different pricing settings by changing the distribution associated to the buyers' demand and the number of arms (prices) over which the optimization is performed. For sake of synthesis, due to space limitation, here we report only the most relevant results that has been obtained in three case studies, provided that experiments on other configurations do not change the final conclusions. In the first two experiments, we compared the proposed algorithms UBC1-O and UCB1-P with the UCB1 policy in terms of their total regret relatively to the total regret suffered by UCB1 ($R_N(UCB1)$), i.e., $R_{\%} = \frac{R_N}{R_N(UCB1)}$. Furthermore, we considered the performance of UCB1-OP, which is a straightforward combination of the other two proposed algorithms considering as upper

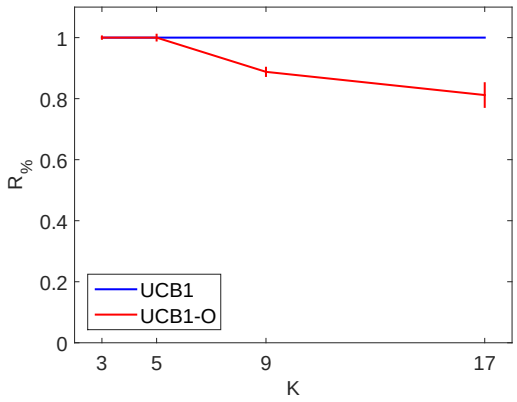

Figure 3: Regret results for S_1 with different number of arms K

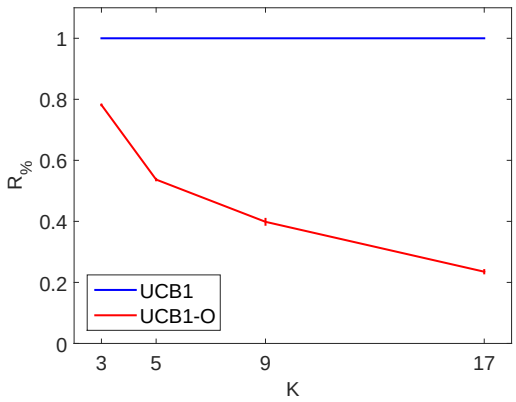

Figure 4: Regret results for S_2 with different number of arms K

bound $u_i = \min_{1\leq j\leq i}\left\{\hat{X}_{ji,t} + \sqrt{\frac{4\mu_{\max}\log(t)}{T_{ji}(t)}}\right\}$, on a third case study where the performance are measured in terms of average regret $\bar{R}_t = \frac{R_t}{t}$. All the algorithms are implemented in `MatLab`.

4.1 Application D1: Correlation Information

Experimental setting We considered 8 different configurations for the distribution S in order to evaluate the integration of the correlation structure information in the algorithms. We considered configurations with different numbers of arms $K \in \{3, 5, 9, 17\}$ (evenly spaced over the interval $[1, 17]$) and with different conversion probabilities, in the specific, $S_1 \sim \mathcal{N}(14, 1)$, representing a situation where the optimal price is high and $S_2 \sim \mathcal{N}(5, 1)$, where the optimal price is the one corresponding to one of the lower arms. Here we do not consider any information about the maximum conversion probability, i.e., we set $\mu_{\max} = 1$. We generate 20 independent sequence of $N = 50,000$ thresholds s from $S1$ and $S2$ and run the UCB1 and UCB1-O policies on them.

Results In Figures 3 and 4 we have the results for $R_{\%}$ for UCB1 (which is always 1 by definition of $R_{\%}$) and for UCB1-O. When the optimal arm is among the higher ones, we have only a little improvement of the performance of UCB1-O w.r.t. UCB1. This is mainly due to the fact that the problem itself favors the arms with high values since the strategy is based on the product of prices and conversion probabilities. Thus, intrinsically even a problem of this kind with a large number of arms would reduce to a selection among a small subset of arms. In the situation where the optimal arm is among the lower ones, we have that the information provided by the ordering is able to induce a significant reduction in the regret, with the regret of UCB1-O that ranges between 30% and 80% of the regret of UCB1. As expected, the advantage increases with the number of arms.

4.2 Application D2: *A Priori* Information

Experimental setting In this second set of experiments, we considered 4 arms $A = \{1, 2, 3, 4\}$, $\mu_{\max} \in \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ and customer thresholds coming from a Gaussian Mixture Model, so that the optimal arm is $\gamma_3 = 3$ and the second best one is $\gamma_1 = 1$ (i.e., having $\mathcal{C} = 0.8 \cdot \mathcal{N}(0.85, 0.38) + 0.2 \cdot \mathcal{N}(3.4, 0.05)$). As in the previous case, results are averaged over 20 repetitions, but the time horizon has been extended to $N = 2,000,000$.

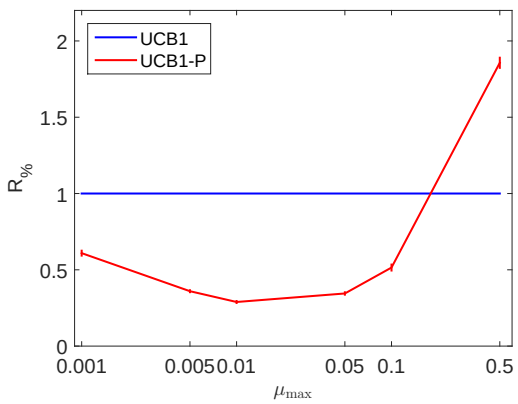

Figure 5: Regret for cases where *a priori* knowledge over $\mu_{\max}$ is available.

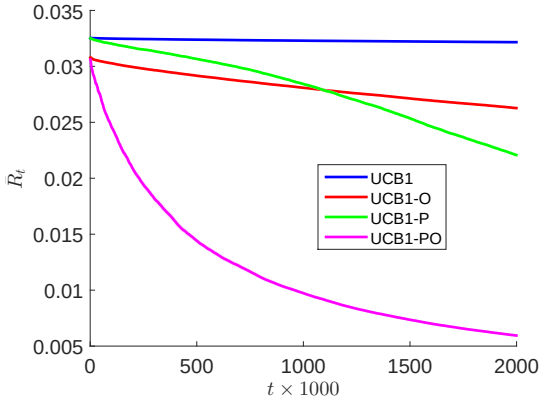

Figure 6: Results for average regret in the realistic pricing problem.

Results The results of the simulations are presented in Figure 5. It is possible to see that, as stated in Section 3, using UCB1-P in settings where arms have high conversion probabilities ($\mu_{\max} > 0.25$) is worse than using UCB1, since the Chernoff bound gets looser than the Hoeffding's one. On the other hand, when we consider cases where $\mu_{\max} \ll 0.25$, the total regret of UCB1-P is between the 30% and the 60% w.r.t. the UCB1 one.

4.3 Application D3: Correlation and *a Priori* Information

Experimental setting At last we considered a realistic scenario of pricing, where both the assumptions hold. Here we, considered $K = 17$ arms (evenly spaced over $[1, 17]$), $\mu_{\max} = 0.01$ and $\mathcal{C} = \mathcal{N}(5, 1)$. We run 20 independent simulations over a time horizon of $N = 2,000,000$.

Results The results of the simulations are presented in Figure 6. All the proposed algorithms provide better results than UCB1, that even after two millions steps exhibits no learning trend. UCB1-O is able to exploit the information about the ordering and provides better performance than UCB1-P in the early stages of the simulations. On the other hand, UCB1-P is able to exploit the *a priori* information over the whole simulation, thus being able to reach a lower $\bar{R}_t$ at the end of the simulation. The combination of the two methodologies in UCB1-OP is here really effective, since it is able to outperform all the other algorithms in terms of average regret during the whole simulation.

5. Conclusion and Future Works

In this paper, we cast the pricing pricing as a multi-armed problem with dependent arms. We presented solutions to incorporate common assumptions in pricing applications, such as the correlation among different arms and *a priori* information about their maximum conversion probabilities, into the design of a UCB-like arm selection policy. We provided a theoretical background to justify the use of these methodologies and presented experimental evidence that their joint use is effective in a realistic scenario. Future developments of this work may explore both the application of such techniques to either other frequentist or Bayesian MAB policies and the inclusion of a strategic component in the problem, for instance by considering also other competing sellers.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *COLT*, volume 2011, pages 133–154, 2011.

Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. *arXiv preprint arXiv:1206.6487*, 2012.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring-classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Dimitris Bertsimas and Georgia Perakis. Dynamic pricing: A learning approach. *Mathematical and computational models for congestion charging*, pages 45–79, 2006.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Herman Chernoff. A note on an inequality involving the normal distribution. *The Annals of Probability*, pages 533–535, 1981.

Eric Cope. Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics (NRL)*, 54(3):265–281, 2007.

Arnoud V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 2015.

Dean P Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *International Conference on Artificial Intelligence and Statistics*, pages 382–390, 2012.

Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 594–605. IEEE, 2003.

Benoît Leloup and Laurent Deveaux. Dynamic pricing on the internet: Theory and simulations. *Electronic Commerce Research*, 1(3):265–276, 2001.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory*, pages 208–223. Springer, 2001.

Pouya Tehrani, Yixuan Zhai, and Qing Zhao. Dynamic pricing under finite space demand uncertainty: a multi-armed bandit with dependent arms. *arXiv preprint arXiv:1206.5345*, 2012.