Proceedings of Machine Learning Research vol 98:1–27, 2019 30th International Conference on Algorithmic Learning Theory

Dynamic Pricing with Finitely Many Unknown Valuations

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Tommaso Cesari TOMMASO.CESARI@UNIMI.IT
Department of Computer Science & DSRC, Università degli Studi di Milano

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
CMLA, ENS Paris-Saclay, Université Paris-Saclay
CNRS, Centre National de la Recherche Scientifique
Criteo AI Lab, Paris

Editors: Aurélien Garivier and Satyen Kale

Abstract

Motivated by posted price auctions where buyers are grouped in an unknown number of latent types characterized by their private values for the good on sale, we investigate regret minimization in stochastic dynamic pricing when the distribution of buyers' private values is supported on an unknown set of points in $[0, 1]$ of unknown cardinality K.
Keywords: Posted price auction, dynamic pricing, regret minimization, multiarmed bandits.

1. Introduction

In the online posted price auction problem, also known as dynamic pricing, an unlimited supply of identical goods is sold to a sequence of buyers. To each buyer in the sequence, the seller makes a take-it-or-leave-it offer for the good at a certain price (which we assume to belong to the unit interval $[0, 1]$). The good is purchased if and only if the offered price is lower or equal to the buyer's private valuation (also assumed to be in $[0, 1]$). At the end of the transaction, the seller's revenue is either zero (if the good is not sold) or equal to the offered price. The buyer's valuation is never observed. Indeed, the seller only learns a single bit for each auction, i.e., whether the good was sold or not at the chosen price. Similarly to previous works (Kleinberg and Leighton, 2003; Blum et al., 2004; Blum and Hartline, 2005), we assume that the price offered to the t-th buyer in the sequence only depends on the past history of observed sales. In particular, we assume that buyers are indistinguishable, and provide no information to the seller other than their willingness to buy at the specified price. For this reason, the seller can post the price for the next buyer publicly, before the buyer shows up.

We evaluate the seller's performance in terms of regret, measuring the difference between the seller's revenue and the revenue achievable by consistently posting the optimal price. The regret in dynamic pricing was initially investigated by Kleinberg and Leighton (2003) under various assumptions on the generation of the buyers' valuations. In the stochastic setting, in which valuations are drawn i.i.d. from a fixed and unknown distribution on $[0, 1]$, they show that no algorithm can achieve a $o(\sqrt{T})$ regret and provide an algorithm achieving regret of order $C\sqrt{T \log T}$, where T is the number of buyers in the sequence and C only depends on the distribution of buyers' valuations. Their upper-bound holds under some assumptions on the demand curve, which is the function D

mapping each price x to the probability $D(x) = \mathbb{P}(V \geq x)$ that the good is sold. Specifically, the revenue function $x \mapsto xD(x)$ is required to have a unique global maximum $x^\star \in (0, 1)$ and be twice differentiable with a negative second derivative at $x^\star$. Without these assumptions, the authors prove a much higher lower bound of order $T^{2/3}$ on the regret. The algorithm achieving the $C\sqrt{T \log T}$ regret under the above assumptions on the demand curve is simple: it runs the UCB1 policy for stochastic bandits (Auer et al., 2002a) on a discretized set of $K = (T/\log T)^{1/4}$ prices.

In this paper, we study the stochastic setting of dynamic pricing under completely different assumptions on the demand curve. Namely, that the distribution of buyers' valuations is supported on an *unknown* set of *unknown* finite cardinality K. This models any setting in which buyers are grouped in an unknown number of latent types, characterized by their private values for the good on sale. In particular, this applies to regret minimization in sellers' repeated second-price auctions with a single relevant buyer. This scenario emerges naturally when a seller and a buyer interact repeatedly, and the valuation of the good depends on contextual information known only to the buyer. For instance, in online advertising each time a user lands on a publisher's website, an impression is put on sale to a set of relevant advertisers through an auction (note that whenever there is a single relevant advertiser for the impression, second-price auctions with reserve price are equivalent to posted price auctions). Now, typically, the advertiser's valuation for the impression depends on which segment the user belongs to, where the finite segmentation is based on private information not accessible to the publisher.

Note that our model is very different from assuming that the seller is restricted to offer prices from a *known* finite set of size K (Rothschild, 1974), which makes dynamic pricing a special case of K-armed stochastic bandits. In our model, the seller does not know the K buyers' valuations, not even their number! So, besides learning which valuation has the highest revenue, the seller must also learn the location of these values. This interplay between noisy search and bandit allocation is one of the main themes of our work.

In contrast with previous approaches, which typically assume parametric (Broder and Rusmevichientong, 2012) or locally smooth (Kleinberg and Leighton, 2003) demand curves, our model with finitely many valuations is equivalent to assuming that the demand curve is piecewise constant with a finite number of discontinuities. Recently, den Boer and Keskin (2018) designed an algorithm for piecewise continuous demand curves achieving an upper bound of order $C\sqrt{T}\log T$ in the piecewise constant case. However, up to constant factors, their hefty leading constant C is at least as big as the maximum between $K^{22}\gamma^{-16}c^{-2}$ and $K^{12}\gamma^{-8}c^{-18}$, where c is the minimum distance between valuations and both K and the smallest drop γ in the demand curve must be known in advance. Although their setting extends ours to certain piecewise *parametric* demand curves, we believe that discontinuities are the real source of additional hardness of this dynamic pricing model with respect to previously studied settings.

Our first result is a lower bound of order $\sqrt{KT}$ on the regret in the distribution-free case (where the regret is maximized over all possible demand curves), which holds even when the seller knows the number and position of buyers' private values in advance. This essentially establishes that our setting is at least as hard as a K-armed bandit problem. Although we build on the stochastic lower bound of Kleinberg and Leighton (2003), our proof is not a simple adaptation of theirs. Indeed, we show that their proof breaks down when K is constant and T grows, which is exactly the regime we are interested in. Then, we present an efficient algorithm achieving a distribution-free upper bound on the regret of order $\sqrt{KT \log T}$ without any additional knowledge of the parameters of the problem.[1] The detailed version of our bound has a significantly better dependence than den Boer and Keskin (2018) on the smallest difference c between two adjacent valuations, and matches—up to logarithmic factors—the lower bound stated above.

In the distribution-dependent case, when the gap Δ between the revenue of the optimal valuation and that of the second-best valuation is constant, we prove the impossibility of obtaining regret bounds of order significantly better than $\sqrt{T}$ even when $K = 3$, thus showing that this setting is strictly harder than K-armed stochastic bandits. Motivated by this impossibility result, we investigate distribution-dependent bounds that rely on additional information about the demand curve. By combining suitable generalizations of UCB1 (Auer et al., 2002a) and the "cautious search" strategy of Kleinberg and Leighton (2003), we obtain an efficient algorithm achieving a regret of order at most $\bigl(1/\Delta + (\log\log T)/\gamma^2\bigr)\bigl(K\log T\bigr)$, where, as before, γ is the smallest drop in the demand curve. Since $(K/\Delta)\log T$ is the regret of K-armed stochastic bandits, this shows that the price of identifying each one of the K valuations is at most $(\log T)(\log\log T)/\gamma^2$, which corresponds (up to $\log\log$ factors) to the known upper bounds for noisy binary search (Karp and Kleinberg, 2007). We conclude the study of the distribution-dependent case by presenting an efficient algorithm with regret of order $(1/\Delta + \log\log T)\log T$ when the number of valuations is known to be at most two. Surprisingly, this bound is the same (up to $\log\log$ terms) as the best possible bound for two-armed stochastic bandits, achievable when not only the number, but also the locations of the valuations are known in advance. In order to prove this result we introduce a novel technique for estimating (up to a multiplicative constant) the expectation μ of any $[0,1]$-valued random variable with probability at least $1-\delta$, using at most $\mathcal{O}\bigl(\frac{1}{\mu}\ln\frac{1}{\delta}\bigr)$ samples, even if the expectation μ is *not* known in advance. We believe this technique may be valuable in its own right.

2. Further related works

The literature on dynamic pricing and online posted price auctions is vast. We address the reader to the excellent survey published by den Boer (2015), providing a comprehensive picture of the state of the art until the end of 2014 —see also the tutorial slides by Slivkins and Zeevi (2015) for a perspective more focused on computer science approaches. An important line of work in dynamic pricing considers a nonstochastic setting in which the sequence of the buyers' private values is deterministic and unknown, and the seller competes against the best fixed price. This model was pioneered by Kleinberg and Leighton (2003), who proved a $\mathcal{O}(T^{2/3})$ upper bound (ignoring logarithmic factors) on the aforementioned notion of regret. Later works (Blum et al., 2004; Blum and Hartline, 2005) show simultaneous multiplicative and additive bounds on the regret when prices have range $[1,h]$. These bounds have the form $\varepsilon\, G_T^\star + \mathcal{O}\bigl((h\ln h)/\varepsilon^2\bigr)$ ignoring $\ln\ln h$ factors, where $G_T^\star$ is the total revenue of the optimal price $p^\star$. Recent improvements on these results are due to Bubeck et al. (2017), who prove that the additive term can be made $\mathcal{O}(p^\star(\ln h)/\varepsilon^2)$, where the linear scaling is now with respect to the optimal price rather than the maximum price h. Other variants consider settings in which the number of copies of the item to sell is limited (Agrawal and Devanur, 2014; Babaioff et al., 2015; Badanidiyuru et al., 2013) or settings in which a returning buyer acts strategically in order to maximize his utility in future rounds (Amin et al., 2013; Devanur et al., 2014)

1. Throughout this paper we assume that the time horizon T is known by the seller in advance. This assumption can be easily removed with a "doubling trick" (see, e.g., (Cesa-Bianchi and Lugosi, 2006)), a standard technique for extending regret bounds to time sequences of unknown length.

Finally, although in this work we focus on the seller's side, regret minimization approaches have been recently applied also on the buyer's side, for example in (McAfee, 2011; Weed et al., 2016).

3. Preliminaries and definitions

We assume all valuations V_t belong to a fixed and unknown finite set $\mathcal{V} = \{v_1, \dots, v_K\} \subset [0,1]$, with $0 = v_0 \le v_1 < \cdots < v_K \le v_{K+1} = 1$. Unless otherwise specified, the sequence $V_1, V_2, \dots$ is assumed to be sampled i.i.d. from a fixed and unknown distribution on $\{v_1, \dots, v_K\}$. Let $p_i = \mathbb{P}(V_1 = v_i)$ and assume (without loss of generality) that $p_i > 0$ for all $i \in \{1, \dots, K\}$. An instance of the posted price problem is then fully specified by the pairs $(v_1, p_1), \dots, (v_K, p_K)$. We assume auctions are implemented according to the following online protocol: for each round $t \in \{1, 2, \dots\}$

1. the seller posts a price $X_t \in [0,1]$
2. buyer's valuation V_t, hidden from the seller, is drawn from $\mathcal{V}$ according to $\{p_1, \dots, p_K\}$
3. the seller observes $\mathbb{I}\{V_t \ge X_t\} \in \{0,1\}$ and computes the revenue $r_t(X_t) = X_t \, \mathbb{I}\{V_t \ge X_t\}$

Note that the expected revenue $\mathbb{E}[r_t(x)] = \mathbb{E}\big[x \, \mathbb{I}\{V_t \ge x\}\big]$ is equal to $x\, D(x)$, where

$$D(x) = \mathbb{P}(V_1 \ge x) = \sum_{k \colon v_k \ge x} p_k \tag{1}$$

is the *demand curve*. Hence the price maximizing the expected revenue $\mathbb{E}[r_t(x)]$ belongs to the set of valuations $\{v_1, \dots, v_K\}$ and we denote one of the possible optimal valuations by $v^\star = v_{i^\star}$. We define the suboptimality gap of v_j with respect to $v^\star$ by $\Delta_j = \mathbb{E}\big[r_1(v^\star) - r_1(v_j)\big]$. The goal of the seller is to minimize the *regret*

$$R_T = \max_{x \in [0,1]} \mathbb{E}\left[\sum_{t=1}^T r_t(x) - \sum_{t=1}^T r_t(X_t)\right] = \mathbb{E}\left[\sum_{t=1}^T r_t(v^\star) - r_t(X_t)\right]$$

where the expectation is understood with respect to any randomness in the generation of $V_1, \dots, V_T$ and $X_1, \dots, X_T$. Formally, a *deterministic seller* is a sequence of functions $X_1, X_2, \dots$ where $X_t = f_t(X_1, Z_1, \dots, X_{t-1}, Z_{t-1})$ is the price posted at time t, the random variable Z_s is the binary feedback $\mathbb{I}\{V_s \ge X_s\}$ received by the seller in at time s, and $f_t \colon \big([0,1] \times \{0,1\}\big)^{t-1} \to [0,1]$ is an arbitrary function. A *randomized seller* is a probability distribution over deterministic sellers.

4. Lower bounds

In this section we discuss some important similarities and differences between dynamic pricing with K valuations and the K-armed bandit problem (proofs are deferred to Appendix A). First, we state that in the distribution-free case the former is at least as difficult as the latter. More precisely, if $T \ge K^3$, no algorithm can have regret better than $\sqrt{KT}$ on dynamic pricing with K valuations.

Theorem 1 *For any number of valuations $K \ge 3$ and all time horizons $T \ge K^3$ there exist K pairs $(v_1, p_1), \dots, (v_K, p_K)$ such that the expected regret of any pricing strategy satisfies $R_T = \Omega\big(\sqrt{KT}\big)$.*

Next, we claim that in the distribution-dependent case, dynamic pricing is strictly harder than multiarmed bandits. More precisely, even if the suboptimality gap Δ is constant and K is small, no dynamic pricing algorithm can have regret better than $\sqrt{T}$, whereas the distribution-dependent regret of multiarmed bandits is $\mathcal{O}(\log T)$.

Algorithm 1:

Input: $T \in \mathbb{N}$, $\delta \in (0,1)$.
Initialization: $\mathcal{K}_1 \leftarrow \{1\}$, $k_1 \leftarrow 1$, $a_1 \leftarrow 0$, $b_1 \leftarrow 1$, $a_0 \leftarrow 0$, $\overline{D}(0) \leftarrow 0$.
for $m = 1, 2, \dots$ **do** // search phase
  if $\{j \in \mathcal{K}_m \mid b_j - a_j > T^{-1/2}\} \neq \varnothing$ **then**
    pick $i_m = \min\{j \in \mathcal{K}_m \mid b_j - a_j > T^{-1/2}\}$;
    offer price $x_m = (a_{i_m} + b_{i_m})/2$ for $\lceil 8\sqrt{T/k_m} \ln \delta^{-1} \rceil$ rounds;
    if $\overline{D}(a_{i_m}) - \overline{D}(x_m) < (k_m/T)^{1/4}/2$ **then** // undershooting
      if $\overline{D}(x_m) - \overline{D}(b_{i_m}) \geq (k_m/T)^{1/4}/2$ **then** // check for fake arms
        update $a_{i_m} \leftarrow x_m$, $\mathcal{K}_{m+1} \leftarrow \mathcal{K}_m$ and $k_{m+1} \leftarrow k_m$;
      else update $\mathcal{K}_{m+1} \leftarrow \mathcal{K}_m \setminus \{i_m\}$ and $k_{m+1} \leftarrow k_m$;
    else if $\overline{D}(a_{i_m}) - \overline{D}(x_m) \geq (k_m/T)^{1/4}/2$ **then** // overshooting
      if $\operatorname{sign}(a_i - x_m)\big(\overline{D}(a_i) - \overline{D}(x_m)\big) \geq (k_m/T)^{1/4}/2$ *for all* i **then** // new arms
        set $a_{k_m+1} \leftarrow x_m$, $b_{k_m+1} \leftarrow b_{i_m}$, $\mathcal{K}_{m+1} \leftarrow \mathcal{K}_m \cup \{k_m+1\}$ and $k_{m+1} \leftarrow k_m + 1$;
      update $b_{i_m} \leftarrow x_m$, $\mathcal{K}_{m+1} \leftarrow \mathcal{K}_m$ and $k_{m+1} \leftarrow k_m$;
  else denote the last macrostep by M and **break**;
end
run the UCB1 algorithm on the set of prices $\{a_j\}_{j \in \mathcal{K}_M}$; // bandit phase

Theorem 2 *If for some constant $c^\star > 0$ a seller algorithm has regret smaller than $c^\star\sqrt{T}$ on any instance of the stochastic dynamic pricing problem with at most three valuations, then there exists an instance with $\Delta = \Theta(1)$ on which the algorithm suffers regret $\Omega(\sqrt{T})$.*

This lower bound shows that $\sqrt{T}$ is best possible in the distribution-dependent case even when K is small and Δ is a constant. In Section 6 we show how regret bounds can be substantially better than $\sqrt{T}$ when the learner knows the value of the smallest drop in the demand curve.

5. Distribution-free bounds

In this section we focus on distribution-free bounds, i.e., bounds that do not depend on the demand curve. The regret bound we prove exceeds the theoretical lower bound stated in Section 4 by a constant term depending only on the distance between adjacent valuations.

Our Algorithm 1 works in two phases: a search phase and a bandit phase. In the search phase a binary search for all "relevant" valuations is performed. By the end of this phase, a tight estimate of all such valuations is determined with high probability. During the bandit phase a stochastic bandit algorithm is run on the estimated valuations. As it turns out, this simple scheme is enough to ensure an optimal $\sqrt{KT}$ convergence up to an additive constant independent of the distribution of buyer's valuations. Notably, the algorithm *does not* need to know K in advance. We call *macrostep* a block of consecutive rounds in which the same price is offered consistently. For each price x we denote by $\overline{D}(x)$ the fraction of accepted offers of x during the last macrostep in which x was offered. At the beginning of the search phase, our algorithm receives as input the time horizon T and a confidence parameter δ. The algorithm then proceeds in macrosteps of length $\lceil 8\sqrt{T/k_m} \ln \delta^{-1} \rceil$, where k_m is the total number of valuations discovered so far. The goal of the search phase is to approximately locate all *relevant* valuations, that is valuations v_i whose associated probability p_i is at least $\sqrt[4]{K/T}$.

Initially, all relevant valuations belong to $[a_1, b_1] = [0, 1]$. The search proceeds as long as there is at least an interval i containing relevant valuations with length larger than $T^{-1/2}$ (line 2). When such an interval i is selected at line 3, a macrostep of binary search is performed and the midpoint price x_m of $[a_i, b_i]$ is offered for $\lceil 8\sqrt{T/k_m} \ln \delta^{-1} \rceil$ rounds (line 4), thus obtaining an estimate of its demand. If the difference in demands (line 5) is smaller than $(k_m/T)^{-1/4}/2$ no new relevant valuation is detected. Before eliminating the lower half of the interval (line 7), a test designed to detect and remove *fake arms* is performed (line 6). We call fake arm an interval containing no relevant valuations. Fake arms might be inadvertently allocated when intervals are too wide. In that case, the comparison between two distant points may reveal a large difference in demands due to the presence of several nonrelevant valuations in between. If that happens, the fake arm is removed when the interval becomes small enough (line 8). When no significant difference is detected between the demands, all relevant valuations in $[a_i, b_i]$ remain in $[x_m, b_i]$ with high probability after the update. If, on the other hand, a difference between demands is detected (line 9), two things happen. First, a test is performed to detect possible new relevant valuations (line 10). If a new relevant valuation is spotted, a new interval $[x_m, b_i]$ is allocated. Second, the upper half of the interval $[a_i, b_i]$ is removed. If $[a_i, b_i]$ is split into $[a_i, x_m]$ and $[x_m, b_i]$, all relevant valuations are split between the two intervals. If $[a_i, b_i]$ is simply updated as $[a_i, x_m]$—since no significant difference was detected between the demands at x_m and b_i—all relevant valuations in $[a_i, b_i]$ remain in $[a_i, x_m]$ with high probability.

When all intervals become smaller than $T^{-1/2}$ (line 13), the search phase ends and all intervals $[a_i, b_i]$ are returned. At this point each relevant valuation is contained in one of the intervals with high probability. Therefore the algorithm has now access to $T^{-1/2}$-close approximations of all of them, and the bandit phase begins. In the bandit phase, the algorithm UCB1 (Auer et al., 2002a) is run on the set of left endpoints of the intervals (line 15).

Theorem 3 *If Algorithm 1 is run on an unknown number K of pairs $(v_1, p_1), \dots, (v_K, p_K)$ with input parameter $\delta = T^{-2}$, then its regret satisfies $R_T = \widetilde{\mathcal{O}}\big(\sqrt{KT}\big) + V(V+1)$ where $V = \max_{i \in \{1,\dots,K\}} v_k^4 (v_i - v_{i-1})^{-5}$.*

We actually prove a slightly improved bound, in which the constant $V(V+1)$ is replaced by the smaller term $K(v_K^4/v_1^4)\big(1 + (v_K^4/c^4)\big)$, where $c = \min_{i \in \{2,\dots,K\}} \{v_i - v_{i-1}\}$. To give a frame of reference, previously known upper bounds for discontinuous demand curves (den Boer and Keskin, 2018) are at best of order $\big(K^{20}/c^{18}\big)\sqrt{T}$, where v_1 is assumed to be bounded away from zero and K needs to be known in advance.

Proof sketch The probability the event $\mathcal{B}$ of making at least one mistake in at least one test of the form

$$\left(\overline{D}(x) - \overline{D}(y) < \sqrt[4]{\frac{k_m}{16T}} \ \wedge \ D(x) - D(y) \geq \sqrt[4]{\frac{k_m}{T}}\right) \text{ or } \left(\overline{D}(x) - \overline{D}(y) \geq \sqrt[4]{\frac{k_m}{16T}} \ \wedge \ D(x) = D(y)\right)$$

is at most $\mathcal{O}\big(\sqrt{K^3/T}\big)$ by Hoeffding's inequality. Assume now that the complement $\overline{\mathcal{B}}$ of $\mathcal{B}$ holds. Then, at most K binary searches are performed and—up to log factors—the regret increases by at most $\sum_{k=1}^{K} \sqrt{T/k} \leq \sqrt{T} \int_0^K x^{-1/2} \mathrm{d}x = 2\sqrt{KT}$. The additive term comes from the two following facts: if v_K is optimal and $v_K \notin \bigcup_{i \in \mathcal{K}_M} [a_i, b_i]$, then it has to have a higher revenue than v_1, which in turn gives $T < K(v_K/v_1)^4$; if any other v_j is optimal and $v_j \notin \bigcup_{i \in \mathcal{K}_M} [a_i, b_i]$, then it has to be better than v_{j+1} and v_1, which in turn gives $T < K v_K^8/(v_1 c)^4$. Finally, running UCB1 on at most K valuations which are $T^{-1/2}$-close approximations of $\{v_1, \dots, v_K\}$ increments the regret by at most $\widetilde{\mathcal{O}}(\sqrt{KT})$. ■

6. Distribution-dependent bounds

In this section we focus on distribution-dependent bounds, i.e., bounds that are parameterized in terms of the demand curve. The algorithm we design ignores the exact number of valuations, but it is given a lower bound γ on the smallest probability $p_{\min}$ (i.e., the smallest drop in the demand). This gives an upper bound on the number of valuations, since $\gamma \leq p_{\min}$ implies $K \leq 1/\gamma$. The regret bound we prove exceeds the distribution-dependent regret $(K \log T)/\Delta$ of standard stochastic bandits by a term of order $K(\log T)(\log\log T)/\gamma^2$. Note that if the number K of valuations (counting only those which are at least T^{-1} apart) is exactly known, it is easy to prove an excess regret bound of order $K\big((\log T)/p_{\min}\big)^2$ even when $p_{\min}$ (or a lower bound on it) is unknown. To see this, consider and algorithm that performs $\mathcal{O}(\log T)$ binary search steps for each one of the K valuations, repeating each step $\mathcal{O}\big((\log T)/\gamma^2\big)$ times and using a value of γ that decreases geometrically until all K valuations are found. A similar argument gives the same regret bound if K is not known exactly, but $\gamma \leq p_{\min}$ and $c \leq \min_k(p_k - p_{k-1})$ are both known.

We define a seller algorithm (Algorithm 2) in which a UCB-like strategy detects promising sub-intervals of $[0,1]$. These sub-intervals are then explored with an extension (to an unknown number of unknown valuations) of the "cautious search" for a single unknown valuation introduced by Kleinberg and Leighton (2003).

The main intuition is very simple: in order to estimate $D(x)$ we divide time in blocks (called again *macrosteps*) of equal length, and build an estimate $\overline{D}(x)$ by consistently posting the same price x within each block. In order to decide which arm i to use in each macrostep, we compute an upper confidence bound U_i on the average demand in the i-th interval, and then select the arm attaining the highest of such bounds. Algorithm 2 receives as input the time horizon T, a lower bound γ on $p_{\min} = \min_i p_i$, and a confidence parameter δ. Given these parameters, the number of macrosteps is defined as the biggest $M_\gamma \in \mathbb{N}$ satisfying $T \geq M_\gamma \lceil 8 \ln(\delta^{-1})/\gamma^2 \rceil$. The fraction of accepted offers of price x during the m-th macrostep (in which x is offered) is denoted by $\overline{D}_m(x)$. In line 3, the selected arm i_m is the one maximizing, over intervals $[a_i, b_i]$, the product $b_i U_i$. The quantity U_i is the upper confidence bound

$$U_i = \widehat{D}_m(i) + \frac{1}{b_i}\sqrt{\frac{\ln(\delta^{-1})}{N_m(i)}}$$

where $N_m(i)$ is $\lceil 8\gamma^{-2} \ln \delta^{-1} \rceil$ (if $i > 1$, which takes into account the macrostep in which interval i was allocated) plus the total number of times that i was picked in the first $m-1$ macrosteps, ignoring the steps occurring in all macrosteps when line 13 was executed. $\widehat{D}_m(i)$ is the fraction of accepted offers during these $N_m(i)$ steps.

The algorithm initially looks for valuation v_1, and then allocates searches for new valuations incrementally. Whenever a new value of the demand curve is observed, providing evidence for the existence of a i-th previously unseen valuation, an interval $[a_i, b_i]$ (which we associate with a bandit arm) and a step size ε_i are allocated. The interval $[a_i, b_i]$ estimates the smallest valuation v_i contained in it. By construction of the algorithm, v_i is never removed from $[a_i, b_i]$ (with high

Algorithm 2:

Input: Time horizon $T \in \mathbb{N}$, confidence parameter $\delta \in (0,1)$.
Initialization: set $\kappa_0 = 1$, $a_1 \leftarrow 0$, $b_1 \leftarrow 1$, $n_1 \leftarrow 1$, $\varepsilon_1 \leftarrow 1/2$, $\overline{D}(a_1) = 1$.

1 **for** $m = 1$ ***to*** M_γ **do**
2   set $\kappa_m \leftarrow \kappa_{m-1}$;
3   compute $i_m \leftarrow \arg\max_{i \le \kappa_m} b_i U_i$; `// greedy pick`
4   **if** $b_{i_m} - a_{i_m} \le 1/T$ **then** post a_{i_m}; `// if` $[a_{i_m}, b_{i_m}]$ `is tiny, keep playing` a_{i_m}
5   **else**
6   post $X_m = a_{i_m} + n_{i_m}\varepsilon_{i_m}$ for $\lceil 8\ln(\delta^{-1})/\gamma^2 \rceil$ rounds and compute $\overline{D}(X_m)$;
7   **if** $\overline{D}(a_{i_m}) - \overline{D}(X_m) < \gamma/2$ **then** `// no new valuations spotted`
8   **if** $X_m + \varepsilon_{i_m} < b_{i_m}$ **then** update $n_{i_m} \leftarrow n_{i_m} + 1$;
9   **else** update $a_{i_m} \leftarrow X_m$, $n_{i_m} \leftarrow 0$, $\varepsilon_{i_m} \leftarrow \varepsilon_{i_m}^2$; `// shrink` $[a_{i_m}, b_{i_m}]$
10   **else** (denoting $a_0 = \overline{D}(0) = 0$)
11   **if** $\forall i \neq i_m$, $\operatorname{sign}(a_i - X_m)\big(\overline{D}(a_i) - \overline{D}(X_m)\big) \ge \gamma/2$ **then** `// new valuation`
12   set $\kappa_m \leftarrow \kappa_{m-1} + 1$, $a_{\kappa_m} \leftarrow X_m$, $b_{\kappa_m} \leftarrow b_{i_m}$, $n_{\kappa_m} \leftarrow 1$, $\varepsilon_{\kappa_m} \leftarrow \varepsilon_{i_m}$;
13   update $a_{i_m} \leftarrow X_m - \varepsilon_{i_m}$, $b_{i_m} \leftarrow X_m$, $n_{i_m} \leftarrow 0$, $\varepsilon_{i_m} \leftarrow \varepsilon_{i_m}^2$; `// shrink` $[a_{i_m}, b_{i_m}]$
14 **end**

probability) when the interval shrinks. This implies that the more $[a_i, b_i]$ shrinks, the closer the upper bound $b_i U_i$ gets to the true revenue $v_i D(v_i)$.

Cautious searches are performed within each interval. At the beginning, all valuations belong to $[a_1, b_1] = [0,1]$. Whenever an interval is selected (line 3), a macrostep of cautious search is performed (lines 4-13). During a cautious search in $[a_i, b_i]$ with step size ε_i, the sequence of values $X_m = a_i + k\varepsilon_i$ for $k \in \{1, 2, \dots\}$ is posted $\lceil 8\gamma^{-2} \ln \delta^{-1} \rceil$ times each, until a change is spotted in the demand or X_m gets within ε_i of b_i. We spot a change in the demand curve when the difference at line 7 becomes bigger than $\gamma/2$. If X_m gets within ε_i of b_i before a change in the demand is discovered, the interval shrinks to $[X_m, b_{i_m}]$ and the step size is refined (line 9). Note that the shrunken interval contains (with high probability) all valuations that were in $[a_i, b_i]$ because no change in the demand was spotted. If, on the other hand, there is evidence of a change in the demand, then the interval shrinks to $[X_m - \varepsilon_i, X_m]$ and the step size is reduced (line 13). If at line 11 the newly discovered value of the demand differs from all previously detected demand values by at least $\gamma/2$, then a new interval $[X_t, b_i]$ is allocated (line 12). Otherwise, the new demand value matches (with high probability) the value of $D(b_i)$ and the shrunken interval contains all valuations that were in $[a_i, b_i]$. This process continues until the length of the feasible interval $[a_j, b_j]$ of the arm j with the highest $b_j U_j$ is less than $1/T$. Afterwards, the seller offers the same price a_j each time that j is selected (line 4).

As time goes by, the number κ of discovered valuations grows until possibly reaching the actual number of valuations K. Simultaneously, each estimate $b_i U_i$ converges to the revenue of the smallest valuation in the interval. After enough macrosteps, picking the interval i with the highest $b_i U_i$ becomes equivalent to choosing a $1/T$-approximation of an optimal valuation.

Without loss of generality, in the analysis of the algorithm we assume all valuations $v_1, \dots, v_K$ are at least $1/T$ apart. Let i_m be the index of the arm chosen at macrostep m (line 3). For any $k = \{1, \dots, K\}$, we denote $\mathcal{M}_k = \big\{ m \le M_\gamma \mid v_k \in [a_{i_m}, b_{i_m}] \big\}$.

Theorem 4 *If Algorithm 2 is run on an unknown number K of pairs $(v_1, p_1) \dots, (v_K, p_K)$ with input parameters $\gamma \le \min_k p_k$ and $\delta = T^{-2}$, then its regret satisfies*

$$R_T \le \sum_{i\colon \Delta_i > 0} \frac{4 \ln T}{\Delta_i} + \mathcal{O}\left(\frac{K \log T}{\gamma^2} \log\log T\right) .$$

To put things into perspective, previously known upper bounds for discontinuous demand curves (den Boer and Keskin, 2018) need at least the additional knowledge of K and are at best of order $\left(K^{24}/\gamma^{16}\right)\sqrt{T}$.

Proof sketch Without loss of generality, assume $M_\gamma B_\gamma = T$ where $B_\gamma \ge 8 \ln(\delta^{-1})/\gamma^2$ is the length of a macrostep. The probability of the event $\mathcal{B}$ of making at least one mistake in at least one test

$$\left(\left|\overline{D}_m(x) - \overline{D}_m(y)\right| < \tfrac{\gamma}{2} \ \wedge\ |D(x) - D(y)| \ge \gamma\right) \text{ or } \left(|\overline{D}_m(x) - \overline{D}_m(y)| > \tfrac{\gamma}{2} \ \wedge\ D(x) = D(y)\right) \tag{2}$$

is at most $4(K+1)M_\gamma\delta$ by Hoeffding's inequality. Assume now that the complement $\overline{\mathcal{B}}$ of $\mathcal{B}$ holds and denote by $v_{\mu(i)}$ the smallest valuation in $[a_i, b_i]$. Since $p_{\mu(i)} \ge \gamma$ by hypothesis, event $X_m > v_{\mu(i)}$ implies that the test in line 7 is false, and therefore line 13 is executed. This implies the following Lemma, that we prove before moving forward with the analysis of Algorithm 2.

Lemma 5 *Pick $k \in \{1, \dots, K\}$ and $n \in \{1, \dots, |\mathcal{M}_k|\}$. Let $[0, 1] \equiv I_1 \supseteq \dots \supseteq I_n \equiv [a'_n, b'_n]$ be the sequence of the first n intervals computed by n steps of a cautious search (see Appendix B) for the single valuation v_k with initial interval $[0, 1]$. If $\overline{\mathcal{B}}$ holds, then $a'_n \le a_{i_m}$ and $b'_n = b_{i_m}$, where m is the n-th smallest value in $\mathcal{M}_k$. Moreover, the price X_m offered by Algorithm 2 at macrostep m is equal to the n-th price offered by the cautious search for the single valuation v_k.*

Proof of Lemma 5 Fix a valuation v_k. Let A be Algorithm 2 and C be the cautious search for v_k. The proof is by induction on n. Since A and C both start with interval $[0, 1]$ and price $1/2$ the statement holds for $n = 1$. Now let m be the $(n+1)$-st smallest value in $\mathcal{T}_k$ and let s be the largest value in $\mathcal{T}_k$ that is smaller than m. Let $I_n \equiv [a'_n, b'_n]$ be the n-th interval computed by C. By induction, $a'_n \le a_{i_s}$, $b'_n = b_{i_s}$, and X_s is offered by both A and C. The only interesting case to discuss is when the test at line 7 is false. There are two subcases: if the test at line 11 is false, then it must be $X_s > v_k$. In this case C overshoots and the interval is updated exactly in the same way by C and A (see line 13). If the test at line 11 is true, then it must be $v_i < X_s \le v_k$. This is not an overshoot for C, so $I_{n+1} \equiv I_n$. A, however, creates a new interval $[a, b]$ —containing v_k— with $a = X_s$, $b = b_{i_s}$, and unchanged step size ε_{i_s}. The next time m this new interval is selected, the price X_m offered by A is the same as the price offered by C because the step size did not change. ■

We now continue with the proof of Theorem 4. Let $n_m(i)$ be the number of macrosteps (in the first $m-1$ macrosteps) where i was picked. Similarly, let $\mathrm{OS}_m(i)$ be the number of macrosteps (in the first $m-1$ macrosteps) when i was picked and $X_m > v_{\mu(i)}$. Then we have $N_m(i) = B_\gamma\big(n_m(i) - \mathrm{OS}_m(i)\big)$. Now note that $\widehat{D}_m(i)$ is the sample mean of a Bernoulli of parameter $D(v_{\mu(i)})$ because it is computed over $N_m(i)$ points sampled between a_i and $v_{\mu(i)}$. Fix a suboptimal valuation v_k and a macrostep m such that $\mu(i_m) = k$. Let $i^\star$ be such that $v^\star \in \left[a_{i^\star}, b_{i^\star}\right]$. Then, $i_m \ne i^\star$

implies

$$b_{i^\star} U_{i^\star} \le b_{i_m} U_{i_m} \iff \left(b_{i^\star} \widehat{D}_m(i^\star) + \sqrt{\frac{\ln(\delta^{-1})}{N_m(i^\star)}} \right) \le \left(b_{i_m} \widehat{D}_m(i_m) + \sqrt{\frac{\ln(\delta^{-1})}{N_m(i_m)}} \right)$$
$$\implies \left(v^\star \widehat{D}_m(i^\star) + \sqrt{\frac{\ln(\delta^{-1})}{N_m(i^\star)}} \right) \le \left(v_k \widehat{D}_m(i_m) + \frac{2B_\gamma}{N_m(i_m)} + \sqrt{\frac{\ln(\delta^{-1})}{N_m(i_m)}} \right)$$

where in the last implication we used Lemma 9 in Appendix B and $n_m(i_m) \ge N_m(i_m)/B_\gamma$. Observe that $\mathbb{E}\big[\widehat{D}_m(i^\star)\big] = D\big(v_{\mu(i^\star)}\big) \ge D(v^\star)$ and $\mathbb{E}\big[\widehat{D}_m(i_m)\big] = D(v_k)$. Moreover, the two quantities $\sqrt{\big(\ln(\delta^{-1})\big)/\big(N_m(i^\star)\big)}$ and $2B_\gamma/N_m(i_m) + \sqrt{\big(\ln(\delta^{-1})\big)/\big(N_m(i_m)\big)}$ play the role of upper confidence bounds for the estimates $v^\star \widehat{D}_m(i^\star)$ and $v_k \widehat{D}_m(i_m)$.

Therefore, we can apply a modification of the analysis of UCB1 (Auer et al., 2002a, Proof of Theorem 1) to K arms with reward expectations $v_k D(v_k)$ for $k \in \{1, \dots, K\}$, and such that the upper confidence bound for any suboptimal arm k is inflated by $2B_\gamma/N_m(i_m)$. (In fact Lemma 10 is stronger than what we need, because $v^\star$ always belongs to some interval $[a_{j^\star}, b_{j^\star}]$ but not all suboptimal valuations v_k are the smallest valuation of the interval $[a_{j_k}, b_{j_k}]$ they belong to.) Recalling that $B_\gamma \ge 8\big(\ln(\delta^{-1})\big)/\gamma^2$, Lemma 10 with $\alpha = 16$ gives

$$B_\gamma \, \mathbb{E}\left[\mathbb{I}\left\{\overline{\mathcal{B}}\right\} \sum_{m \colon \mu(i_m)=k} \mathbb{I}\left\{ i_m \ne i^\star \right\} \right] \le 1 + \left((\delta T)^2 + \frac{64}{\gamma^2} \right) 2K \ln(\delta^{-1}) + \sum_{k \colon \Delta_k > 0} \frac{4\ln(\delta^{-1})}{\Delta_k}$$

where $\Delta_k = v^\star D(i^\star) - v_k D(v_k) > 0$.[2] The fact that we prevent the algorithm from switching arms within each macrostep is not an issue. Indeed, the proof of the lemma works irrespective of whether the decision of pulling a different arm is made at every macrostep as opposed to every step. In particular, the proof establishes that after each suboptimal arm is selected order of $(\ln T)/\gamma^2$ times, corresponding to a constant number of macrosteps, the probability of pulling any suboptimal arm ever again becomes tiny, of order T^{-2}.

Applying Lemma 5 and Lemma 8 to the macrosteps of Algorithm 2, we obtain

$$\sum_{m \in \mathcal{M}_k} \big(r_t(v_k) - r_t(X_t) \big) \le \big(3 \ln\ln |\mathcal{M}_k| \big) + 8 \, . \tag{3}$$

2. The factor $\mathbb{I}\left\{\overline{\mathcal{B}}\right\}$ inside the expectation is needed to reduce the problem to an instance of a standard stochastic bandit. It can be conveniently dropped in the analysis of Lemma 10.

Therefore, with probability at least $1 - 4(K+1)M_\gamma\delta$, the regret over the T steps (recall that we repeatedly post the same price in each step of a macrostep) is bounded by

$$
\begin{aligned}
&B_\gamma \mathbb{E}\left[\sum_{m=1}^{M_\gamma}\Big(v^\star D(v^\star) - X_m D(X_m)\Big)\right] \\
&\quad \le B_\gamma \mathbb{E}\left[\sum_{k=1}^{K}\sum_{m\,:\,\mu(i_m)=k}\Big(v^\star D(v^\star) - v_k D(v_k)\Big) + \sum_{k=1}^{K}\sum_{m\in\mathcal{M}_k}\Big(v_k D(v_k) - X_m D(X_m)\Big)\right] \\
&\quad \le B_\gamma \sum_{k=1}^{K}\Delta_k \mathbb{E}\left[\mathbb{I}\left\{\overline{\mathcal{B}}\right\}\sum_{m\,:\,\mu(i_m)=k}\mathbb{I}\left\{i_m \neq i^\star\right\}\right] + T\mathbb{P}(\mathcal{B}) + B_\gamma\sum_{k=1}^{K}\big(3\ln\ln|\mathcal{M}_k| + 8\big) \\
&\qquad\qquad \text{(using (3))} \\
&\quad \le 1 + \left((\delta T)^2 + \frac{64}{\gamma^2}\right)2K\ln(\delta^{-1}) + \sum_{k:\,\Delta_k>0}\frac{4\ln(\delta^{-1})}{\Delta_k} + T\mathbb{P}(\mathcal{B}) + B_\gamma K\big(3\ln\ln T + 8\big)\ .
\end{aligned}
$$

Finally, in order to bound $T\mathbb{P}(\mathcal{B}) \le 4(K+1)TM_\gamma\delta = (K+1)(T\gamma)^2\delta/(2\ln\delta^{-1})$, set $\delta = T^{-2}$. ■

We conclude this section by discussing the case of at most two valuations. We design an algorithm with regret of order $\log(T)/\Delta + \log(T)\log\log(T)$, which is (up to the $\log\log$ term) as if the exact values of v_1 and v_2 were known in advance! This is achieved by leveraging some properties of the smallest and the biggest valuation. For example, any offer of a price lower or equal to v_1 is deterministically accepted and all offers above v_2 are always rejected. If on the other hand a price $x \in (v_1, v_2]$ is offered, the probability that that price is accepted is exactly p_2, which is enough to reconstruct the entire distribution (p_1, p_2) on $\{v_1, v_2\}$. Furthermore, the suboptimality gap Δ is always equal to $|v_1 - p_2 v_2|$.

Other than the result itself, we believe the techniques used in designing and analyzing the algorithm could be of interest on their own. Theorem 13 in particular gives a way to compute a high-probability multiplicative estimate of the unknown expectation $\mu > 0$ of any $[0,1]$-valued random variable using only $\mathcal{O}\big(\frac{1}{\mu}\big)$ samples. We now state the result. All the details about the algorithm and its subroutines, their pseudocodes, and the remaining theoretical results are presented in Appendix D.

Theorem 6 *If Algorithm 7 (see Appendix D) is run with input parameter $\delta = T^{-2}$ on an unknown instance (v_1, p_1) and (v_2, p_2), then its regret satisfies $R_T = \mathcal{O}\big(\log(T)/\Delta + (\log T)(\log\log T)\big)$, where the first term is zero when $\Delta = |p_2 v_2 - v_1|$ is zero.*

7. Open problems

Our work leaves some interesting questions open. Can we prove a distribution-free upper bound of order $\sqrt{KT}$ that does not depend on the locations of buyers' valuations? Can we prove a distribution-dependent upper bound without any prior knowledge at all for K larger than two? Can we obtain a $\sqrt{KT}$ regret bound in the nonstochastic setting when $K \ge 2$ and all valuations are unknown?

Acknowledgments

Nicolò Cesa-Bianchi and Tommaso Cesari gratefully acknowledge the support of Criteo AI Lab through a Faculty Research Award. Vianney Perchet acknowledges the support of the FMJH Program Gaspard Monge in Optimization and operations research (supported in part by EDF) and the CNRS through the PEPS program.

References

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006. ACM, 2014.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2013.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. on Computing*, 32(1):48–77, 2002b.

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):4, 2015.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

Quentin Berthet and Vianney Perchet. Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems*, pages 2222–2231, 2017.

Avrim Blum and Jason D. Hartline. Near-optimal online auctions. In *Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1156–1163. Society for Industrial and Applied Mathematics, 2005.

Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2-3):137–146, 2004.

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

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134, 2013.

Sebastien Bubeck, Nikhil R Devanur, Zhiyi Huang, and Rad Niazadeh. Online auctions and multi-scale online learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 497–514. ACM, 2017.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Tommaso R. Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. *arXiv preprint arXiv:1807.03288*, 2018.

Alexandre Cotarmanac'h. Auction mechanics: A buyer's perspective. Blogpost, 2017. URL https://goo.gl/7Nymnt.

Arnoud V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

Arnoud V. den Boer and N. Bora Keskin. Discontinuous demand functions: Estimation and pricing. Technical report, Available at SSRN, 2018. URL https://ssrn.com/abstract=3003984.

Nikhil Devanur, Yuval Peres, and Balasubramanian Sivan. Perfect Bayesian equilibria in repeated sales. In *Proceedings of the 26th annual ACM-SIAM symposium on Discrete algorithms*, pages 983–1002. SIAM, 2014.

Richard M. Karp and Robert Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 881–890. Society for Industrial and Applied Mathematics, 2007.

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

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory*, pages 1–9, 2009.

R. Preston McAfee. The design of advertising exchanges. *Review of Industrial Organization*, 39 (3):169–185, 2011.

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

Aleksandrs Slivkins and Assaf Zeevi. Dynamic Pricing Under Model Uncertainty. Tutorial given at the 16th ACM Conference on Economics and Computation, 2015.

M. Wedel and W.A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. International Series in Quantitative Marketing. Springer US, 2012.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583, 2016.

Appendix A. Lower Bounds

In this section we prove the lower bounds (Theorems 1 and 2) stated in Section 4. Kleinberg and Leighton (2003) showed that $R_T = \Omega(T^{2/3})$ if $T \le K^3$ by building a distribution over a set of ε-spaced valuations $v_1, \dots, v_K \in \left[\frac{1}{2}, 1\right]$. A key technical property needed in their proof is that $\mathrm{KL}\left(\frac{1}{2v}, \frac{9}{10}\frac{1}{2v} + \frac{1}{10}\frac{1}{2(v-\varepsilon)}\right) \le c\varepsilon^2$ for some constant c independent of ε and for all $v \ge 3/4$. A long and tedious computation shows that such construction only works if K is large compared to T.

Lemma 7 *For all $K \ge 1$, for all $\varepsilon \in \left(0, \frac{1}{2K}\right]$, and for all $k \in \{1, \dots, K\}$, denoting $v = \frac{1}{2} + k\varepsilon$,*

$$\mathrm{KL}\left(\frac{1}{2v} \,\middle\|\, \frac{9}{10}\frac{1}{2v} + \frac{1}{10}\frac{1}{2(v-\varepsilon)}\right) > \frac{\varepsilon}{800k} .$$

Even if the technique used by Kleinberg and Leighton (2003) fails in our setting, it is still possible to prove the following lower bound by changing some key aspects of their analysis, which in turn is based on the lower bound analysis of (Auer et al., 2002b). First, valuations need to be distanced as much as possible —this is the exact opposite of their construction, where valuations were placed ε-close to each others. Second, the base distribution is only perturbed by an appropriate small constant. Third, the "good valuation" is drawn from a sensible proper subset of valuations. We now restate and prove Theorem 1.

Theorem 1 *For any number of valuations $K \ge 3$ and all time horizons $T \ge K^3$ there exist K pairs $\big(v_1, p(v_1)\big), \dots, \big(v_K, p(v_K)\big)$ such that the expected regret of any pricing strategy satisfies*

$$R_T \ge \frac{1}{375}\sqrt{KT} .$$

Proof For notational convenience, fix $K \ge 2$ and define the set $\{v_0, \dots, v_K\}$ of $K+1$ valuations by

$$v_i = \frac{1}{2} + \frac{i}{2K}, \qquad \forall i \in \{0, \dots, K\} .$$

Define the distribution p_0 on $\{v_0, \dots, v_K\}$ of the random variable V_0 by

$$\mathbb{P}(V_0 \ge v) = \sum_{i\colon v_i \ge v} p_0(v_i) = \frac{1}{2v}, \qquad \forall v \in \{v_0, \dots, v_K\} .$$

With this choice of demand curve, $v\mathbb{P}(V_0 \ge v) = 1/2$, i.e., each valuation v has the same expected revenue. Furthermore, the distribution $v \mapsto p_0(v)$ satisfies the following: $p_0(v_0) = \frac{1}{K+1}$; p_0 decreases monotonically on $\{v_0, \dots, v_{K-1}\}$, $p_0(v_{K-1}) = \frac{1}{2K-1}$, and $p_0(v_K) = 1/2$. Therefore

$$\frac{1}{2K} \le p_0(v) \le \frac{1}{K}, \qquad \forall v \in \{v_0, \dots, v_{K-1}\} . \tag{4}$$

Now, for each $j \in \big\{\lceil K/2 \rceil, \dots, K\big\}$, define the distribution p_j by slightly lowering the probability of v_{j-1} and upping the probability of v_j by the same amount:

$$p_j(v_i) = \begin{cases} p_0(v_i), & i \in \{0, \dots, K\} \setminus \{j-1, j\}, \\ (1 - 4K\varepsilon)p_0(v_{j-1}), & i = j-1, \\ p_0(v_j) + 4K\varepsilon p_0(v_{j-1}), & i = j, \end{cases} \tag{5}$$

where $\varepsilon \in \left(0, \frac{1}{40}\right)$ is a small constant determined below. Note that if the buyers' valuations were distributed as p_j, all valuations $v \neq v_j$ would have expected revenue $\frac{1}{2}$, but v_j whould have expected revenue at least $\frac{1}{2} + \varepsilon$ because of (4) and (5). In order to define the distribution of buyers' valuations $V = \left(V_1, \dots, V_T\right)$, let J be uniformly distributed over $\left\{\lceil K/2 \rceil, \dots, K\right\}$ (that is, the set of indices $i \in \{1, \dots, K\}$ such that $v_i \geq \frac{3}{4}$). The value of J will give the "good valuation", that is the valuation with the highest expected revenue. For all t, the distribution of V_t is determined by

$$\mathbb{P}\left(V_t = v_i \mid J = j\right) = p_j(v_i), \qquad \forall i \in \{0, \dots, K\}, \forall j \in \left\{\lceil K/2 \rceil, \dots, K\right\} .$$

Denoting the seller's randomized strategy by $X = (X_1, \dots, X_T)$ and applying Fubini's theorem, we obtain

$$R_T = \max_{k \in \{0, \dots, K\}} \mathbb{E}_X \mathbb{E}_{J,V} \left[\sum_{t=1}^{T} r_t(v_k) - \sum_{t=1}^{T} r_t(X_t) \right] .$$

According to the previous identity, we can (an will!) lower bound the internal expectation assuming that the seller's strategy is deterministic. Furthermore, assume that the seller's pricing strategy only offers prices in $\{v_{\lceil K/2 \rceil}, \dots, v_K\}$ —since it is counterproductive to offer a price outside of it as all other valuations ($v_1, \dots, v_{\lceil K/2 \rceil - 1}$ in particular) have smaller expected revenues. Now let N_i be the number of times the seller offer valuation v_i,

$$N_i = \sum_{t=1}^{T} \mathbb{I}\{X_t = v_i\} .$$

By construction, each time the seller picks the "good valuation", no regret is accrued; all other times at least ε is lost. Therefore

$$\mathbb{E}_{J,V} \left[\sum_{t=1}^{T} r_t(v_k) - \sum_{t=1}^{T} r_t(X_t) \right] \geq \varepsilon\left(T - \mathbb{E}_{J,V}[N_J]\right) . \tag{6}$$

Denote by Y_t the Bernoulli random variable $\mathbb{I}\{V_t \geq X_t\}$ which is 1 if and only if the t-th buyer accepted the price offered, $Y^t = (Y_1, \dots, Y_t)$, and $Y = Y^T$. Denote by q_0 the distribution of Y if buyer's valuations were distributed as p_0 and by q_i the distribution of Y if buyer's valuations were distributed as p_i. For any deterministic function $f \colon \{0,1\}^T \to [0, M]$,

$$\begin{aligned}
\mathbb{E}_V\left[f(Y) \mid J = i\right] - \mathbb{E}_0[f(Y)] &= \sum_{b^T \in \{0,1\}^T} f(b^T)\left(q_i(b^T) - q_0(b^T)\right) \\
&\leq \sum_{\substack{b^T \in \{0,1\}^T \\ q_i(b^T) > q_0(b^T)}} f(b^T)\left(q_i(b^T) - q_0(b^T)\right) \\
&\leq M \sum_{\substack{b^T \in \{0,1\}^T \\ q_i(b^T) > q_0(b^T)}} \left(q_i(b^T) - q_0(b^T)\right) \\
&\leq M \sqrt{\frac{1}{2} \mathrm{KL}(q_0 \parallel q_i)}
\end{aligned}$$

where $\mathbb{E}_0$ is the expectation with respect to distribution p_0 and in the last step we used Pinsker's inequality. Let $q_i(b_t \mid b^{t-1}) = p_i\left(Y_t = b_t \mid Y_1 = b_1, \dots, Y_{t-1} = b_{t-1}\right)$ and let $q_0(b_t \mid b^{t-1})$ be defined similarly. By the chain rule of the relative entropy

$$\begin{aligned}
\mathrm{KL}(q_0 \parallel q_i) &= \sum_{t=1}^{T} q_0(b^{t-1}) \sum_{b^{t-1} \in \{0,1\}^{t-1}} \mathrm{KL}\big(q_0(b_t \mid b^{t-1}) \parallel q_i(b_t \mid b^{t-1})\big) \\
&= \sum_{t=1}^{T} q_0(b^{t-1}) \sum_{b^{t-1} :\, X_t(b^{t-1}) \neq v_i} \underbrace{\mathrm{KL}\big(q_0(b_t \mid b^{t-1}) \parallel q_i(b_t \mid b^{t-1})\big)}_{=0} \\
&\quad + \sum_{t=1}^{T} q_0(b^{t-1}) \sum_{b^{t-1} :\, X_t(b^{t-1}) = v_i} \mathrm{KL}\big(q_0(b_t \mid b^{t-1}) \parallel q_i(b_t \mid b^{t-1})\big)
\end{aligned}$$

where the relative entropy is zero when $X_t \neq v_i$ because in that case $p_i(Y_t = 1) = p_0(Y_t = 1)$. If on the other hand, $X_t = v_i$, for all $v_i \geq \frac{3}{4}$,

$$\mathrm{KL}\big(q_0(b_t \mid b^{t-1}) \parallel q_i(b_t \mid b^{t-1})\big) = \mathrm{KL}\left(\frac{1}{2v_i} \,\Big\|\, \frac{1}{2v_i} + 4K\varepsilon p_0(v_{j-1})\right) \leq 108\varepsilon^2$$

where the last inequality follows by (4) and $\mathrm{KL}\big(x \parallel x + \alpha\big) \leq \alpha^2 (x+\alpha)^{-1}(1-x-\alpha)^{-1}$, with $x = \frac{1}{2v_i} \in \left[\frac{1}{2}, \frac{2}{3}\right]$ and $\alpha = 4K\varepsilon p_0(v_{j-1}) \in [2\varepsilon, 4\varepsilon]$. Therefore

$$\mathrm{KL}(q_0 \parallel q_i) \leq 108\varepsilon^2 \sum_{t=1}^{T} q_0(b^{t-1}) \sum_{b^{t-1} :\, X_t(b^{t-1}) = v_i} 1 = 108\varepsilon^2 \sum_{t=1}^{T} p_0(X_t = v_i) = 108\varepsilon^2 \mathbb{E}_0[N_i] \, ,$$

where again, $\mathbb{E}_0$ is the expectation with respect to distribution p_0. This gives

$$\mathbb{E}_V[f(Y) \mid J = i] \leq \mathbb{E}_0[f(Y)] + \varepsilon M \sqrt{54\mathbb{E}_0[N_i]} \, .$$

Then, being for any deterministic online pricing strategy the random variable N_i a deterministic function of Y, $\mathbb{E}_V[N_i \mid J = i] \leq \mathbb{E}_0[N_i] + \varepsilon T \sqrt{54\mathbb{E}_0[N_i]}$. Thus, using Jensen inequality, $\mathbb{E}_{J,V}[N_i] \leq \mathbb{E}_J \mathbb{E}_0[N_J] + \varepsilon T \sqrt{54\mathbb{E}_J\mathbb{E}_0[N_J]}$. Using again Jensen inequality, Fubini's Theorem, and inequality (6),

$$\mathbb{E}_{J,V}\mathbb{E}_X\left[\sum_{t=1}^{T} r_t(v_k) - \sum_{t=1}^{T} r_t(X_t)\right] \geq \varepsilon\left(T - \mathbb{E}_J\mathbb{E}_0\mathbb{E}_X[N_J] - \varepsilon T\sqrt{54\mathbb{E}_J\mathbb{E}_0\mathbb{E}_X[N_J]}\right) \, .$$

Since $\sum_{i=\lceil K/2 \rceil}^{K} N_i = T$, we also have $\sum_{i=\lceil K/2 \rceil}^{K} \mathbb{E}_0\mathbb{E}_X[N_i] = T$. Using the fact that $K - \lceil K/2 \rceil + 1 \geq \max\{3/2, K/2\}$, this implies

$$\mathbb{E}_J\mathbb{E}_0\mathbb{E}_X[N_J] = \frac{1}{K - \lceil K/2 \rceil + 1} \sum_{i=\lceil K/2 \rceil}^{K} \mathbb{E}_0\mathbb{E}_X[N_i] \leq \min\left\{\frac{2}{3}, \frac{2}{K}\right\} T \, .$$

Putting everything together, we get

$$R_T \geq \varepsilon\left(T - \frac{2}{3}T - \varepsilon T\sqrt{\frac{108T}{K}}\right) = \varepsilon T\left(\frac{1}{3} - \varepsilon\sqrt{\frac{108T}{K}}\right) \, ,$$

which picking $\varepsilon = \frac{1}{6\sqrt{108}}\sqrt{K/T}$ so that $\varepsilon\sqrt{108T/K} = 1/6$, gives

$$R_T \geq \frac{1}{375}\sqrt{KT}$$

as desired. ■

We now move on to proving Theorem 2, that we restate below.

Theorem 2 *If for some constant $c^\star > 0$ a seller algorithm has regret smaller than $c^\star\sqrt{T}$ on any instance of the stochastic dynamic pricing problem with at most three valuations, then there exists an instance with $\Delta = \Theta(1)$ on which the algorithm suffers regret $\Omega(\sqrt{T})$.*

Proof We consider two instances. The first has $\Delta = \frac{1}{4}$ and the second has $\Delta = \mathcal{O}(1/\sqrt{T})$. We prove that if the algorithm has regret $\mathcal{O}(\sqrt{T})$ on both instances, then it must have regret $\Omega(\sqrt{T})$ on the first instance. The two instances are defined as follows.

Instance 1

$v_1^{(1)} = 0$	$D^{(1)}(0) = 1$	$r^{(1)}(0) = 0$
$v_2^{(1)} = \frac{1}{2}$	$D^{(1)}\left(\frac{1}{2}\right) = \frac{1}{2}$	$r^{(1)}\left(\frac{1}{2}\right) = \frac{1}{4}$

Instance 2

$v_1^{(2)} = 0$	$D^{(2)}(0) = 1$	$r^{(2)}(0) = 0$
$v_2^{(2)} = \frac{1-\eta}{2}$	$D^{(2)}\left(\frac{1-\eta}{2}\right) = \frac{1}{2} + \eta$	$r^{(2)}\left(\frac{1-\eta}{2}\right) = \frac{1+\eta-2\eta^2}{4}$
$v_3^{(2)} = \frac{1}{2}$	$D^{(2)}\left(\frac{1}{2}\right) = \frac{1}{2}$	$r^{(2)}\left(\frac{1}{2}\right) = \frac{1}{4}$

In Instance 1 the optimal price is $v_2^{(1)} = \frac{1}{2}$ with revenue $\frac{1}{4}$. In Instance 2 the optimal price is $v_2^{(2)} = \frac{1-\eta}{2}$ with revenue $\frac{1+\eta-2\eta^2}{4} \geq \frac{1}{4} + \frac{\eta}{8}$ for $\eta \leq \frac{1}{4}$. Without loss of generality, we can assume that the seller algorithm only posts prices in the set $\left\{0, \frac{1-\eta}{2}, \frac{1}{2}\right\}$. Let $N_\eta(t)$ be the number of times that the price $\frac{1-\eta}{2}$ is posted and let $\nu_t^{(i)}$ be the law of observed rewards up to time t in Instance $i \in \{1, 2\}$. Since prices 0 and $\frac{1}{2}$ are uninformative (because demand and revenue do no change across the two instances), it follows from standard calculations that the KL divergence between $\nu_t^{(1)}$ and $\nu_t^{(2)}$ is upper bounded by the KL between two Bernoulli of parameter $\frac{1}{2}$ and $\frac{1}{2} + \eta$ times the expected number of times v_2 is chosen under Instance 1,

$$\mathrm{KL}\big(\nu_t^{(1)} \,\|\, \nu_t^{(2)}\big) \leq \mathrm{KL}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + \eta\right) \mathbb{E}_1\big[N_\eta(t)\big] \leq 4\eta^2\, \mathbb{E}_1\big[N_\eta(t)\big] \quad \text{ if } \eta \leq \frac{1}{4}$$

where $\mathbb{E}_1$ denotes expectation under Instance 1. Let $R_T^{(i)}$ be the regret under Instance $i \in \{1, 2\}$. Since $r^{(1)}\left(\frac{1-\eta}{2}\right) = \frac{1-\eta}{2} D^{(1)}\left(\frac{1-\eta}{2}\right) = \frac{1-\eta}{4}$, we have $R_T^{(1)} \geq \frac{\eta}{4}\mathbb{E}_1\big[N_\eta(T)\big]$. Using the assumption that the seller's algorithm has a regret smaller than $c^\star\sqrt{T}$, and adapting an argument of Bubeck et al. (2013, Proof of Theorem 5), we can write

$$\frac{\eta}{4}\frac{T}{4} \exp\Big(-4\eta^2\mathbb{E}_1\big[N_\eta(T)\big]\Big) \leq \max\Big\{R_T^{(1)}, R_T^{(2)}\Big\} \leq c^\star\sqrt{T}\,.$$

Hence, for $\eta = \frac{32c^\star}{\sqrt{T}}$, it must hold that $\mathbb{E}_1\big[N_\eta(t)\big] \geq \frac{\ln 2}{4\eta^2}$, which implies that $R_T^{(1)} \geq \frac{\ln 2}{512c^\star}\sqrt{T}$. ■

Theorem 2 can be extended to the case when K is known to the seller. This can be done by adding an extra valuation $v_3^{(1)} > v_2^{(1)}$ to Instance 1 which has either vanishing probability p_3 or vanishing distance $v_3^{(1)} - v_2^{(1)}$ from $v_2^{(1)}$. (In the latter case the value of $v_3^{(1)}$ depends on the algorithms.) In both cases the seller algorithm is unlikely to detect the presence of this extra valuation, and a slight modified proof of Theorem 2 can be applied.

Algorithm 3: Cautious search

Input: Time horizon $T \in \mathbb{N}$.

Initialization: set $a \leftarrow 0$, $b \leftarrow 1$, $n \leftarrow 1$, $\varepsilon \leftarrow 1/2$.

for $t \in \{1, \dots T\}$ **do**
  post $X_t = a + n\varepsilon$ and get feedback $Z_t = \mathbb{I}\{X_t \leq v\}$;
  if $Z_t = 1$ **then** // undershooting
    if $X_t + \varepsilon < b$ **then** update $n \leftarrow n + 1$;
    else update $a \leftarrow X_t$, $n \leftarrow 1$, $\varepsilon \leftarrow \varepsilon^2$; // shrink the interval
  else if $Z_t = 0$ **then** // overshooting
    update $a \leftarrow X_t - \varepsilon$, $b \leftarrow X_t$, $n \leftarrow 1$, $\varepsilon \leftarrow \varepsilon^2$; // shrink the interval
end

Appendix B. Cautious search

Kleinberg and Leighton (2003) were first to introduce a "cautious search" as an optimal algorithm for posted price with a single unknown evaluation. Similarly, our cautious search (Algorithm 3) proceeds in phases $s \in \{1, 2, \dots\}$ in which an interval $[a_s, b_s]$ (initialized to $[0, 1]$) and a step size ε_s (initialized to $1/2$) are maintained. In a given phase s of the algorithm, prices $a_s + \varepsilon_s$, $a_s + 2\varepsilon_s$, $a_s + 3\varepsilon_s$, $\dots$ are posted until one of them, say X_s, becomes bigger than the hidden evaluation (overshooting). At this point a new phase begins: the interval becomes $[a_{s+1}, b_{s+1}] = [X_s - \varepsilon_s, X_s]$, and the new step size becomes $\varepsilon_{s+1} = \varepsilon_s^2$. This process continues until the length of the interval is less than $1/T$. Then the left endpoint of the interval is picked for all remaining rounds. We now state two lemmas about the behavior of cautious search. The first one is proven in (Kleinberg and Leighton, 2003, Theorem 2.1).

Lemma 8 *The regret of Algorithm 3 satisfies* $\mathbb{E}\left[\sum_{t=1}^{T} r_t(v) - \sum_{t=1}^{T} r_t(X_t)\right] \leq 3 \ln\ln(T) + 8$. *Moreover, the number of overshootings is upper bounded by* $\log\log T$.

The second lemma bounds the size of the interval as a function of the number of steps.

Lemma 9 *For all m, the size of an interval $[a_s, b_s]$ after m steps of Algorithm 3 satisfies*

$$b_s - a_s \leq \frac{2}{m} \, .$$

Proof The worst case happens when the sequence $(b_1 - a_1, b_2 - a_2, \dots)$ of interval endpoints takes values

$$\left(1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \dots, \frac{1}{2^{2^n}}, \dots, \frac{1}{2^{2^n}}, \dots\right) \tag{7}$$

where the general term $1/2^{2^n}$ is repeated 2^{2^n} times. It is then sufficient to show that the inequality holds for all values before a switch. Formally, that for all $n \in \{0, 1, 2, \dots\}$

$$\frac{1}{2^{2^n}} \leq \frac{2}{2 + \sum_{j=0}^{n} 2^{2^j}} \qquad \text{or, equivalently,} \qquad 2 + \sum_{j=0}^{n} 2^{2^j} \leq 2 \cdot 2^{2^n} \, .$$

We prove this by induction on n. The case $n = 0$ is trivial. If the inequality holds for $n \in \{0, 1, \dots\}$, then

$$2 + \sum_{j=0}^{n+1} 2^{2^j} = 2 + \sum_{j=0}^{n} 2^{2^j} + 2^{2^{n+1}} \le 2 \cdot 2^{2^n} + 2^{2^{n+1}} = 2^{2^n}\left(2 + 2^{2^n}\right) \le 2 \cdot 2^{2^{n+1}} .$$

This concludes the proof. ■

The previous bound is unimprovable. Indeed in scenario (7), for all $n \in \{0, 1, \dots\}$

$$2^{2^n} < 2 + \sum_{j=0}^{n} 2^{2^j} \le 2 \cdot 2^{2^n}$$

and the second inequality is actually an equality for $n = 0$.

Appendix C. UCB with inflated confidence bounds

In this section we prove a regret bound for UCB1 run with an oracle that systematically inflates the upper confidence bounds for suboptimal arms.

Lemma 10 *Consider a stochastic bandit problem with K arms, i.i.d. rewards $X_t(k) \in [0, 1]$ from each arm k, and average rewards $\mu_1, \dots, \mu_K$. Let $\Delta_k = \mu^\star - \mu_k$ where $\mu^\star = \mu_{i^\star}$ and $i^\star$ is the index of an optimal arm. Consider a UCB policy that at round t selects arm I_t defined by*

$$I_t = \operatorname*{arg\,max}_{k \in \{1,\dots,K\}} \left(\widehat{X}_t(k) + c\big(N_t(k), k\big)\right)$$

(ties broken arbitrarily), where $\widehat{X}_t$ is the sample average of the rewards obtained from arm k over the $N_t(k)$ times when the arm was chosen in rounds $1, \dots, t-1$ (initially, $N_1(k) = 0$ for all arms) and

$$c(s,k) = \begin{cases} \dfrac{\alpha \ln(\delta^{-1})}{\gamma^2 s} + \sqrt{\dfrac{\ln(\delta^{-1})}{s}} & \text{if } k \text{ is suboptimal,} \\ \sqrt{\dfrac{\ln(\delta^{-1})}{s}} & \text{otherwise,} \end{cases}$$

with $\alpha \ge 0$ and $c(s,k) = +\infty$ if $s = 0$. Then

$$R_T \le 1 + \left(2(\delta T)^2 + \frac{8\alpha \ln(\delta^{-1})}{\gamma^2}\right) K + \sum_{k \colon \Delta_k > 0} \frac{4 \ln(\delta^{-1})}{\Delta_k} .$$

Proof Pick any suboptimal arm k and $t \ge 2$. Note that $I_t = k$ implies

$$\widehat{X}_t(i^\star) + c\big(N_t(i^\star), i^\star\big) \le \widehat{X}_t(k) + c\big(N_t(k), k\big)$$

which in turn imply

$$\left(\widehat{X}_t(i^\star) \le \mu^* - c\big(N_t(i^\star), i^\star\big)\right) \vee \left(\widehat{X}_t(k) \ge \mu_k + c\big(N_t(k), k\big)\right) \vee \left(c\big(N_t(k), k\big) > \Delta_k/2\right) .$$

Using standard Chernoff bounds, we can write

$$\sum_{t=2}^{T} \mathbb{P}\Big(\widehat{X}_t(i^\star) \le \mu^* - c\big(N_t(i^\star), i^\star\big)\Big) \le \sum_{t=2}^{T} \mathbb{P}\Big(\exists s \in \{1, \dots, t-1\},\ \widehat{X}_t(i^\star) \le \mu^* - c(s, i^\star)\Big)$$
$$\le \sum_{t=2}^{T} \sum_{s=1}^{t-1} \exp\left(-2s \frac{\ln(\delta^{-1})}{s}\right) \le T^2 \delta^2$$

and

$$\sum_{t=2}^{T} \mathbb{P}\Big(\widehat{X}_t(k) \ge \mu_k + c\big(N_t(k), k\big)\Big) \le \sum_{t=2}^{T} \mathbb{P}\Big(\exists s \in \{1, \dots, t-1\},\ \widehat{X}_t(k) \ge \mu_k + c(s, k)\Big)$$
$$\le \sum_{t=2}^{T} \mathbb{P}\left(\exists s \in \{1, \dots, t-1\},\ \widehat{X}_t(k) \ge \mu_k + \sqrt{\frac{2\ln(\delta^{-1})}{s}}\right)$$
$$\le \sum_{t=2}^{T} \sum_{s=1}^{t-1} \exp\left(-2s \frac{\ln(\delta^{-1})}{s}\right) \le T^2 \delta^2 \ .$$

It remains to control $\mathbb{I}\big\{c\big(N_t(k), k\big) > \Delta_k/2\big\}$ when $I_t = k$. We now show that

$$\sum_{t=2}^{T} \mathbb{I}\big\{c\big(N_t(k), k\big) > \Delta_k/2\big\} \le 4\left(\frac{2\alpha}{\gamma^2 \Delta_k} + \frac{1}{\Delta_k^2}\right) \ln(\delta^{-1}) \ .$$

If k is chosen $s > 0$ times in the first $t-1$ steps, then $N_t(k) = s$. Thus $c(s, k) > \Delta_k/2$ implies

$$\frac{\alpha \ln(\delta^{-1})}{\gamma^2 s} + \sqrt{\frac{\ln(\delta^{-1})}{s}} > \frac{\Delta_k}{2} \ . \tag{8}$$

We now prove that s must be smaller than

$$4\left(\frac{2\alpha}{\gamma^2 \Delta_k} + \frac{1}{\Delta_k^2}\right) \ln(\delta^{-1})$$

for this to happen. If $\alpha = 0$ this is trivially true. To see that this still true for $\alpha > 0$, note that with this assumption (8) is equivalent to

$$\sqrt{\frac{\ln(\delta^{-1})}{s}} > \frac{-1 + \sqrt{1 + 2\Delta_k \alpha/\gamma^2}}{2\alpha/\gamma^2} \ .$$

Setting $x = 2\Delta_k \alpha/\gamma^2 > 0$, solving for s, and using $\frac{x}{\sqrt{1+x}-1} \le 2\sqrt{1+x}$ proves the claim. Picking $\delta = T$, the regret is therefore bounded as follows

$$R_T \le 1 + \sum_{k:\ \Delta_k > 0} \Delta_k \sum_{t=2}^{T} \mathbb{P}(I_t = k) \le 1 + 2KT^2\delta^2 + \frac{8\alpha K}{\gamma^2} \ln(\delta^{-1}) + \sum_{k:\ \Delta_k > 0} \frac{4\ln(\delta^{-1})}{\Delta_k} \ .$$

This concludes the proof. ∎

Algorithm 4: Noisy Cautious Search

Input: confidence parameter $\delta \in (0,1)$, valuation index $i \in \{1,2\}$, lower bound $\gamma_i \in (0,1)$.
Initialization: set $a \leftarrow 0$, $b \leftarrow 1$.

for $s \in \big\{0,1,\dots,\lceil \log_2 \log_2 T \rceil\big\}$ **do** // phases
  set $n \leftarrow 1$, $\varepsilon_s \leftarrow 2^{-2^s}$, $\overline{D} \leftarrow 1$;
  while $(a + n\varepsilon_s < b) \wedge \big[(i = 1 \wedge \overline{D} = 1) \vee (i = 2 \wedge \overline{D} > 0)\big]$ **do**
    offer price $a + n\varepsilon_s$ for $\big\lceil \ln(\delta)/\ln(1-\gamma_i) \big\rceil$ rounds; // a macrostep
    update $n \leftarrow n+1$ and the sample mean $\overline{D}$ of $D\big(a + (n-1)\varepsilon_s\big)$;
  end
  update $a \leftarrow a + (n-1)\varepsilon_s$, $b \leftarrow a + n\varepsilon_s$;
end
offer a for all remaining rounds;

Appendix D. Two valuations

In this section we present all key results related to subroutines of Algorithm 7 and give a formal proof of Theorem 6.

Noisy Cautious Search

This procedure is a variant of the cautious search described in Appendix B. It identifies the location of a valuation v_i with high probability and low regret whenever a lower bound γ_i on its probability p_i is known in advance. During the search, each price is posted for $\big\lceil \ln(\delta)/\ln(1-\gamma_i) \big\rceil$ times in a row, where δ is a confidence parameter. We call such a sequence of consecutive rounds a *macrostep*. For $i = 1$, we say that a macrostep is a *failure* if at least one price is rejected, it is a *success* if all prices are accepted, and the algorithm makes a *mistake* if the macrostep is a success but the price offered is strictly bigger than v_1. For $i = 2$, we say that a macrostep is a *failure* if no price is accepted, it is a *success* if at least one price is accepted, and the algorithm makes a *mistake* if the macrostep is a failure but the price offered is at most v_2.

The Noisy Cautious Search for a valuation v_i proceeds in phases and begins by offering $1/2$ during the first macrostep. During each phase $n \geq 0$, if the last macrostep was a success, the price offered is increased by 2^{-2^n}. As soon as a macrostep is a failure, phase n ends and phase $n+1$ begins by offering the price of the last successful macrostep, plus $2^{-2^{n+1}}$. After $\lceil \log_2 \log_2 T \rceil$ phases, the price of the last successful macrostep is offered for all remaining rounds.

Lemma 11 *The Noisy Cautious Search for v_i with parameters i, δ, γ_i satisfies the following:*

1. *the price offered during each macrostep m is $2/m$-close to v_i with probability at least $1 - m\delta$;*
2. *the total reward accumulated by the end of macrostep m is at least*

$$\big(m v_i D(v_i) - 3(\ln \ln T) - 8\big) \frac{\ln \delta}{\ln(1-\gamma_i)}$$

with probability at least $1 - m\delta$.

Proof Claim 1 follows by Lemma 9 and the fact that the probability of making a mistake during each macrostep is at most δ by Chernoff inequality for Bernoulli random variables. Similarly, claim 2 follows by Lemma 8 and, again, Chernoff inequality. ■

Capped Mean Estimation

We begin this section by providing a method to find a high-confidence multiplicative estimate of the expectation μ of any $[0,1]$-valued random variable, using only $\mathcal{O}\big(\ln(1/\delta)/\mu\big)$ samples. Most notably, the expectation μ need *not* be known in advance. With our novel technique, we improve upon Berthet and Perchet (2017, Lemma 13), that proved a similar risult using $\mathcal{O}\big(\ln(1/\delta)/\mu^2\big)$ samples. This result will be pivotal for our analysis and we believe it will also be valuable in its own right. For any set $X_1, \dots, X_T$ of random variables, we denote by

$$\overline{X}_t = \frac{1}{t}\sum_{s=1}^{t} X_s \qquad \text{and} \qquad S_t^2 = \frac{1}{t-1}\sum_{s=1}^{t} \big(X_s - \overline{X}_t\big)^2$$

the *sample mean* and the *sample variance* of the first t random variables. The following result is a straightforward consequence of the empirical Bernstein bound and the confidence bound for standard deviation proven in (Maurer and Pontil, 2009, Theorems 4, 10).

Theorem 12 *Let $X_1, \dots, X_T$ be a set of $[0,1]$-valued i.i.d. random variables with expectation μ and standard deviation σ. For all $\delta \in (0,1)$ and all $t \in \{2, \dots, T\}$, the two following conditions hold simultaneously with probability at least $1-3\delta$*

$$\big|\overline{X}_t - \mu\big| \le \sqrt{2} S_t \left(\frac{\ln(1/\delta)}{t}\right)^{1/2} + \frac{7}{3}\frac{\ln(1/\delta)}{t-1} \qquad \textit{and} \qquad S_t \le \sigma + \sqrt{2}\left(\frac{\ln(1/\delta)}{t-1}\right)^{1/2} .$$

We can now prove our multiplicative mean estimation theorem.

Theorem 13 (Multiplicative mean estimation) *Let $X_1, \dots, X_T$ be a set of $[0,1]$-valued i.i.d. random variables with expectation $\mu > 0$ and standard deviation σ. For all $\delta \in (0,1)$ and all $\alpha \ge 0$, if $T \ge t_0$, where*

$$t_0 = \left\lceil \frac{\alpha+2}{3\mu} \ln\left(\frac{1}{\delta}\right) \left(\sqrt{9\alpha^2 + 114\alpha + 192} + 3\alpha + 19\right) \right\rceil + 2 = \mathcal{O}\left(\frac{\alpha^2}{\mu} \ln\frac{1}{\delta}\right)$$

and $\tau = \tau(T, \delta, \alpha)$ is the smallest time $t \in \{2, \dots, T\}$ such that

$$\frac{\overline{X}_t}{\alpha+1} \ge \sqrt{2} S_t \left(\frac{\ln(1/\delta)}{t}\right)^{1/2} + \frac{7}{3}\frac{\ln(1/\delta)}{t-1} \tag{9}$$

then, with probability at least $1 - 3(T-1)\delta$,

1. *$\tau \le t_0$,*
2. *for all $t \in \{2, \dots, T\}$ such that (9) holds,*

$$\left(\frac{\alpha}{\alpha+1}\right)\overline{X}_t < \mu < \left(\frac{\alpha+2}{\alpha+1}\right)\overline{X}_t \,. \tag{10}$$

Proof Denote for all $t \in \{2, \dots, T\}$, $c_t = \sqrt{2\, S_t^2 \ln(1/\delta)/t} + (7/3)\ln(1/\delta)/(t-1)$. By Theorem 12, the *good* event

$$G = \Big\{\forall t \in \{2, \dots, T\}, \quad \overline{X}_t - c_t < \mu < \overline{X}_t + c_t \quad \text{and} \quad S_t \le \sigma + \sqrt{2\ln(1/\delta)/(t-1)}\Big\}$$

Algorithm 5: Capped Mean Estimation

Input: $x_1, x_2, \ldots \in [0,1]$, $\theta \in [0,1]$, $\delta \in (0,1)$, $\rho \in \{0,1\}$.

Initialization: set $t \leftarrow 3$ and $\widehat{D}_s = (1-\rho)\mathbb{I}\{V_s \geq x_s\} + \rho(1 - \mathbb{I}\{V_s \geq x_s\})$ for all s.

offer x_1 and x_2 once each;

set $\overline{D} \leftarrow \frac{1}{2}\sum_{s=1}^{2} \widehat{D}_s$ and $S^2 \leftarrow \sum_{s=1}^{2} \big(\widehat{D}_s - \overline{D}\big)^2$;

while $\big[t \leq \lceil 40 \ln(1/\delta)/\theta \rceil + 2\big] \wedge \big[\overline{D} < \sqrt{8S^2 \ln(1/\delta)/t} + (14/3)\ln(1/\delta)/(t-1)\big]$ **do**

  offer price x_t once;

  update $\overline{D} \leftarrow \big(\overline{D}(t-1) + \widehat{D}_t\big)/t$, $S^2 \leftarrow \big(S^2(t-2) + (\widehat{D}_t - \overline{D})^2\big)/(t-1)$, and $t \leftarrow t+1$;

end

if $t > \lceil 40 \ln(1/\delta)/\theta \rceil + 2$ **then** return that $\mu \leq \theta$;

else return $\overline{D}/2$;

has probability $\mathbb{P}(G) \geq 1 - 3(T-1)\delta$. For all outcomes in G and all $t \in \{2, \ldots, T\}$,

$$\overline{X}_t < (\alpha+1)c_t \iff \mu - c_t < \overline{X}_t < (\alpha+1)c_t \implies \mu < (\alpha+2)c_t \implies t < t_0$$

hence $\tau \leq t_0$. This implies that for all outcomes in G and all $t \in \{1, \ldots, T\}$ such that $\overline{X}_t \geq (\alpha+1)c_t$,

$$\left(\frac{\alpha}{\alpha+1}\right)\overline{X}_t = \overline{X}_t - \frac{\overline{X}_t}{\alpha+1} \leq \overline{X}_t - c_t < \mu < \overline{X}_t + c_t \leq \overline{X}_t + \frac{\overline{X}_t}{\alpha+1} = \left(\frac{\alpha+2}{\alpha+1}\right)\overline{X}_t.$$

■

The following capped version of the previous theorem interrupts the process if during the multiplicative mean estimation it is learned that μ is smaller than some threshold parameter θ.

Corollary 1 (Capped Mean Estimation) *For any threshold parameter $\theta \in [0,1]$, under the same assumptions of Theorem 13, define $\tau_\theta = \min\{\tau, t_\theta\}$, where*

$$t_\theta = \left\lceil \frac{\alpha+2}{3\theta} \ln\left(\frac{1}{\delta}\right)\left(\sqrt{9\alpha^2 + 114\alpha + 192} + 3\alpha + 19\right)\right\rceil + 2 = \mathcal{O}\left(\frac{\alpha^2}{\theta}\ln\frac{1}{\delta}\right) .$$

With probability at least $1 - 3(T-1)\delta$,

1. *if $\tau_\theta = \tau$, then for all $t \in \{2, \ldots, T\}$ such that (9) holds, inequalities (10) also hold;*

2. *if $\tau_\theta = t_\theta$, then $\mu \leq \theta$.*

Our Capped Mean Estimation is defined as the Capped Mean Estimation of the demand curve (or one minus the demand curve if $\rho = 1$) at a given sequence of prices[3] $x_1, x_2, \ldots$, with threshold $\theta \in [0,1]$ (where $1/\theta$ is interpreted as ∞ when $\theta = 0$), confidence parameter $\delta \in (0,1)$, reverse parameter ρ (that regulates if $D(x_1)$ or $1 - D(x_1)$ is being estimated) and $\alpha = 1$ (Algorithm 5).

3. This algorithm is only used for prices $x_1, x_2, \ldots$ such such that $D(x_s) = D(x_t)$ for all s, t.

VARIANT: JOINT CAPPED MEAN ESTIMATION

We call (w, θ, δ)-Joint Capped Mean Estimation a variant of Algorithm 5 in which $x_t = w$ for all t and estimations for both $\rho = 0$ and $\rho = 1$ are carried on at the same time; i.e., where both $\overline{D}$ (sample mean for $\rho = 0$) and $\overline{D}' = 1 - \overline{D}$ (sample mean for $\rho = 1$), as well as their respective sample variances S^2 and $(S')^2$ are maintained; the condition $\big[\overline{D} \le \sqrt{8S^2 \ln(1/\delta)/t} + (14/3)\ln(1/\delta)/(t-1)\big]$ in the **while** loop is replaced by

$$(A \vee A') = \left(\left[\overline{D} < \sqrt{\frac{8S^2}{t}\ln\frac{1}{\delta}} + \frac{14}{3(t-1)}\ln\frac{1}{\delta}\right] \vee \left[\overline{D}' < \sqrt{\frac{8(S')^2}{t}\ln\frac{1}{\delta}} + \frac{14}{3(t-1)}\ln\frac{1}{\delta}\right]\right)$$

and at the end, we return $\overline{D}/2$ (resp., $\overline{D}'/2$) and we say that $D(w)$ (resp., $1-D(w)$) is *well-estimated* if and only if A (resp., A') is false; if A (resp., A') is true we return that $D(w)$ (resp., $1 - D(w)$) is at most θ.

VARIANT: CAPPED MEAN ESTIMATION ON NOISY CAUTIOUS SEARCH

With a slight abuse of notation, we say that a (θ, δ, ρ)-Capped Mean Estimation is run on a (δ, i, γ_i)-Noisy Cautious Search if $x_1, x_2, \dots$ are the prices offered during the first successful macrosteps of a (δ, i, γ_i)-Noisy Cautious Search run for $\Theta\big(\frac{1}{D(x_1)}\big)$ macrosteps (resp., $\Theta\big(\frac{1}{1-D(x_1)}\big)$ macrosteps); i.e., while the Noisy Cautious Search proceeds, an increasingly accurate estimate $\widehat{p}$ of $D(x_1)$ (resp., $1 - D(x_1)$) is maintained at the same time using samples from successful macrosteps; as soon as the stopping criterion for the Capped Mean Estimation is met, the estimation stops while the Noisy Cautious Search proceeds until it reaches $\lceil 6/\widehat{p}\rceil$ macrosteps, at which point the whole process ends returning $\widehat{p}$ and the price $\widehat{v}_i$ offered during the last succesful Noisy Cautious Search macrostep.

Cautious Mean Estimation

The main idea of this section is that the problem for $K = 2$ is completely solved by determining v_1, v_2, and p_2. This suggests that computing an high-confidence estimate p_2 once a value $w \in (v_1, v_2]$ is located might be a good idea. Sadly, it is not. The problem with this approach is that if p_2 is very small an arbitrary high regret may be incurred in doing so. On the other hand, the more evidence is gathered that p_2 is very small, the less likely it is that v_2 is optimal. For these and other more subtle reasons, a great deal of caution is needed in order to obtain estimate of p_2 that is just good enough to use.

The algorithm we present for dealing with these issues is called Cautious Mean Estimation and it receives as an input a price $w \in (v_1, v_2]$ (i.e., that can be used to estimate p_2), as well as a confidence parameter δ. The routine begins by determining if p_1 and p_2 are both bigger than $1/4$ by using a Joint Capped Mean Estimation and invoking Corollary 1. If this is true, it simply returns the estimates of p_1 and p_2 to the main routine; otherwise it behaves differently depending on which one is true: $p_2 \le 1/4$ or $p_2 \ge 3/4$, which can be checked invoking again Corollary 1. If $p_2 \le 1/4$, it proceeds in phases. In each phase s, it checks if $v_1 \ge 2^{-s}$ by offering 2^{-s} a small number of times, in which case it halts returning that v_1 is the optimum. If it is not, it determines if p_1 and p_2 are bigger than $2^{-(k+1)}$ by using one more time Corollary 1, in which case it returns their estimates to the main routine. If they are not, it moves on to phase $k + 1$. If on the other hand p_2 was bigger than $3/4$, it performs a Noisy Cautious Search for v_2, while at the same time collecting samples to estimate p_1, returning estimates $\widehat{v}_2$ and $\widehat{p}_1$. Then it first checks if $v_1 \le \widehat{v}_2(1 - \widehat{p}_1) - \widehat{p}_1$ by posting

Algorithm 6: Cautious Mean Estimation

Input: price $w \in [0, 1]$, confidence parameter $\delta \in (0, 1)$.

1 run a $(w, 2^{-2}, \delta)$-Joint Capped Mean Estimation, returning $\widehat{p}_1, \widehat{p}_2$;
2 **if** *$D(w)$ and $1 - D(w)$ are both well-estimated* **then** // $1/4 \le p_1, p_2 \le 3/4$
3   return $\widehat{p}_1, \widehat{p}_2$;
4 **else if** *$1 - D(w)$ is well-estimated* **then** // $p_1 > 3/4$
5   **for** $s \in \{2, 3, \dots\}$ **do**
6   offer 2^{-s} for $\lceil \ln(\delta)/\ln(3/4) \rceil$ rounds;
7   **if** *all offers are accepted* **then break** and return that v_1 is optimal;
8   **else**
9   continue the Joint Capped Mean Estimation with new parameters $w, 2^{-(s+1)}, \delta$;
10   **if** *p_1 and p_2 are both well-estimated* **then break** and return $\widehat{p}_1, \widehat{p}_2$;
11   **end**
12 **else if** *$D(w)$ is well-estimated* **then** // $p_2 > 3/4$
13   run $(0, \delta, 1)$-Capped Mean Estimation on $\big(\delta, 2, \frac{3}{4}\big)$-Noisy Cautious Search, returning $\widehat{p}_1, \widehat{v}_2$;
14   offer $\widehat{v}_2 \widehat{q}_2 - \widehat{p}_1$ for $\lceil \ln(1/\delta)/\widehat{p}_1 \rceil$ rounds, where $\widehat{q}_2 \leftarrow 1 - \widehat{p}_1$;
15   **if** *at least one offer is rejected* **then** return that v_2 is optimal;
16   **else** return $\widehat{p}_1, \widehat{p}_2$;

the latter for $\lceil \ln(1/\delta)/\widehat{p}_1 \rceil$ rounds. If the test is positive, it halts returning that v_2 is the optimum. Otherwise it returns $\widehat{p}_1$ and $\widehat{p}_2$ to the main routine.

Lemma 14 *For all $w \in (v_1, v_2]$ and all $\delta \in (0, 1)$, the Cautious Mean Estimation run with parameters w, δ satisfies the following with probability at least $1 - (15T - 13)\delta$:*

1. *if the algorithm returns that v_1 or v_2 is optimal, then it is correct;*
2. *if the algorithm returns $\widehat{p}_1$ and $\widehat{p}_2$, then both satisfy $p_i/3 < \widehat{p}_i < p_i$;*
3. *the regret of the algorithm it at most $(13)^2 \ln(1/\delta) + 6$.*

Proof By definition of Joint Capped Mean Estimation, line 6 lasts for at most $\lceil 160 \ln(1/\delta) \rceil + 2$ rounds, which upper bounds the regret accrued during those time steps. Denote G the *good* event in which which items 1 and 2 of Corollary 1 hold simultaneously for both the estimate of p_1 and p_2. To prove the result, we can (and do!) restrict our analysis to *good* outcomes, i.e., outcomes belonging in G. Indeed, Corollary 1 implies that one and only one of the three conditions at lines 2, 4, and 12 is executed with probability at least $\mathbb{P}(G) \ge 1 - 6(T - 1)\delta$ and we will show that the result holds in all three cases.

If the condition at line 2 is true, then the result follows immediately by Corollary 1.

Assume now that the condition at line 4 is true and fix $k \in \mathbb{N}$ such that $2^{-k} \le \max\{v_1, p_2\} \le 2^{-(k-1)}$. Note that if $v_1 \ge p_2$, the loop at line 5 will break with probability at least $1-\delta$ (by Chernoff inequality) at line 7 as soon as $s = k$; this proves point 1 for v_1. If on the other hand $v_1 < p_2$, the loop will break with probability at least $1 - 6(T - 1)\delta$ (by Corollary 1) at line 10 as soon as $s = k-1$; this proves point 2. In any case, then, at most $k-1$ cycles of the loop are performed with probability at least $1 - (6T - 5)\delta$. If $s \le k$, line 6 is performed at most $k - 1$ times and since the cost of sampling is at most v_1 (if v_1 is optimal) or p_2 (if v_2 is optimal), than the total regret accrued by executing line 6 is at most $(k-1)\lceil \ln(\delta)/\ln(3/4)\rceil \max\{v_1, p_2\} \le (e \ln 2)^{-1}\lceil \ln(\delta)/\ln(3/4)\rceil$, where we used $x \log_2(1/x) \le (e \ln 2)^{-1}$, for all $x > 0$. On the other hand, by the end of phase k the Joint Capped Mean Estimation at lines 6, 9 has offered w for at most $\lceil 2^{k+1} 40 \ln(1/\delta)\rceil + 2$ accruing at most $40 \ln(1/\delta) + 3$ regret. This proves point 3.

Finally, consider the case in which the condition at line 12 is true. The Noisy Cautious Search at line 13 stops after at most $\lceil 40 \ln(1/\delta)/p_1\rceil + 2$ rounds, returning $\widehat{p_i} \in (p_i/3, p_i)$, with probability at least $1-(7T-6)\delta$ by the fact that it makes a mistake with probability at most δ and Theorem 13. This proves point 2. If v_2 is optimal, Lemma 11 shows that the regret of the Noisy Cautious Search is at most $\big(3(\ln\ln T) + 8\ln(1/\delta)\big)\ln(4/3)$ with probability at least $1 - T\delta$. If v_1 is optimal, the additional regret is at most $\big(\lceil 40\ln(1/\delta)/p_1\rceil + 2\big)(v_1 - wp_2) \le 40\ln(1/\delta) + 3$.

Consider now lines 14-15. Since $p_1/3 < \widehat{p}_1 < p_1$, then $p_2 < \widehat{q}_2 < p_2 + (2/3)p_1$. Furthermore, $v_2 - p_1 \le \widehat{v}_2 \le v_2$ with probability at least $1 - T\delta$ by Lemma 11. If the test at line 15 is true, then $v_1 < v_2 p_2$ and v_2 is optimal with probability at least $1-\delta$; this proves point 1 for v_2. To compute the regret accumulated at line 14, assume first that v_1 is optimal; then necessarily $v_1 \ge \widehat{v}_2\widehat{q}_2 - \widehat{p}_1$ and the regret of line 14 is at most $(v_1 - \widehat{v}_2\widehat{q}_2 + \widehat{p}_1)\lceil 3\ln(1/\delta)/p_1\rceil \le 9\ln(1/\delta) + 3$. If on the other hand v_2 is optimal, then the regret of line 14 is at most $(p_2 v_2 - \widehat{v}_2\widehat{q}_2 + \widehat{p}_1)\lceil 3\ln(1/\delta)/p_1\rceil \le 6\ln(1/\delta) + 2$. This proves point 3 and concludes the proof. ■

2-UCB

This subroutine is a slightly modified version of Algorithm 2. The only differences are that two feasible intervals are initialized at the beginning, each valuation v_i gets a personalized number of rounds $\lceil 8\ln(\delta)/\ln(1-\gamma_i)\rceil$ at line 6, and the test at line 11 need not be executed as it is known in advance that $K = 2$.

The following result is a straightforward adaptation of Theorem 4. As such, the proof is omitted.

Lemma 15 *If $\Delta = |p_2 v_2 - v_1|$, $\gamma_1 \le p_1$, $\gamma_2 \le p_2$, and 2-UCB run with $\delta = T^{-2}$, it incurs a regret*

$$\mathcal{O}\left(\frac{\ln T}{\Delta} + (\ln T)(\ln\ln T)\left(\frac{\alpha_1}{-\ln(1-\gamma_1)} + \frac{\alpha_2}{-\ln(1-\gamma_2)}\right)\right),$$

where $(\alpha_1, \alpha_2) = (v_1 - v_1 p_2, v_1)$ if v_1 is optimal, $(\alpha_1, \alpha_2) = (v_2 p_2 - v_1 p_2, p_2 v_2)$ if v_2 is optimal, and the first term is absent if $\Delta = 0$.

Proof of Theorem 6

We finally have all the instruments to prove Theorem 6, that we restate for completeness.

Theorem 6 *If Algorithm 7 is run on two unknown pairs (v_1, p_1) and (v_2, p_2) with input parameter $\delta = T^{-2}$, then its regret satisfies*

$$R_T = \mathcal{O}\left(\frac{\log T}{\Delta} + (\log T)(\log\log T)\right),$$

where the first term is absent if $\Delta = |p_2 v_2 - v_1|$ is zero.

Proof Putting together the proofs of all previous lemmas, the probability of making a mistake in at least a test of at least a routine is upper bounded by $\mathcal{O}(T\delta)$. For this reason, we can (an do) assume that no mistakes happen. We divide the proof into three different cases.

Algorithm 7:

Input: Confidence parameter $\delta \in (0,1)$.

```
run a Binary Search, returning [a_1, a_2];                                         // phase 1
run a Capped Mean Estimation of the demand at a_2 with parameter θ = a_1, returning p̃_2;
if p̃_2 > 0 then set w ← a_2 ;
else
    offer price a_1 until it is rejected;                       // check if a_1 < v_1 ≤ v_2 < a_2
    set w ← a_1;
run a Cautious Mean Estimation of the demand at w, returning p̂_1 and p̂_2;          // phase 2
if the Cautious Mean Estimation was halted because v_1 or v_2 is the obvious optimum then
    run a Cautious Search for the optimal valuation with lower bound 1/2;
else run 2-UCB with parameters γ_1 = p̂_1 and γ_2 = p̂_2 ;                 // shrink the interval
```

Case 1 Assume that during phase 1 all offers of a_2 are rejected and all offers of a_1 are accepted. Consider the following four subcases. If $a_1 \leq v_1 \leq v_2 \leq a_2$, the regret is at most $\mathcal{O}(\log T)$. Assume now that $v_1 \leq a_1 \leq v_2 \leq a_2$. If v_2 is optimal, then the regret is at most $\mathcal{O}\big(\log(T)/a_1\big) = \mathcal{O}\big(\log(T)/\Delta\big)$. If v_1 is optimal, then the regret is at most $\mathcal{O}\big((v_1 - a_1 p_2)T + v_1 \ln(T)/a_1\big) = \mathcal{O}\big(a_1 p_1 T + \ln(T)\big)$. Note that this case only happens with probability $p_2^{\mathcal{O}(T - \ln(T)/a_1)}$, which is at least $1/T$ only if $p_1 = \mathcal{O}\big(\frac{\log T}{T - \log(T)/a_1}\big)$. Now, if $a_1 = \Omega\big(\log(T)/T\big)$ then the regret is at most $\mathcal{O}(\log T)$; otherwise it is at most $\mathcal{O}(\log T)$ because v_1 is small. If $a_1 \leq v_1 \leq a_2 \leq v_2$, the regret is at most $\mathcal{O}\big((\max\{v_1, p_2 v_2\} - a_1)T + \max\{v_1, p_2 v_2\}\log(T)/a_1\big)$. Since all offers of a_2 were rejected, Corollary 1 implies that $p_2 \leq a_1$, then $p_2 v_2 \leq a_1 \leq v_1$, hence v_1 is optimal. The regret is therefore at most $\mathcal{O}(\log T)$. Finally, assume that $v_1 \leq a_1 \leq a_2 \leq v_2$. Combining the same arguments as above, v_1 is optimal but p_1 is small and the total regret is at most $\mathcal{O}(\log T)$.

Case 2 Assume that during phase 1 some offers of a_2 are accepted. Corollary 1 implies that the first Capped Mean Estimation lasts at most $\mathcal{O}\big(\log(T)/\max\{a_1, p_2\}\big)$ rounds, hence its regrets is at most $\mathcal{O}(\log T)$. Lemma 14 implies that the Cautious Mean Estimation has a regret at most $\mathcal{O}(\log T)$. If the cautious mean estimation is halted returning that v_1 or v_2 is optimal, then Lemma 11 implies that the regret is at most $\mathcal{O}\big((\log T)(\log\log T)\big)$. Assume now that the cautious mean estimation returns $\widehat{p}_1, \widehat{p}_2$. By construction, if $p_2 \leq 1/4$, then necessarily $v_1 \leq 2p_2$, thus $\Delta \leq 2p_2$. On the other end, if $p_2 \geq 3/4$, then necessarily $v_1 \geq p_2 v_2 - 2p_1$ thus $\Delta \leq 2p_1$ if v_2 is optimal. Using Lemma 15 and plugging in the above upper bounds gives the result.

Case 3 Assume that during phase 1 all offers of a_2 are rejected and some offers of a_1 are rejected. The proof of this case is the same as the previous one, except that sampling a_2 has an extra regret cost. If v_1 is optimal, then the additional regret is at most $\mathcal{O}\big(v_1 \log(T)/a_1\big) = \mathcal{O}(\ln T)$ because $v_1 < a_1$. Finally, assume that v_2 is optimal. If $v_2 \leq a_2$, the additional cost is at most $\big(\ln(T)/a_1\big)p_2 v_2 = \mathcal{O}(\ln T)$. If $a_2 < v_2$, then $p_2 \leq a_1$ by Corollary 1 and the additional cost is at most $\mathcal{O}\big((p_2 v_2 - p_2 a_2)\log(T)/a_1\big) = \mathcal{O}(\log T)$. ■