Dynamic Pricing with Limited Supply (extended abstract)*

Moshe Babaioff[†] Shaddin Dughmi[‡] Robert Kleinberg[§] Aleksandrs Slivkins[†]

May 2012

Abstract

We consider the problem of designing revenue maximizing online posted-price mechanisms when the seller has limited supply. A seller has k identical items for sale and is facing n potential buyers ("agents") that are arriving sequentially. Each agent is interested in buying one item. Each agent's value for an item is an independent sample from some fixed (but unknown) distribution with support $[0,1]$. The seller offers a take-it-or-leave-it price to each arriving agent (possibly different for different agents), and aims to maximize his expected revenue.

We focus on mechanisms that do not use any information about the distribution; such mechanisms are called *prior-independent*. They are desirable because knowing the distribution is unrealistic in many practical scenarios. We study how the revenue of such mechanisms compares to the revenue of the optimal offline mechanism that knows the distribution ("offline benchmark").

We present a prior-independent mechanism whose revenue is at most $O((k \log n)^{2/3})$ less than the offline benchmark, for every distribution that is regular. This guarantee holds without *any* assumptions if the benchmark is relaxed to fixed-price mechanisms. Further, we prove a matching lower bound.

On a technical level, we exploit the connection to multi-armed bandits (MAB). While dynamic pricing with unlimited supply can easily be seen as an MAB problem, the intuition behind MAB approaches breaks when applied to the setting with limited supply. Our high-level conceptual contribution is that even the limited supply setting can be fruitfully treated as a bandit problem.

1. Introduction

Consider an airline that is interested in selling k tickets for a given flight. The seller is interested in maximizing her revenue from selling these tickets, and is offering the tickets on a website such as Expedia. Potential buyers ("agents") arrive one after another, each with the goal of purchasing a ticket if the price is smaller than the agent's valuation. The seller expects n such agents to arrive. Whenever an agent arrives the seller presents to him a take-it-or-leave-it price (*posted price*), and the agent makes a purchasing decision according to that price. The seller can update the price taking into account the observed history and the number of remaining items and agents.

Posted price mechanisms are commonly used in practice, and are appealing for several reasons. First, an agent only needs to evaluate her offer rather than compute her private value exactly. Human agents tend to find the former task much easier than the latter. Second, agents do not reveal their entire private information to the seller: rather, they only reveal whether their private value is larger than the posted price. Third, posted-price mechanisms are truthful (in dominant strategies) and moreover also group strategyproof (a notion of collusion resistance when side payments are not allowed). Further, prior-independent posted-price mechanisms are particularly useful in practice as the seller is not required to estimate the demand distribution in advance. Similar arguments can be found in prior work, e.g. (Chawla et al., 2010).

We adopt a Bayesian view that the valuations of the buyers are IID samples from a fixed distribution, called *demand distribution*. A standard assumption in a Bayesian setting is that the demand distribution is known to the seller, who can design a specific mechanism tailored to this knowledge. (For example, the Myerson optimal auction for one item sets a reserve price that is a function of the distribution). However, in some settings this assumption is very strong, and should be avoided if possible. For example, when the

*The full paper (with more results) will be published in *ACM EC 2012*, and is available on arxiv.org.

[†]Microsoft Research Silicon Valley, Mountain View CA, USA. Email: {moshe,slivkins} @ microsoft.com.

[‡]Microsoft Research, Redmond WA, USA. Email: shaddin@microsoft.com.

[§]Department of Computer Science, Cornell University, Ithaca NY, USA. Email: rdk@cs.cornell.edu.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

seller enters a new market, she might not know the demand distribution, and learning it through market research might be costly. Likewise, when the market has experienced a significant recent change, the new demand distribution might not be easily derived from the old data.

We would like to design mechanisms that perform well for any demand distribution, and yet do not rely on knowing it. Such mechanisms are called *prior-independent*. Learning about the demand distribution is then an integral part of the problem. The performance of such mechanisms is compared to a benchmark that *does* depend on the specific demand distribution, as in (Kleinberg & Leighton, 2003; Hartline & Roughgarden, 2008; Besbes & Zeevi, 2009; Dhangwatnotai et al., 2010) and many other papers.

2. Our model and contributions

We consider the following limited supply auction model, which we term *dynamic pricing with limited supply*. A seller has k items she can sell to a set of n agents (potential buyers), aiming to maximize her expected revenue. The agents arrive sequentially to the market and the seller interacts with each agent before observing future agents. We make the simplifying assumption that each agent interacts with the seller only once, and the timing of the interaction cannot be influenced by the agent. (This assumption is also made in other papers that consider our problem for special supply amounts (Kleinberg & Leighton, 2003; Babaioff et al., 2011; Besbes & Zeevi, 2009).) Each agent i ($1 \leq i \leq n$) is interested in buying one item, and has a private value v_i for an item. The private values are independently drawn from the same *demand distribution* F. The F is *unknown* to the seller, but it is known that F has support in $[0,1]$.[1] Letting $F(p)$ denote the c.d.f., $S(p) \triangleq 1 - F(p)$ is called *survival rate*, which in our setting means is the the probability of a sale at price p.

Whenever agent i arrives to the market the seller offers him a price p_i for an item. The agent buys the item if and only if $v_i \geq p_i$, and in case she buys the item she pays p_i (so the mechanism is incentive-compatible). The seller never learns the exact value of v_i, she only observes the agent's binary decision to buy the item or not. The seller selects prices p_i using an online algorithm, that we henceforth call *pricing strategy*. We are interested in designing pricing strategies with high revenue compared to a natural benchmark, with minimal assumptions on the demand distribution.

Our main benchmark is the maximal expected revenue of an offline mechanism that is allowed to use the demand distribution; henceforth, we will call it *offline benchmark*. This is a very strong benchmark, as it has the following advantages over our mechanism: it is allowed to use the demand distribution, it is not constrained to posted prices and is not constrained to run online. It is realized by a well-known Myerson Auction (Myerson, 1981) (which *does* rely on knowing the demand distribution).

Theorem 1. *There exists a prior-independent pricing strategy such that for any regular demand distribution its expected revenue is at least the offline benchmark minus* $O((k \log n)^{2/3})$.

Regularity is a mild and standard condition in the Mechanism Design literature.[2] The pricing strategy in Theorem 1 is deterministic and (trivially) runs in polynomial time. The resulting mechanism is incentive-compatible as it is a posted price mechanism. The specific bound $O((k \log n)^{2/3})$ is most informative when $k \gg \log n$, so that the dependence on n is insignificant; the focus here is to optimize the power of k.

The proof of Theorem 1 consists of two stages. The first stage (immediate from (Yan, 2011)) reduces the problem to the *fixed-price benchmark*: the expected revenue of the best fixed-price strategy[3] for a given distribution. We observe that for any regular demand distribution, the fixed-price benchmark is close to the offline benchmark. The second stage, which is our main technical contribution, is to show that our pricing strategy achieves expected revenue that is close to the fixed-price benchmark. Surprisingly, this holds without *any* assumptions on the demand distribution.

Theorem 2. *There exists a prior-independent pricing strategy whose expected revenue is at least the fixed-price benchmark minus* $O((k \log n)^{2/3})$. *This result holds for every demand distribution. Moreover, this result is the best possible up to a factor of* $O(\log n)$.

If the demand distribution is regular and moreover the ratio $\frac{k}{n}$ is sufficiently small then the guarantee in Theorem 1 can be improved to $O(\sqrt{k} \log n)$, with a distribution-specific constant.

Theorem 3. *There exists a detail-free pricing strategy whose expected revenue, for any regular demand distribution* F*, is at least the offline benchmark minus* $O(c_F \sqrt{k} \log n)$ *whenever* $\frac{k}{n} \leq s_F$*, where* c_F *and* s_F *are positive constants that depend only on* F.

The bound in Theorem 3 is achieved using the pricing strategy from Theorem 1 with a different parameter. Varying this parameter, we obtain a family of strategies that improve over the bound in Theorem 1 in the "nice" setting of

[1] Assuming that $\texttt{support}(F) \subset [0,1]$ is w.l.o.g. (by normalizing) as long as the seller knows an upper bound on the support.

[2] The demand distribution F is called *regular* if $F(\cdot)$ is twice differentiable and $R(p) = p\,S(p)$ is concave: $R''(\cdot) \leq 0$.

[3] A fixed-price strategy is a pricing strategy that offers the same price to all agents, as long as it has items to sell.

Theorem 3, and moreover have non-trivial additive guarantees for arbitrary demand distributions. However, we cannot match both theorems with the same parameter.

Note that the rate-$\sqrt{k}$ dependence on k in Theorem 3 contains a distribution-dependent constant c_F (which can be arbitrarily large, depending on F), and thus is not directly comparable to the rate-$k^{2/3}$ dependence in Theorem 2. The distinction (and a significant gap) between bounds with and without distribution-dependent constants is not uncommon in the literature on sequential decision problems, e.g. in (Auer et al., 2002a; Kleinberg & Leighton, 2003; Kleinberg et al., 2008).[4]

In fact, we show that the $c_F\,\sqrt{k}$ dependence on k is essentially the best possible.[5] We focus on the fixed-price benchmark (which is a weaker benchmark, so it gives to a stronger lower bound). Following the literature, we define *regret* as the fixed-price benchmark minus the expected revenue of our pricing strategy.

Theorem 4. *For any $\gamma < \frac{1}{2}$, no detail-free pricing strategy can achieve regret $O(c_F\, k^\gamma)$ for all demand distributions F and arbitrarily large k, n, where the constant c_F can depend on F.*

3. High-level discussion

Absent the supply constraint, our problem fits into the *multi-armed bandit* (MAB) framework (Cesa-Bianchi & Lugosi, 2006): in each round, an algorithm chooses among a fixed set of alternatives ("arms") and observes a payoff, and the objective is to maximize the total payoff over a given time horizon.[6] Our setting corresponds to (prior-free) MAB with *stochastic payoffs* (Lai & Robbins, 1985): in each round, the payoff is an independent sample from some unknown distribution that depends on the chosen "arm" (price). This connection is exploited in (Kleinberg & Leighton, 2003; Blum et al., 2003) for the special case of unlimited supply ($k = n$). The authors use a standard algorithm for MAB with stochastic payoffs, called `UCB1` (Auer et al., 2002a). Specifically, they focus on the prices $\{i\delta : i \in \mathbb{N}\}$, for some parameter δ, and run `UCB1` with these prices as "arms". The analysis relies on the regret bound from (Auer et al., 2002a).

However, neither the analysis nor the intuition behind `UCB1` and similar MAB algorithms is directly applicable for the setting with limited supply. Informally, the goal of an MAB algorithm would be to converge to a price p that maximizes the expected per-round revenue $R(p) \triangleq p\,S(p)$. This is, in general, a wrong approach if the supply is limited: indeed, selling at a price that maximizes $R(\cdot)$ may quickly exhaust the inventory, in which case a higher price would be more profitable.

Our high-level conceptual contribution is showing that even the limited supply setting can be fruitfully treated as a bandit problem. The MAB perspective here is that we focus on the trade-off between *exploration* (acquiring new information) and *exploitation* (taking advantage of the information available so far). In particular, we recover an essential feature of `UCB1` that it does not separate exploration and exploitation, and instead explores arms (prices) according to a schedule that unceasingly adapts to the observed payoffs. This feature results, both for `UCB1` and for our algorithm, in a much more efficient exploration of suboptimal arms: very suboptimal arms are chosen very rarely even while they are being "explored".

4. Our approach

We use an "index-based" algorithm where each arm is deterministically assigned a numerical score ("index") based on the past history, and in each round an arm with a maximal index is chosen; the index of an arm depends on the past history of this arm (and not on other arms). One key idea is that we define the index of an arm according to the estimated expected total payoff from this arm given the known constraints, rather than according to its estimated expected payoff in a single round. This idea leads to an algorithm that is simple and (we believe) very natural. However, while the algorithm is simple its analysis is not: some new ideas are needed, as the elegant tricks from prior work do not apply.

We apply the above idea to `UCB1`. The index in `UCB1` is, essentially, the best available Upper Confidence Bound (UCB) on the expected single-round payoff from a given arm. Accordingly, we define a new index, so that the index of a given price corresponds to a UCB on the expected total payoff from this price (i.e., from a fixed-price strategy with this price), given the number of agents and the inventory size. Such index takes into account both the average payoff from this arm ("exploitation") and the number of samples for this arm ("exploration"), as well as the supply constraint. In particular we recover the appealing property of `UCB1` that it does not separate "exploration" and "exploitation", and instead explores arms (prices) according to

[4] For a particularly pronounced example, for the K-armed bandit problem with stochastic payoffs the best possible rates for regret with and without a distribution dependent constant are respectively $O(c_F \log n)$ and $O(\sqrt{Kn})$ (Auer et al., 2002a;b; Audibert & Bubeck, 2010).

[5] However, the lower bound in Theorem 4 does not match the upper bound in Theorem 3 since the latter assumes regularity.

[6] To avoid a possible confusion, let us note that our supply constraint is very different from the "budget constraint" in line of work on *budgeted MAB* (see (Bubeck et al., 2009; Goel et al., 2009) for details and further references). The latter consraint is essentially the duration of the experimentation phase (n), rather than the number of rounds with positive reward (k).

a schedule that unceasingly adapts to the observed payoffs.

There are several steps to make this approach more precise. First, while it is tempting to use the current values for the number of agents and the inventory size to define the index, we adopt a non-obvious (but more elegant) design choice to use the original values, i.e. the n and the k. Second, since the exact expected total revenue for a given price p is hard to quantify, we will instead use what we prove is a good approximation thereof:

$$\nu(p) = p\,\min(k, nS(p)), \tag{1}$$

where $S(p)$ is the survival rate. That is, our index will be a UCB on $\nu(p)$. More specifically, we define

$$I_t(p) \triangleq p \cdot \min(k,\, n\, S_t^{\mathtt{UB}}(p)), \tag{2}$$

where $S_t^{\mathtt{UB}}(p)$ is a UCB on $S(p)$. Third, in specifying $S_t^{\mathtt{UB}}(p)$ we will use a non-standard estimator from (Kleinberg et al., 2008) to better handle prices with very low survival rate (see the full version for the details).

The main technical hurdle in the analysis is to "charge" each suboptimal price for each time that it is chosen, in a way that the total regret is bounded by the sum of these charges and this sum can be usefully bounded from above.

An additional difficulty comes from the probabilistic nature of the analysis. To this end, we cleanly decouple the analysis into "probabilistic" and "deterministic" parts. While we use a well-known trick – we define some high-probability events and assume that these events hold deterministically in the rest of the analysis – identifying an appropriate collection of events is non-trivial. Proving that these events indeed hold with high probability relies on some non-standard tail bounds from prior work.

5. Our pricing strategy: `CappedUCB`

The pricing strategy is initialized with a set $\mathcal{P}$ of "active prices". In each round t, some price $p \in \mathcal{P}$ is chosen. Namely, for each price $p \in \mathcal{P}$ we define a numerical score, called *index*, and we pick a price with the highest index, breaking ties arbitrarily. Once k items are sold, `CappedUCB` sets the price to ∞ and never sells any additional item.

Recall that the total expected revenue from the fixed-price strategy with price p is approximated by (1). In each round t, we define the *index* $I_t(p)$ as a UCB on $\nu(p)$ as in (2).

For each $p \in \mathcal{P}$ and time t, let $N_t(p)$ be the number of rounds before t in which price p has been chosen, and let $k_t(p)$ be the number of items sold in these rounds. Then $\widehat{S}_t(p) \triangleq k_t(p)/N_t(p)$ is the current average survival rate. (Define $\widehat{S}_t(p)$ to be equal to 1 when $N_t(p) = 0$.)

Mechanism 1 `CappedUCB` for n agents and k items

Parameter: $\delta \in (0,1)$
1: $\mathcal{P} \leftarrow \{\delta(1+\delta)^i \in [0,1] : i \in \mathbb{N}\}$ {"active prices"}
2: While there is at least one item left,
    in each round t,
        pick any price $p \in \operatorname{argmax}_{p\in\mathcal{P}} I_t(p)$,
            where $I_t(p)$ is the "index" given by (5).
3: For all remaining agents, set price $p = \infty$.

A *confidence radius* is some number $r_t(p)$ such that

$$|S(p) - \widehat{S}_t(p)| \le r_t(p) \quad (\forall\, p \in \mathcal{P}, t \le n). \tag{3}$$

holds w.h.p., namely with probability at least $1 - n^{-2}$.

We need to define a suitable confidence radius $r_t(p)$, which we want to be as small as possible subject to (3). Note that $r_t(p)$ must be defined in terms of quantities that are observable at time t, such as $N_t(p)$ and $\widehat{S}_t(p)$. A standard confidence radius used in the literature is (essentially)

$$r_t(p) = \sqrt{\tfrac{\Theta(\log n)}{N_t(p)+1}}.$$

Instead, we use a more elaborate confidence radius from (Kleinberg et al., 2008):

$$r_t(p) \triangleq \frac{\alpha}{N_t(p)+1} + \sqrt{\frac{\alpha\, \widehat{S}_t(p)}{N_t(p)+1}}, \tag{4}$$

for some $\alpha = \Theta(\log n)$.

The reason for using the confidence radius in (4) is that performs as well as the standard one in the worst case: $r_t(p) \le \sqrt{\frac{O(\log n)}{N_t(p)+1}}$, and much better for very small survival rates: $r_t(p) \le \frac{O(\log n)}{N_t(p)+1}$. (See (7) for the precise statement.)

Now we are ready to define the index:

$$I_t(p) \triangleq p \cdot \min(k,\, n\,(\widehat{S}_t(p) + r_t(p))). \tag{5}$$

Finally, the active prices are given by

$$\mathcal{P} = \{\delta(1+\delta)^i \in [0,1] : i \in \mathbb{N}\}, \tag{6}$$

where $\delta \in (0,1)$ is a parameter to be adjusted. See Mechanism 1 for the pseudocode.

All proofs can be found in the full version. For an interested reader, we include the proof of the main technical result (Theorem 2) in the appendix.

6. Related work

Dynamic pricing problems and, more generally, revenue management problems, have a rich literature in Operations

Research. A proper survey of this literature is beyond our scope; see (Besbes & Zeevi, 2009) for an overview. The main focus is on parameterized demand distributions, with priors on the parameters.

The study of dynamic pricing with *unknown* demand distribution has been initiated in (Blum et al., 2003; Kleinberg & Leighton, 2003). Several special cases of our setting have been studied in (Kleinberg & Leighton, 2003; Babaioff et al., 2011; Besbes & Zeevi, 2009), detailed below. First, (Kleinberg & Leighton, 2003) consider the unlimited supply case (building on the earlier work (Blum et al., 2003)). Among other results, they study IID valuations, i.e. our setting with $k = n$. They provide an $O(n^{2/3} \log n)$ upper bound on regret, and prove a matching lower bound. On the other extreme, (Babaioff et al., 2011) consider the case that the seller has only one item to sell ($k = 1$). They provide a super-constant multiplicative lower bound for unrestricted demand distribution (with respect to the online optimal mechanism), and a constant-factor approximation for monotone hazard rate distributions. (Besbes & Zeevi, 2009) consider a continuous-time version which (when specialized to discrete time) is essentially equivalent to our setting with $k = \Omega(n)$. They prove a number of upper bounds on regret with respect to the fixed-price benchmark, with guarantees that are inferior to ours. The key distinction is that their pricing strategies separate exploration and exploitation.

The study of online mechanisms was initiated by (Lavi & Nisan, 2000), who unlike us consider the case that each agent is interested in multiple items, and provide a logarithmic multiplicative approximation. Below we survey only the most relevant papers in this line of work, in addition to the special cases of our setting that we have already discussed. Several papers (Bar-Yossef et al., 2002; Blum et al., 2003; Kleinberg & Leighton, 2003; Blum & Hartline, 2005) consider online mechanisms with unlimited supply and adversarial valuations (as opposed to limited supply and IID valuations in our setting). (Hajiaghayi et al., 2004; Devanur & Hartline, 2009) study online mechanisms for limited supply and IID valuations (same as us), but their mechanisms are not posted-price.

MAB has a rich literature in Statistics, Operations Research, Computer Science and Economics; a reader can refer to (Cesa-Bianchi & Lugosi, 2006; Bergemann & Välimäki, 2006) for background. Most relevant to our specific setting is the work on (prior-free) MAB with stochastic payoffs, e.g. (Lai & Robbins, 1985; Auer et al., 2002a), and MAB with Lipschitz-continuous stochastic payoffs, e.g. (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Kleinberg et al., 2008; Bubeck et al., 2011). The posted-price mechanisms in (Blum et al., 2003; Kleinberg & Leighton, 2003; Blum & Hartline, 2005) mentioned above are based on a well-known MAB algorithm (Auer et al., 2002b) for adversarial payoffs. The connection between reinforcement learning and mechanism design has been explored in a number of other papers, including (Nazerzadeh et al., 2008; Devanur & Kakade, 2009; Babaioff et al., 2009; 2010).

7. Conclusions and open questions

We consider dynamic pricing with limited supply and achieve near-optimal performance using an index-based bandit-style algorithm. A key idea in designing this algorithm is that we define the index of an arm (price) according to the estimated expected *total payoff* from this arm given the known constraints.

It is worth noting that a good index-based algorithm did not *have* to exist in our setting. Indeed, many bandit algorithms in the literature are not index-based, e.g. `EXP3` (Auer et al., 2002b) and "zooming algorithm" (Kleinberg et al., 2008) and their respective variants. The fact that Gittins algorithm (Gittins, 1979) and `UCB1` (Auer et al., 2002a) achieve (near-)optimal performance with index-based algorithms was widely seen as an impressive contribution.

While in this paper we apply the above key idea to a specific index-based algorithm (`UCB1`), it can be seen as an (informal) general reduction for index-based algorithms for dynamic pricing, from unlimited supply to limited supply. This reduction may help with more general dynamic pricing settings (more on that below), and moreover it can be extended to other bandit-style settings where the "best arm" is *not* an arm with the best expected per-round payoff. In particular, an ongoing project (Abraham et al., 2012) uses this reduction in the context of adaptive crowd-selection in crowdsourcing.

It is an interesting open question whether a reduction such as above can be made more formal, and which algorithms and which settings it can be applied to. An ambitions conjecture for our setting is that there is a simple black-box reduction from unlimited supply to limited supply that applies to arbitrary "reasonable" algorithms. In the full generality this conjecture appears problematic; e.g. some reasonable bandit algorithms such as `EXP3` are hard-coded to spend a prohibitively large amount of time on exploration.

This paper gives rise to a number of more concrete open questions. First, it is desirable to extend Theorem 1 to possibly irregular distributions, i.e. obtain non-trivial regret bounds with respect to the offline benchmark. Second, one wonders whether the optimal $O(c_F \sqrt{k})$ regret rate from Theorem 3 can be extended to all regular demand distributions. Third, it is open whether our lower bounds can be strengthened to regular demand distributions.

Further, it is desirable to extend dynamic pricing with limited supply beyond IID valuations. A recent result in this direction is (Besbes & Zeevi, 2011), where the demand distribution can change exactly once, at some point in time that is unknown to the mechanism. Natural specific targets for further work are slowly changing valuations and adversarial valuations. One promising approach for slowly changing valuations is to apply the reduction from this paper to index-based algorithms for the corresponding bandit setting (Slivkins & Upfal, 2008; Slivkins, 2011).

References

Abraham, Ittai, Alonso, Omar, Kandylas, Vasilis, and Slivkins, Aleksandrs. Adaptive Algorithms for Crowdsourcing, 2012. Ongoing project.

Agrawal, Rajeev. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, 1995.

Audibert, J.Y. and Bubeck, S. Regret Bounds and Minimax Policies under Partial Monitoring. *J. of Machine Learning Research (JMLR)*, 11:2785–2836, 2010. A preliminary version has been published in *COLT 2009*.

Auer, Peter, Cesa-Bianchi, Nicolò, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. Preliminary version in *15th ICML*, 1998.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Auer, Peter, Ortner, Ronald, and Szepesvári, Csaba. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *20th Conf. on Learning Theory (COLT)*, pp. 454–468, 2007.

Babaioff, Moshe, Sharma, Yogeshwer, and Slivkins, Aleksandrs. Characterizing truthful multi-armed bandit mechanisms. In *10th ACM Conf. on Electronic Commerce (EC)*, pp. 79–88, 2009.

Babaioff, Moshe, Kleinberg, Robert, and Slivkins, Aleksandrs. Truthful mechanisms with implicit payment computation. In *11th ACM Conf. on Electronic Commerce (EC)*, pp. 43–52, 2010. Best Paper Award.

Babaioff, Moshe, Blumrosen, Liad, Dughmi, Shaddin, and Singer, Yaron. Posting prices with unknown distributions. In *Symp. on Innovations in CS*, 2011.

Bar-Yossef, Z., Hildrum, K., and Wu, F. Incentive-compatible online auctions for digital goods. In *13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2002.

Bergemann, Dirk and Välimäki, Juuso. Bandit Problems. In Durlauf, Steven and Blume, Larry (eds.), *The New Palgrave Dictionary of Economics, 2nd ed.* Macmillan Press, 2006.

Besbes, Omar and Zeevi, Assaf. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57:1407–1420, 2009.

Besbes, Omar and Zeevi, Assaf. On the minimax complexity of pricing in a changing environment. *Operations Reseach*, 59: 66–79, 2011.

Blum, Avrim and Hartline, Jason. Near-optimal online auctions. In *16th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2005.

Blum, Avrim, Kumar, Vijay, Rudra, Atri, and Wu, Felix. Online learning in online auctions. In *14th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pp. 202–204, 2003.

Bubeck, Sébastien, Munos, Rémi, and Stoltz, Gilles. Pure Exploration in Multi-Armed Bandit Problems. In *20th Intl. Conf. on Algorithmic Learning Theory (ALT)*, 2009.

Bubeck, Sébastien, Munos, Rémi, Stoltz, Gilles, and Szepesvari, Csaba. Online Optimization in X-Armed Bandits. *J. of Machine Learning Research (JMLR)*, 12:1587–1627, 2011. Preliminary version in *NIPS 2008*.

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

Chawla, Shuchi, Hartline, Jason D., Malec, David L., and Sivan, Balasubramanian. Multi-parameter mechanism design and sequential posted pricing. In *ACM Symp. on Theory of Computing (STOC)*, pp. 311–320, 2010.

Devanur, Nikhil and Hartline, Jason. Limited and online supply and the bayesian foundations of prior-free mechanism design. In *ACM Conf. on Electronic Commerce (EC)*, 2009.

Devanur, Nikhil and Kakade, Sham M. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pp. 99–106, 2009.

Dhangwatnotai, Peerapong, Roughgarden, Tim, and Yan, Qiqi. Revenue maximization with a single sample. In *ACM Conf. on Electronic Commerce (EC)*, pp. 129–138, 2010.

Gittins, J. C. Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B*, 41:148–177, 1979.

Goel, Ashish, Khanna, Sanjeev, and Null, Brad. The Ratio Index for Budgeted Learning, with Applications. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pp. 18–27, 2009.

Hajiaghayi, Mohammad T., Kleinberg, Robert, and Parkes, David C. Adaptive limited-supply online auctions. In *Proc. ACM Conf. on Electronic Commerce*, pp. 71–80, 2004.

Hartline, J.D. and Roughgarden, T. Optimal mechanism design and money burning. In *ACM Symp. on Theory of Computing (STOC)*, 2008.

Kleinberg, Robert. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems (NIPS)*, 2004.

Kleinberg, Robert, Slivkins, Aleksandrs, and Upfal, Eli. Multi-Armed Bandits in Metric Spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pp. 681–690, 2008.

Kleinberg, Robert D. and Leighton, Frank T. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2003.

Lai, T.L. and Robbins, Herbert. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Lavi, Ron and Nisan, Noam. Competitive analysis of incentive compatible on-line auctions. In *ACM Conference on Electronic Commerce*, pp. 233–241, 2000.

Myerson, R. B. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Nazerzadeh, Hamid, Saberi, Amin, and Vohra, Rakesh. Dynamic cost-per-action mechanisms and applications to online advertising. In *17th Intl. World Wide Web Conf. (WWW)*, 2008.

Slivkins, Aleksandrs. Contextual Bandits with Similarity Information. In *24th Conf. on Learning Theory (COLT)*, 2011.

Slivkins, Aleksandrs and Upfal, Eli. Adapting to a Changing Environment: the Brownian Restless Bandits. In *21st Conf. on Learning Theory (COLT)*, pp. 343–354, 2008.

Yan, Qiqi. Mechanism design via correlation gap. In *22nd ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2011.

Appendix A: Proof of Theorem 2

We prove that `CappedUCB` achieves regret $O(k \log n)^{2/3}$, given parameter $\delta = k^{-1/3} (\log n)^{2/3}$.

Since this regret bound is trivial for $k < \log^2 n$, we will assume that $k \geq \log^2 n$ from now on.

Note that `CappedUCB` "exits" (sets the price to ∞) after it sells k items. For a thought experiment, consider a version of this pricing strategy that does not "exit" and continues running as if it has unlimited supply of items; let us call this version `CappedUCB`$'$. Then the realized revenue of `CappedUCB` is exactly equal to the realized revenue obtained by `CappedUCB`$'$ from selling the first k items. Thus from here on we focus on analyzing the latter.

We will use the following notation. Let X_t be the indicator variable of the random event that `CappedUCB`$'$ makes a sale in round t. Note that X_t is a 0-1 random variable with expectation $S(p_t)$, where p_t depends on $X_1, \ldots, X_{t-1}$. Let $X \triangleq \sum_{t=1}^{n} X_t$ be the total number of sales if the inventory were unlimited. Note that $\mathbb{E}[X] = S \triangleq \sum_{t=1}^{n} S(p_t)$.

Going back to our original algorithm, let $\widehat{\texttt{Rev}}$ denote the realized revenue of `CappedUCB` (revenue that is realized in a given execution). Then $\widehat{\texttt{Rev}} = \sum_{t=1}^{N} p_t\, X_t$, where N is the largest integer such that $N \leq n$ and $\sum_{t=1}^{N} X_t \leq k$.

High-probability events. We tame the randomness inherent in the sales X_t by setting up three high-probability events, as described below. In the rest of the analysis, we will argue deterministically under the assumption that these three events hold. It suffices because the expected loss in revenue from the low-probability failure events will be negligible. The three events are summarized as follows:

Claim 5. *With probability at least $1 - n^{-2}$ holds, for each round t and each price $p \in \mathcal{P}$:*

$$\begin{aligned} |S(p) - \widehat{S}_t(p)| &\leq r_t(p) \\ &\leq 3 \left(\tfrac{\alpha}{N_t(p)+1} + \sqrt{\tfrac{\alpha\, S_t(p)}{N_t(p)+1}} \right), \end{aligned} \tag{7}$$

$$|X - S| < O(\sqrt{S \log n} + \log n), \tag{8}$$

$$\left|\textstyle\sum_{t=1}^{n} p_t (X_t - S(p_t))\right| < O(\sqrt{S \log n} + \log n). \tag{9}$$

In the first event, the left inequality asserts that $r_t(p)$ is a confidence radius, and the right inequality gives the performance guarantee for it. The other two events focus on `CappedUCB`$'$, and bound the deviation of the total number of sales (X) and the realized revenue ($\sum_{t=1}^{n} p_t\, X_t$) from their respective expectations; importantly, these bound are in terms of $\sqrt{S}$ rather than $\sqrt{n}$.

The proof of Claim 5 can be found in the full version. In the rest of the analysis we will assume that the three events in Claim 5 hold deterministically.

Single-round analysis. Let us analyze what happens in a particular round t of the pricing strategy. Let p_t be the price chosen in round t. Let $p^*_{\text{act}} \in \operatorname{argmax}_{p \in \mathcal{P}} \nu(p)$ be the best active price according to $\nu(\cdot)$, and let $\nu^*_{\text{act}} \triangleq \nu(p^*_{\text{act}})$. Let $\Delta(p) \triangleq \max(0, \frac{1}{n}\, \nu^*_{\text{act}} - p\, S(p))$ be our notion of "badness" of price p, compared to the optimal approximate revenue ν^*. We will use this notation throughout the analysis, and eventually we will bound regret in terms of $\sum_{p \in \mathcal{P}} \Delta(p)\, N(p)$, where $N(p)$ is the total number of times price p is chosen.

Claim 6. *For each price $p \in \mathcal{P}$ it holds that*

$$N(p)\, \Delta(p) \leq O(\log n) \left(1 + \tfrac{k}{n}\, \tfrac{1}{\Delta(p)}\right). \tag{10}$$

Proof. By definition (3) of the confidence radius, for each price $p \in \mathcal{P}$ and each round t we have

$$\nu(p) \leq I_t(p) \leq p \cdot \min\left(k,\, n\, \left(S(p) + 2\, r_t(p)\right)\right). \tag{11}$$

Let us use this to connect each choice p_t with ν^*_{act}:

$$\begin{cases} I_t(p_t) \geq I_t(p^*_{\text{act}}) \geq \nu(p^*_{\text{act}}) \triangleq \nu^*_{\text{act}} \\ I_t(p_t) \leq p_t \cdot \min\left(k,\, n\, \left(S(p_t) + 2\, r_t(p_t)\right)\right). \end{cases}$$

Combining these two inequalities, we obtain the key inequality:

$$\tfrac{1}{n}\, \nu^*_{\text{act}} \leq p_t \cdot \min\left(\tfrac{k}{n},\, S(p_t) + 2\, r_t(p_t)\right). \tag{12}$$

There are several consequences for p_t and $\Delta(p_t)$:

$$\begin{cases} p_t & \geq & \frac{1}{k}\,\nu^*_{\text{act}} \\ \Delta(p_t) & \leq & 2\,p_t\,r_t(p_t) \\ \Delta(p_t) > 0 & \Rightarrow & S(p_t) < \frac{k}{n} \end{cases}. \tag{13}$$

The first two lines in (13) follow immediately from (12). To obtain the third line, note that $\Delta(p_t) > 0$ implies $p_t\,k \geq \nu^*_{\text{act}} > n\,p_t\,S(p_t)$, which in turn implies $S(p_t) < \frac{k}{n}$.

Note that we have not yet used the definition (4) of the confidence radius. For each price $p = p_t$, let t be the last round in which this price has been selected by the pricing strategy. Note that $N(p)$ (the total number of times price p is chosen) is equal to $N_t(p) + 1$. Then using the second line in (13) to bound $\Delta(p)$, Eq. (7) to bound the confidence radius $r_t(p)$, and the third line in (13) to bound the survival rate, we obtain:

$$\Delta(p) \leq O(p) \times \max\left(\tfrac{\log n}{N(p)},\ \sqrt{\tfrac{k}{n}\,\tfrac{\log n}{N(p)}}\right).$$

Rearranging the terms, we can bound $N(p)$ in terms of $\Delta(p)$ and obtain (10). $\square$

Analyzing the total revenue. A key step is the following claim that allows us to consider $\sum_{t=1}^n p_t\,S(p_t)$ instead of the realized revenue $\widehat{\texttt{Rev}}$, effectively ignoring the capacity constraint. This is where we use the high-probability events (8) and (9). For brevity, let us denote $\beta(S) = O(\sqrt{S\log n} + \log n)$.

Claim 7. *$\widehat{\texttt{Rev}} \geq \min(\nu^*_{act},\ \sum_{t=1}^n p_t\,S(p_t)) - \beta(k)$.*

Proof. Recall that $p_t \geq \frac{1}{k}\nu^*_{\text{act}}$ by (13). It follows that $\widehat{\texttt{Rev}} \geq \nu^*_{\text{act}}$ whenever $\sum_{t=1}^n X_t > k$. Therefore, if $\widehat{\texttt{Rev}} < \nu^*_{\text{act}}$ then $\sum_{t=1}^n X_t \leq k$ and so $\widehat{\texttt{Rev}} = \sum_{t=1}^n p_t\,X_t$. Thus, by (9) it holds that

$$\begin{aligned}\widehat{\texttt{Rev}} &\geq \min\left(\nu^*_{\text{act}},\ \textstyle\sum_{t=1}^n p_t\,X_t\right)\\ &\geq \min\left(\nu^*_{\text{act}},\ \textstyle\sum_{t=1}^n p_t\,S(p_t) - \beta(S)\right).\end{aligned}$$

So the claim holds when $S \leq k$. On the other hand, if $S > k$ then by (8) it holds that

$$\begin{aligned} X &\geq S - \beta(S) \geq k - \beta(k)\\ \widehat{\texttt{Rev}} &\geq \min(k, X)\left(\tfrac{1}{k}\,\nu^*_{\text{act}}\right) \geq \nu^*_{\text{act}} - \beta(k). \end{aligned} \quad \square$$

In light of Claim 7, we can now focus on $\sum_{t=1}^n p_t\,S(p_t)$.

$$\begin{aligned}\textstyle\sum_{t=1}^n p_t\,S(p_t) &\geq \textstyle\sum_{t=1}^n \frac{1}{n}\,\nu^*_{\text{act}} - \Delta(p_t)\\ &= \nu^*_{\text{act}} - \textstyle\sum_{t=1}^n \Delta(p_t)\\ &= \nu^*_{\text{act}} - \textstyle\sum_{p\in\mathcal{P}} \Delta(p)\,N(p). \end{aligned} \tag{14}$$

Fix a parameter $\epsilon > 0$ to be specified later, and denote

$$\begin{cases}\mathcal{P}_{\text{sel}} & \triangleq \{p \in \mathcal{P}:\ N(p) \geq 1\}\\ \mathcal{P}_{\epsilon} & \triangleq \{p \in \mathcal{P}_{\text{sel}} : \Delta(p) \geq \epsilon\}\end{cases}$$

to be, respectively, be the set of prices that have been selected at least once and the set of prices of badness at least ϵ that have been selected at least once. Plugging (10) into (14):

$$\begin{aligned}&\textstyle\sum_{p\in\mathcal{P}} \Delta(p)\,N(p)\\ &\quad\leq \textstyle\sum_{p\in\mathcal{P}_{\text{sel}}\setminus\mathcal{P}_\epsilon} \Delta(p)\,N(p) + \sum_{p\in\mathcal{P}_\epsilon} \Delta(p)\,N(p)\\ &\quad\leq \epsilon n + O(\log n)\textstyle\sum_{p\in\mathcal{P}_\epsilon}\left(1 + \frac{k}{n}\,\frac{1}{\Delta(p)}\right)\\ &\quad\leq \epsilon n + O(\log n)\left(|\mathcal{P}_\epsilon| + \frac{k}{n}\textstyle\sum_{p\in\mathcal{P}_\epsilon}\frac{1}{\Delta(p)}\right).\end{aligned} \tag{15}$$

Combining (14), (15) and Claim 7 we obtain that

$$\begin{aligned}\nu^*_{\text{act}} - \mathbb{E}[\widehat{\texttt{Rev}}] &\leq \epsilon n + \beta(k) +\\ &\quad + O(\log n)\left(|\mathcal{P}_\epsilon| + \tfrac{k}{n}\textstyle\sum_{p\in\mathcal{P}_\epsilon}\frac{1}{\Delta(p)}\right).\end{aligned}$$

The above fact summarizes our findings so far. Interestingly, it holds for any set of active prices.

The following claim, however, takes advantage of the fact that the active prices are given by (6).

Claim 8. *$\nu^*_{act} \geq \nu^* - \delta k$, where $\nu^* \triangleq \max_p \nu(p)$.*

Proof. Let $p^* \in \operatorname{argmax}_p \nu(p)$ denote the best fixed price with respect to $\nu(\cdot)$, ties broken arbitrarily. If $p^* \leq \delta$ then $\nu^* \leq \delta k$. Else, letting $p_0 = \max\{p \in \mathcal{P} :\ p \leq p^*\}$ we have $p_0/p \geq \frac{1}{1+\delta} \geq 1 - \delta$, and so

$$\nu^*_{\text{act}} \geq \nu(p_0) \geq \tfrac{p_0}{p^*}\,\nu(p^*) \geq \nu^*(1-\delta) \geq \nu^* - \delta k. \quad \square$$

It follows that for any $\epsilon > 0$ and $\delta \in (0, 1)$ we have:

$$\begin{aligned}\texttt{Regret} &\leq O(\log n)\left(|\mathcal{P}_\epsilon| + \tfrac{k}{n}\textstyle\sum_{p\in\mathcal{P}_\epsilon}\frac{1}{\Delta(p)}\right) && (16)\\ &\quad + \epsilon n + \delta k + \beta(k). && (17)\end{aligned}$$

The rest is a standard computation. Plugging in $\Delta(p) \geq \epsilon$ for each $p \in \mathcal{P}_\epsilon$ in (16), we obtain:

$$\texttt{Regret} \leq O(|\mathcal{P}_\epsilon|\log n)\left(1 + \tfrac{1}{\epsilon}\,\tfrac{k}{n}\right) + \epsilon n + \delta k + \beta(k).$$

Note that $|\mathcal{P}| \leq \frac{1}{\delta}\,\log n$. To simplify the computation, we will assume that $\delta \geq \frac{1}{n}$ and $\epsilon = \delta\,\frac{k}{n}$. Then

$$\texttt{Regret} \leq O\left(\delta k + \tfrac{1}{\delta^2}(\log n)^2 + \sqrt{k\log n}\right). \tag{18}$$

Finally, it remains to pick δ to minimize the right-hand side of (18). Let us simply take δ such that the first two summands are equal: $\delta = k^{-1/3}\,(\log n)^{2/3}$. Then the two summands are equal to $O(k\log n)^{2/3}$.