Where to Sell:
Simulating Auctions From Learning Algorithms

Hamid Nazerzadeh[1], Renato Paes Leme[2], Afshin Rostamizadeh[2], and Umar Syed[2]

[1]USC Marshall School of Business
[2]Google Research NYC

June 1, 2016

Abstract

Ad Exchange platforms connect online publishers and advertisers and facilitate selling billions of impressions every day. We study these environments from the perspective of a publisher who wants to find the profit maximizing exchange to sell his inventory. Ideally, the publisher would run an auction among exchanges. However, this is not possible due to technological and other practical considerations. The publisher needs to send each impression to one of the exchanges with an asking price. We model the problem as a variation of multi-armed bandits where exchanges (arms) can behave strategically in order to maximizes their own profit. We propose mechanisms that find the best exchange with sub-linear regret and have desirable incentive properties.

1 Introduction

We investigate a setting in which running an auction would be desirable but practical business considerations prevent it. Instead, we seek to simulate the auction outcome using online learning algorithms. This problem is motivated in part by the applications in Internet advertising. Publishers sell the space on their webpages, often called slots, to advertisers. The values of different slots varies a lot and range from highly desirably premium inventory such as the front page of New York Times to very specialized properties, such as small blogs. Instead of selling the rights to advertise in those slots directly to advertisers, some publishers send their inventory (ad impressions) to advertising exchanges. Advertisement exchanges are auction platforms that connects publishers and advertisers. Examples of major exchanges include Google AdExchange, AppNexus, Rubicon, and Facebook Exchange. They sell billions of impressions every day [19, 23].

From the perspective of the publisher, his ideal world would be one in which there is a single exchange in which he has access to all advertisers interested in his impressions. This would generate a sufficiently competitive market that would allow him to extract the fair price for his inventory. However, unfortunately, the proliferation of advertisement exchanges has caused the market to be fragmented. For each ad impression, the publisher needs to decide to which exchange to send this impression and which reserve price to submit. A key question we aim to answer in this paper is the following: can a seller emulate a competitive market through online learning?

Our Model & Results We model the publisher's problem of finding the best exchange in a multi-armed bandit (MAB) setting. From this point on, we refer to the publisher as the *seller* and to each exchange as a *buyer*. The seller in each timestep chooses a buyer and offers the impression to him at a certain price. In MAB language, this correspond to pulling an arm that consists of a pair of a buyer and a price. The buyer then decides whether to accept or reject the seller's offer. If the buyer accepts and purchases the impression, the seller receives revenue equal to the price he quoted. Otherwise, she gets revenue zero.

So far, this is a standard multi-armed bandits problem for which standard algorithms provide already sublinear regret. The challenging aspect is that we are deploying this algorithm in a *market setting*, where buyers (arms) are strategic economic agents. Therefore, any successful algorithm must take into account incentives of the buyers. To this aim, we consider two types of buyers: *myopic* and *strategic*. A myopic buyer purchases an impression only when his valuation is above the current asking price. On the other hand, a strategic buyer may use complicated strategies in order to maximize his long-term utility. The seller does not know if a buyer is myopic or strategic. Since the ads ecosystem has buyers with different levels of sophistication, it is important for any practical algorithm to be agnostic to the type of the buyer. Having an algorithm that works for a mixture of myopic and strategic buyers will ensure that we will correctly deal with incentives, but will also prevent us from relying too much on perfect rationality of the buyers.

We observe that buyers' strategic behavior can affect the seller's revenue in two opposing directions. A more familiar aspect is similar to bid shading in (first-price) auctions. A buyer may not purchase impressions that he values above the current asking price, because he worries that the seller may learn that the buyer has high valuations and increases the price in the future.

On the flip side, the strategic behavior of the buyer may in fact increase the revenue of the seller. Namely, the buyer may purchase impressions at a loss in the hope of receiving more impressions in the future. The intuition is as follows. Any learning algorithms that suffers small-regret almost always sends impressions to a buyer from which it perceives that it can extract highest revenue (i.e., a good arm is pulled more often). In response to this, the buyer may accept seller's offers at a higher prices (even if they get negative utility for those particular impressions) to incentivize the seller to send more impressions to him in the future. At first glance, this phenomenon might appear as an artifact that comes out of equilibrium analysis. The effect, however, is real and measurable in the advertising exchange business. The determinant factor of a successful exchange is the ability to attract inventory. It is easier for an exchange with a large availability of inventory to attract buyers than the other way round. Given this fact, it is only natural that an exchange would accept higher prices for certain items in the hope of continuing to receive inventory from that particular seller.

In our setting, the learning algorithm designed by the seller induces a game among the strategic buyers. Our goal is to design a learning algorithm with sublinear regret for the seller when buyers play an ϵ-approximate equilibrium of the induced game.

As previously discussed, traditional MAB algorithms identify the arms (buyer and price pair) that generate higher revenue and pull them in most of the rounds. This corresponds to a first-price auction behavior, which given incentives to buyers to bid less. To address this issue, we propose two such mechanisms that combine standard MAB algorithms and the second price auction. Our first algorithm, called *Second Price Histogram*, consists of two phases, exploration and exploitation. During exploration, each arm is pulled a few times in order to estimate the distribution of the valuations of the buyers. Then, during exploitation, the item is assigned to the buyer with an arm that generates the highest estimated revenue. In order to induce an approximate equilibrium, we charge the buyer a price that generates the revenue equal to the highest revenue that can be obtained from the other buyers.

This design doesn't address that issue that buyers might behave in a certain way during the exploration and change their behavior during exploitation. In order to address that, we introduce the notion of "consistency checks". To make the algorithm robust with respect to deviations to dynamic strategies, we check for each arm if it is behaving in a way that is consistent with a static (historyindependent) strategy. If ever we realize that the behavior is not consistent, we never pull that arm again. The intuition is that consistency check basically eliminates the utility that can be obtained from deviation strategies where a buyer would pretend to have high valuations during exploration and then reduces the purchase rate, and subsequently the generated revenue, during exploitation. The mechanism may "mistakenly" stops allocating the items, but that happens in equilibrium with very small probability.

We show that a simple strategy, called *aggressive strategy*, is an $\widetilde{\mathcal{O}}(T^{-1/4})$-dominant strategy for the buyers, where T is the length of the time horizon. Under the aggressive strategy, a buyer accepts all prices below his expected value, even if the current realized valuations is below the offered price. We show that no other (possibly quite complicated) strategy can improve the expected average utility of the buyer by more than $\widetilde{\mathcal{O}}(T^{-1/4})$. Furthermore, the seller's regret, compared with the second-highest price benchmark, when all buyers play the aggressive strategy is at most $\widetilde{\mathcal{O}}(T^{3/4})$.

Our second mechanism is a variation of the UCB algorithm. The algorithm at each step keeps an estimate and a confidence interval for each arm and chooses to pull an arm that maximizes the upper confidence bound (UCB), which is the estimated expected value plus an error term. Similar to the previous algorithm, we charge the buyer the second highest UCB. More precisely, we charge the buyer the lowest among his prices where the UCB is still above the highest UCB of all other buyers.

We show that the mechanism induces an ϵ-approximate equilibrium for the buyers, for $\epsilon = \widetilde{\mathcal{O}}(T^{-1/6})$. Under this (aggressive) strategy profile, the mechanism has regret at most $\widetilde{\mathcal{O}}(T^{2/3})$.

Related Work The literature on pricing using learning algorithms has been growing over the past few year. [17] propose one of the first algorithm of this kind in a setting where the goal is to sell items to customers that arrive over time using posted-prices. The algorithm is a variation of UCB algorithm [3] where each arm corresponds to a posted-price. Under regularity assumptions, their algorithm obtains sub-linear (optimal) regret. This result has been extended to more general settings; see [1, 4, 7, 9, 26].

In the context of online advertising, [5, 6] and [13] study multi-armed bandit settings where each arm corresponds to an advertiser. Each advertiser knows the value they obtained from each click but *not* the probability of clicks (i.e., click-through rate). Each advertiser reports his private information (i.e., vale per click) at the very beginning to the mechanism and the MAB algorithms are used to learn the probably of the clicks. See [8] and [15] on game-theoretic Bayesian multi-armed bandit settings.

Another line of research related to ours is reserve-price optimization in repeated auctions. [12] and [20] look at the algorithmic aspects of optimizing reserve prices but they do not consider strategic behavior of the buyers. With this motivation, [2, 21] study the problem of selling items to a single strategic buyer repeatedly over time. However they assume that the buyer is impatient and has a time discounted utility compared to the seller. In a multi-buyer setting, [16] show that if the distributions of the valuations are correlated, then setting reserve price dynamically can in fact increase the revenue of the seller even if the buyers are strategic and patient.

2 Preliminaries

Consider a seller, a set of buyers B, with $n = |B|$, and a horizon of length T.[1]

For each buyer $b \in B$, his valuation at each timestep $t \in [T]$, denoted by $v_{b,t}$, is drawn independently from distribution $\mathbf{D}_b$ with support in the $[0, 1]$ interval and mean equal to $\mu_b = \mathrm{E}[v_{b,t}]$. The distributions of valuations are unknown to the seller.

[1] To simplify the presentation, we assume that T is known in advance. This assumption can be relaxed using standard techniques [22].

The decision faced by the seller in each timestep t is to choose a buyer $b_t \in B$ and a price $p_t \in [0,1]$. After the impression is offered to buyer b_t, he decides whether to accept or to reject the price. If he accepts, the seller receives revenue p_t and the buyer obtains utility $v_{b,t} - p_t$. To map this setting to our motivating application, suppose exchange b allocates the impression using an auction among the advertisers in this exchange. After receiving the publisher's price p, exchange b collects bids from the advertisers and runs an auction. The value $v_{b,t}$ of the exchange for this impression corresponds to the revenue that the exchange can obtain from it advertisers. The exchange can decide them either to accept or not the price, and upon accepting, the exchange pays p_t to the seller.[2]

Let $\mathbf{A}_t$ denote the event that the buyer purchases the impression. Hence, the total revenue of the seller is equal to:

$$\textsc{Rev} = \mathrm{E}\left[\sum_{t=1}^{T} p_t \cdot \mathbf{1}\{\mathbf{A}_t\}\right]$$

The seller's objective is maximize his total revenue. But he needs to take into account the buyers' incentives. We now look at the buyer's problem.

2.1 Buyer strategies, equilibria and ϵ-dominance

Let u_b denote the average utility of the buyer; namely,

$$u_b = \left[\frac{1}{T}\sum_{t=1}^{T}(v_{b,t} - p_t) \cdot \mathbf{1}\{b_t = b \text{ and } \mathbf{A}_t\}\right]$$

We consider two types of buyers, myopic and strategic. The type of the buyer in unknown to the seller.

Definition 1 (**Myopic Buyers**). *Myopic Buyers aim to maximize their profit form each impression, without taking into account the effects of their current action on the future allocations and prices. Myopic buyers simply purchase an impression whenever $p_t \le v_{t,b}$.*

Definition 2 (**Strategic Buyers**). *A* strategic buyer *tries to find a strategy that maximizes their long-term utility. A strategy determines buyer's policy on whether to accept of reject the seller's offer in response to the seller's mechanism and possibly other buyer's strategies. We assume a strategic buyer knows his distribution of valuations, $\mathbf{D}_b$, and hence μ_b.*

A buyer could deploy complicated history-dependent strategies. However, buyers may prefer simple strategies if they are near-optimal. We say that buyer b employs a **static policy** if his decision to purchase depends only on the price offered and his valuation. We define a special static policy which we call the **aggressive policy**, in which the buyers purchases an impression whenever $p_t \le \mu_b$.

We now define the notion of equilibrium.

Definition 3 (**ϵ-equilibrium**). *A profile Ω of buyers' strategies define an ϵ-equilibrium if no strategic buyer can change his policy to* any *other (possibly non-static) policy and improve his average utility by more than ϵ. More precisely, for any buyer b, we should have*

$$u_b(\Omega_b, \Omega_{-b}) \ge u_b(\Omega'_b, \Omega_{-b}) - \epsilon, \forall \Omega'_b$$

where Ω_b, Ω'_b, and Ω_{-b} respectively correspond to buyer b's equilibrium strategy, any possible deviation for buyer b, and strategies of other buyers.

[2] An alternative setting would be one in which the exchange may pay the publisher any amount higher than the price quoted; for instance, the second highest price (minus a revenue-share cut) if its higher than the price. Although we do not formally study this alternative model, our results can be extended there. Furthermore, we point out ad auctions are often thin and effectively have one buyer (cf., [11]); such environments fit our model well.

In this paper we will be typically interested in $o(T^{-\alpha})$-equilibria for $\alpha \in (0, 1)$. See [14, 24, 25] for further discussions on approximate and asymptotic notions of equilibrium in similar settings.

A stronger notion than ϵ-equilibrium is ϵ-dominance.

Definition 4 (**ϵ-dominance**). *We say that a strategy Ω_b is ϵ-dominant for buyer b if no matter what strategies the other buyers are employing, buyer b cannot improve his average utility by more then ϵ by deviating to any other (possibly non-static) policy. More precisely, for any buyer b, we should have*

$$u_b(\Omega_b, \Omega_{-b}) \geq u_b(\Omega'_b, \Omega_{-b}) - \epsilon, \forall \Omega'_b, \forall \Omega_{-b}$$

where Ω_b, Ω'_b, and Ω_{-b} respectively correspond to buyer b's equilibrium strategy, any possible deviation for buyer b, and strategies of other buyers.

Note that if every strategy in a profile is ϵ-dominant then this profiles forms an ϵ-equilibrium.

2.2 Revenue Benchmark

The maximum per-timestep revenue that can be extracted from a myopic buyer is $\bar{\rho}_b = \max_p p \cdot \Pr[v_b \geq p]$. For a strategic buyer, we will us his expected surplus per period as an upper bound $\bar{\rho}_b = \mu_b$. A natural upper bound on the total revenue is $T \times \max_b\{\bar{\rho}_b\}$. It is certainly possible to achieve sublinear regret with respect to this policy if all buyers are myopic. In the language of auction theory this corresponds to a first-price auction type of benchmark, which is known to not be achievable in strategic settings. Indeed, a buyer with large $\bar{\rho}_b$ will pretend that his value is lower to prevent the seller from extracting revenue from him, cf. [2, 16]. Inspired by the second-price auction, we choose the second-best solution as our benchmark; namely, the second highest value in $\{\bar{\rho}_b\}$. Assuming that the buyers are sorted such that

$$\bar{\rho}_1 \geq \bar{\rho}_2 \geq \ldots \geq \bar{\rho}_n$$

we denote the second highest value in $\{\bar{\rho}_b\}$ by $\bar{\rho}_2$. Another natural benchmark would have been the second highest $v_{b,t}$ which can be obtained *if* we could bring together all the buyers. However, this benchmark is infeasible in our setting. The main reason is that when the publisher offers an impression to an exchange, he cannot renege after the exchange accepts the impression and has to allocate. Therefore, the publisher cannot observe the realizations of $v_{b,t}$ and has to make decisions based on the estimated distributions or simply expected values of $v_{b,t}$. In appendix A, we discuss in detail the relation with this and other benchmarks.

Our main goal is to achieve sublinear regret with respect to this benchmark (this is often called *pseudo-regret*).

Definition 5 (**Regret**). *Given a strategy profile of the buyers, the regret is defined as:*

$$\textsc{Regret} = T \cdot \bar{\rho}_2 - \mathrm{E}\left[\sum_{t=1}^{T} p_t \cdot \mathbf{1}\{\mathbf{A}_t\}\right]$$

Formally, our goal is to design a learning algorithm for which there is a profile of policies in $o(T^{-\alpha})$-equilibrium such that $\textsc{Regret} \leq o(T)$.

2.3 Upper Confidence Bound (UCB) Algorithms

The algorithms we discuss in this paper are based on the concept of Upper Confidence Bound (UCB). Given s iid drawns $X_1, \ldots, X_s$ from a random variable with mean μ and support in $[0, 1]$, Hoeffding's inequality guarantees that:

$$\Pr\left[|\hat{\mu} - \mu| \geq \lambda/\sqrt{s}\right] \leq O(e^{-c\lambda^2}) \tag{HI}$$

for $\hat{\mu} = \frac{1}{s}\sum_1^s X_s$ for any $\lambda > 0$. In particular, taking $\lambda = \sqrt{a \cdot \log T}$ for some constant $a > 0$, we get that:

$$\Pr\left[|\hat{\mu} - \mu| \geq \sqrt{\tfrac{a\log(T)}{s}}\right] \leq O(T^{-ac}).$$

For any given algorithm, if buyers are employing a static policy, then the event of buyer b accepting price p is iid across timesteps. Therefore, we can build an estimate $\hat{r}_{b,p,t}$ of the revenue that can be collected from buyer b at time t. If we offered buyer b the item at price p a number of times $s_{b,p,t}$ before time t and from those he accepted $y_{b,p,t}$ impression, we can build the estimate

$$\hat{r}_{b,p,t} = p \cdot y_{b,p,t}/s_{b,p,t}$$

with error

$$\hat{\sigma}_{b,p,t} = \sqrt{\frac{a\log(T)}{s_{b,p,t}}}$$

and the confidence interval:

$$I_{b,p,t} = [\hat{r}_{b,p,t} - \hat{\sigma}_{b,p,t}, \hat{r}_{b,p,t} + \hat{\sigma}_{b,p,t}]$$

which holds with probability $1 - O(T^{-ac})$. We denote by $\textsc{Ucb}(b,p,t)$ and $\textsc{Lcb}(b,p,t)$ the upper and lower ends of the interval $I_{b,p,t}$. We omit the index t whenever it is clear from the context. Also, given a confidence interval I, we will often denote $b(I)$ and $p(I)$ for the buyer and price associated with it.

2.4 $\widetilde{\mathcal{O}}$-notation

In some of our results to improve readability we use the notation $\widetilde{\mathcal{O}}(T^\beta)$ to highlight the polynomial dependence of a certain expression with respect to T. This notation hides constants and dependencies poly-log terms in T. Formally, we say that $f = \widetilde{\mathcal{O}}(T^\beta)$ if $f = O(T^\beta \log^\gamma(T))$ for some constants γ and τ.

3 Histograms with Consistency Checks

3.1 Second Price Histogram Algorithm

We design a simple learning algorithm with incentive properties similar to those of the second price auction. Before we describe our algorithm, consider a version of this problem where incentives are ignored. Fix a static strategy for each buyer and a discretization parameter k. Based on k, construct a set of prices $P = \{\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}, 1\}$. Now, treat the problem as a (stochastic) multi-armed bandit problem in which each pair (b,p) with $b \in B$ and $p \in P$ corresponds to an arm. Also, the reward associated with (b,p) is the revenue obtained from offering price p to buyer b.

Our first algorithm for this setting, which we call *Histogram* consists of two phases. In the *exploration phase*, the algorithm pulls each arm (b,p) for h rounds. We can use the average reward obtained in those h rounds to build an estimate $\hat{r}_{b,p}$ of the reward that can be obtained from that arm. In the *exploitation phase*, the algorithm pulls the arm with the best estimated reward. If $hk = o(T)$, there is a single arm (b^*,p^*) that is pulled in all but sublinearly many rounds and this is the arm with largest empirical revenue.

The seller, therefore, identifies the arm that generates largest possible revenue and pulls it for the remainder of the algorithm. This corresponds, in auction theory language, to running a first-price auction. Similar to bid shading in first price auctions, in our setting charging the highest possible price would incentivize buyers to pretend to have lower valuations.

To address this issue we borrow ideas from the second price auction, which allocates the item to the highest bidder but charges him only the second highest bid. Making the price paid by a certain agent not depend on his actual bid is they key to design incentive compatible mechanisms. This idea is often called the *taxation principle*, where data from all buyers except b are used to determine the price offered to b. It can be shown that a mechanism is incentive compatible if and only if it can be described in terms of the taxation principle.

We propose a second price auction version of the Histogram algorithm: the algorithm first chooses the buyer b^* with largest estimated revenue, but offers him the smallest price p such that the estimated revenue is larger than the estimated revenue of any other buyer. Therefore even though we use the estimation of a buyer to choose the winner, we determine his price based on the estimation of the buyer with the second highest estimation.

```
Second Price Histogram
1: Pull $h$ times each arm $(b, p)$. Let $\hat{r}_{b,p}$ be the average reward obtained.
2: $(b^*, p^*) = \text{argmax}_{b \in B, p \in P}\{\hat{r}_{b,p}\}$.
3: $L = \max_{b \neq b^*, p}\{\hat{r}_{b,p}\}$.
4: $p' = \min\{p; \hat{r}_{b^*,p} \geq L\}$.
5: Pull arm $(b^*, p')$ for the remaining rounds.
```

The second price modification clearly does not address all incentive issues. For example, why shouldn't buyers behave in a certain way during the exploration phase and then in a different way during exploitation ? We will come back to this issue later. But before that, let's assume that buyers play a static strategy, i.e., their decisions on whether to accept the price or not depend only on the price offered in this timestep and the value in this timestep and *not* on the history of the auction. Our first instinct is to believe in such condition, myopic is an (at least approximately) optimal static policy, i.e., no other *static* policy would provide significant improvement over accepting whenever $p_t \leq v_t$. This is however not the case.

Consider two strategic buyers where the first has uniform valuation in $[0, 1]$ and the second one has valuation equal to $1/3$ deterministically. If both buyers respond myopically, then the algorithm will estimate the maximum revenue from buyer 1 to be around $1/4$ (when pricing at $p = 1/2$ and the maximum revenue from buyer 2 to be around $1/3$. It will cause the algorithm to choose buyer 2 in all but in a sublinear number of rounds, leaving buyer 1 with average utility $o(T)/T$. A good strategy for buyer 1 in this example is to accept all offers below $1/2$. This will entail accepting some offers below his value, but will cause the seller to have an estimate of $1/2$ for his revenue at $1/2$. Since the price that he will be offered will be around $1/3$, he will have average utility around $1/6 \pm o(T)/T$.

The aggressive policy turns out to be approximately optimal among static policies:

Theorem 1. *In the Second Price Histogram algorithm, the aggressive strategy is ϵ-dominant among* static *strategies for* $\epsilon = \widetilde{\mathcal{O}}\left(\frac{hk}{T} + \frac{1}{\sqrt{h}} + \frac{1}{k}\right)$ *. In other words, regardless of the strategies of other buyers, no buyer can improve his average utility by more than ϵ by deviating to another static strategy.*

Moroever, if all strategic buyers play aggressive strategies, the regret of the seller is bounded by $\widetilde{\mathcal{O}}\left(hk + \frac{T}{\sqrt{h}} + \frac{T}{k}\right)$.

Corollary 1. *For $h = \sqrt{T}$ and $k = T^{1/4}$, then it is an $\widetilde{\mathcal{O}}(T^{-1/4})$-dominant strategy for buyers to play the aggressive strategy and the seller's regret is at most $\widetilde{\mathcal{O}}(T^{3/4})$.*

3.2 Consistency Checks

The previous results show that unlike the standard Histogram algorithm, the Second Price Histogram guarantees that the aggressive strategy is an ϵ-equilibrium with respect to static policies. This algorithm, however, does not preclude buyers from pretending they can generate a high value in the exploration phase and once the buyer is chosen as b^* to switch to the myopic policy. If buyers were to play such non-static policies, the seller's regret could be arbitrarily bad. In order to address this issue, we introduce the notion of *consistency checks*.[3]

The idea of consistency checks is to force the buyer to play a strategy resembling a static strategy. The idea is as follows: if all buyers are playing static strategies, each arm has a well-defined average reward $\bar{r}_{b,p}$ and in each timestep t, if the arm has been pulled $s_{b,p,t}$ times and the price was accepted $y_{b,p,t}$ times, then with very high probability the average reward is in the interval: $I_{b,p,t} = [\hat{r}_{b,p,t} - \hat{\sigma}_{b,p,t}, \hat{r}_{b,p,t} + \hat{\sigma}_{b,p,t}]$ for $\hat{r}_{b,p,t} = p y_{b,p,t}/s_{b,p,t}$ and $\hat{\sigma}_{b,p,t} = \sqrt{\frac{a \log T}{s_{b,p,t}}}$. Therefore, if all buyer strategies are static, then with very high probability, the intersection of all confidence intervals for each arm $\cap_{t=1}^{T} I_{b,p,t}$ is non-empty since it contains $\bar{r}_{b,p}$.

We augment the algorithm by checking in each iteration t if $\cap_{\tau=1}^{t} I_{b,p,\tau} \neq \emptyset$. If so, we say that arm (b,p) is consistent at time t. If in any iteration we realize that the chosen arm (b^*, p') is no longer consistent, we stop allocating the item.

```
Consistent Second Price Histogram
Pull $h$ times each arm $(b,p)$. Let $\hat{r}_{b,p}$ be the average reward obtained.
$(b^*, p^*) = \text{argmax}_{b,p} \hat{r}_{b,p}$.
$L = \max_{b \neq b^*, p} \hat{r}_{b,p}$.
$p' = \min\{p; \hat{r}_{b^*,p} \geq L\}$.
While $\cap_{\tau=1}^{t} I_{b^*,p',\tau} \neq \emptyset$, pull arm $(b^*, p')$. If the intersection ever becomes empty,
    stop allocating the item altogether.
```

Theorem 2. *In the* Consistent *Second Price Histogram algorithm, the aggressive strategy is ϵ-dominant for $\epsilon = \widetilde{\mathcal{O}}\left(\frac{hk}{T} + \frac{1}{\sqrt{h}} + \frac{1}{k}\right)$. In other words, regardless of the strategies of other buyers, no buyer can improve his average utility by more than ϵ by deviating to another (possibly non-static) strategy.*

Moreover, if all strategic buyers play aggressive strategies, the regret of the seller is bounded by $\widetilde{\mathcal{O}}\left(hk + \frac{T}{\sqrt{h}} + \frac{T}{k}\right)$.

3.3 Splitting the probability space

In this section, we describe a common tool in the analysis of the stochastic bandits mechanisms proposed in the previous section. The execution of any learning algorithm on a fixed set of buyer policies is a random process: the randomness comes from the valuations of the buyers that are drawn randomly in each iteration and possibly from the policies employed by the buyers which can itself be randomized. Despite the randomness, the analysis of both regret and equilibrium will be mostly deterministic. This is accomplished by splitting the probability space in two: one part called Nice in which the random variables of interest respect appropriate confidence intervals and a part called Nasty which occurs with very small probability.

[3] See [10] who use consistency check ideas to design bandit mechanisms that perform well both in stochastic and adversarial settings.

Fix a profile of <u>*static*</u> policies for the buyers and let $z_{b,p,t}$ be the revenue obtained by pulling arm (b,p) at time t. For example, if buyer b is myopic, $z_{b,p,t} = p \cdot \mathbf{1}\{v_{b,t} \geq p\}$. If buyer b is strategic and employing an aggressive policy, $z_{b,p,t} = p \cdot \mathbf{1}\{v_{b,t} \geq \mu_b\}$. Since we assumed that the policies are static, then for fixed (b,p), the family $\{z_{b,p,t}\}_{t=1}^T$ consists of iid random variables. Now we are ready to define the average reward and estimated reward formally in terms of z.

The real average reward is given $\bar{r}_{b,p} = \mathrm{E}[z_{b,p,t}]$. In order to define the estimated reward, let $\tau_{b,p}^j$ be a random variable indicating the j-th time arm (b,p) is pulled by the algorithm. Recall that $s_{b,p,t}$ denotes the number of times that the arm is pulled and let $\bar{s}_{b,p} = \max_t s_{b,p,t}$ be the random variable indicating the total number of times this arm is pulled in the course of the algorithm. The estimated reward at time t is given by:

$$\hat{r}_{b,p,t} = \frac{1}{s_{b,p,t}} \sum_{j=1}^{s_{b,p,t}} z_{b,p,\tau_{b,p}^j}$$

Now, we are ready to define the event Nice as the event such that for all (b,p) and for all $s \leq \bar{s}_{b,p}$, it holds that:

$$\left|\bar{r}_{b,p} - \tfrac{1}{s}\textstyle\sum_{j=1}^{s} z_{b,p,\tau_{b,p}^j}\right| \leq \sqrt{\tfrac{a \log T}{s}} \tag{N1}$$

and:

$$\left|\mu_b - \tfrac{1}{s}\textstyle\sum_{j=1}^{s} v_{b,\tau_{b,p}^j}\right| \leq \sqrt{\tfrac{a \log T}{s}} \tag{N2}$$

We denote by Nasty the complement of Nice in the probability space. Notice that Nasty happens when *at least one* of the confidence intervals is not satisfied. The following result follows directly from Hoeffding's inequality (HI) in Section 2 and the Union Bound:

Lemma 1. $\Pr[\text{Nasty}] \leq O(nk/T^2)$ *when* $a = 4/c$*, where* c *is the constant in inequality (HI).*

A note on non-static buyers The events Nice and Nasty are defined when all buyers use *static* strategies. When we analyze a situation in which not all buyers are static, we abuse notation and still refer to Nice and Nasty meaning that inequality (N1) hold for all buyers that are using static strategies, if any, and inequality (N2) holds for all buyers. Lemma 1 still holds in this setting.

3.4 Proof of Regret in Theorems 1 and 2

We now prove the regret part of Theorems 1 and 2. Assume that all strategic buyers are playing aggressive strategies. First we consider the loss from discretizing the space of prices:

Loss from discretizing prices. Let $\tilde{\rho}_b = \max_p \bar{r}_{b,p}$. If we had infinitely many arms, one for each price $p \in [0,1]$, $\tilde{\rho}_b$ would be equal to $\bar{\rho}_b$ in the benchmark. Since we are only considering $p \in P$, we have potentially an error of at most $1/k$, i.e., $|\bar{\rho}_b - \tilde{\rho}_n| \leq 1/k$. Re-sorting the buyers such that $\tilde{\rho}_1 \geq \tilde{\rho}_2 \geq \tilde{\rho}_3 \geq \ldots$, we will define the discrete-regret as the difference between $T\tilde{\rho}_2$ and the revenue obtained by the algorithm. The regret is at most the discrete-regret plus T/k.

Loss from exploration rounds. Since we pull every arm h times, using the trivial bound for the loss in each iteration we have loss at most nkh across all exploration rounds.

Splitting the probability space. Now, we can bound the expectation of the discrete-regret by conditioning on Nice and Nasty

$$\mathrm{E}[\text{Regret}] = \mathrm{E}[\text{Regret}|\text{Nasty}] \cdot \Pr[\text{Nasty}] + \mathrm{E}[\text{Regret}|\text{Nice}] \cdot \Pr[\text{Nice}]$$

We use the crude bound of T for E[Regret|Nasty] since Nasty happens with negligible probability. By Lemma 1, the total contribution of Nasty to the regret is $O(nk/T) = \widetilde{\mathcal{O}}(1)$, for $k = \widetilde{\mathcal{O}}(T)$. Using the trivial bound for Pr[Nice] we get:

$$\mathrm{E}[\text{Regret}] \leq \widetilde{\mathcal{O}}(1) + \mathrm{E}[\text{Regret}|\text{Nice}]$$

Therefore, we ignore Nasty from now on and focus on bounding E[Regret|Nice].

Conditioning on Nice. Conditioned on Nice, no arm ever becomes inconsistent, so the Second Price Histogram algorithm and the Consistent Second Price Algorithm are identical. Notice that each buyer b has an arm (b, p_b) such that $\tilde{\rho}_b = \bar{r}_{b,p_b}$ and since we are conditioning on Nice, $\hat{r}_{b,p_b} \geq \bar{r}_{b,p_b} - \sqrt{\frac{a\log(T)}{h}}$. Therefore in the description of the algorithm we must have $L \geq \tilde{\rho}_2 - \sqrt{\frac{a\log(T)}{h}}$.

Since the arm (b^*, p') chosen by the algorithm has $\hat{r}_{b^*,p'} \geq L$ at the end of the exploration round, the average reward of this arm by the end of the algorithm must be at least $L - \sqrt{\frac{a\log(T)}{\bar{s}_{b^*,p'}}} \geq L - \sqrt{\frac{a\log(T)}{h}}$ since we are conditioning on Nice. Therefore the total loss per round is at most $2\sqrt{\frac{a\log(T)}{h}}$ which is a total loss of $\widetilde{\mathcal{O}}\left(\frac{T}{\sqrt{h}}\right)$.

Combining all losses. Combining the loss of $\widetilde{\mathcal{O}}(T/k)$ from discretization, the loss of nhk from the exploration rounds and the loss of $\widetilde{\mathcal{O}}(T/\sqrt{h})$ from the exploration rounds we get the regret in Theorems 1 and 2.

3.5 Proof of ϵ-dominance in Theorems 1 and 2

We now show that the aggressive strategy is ϵ-dominant, i.e., regardless of the strategies employed by other players, any given player can't improve his average utility by more than $\epsilon = \widetilde{\mathcal{O}}\left(\frac{hk}{T} + \frac{1}{\sqrt{h}} + \frac{1}{k}\right)$ by deviating from the aggressive strategy. First we prove this for the Consistent Second Price Histogram (Theorem 2) and remark that Theorem 1 is a special case.

First we bound the utility that buyer b can get by playing the aggressive strategy:

Lemma 2. *Fix an arbitrary strategy profile for players $b' \neq b$ and let $\theta = \max_{b'\neq b,p} \hat{r}_{b',p}$ be the random variable indicating the maximum estimated revenue for all buyers except b and let $\delta = \sqrt{\frac{a\log(T)}{h}} + \frac{1}{k}$. Then the average utility of buyer b by playing the aggressive strategy is at least:* $\mathrm{E}\left[\mu_b - \theta - \delta\right]^+ - 2\delta - \widetilde{\mathcal{O}}(hk/T)$.

Proof. The total utility of buyer b is at least -1 in each of the hk exploration steps. To bound the expected total utility of buyer b in the remaining timesteps, notice that the expected utility he can get on Nasty is negligible (using the same argument used for regret in the previous subsection), so we bound his expected utility conditioned on Nice. Further condition on θ. One of two things can happen:

Case 1 Buyer b is selected as b^*. The price charged by the algorithm in exploration is therefore at most $\theta + \frac{1}{k}$, so the total utility of the buyer per round during the exploitation phase is at least the sum of his values during this phase minus the product of $\theta + \frac{1}{k}$ and the number of rounds in this phase. Since we are conditioning on Nice, we can use condition (N2) for arm (b^*, p') and $s = \bar{s}_{b^*,p'}$ the number of times arm (b^*, p') has been pulled. This condition implies that the average value of the buyer per round is at least $\mu_b - \sqrt{\frac{a\log(T)}{h}}$. Conditioned on this case, the total utility is at least: $-hk + T(\mu_b - \sqrt{\frac{a\log(T)}{h}} - (\theta + \frac{1}{k})) = -hk + T(\mu_b - \theta - \delta)$.

Since the buyer is selected, we must have $\mu_b \geq \theta - \delta$, otherwise the buyer b couldn't have been selected since we are conditioning on Nice. Therefore the total utility is at least $-hk - 2\delta T$. Combining those facts we get that the total utility is at least $-hk + T \cdot [(\mu_b - \theta - \delta)^+ - 2\delta]$, since when $\mu_b \geq \theta + \delta$ we can use the bound $-hk + T(\mu_b - \theta - \delta)$ and when $\mu_b \leq \theta + \delta$ we can use the bound $-hk - 2\delta T$.

Case 2 Buyer b is not selected as b^*. Then it must be that $\mu_b \leq \theta + \delta$, otherwise, since we are conditioning on Nice, buyer b would have been selected (despite discretization and sampling errors). Therefore, the total utility of the buyer is at least $-hk$ and $\mu_b - \theta - \delta \leq 0$. So the total utility is at least $-hk = -hk + T \cdot (\mu_b - \theta - \delta)^+ \geq -hk + T \cdot [(\mu_b - \theta - \delta)^+ - 2\delta]$.

Therefore in either case, the total utility is at least $-hk + T \cdot [(\mu_b - \theta - \delta)^+ - 2\delta]$. The lemma follows by taking expectations over θ and dividing by T to obtain the average utility. □

Lemma 3. *Fix an arbitrary strategy profile for players $b' \neq b$ and let θ and δ be as in the previous lemma. Then the utility of the buyer by playing any (possibly non-static) strategy is at most* $\mathrm{E}[(\mu_b - \theta + \delta)^+ + 2\delta] + \widetilde{O}(hk/T)$.

Proof. Fix some arbitrary, possibly non-static, strategy for buyer b. We will now upper bound his total utility for this deviation. Again, we ignore the utility that the buyer can get at Nasty since it is negligible, so we focus on Nice. Conditioning on θ, we have that either:

Case 1 Buyer b is selected as b^*. In such case, the estimation $\hat{r}_{b^*,p'} \geq \theta$, so since the confidence interval for arm (b^*, p') must have radius $\sqrt{\frac{a\log(T)}{h}}$, all the points in the confidence interval must be above $\theta - \sqrt{\frac{a\log(T)}{h}}$. Let s be the number of times the arm has been pulled throughout the algorithm. By the consistency rule, there must be x in the intersection of all of the confidence intervals before the last time the arm was pulled. Since the confidence interval just after exploration lies above $\theta - \sqrt{\frac{a\log(T)}{h}}$, then we must have $x > \theta - \sqrt{\frac{a\log(T)}{h}}$.

In particular x is in the confidence interval of arm (b^*, p') just before it is pulled. Therefore, the empirical average revenue from this arm must be at most $\sqrt{\frac{a\log T}{s-1}}$ away from x. In particular in the notation of equation (N1):

$$\left| \frac{1}{s-1} \sum_{j=1}^{s-1} z_{b^*,p',\tau^j_{b^*,p'}} - x \right| \leq \sqrt{\frac{a\log T}{s-1}}$$

Therefore the total payment of buyer b^* across all times arm (b^*, p') was pulled is at least $(s-1)(x - \sqrt{\frac{a\log T}{s-1}}) \geq (s-1)(\theta - \sqrt{\frac{a\log(T)}{h}} - \sqrt{\frac{a\log T}{s-1}}) \geq (s-1)(\theta - 2\sqrt{\frac{a\log(T)}{h}})$. We can now use condition (N2) at the last time s pulled to claim that the total value obtained from the buyer to those items is at most $s\left(\mu_b + \sqrt{\frac{a\log T}{s}}\right)$. So the total utility from pulling arm (b^*, p') is at most $s(\mu_b - \theta) + 3\sqrt{\frac{a\log T}{h}} + 1 \leq T(\mu_b - \theta + 3\delta) + 1$. The utility he can get from other arms (b^*, p) for $p \neq p'$ in exploration is at most $h(k-1)$.

Case 2 Buyer b is not selected as b^*. He can get utility at most hk from the exploration phase, since he won't be selected in exploration.

In either case, the utility of the buyer is at most $T[(\mu_b - \theta + \delta)^+ + 2\delta] + hk$. Dividing by T and taking expectations over T we obtained the result in the lemma. □

Now we are ready to prove the incentives part of Theorems 1 and 2:

Proof of ϵ-dominance in Theorems 1 and 2. By switching from the aggressive strategy to any other strategy, the gain in utility is at most $\Big[\mathrm{E}[(\mu_b - \theta + \delta)^+ + 2\delta] + \widetilde{\mathcal{O}}(hk/T)\Big] - \Big[\mathrm{E}[(\mu_b - \theta - \delta)^+ - 2\delta] - \widetilde{\mathcal{O}}(hk/T)\Big] = \widetilde{\mathcal{O}}\left(\delta + \frac{hk}{T}\right) = \widetilde{\mathcal{O}}\left(\frac{hk}{T} + \frac{1}{\sqrt{h}} + \frac{1}{k}\right)$. □

4 Second UCB Auction

In this section, we design a learning algorithm that combines the learning properties of the standard UCB algorithm with the incentive properties of a second price auction. The algorithm maintains an estimate and a confidence interval for each buyer-price pair. At each time step, the algorithm first chooses buyer b^* with the largest upper bound for any of its confidence intervals (a.k.a, upper confidence bound, or UCB), but offers him the smallest price p such that $\textsc{Ucb}(b^*, p)$ is larger than the UCB of any other buyer. Therefore even though we use the UCB of a buyer to choose the winner, we determine his price based on the UCB of the buyer with second highest UCB. As in a second price auction, only offering buyers prices determined by other buyers helps to address incentive issues, while continually updating the confidence intervals leads to lower regret than the histogram algorithm from the previous section.

```
Second UCB Auction
$k \leftarrow \left(\frac{T}{n \log T}\right)^{1/3}$ and $P \leftarrow \left\{\frac{1}{k}, \frac{2}{k}, \ldots, 1\right\}$.
For each $(b, p) \in B \times P$ let $b_t = b$ and $p_t = p$ for $T^{1/3}$ time steps.
for $t = nkT^{1/3} + 1, \ldots, T$ do
    $b^* = \arg\max_b \max_{p \in P} \textsc{Ucb}(b, p, t)$.
    $L_t = \max_{b \neq b^*, p \in P} \textsc{Ucb}(b, p, t)$.
    $p^- = \arg\min\{p \in P; \textsc{Ucb}(b^*, p, t) \geq L_t\}$.
    Let $b_t = b^*$ and $p_t = p^-$.
    If $\cap_{\tau=1}^{t} I_{b^*, p^-, \tau} = \emptyset$ then stop allocating the item altogether.
end for
```

Like in the Histogram algorithm, myopic strategy is not an optimal policy for strategic buyers. We will show that the policy in which buyers apply an aggressive strategy (i.e. they accept all prices with $p_t \leq \mu_b$) is an $\widetilde{\mathcal{O}}(T^{-1/6})$-equilibrium. Before discussing incentives, we show that under this policy, the algorithm has sublinear regret.

4.1 Regret Analysis

Theorem 3. *If strategic buyers play aggressive strategies, then the* Second UCB Auction *algorithm has regret bounded by $\widetilde{\mathcal{O}}(T^{2/3})$.*

Proof. Let $H = nkT^{1/3} + 1$ be the first time step of the algorithm's **for** loop, and $\tilde{\rho}_b = \max_{p \in P} \bar{r}_{b,p}$, where P is the set of prices used by the algorithm. We have

$$\mathrm{E}[\textsc{Regret}] \triangleq \mathrm{E}\left[\sum_{t=1}^{T} (\bar{\rho}_2 - p_t) \mathbf{1}\{\mathbf{A}_t\}\right] \leq T|\bar{\rho}_2 - \tilde{\rho}_2| + H + \mathrm{E}\left[\sum_{t=H}^{T} (\tilde{\rho}_2 - p_t) \mathbf{1}\{\mathbf{A}_t\}\right], \tag{1}$$

where the last sum is the algorithm's 'discrete regret', which can be decomposed into two terms based on whether the event Nice occurs:

$$\begin{aligned}
\mathrm{E}\left[\sum_{t=H}^{T}(\tilde{\rho}_2 - p_t)\mathbf{1}\left\{\mathbf{A}_t\right\}\right] &\leq \mathrm{E}\left[\sum_{t=H}^{T}(\tilde{\rho}_2 - p_t)\mathbf{1}\left\{\mathbf{A}_t\right\} \;\middle|\; \textsc{Nice}\right]\Pr[\textsc{Nice}] + T\Pr[\textsc{Nasty}] \\
&\leq \mathrm{E}\left[\sum_{t=H}^{T}(\tilde{\rho}_2 - p_t)\mathbf{1}\left\{\mathbf{A}_t\right\} \;\middle|\; \textsc{Nice}\right] + O(1). \qquad (2)
\end{aligned}$$

The first inequality used $\sum_{t=H}^{T}(\tilde{\rho}_2 - p_t)\mathbf{1}\left\{\mathbf{A}_t\right\} \leq T$ and the second inequality follows from $\Pr[\textsc{Nasty}] \leq O(\frac{1}{T})$, which we proved in Section 3.3. Now we can bound the discrete regret of the algorithm conditioned on Nice as follows:

$$\begin{aligned}
\mathrm{E}\left[\sum_{t=H}^{T}(\tilde{\rho}_2 - p_t)\mathbf{1}\left\{\mathbf{A}_t\right\} \;\middle|\; \textsc{Nice}\right] &= \mathrm{E}\left[\sum_{b,p}\sum_{t=H}^{T}(\tilde{\rho}_2 - p)\mathbf{1}\left\{\mathbf{A}_t, b_t = b, p_t = p\right\} \;\middle|\; \textsc{Nice}\right] \\
&\leq \sum_{b,p}\Delta_{b,p}E\left[\sum_{t=1}^{T}\mathbf{1}\left\{b_t = b, p_t = p\right\} \;\middle|\; \textsc{Nice}\right] \qquad (3)
\end{aligned}$$

where the inequality follows from the definition $\Delta_{b,p} \triangleq \max\{0, \tilde{\rho}_2 - \bar{r}_{b,p}\}$ and the fact that $\mathrm{E}[(\tilde{\rho}_2 - p)\mathbf{1}\left\{\mathbf{A}_t, b_t = b, p_t = p\right\} \mid \textsc{Nice}] \leq \tilde{\rho}_2 - \mathrm{E}[p\mathbf{1}\left\{\mathbf{A}_t, b_t = b, p_t = p\right\} \mid \textsc{Nice}] = \tilde{\rho}_2 - \bar{r}_{b,p}$.

We will now upper bound Eq. (3). Observe that if the event Nice occurs we have

$$\bar{r}_{b,p} \leq \textsc{Ucb}(b,p,t) \leq \bar{r}_{b,p} + 2\sqrt{\frac{a\log(T)}{s_{b,p,t}}}$$

for all (b,p,t). This implies $\max_{p\in P}\textsc{Ucb}(b,p,t) \geq \max_{p\in P}\bar{r}_{b,p} \triangleq \tilde{\rho}_b$, and thus

$$L_t \triangleq \max_{b\neq b^*, p\in P}\textsc{Ucb}(b,p,t) \geq \tilde{\rho}_2 \qquad (4)$$

for any buyer b^*. Also, if the event Nice occurs and $s_{b,p,t} > 4a\log(T)/\Delta_{b,p}^2$ then

$$\textsc{Ucb}(b,p,t) \leq \bar{r}_{b,p} + 2\sqrt{\frac{a\log(T)}{s_{b,p,t}}} < \bar{r}_{b,p} + \Delta_{b,p} \triangleq \tilde{\rho}_2. \qquad (5)$$

By the definition of the algorithm,

$$\textsc{Ucb}(b_t, p_t, t) \geq L_t \qquad (6)$$

for each time step t. Recall that $s_{b,p,t} = \sum_{\tau=1}^{t}\mathbf{1}\left\{b_\tau = b, p_\tau = p\right\}$. Combining (4), (5) and (6) we have

$$\begin{aligned}
&E\left[\sum_{t=1}^{T}\mathbf{1}\left\{b_t = b, p_t = p\right\} \;\middle|\; \textsc{Nice}\right] = E\left[\sum_{t=1}^{T}\mathbf{1}\left\{b_t = b, p_t = p, \textsc{Ucb}(b,p,t) \geq L_t\right\} \;\middle|\; \textsc{Nice}\right] \\
&\leq \ell + E\left[\sum_{t=\ell}^{T}\mathbf{1}\left\{b_t = b, p_t = p, \textsc{Ucb}(b,p,t) \geq L_t, s_{b,p,t} \geq \ell\right\} \;\middle|\; \textsc{Nice}\right] \leq \frac{4a\log T}{\Delta_{b,p}^2} \qquad (7)
\end{aligned}$$

where $\ell = 4a\log(T)/\Delta_{b,p}^2$.

Now let $A^- = \{(b,p) \in B \times P; \Delta_{b,p} < \Delta\}$ and $A^+ = \{(b,p) \in B \times P; \Delta_{b,p} \geq \Delta\}$, for a constant $\Delta > 0$ to be chosen later. Eq (7) implies

$$\sum_{b,p} \Delta_{b,p} E\left[\sum_{t=1}^{T} \mathbf{1}\{b_t = b, p_t = p\} \;\middle|\; \textsc{Nice}\right] \leq \sum_{(b,p)\in A^-} s_{b,p,T}\Delta + \sum_{(b,p)\in A^+} \frac{4a\log T}{\Delta} \tag{8}$$

Finally, combining (1), (2), (3) and (8) we have

$$\mathrm{E}[\textsc{Regret}] \leq T|\bar{\rho}_2 - \tilde{\rho}_2| + nkT^{1/3} + T\Delta + \frac{nk4a\log T}{\Delta} + O(1)$$

Choosing $\Delta = \sqrt{nk\log(T)/T}$, and observing that $k = \left(\frac{T}{n\log T}\right)^{1/3}$ and $|\bar{\rho}_2 - \tilde{\rho}_2| \leq \frac{1}{k}$, proves the theorem. □

4.2 Equilibrium Analysis

We use similar techniques as the one used in Section 3 to show that it is an ϵ-equilibrium for buyers to play the aggressive strategy. We do so by bounding the utility a strategic buyer can obtain by playing the aggressive strategy and then using consistency checks to argue that they can't improve their utility by much by deviating. The proof can be found in Appendix B.

Theorem 4. *The profile of buyer policies in which all strategic buyers play an aggressive strategy is an $\widetilde{O}(T^{-1/6})$-equilibrium.*

5 Discussion and Future Directions

In this paper, we showed that a UCB learning algorithm for optimizating the seller's revenue can be modified in such a way that simple buyer strategies will induce approximate-equilibria. An alternative question would be to analyze the equilibria of the standard UCB or other common learning algorithms. This would be the learning theoretic equivalent of studying the set of equilibria of first price auctions.

From a practical perspective, an important generalization would be the case where the publisher can send the impression to another exchange, if the selected exchange rejects the offered price. Since the publisher must display an ad in milliseconds, the publisher can try a very small number of exchanges. We believe the ideas we developed in this paper can pave the way for more general settings.

In the following subsections we discuss in more detail an important direction for future research, namely, characterizing the trade-off between the seller's regret and buyers' incentives.

Buyer-Seller Trade-offs

An important avenue of investigation is to study the trade-offs between seller's regret and buyer's utility. In the previous sections, we evaluated our algorithms with respect to their regret and buyers incentives, $O(T^\alpha)$-regret and $O(T^{\beta-1})$-equilibrium for respectively $(\alpha,\beta) = (3/4, 3/4)$ and $(\alpha,\beta) = (2/3, 5/6)$. A major open problem is to characterize the pairs (α,β) for which learning algorithms exist with the desired regret and incentive properties.

In this section, we discuss an additional formulation in terms of *buyer's penalty*: we establish a benchmark for buyer's utility and measure the loss that each learning algorithm induces for each buyer according to this benchmark. We establish a trade-off between those quantities:

Definition 6 (Buyer Penalty). *Given buyer b with highest μ_b playing a fixed policy, let p^* denote the price at which the buyer generate the second-highest revenue benchmark:* $\mathrm{E}\left[p^* \cdot \mathbf{1}\{\mathbf{A}_t\} \mid b_t = b\right] = \bar{\rho}_2$. *We define the* buyer penalty*, with respect to a seller mechanism M that pulls the arms (b_t, p_t) at iteration t, to be*

$$\mathrm{E}\Big[\sum_{t=1}^{T}(v_{b,t} - p^*) - \sum_{t=1}^{T}(v_{b,t} - p_{b,t}) \cdot \mathbf{1}\{b_t = b \wedge \mathbf{A}_t\}\Big]\,.$$

In other words, this is the difference between the utility gained by the buyer that is asked to generate the second-highest revenue benchmark on every round in expectation and the utility gained in the presence of the seller mechanism M.

The following theorem can be used to show a trade-off between seller regret and buyer penalty. The main idea of the proof (found in the appendix) is to use an anti-concentration bound for the binomial distribution to show that at least a certain number of samples from the second highest buyer are necessary to build a good estimate of the benchmark ρ_2.

Theorem 5. *Let B be a set that contains a mixture of myopic buyers and strategic buyers with value distributions that have support over $[0, 1]$. Then for any seller mechanism, there exists a setting where at least one of the following holds for any $0 < \alpha \leq 1/3$ with probability at least δ:*

1. *The seller incurs a regret of $\Omega(T^{1-\alpha})$.*
2. *The top buyer suffers a buyer penalty of $\Omega(\log(1/\delta)T^{2\alpha})$.*
3. *At least one buyer is not playing an aggressive strategy.*

The main implication of this theorem is that if all strategic buyers are playing the aggressive policy at an approximate equilibrium, then it cannot be the case that both seller regret and buyer penalty are small. In particular, if the seller mechanism incurs a regret of at most $o(T^{1-\alpha})$ and strategic buyers play the aggressive strategy at equilibrium, it must be the case that a winning strategic buyer is willing to accept a buyer penalty of at least $\Omega(T^{2\alpha})$. Conversely, if a strategic buyer allows for no more than $o(T^{2\alpha})$ buyer penalty before deviating when playing the aggressive policy, then the seller must necessarily suffer $\Omega(T^{1-\alpha})$ regret if the mechanism wishes to induce an approximate equilibirum where strategic buyers use the aggressive policy.

Acknowledgment Hamid Nazerzadeh's work was supported by a Faculty Research Award.

References

[1] Shipra Agrawal and Nikhil R. Devanur. 2014. Bandits with Concave Rewards and Convex Knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. 989–1006.

[2] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. 2013. Learning Prices for Repeated Auctions with Strategic Buyers. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*. 1169–1177.

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

[4] Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. 2012. Dynamic Pricing with Limited Supply. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. 74–91.

[5] Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. 2010. Truthful Mechanisms with Implicit Payment Computation. In *ACM Conference on Electronic Commerce*.

[6] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2009. Characterizing truthful multi-armed bandit mechanisms. In *ACM Conference on Electronic Commerce*. 79–88.

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

[8] Dirk Bergemann and Juuso Välimäki. 2010. The Dynamic Pivot Mechanism. *Econometrica* 78 (2010), 771–789.

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

[10] Sébastien Bubeck and Aleksandrs Slivkins. 2012. The Best of Both Worlds: Stochastic and Adversarial Bandits. In *The 25th Annual Conference on Learning Theory (COLT), June 25-27, 2012, Edinburgh, Scotland*. 1–23.

[11] L. Elisa Celis, Gregory Lewis, Markus Mobius, and Hamid Nazerzadeh. 2014. Buy-it-Now or Take-a-Chance: Price Discrimination through Randomized Auctions. *Management Science* (2014).

[12] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2013. Regret Minimization for Reserve Prices in Second-Price Auctions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1190–1204.

[13] Nikhil R. Devanur and Sham M. Kakade. 2009. The price of truthfulness for pay-per-click auctions. In *ACM Conference on Electronic Commerce*. 99–106.

[14] Jason Hartline, Vasilis Syrgkanis, and Eva Tardos. 2015. No-Regret Learning in Bayesian Games. In *Advances in Neural Information Processing Systems*. 3043–3051.

[15] Sham M. Kakade, Ilan Lobel, and Hamid Nazerzadeh. 2013. Optimal Dynamic Mechanism Design and the Virtual Pivot Mechanism. *Operations Research* 61, 4 (2013), 837–854.

[16] Yash Kanoria and Hamid Nazerzadeh. 2014. Dynamic Reserve Prices for Repeated Auctions: Learning from Bids - Working Paper. In *Web and Internet Economics - 10th International Conference (WINE)*. 232.

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

[18] Jiří Matoušek and Jan Vondrák. 2001. The probabilistic method. *Lecture Notes, Department of Applied Mathematics, Charles University, Prague* (2001).

[19] R Preston McAfee and Sergei Vassilvitskii. 2012. An overview of practical exchange design. *Current Science(Bangalore)* 103, 9 (2012), 1056–1063.

[20] Mehryar Mohri and Andres Muñoz Medina. 2014a. Learning Theory and Algorithms for revenue optimization in second price auctions with reserve. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*. 262–270.

[21] Mehryar Mohri and Andres Muñoz Medina. 2014b. Revenue Optimization in Posted-Price Auctions with Strategic Buyers. *NIPS* (2014).

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of machine learning.* MIT press.

[23] S. Muthukrishnan. 2009. Ad Exchanges: Research Issues. In *Internet and Network Economics, 5th International Workshop (WINE).* 1–12.

[24] Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. 2013. Dynamic Cost-Per-Action Mechanisms and Applications to Online Advertising. *Operations Research* 61, 1 (2013), 98–111.

[25] Denis Nekipelov, Vasilis Syrgkanis, and Eva Tardos. 2015. Econometrics for learning agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation.* ACM, 1–18.

[26] Zizhuo Wang, Shiming Deng, and Yinyu Ye. 2014. Close the Gaps: A Learning-While-Doing Algorithm for Single-Product Revenue Management Problems. *Operations Research* 62, 2 (2014), 318–331.

A Comparison to Other Revenue Benchmarks

A typical benchmark in Online Learning is to compare against the optimal arm, i.e., $\bar{\rho}_1$. Such benchmark, however, is not achievable in a strategic setting. Even when all buyers have deterministic valuations $v_{b,t} = \mu_b$, if $\mu_1 \gg \mu_2$, then buyer 1 can act as if his value were $\mu_2 + \epsilon$ and any algorithm with sublinear regret must allocate to buyer 1 all but a sublinear number of times charging him at most $\mu_2 + \epsilon$.

Another benchmark is the revenue that is obtained by the second-price auction *if* we could bring together all the buyers, which is $T \cdot \mathrm{E}[\textsc{SMax}_b v_b]$, where $\textsc{SMax}$ corresponds to the second maximum valuation. Such a benchmark could be too strong in our setting (in particular when the number of the buyers is large). The main reason that such benchmark is infeasible is that after the publisher offers an impression to an exchange and exchanges accepts, the publisher cannot reneg and not allocate. Therefore, the seller cannot observe the realization of $v_{b,t}$ but has to make decisions based on the estimated distribution or simply expected value of $\mathrm{E}[v_{b,t}]$. : consider for example n buyers with uniform valuations over $[0, 1]$. The expected revenue of a the second price auction is $\mathrm{E}[\textsc{SMax}_b v_b] = 1 - O(1/n)$. In our setting, however, since the seller chooses the buyers to offer the good before observing valuations (which are drawn independently of the seller's decision), the overall revenue from any algorithm can be at most the sum of the valuations of the selected buyers, which is $T \cdot \mathrm{E}[v_{b,t}] = \frac{1}{2} \cdot T$.

If all buyers are strategic, our benchmark becomes $T \cdot \textsc{SMax}\, \mathrm{E}[v_b]$. This benchmark is incomparable to the second price auction benchmark $T \cdot \mathrm{E}[\textsc{SMax} v_b]$. The previous paragraph shows an example where $\mathrm{E}[\textsc{SMax} v_b] \geq \textsc{SMax}\, \mathrm{E}[v_b]$. For an example where the opposite inequality holds, consider two buyers with iid valuations $v = 1$ with probability p and $v = p$ with probability $1 - p$. Then: $\textsc{SMax}\, \mathrm{E}[v_b] = p + (1-p)p \geq p^2 + (1-p^2)p = \mathrm{E}[\textsc{SMax} v_b]$ for any $0 < p < 1$.

B Proof of Theorem 4

Proof of Theorem 4. Consider a strategy profile $\Omega = (\Omega_b, \Omega_{-b})$ in which all strategic buyers other than

b play the aggressive strategy and buyer b plays an arbitrary (and possibly non-static) strategy. Define

$$U_{b,p} \triangleq \sum_{t=1}^{T} (v_{b,t} - p) \cdot \mathbf{1}\left\{b_t = b, p_t = p, \mathbf{A}_t\right\}$$

to be the utility that buyer b accrues from timesteps in which the algorithm offers price p to buyer b. We will begin by proving an upper bound on $\mathrm{E}[U_{b,p}]$, for each $(b,p) \in B \times P$.

Define $\tilde{\rho}_b = \max_{p \in P} \bar{r}_{b,p}$ for any buyer $b \in B$. For each $(b,p) \in B \times P$ choose the largest $x_{b,p} \in \cap_{t=1}^{T} I_{b,p,t}$, and define the event

$$\textsc{Near}(b,p) \triangleq \{x_{b,p} + 2\delta \geq \tilde{\rho}_2\},$$

where $\delta > 0$ is a constant that will be chosen later. Also, let $\textsc{Far}(b,p)$ be the complement of $\textsc{Near}(b,p)$.

We can decompose the expected utility $\mathrm{E}[U_{b,p}]$ as follows:

$$\begin{aligned}
\mathrm{E}[U_{b,p}] &\leq \mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Near}(b,p)] + \mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Far}(b,p)] \\
&\quad + \mathrm{E}[U_{b,p} \mid \textsc{Nasty}] \Pr[\textsc{Nasty}] \\
&\leq \mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Near}(b,p)] + \mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Far}(b,p)] + O(1) \qquad (9)
\end{aligned}$$

where the inequality used $U_{b,p} \leq T$ and $\Pr[\textsc{Nasty}] \leq O(\frac{1}{T})$, which we proved in Section 3.3. We will first upper bound $\mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Near}(b,p)]$, and then upper bound $\mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Far}(b,p)]$.

We have

$$\begin{aligned}
U_{b,p} &= \sum_{t=1}^{T} v_{b,t} \mathbf{1}\left\{b_t = b, p_t = p\right\} - \sum_{t=1}^{T} p \cdot \mathbf{1}\left\{b_t = b, p_t = p, \mathbf{A}_t\right\} \\
&= s_{b,p,T} \hat{v}_{b,p,T} - s_{b,p,T} \hat{r}_{b,p,T} \qquad (10)
\end{aligned}$$

Now we will upper bound $s_{b,p,T} \hat{v}_{b,p,T}$ and lower bound $s_{b,p,T} \hat{r}_{b,p,T}$, conditioned on $\textsc{Near}(b,p)$ and $\textsc{Nice}$ occuring.

If $\textsc{Nice}$ occurs then $\hat{v}_{b,p,T} \leq \mu_b + \sigma_{b,p,T}$, which implies

$$s_{b,p,T} \hat{v}_{b,p,T} \leq s_{b,p,T}(\mu_b + \sigma_{b,p,T}) \qquad (11)$$

Both $\hat{r}_{b,p,t} \in I_{b,p,t}$ and $x_{b,p} \in I_{b,p,t}$, by definition. Recalling that $|I_{b,p,t}| = 2\sigma_{b,p,t}$, this implies $\hat{r}_{b,p,t} \geq x_{b,p} - 2\sigma_{b,p,t}$. Thus

$$s_{b,p,T} \hat{r}_{b,p,T} \geq s_{b,p,T}(x_{b,p} - 2\sigma_{b,p,T}) \geq s_{b,p,T}(\tilde{\rho}_2 - 2\delta - 2\sigma_{b,p,T}) \qquad (12)$$

where the last inequality follows if $\textsc{Near}(b,p)$ occurs.

Combining Eq. (10), (11) and (12), and recalling that $\sigma_{b,p,T} = \sqrt{\frac{a \log T}{s_{b,p,T}}}$, we have

$$\begin{aligned}
&E\left[U_{b,p} \mid \textsc{Nice}, \textsc{Near}(b,p)\right] \\
&\leq E\left[s_{b,p,T}(\mu_b - \tilde{\rho}_2) + 3\sqrt{a s_{b,p,T} \log T} + 2 s_{b,p,T} \delta \mid \textsc{Nice}, \textsc{Near}(b,p)\right] \qquad (13)
\end{aligned}$$

Now to upper bound $\mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Far}(b,p)]$. We know that if $\textsc{Nice}$ occurs then $\bar{r}_{b,p} \in I_{b,p,t}$, which implies by the choice of $x_{b,p}$ that $\bar{r}_{b,p} \leq x_{b,p}$. Moreover, if $\textsc{Far}(b,p)$ occurs then $x_{b,p} + 2\delta < \tilde{\rho}_2$, which implies that $\Delta_{b,p} > 2\delta$. By Eq. (7) of Theorem 3 we have

$$\mathrm{E}[U_{b,p} \mid \textsc{Nice}, \textsc{Far}(b,p)] \leq \mathrm{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{b_t = b, p_t = p\right\} \;\middle|\; \textsc{Nice}, \textsc{Far}(b,p)\right] \leq \frac{4a \log T}{\delta^2} \qquad (14)$$

Combining Eq. (9), (13) and (14), and setting $\delta = \sqrt{\frac{a \log T}{T^{4/9}}}$, we have

$$E[U_{b,p}] \leq \mathrm{E}\left[\max\left\{\widetilde{\mathcal{O}}(T^{4/9}),\ \bar{s}_{b,p}(\mu_b - \tilde{\rho}_2 + \widetilde{\mathcal{O}}(T^{-2/9}) + \widetilde{\mathcal{O}}(\sqrt{\bar{s}_{b,p}})\right\}\right]$$

Summing this bound over all $p \in P$, and using the fact that $|P| = k$ and $\sum_p \bar{s}_{b,p} \leq T$, we have that $\mathrm{E}[u_b] \triangleq \frac{1}{T}\sum_{p\in P}\mathrm{E}[U_{b,p}]$ satisfies

$$\mathrm{E}[u_b] \leq (\mu_b - \tilde{\rho}_2) + \widetilde{\mathcal{O}}(T^{-1/6}) \tag{15}$$

From Eq.(15), no buyer $b > 1$ have a deviation that improves his average utility by more than $\widetilde{\mathcal{O}}(T^{-1/6})$. We are left to prove that the first buyer (assuming he is strategic) has no deviation improving his utility by more than $\widetilde{\mathcal{O}}(T^{-1/6})$.

If we show that the average utility of buyer 1, assuming he is strategic, under the equilibrium strategy is at least $\mu_1 - \tilde{\rho}_2 - \widetilde{\mathcal{O}}(T^{-1/6})$ then we are done. Consider two cases: in the first case $\mu_1 - \tilde{\rho}_2 \leq \widetilde{\mathcal{O}}(T^{-1/6})$. In such case, the utility of buyer 1 for any strategy must be at most $\widetilde{\mathcal{O}}(T^{-1/6})$ by Eq. (15), so in particular, it must be also so for the aggressive strategy. In the second case, $\mu_1 - \tilde{\rho}_2 \geq \widetilde{\mathcal{O}}(T^{-1/6})$. So conditioned on NICE all arms have confidence intervals of length $\widetilde{\mathcal{O}}(T^{-1/6})$, so an arm of buyer 1 will be always picked. Moreover, buyer 1 has a price p with $\tilde{\rho}_2 + \widetilde{\mathcal{O}}(T^{-1/6}) \leq p \leq \tilde{\rho}_2 + \widetilde{\mathcal{O}}(T^{-1/6}) + 1/k \leq \mu_1$. By the definition of the Second UCB Auction, either this arm or an arm with lower price will be chosen, generating average utility at least $\mu_1 - \tilde{\rho}_2 - \widetilde{\mathcal{O}}(T^{-1/6})$. □

C Proofs Omitted From Section 5

The following lemma, which will be used in the trade-off analysis, establishes a lower bound needed on the number of samples needed to accurately estimate the mean of certain binomial random variables. The proof, which borrows heavily from Proposition 7.3.2 of [18], can be found in the appendix.

Lemma 4. *Let $\epsilon \in [0, \frac{1}{8}]$ and $0 \leq \nu \leq \epsilon$. Then for a binomial variable $X \sim B(n, \frac{1}{2} - \nu)$ the following inequality holds:*

$$\Pr\left(X \geq (1/2 - \nu)n + \epsilon n\right) \geq \frac{1}{20}e^{-36n\epsilon^2}\ .$$

Now we ready for the main result of the section. The implication of this theorem are discussed further after the proof.

Proof. We first let $t = \epsilon n$ (for simplicity of presentation assume t and $n/2$ are integers) and expand the expression

$$\begin{aligned}
\Pr\Big(X \geq (1/2-\nu)n+\epsilon n\Big) &\geq \Pr\Big(X \geq \frac{n}{2}+t\Big) \\
&= \sum_{i=\frac{n}{2}+t}^{n} \binom{n}{i} (\frac{1}{2}-\nu)^i (\frac{1}{2}+\nu)^{n-i} \\
&= \frac{1}{2^n} \sum_{i=\frac{n}{2}+t}^{n} \binom{n}{i} (1-2\nu)^i (1+2\nu)^{n-i} \\
&\geq \frac{1}{2^n} \sum_{i=\frac{n}{2}+t}^{\frac{n}{2}+2t} \binom{n}{i} (1-2\nu)^i (1+2\nu)^{n-i} \\
&\geq \min_{j\in[t,2t]} \Big\{(1-2\nu)^{n/2+j}(1+2\nu)^{n/2-j}\Big\} \frac{1}{2^n} \sum_{i=\frac{n}{2}+t}^{\frac{n}{2}+2t} \binom{n}{i}.
\end{aligned}$$

The min term can be further bound

$$\begin{aligned}
\min_{j\in[t,2t]} \Big\{(1-2\nu)^{n/2+j}(1+2\nu)^{n/2-j}\Big\} &= (1-2\nu)^{n/2+2t}(1+2\nu)^{n/2-2t} \\
&= (1-4\nu^2)^{n/2-2t}(1-2\nu)^{4t} \\
&\geq e^{-8\nu^2(n/2-2t)} e^{-16\nu t} \geq e^{-4\nu^2 n - 16\nu t},
\end{aligned}$$

where the penultimate inequality follows from $1-x \geq e^{-2x}$ for $0 \leq x \leq 1/2$.

The sum term can be bound as follows

$$\begin{aligned}
\sum_{i=\frac{n}{2}+t}^{\frac{n}{2}+2t} \binom{n}{i} &\geq t\binom{n}{\frac{n}{2}+2t} \\
&= t\binom{n}{\frac{n}{2}} \frac{n/2-2t+1}{n/2+1} \cdot \frac{n/2-2t+2}{n/2+2} \cdots \frac{n/2}{n/2+2t} \\
&\geq \frac{2^n t}{\sqrt{2n}} \prod_{i=1}^{2t} \Big(1 - \frac{2t}{n/2+i}\Big) \\
&\geq \frac{2^n t}{\sqrt{2n}} \Big(1-\frac{2t}{n/2}\Big)^{2t} \geq \frac{2^n t}{\sqrt{2n}} e^{-16t^2/n},
\end{aligned}$$

where the inequality $\binom{n}{\frac{n}{2}} \geq \frac{2^n}{\sqrt{2n}}$ follows from Stirling's approximation and the final inequality again follows from $1-x \geq e^{-2x}$ for $0 \leq x \leq 1/2$.

Combining these intermediate results we have

$$\begin{aligned}
\Pr\Big(X \geq (1/2-\nu)n+\epsilon n\Big) &\geq \frac{t}{\sqrt{2n}} e^{-4\nu^2 n - 16(t^2/n+\nu t)} \\
&= \epsilon\sqrt{\frac{n}{2}} e^{-4\nu^2 n - 16n(\epsilon^2+\nu\epsilon)} \\
&\geq \epsilon\sqrt{\frac{n}{2}} e^{-36\epsilon^2 n} \geq \begin{cases} \frac{1}{12}e^{-36\epsilon^2 n}, & \text{if } \epsilon > \frac{1}{12}\sqrt{\frac{2}{n}} \\ \frac{1}{12}e^{-1/2}, & \text{if } 0 \leq \epsilon \leq \frac{1}{12}\sqrt{\frac{2}{n}} \end{cases}.
\end{aligned}$$

Note that $\frac{1}{12}e^{-1/2} \geq \frac{1}{20}$ and so for all $\epsilon \in [0, \frac{1}{8}]$ we have $\Pr\Big(X \geq (1/2-\nu)n + \epsilon n\Big) \geq \frac{1}{20}e^{-36n\epsilon^2}$, which completes the lemma. □

Proof of Theorem 5. Let $r_{b,p}$ denote the expected revenue generated by buyer b at price p. Now, consider a set of two buyers $B = \{b, b'\}$. Let b be a strategic buyer that statically follows the *aggressive policy* and has $\mu_b = 1$. Note, if the buyer deviates from the aggressive policy, then outcome (3) holds and we are done. Let b' be a myopic buyer with a Bernoulli value distribution with parameter $\beta_{b'}$.

Let us consider two possible scenarios differentiated by the setting of the myopic buyer's parameter $\beta_{b'}$:

Scenario A $\beta_{b'} = \frac{1}{2}$, which implies

$$\max_p r_{b',p} = \max_p p \Pr_{v\sim \mathbf{D}_{b'}}(p \leq v) = \max_p p\beta_{b'} = \frac{1}{2} < 1 = r_{b,1} = \max_p r_{b,p}\,,$$

as well as $r_{b,\frac{1}{2}} = \frac{1}{2} = r_{b',1} = \max_p r_{b',p}$.

Scenario B $\beta_{b'} = \frac{1}{2} - \frac{1}{T^\alpha}$, which implies

$$\max_p r_{b',p} = \max_p p \Pr_{v\sim \mathbf{D}_{b'}}(p \leq v) = \max_p p\beta_{b'} = \frac{1}{2} - \frac{1}{T^\alpha} < 1 = r_{b,1} = \max_p r_{b,p}\,,$$

as well as $r_{b,\frac{1}{2}-\frac{1}{T^\alpha}} = \frac{1}{2} - \frac{1}{T^\alpha} = r_{b',1} = \max_p r_{b',p}$.

Thus, in both scenarios buyer, b is able to generate the highest revenue and buyer b' sets the second-highest revenue benchmark. Also, in both cases, there is a price ($p = \beta_{b'}$) at which the first buyer can generate exactly this benchmark revenue.

Let $T_{b,p}$ denote the set of iterations where buyer b is offered price p and let $\hat{r}_{b,p} = \frac{1}{|T_{b,p}|}\sum_{t\in T_{b,p}} p \cdot \mathbf{1}\{\mathbf{A}_t\}$ denote the empirical estimate of $r_{b,p}$. Then, note that Lemma 4 implies that in Scenario A

$$\begin{aligned}\Pr\Big(r_{b',p} - \hat{r}_{b',p} \geq \frac{1}{T^\alpha}\Big) &= \Pr\Big(\Pr(p \leq v) - \frac{1}{|T_{b',p}|}\sum_{t\in T_{b',p}} \mathbf{1}\{p \leq v_{b,t}\} \geq \frac{1}{pT^\alpha}\Big)\\ &= \Pr_{X\sim B(|T_{b',p}|,\frac{1}{2})}\Big(\frac{1}{2} - \frac{1}{|T_{b',p}|}X \geq \frac{1}{pT^\alpha}\Big)\\ &= \Pr_{X\sim B(|T_{b',p}|,\frac{1}{2})}\Big(X \leq (\frac{1}{2} - \frac{1}{pT^\alpha})|T_{b',p}|\Big)\\ &= \Pr_{X\sim B(|T_{b',p}|,\frac{1}{2})}\Big(X \geq (\frac{1}{2} + \frac{1}{pT^\alpha})|T_{b',p}|\Big)\\ &\geq \frac{1}{20}\exp\Big(-\frac{36|T_{b',p}|}{p^2T^{2\alpha}}\Big)\,,\end{aligned}$$

where we've used $\epsilon = \frac{1}{pT^\alpha}$, $\nu = 0$, and $n = |T_{b',p}|$. Similarly in Scenario B we have

$$\begin{aligned}\Pr\Big(\hat{r}_{b',p} - r_{b',p} \geq \frac{1}{T^\alpha}\Big) &= \Pr_{X\sim B(|T_{b',p}|,\frac{1}{2}-\frac{1}{T^\alpha})}\Big(X \geq (\frac{1}{2} - \frac{1}{T^\alpha} + \frac{1}{pT^\alpha})|T_{b',p}|\Big)\\ &\geq \frac{1}{20}\exp\Big(-\frac{36|T_{b',p}|}{p^2T^{2\alpha}}\Big)\,,\end{aligned}$$

where again we have applied Lemma 4, with $\epsilon = \frac{1}{pT^\alpha}$, $\nu = \frac{1}{T^\alpha}$, and $n = |T_{b',p}|$.

Note, in either scenario, if for all prices p we have $|T_{b',p}| < \Omega(\log(1/\delta)T^{2\alpha})$, then with probability at least δ we cannot correctly determine whether $\beta_{b'} = \frac{1}{2}$ or $\beta_{b'} = \frac{1}{2} + \frac{1}{T^\alpha}$. In other words, with probability at least δ such a seller mechanism cannot distinguish between Scenario A and Scenario B and will behave the same (in expectation) in both scenarios. Now, let $\hat{p}$ be the average price offered to buyer b on the more than $T - T^{2\alpha} = \Omega(T)$ rounds that the seller mechanism offers a price buyer b. If $\hat{p} \leq \frac{1}{2} - \frac{1}{2T^\alpha}$, then in Scenario A the buyer suffers regret more than $\Omega(T) \cdot \frac{1}{2T^\alpha} = \Omega(T^{1-\alpha})$ and outcome (1) is achieved. Similarly, if $\hat{p} > \frac{1}{2} - \frac{1}{2T^\alpha}$, then in Scenario B the top buyer suffers a penalty of $\Omega(T^{1-\alpha}) \geq \Omega(T^{2\alpha})$ (for $0 < \alpha \leq 1/3$) and outcome (2) is achieved.

Finally, consider a seller mechanism that selects a price p such that $|T_{b',p}| \geq \Omega(\log(1/\delta)T^{2\alpha})$. Then the first buyer b suffers a buyer penalty of at least $\Omega(\log(1/\delta)T^{2\alpha})$ since it makes no utility on these rounds and. Therefore, outcome (2) is achieved and we are done. □