Low-Regret Algorithms for Strategic Buyers with Unknown Valuations in Repeated Posted-Price Auctions

Jason Rhuggenaath(✉), Paulo Roberto de Oliveira da Costa, Yingqian Zhang, Alp Akcay, and Uzay Kaymak

Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands
{j.s.rhuggenaath,p.r.d.oliveira.da.costa,yqzhang,a.e.akcay, u.kaymak}@tue.nl

Abstract. We study repeated posted-price auctions where a single seller repeatedly interacts with a single buyer for a number of rounds. In previous works, it is common to consider that the buyer knows his own valuation with certainty. However, in many practical situations, the buyer may have a stochastic valuation. In this paper, we study repeated posted-price auctions from the perspective of a utility maximizing buyer who does not know the probability distribution of his valuation and only observes a sample from the valuation distribution after he purchases the item. We first consider non-strategic buyers and derive algorithms with sublinear regret bounds that hold irrespective of the observed prices offered by the seller. These algorithms are then adapted into algorithms with similar guarantees for strategic buyers. We provide a theoretical analysis of our proposed algorithms and support our findings with numerical experiments. Our experiments show that, if the seller uses a low-regret algorithm for selecting the price, then strategic buyers can obtain much higher utilities compared to non-strategic buyers. Only when the prices of the seller are not related to the choices of the buyer, it is not beneficial to be strategic, but strategic buyers can still attain utilities of about 75% of the utility of non-strategic buyers.

Keywords: Online learning · Posted-price auctions · No-regret learning

1 Introduction

A growing fraction of online advertisements are sold via ad exchanges. In an ad exchange, after a visitor arrives on a webpage, advertisers compete in an auction to win the impression (the right to deliver an ad to that visitor). Typically, these auctions are second-price auctions, where the winner pays the second highest bid or a reserve price (whichever is larger), and no sale occurs if all of the bids are lower than the reserve price. However, as indicated by e.g. [2,3,21], a non-trivial fraction of auctions only involve a single bidder and this reduces to a

F. Hutter et al. (Eds.): ECML PKDD 2020, LNAI 12458, pp. 416–436, 2021.
https://doi.org/10.1007/978-3-030-67661-2_25

posted-price auction [20] when reserve prices known: the seller sets a reserve price and the buyer decides whether to accept or reject it. A single publisher can track a large number of visitors with similar properties over time and sell the impressions generated by these visitors to buyers. As buyers typically are involved in a large number of auctions, there is an incentive for them to act strategically [2,3,16,21]. These observations have led to the study of repeated posted-price auctions between a single seller and strategic buyer.

In this paper we consider a repeated posted-price auction between a single seller and a single buyer, similar to that considered in [2,21]. In every round, the seller posts a price and the buyer decides to buy or not at that price. The buyer does not know the distribution of his valuation, the seller's pricing algorithm or the seller's price set. Furthermore, the seller does not know the valuation distribution and needs to learn how to set the price over time. There are a number of differences between this paper and previous work on repeated posted-price auction such as [2,3,21]. First, unlike in previous work, we study the problem from the perspective of a buyer that aims to maximize his expected utility or surplus, instead of the perspective of the seller that aims to maximize his revenue. Second, previous papers assume that the buyer knows his valuation in each round. In this paper, we relax this assumption and assume the buyer does not know the distribution of his valuation and the valuation is only revealed after he buys the item. This is motivated by applications in online advertising where the buyer (advertiser) does not know the exact value of showing the ads to a set of users: some users may click on the ad and in some cases the ad may lead to a sale, but the buyer only observes a response after he displays the advertisement to the user.

As the valuation distribution is unknown, buyers face an exploration and exploitation trade-off and their decisions lead to regret: (i) accepting a price that is at most the mean valuation leads to positive expected utility and accepting a price above it leads to negative utility; (ii) buying the item leads to additional information about the mean valuation (at the risk of negative utility), but by not buying there is a risk of missing out on positive utility. We study two types of buyers: strategic buyers and non-strategic buyers. Non-strategic buyers are only interested achieving sub-linear regret given the prices that are observed and do not attempt to manipulate or influence the observed prices. Strategic buyers are also interested in sub-linear regret given the observed prices, but they also actively attempt to influence future prices that will be offered. If non-strategic buyers knew the mean valuation they would use the following rule: always accept a price that is at most the mean valuation and always reject a price above it. Strategic buyers on the other hand, would sometimes deviate from this rule in an attempt to influence future prices that will be offered. If non-strategic buyers knew the mean valuation, then their decisions would have low regret but the seller could learn to ask a price very close to the mean valuation, resulting in low utility for the buyer [2,21]. Strategic buyers attempt to influence the learning process of the seller in order to lower the price and to increase the utility. However, as these attempts are not guaranteed to succeed (as buyers

don't know the seller's pricing algorithm or price set), strategic buyers still want to ensure sub-linear regret for all possible prices sequences.

In our setting, the seller needs to learn to set his prices because he does not know the valuation distribution. To the best of our knowledge, there are no existing 'optimal' algorithms with performance guarantees (specifically) for repeated posted-price auctions with a single seller and a single strategic buyer that doesn't know his valuation: existing algorithms (e.g., [2,3,14,15,21,25]) assume that buyers know their valuation and thus lose their performance guarantees. In our experiments (see Sect. 5) we therefore assume that the seller uses an off-the-shelf low-regret learning algorithm for adaptive adversarial bandit feedback as these have known performance guarantees [10,20,22].

Our main contributions are as follows. First, to the best of our knowledge, we are the first to study repeated posted-price auctions in strategic settings from the perspective of the buyer. We do not assume that the buyer knows his valuation distribution. Second, we construct algorithms with sub-linear (in the problem horizon) regret for both non-strategic and strategic buyers by using ideas from popular multi-armed bandit algorithms UCB1 [5] and Thompson Sampling [1]. Our algorithms do not require knowledge about the seller's pricing algorithm or price set. Third, we use experiments to support our theoretical findings. Using experiments we show that, if the seller is using a low-regret learning algorithm based on weights updating (such as EXP3.P [4,10]), then strategic buyers can obtain much higher utilities compared to non-strategic buyers.

The remainder of this paper is organized as follows. In Sect. 2 we discuss the related literature. Section 3 provides a formal description of the problem. In Sect. 4 we present the our proposed algorithms and provide a theoretical analysis. In Sect. 5 we perform experiments in order to assess the quality of our proposed algorithms. Section 6 concludes our work and provides some directions for further research.

2 Related Literature

The work in this paper is mainly related to the following areas of the literature: posted-price auctions, low-regret learning by sellers and buyers, and decision making for buyers in auctions. We discuss these areas in more detail below.

Repeated posted-price auctions with the goal of maximizing revenue for the seller and assuming that the feedback from buyers is i.i.d. distributed was studied in [20]. Other works [2,3,14,15,19,21,25] instead study repeated posted-price auctions with strategic buyers. However, these papers all study the seller side of the problem and assume that buyers know their valuations in each round.

On a high level this paper is related to works that study repeated auctions where either the seller and/or the buyer is running a low-regret learning algorithm [8,9,11] and the interaction between bandit algorithms and incentives of buyers [6,7,12,18]. The goal in such studies is to design (truthful) mechanisms that either maximize revenue of the seller or welfare, when decision are made based on low-regret algorithms. This is not the focus of our paper.

The aforementioned works focus on either the seller side or on mechanism design, but there is also work that considers the perspective of buyers or bidders. In [13,23,24] the focus is on maximizing clicks when click-through-rates are unknown and typically with budget constraints. In this paper, rewards for buyers are not determined by the number of clicks, instead the buyer aims to maximize cumulative utilities or his net surplus as in e.g., [2,3,21]. In [17] the focus is on designing bidding strategies for buyers that compete against each other and where the buyer valuation is unknown. However, these studies do not focus on repeated posted-price auctions and strategic behaviour of buyers is not considered.

3 Problem Formulation

We consider a single buyer and a single seller that interact for T rounds. An item, such as an advertisement space, is repeatedly offered for sale by the seller to the buyer over these T rounds. In each round $t \in \mathcal{T} = \{1, \ldots, T\}$, a price $p_t \in \mathcal{P}$ is offered by the seller and a decision $a_t \in \{0, 1\}$ is made by the buyer: $a_t = 1$ when the buyer accepts to buy at that price, $a_t = 0$ otherwise. The buyer holds a private valuation $v_t \in [0, 1]$ for the item in round t. The value of v_t is an i.i.d. draw from a distribution $\mathcal{D}$ and has expectation $\nu = \mathbb{E}\{v_t\}$. The buyer does not know $\mathcal{D}$ and ν. Also, the buyer does not know $\mathcal{P}$ or the seller's pricing algorithm. The value v_t is only revealed to the buyer if he buys the item in round t, i.e., the buyer only observes the value after he buys the item. The seller also does not know $\mathcal{D}$ or ν and does not observe v_t.

The utility of the buyer in round t is given by $u_t = a_t \cdot (v_t - p_t)$. In other words, if the buyer purchases the item the utility is the difference between the valuation and the price. Otherwise, the utility is zero. For a fixed sequence $\overrightarrow{p} = p_1, \ldots, p_T$ of observed prices and a fixed sequence of decisions $a_1, \ldots, a_T$ by the buyer, the pseudo-regret of the buyer over T rounds is defined as $R_T(\overrightarrow{p}) = \sum_{t=1}^{T} \max\{\nu - p_t, 0\} - \sum_{t=1}^{T} a_t \cdot (\nu - p_t)$. The term $\max\{\nu - p_t, 0\}$ represents the expected utility of the optimal decision in round t and the term $a_t \cdot (\nu - p_t)$ represents the expected utility of the actual decision that is made by the buyer in round t. The expected pseudo-regret over T rounds is defined as $\mathcal{R}_T(\overrightarrow{p}) = \mathbb{E}\{R_T(\overrightarrow{p})\}$, where the expectation is taken with respect to possible randomization in the selection of the actions $a_1, \ldots, a_T$. In the remainder, the expected pseudo-regret will simply be referred to as the regret. The notation using $\overrightarrow{p}$ makes it clear that the regret depends on the sequence of observed prices. We will omit this dependence when the meaning is clear from the context or when a relation is understood to hold for all possible price sequences. For example, we write $\mathcal{R}_T \leq O(\sqrt{T \log T})$ when $\mathcal{R}_T(\overrightarrow{p}) \leq O(\sqrt{T \log T})$ for all choices of $\overrightarrow{p}$.

We consider two types of buyers: non-strategic buyers and strategic buyers. Non-strategic buyers are interested in achieving sub-linear regret for all possible price sequences, but they treat the price sequence as exogenous. That is, if non-strategic buyers knew ν, then they would follow this rule: buy if and only if $p_t \leq \nu$. Strategic buyers also want sub-linear regret for all possible prices

sequences, but they would sometimes deviate from this rule in an attempt to influence (i.e., lower) future prices that will be offered. If non-strategic buyers knew ν, then their decisions would have low regret but the seller could learn to ask a price just below ν, resulting in low utility for the buyer [2,21]. Strategic buyers actively attempt to influence the learning process of the seller in order to lower the price and to increase the utility. However, as these attempts are not guaranteed to succeed (recall that buyers do not know the seller's pricing algorithm or $\mathcal{P}$), strategic buyers still want to ensure sub-linear regret for all possible prices sequences. The seller does not know $\mathcal{D}$ or ν and does not observe v_t, and so he has to *learn* how to set his price over time under bandit feedback. This paper focuses on the buyer side and the regret bounds that we derive do not depend on the seller's pricing algorithm. However, in order to test our algorithms, some assumption about the seller's algorithm is required. To the best of our knowledge, there are no existing 'optimal' algorithms for sellers with performance guarantees (specifically) for repeated posted-price auctions with a single seller and a single strategic buyer that doesn't know his valuation: existing algorithms (e.g., [2,3,14,15,19,21,25]) assume that buyers know v_t and thus lose their performance guarantees. In our experiments (see Sect. 5) we therefore assume that the seller uses an off-the-shelf low-regret learning algorithm for adaptive adversarial bandit feedback as these have known performance guarantees [10,20,22].

Algorithm 1: UCB-NS

Input: $N \in \mathbb{N}$, T.
Set $\mathcal{V} = \varnothing$. Set $t = 1$. ;
Set $n = 1$. ;
Buy item at price p_t. ;
Observe v_t. ;
Set $\mathcal{V} = \mathcal{V} \cup \{v_t\}$. ;
for $t \in \{2, \ldots, T\}$ **do**
    Set $n_t = n$. ;
    Set $\bar{v}_t = \frac{1}{n_t} \sum_{v \in \mathcal{V}} v$. ;
    Set $r_t = \sqrt{(2 \log t)/n_t}$. ;
    Set $I_t = \bar{v}_t + r_t$. ;
    if $I_t \geq p_t$ **then**
        Buy item at price p_t. ;
        Observe v_t. ;
        Set $\mathcal{V} = \mathcal{V} \cup \{v_t\}$. ;
        Set $n = n + 1$. ;
    end
end

Algorithm 2: TS-NS

Input: $N \in \mathbb{N}$, T.
Set $\mathcal{V} = \varnothing$. Set $t = N$. ;
Set $n = N$. ;
Buy item in first N rounds. ;
Observe $\mathcal{V}^N = \cup_{k=1}^{N} \{v_k\}$. ;
Set $\mathcal{V} = \mathcal{V} \cup \mathcal{V}^N$. ;
for $t \in \{N + 1, \ldots, T\}$ **do**
    Set $n_t = n$. ;
    Set $\bar{v}_t = \frac{1}{n_t} \sum_{v \in \mathcal{V}} v$. ;
    Sample $I_t \sim \mathcal{N}(\bar{v}_t, \frac{1}{n_t})$. ;
    if $I_t \geq p_t$ **then**
        Buy item at price p_t. ;
        Observe v_t. ;
        Set $\mathcal{V} = \mathcal{V} \cup \{v_t\}$. ;
        Set $n = n + 1$. ;
    end
end

4 Algorithms and Analysis

In this section we present our proposed algorithms for strategic and non-strategic buyers and we provide a theoretical analysis of these algorithms.

4.1 Non-strategic Buyers

We provide two algorithms for non-strategic buyers that have sub-linear regret. The first algorithm, UCB-NS, is based on UCB (upper confidence bound) style bandit algorithms [5] and the second algorithm, TS-NS, is based on the Thompson Sampling principle [1]. In every round, UCB-NS maintains an optimistic estimate of the unknown mean ν and decides to buy the item if the estimate is at least as large as the offered price p_t. TS-NS samples from a posterior distribution and decides to buy the item if the sampled value is at least as large as the offered price p_t. Proposition 1 and 2 bound the regret of UCB-NS and TS-NS, respectively.

Proposition 1. *If Algorithm 1 is run with inputs: T, then $\mathcal{R}_T \leq O(\sqrt{T \log T})$.*

Proof. If $\mathbb{I}\{\nu > p_t > I_t\} = 1$ then the buyer did not buy the item when instead he should have bought it. Similarly, if $\mathbb{I}\{\nu < p_t \leq I_t\} = 1$, then the buyer did buy the item when instead he should not have bought it.

Note that we can bound the regret as follows

$$\begin{aligned}\mathcal{R}_T \leq 1 &+ \sum_{t=1}^{T} \mathbb{E}\left\{(\nu - p_t) \cdot \mathbb{I}\{\nu > p_t > I_t\}\right\} \\ &+ \sum_{t=1}^{T} \mathbb{E}\left\{(p_t - \nu) \cdot \mathbb{I}\{\nu < p_t \leq I_t\}\right\}.\end{aligned}$$

Define $A = \sum_{t=1}^{T} \mathbb{E}\left\{(\nu - p_t) \cdot \mathbb{I}\{\nu > p_t > I_t\}\right\}$ and
$B = \sum_{t=1}^{T} \mathbb{E}\left\{(p_t - \nu) \cdot \mathbb{I}\{\nu < p_t \leq I_t\}\right\}$. We will bound each term separately.

Define the following events $F_t = \{\nu > p_t > I_t\}$, $E_t = \{I_t > \nu\}$, $H_t = \{|\bar{v}_t - \nu| \leq \sqrt{\frac{2 \log T}{n_t}}\}$ and $H_t^C = \{|\bar{v}_t - \nu| > \sqrt{\frac{2 \log T}{n_t}}\}$.

For term A we have,

$$\begin{aligned}A &\leq \sum_{t=1}^{T} \mathbb{E}\left\{(\nu - p_t) \cdot \mathbb{I}\{F_t\}\right\} \leq \sum_{t=1}^{T} \mathbb{E}\left\{1 \cdot \mathbb{I}\{F_t\}\right\} \\ &\leq \sum_{t=1}^{T} \mathbb{P}\{F_t\} \leq \sum_{t=1}^{T} \mathbb{P}\{\nu > I_t\}\end{aligned}$$

Using Hoeffding's inequality (and a union bound) we obtain $\mathbb{P}\{\nu > I_t\} \leq \frac{1}{t^3} \leq \frac{1}{t^2}$. Therefore, we conclude that $\sum_{t=1}^{T} \mathbb{P}\{\nu > I_t\} \leq \frac{\pi^2}{6}$.

Define $\mathcal{B} = \{t \in \mathcal{T} \mid I_t \geq p_t\}$. For term B we have,

$$B \leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(p_t - \nu)\cdot \mathbb{I}\left\{\nu < p_t \leq I_t\right\}\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{I_t > \nu\right\}\right\}$$
$$\leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t\right\}\right\} + \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t^C\right\}\right\}.$$

Define $B_1 = \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t\right\}\right\}$. We bound B_1 as follows:

$$B_1 \leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{|I_t - \nu|\cdot \mathbb{I}\left\{H_t\right\}\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\nu - I_t| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\}$$
$$\leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\nu - I_t| \;\Big|\; H_t \Big\} \leq \sum_{t\in\mathcal{B}} 2\sqrt{\frac{2\log T}{n_t}}$$
$$\leq \sum_{t\in\mathcal{T}} 2\sqrt{\frac{2\log T}{t}} \leq 2\int_0^T \sqrt{\frac{2\log T}{t}}dt \leq 4\sqrt{2\log T}\sqrt{T}.$$

Define $B_2 = \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t^C\right\}\right\}$. We bound B_2 as follows:

$$B_2 \leq \sum_{t\in\mathcal{B}} \mathbb{P}\left\{H_t^C\right\}$$
$$\leq \sum_{t\in\mathcal{B}} \mathbb{P}\left\{|\bar{v}_t - \nu| > \sqrt{\frac{2\log T}{n_t}}\right\} \overset{(a)}{\leq} T\cdot\frac{2}{T^4}.$$

Inequality (a) follows from applying Hoeffding's inequality and from the fact that $|\mathcal{B}| \leq T$.

Putting everything together we obtain $\mathcal{R}_T \leq 1 + \frac{\pi^2}{6} + 4\sqrt{2\log T}\sqrt{T} + T\cdot\frac{2}{T^4}$. Therefore, we conclude that $\mathcal{R}_T \leq O(\sqrt{T\log T})$. □

Proposition 2. *If Algorithm 2 is run with inputs: T and $N = \lceil c_N \cdot T^{\frac{2}{3}}\rceil$, then $\mathcal{R}_T \leq O(T^{\frac{2}{3}}\sqrt{\log T})$.*

Proof. The proof can be found in the Appendix. □

4.2 Strategic Buyers

In this section we show how the algorithms for non-strategic buyers can be converted into algorithms for strategic buyers with the same growth rate (up to constant factors) for the regret. Our proposed approach BUYER-STRAT is presented in Algorithm 3. The main idea behind BUYER-STRAT is to take a base algorithm $\mathcal{A}_{base}$ for non-strategic buyers (e.g. UCB-NS or TS-NS) and modify it using what we refer to as *strategic cycles*.

We now give a description of how Algorithm 3 works. In BUYER-STRAT the buyers make decisions according to $\mathcal{A}_{base}$ for the first N_1 rounds. Afterwards, in

the next N_2 rounds, we enter a so-called strategic cycle. In this strategic cycle, the buyer only buys the item if the price is below some threshold, that is, if $p_t \leq v^* - c_1$. Here v^* is an estimate of the unknown mean ν and $0 < c_1 < 1$ is a parameter chosen by the buyer (e.g. $c_1 = 0.1$). The purpose of this strategic cycle is to entice the seller into asking prices that are lower than ν. After this strategic cycle comes to an end, we start another strategic cycle of length L with some small probability p_{cycle}. If another strategic cycle has been triggered, we set a new parameter $0 < c_{target} < 1$ and only prices $p_t \leq v^* - c_{target}$ are accepted. If no strategic cycle is triggered, the buyer makes decisions according to $\mathcal{A}_{base}$. In the next round, we start a strategic cycle of length L with probability p_{cycle} and the aforementioned process is repeated.

Algorithm 3 makes use of the functions F_1, F_2, F_3, F_4, F_5, F_6. The intuition behind these functions is as follows. In every strategic cycle, only prices that satisfy $p_t \leq v^* - c$ are accepted, where $c \in \mathcal{C}$ for some set $\mathcal{C}$. The value of v^* is selected using the function $F_5(\cdot)$ which takes as input a base algorithm $\mathcal{A}_{base}$. The value $c \in \mathcal{C}$ is selected by using the function $F_1(\cdot)$ which depends on a counter of the number strategic cycles that have passed C_{phase}. Initially, the number of strategic cycles in which values $c \in \mathcal{C}$ are used, is equal to N_{phase}. When $F_2(x) = 1$, this indicates that the last strategic cycle in which a value $c \in \mathcal{C}$ is used has just been completed, and the function $F_3(\cdot)$ is used to collect information about the price trajectory. When $F_6(x) = 1$, a final value for p_{target} is chosen (using $F_4(\cdot)$) and only prices with $p_t \leq p_{target}$ are accepted in all subsequent strategic cycles. In Sect. 5 we discuss these functions in more detail and give specific examples that are used in our experiments.

The key parameters to control the regret of Algorithm 3 are the cycle probability p_{cycle} and the cycle length L. Proposition 3 shows that BUYER-STRAT with $\mathcal{A}_{base}$ chosen as UCB-NS has regret of order $O(\sqrt{T \log T})$ if the probability p_{cycle} and the cycle length L is carefully chosen. Proposition 4 shows an analogous result for BUYER-STRAT with TS-NS.

Proposition 3. *Let A_p, A_L and A_N be positive real constants. Assume that Algorithm 3 is run with $\mathcal{A}_{base}$ chosen as UCB-NS and with inputs: T, $N_1 = \lceil T^{\frac{2}{3}} (\log T)^{\frac{1}{2}} \rceil$, $N_2 = \lceil A_N \sqrt{T \log T} \rceil$, $p_{cycle} = A_p T^{-\frac{1}{2}}$ and $L = A_L \sqrt{\log T}$, then $\mathcal{R}_T \leq O(\sqrt{T \log T})$.*

Proof. We will decompose the regret in two parts: the regret incurred in rounds that are part of strategic cycles and rounds that are not. For an arbitrary subset $\mathcal{T}^* \subseteq \mathcal{T}$, let $\mathcal{R}_{T,\mathcal{T}^*} = \sum_{t \in \mathcal{T}^*} \mathbb{E}\{(\nu - p_t) \cdot \mathbb{I}\{\nu > p_t > I_t\}\} + \sum_{t \in \mathcal{T}^*} \mathbb{E}\{(p_t - \nu) \cdot \mathbb{I}\{\nu < p_t \leq I_t\}\}$. Let $\mathcal{T}_S \subseteq \mathcal{T}$ denote the indices of the rounds that are part of strategic cycles and let $\mathcal{T}_{NS} = \mathcal{T} \setminus \mathcal{T}_S$ denote the indices of the rounds that are not. Then we can write, $\mathcal{R}_T = \mathcal{R}_{T,\mathcal{T}_{NS}} + \mathcal{R}_{T,\mathcal{T}_S}$.

For $\mathcal{R}_{T,\mathcal{T}_S}$ we have that $\mathcal{R}_{T,\mathcal{T}_S} \leq N_2 + T \cdot p_{cycle} \cdot L$. This follows from the fact that the expected number of triggered strategic cycles (after round $N_1 + N_2$) is at most $T \cdot p_{cycle}$ and the regret in every such cycle is at most L. Furthermore, the first strategic cycle has length N_2. For $\mathcal{R}_{T,\mathcal{T}_{NS}}$ we have that $\mathcal{R}_{T,\mathcal{T}_{NS}} \leq 5 + 4\sqrt{2 \log T}\sqrt{T}$. This follows from the fact that $\mathcal{R}_{T,\mathcal{T}_{NS}}$ represents the regret after $|\mathcal{T}_{NS}| \leq T$ rounds in a problem with horizon T, and by Proposition 1,

this quantity is bounded by $5 + 4\sqrt{2\log T}\sqrt{T}$. By plugging in the values we get $\mathcal{R}_T = \mathcal{R}_{T,\mathcal{I}_{NS}} + \mathcal{R}_{T,\mathcal{I}_S} \leq O(\sqrt{T\log T})$. □

Proposition 4. *Let A_p, A_L and A_N be positive real constants. Assume that Algorithm 3 is run with $\mathcal{A}_{base}$ chosen as TS-NS and with inputs: T, $N_1 = \lceil T^{\frac{2}{3}}(\log T)^{\frac{1}{2}} \rceil$, $N_2 = \lceil A_N\sqrt{T\log T} \rceil$, $p_{cycle} = A_p T^{-\frac{1}{2}}$ and $L = A_L\sqrt{\log T}$. Assume that TS-NS is run with inputs: T and $N = \lceil c_N \cdot T^{\frac{2}{3}} \rceil$. Then $\mathcal{R}_T \leq O(T^{\frac{2}{3}}\sqrt{\log T})$.*

Proof. The proof uses similar arguments as the proof of Proposition 3 and is omitted. A complete proof can be found in the Appendix. □

Algorithm 3: BUYER-STRAT

Input: F_1, F_2, F_3, F_4, F_5, F_6, L, p_{cycle}, N_{phase}, N_1, N_2, c_1, T, $\mathcal{A}_{base}$.
Set $L_p = \varnothing$, $L_{target} = \varnothing$, $C_{phase} = 0$, $t = 1$.;
for $t = 1, \ldots, N_1$ **do**
  Observe price p_t. Choose to buy or not based on $\mathcal{A}_{base}$.;
end
$v^* = F_5(\mathcal{A}_{base})$.;
for $t = N_1 + 1, \ldots, N_1 + N_2$ **do**
  Observe price p_t. Buy if $p_t \leq v^* - c_1$.;
end
while $t \in \{N_1 + N_2 + 1, \ldots, T\}$ **do**
  Draw D from Bernoulli distribution with success parameter p_{cycle}.;
  if $D = 1$ **then**
    $v^* = F_5(\mathcal{A}_{base})$.;
    if $C_{phase} \leq N_{phase}$ **then**
      Set $c_{target} = F_1(C_{phase})$. Set $p_{target} = v^* - c_{target}$.;
    end
    for $l \in \{1, \ldots, L\}$ **do**
      Observe price p_t.;
      $L_p = L_p \cup \{p_t\}$.;
      Buy if $p_t \leq p_{target}$.;
      Set $t = t + 1$.;
    end
    if $F_2(C_{phase}) = 1$ **then**
      Set $c_e = F_3(L_p)$. Set $L_{target} = L_{target} \cup \{c_e\}$.;
      Set $C_{phase} = C_{phase} + 1$.;
      if $F_6(C_{phase}) = 1$ **then**
        $p_{target} = F_4(L_{target})$;
      end
    end
    if $D = 0$ **then**
      Observe price p_t.;
      Choose to buy or not based on $\mathcal{A}_{base}$.;
      Set $t = t + 1$.;
    end
  end
end

Remark 5. In order to derive the results of Proposition 3 and 4, we only used the fact that the regret for $\mathcal{A}_{base}$ is bounded by $O(\sqrt{T \log T})$ or $O(T^{\frac{2}{3}}\sqrt{\log T})$. The same proof is also valid for any other base algorithm that satisfies these bounds. Also, the exact choices for functions F_1, F_2, F_3, F_4, F_5, F_6 do not effect the regret guarantee (in Sect. 5 we discuss these functions in more detail).

In which setting is BUYER-STRAT useful? As the seller does not know $\mathcal{D}$, it is reasonable to assume (as argued in Sect. 3) that the seller uses a low-regret algorithm to *learn* how to set prices. Note that many online learning algorithms (e.g. EXP3 and its variants) are *weight-based* algorithms: at round t, there are weights $w_{k,t}, \ldots, w_{K,t}$ and an action $k \in \{1, \ldots, K\}$ is chosen with probability $w_{k,t} / \sum_{k=1}^{K} w_{k,t}$. We call an algorithm a *pure weight-based* algorithm if in round t, only the weight of the selected action gets updated and if weights can only increase due to positive rewards (note that EXP3 is an example, see the Appendix for a general definition). Proposition 6 shows that, if the seller uses a pure weight-based algorithm, then BUYER-STRAT tends to encourage lower prices by using strategic cycles.

Proposition 6. *Assume that the buyer uses Algorithm 3, that the seller is using a pure weight-based algorithm and that the price set $\mathcal{P}$ is finite. Suppose that a strategic cycle runs from round $t+1$ to round $t+L$ with p_{target}, then $\mathbb{P}\{p_{t+L+1} \leq p_{target}\} \geq \mathbb{P}\{p_{t+1} \leq p_{target}\}$.*

Proof. The proof can be found in the Appendix. □

5 Experiments

In this section we verify the theoretical results that were derived and investigate the effects of strategic behaviour on the regret in different scenarios.

5.1 Setup of Experiments

In the experiments v_t is drawn from an uniform distribution on $[a - 0.3, a + 0.3]$, where a is drawn from an uniform distribution on $[0.4, 0.7]$ independently for each run. We consider two settings for the set of prices used by the seller and these are given by $\mathcal{P}_1$ and $\mathcal{P}_2$: $\mathcal{P}_1 = \{a + x \mid x \in \{-0.35, -0.3, -0.25, -0.2, -0.1, -0.05, -0.02, 0.0, 0.1, 0.3\}\}$, $\mathcal{P}_2 = \{a + x \mid x \in \{-0.05, -0.02, 0.0, 0.1, 0.3\}\}$. We will use the following abbreviations: P1 and P2. The abbreviation P1 means that $\mathcal{P}_1$ is used. The other abbreviations have a similar interpretation.

We consider three options for the seller pricing algorithm: (i) the seller chooses a price at random from the price set (RAND seller); (ii) the seller uses the low-regret learning algorithm EXP3.P (EXP3.P seller); (iii) the seller uses the full-information algorithm HEDGE (HEDGE seller). RAND seller is included because it models a situation where the buyer has no influence over the prices. EXP3.P seller is included because it is a bandit algorithm designed

for adaptive adversaries and it enjoys high-probability regret bounds [4,10]. It models a seller that is learning which prices to use based on bandit feedback that is non-stochastic. HEDGE seller is included in order to investigate whether the restriction to bandit feedback has a major impact on the performance of BUYER-STRAT. HEDGE seller is tuned according to Remark 5.17 in [22] and EXP3.P according to Theorem 3.2 in [10].

In the experiments, BUYER-STRAT is tuned with $N_1 = \lceil T^{\frac{2}{3}} \log T \rceil$, $N_2 = \lceil 2\sqrt{T \log T} \rceil$, $L = \lfloor 25\sqrt{\log T} \rfloor$, $p_{cycle} = \frac{5}{\sqrt{T}}$, $c_1 = 0.1$. We set $N_{phase} = 4 \cdot N_3$, where $N_3 = \lceil 0.1 \cdot \sqrt{T} \rceil$. TS-NS is tuned with $N = \lceil 0.005 \cdot T^{\frac{2}{3}} \rceil$. We will refer to BUYER-STRAT with $\mathcal{A}_{base}$ chosen as UCB-NS, as UCB-S (Upper Confidence Bound Strategic). Similarly, We will refer to BUYER-STRAT with $\mathcal{A}_{base}$ chosen as TS-NS, as TS-S (Thompson Sampling Strategic). The functions F_1, F_2, F_3, F_4, F_5, F_6 are chosen as follows.

$$F_1(x) = \begin{cases} 0.2 & \text{if } x \leq N_3 \\ 0.3 & \text{if } 1 \cdot N_3 < x \leq 2 \cdot N_3 \\ 0.4 & \text{if } 2 \cdot N_3 < x \leq 3 \cdot N_3 \\ 0.5 & \text{if } 3 \cdot N_3 < x \leq 4 \cdot N_3 \end{cases} \tag{1}$$

For $F_2(x)$ we take $F_2(x) = \mathbb{I}\{x \in \{N_3, 2 \cdot N_3, 3 \cdot N_3, 4 \cdot N_3\}\}$. The function $F_3(L_p)$ takes the last 100 elements added to the input list L_p and then calculates the 25-th percentile of these 100 values. The function $F_4(\cdot)$ is defined as $F_4(L_{target}) = \min\{L_{target}\} + \varepsilon$. The function $F_4(L_{target})$ takes the smallest number in the set L_{target} and adds a small value to it. In our experiments we use $\varepsilon = 0.005$. The function $F_5(\cdot)$ takes as input a base algorithm and returns the value of $\bar{v}_t$ in the base algorithm. For $F_6(x)$ we take $F_6(x) = \mathbb{I}\{x = 4 \cdot N_3\}$.

The intuition behind these choices is as follows. In every strategic cycle, only prices that satisfy $p_t \leq v^* - c$ are accepted, where $c \in \mathcal{C} = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and where c is chosen in increasing order (to try to reduce the price in stages) as the number of strategic cycles increases (this is specified by the function $F_1(\cdot)$). Initially, the number of strategic cycles in which every $c \in \mathcal{C}$ is used, is proportional to N_3. When $F_2(x) = 1$, this indicates that the last strategic cycle in which $c = x$ has just been completed, and the function $F_3(\cdot)$ is used to collect information about the price trajectory. When $F_6(x) = 1$, a final value for p_{target} is chosen (using $F_4(\cdot)$) and this value is used in all subsequent strategic cycles.

We perform 100 independent simulation runs in order to calculate our performance metrics. We use three performance metrics in order to evaluate our algorithm. In each run, we calculate the cumulative regret $R_T = \sum_{t=1}^{T} \max\{\nu - p_t, 0\} - \sum_{t=1}^{T} a_t \cdot (\nu - p_t)$, the cumulative utility $U_T = \sum_{t=1}^{T} a_t \cdot (\nu - p_t)$ and the scaled cumulative regret $R_T^S = R_T / \sum_{t=1}^{T} \max\{\nu - p_t, 0\}$. In the experiments we set $T \in \{25000, 50000, 75000, 100000, 200000, \ldots, 1000000\}$.

5.2 Results: Non-strategic Buyers vs. Strategic Buyers

Non-strategic Buyers. In Figs. 1 and 4 the cumulative regret is shown for different experimental settings and different values for the problem horizon. Each point in the graph shows the cumulative regret over T rounds for a problem of horizon T averaged over 100 simulations. In all figures, the lines indicate the mean and the shaded region indicates a 95% confidence interval. The results indicate that the expected regret indeed grows as a sub-linear function of T and that this pattern holds for both RAND seller and EXP3.P seller. An interesting finding is that the regret for TS-NS is lower than UCB-NS: based on the theoretical analysis one would expect the opposite pattern. Figures 3 and 6 show the scaled cumulative regret and provides further evidence that the expected regret is a sub-linear function of the horizon T, as the curve shows a monotonically decreasing pattern. Figures 2 and 5 show the cumulative utility against different sellers. Here we observe that the utility tends to be higher if the seller uses $\mathcal{P}_1$, which makes intuitive sense as this price set contains lower prices.

Strategic Buyers. Figures 7, 8, 9, 10, 11 and 12 show the same performance metrics as for the non-strategic bidders. Figures 1 and 4 show that the level of the expected regret for strategic bidders is higher compared to the non-strategic bidders. Figures 9 and 12 again indicate that the expected regret is sub-linear in T, as the curves show a monotonically decreasing pattern (from Fig. 7 it is hard to tell). Thus, we observe sub-linear regret for both UCB-S and TS-S regardless of the seller algorithm and this is in line with the theoretical analysis. If we compare the cumulative utility in Figs. 8 and 11 with those in Figs. 2 and 5, then we observe some interesting results. First, when strategic buyers are facing RAND seller (Fig. 8), then we see that the cumulative utility is about 70%–80% of the cumulative utility if non-strategic buyers are facing RAND seller (Fig. 2). Second, we see that if the seller is using EXP3.P (i.e, a low-regret learning algorithm), then the cumulative utility for strategic buyers is much higher compared to the cumulative utility for non-strategic buyers. In scenario P1 utilities are about 2.5–3 times higher and in scenario P2 utilities are about 2 times higher. The results for scenario P2 imply that, even when the lowest price is very close to the unknown mean valuation (absolute distance at most 0.05), it is still beneficial to act strategically. Additional experimental results when the seller uses EXP3.S [4] can be found in the Appendix.

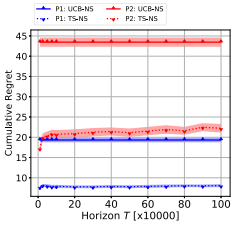

Fig. 1. R_T with RAND seller.

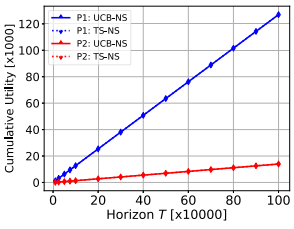

Fig. 2. U_T with RAND seller.

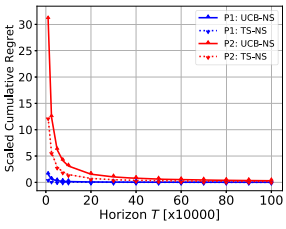

Fig. 3. R_T^S with RAND seller.

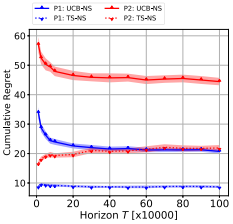

Fig. 4. R_T with EXP3.P seller.

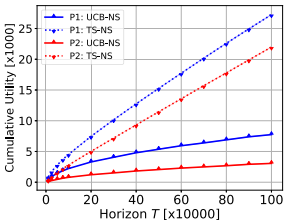

Fig. 5. U_T with EXP3.P seller.

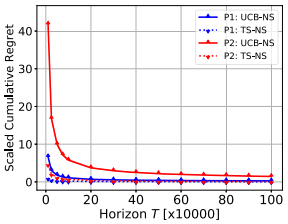

Fig. 6. R_T^S with EXP3.P seller.

5.3 Explanation of Differences

In order to study the impact of the quality of feedback that the seller observes, we give the seller full-information feedback instead of bandit feedback. More specifically, we assume the seller uses the algorithm HEDGE. Figures 13, 14 and 15 show results for TS-S and TS-NS against HEDGE seller. Even with full-information the results are qualitatively similar as before: the regret for the strategic buyers is sub-linear and cumulative utility is much higher for strategic buyers. Thus, the results indicate that the feedback type is not the main driver for the observed patterns.

Figures 16 and 17 display the gap $\nu - p_t$ for a problem with horizon $T = 200000$ averaged over the 100 simulation runs. If the seller is using a low-regret algorithm in order to set prices and buyers are non-strategic, then we observe that prices tend to increase towards the mean valuation ν. This effect is stronger for HEDGE seller compared to EXP3.P seller and this is in line with expectations as HEDGE uses full-information feedback. Furthermore, we see a qualitatively similar pattern for the price sets $\mathcal{P}_1$ and $\mathcal{P}_2$, although the increase in price with $\mathcal{P}_2$ is slightly larger. For HEDGE seller, we hardly see any difference for different price sets. If buyers are strategic then we see the opposite pattern. The algorithms for strategic buyers tend to lower the price over time and the magnitude of this reduction depends on the price set of the seller (reduction for $\mathcal{P}_1$ is larger than for $\mathcal{P}_2$).

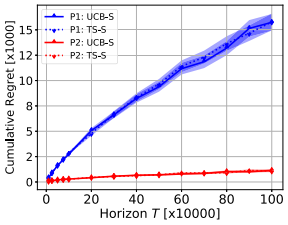

Fig. 7. R_T with RAND seller.

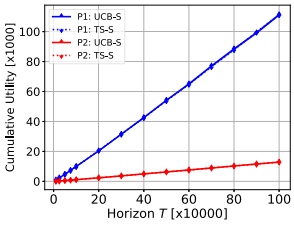

Fig. 8. U_T with RAND seller.

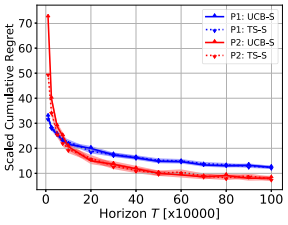

Fig. 9. R_T^S with RAND seller.

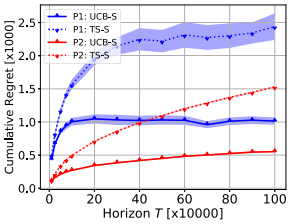

Fig. 10. R_T with EXP3.P seller.

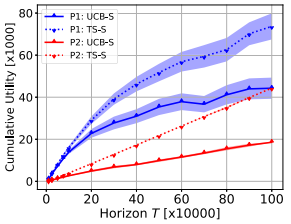

Fig. 11. U_T with EXP3.P seller.

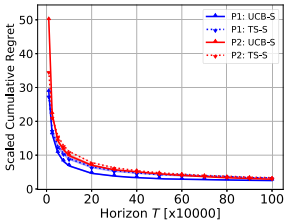

Fig. 12. R_T^S with EXP3.P seller.

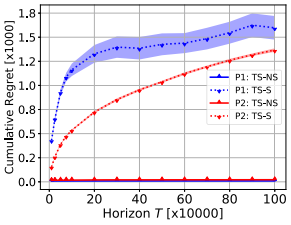

Fig. 13. R_T with HEDGE seller.

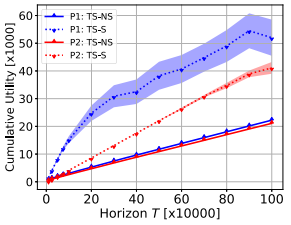

Fig. 14. U_T with HEDGE seller.

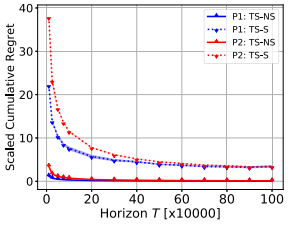

Fig. 15. R_T^S with HEDGE seller.

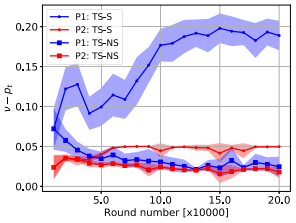

Fig. 16. $\nu - p_t$ with EXP3.P seller.

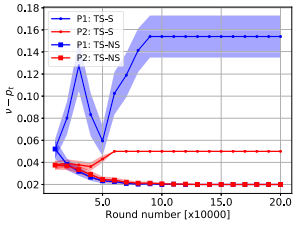

Fig. 17. $\nu - p_t$ with HEDGE seller.

However, even with price set $\mathcal{P}_2$ where the lowest prices are very close to ν, strategic behaviour is beneficial and strategic buyers can induce prices that are almost twice as far from ν.

6 Conclusion

This is paper we study repeated posted-price auctions with a single seller from the perspective of a utility maximizing buyer that does not know the distribution of his valuation. Previous work has only focused on the seller side and does not study how buyers should make decisions, hence in this paper, we address this gap in the literature. We study two types of buyers (strategic and non-strategic) and derive sub-linear regret bounds that hold for all possible sequences of observed prices. Our algorithms are based on ideas from UCB-type bandit algorithms and

Thompson Sampling. Our experiments we show that, if the seller is using a low-regret learning algorithm based on weights updating, then strategic buyers can obtain much higher utilities compared to non-strategic buyers.

In practice, buyers have limited budgets for purchasing items. One direction for future work is to investigate the impact of this on the problem. In particular, it would be interesting to analyze how a budget constraint would affect the regret guarantees derived in this paper and whether budget constraints make it easier or harder to engage in strategic behavior.

Appendix

Section A contains proofs that are omitted from the main text. Section B presents some additional experimental results.

A Proofs for Sect. 4

A.1 Proof of Proposition 2

Proof. We can bound the regret as follows $\mathcal{R}_T \leq N \cdot 1 + \sum_{t=N+1}^{T} \mathbb{E}\{(\nu - p_t) \cdot \mathbb{I}\{\nu > p_t > I_t\}\} + \sum_{t=N+1}^{T} \mathbb{E}\{(p_t - \nu) \cdot \mathbb{I}\{\nu < p_t \leq I_t\}\}$.

Define $A = \sum_{t=N+1}^{T} \mathbb{E}\{(\nu - p_t) \cdot \mathbb{I}\{\nu > p_t > I_t\}\}$ and

$B = \sum_{t=N+1}^{T} \mathbb{E}\{(p_t - \nu) \cdot \mathbb{I}\{\nu < p_t \leq I_t\}\}$. We will bound each term separately. Define the event $F_t = \{\nu > p_t > I_t\}$.

$$
\begin{aligned}
A &\leq \sum_{t=N+1}^{T} \mathbb{E}\{(\nu - p_t) \cdot \mathbb{I}\{F_t\}\} \leq \sum_{t=N+1}^{T} \mathbb{E}\{(\nu - I_t) \cdot \mathbb{I}\{F_t\}\} \\
&\leq \sum_{t=N+1}^{T} \mathbb{E}\{|\nu - I_t| \cdot \mathbb{I}\{F_t\}\} \leq \sum_{t=N+1}^{T} \mathbb{E}\Big\{ |(\nu - \bar{v}_t) - (I_t - \bar{v}_t)| \;\Big|\; F_t \Big\} \cdot \mathbb{P}\{F_t\} \\
&\leq \sum_{t=N+1}^{T} \mathbb{E}\Big\{ |\nu - \bar{v}_t| \;\Big|\; F_t \Big\} \cdot \mathbb{P}\{F_t\} + \sum_{t=N+1}^{T} \mathbb{E}\Big\{ |I_t - \bar{v}_t| \;\Big|\; F_t \Big\} \cdot \mathbb{P}\{F_t\} \\
&\leq \sum_{t=N+1}^{T} \mathbb{E}\{|\nu - \bar{v}_t|\} + \sum_{t=N+1}^{T} \mathbb{E}\{|I_t - \bar{v}_t|\}
\end{aligned}
$$

Using Hoeffding's inequality we obtain, for $t > N$, that $\mathbb{E}\{|\nu - \bar{v}_t|\} \leq \frac{2}{T^4} + 2\sqrt{\frac{2 \log T}{N}}$. Using the fact that $N = \lceil c_N \cdot T^{\frac{2}{3}} \rceil$ and that $T - (N+1) \leq T$, this yields $\sum_{t=N+1}^{T} \mathbb{E}\{|\nu - \bar{v}_t|\} \leq \frac{2}{T^3} + T^{\frac{2}{3}} 2\sqrt{\frac{2 \log T}{c_N}}$. Using the fact that, for $t > N$, $I_t - \bar{v}_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \frac{1}{n_t} \leq \frac{1}{N}$, we obtain that $\mathbb{E}\{|I_t - \bar{v}_t|\} \leq \sqrt{\frac{2}{\pi \cdot c_N}} T^{-\frac{1}{3}}$ and this yields $\sum_{t=N+1}^{T} \mathbb{E}\{|I_t - \bar{v}_t|\} \leq \sqrt{\frac{2}{\pi \cdot c_N}} T^{\frac{2}{3}}$.

Define $\mathcal{B} = \{t \in \mathcal{T} \mid t > N, I_t \geq p_t\}$. Let $E_t = \{I_t > \nu\}$, let $H_t = \{|\bar{v}_t - \nu| \leq \sqrt{\frac{2\log T}{n_t}}\}$ and let $H_t^C = \{|\bar{v}_t - \nu| > \sqrt{\frac{2\log T}{n_t}}\}$. Let $\hat{v}_s$ denote the sample mean of s i.i.d. draws from distribution $\mathcal{D}$ and let $\hat{I}_s \sim \mathcal{N}(\hat{v}_s, \frac{1}{s})$. For term B we have,

$$\begin{aligned} B &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(p_t - \nu)\cdot \mathbb{I}\left\{\nu < p_t \leq I_t\right\}\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{I_t > \nu\right\}\right\} \\ &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t\right\}\right\} + \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t^C\right\}\right\}. \end{aligned}$$

Define $B_1 = \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t\right\}\right\}$. We bound B_1 as follows:

$$\begin{aligned} B_1 &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\left\{|I_t - \nu|\cdot \mathbb{I}\left\{H_t\right\}\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\nu - I_t| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\} \\ &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |I_t - \bar{v}_t| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\} + \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\bar{v}_t - \nu| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\}. \end{aligned}$$

Define $B_{11} = \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |I_t - \bar{v}_t| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\}$ and
$B_{12} = \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\bar{v}_t - \nu| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\}$.

We bound B_{11} as follows:

$$\begin{aligned} B_{11} &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |I_t - \bar{v}_t| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\} + \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |I_t - \bar{v}_t| \;\Big|\; H_t^C \Big\}\cdot \mathbb{P}\left\{H_t^C\right\} \\ &= \sum_{t\in\mathcal{B}} \mathbb{E}\{|I_t - \bar{v}_t|\} = \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{|\hat{I}_{n_t} - \hat{v}_{n_t}|\Big\} \leq \sum_{t\in\mathcal{T}} \mathbb{E}\Big\{|\hat{I}_t - \hat{v}_t|\Big\} \\ &\leq \sum_{t\in\mathcal{T}} \sqrt{\frac{2}{\pi t}} \leq \int_0^T \sqrt{\frac{2}{\pi t}}dt = 2\sqrt{\frac{2}{\pi}T}. \end{aligned}$$

We bound B_{12} as follows:

$$\begin{aligned} B_{12} &\leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\bar{v}_t - \nu| \;\Big|\; H_t \Big\}\cdot \mathbb{P}\left\{H_t\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{E}\Big\{ |\bar{v}_t - \nu| \;\Big|\; H_t \Big\} \\ &\leq \sum_{t\in\mathcal{T}} \mathbb{E}\left\{ |\hat{v}_t - \nu| \;\middle|\; |\hat{v}_t - \nu| \leq \sqrt{\frac{2\log T}{t}} \right\} \\ &\leq \sum_{t\in\mathcal{T}} \sqrt{\frac{2\log T}{t}} \leq \int_0^T \sqrt{\frac{2\log T}{t}}dt \leq 2\sqrt{2\log T}\sqrt{T}. \end{aligned}$$

Define $B_2 = \sum_{t\in\mathcal{B}} \mathbb{E}\left\{(I_t - \nu)\cdot \mathbb{I}\left\{E_t \cap H_t^C\right\}\right\}$. We bound B_2 as follows:

$$\begin{aligned} B_2 &\leq \sum_{t\in\mathcal{B}} \mathbb{P}\left\{H_t^C\right\} \leq \sum_{t\in\mathcal{B}} \mathbb{P}\left\{|\hat{v}_{n_t} - \nu| > \sqrt{\frac{2\log T}{n_t}}\right\} \\ &\leq \sum_{t\in\mathcal{T}} \mathbb{P}\left\{|\hat{v}_t - \nu| > \sqrt{\frac{2\log T}{t}}\right\} \leq T\cdot\frac{2}{T^4}. \end{aligned}$$

Putting everything together we obtain $\mathcal{R}_T \leq N \cdot 1 + \frac{2}{T^3} + 2T^{\frac{2}{3}}\sqrt{\frac{2\log T}{c_N}} + \sqrt{\frac{2}{\pi \cdot c_N}}T^{\frac{2}{3}} + 2\sqrt{\frac{2T}{\pi}} + 2\sqrt{2T\log T} + T \cdot \frac{2}{T^4}$. So, we conclude that $\mathcal{R}_T \leq O(T^{\frac{2}{3}}\sqrt{\log T})$. □

A.2 Proof of Proposition 4

Proof. We will decompose the regret in two parts: the regret incurred in rounds that are part of strategic cycles and rounds that are not. For an arbitrary subset $\mathcal{T}^* \subseteq \mathcal{T}$, let $\mathcal{R}_{T,\mathcal{T}^*} = \sum_{t\in\mathcal{T}^*}\mathbb{E}\left\{(\nu - p_t)\cdot\mathbb{I}\left\{\nu > p_t > I_t\right\}\right\} + \sum_{t\in\mathcal{T}^*}\mathbb{E}\left\{(p_t - \nu)\cdot\mathbb{I}\left\{\nu < p_t \leq I_t\right\}\right\}$. Let $\mathcal{T}_S \subseteq \mathcal{T}$ denote the indices of the rounds that are part of strategic cycles and let $\mathcal{T}_{NS} = \mathcal{T} \setminus \mathcal{T}_S$ denote the indices of the rounds that are not. Then we can write, $\mathcal{R}_T = \mathcal{R}_{T,\mathcal{T}_{NS}} + \mathcal{R}_{T,\mathcal{T}_S}$.

For $\mathcal{R}_{T,\mathcal{T}_S}$ we have that $\mathcal{R}_{T,\mathcal{T}_S} \leq N_2 + T \cdot p_{cycle} \cdot L$. This follows from the fact that the expected number of triggered strategic cycles (after round $N_1 + N_2$) is $T \cdot p_{cycle}$ and the regret in every such cycle is at most L. Furthermore, the first strategic cycle has length N_2. For $\mathcal{R}_{T,\mathcal{T}_{NS}}$ we have that $\mathcal{R}_{T,\mathcal{T}_{NS}} \leq O(T^{\frac{2}{3}}\sqrt{\log T})$. This follows from the fact that $\mathcal{R}_{T,\mathcal{T}_{NS}}$ represents the regret after $|\mathcal{T}_{NS}| \leq T$ rounds in a problem with horizon T, and by Proposition 2, this quantity is bounded by $O(T^{\frac{2}{3}}\sqrt{\log T})$. By plugging in the values we get $\mathcal{R}_T = \mathcal{R}_{T,\mathcal{T}_{NS}} + \mathcal{R}_{T,\mathcal{T}_S} \leq O(T^{\frac{2}{3}}\sqrt{\log T})$. □

A.3 Proof of Proposition 6

In this section we give a proof of Proposition 6. We first give a definition of a pure weight-based algorithm.

Definition 7. *Let there be K actions in total and let $\mathcal{K} = \{1, \ldots, K\}$. Let $w_{k,t} \in \mathbb{R}$ denote the weight of action k at the beginning of round t. Suppose that action j is selected in round t and that the observed reward for action j in round t equals $r_{j,t}$. Let $\hat{p}_{k,t}$ denote the probability that action k is selected in round t. An algorithm $\mathcal{A}$ is called a pure weight-based algorithm if the following conditions are satisfied:*

1. *if $r_{j,t} > 0$, then $w_{j,t+1} > w_{j,t}$.*
2. *if $r_{j,t} = 0$, then $w_{j,t+1} = w_{j,t}$.*
3. *if $k \neq j$, then $w_{k,t+1} = w_{k,t}$.*
4. *$\sum_{k\in\mathcal{K}^*}\hat{p}_{k,t} = F(\sum_{k\in\mathcal{K}^*}w_{k,t}/\sum_{k=1}^{K}w_{k,t})$ for all subsets $\mathcal{K}^* \subseteq \mathcal{K}$, where $F(\cdot)$ is an increasing function. That is, for all subsets $\mathcal{K}^* \subseteq \mathcal{K}$, if $a = \sum_{k\in\mathcal{K}^*}w_{k,t}/\sum_{k=1}^{K}w_{k,t}$, $b = \sum_{k\in\mathcal{K}^*}w'_{k,t}/\sum_{k=1}^{K}w'_{k,t}$ and $a > b$, then $F(a) > F(b)$.*

Note that if $\hat{p}_{k,t} = w_{k,t}/\sum_{k=1}^{K}w_{k,t}$ then condition 4 in Definition 7 is satisfied. Also note that EXP3 of [4] uses $\hat{p}_{k,t} = (1-\gamma)w_{k,t}/\sum_{k=1}^{K}w_{k,t} + \gamma/K$ and this choice also satisfies condition 4 in Definition 7.

Proof (of of Proposition 6). Let $|\mathcal{P}| = K$, $\mathcal{K} = \{1, \ldots, K\}$, $p_{max} = \max\{\mathcal{P}\}$ and $p_{min} = \min\{\mathcal{P}\}$. Assume, without loss of generality, that $\mathcal{P} = \{p^1, \ldots, p^K\}$ and that $0 < p_{min} = p^1 \leq p^2 \leq \cdots \leq p^{K-1} \leq p^K = p_{max}$. Let $\hat{\mathcal{P}} = \{p \in \mathcal{P} |\ p \leq p_{target}\}$. Let $\bar{\mathcal{P}} = \{p \in \mathcal{P} |\ p > p_{target}\}$. Let $\hat{\mathcal{K}} = \{k \in \mathcal{K} |\ p^k \in \hat{\mathcal{P}}\}$. Let $\bar{\mathcal{K}} = \{k \in \mathcal{K} |\ p^k \in \bar{\mathcal{P}}\}$. Let $w_{k,t}$ denote the weight of action k at the beginning of round t.

We now proceed to prove the statement in the Proposition. We prove the Proposition for $L = 1$. The case for general L follows by repeatedly applying the result for $L = 1$.

We distinguish the following cases. Case 1: $p_{target} \geq p_{max}$. Case 2: $p_{target} < p_{min}$. Case 3: $p_{min} \leq p_{target} < p_{max}$.

- Case 1: $p_{target} \geq p_{max}$. In this case, $p \leq p_{target}$ for all $p \in \mathcal{P}$. Therefore, $\mathbb{P}\{p_{t+1} \leq p_{target}\} = 1$ and $\mathbb{P}\{p_{t+2} \leq p_{target}\} = 1$ and the statement in the Proposition holds.
- Case 2: $p_{target} < p_{min}$. In this case, $p > p_{target}$ for all $p \in \mathcal{P}$. Therefore, $\mathbb{P}\{p_{t+1} \leq p_{target}\} = 0$ and $\mathbb{P}\{p_{t+2} \leq p_{target}\} = 0$ and the statement in the Proposition holds.
- Case 3: $p_{min} \leq p_{target} < p_{max}$. There are 2 subcases to consider. Case A: $p_{t+1} > p_{target}$ and Case B: $p_{t+1} \leq p_{target}$.
  - In Case A, none of the weights get updated. This is true because none of the prices in $\hat{\mathcal{P}}$ are selected since $p_{t+1} > p_{target}$. By condition 3 in Definition 7, it follows that none of the weights corresponding to the prices in $\hat{\mathcal{P}}$ will get updated.
    Also, none of the prices in $\bar{\mathcal{P}}$ will get a positive reward because they will all be rejected by the buyer. By condition 2 in Definition 7, it follows that none of the weights corresponding to the prices in $\bar{\mathcal{P}}$ will get updated. As none of the weights will get updated after round $t+1$ is completed, we have that $\mathbb{P}\{p_{t+2} \leq p_{target}\} = \mathbb{P}\{p_{t+1} \leq p_{target}\}$. So we conclude that the statement in the Proposition holds.
  - In Case B, there exists a $j \in \{1, \ldots, K\}$ such that $p_{t+1} = p^j$ and the reward for action j satisfies $r_{j,t+1} > 0$. This is true because the price $p_{t+1} = p^j$ will be accepted by the buyer and the reward equals the price p^j which (by assumption) satisfies $p^j \geq p_{min} > 0$.
    By condition 1 in Definition 7, it follows that $w_{j,t+2} > w_{j,t+1}$. By condition 3 in Definition 7, it follows that $w_{k,t+2} = w_{k,t+1}$ for all $k \neq j$, since these prices/actions are not selected in round $t+1$.
    This yields the following:

$$\sum_{k \in \hat{\mathcal{K}}} w_{k,t+2} > \sum_{k \in \hat{\mathcal{K}}} w_{k,t+1} \tag{2}$$

$$\sum_{k \in \bar{\mathcal{K}}} w_{k,t+2} = \sum_{k \in \bar{\mathcal{K}}} w_{k,t+1} \tag{3}$$

$$\sum_{k=1}^{K} w_{k,t+2} > \sum_{k=1}^{K} w_{k,t+1} \tag{4}$$

Note that we also have:

$$\mathbb{P}\left\{p_{t+2} \leq p_{target}\right\} = 1 - \mathbb{P}\left\{p_{t+2} > p_{target}\right\}, \tag{5}$$

$$\mathbb{P}\left\{p_{t+1} \leq p_{target}\right\} = 1 - \mathbb{P}\left\{p_{t+1} > p_{target}\right\}. \tag{6}$$

By combining (3) and (4), and by condition 4 in Definition 7, we obtain that $\mathbb{P}\left\{p_{t+2} > p_{target}\right\} < \mathbb{P}\left\{p_{t+1} > p_{target}\right\}$. As a consequence, by using (5) and (6), it follows that $\mathbb{P}\left\{p_{t+2} \leq p_{target}\right\} > \mathbb{P}\left\{p_{t+1} \leq p_{target}\right\}$. So we conclude that the statement in the Proposition holds.

The case for general L follows from repeatedly applying the above argument. Note that the argument above works every initial weight vector. By repeatedly applying the above argument, one can show that $\mathbb{P}\left\{p_{t+1} \leq p_{target}\right\} \leq \mathbb{P}\left\{p_{t+2} \leq p_{target}\right\} \leq \cdots \leq \mathbb{P}\left\{p_{t+L} \leq p_{target}\right\} \leq \mathbb{P}\left\{p_{t+L+1} \leq p_{target}\right\}$. □

B Additional experiments

This part contains additional results related to the experiments in the main text. We show results for non-strategic and strategic buyers against another (more powerful) seller algorithm. We assume the seller uses the EXP3.S algorithm from [4]. We will refer to this as EXP3.S Seller. This algorithm has sub-linear regret with respect to action sequences with at most S switches. EXP3.S Seller is tuned according to Corollary 8.2 in [4].

Figures 18, 19 and 20 display the results for non-strategic buyers and Figs. 21, 22 and 23 display the results for strategic buyers. In all figures, the lines indicate the mean and the shaded region indicates a 95% confidence interval. The results are qualitatively similar to those reported in the main text. The results indicate that the proposed algorithms for strategic and non-strategic buyers have sub-linear regret in all cases considered.

In scenario P1 utilities are about 2.0–2.5 times higher. In scenario P2 the differences are smaller, which is in line with expectations since the lowest price of the seller is very close to the unknown mean valuation. In general, the strategic buyers tend have higher utilities.

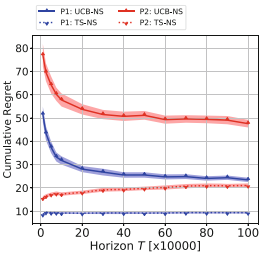

Fig. 18. R_T with EXP3.S seller.

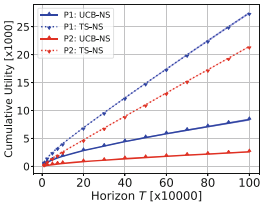

Fig. 19. U_T with EXP3.S seller.

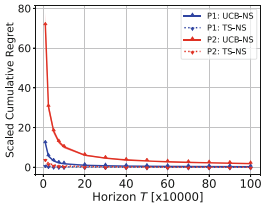

Fig. 20. R_T^S with EXP3.S seller.

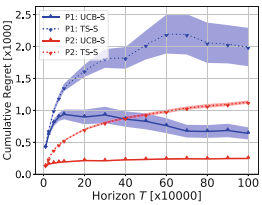

Fig. 21. R_T with EXP3.S seller.

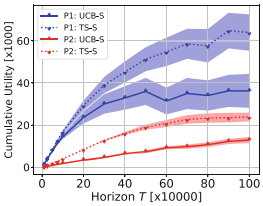

Fig. 22. U_T with EXP3.S seller.

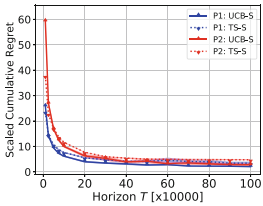

Fig. 23. R_T^S with EXP3.S seller.

References

1. Agrawal, S., Goyal, N.: Further optimal regret bounds for Thompson sampling. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, vol. 31, pp. 99–107. PMLR (2013)
2. Amin, K., Rostamizadeh, A., Syed, U.: Learning prices for repeated auctions with strategic buyers. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, pp. 1169–1177. Curran Associates Inc. (2013)
3. Amin, K., Rostamizadeh, A., Syed, U.: Repeated contextual auctions with strategic buyers. Adv. Neural Inf. Process. Syst. **27**, 622–630 (2014)
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.: The nonstochastic multiarmed bandit problem. SIAM J. Comput. **32**(1), 48–77 (2002)
5. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2), 235–256 (2002). https://doi.org/10.1023/A:1013689704352
6. Babaioff, M., Kleinberg, R.D., Slivkins, A.: Truthful mechanisms with implicit payment computation. In: Proceedings of the 11th ACM Conference on Electronic Commerce, pp. 43–52. Association for Computing Machinery (2010)
7. Babaioff, M., Sharma, Y., Slivkins, A.: Characterizing truthful multi-armed bandit mechanisms. SIAM J. Comput. **43**(1), 194–230 (2014)
8. Blum, A., Kumar, V., Rudra, A., Wu, F.: Online learning in online auctions. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 202–204. SIAM (2003)
9. Braverman, M., Mao, J., Schneider, J., Weinberg, M.: Selling to a no-regret buyer. In: Proceedings of the 2018 ACM Conference on Economics and Computation, pp. 523–538. ACM (2018)

10. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found. Trends® Mach. Learn. **5**(1), 1–122 (2012)
11. Deng, Y., Schneider, J., Sivan, B.: Prior-free dynamic auctions with low regret buyers. Adv. Neural Inf. Proces. Syst. **32**, 4803–4813 (2019)
12. Devanur, N.R., Kakade, S.M.: The price of truthfulness for pay-per-click auctions. In: Proceedings of the 10th ACM Conference on Electronic Commerce, pp. 99–106 (2009)
13. Ding, W., Qiny, T., Zhang, X.D., Liu, T.Y.: Multi-armed bandit with budget constraint and variable costs. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 232–238. AAAI Press (2013)
14. Drutsa, A.: Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In: Proceedings of the 26th International Conference on World Wide Web, pp. 33–42 (2017)
15. Drutsa, A.: Weakly consistent optimal pricing algorithms in repeated posted-price auctions with strategic buyer. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 1319–1328. PMLR, 10–15 July 2018
16. Edelman, B., Ostrovsky, M.: Strategic bidder behavior in sponsored search auctions. Decis. Support Syst. **43**(1), 192–198 (2007)
17. Feng, Z., Podimata, C., Syrgkanis, V.: Learning to bid without knowing your value. In: Proceedings of the 2018 ACM Conference on Economics and Computation, pp. 505–522 (2018)
18. Gatti, N., Lazaric, A., Trovò, F.: A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In: Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 605–622. ACM (2012)
19. Immorlica, N., Lucier, B., Pountourakis, E., Taggart, S.: Repeated sales with multiple strategic buyers. In: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 167–168 (2017)
20. Kleinberg, R., Leighton, T.: The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, pp. 594 (2003)
21. Mohri, M., Medina, A.M.N.: Optimal regret minimization in posted-price auctions with strategic buyers. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 1871–1879 (2014)
22. Slivkins, A.: Introduction to multi-armed bandits. Found. Trends® Mach. Learn. **12**(1–2), 1–286 (2019)
23. Tran-Thanh, L., Chapman, A., Rogers, A., Jennings, N.R.: Knapsack based optimal policies for budget-limited multi-armed bandits. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 1134–1140. AAAI Press (2012)
24. Trovò, F., Paladino, S., Restelli, M., Gatti, N.: Budgeted multi-armed bandit in continuous action space. In: Proceedings of the Twenty-Second European Conference on Artificial Intelligence, pp. 560–568. IOS Press (2016)
25. Vanunts, A., Drutsa, A.: Optimal pricing in repeated posted-price auctions with different patience of the seller and the buyer. In: Advances in Neural Information Processing Systems, vol. 32 (2019)