Repeated Contextual Auctions with Strategic Buyers

Kareem Amin
University of Pennsylvania
akareem@cis.upenn.edu

Afshin Rostamizadeh
Google Research
rostami@google.com

Umar Syed
Google Research
usyed@google.com

Abstract

Motivated by real-time advertising exchanges, we analyze the problem of pricing inventory in a repeated posted-price auction. We consider both the cases of a *truthful* and *surplus-maximizing* buyer, where the former makes decisions myopically on every round, and the latter may strategically react to our algorithm, forgoing short-term surplus in order to trick the algorithm into setting better prices in the future. We further assume a buyer's valuation of a good is a function of a context vector that describes the good being sold. We give the first algorithm attaining sublinear ($\widetilde{O}(T^{2/3})$) regret in the contextual setting against a surplus-maximizing buyer. We also extend this result to repeated second-price auctions with multiple buyers.

1 Introduction

A growing fraction of Internet advertising is sold through automated *real-time ad exchanges*. In a real-time ad exchange, after a visitor arrives on a webpage, information about that visitor and webpage, called the *context*, is sent to several advertisers. The advertisers then compete in an auction to win the *impression*, or the right to deliver an ad to that visitor. One of the great advantages of online advertising compared to advertising in traditional media is the presence of rich contextual information about the impression. Advertisers can be particular about whom they spend money on, and are willing to pay a premium when the right impression comes along, a process known as *targeting*. Specifically, advertisers can use context to specify which auctions they would like to participate in, as well as how much they would like to bid. These auctions are most often *second-price* auctions, wherein the winner is charged either the second highest bid or a prespecified *reserve price* (whichever is larger), and no sale occurs if the reserve price isn't cleared by one of the bids.

One side-effect of targeting, which has been studied only recently, is the tendency for such exchanges to generate many auctions that are rather uncompetitive or *thin*, in which few advertisers are willing to participate. Again, this stems from the ability of advertisers to examine information about the impression before deciding to participate. While this selectivity is clearly beneficial for advertisers, it comes at a cost to webpage publishers. Many auctions in real-time ad exchanges ultimately involve just a single bidder, in which case the publisher's revenue is entirely determined by the selection of reserve price. Although a lone advertiser may have a high valuation for the impression, a low reserve price will fail to extract this as revenue for the seller if the advertiser is the only participant in the auction.

As observed by [2], if a single buyer is repeatedly interacting with a seller, selecting revenue-maximizing reserve prices (for the seller) reduces to revenue-maximization in a repeated *posted-price* setting: On each round, the seller offers a good to the buyer at a price. The buyer observes her *value* for the good, and then either accepts or rejects the offer. The seller's price-setting algorithm is known to the buyer, and the buyer behaves to maximize her (time-discounted) cumulative *surplus*, i.e., the total difference between the buyer's value and the price on rounds where she accepts the offer. The goal of the seller is to extract nearly as much revenue from the buyer as would have been

possible if the process generating the buyer's valuations for the goods had been known to the seller before the start of the game. In [2] this goal is called minimizing *strategic regret*.

Online learning algorithms are typically designed to minimize regret in *hindsight*, which is defined as the difference between the loss of the best action and the loss of the algorithm given the observed sequence of events. Furthermore, it is assumed that the observed sequence of events are generated adversarially. However, in our setting, the buyer behaves self-interestedly, which is not necessarily the same as behaving adversarially, because the interaction between the buyer and seller is not zero-sum. A seller algorithm designed to minimize regret against an adversary can perform very suboptimally. Consider an example from [2]: a buyer who has a large valuation v for every good. If the seller announces an algorithm that minimizes (standard) regret, then the buyer should respond by only accepting prices below some $\epsilon \ll v$. In hindsight, posting a price of ϵ in every round would appear to generate the most revenue for the seller given the observed sequence of buyer actions, and therefore ϵT cumulative revenue is "no-regret". However, the seller was tricked by the strategic buyer; there was $(v - \epsilon)T$ revenue left on the table. Moreoever, this is a good strategy for the buyer (it must have won the good for nearly nothing on $\Omega(T)$ rounds).

The main contribution of this paper is extending the setting described above to one where the buyer's valuations in each round are a function of some context observed by both the buyer and seller. While [2] is motivated by our same application, they imagine an overly simplistic model wherein the buyer's value is generated by drawing an independent v_t from an unknown distribution $\mathcal{D}$. This ignores that v_t will in reality be a function of contextual information $\mathbf{x}_t$, information that is available to the seller, and the entire reason auctions are thin to begin with (without $\mathbf{x}_t$ there would be no targeting). We give the first algorithm that attains sublinear regret in the contextual setting, against a surplus-maximizing buyer. We also note that in the non-contextual setting, regret is measured against the revenue that could have been made if $\mathcal{D}$ were known, and the single fixed optimal price were selected. Our comparator will be more challenging as we wish to compete with the best *function* (in some class) from contexts $\mathbf{x}_t$ to prices.

The rest of the paper is organized as follows. We first introduce a linear model by which values v_t are derived from contexts $\mathbf{x}_t$. We then demonstrate an algorithm based on stochastic gradient descent (SGD) which achieves sublinear regret against an truthful buyer (one that accepts price p_t iff $p_t \leq v_t$ on every round t). The analysis for the truthful buyer uses prexisting high probability bounds for SGD when minimizing strongly convex functions [22]. Our main result requires an extension of this analysis to cases in which "incorrect" gradients are occasionally observed. This lets us study a buyer that is allowed to best-respond to our algorithm, possibly rejecting offers that the truthful buyer would not, in order to receive better offers on future rounds. We also adapt our algorithm to non-linear settings via a kernelized version of the algorithm. Finally, we extend our results to second-price auctions with multiple buyers.

Related Work: The pricing of digital good in repeated auctions has been considered by many other authors, including [2, 17, 4, 3, 6, 19]. However, most of these papers do not consider a buyer who behaves strategically *across rounds*. Buyers either behave randomly [19], or only participate in a single round [17, 4, 3, 6], or participate in multiple rounds but only desire a single good [20, 12] and therefore, in each of these cases, are not incentivized to manipulate the seller's behavior on future rounds. In reality buyers repeatedly interact with the same seller. There is empirical evidence suggesting that buyers are not myopic, and do in fact behave strategically to induce better prices in the future [9], as well as literature studying different strategies for strategic buyers [5, 15, 16].

Repeated posted price actions against the same strategic buyer have been considered in the economics literature under the heading of *behavior-based price discrimination* (BBPD) by [13, 23, 1, 11], and more recently by [8]. These works differ from ours in two key ways. First, all these works imagine that the buyer's type is drawn from some fixed publicly available distribution. Therefore learning $\mathcal{D}$ is not at issue. In contrast, we argue that access to an accurate prior is particularly problematic in these settings. After all, the seller cannot expect to reliably estimate $\mathcal{D}$ from data when the buyer is explicitly incentivized to hide its type (as illustrated in the Introduction; see also [2]). This tension between learning and buyer truthfulness is in many ways central to our study.

Secondly, given a fixed prior, the most common solution concept in the BBPD literature is a perfect Bayes-Nash equilibrium, in which both the seller and buyer strategies are best responses to each other. However, in the context of Internet advertising, a seller must first deploy an algorithm which

automates the pricing strategy, and buyers subsequently react to the observed behavior of the pricing algorithm. Any modifications the seller wishes to make to the pricing algorithm will typically require changes to the end-user licensing agreement, which the seller will not want to do too frequently. Therefore, in this paper, we make a commitment assumption on the seller: the seller acts first, announcing its pricing strategy, after which the buyer plays a best response strategy. Such Stackleberg models of commitment [10] have sparked a great deal of recent interest due to their success in security games (see [7] and [18] for an overview), including practical deployment [21, 14].

2 Preliminaries

Throughout this work, we will consider a repeated auction where at every round a single seller prices an item to sell to a single buyer (extensions to multiple buyers are discussed in Section 5). The good sold at step t in the repeated auction is represented by a context (feature) vector $\mathbf{x}_t \in \mathcal{X} = \{\mathbf{x} \colon \|\mathbf{x}\|_2 \leq 1\}$ and is drawn according a fixed distribution $\mathcal{D}$, which is unknown to the seller. The good has a value v_t that is a linear function of a parameter vector $\mathbf{w}^*$, also unknown to the seller, $v_t = {\mathbf{w}^*}^\top \mathbf{x}_t$ (extensions to non-linear functions of the context are considered in Section 5). We assume that $\mathbf{w}^* \in \mathcal{W} = \{\mathbf{w} \colon \|\mathbf{w}\|_2 \leq 1\}$ and also that $0 \leq {\mathbf{w}^*}^\top \mathbf{x} \leq 1$ with probability one with respect to $\mathcal{D}$.

For rounds $t = 1, \ldots, T$ the repeated posted-price auction is defined as follows: (1) The buyer and seller both observe $\mathbf{x}_t \sim \mathcal{D}$. (2) The seller offers a price p_t. (3) The buyer selects $a_t \in \{0, 1\}$. (4) The seller receives revenue $a_t p_t$.

Here, a_t is an indicator variable that represents whether or not the buyer accepted the offered price (1 indicates yes). The goal of the seller is to select a price p_t in each round t such that the expected *regret* $R(T) = \mathrm{E}\left[\sum_{t=1}^{T} v_t - a_t p_t\right]$ is $o(T)$. The choice of a_t will depend on the buyer's behavior. We will analyze two types of buyers in the subsequent sections of the paper: *truthful* and *surplus-maximizing* buyers, and will attempt to minimize *regret against the truthful buyer* and *regret against the surplus-maximizing buyer*. Note, the regret is the difference between the maximum revenue possible and the amount made by the algorithm that offers prices to the buyer.

3 Truthful Buyer

In this section we introduce the Learn-Exploit Algorithm for Pricing (LEAP), which we show has regret of the form $O(T^{2/3}\sqrt{\log(T)})$ against a *truthful* buyer. A buyer is truthful if she accepts any offered price that gives a non-negative *surplus*, which is defined as the difference between the buyer's value for the good minus the price paid: $v_t - p_t$. Therefore, for a truthful buyer we define $a_t = \mathbf{1}\{p_t \leq v_t\}$.

At this point, we note that the loss function $v_t - \mathbf{1}\{p_t \leq v_t\}p_t$, which we wish to minimize over all rounds, is not convex, differentiable or even continuous. If the price is even slightly above the truthful buyers valuation it is rejected and the seller makes zero revenue. To circumvent this our algorithm will attempt to learn $\mathbf{w}^*$ directly by minimizing a surrogate loss function for which $\mathbf{w}^*$ in the minimizer. Our analysis hinges on recent results [22] which give optimal rates for gradient descent when the function being minimized is strongly convex. Our key trick is to offer prices so that, in each round, the buyer's behavior reveals the gradient of the surrogate loss at our current estimate for $\mathbf{w}^*$. Below we define the LEAP algorithm (Algorithm 1), which we show addresses these difficulties in the online setting.

The algorithm depends on input parameters α, ϵ and λ. The α parameter determines what fraction of rounds are spent in the learning phase as oppose to the exploit phase. During the learning phase, uniform random prices are offered and the model parameters are updated as a function of the feedback given by the buyer. During the exploit phase, the model parameters are fixed and the offered price is computed as a linear function of these parameters minus the value of the ϵ parameter. The ϵ parameter can be thought of as inversely proportional to our confidence in the fixed model parameters and is used to hedge against the possibility of over-estimating the value of a good. The λ parameter is a learning-rate parameter set according to the minimum eigenvalue of the covariance matrix, and is defined below in Assumption 1. In order to prove a regret bound, we first show that

Algorithm 1 LEAP algorithm

- Let $0 \leq \alpha \leq 1$, $\mathbf{w}_1 = \mathbf{0} \in \mathcal{W}$, $\epsilon \geq 0$, $\lambda > 0$, $T_\alpha = \lceil \alpha T \rceil$.
- For $t = 1, \ldots, T_\alpha$ (Learning phase)
  - Offer $p_t \sim U$, where U is the uniform distribution on the interval $[0, 1]$.
  - Observe a_t.
  - $\tilde{\mathbf{g}}_t = 2\big(\mathbf{w}_t \cdot \mathbf{x}_t - a_t\big)\mathbf{x}_t$.
  - $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \frac{1}{\lambda t}\tilde{\mathbf{g}}_t)$.
- For $t = T_\alpha + 1, \ldots, T$ (Exploit phase)
  - Offer $p_t = \mathbf{w}_{T_\alpha + 1} \cdot \mathbf{x}_t - \epsilon$.

the learning phase of the algorithm is minimizing a strongly convex surrogate loss and then show that this implies the buyer enjoys near optimal revenue during the exploit phase of the algorithm.

Let $\mathbf{g}_t = 2(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{1}\{p_t \leq v_t\})\mathbf{x}_t$ and $F(\mathbf{w}) = \mathrm{E}_{\mathbf{x}\sim\mathcal{D}}\big[(\mathbf{w}^{*\top}\mathbf{x} - \mathbf{w}^\top\mathbf{x})^2\big]$. Note that when the buyer is truthful $\tilde{\mathbf{g}}_t = \mathbf{g}_t$. Against a truthful buyer, $\mathbf{g}_t$ is an unbiased estimate of the gradient of F.

Proposition 1. *The random variable $\mathbf{g}_t$ satisfies $\mathrm{E}[\mathbf{g}_t \mid \mathbf{w}_t] = \nabla F(\mathbf{w}_t)$. Also, $\|\mathbf{g}_t\| \leq 4$ with probability 1.*

Proof. First note that $\mathrm{E}[\mathbf{g}_t \mid \mathbf{w}_t] = \mathrm{E}_{\mathbf{x}_t}\big[2\big(\mathbf{w}_t \cdot \mathbf{x}_t - \mathrm{E}_{p_t}[\mathbf{1}\{p_t \leq v_t\}]\big)\big] = \mathrm{E}_{\mathbf{x}_t}\big[2\big(\mathbf{w}_t \cdot \mathbf{x}_t - \Pr_{p_t}(p_t \leq v_t)\big)\big]$. Since p_t is drawn uniformly from $[0, 1]$ and v_t is guaranteed to lie in $[0, 1]$ we have that $\Pr(p_t \leq v_t) = \int_0^1 \mathbf{1}\{p_t \leq v_t\} dp_t = v_t$. Plugging this back into $\mathbf{g}_t$ gives us exactly the expression for $\nabla F(\mathbf{w}_t)$. Furthermore, $\|\mathbf{g}_t\| = 2|\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{1}\{p_t \leq v_t\}|\,\|\mathbf{x}_t\| \leq 4$ since $|\mathbf{w}_t^\top \mathbf{x}_t| \leq \|\mathbf{w}_t\|\|\mathbf{x}_t\| \leq 1$ and $\|\mathbf{x}_t\| \leq 1$ □

We now introduce the notion of *strong convexity*. A twice-differentiable function $H(\mathbf{w})$ is λ-strongly convex if and only if the Hessian matrix $\nabla^2 H(\mathbf{w})$ is full rank and the minimum eigenvalue of $\nabla^2 H(\mathbf{w})$ is at least λ. Note that the function F is strongly convex if and only if the covariance matrix of the data is full-rank, since $\nabla^2 F(\mathbf{w}) = 2\mathrm{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$. We make the following assumption.

Assumption 1. *The minimum eigenvalue of $2\mathrm{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$ is at least $\lambda > 0$.*

Note that if this is not the case then there is redundancy in the features and the data can be projected (for example using PCA) into a lower dimensional feature space with a full-rank covariance matrix and without any loss in information. The seller can compute an offline estimate of both this projection and λ by collecting a dataset of context vectors before starting to offer prices to the buyer.

Thus, in view of Proposition 1 and the strong convexity assumption, we see the learning phase of the LEAP algorithm is conducting a stochastic gradient descent to minimize the λ-strongly convex function F, where at each time step we update $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \frac{1}{\lambda t}\tilde{\mathbf{g}}_t)$ and $\tilde{\mathbf{g}}_t = \mathbf{g}_t$ is an unbiased estimate of the gradient. We now make use of an existing bound ([22]) for stochastic gradient descent on strongly convex functions.

Lemma 1 ([22] Proposition 1)**.** *Let $\delta \in (0, 1/e)$, $T_\alpha \geq 4$ and suppose F is λ-strongly convex over the convex set $\mathcal{W}$. Also suppose $\mathrm{E}[\mathbf{g}_t \mid \mathbf{w}_t] = \nabla F(\mathbf{w})$ and $\|\mathbf{g}_t\|^2 \leq G^2$ with probability 1. Then with probability at least $1 - \delta$ for any $t \leq T_\alpha$ it holds that*

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq \frac{(624\log(\log(T_\alpha)/\delta) + 1)G^2}{\lambda^2 t} \quad \textit{where} \quad \mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}} F(\mathbf{w})\,.$$

This guarantees that, with high probability, the distance between the learned parameter vector $\mathbf{w}_t$ and the target weight vector $\mathbf{w}^*$ is bounded and decreasing as t increases. This allows us to carefully tune the ϵ parameter that is used in the exploit phase of the algorithm (see Lemma 6 in the appendix). We are now equipped to prove a bound on the regret of the LEAP algorithm.

Theorem 1. *For any $T > 4$, $0 < \alpha < 1$ and assuming a truthful buyer, the LEAP algorithm with $\epsilon = \sqrt{\frac{(624\log(\sqrt{T_\alpha}\log(T_\alpha))+1)G^2}{\lambda^2 T_\alpha}}$, where $G = 4$, has regret against a truthful buyer at most*

$R(T) \leq 2\alpha T + 4\sqrt{\frac{T}{\alpha}}\sqrt{\frac{(624\log(\sqrt{T_\alpha}\log(T_\alpha))+1)G^2}{\lambda^2}}$, *which implies for* $\alpha = T^{-1/3}$ *a regret at most*

$$R(T) \leq 2T^{2/3} + 4T^{2/3}\sqrt{\frac{(624\log(T^{1/3}\log(T^{2/3}))+1)G^2}{\lambda^2}} = O\Big(T^{2/3}\sqrt{\log(T)}\Big)\,.$$

Proof. We first decompose the regret

$$\mathrm{E}\Big[\sum_{t=1}^{T} v_t - a_t p_t\Big] = \mathrm{E}\Big[\sum_{t=1}^{T_\alpha} v_t - a_t p_t\Big] + \mathrm{E}\Big[\sum_{t=T_\alpha+1}^{T} v_t - a_t p_t\Big] \leq T_\alpha + \sum_{t=T_\alpha+1}^{T} \mathrm{E}\Big[v_t - a_t p_t\Big]\,, \quad (1)$$

where we have used the fact $|v_t - a_t p_t| \leq 1$. Let A denote the event that, for all $t \in \{T_\alpha+1, \ldots, T\}$, $a_t = 1 \wedge v_t - p_t \leq \epsilon$. Lemma 6 (see Appendix, Section A.1) proves that A occurs with probability at least $1 - T_\alpha^{-1/2}$. For brevity let $N = \sqrt{(624\log(\sqrt{T_\alpha}\log(T_\alpha))+1)G^2/\lambda^2}$, then we can decompose the expectation in the following way:

$$\begin{aligned}\mathrm{E}\Big[v_t - a_t p_t\Big] &= \Pr[A]E[v_t - a_t p_t|A] + (1-\Pr[A])E[v_t - a_t p_t|\neg A] \\ &\leq \Pr[A]\epsilon + (1-\Pr[A]) \leq \epsilon + T_\alpha^{-1/2} = \sqrt{\frac{N}{T_\alpha}} + \sqrt{\frac{1}{T_\alpha}} \leq 2\sqrt{\frac{N}{T_\alpha}}\,,\end{aligned}$$

where the inequalities follow from the definition of A, Lemma 6, and the fact that $|v_t - a_t p_t| < 1$. Plugging this back into equation (1) gives $T_\alpha + \sum_{t=T_\alpha+1}^{T} \mathrm{E}[v_t - a_t p_t] \leq T_\alpha + \frac{\lceil (1-\alpha)T \rceil}{\sqrt{T_\alpha}} 2\sqrt{N} \leq 2\alpha T + 4\sqrt{\frac{T}{\alpha}}\sqrt{N}$, proving the first result of the theorem. $\alpha = T^{-1/3}$ gives the final expression. □

In the next section we consider the more challenging setting of a surplus-maximizing buyer, who may accept/reject prices in a manner meant to lower the prices offered.

4 Surplus-Maximizing Buyer

In the previous section we considered a truthful buyer who myopically accepts every price below her value, i.e., she sets $a_t = \mathbf{1}\{p_t \leq v_t\}$ for every round t. Let $S(T) = E\left[\sum_{t=1}^{T} \gamma_t a_t (v_t - p_t)\right]$ be the buyer's cumulative discounted surplus, where $\{\gamma_t\}$ is a decreasing discount sequence, with $\gamma_t \in (0,1)$. When prices are offered by the LEAP algorithm, the buyer's decisions about which prices to accept during the learning phase have an influence on the prices that she is offered in the exploit phase, and so a surplus-maximizing buyer may be able to increase her cumulative discounted surplus by occasionally behaving untruthfully. In this section we assume that the buyer knows the pricing algorithm and seeks to maximize $S(T)$.

Assumption 2. *The buyer is surplus-maximizing, i.e., she behaves so as to maximize* $S(T)$*, given the seller's pricing algorithm.*

We say that a *lie* occurs in any round t where $a_t \neq \mathbf{1}\{p_t \leq v_t\}$. Note that a surplus-maximizing buyer has no reason to lie during the exploit phase, since the buyer's behavior during exploit rounds has no effect on the prices offered. Let $\mathcal{L} = \{t : 1 \leq t \leq T_\alpha \wedge a_t \neq \mathbf{1}\{p_t \leq v_t\}\}$ be the set of learning rounds where the buyer lies, and let $L = |\mathcal{L}|$ be the number of lies. Observe that $\tilde{\mathbf{g}}_t \neq \mathbf{g}_t$ in any lie round (recall that $\mathrm{E}[\mathbf{g}_t \mid \mathbf{w}_t] = \nabla F(\mathbf{w}_t)$, i.e., $\mathbf{g}_t$ is the stochastic gradient in round t).

We take a moment to note the necessity of the discount factor γ_t. This essentially models the buyer as valuing surplus at the current time step more than in the future. Another way of interpreting this, is that the seller is more "patient" as compared to the buyer. In [2] the authors show a lower bound on the regret against a surplus-maximizing buyer in the contextless setting of the form $O(T_\gamma)$, where $T_\gamma = \sum_{i=1}^{T} \gamma_t$. Thus, if no decreasing discount factor is used, i.e. $\gamma_t = 1$, then sublinear regret is not possible. Note, the lower bound of the contextless setting applies here as well, since the case of a distribution $\mathcal{D}$ that induces a fixed context $\mathbf{x}^*$ on every round is a special case of our setting. In that case the problem reduces to the fixed unknown value setting since on every round $v_t = \mathbf{w}^{*\top}\mathbf{x}^*$.

In the rest of this section we prove an $O\big(T^{2/3}\sqrt{\log(T)(1+1/\log(1/\gamma))}\big)$ bound on the seller's regret under the assumption that the buyer is surplus-maximizing and that her discount sequence is

$\gamma_t = \gamma^{t-1}$ for some $\gamma \in (0,1)$. The idea of the proof is to show that the buyer incurs a cost for telling lies, and therefore will not tell very many, and thus the lies she does tell will not significantly affect the seller's estimate of $\mathbf{w}^*$.

Bounding the cost of lies: Observe that in any learning round where the surplus-maximizing buyer tells a lie, she loses surplus in that round relative to the truthful buyer, either by accepting a price higher than her value (when $a_t = 1$ and $v_t < p_t$) or by rejecting a price less than her value (when $a_t = 0$ and $v_t > p_t$). This observation can be used to show that lies result in a substantial loss of surplus relative to the truthful buyer, provided that in most of the lie rounds there is a nontrivial gap between the buyer's value and the seller's price. Because prices are chosen uniformly at random during the learning phase, this is in fact quite likely, and with high probability the surplus lost relative to the truthful buyer during the learning phase grows exponentially with the number of lies. The precise quantity is stated in the Lemma below. A full proof appears in the appendix, Section A.3.

Lemma 2. *Let the discount sequence be defined as $\gamma_t = \gamma^{t-1}$ for $0 < \gamma < 1$ and assume the buyer has told L lies. Then for $\delta > 0$ with probability at least $1 - \delta$ the buyer loses a surplus of at least $\frac{\gamma^{-L+3}-1}{8T_\alpha \log(\frac{1}{\delta})}\left(\frac{\gamma^{T_\alpha}}{1-\gamma}\right)$ relative to the truthful buyer during the learning phase.*

Bounding the number of lies: Although we argued in the previous lemma that lies during the learning phase cause the surplus-maximizing buyer to lose surplus relative to the truthful buyer, those lies may result in lower prices offered during the exploit phase, and thus the overall effect of lying may be beneficial to the buyer. However, we show that there is a limit on how large that benefit can be, and thus we have the following high-probability bound on the number of learning phase lies.

Lemma 3. *Let the discount sequence be defined as $\gamma_t = \gamma^{t-1}$ for $0 < \gamma < 1$. Then for $\delta > 0$ with probability at least $1 - \delta$, the number of lies $L \leq \frac{\log(32T_\alpha \frac{1}{\delta}\log(\frac{2}{\delta})+1)}{\log(1/\gamma)}$.*

The full proof is found in the appendix (Section A.4), and we provide a proof sketch here. The argument proceeds by comparing the amount of surplus lost (compared to the truthful buyer) due to telling lies in the learning phase to the amount of surplus that could hope to be gained (compared to the truthful buyer) in the exploit phase. Due to the discount factor, the surplus lost will eventually outweigh the surplus gained as the number of lies increases, implying a limit to the number of lies a surplus maximizing buyer can tell.

Bounding the effect of lies: In Section 3 we argued that if the buyer is truthful then, in each learning round t of the LEAP algorithm, $\tilde{\mathbf{g}}_t$ is a stochastic gradient with expected value $\nabla F(\mathbf{w}_t)$. We then use the analysis of stochastic gradient descent in [22] to prove that $\mathbf{w}_{T_\alpha+1}$ converges to $\mathbf{w}^*$ (Lemma 1). However, if the buyer can lie then $\tilde{\mathbf{g}}_t$ is not necessarily the gradient and Lemma 1 no longer applies. Below we extend the analysis in Rakhlin et al. [22] to a setting where the gradient may be corrupted by lies up to L times.

Lemma 4. *Let $\delta \in (0, 1/e)$, $T_\alpha \geq 2$. If the buyer tells L lies then with probability at least $1 - \delta$, $\|\mathbf{w}_{T_\alpha+1} - \mathbf{w}^*\|^2 \leq \frac{1}{T_\alpha+1}\left(\frac{(624\log(\log(T_\alpha)/\delta)+e^2)G^2}{\lambda^2} + \frac{4e^2 L}{\lambda}\right)$.*

The proof of the lemma is similar to that of Lemma 1, but with extra steps needed to bound the additional error introduced due to the erroneous gradients. Due to space constraints, we present the proof in the appendix, Section A.6. Note that, modulo constants, the bound only differs by the additive term L/T_α. That is, there is an extra additive error term that depends on the ratio of lies to number of learning rounds. Thus, if no lies are told, then there is no additive error. While if many lies are told, e.g. $L = T_\alpha$, then the bound become vacuous.

Main result: We are now ready to prove an upper bound on the regret of the LEAP algorithm when the buyer is surplus-maximizing.

Theorem 2. *For any $0 < \alpha < 1$ (such that $T_\alpha \geq 4$), $0 < \gamma < 1$ and assuming a surplus-maximizing buyer with exponential discounting factor $\gamma_t = \gamma^{t-1}$, then the LEAP algorithm using parameter $\epsilon = \sqrt{\frac{1}{T_\alpha}\left(\frac{(624\log(2\sqrt{T_\alpha}\log(T_\alpha))+e^2)G^2}{\lambda^2} + \frac{4e^2\log(128\sqrt{T_\alpha}\log(4\sqrt{T_\alpha})+1)}{\lambda\log(1/\gamma)}\right)}$, where $G = 4$, has regret against a surplus-maximizing buyer at most*

$$R(T) \leq 2\alpha T + 4\sqrt{\frac{T}{\alpha}}\sqrt{\frac{(624\log(2\sqrt{T_\alpha}\log(T_\alpha)) + e^2)G^2}{\lambda^2} + \frac{4e^2\log(128\sqrt{T_\alpha}\log(4\sqrt{T_\alpha}) + 1)}{\lambda\log(1/\gamma)}},$$

which for $\alpha = T^{-1/3}$ *implies* $R(T) \leq O\left(T^{2/3}\sqrt{\log(T)\left(1+\frac{1}{\log(1/\gamma)}\right)}\right)$.

Proof. Taking the high probability statements of Lemma 3 and Lemma 4 with $\delta/2 \in [0, 1/e]$ tells us that with probability at least $1-\delta$, $\|\mathbf{w}_{T_\alpha} - \mathbf{w}^*\|^2 \leq \frac{1}{T_\alpha}\left(\frac{(624\log(2\log(T_\alpha)/\delta)+e^2)G^2}{\lambda^2} + \frac{4e^2\log(64T_\alpha\frac{1}{\delta}\log(\frac{4}{\delta})+1)}{\lambda\log(1/\gamma)}\right)$.

Since we assume $T_\alpha \geq 4$, if we set $\delta = T_\alpha^{-1/2}$ it implies $\delta/2 = T_\alpha^{-1/2}/2 \leq 1/e$, which is required for Lemma 4 to hold. Thus, if we set the algorithm parameter ϵ as indicated in the statement of theorem, we have that with probability at least $1 - T_\alpha^{-1/2}$ for all $t \in \{T_\alpha + 1, \ldots, T\}$ that $a_t = 1$ and $v_t - p_t \leq \epsilon$, which follows from the same argument used for Lemma 6.

Finally, the same steps as in the proof of Theorem 1 we can be used to show the first inequality. Setting $\alpha = T^{-1/3}$ shows the second inequality and completes the theorem. □

Note that the bound shows that if $\gamma \to 1$ (i.e. no discounting) the bound becomes vacuous, which is to be expected since the $\Omega(T_\gamma)$ lower bound on regret demonstrates the necessity of a discounting factor. If $\gamma \to 0$ (i.e. buyer become myopic, thereby truthful), then we retrieve the truthful bound modulo constants. Thus for any $\gamma < 1$, we have shown the first sublinear bound on the seller's regret against a surplus-maximizing buyer in the contextual setting.

5 Extensions

Doubling trick: A drawback of Theorem 2 is that optimally tuning the parameters ϵ and α requires knowledge of the horizon T. The usual way of handling this problem in the standard online learning setting is to apply the 'doubling trick': If a learning algorithm that requires knowledge of T has regret $O(T^c)$ for some constant c, then running independent instances of the algorithm during consecutive phases of exponentially increasing length (i.e., the ith phase has length 2^i) will also have regret $O(T^c)$. We can also apply the doubling trick to our strategic setting, but we must exercise caution and argue that running the algorithm in phases does not affect the behavior of a surplus-maximizing buyer in a way that invalidates the proof of Theorem 2. We formally state and prove the relevant corollary in Section A.8 of the Appendix.

Kernelized Algorithm: In some cases, assuming that the value of a buyer is a linear function of the context may not be accurate. In Section A.7 of the Appendix we describe a kernelized version of LEAP, which allows for a non-linear model of the buyer value as a function of the context x. At the same time, the regret guarantees provided in the previous sections still apply since we can view the model as linear function of the induced features $\phi(x)$, where $\phi(\cdot)$ is a non-linear map and the kernel function K is used to compute the inner product in this induced feature space: $K(x, x') = \phi(x)^\top \phi(x')$.

Multiple Buyers: So far we have assumed that the seller is interacting with a single buyer across multiple posted price auctions. Recall that the motivation for considering this setting was repeated *second price* auctions against a *single* buyer, a situation that happens often in online advertising because of targetting. One might nevertheless wonder whether the algorithm can be applied to a setting where there can be multiple buyers, and whether it remains robust in such a setting. We describe a way in which the analysis for the posted-price setting can carry over to multiple buyers.

Formally, suppose there are K buyers, and on round t, buyer k receives a valuation of $v_{k,t}$. We let $k^{\text{val}}(t) = \arg\max_k v_{k,t}$, $v_t^+ = v_{k^{\text{val}}(t),t}$, and $v_t^- = \max_{k \neq k^{\text{val}}(t)} v_{k,t}$: the buyer with the highest valuation, the highest valuation itself, and the second-highest valuation respectively. In a second price auction, each buyer also submits a bid $b_{k,t}$, and we define $k^{\text{bid}}(t)$, b_t^+ and b_t^- analogously to $k^{\text{val}}(t)$, v_t^+, v_t^-, corresponding to the highest bidder, the largest bid, and the second-largest bid. After the seller announces a reserve price p_t, buyers submit their bids $\{b_{k,t}\}$, and the seller receives round t revenue of $r_t = \mathbf{1}\{p_t \leq b_t^+\}\max\{b_t^-, p_t\}$. The goal of the seller is to minimize $R(T) = \mathrm{E}[\sum_{t=1}^T v_t^+ - r_t]$. We assume that buyers are surplus-maximizing, and select a strategy that maps previous reserve prices $p_1, ..., p_{t-1}, p_t$, and $v_{k,t}$ to a choice of bid on round t.

We call v_t^+ the *market valuation* for good t. The key to extending the LEAP algorithm to the multiple buyer setting will be to treat market valuations in the same way we treated the individual buyer's valuation in the single-buyer setting. In order to do so, we make an analogous modelling assumption to that of Section 2. Specifically, we assume that there is some $\mathbf{w}^*$ such that $v_t^+ = {\mathbf{w}^*}_t^\top \mathbf{x}_t$.[1] Note that we assume a model on the *market price* itself.

At first glance, this might seem like a strange assumption since v_t^+ is itself the result of a maximization over $v_{k,t}$. However, we argue that it's actually rather unrestrictive. In fact the individual valuations $v_{k,t}$ can be generated arbitrarily so long as $v_{k,t} \leq {\mathbf{w}^*}_t^\top \mathbf{x}_t$ and equality holds for some k. In other words, we can imagine that nature first computes the market valuation v_t^+, then arbitrarily (even adversarialy) selects which buyer gets this valuation, and the other buyer valuations.

Now we can define $a_t = \mathbf{1}\{p_t \leq b_t^+\}$, whether the largest bid was greater than the reserve, and consider running the LEAP algorithm, but with this choice of a_t. Notice that for any t, $a_t p_t \leq r_t$, thereby giving us the following key fact: $R(T) \leq R'(T) \triangleq \mathrm{E}[\sum_{t=1}^T v_t^+ - a_t p_t]$. We also redefine L to be the number of *market lies*: rounds $t \leq T_\alpha$ where $a_t \neq \mathbf{1}\{p_t \leq v_t^+\}$. Note the market tells a lie if either all valuations were below p_t, but somebody bid over p_t anyway, or if some valuation was above p_t but no buyer decided to outbid p_t. With this choice of L, Lemma 4 holds exactly as written but in the multiple buyer setting.

It's well-known [24] that single-shot second price auctions are *strategy-proof*. Therefore, during the exploit phase of the algorithm, all buyers are incentivized to bid truthfully. Thus, in order to bound $R'(T)$ and therefore $R(T)$, we need only rederive Lemma 3 to bound the number of *market lies*. We begin partitioning the market lies. Let $\mathcal{L} = \{t : t \leq T_\alpha, \mathbf{1}\{p_t \leq v_t^+\} \neq \mathbf{1}\{p_t \leq b_t^+\}\}$, while letting $\mathcal{L}_k = \{t : t \leq T_\alpha, v_t^+ < p_t^+ \leq b_t^+, k^{\mathrm{bid}}(t) = k\} \cup \{t \leq T_\alpha, b_t^+ < p_t \leq v_t^+, k^{\mathrm{val}}(t) = k\}$. In other words, we attribute a lie to buyer k if (1) the reserve was larger than the market value, but buyer k won the auction anyway, or (2) buyer k had the largest valuation, but nobody cleared the reserve. Checking that $\mathcal{L} = \cup_k \mathcal{L}_k$ and letting $L_k = |\mathcal{L}_k|$ tells us that $L \leq \sum_{k=1}^K L_k$. Furthermore, we can bound L_k using nearly identical arguments to the posted price setting, giving us the subsequent Corollary for the multiple buyer setting.

Lemma 5. *Let the discount sequence be defined as $\gamma_t = \gamma^{t-1}$ for $0 < \gamma < 1$. Then for $\delta > 0$ with probability at least $1 - \delta$, $L_k \leq \frac{\log(32T_\alpha/\delta+1)}{\log(1/\gamma)}$, and $L \leq KL_k$.*

Proof. We first consider the surplus buyer k loses during learning rounds, compared to if he had been truthful. Suppose buyer k unilateraly switches to always bidding his value (i.e. $b_{k,t} = v_{k,t}$). For a single-shot second price auction, being truthful is a dominant strategy and so he would only increase his surplus on learning rounds. Furthermore, on each round in $\mathcal{L}_k$ he would increase his (undiscounted) surplus by at least $|v_{k,t} - p_t|$. Now the analysis follows as in Lemmas 2 and 3. □

Corollary 1. *In the multiple surplus-maximizing buyers setting the LEAP algorithm with $\alpha = T^{-1/3}$, $\epsilon = \sqrt{\frac{1}{T_\alpha}\left(\frac{(624\log(2\sqrt{T_\alpha}\log(T_\alpha))+e^2)G^2}{\lambda^2} + \frac{4e^2K\log(128\sqrt{T_\alpha}\log(4\sqrt{T_\alpha})+1)}{\lambda\log(1/\gamma)}\right)}$, has regret $R(T) \leq R'(T) \leq O\left(T^{2/3}\sqrt{\log(T) + \frac{K\log(T)}{\log(1/\gamma)}}\right)$*

6 Conclusion

In this work, we have introduced the scenario of contextual auctions in the presence of surplus-maximizing buyers and have presented an algorithm that is able to achieve sublinear regret in this setting, assuming buyers receive a discounted surplus. Once again, we stress the importance of the contextual setting, as it contributes to the rise of targeted bids that result in auction with single high-bidders, essentially reducing the auction to the posted-price scenario studied in this paper. Future directions for extending this work include considering different surplus discount rates as well as understanding whether small modifications to standard contextual online learning algorithms can lead to no-strategic-regret guarantees.

[1] Note that we could also apply the kernelized LEAP algorithm in the multiple buyer setting.

References

[1] Alessandro Acquisti and Hal R Varian. Conditioning prices on purchase history. *Marketing Science*, 24 (3):367–381, 2005.

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

[3] Ziv Bar-Yossef, Kirsten Hildrum, and Felix Wu. Incentive-compatible online auctions for digital goods. In *Proceedings of Symposium on Discrete Algorithms*, pages 964–970. SIAM, 2002.

[4] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. In *Proceedings Symposium on Discrete algorithms*, pages 202–204. SIAM, 2003.

[5] Matthew Cary, Aparna Das, Ben Edelman, Ioannis Giotis, Kurtis Heimerl, Anna R Karlin, Claire Mathieu, and Michael Schwarz. Greedy bidding strategies for keyword auctions. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 262–271. ACM, 2007.

[6] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Symposium on Discrete Algorithms*. SIAM, 2013.

[7] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90. ACM, 2006.

[8] Nikhil R Devanur, Yuval Peres, and Balasubramanian Sivan. Perfect bayesian equilibria in repeated sales. *arXiv preprint arXiv:1409.3062*, 2014.

[9] Benjamin Edelman and Michael Ostrovsky. Strategic bidder behavior in sponsored search auctions. *Decision support systems*, 43(1):192–198, 2007.

[10] Drew Fudenberg and Jean Tirole. Game theory. *MIT Press Books*, 1, 1991.

[11] Drew Fudenberg and J Miguel Villas-Boas. Behavior-based price discrimination and customer recognition. *Handbook on economics and information systems*, 1:377–436, 2006.

[12] Mohammad Taghi Hajiaghayi, Robert Kleinberg, and David C Parkes. Adaptive limited-supply online auctions. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 71–80. ACM, 2004.

[13] Oliver D Hart and Jean Tirole. Contract renegotiation and coasian dynamics. *The Review of Economic Studies*, 55(4):509–540, 1988.

[14] Manish Jain, Jason Tsai, James Pita, Christopher Kiekintveld, Shyamsunder Rathi, Milind Tambe, and Fernando Ordóñez. Software assistants for randomized patrol planning for the lax airport police and the federal air marshal service. *Interfaces*, 40(4):267–290, 2010.

[15] Brendan Kitts and Benjamin Leblanc. Optimal bidding on keyword auctions. *Electronic Markets*, 14(3): 186–201, 2004.

[16] Brendan Kitts, Parameshvyas Laxminarayan, Benjamin Leblanc, and Ryan Meech. A formal analysis of search auctions including predictions on click fraud and bidding tactics. In *Workshop on Sponsored Search Auctions*, 2005.

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

[18] Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. Stackelberg vs. nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *J. Artif. Intell. Res.(JAIR)*, 41:297–327, 2011.

[19] Andres Munoz Medina and Mehryar Mohri. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *Proceedings of The 31st International Conference on Machine Learning*, pages 262–270, 2014.

[20] David C Parkes. Online mechanisms. In Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[21] James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Deployed armor protection: the application of a game theoretic model for security at the los angeles international airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, pages 125–132. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[22] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[23] Klaus M Schmidt. Commitment through incomplete information in a simple repeated bargaining game. *Journal of Economic Theory*, 60(1):114–139, 1993.

[24] Hal R Varian and Jack Repcheck. *Intermediate microeconomics: a modern approach*, volume 6. WW Norton & Company New York, NY, 2010.

A Appendix

A.1 Selecting the ϵ parameter

Lemma 6. *Assume $T_\alpha \geq 4$. Then using the LEAP algorithm, in the presence of a truthful buyer, ensures that with probability at least $1 - T_\alpha^{-1/2}$ for all $t \in \{T_\alpha + 1, \ldots, T\}$ we have $a_t = 1$ and $v_t - p_t \leq \epsilon = \sqrt{\frac{(624 \log(\sqrt{T_\alpha} \log(T_\alpha)) + 1) G^2}{\lambda^2 T_\alpha}}$.*

Proof. Using Lemma 1, we have with probability at least $1 - T_\alpha^{-1/2}$ for $\mathbf{x} \in \mathcal{X}$

$$|\mathbf{w}^* \cdot \mathbf{x} - \mathbf{w}_{T_\alpha} \cdot \mathbf{x}| = |(\mathbf{w}^* - \mathbf{w}_{T_\alpha}) \cdot \mathbf{x}| \leq \|\mathbf{w}^* - \mathbf{w}_{T_\alpha}\| \|\mathbf{x}\| \leq \|\mathbf{w}^* - \mathbf{w}_{T_\alpha}\| \leq \sqrt{\frac{(624 \log(\sqrt{T_\alpha} \log(T_\alpha)) + 1) G^2}{\lambda^2 T_\alpha}}\,.$$

Therefore with probability $1 - T_\alpha^{-1/2}$ for all $t \in \{T_\alpha + 1, \ldots, T\}$

$\mathbf{w}^* \cdot \mathbf{x}_t - \mathbf{w}_{T_\alpha} \cdot \mathbf{x}_t + \epsilon \geq 0 \iff a_t = 1 \quad \text{and} \quad \mathbf{w}^* \cdot \mathbf{x}_t - \mathbf{w}_{T_\alpha} \cdot \mathbf{x}_t - \epsilon \leq 0 \iff v_t - p_t \leq \epsilon,$

which completes the lemma. □

A.2 Chernoff-style bound.

Lemma 7. *Let $S = \sum_{i=1}^n x_i$, where each $x_i \in \{0, 1\}$ is an independent random variable. Then the following inequality holds for any $0 < \epsilon < 1$.*

$$\Pr(S > (1 + \epsilon)\mathrm{E}[S]) \leq \frac{e^{\epsilon \mathrm{E}[S]}}{(1 + \epsilon)^{(1+\epsilon)\mathrm{E}[S]}} \leq \exp\Big(\frac{-\epsilon^2 \mathrm{E}[S]}{4}\Big)\,.$$

Proof. In what follows denote $\Pr(x_i = 1) = p_i$. To show the first inequality, we follow standard steps for arriving at a multiplicative Chernoff bound. For any $t > 0$ and using Markov's inequality, we have

$$\Pr(S > (1 + \epsilon)\mathrm{E}[S]) = \Pr(\exp(tS) > \exp(t(1 + \epsilon)\mathrm{E}[S])) \leq \frac{\mathrm{E}[\exp(tS)]}{\exp(t(1 + \epsilon)\mathrm{E}[S])}. \tag{2}$$

Now, noting that the random variables are independent, the numerator of this expression can be bounded as follows

$$\begin{aligned}\mathrm{E}[\exp(tS)] = \mathrm{E}\Big[\prod_{i=1}^n \exp(tx_i)\Big] = \prod_{i=1}^n \mathrm{E}[\exp(tx_i)] = \prod_{i=1}^n p_i e^t + (1 - p_i) = \prod_{i=1}^n p_i(e^t - 1) + 1 \\ \leq \prod_{i=1}^n \exp(p_i(e^t - 1)) = \exp\Big((e^t - 1)\sum_{i=1}^n p_i\Big) = \exp((e^t - 1)\mathrm{E}[S])\,,\end{aligned}$$

where the inequality uses the fact $1 + x \leq e^x$. Plugging this back into (2) and setting $t = \log(1 + \epsilon)$ results in

$$\Pr(S > (1 + \epsilon)\mathrm{E}[S]) \leq \frac{\exp((e^t - 1)\mathrm{E}[S])}{\exp(t(1 + \epsilon)\mathrm{E}[S])} = \frac{\exp((1 + \epsilon - 1)\mathrm{E}[S])}{(1 + \epsilon)^{(1+\epsilon)\mathrm{E}[S]}} = \frac{e^{\epsilon \mathrm{E}[S]}}{(1 + \epsilon)^{(1+\epsilon)\mathrm{E}[S]}}\,,$$

which proves the first inequality. To prove the second inequality, it suffices to show that

$$\begin{aligned}(1 + \epsilon)^{-(1+\epsilon)\mathrm{E}[S]} = \exp(-\log(1 + \epsilon)(1 + \epsilon)\mathrm{E}[S]) \leq \exp\Big(-\epsilon \mathrm{E}[S] - \frac{\epsilon^2 \mathrm{E}[S]}{4}\Big) \\ \iff \log(1 + \epsilon)(1 + \epsilon) \geq \epsilon + \frac{\epsilon^2}{4}\,.\end{aligned} \tag{3}$$

To prove this, note that for $f(\epsilon) = \log(1 + \epsilon)(1 + \epsilon) - \epsilon - \epsilon^2/4$, we have

$$\begin{aligned}f(0) = 0 \\ \forall \epsilon \in [0, 1], \;\; f'(\epsilon) = \log(1 + \epsilon) - \epsilon/2 \geq \epsilon - \epsilon^2/2 - \epsilon/2 > 0\,.\end{aligned}$$

Thus, the function f is zero at zero and increasing between values zero and one, implying it is positive between values zero and one and which proves the inequality in (3) and completes the lemma. □

A.3 Proof of Lemma 2

Before we present the proof of Lemma 2 we define a couple variables and also present an intermediate lemma. Define the variable

$$M_\rho = \sum_{t=1}^{T_\alpha} \mathbf{1}\{|v_t - p_t| < \rho\}, \tag{4}$$

as the number of times that the gap between the price offered and the buyer's value is less than ρ. For $\delta > 0$, let

$$\mathcal{E}_{\delta,\rho} = \left\{ M_\rho \le 2\rho T_\alpha + \sqrt{8\rho T_\alpha \log \frac{1}{\delta}} \right\}, \tag{5}$$

denote the event that there are not too many rounds on which this gap is smaller than ρ. We first prove the following lemma:

Lemma 8. *For any $\delta > 0$ and $0 < \rho < 1$ we have $P(\mathcal{E}_{\delta,\rho}) \ge 1 - \delta$.*

Proof. First notice that on lie rounds, the (undiscounted) surplus lost compared to the truthful buyer is

$$\underbrace{\mathbf{1}\{p_t \le v_t\}(v_t - p_t)}_{\text{truthful surplus}} - \underbrace{\mathbf{1}\{p_t > v_t\}(v_t - p_t)}_{\text{untruthful surplus}} = |v_t - p_t| \,.$$

Since each value $v_t \in [0,1]$ and price $p_t \in [0,1]$ is chosen i.i.d. during the first T_α rounds of the algorithm and furthermore p_t is chosen uniformly at random, we have that on any round $\Pr(|v_t - p_t| < \rho) \le 2\rho$. Using this, we note

$$\mathrm{E}[M_\rho] = \mathrm{E}\left[\sum_{t=1}^{T_\alpha} \mathbf{1}\{|v_t - p_t| < \rho\}\right] = \sum_{t=1}^{T_\alpha} \mathrm{E}[\mathbf{1}\{|v_t - p_t| < \rho\}] = \sum_{t=1}^{T_\alpha} \Pr(|v_t - p_t| < \rho) \le 2\rho T_\alpha \,.$$

Now, since M_ρ is a sum of T_α independent random variables taking values in $\{0,1\}$, Lemma 7 (in the appendix) implies

$$\Pr[M_\rho \ge (1+\epsilon)\mathrm{E}[M_\rho]] \le \exp\left(\frac{-\epsilon^2 \mathrm{E}[M_\rho]}{4}\right).$$

After setting the right hand side equal to δ and solving for ϵ, we have with probability at least $1 - \delta$,

$$M_\rho \le \mathrm{E}[M_\rho]\left(1 + \sqrt{\frac{4}{\mathrm{E}[M_\rho]} \log \frac{1}{\delta}}\right) = \mathrm{E}[M_\rho] + \sqrt{4\mathrm{E}[M_\rho] \log \frac{1}{\delta}} \le 2\rho T_\alpha + \sqrt{8\rho T_\alpha \log \frac{1}{\delta}} \,,$$

which completes the proof of the intermediate lemma. □

We can now give the proof of Lemma 2, which shows if we select

$$\rho^* = 1/(8T_\alpha \log(1/\delta)), \tag{6}$$

and the event $\mathcal{E}_{\delta,\rho^*}$ occurs, then at least $\frac{\gamma^{-L+3}-1}{8T_\alpha \log(\frac{1}{\delta})}\left(\frac{\gamma^{T_\alpha}}{1-\gamma}\right)$ surplus is lost compared to the truthful buyer.

Proof of Lemma 2. Let $M' = \left\lceil 2\rho T_\alpha + \sqrt{8\rho T_\alpha \log 1/\delta} \right\rceil$. Lemma 8 guarantees that with at least probability $1 - \delta$, M' is the maximum number of rounds where $|v_t - p_t| \le \rho$ occurs. Thus, on at least $L_\rho = L - M'$ of the lie rounds, at least ρ (undiscounted) surplus is lost compared to the truthful buyer. Let $\mathcal{L}_\rho$ denote the set of rounds where these events occur (so that $|\mathcal{L}_\rho| = L_\rho$), then since the discount sequence is decreasing the disounted surplus lost is at least

$$\sum_{t \in \mathcal{L}_\rho} \gamma_t |v_t - p_t| \ge \rho \sum_{t \in \mathcal{L}_\rho} \gamma_t \ge \rho \sum_{t=T_\alpha - L_\rho}^{T_\alpha} \gamma_t \,.$$

We can continue to lower bound this quantity:

$$\sum_{t=T_\alpha-L_\rho}^{T_\alpha} \gamma_t \geq \sum_{t=0}^{T_\alpha-1} \gamma^t - \sum_{t=0}^{T_\alpha-L_\rho-1} \gamma^t = \frac{1-\gamma^{T_\alpha}}{1-\gamma} - \frac{1-\gamma^{T_\alpha-L_\rho}}{1-\gamma} = (\gamma^{-L_\rho}-1)\frac{\gamma^{T_\alpha}}{1-\gamma}\,.$$

We also have that:

$$L_\rho \geq L - \lceil 2\rho T_\alpha + \sqrt{8\rho T_\alpha \log(1/\delta)}\rceil \geq L - 2\rho T_\alpha - \sqrt{8\rho T_\alpha \log(1/\delta)} - 1$$

where the first inequality follows from the definition of L_ρ, the second from the fact that $\lceil n\rceil \leq n+1$. Therefore, defining $L'_\rho = L - 2\rho T_\alpha - \sqrt{8\rho T_\alpha \log(1/\delta)} - 1$, gives us that for any $0 < \rho < 1/2$:

$$\sum_{t=T_\alpha-L_\rho}^{T_\alpha} \gamma_t \geq (\gamma^{-L'_\rho}-1)\frac{\gamma^{T_\alpha}}{1-\gamma}\,.$$

Selecting $\rho = 1/(8T_\alpha \log(1/\delta))$ gives us:

$$\rho\Big(\gamma^{-L'_\rho}-1\Big)\frac{\gamma^{T_\alpha}}{1-\gamma} \geq (8\log(1/\delta))^{-1}\frac{1}{T_\alpha}\left(\gamma^{-L+3}-1\right)\frac{\gamma^{T_\alpha}}{1-\gamma}\,,$$

which completes the lemma. □

A.4 Proof of Lemma 3

Proof. Let S_1 and S_2 be the excess surplus that a surplus-maximizing buyer earns over the truthful buyer during the learning and exploit phase of the LEAP algorithm, respectively. We have

$$S_2 \leq \sum_{t=T_\alpha+1}^{T} \gamma^{t-1} = \gamma^{T_\alpha} \sum_{t=0}^{T-T_\alpha-1} \gamma^t = \frac{\gamma^{T_\alpha}}{1-\gamma}(1-\gamma^{T-T_\alpha})\,. \tag{7}$$

Indeed, this an upper bound on the total surplus any buyer can hope to achieve in the second phase. Now observe that for any constants $C > 0$, $\delta_0 > 0$ and ρ^* as defined in equation (6), we have

$$\begin{aligned}
E[S_1] &= \Pr[\mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C]E[S_1 \mid \mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C] + \Pr[\neg\mathcal{E}_{\delta_0,\rho^*} \vee L < C]E[S_1 \mid \neg\mathcal{E}_{\delta_0,\rho^*} \vee L < C] \\
&\leq \Pr[\mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C]E[S_1 \mid \mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C] \\
&= \Pr[\mathcal{E}_{\delta_0,\rho^*}]\Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}]E[S_1 \mid \mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C] \\
&\leq -(1-\delta_0)\Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}]\frac{\gamma^{-C+3}-1}{8T_\alpha\log(1/\delta_0)}\left(\frac{\gamma^{T_\alpha}}{1-\gamma}\right)
\end{aligned}$$

The steps follow respectively by the law of iterated expectation; because $S_1 \leq 0$ with probability 1, since the truthful buyer strategy gives maximal revenue during the non-adaptive first phase; definition of conditional probability; and finally, applying Lemma 8 to lower bound $\Pr[\mathcal{E}_{\delta_0,\rho^*}]$ and the second half of the proof of Lemma 2 (shown in Section A.3) to upper bound $E[S_1 \mid \mathcal{E}_{\delta_0,\rho^*} \wedge L \geq C]$ (which is a negative quantity).

Note, since we are assuming a surplus maximizing buyer, it must be the case that $0 \leq E[S_1 + S_2]$. Thus, using the upper bound on S_2 and the upper bound on $E[S_1]$, we can rewrite the fact $0 \leq E[S_1 + S_2]$ as:

$$\begin{aligned}
&Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}](1-\delta_0)\frac{\gamma^{-C+3}-1}{8T_\alpha\log(1/\delta_0)}\left(\frac{\gamma^{T_\alpha}}{1-\gamma}\right) \leq \frac{\gamma^{T_\alpha}}{1-\gamma}(1-\gamma^{T-T_\alpha}) \\
\iff\ &Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}] \leq 8T_\alpha\log(1/\delta_0)(1-\gamma^{T-T_\alpha})/((1-\delta_0)(\gamma^{-C+3}-1))
\end{aligned}$$

Therefore, when

$$C = \frac{\log\left(\frac{\left(1-\gamma^{T-T_\alpha}\right)8T_\alpha\log(1/\delta_0)}{\delta_0(1-\delta_0)}+1\right)}{\log(1/\gamma)} - 3 \qquad \text{we have} \qquad \Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}] \leq \delta_0\,.$$

Fixing this choice of C, lets us conclude:

$$\begin{aligned}\Pr[L \geq C] &= \Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}] \Pr[\mathcal{E}_{\delta_0,\rho^*}] + \Pr[L \geq C \mid \neg\mathcal{E}_{\delta_0,\rho^*}] \Pr[\neg\mathcal{E}_{\delta_0,\rho^*}] \\ &\leq \Pr[L \geq C \mid \mathcal{E}_{\delta_0,\rho^*}] + \Pr[\neg\mathcal{E}_{\delta_0,\rho^*}] \leq \delta_0 + \delta_0\end{aligned}$$

Thus, setting $\delta_0 = \delta/2$ tells us that $\Pr[L < C] \geq 1 - \delta$. Finally, to complete the lemma, we upper bound C by dropping the terms $(1-\gamma^{T-T_\alpha})$ and -3, and using $1/(\delta_0(1-\delta_0)) = 2/(\delta(1-\delta/2)) \leq 4/\delta$. □

A.5 Results from Rakhlin et al. [22]

Let $Z_t = (\nabla F(\mathbf{w}_t) - \mathbf{g}_t)^\top (\mathbf{w}_t - \mathbf{w}^*)$ and

$$Z(T) = \frac{2}{\lambda} \sum_{t=2}^{T} \frac{Z_t}{t} \prod_{t'=t+1}^{T} \left(1 - \frac{2}{t'}\right). \tag{8}$$

Rakhlin et al. [22] proved the following upper bound on $Z(T)$ in the last half of the proof of their Proposition 1. For convenience, we isolate it into a separate lemma.

Lemma 9. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_T$ be any sequence of weight vectors. If $E[\mathbf{g}_t] = \nabla F(\mathbf{w}_t)$ and $\|\mathbf{g}_t\|^2 \leq G^2$ then for any $\delta < 1/e$ and $T \geq 2$*

$$Z(T) \leq \frac{16G\sqrt{\log(\log(T)/\delta)}}{\lambda(T-1)T} \sqrt{\sum_{t=2}^{T}(t-1)^2\|\mathbf{w}_t - \mathbf{w}^*\|^2} + \frac{16G^2\log(\log(T)/\delta)}{\lambda^2 T}.$$

Importantly, for the previous lemma to hold it is *not* necessary for the $\mathbf{w}_t$'s to have been generated by stochastic gradient descent. The same remark applies to the next lemma, which gives a recursive upper bound on $\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2$, and which was also proven by Rakhlin et al. [22] in the last half of the proof of their Proposition 1.

Lemma 10. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$ be any sequence of weight vectors. Suppose the following three conditions hold:*

1. $\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq \frac{a}{t}$ *for* $t \in \{1, 2\}$,
2. $\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \leq \frac{b}{(t-1)t}\sqrt{\sum_{i=2}^{t}(i-1)^2\|\mathbf{w}_i - \mathbf{w}^*\|^2} + \frac{c}{t}$ *for* $t \in \{2, \ldots, T\}$, *and*
3. $a \geq \frac{9b^2}{4} + 3c$.

Then $\|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2 \leq \frac{a}{(T+1)}$.

A.6 Proof of Lemma 4

Proof. Recall that F is λ-strongly convex. A well-known property of λ-strongly convex functions is that

$$\nabla F(\mathbf{w}')^\top(\mathbf{w}' - \mathbf{w}'') \geq F(\mathbf{w}') - F(\mathbf{w}'') + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}''\|^2 \tag{9}$$

for any weight vectors $\mathbf{w}', \mathbf{w}''$ (for example, see [22]). Letting $\mathbf{w}' = \mathbf{w}^*$ and $\mathbf{w}'' = \mathbf{w}$ in Eq. (9) we have

$$\begin{aligned}0 = \nabla F(\mathbf{w}^*)^\top(\mathbf{w}^* - \mathbf{w}) &\geq F(\mathbf{w}^*) - F(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}^* - \mathbf{w}\|^2 \\ \Rightarrow F(\mathbf{w}) - F(\mathbf{w}^*) &\geq \frac{\lambda}{2}\|\mathbf{w}^* - \mathbf{w}\|^2 \end{aligned} \tag{10}$$

where we used the fact that $\mathbf{w}^*$ minimizes F, and thus $\nabla F(\mathbf{w}^*) = \mathbf{0}$. Now letting $\mathbf{w}' = \mathbf{w}$ and $\mathbf{w}'' = \mathbf{w}^*$ in Eq. (9) and applying Eq. (10) proves

$$\nabla F(\mathbf{w})^\top(\mathbf{w} - \mathbf{w}^*) \geq \lambda\|\mathbf{w} - \mathbf{w}^*\|^2. \tag{11}$$

Note that $\tilde{\mathbf{g}}_t = \mathbf{g}_t \pm \mathbf{1}\{t \in \mathcal{L}\}\mathbf{x}_t$, where the $\pm$ depends on the value of a_t. Let $Z_t = (\nabla F(\mathbf{w}_t) - \mathbf{g}_t)^\top (\mathbf{w}_t - \mathbf{w}^*)$. We have

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &= \|\mathbf{w}_t - \eta_t \tilde{\mathbf{g}}_t - \mathbf{w}^*\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \tilde{\mathbf{g}}_t^\top (\mathbf{w}_t - \mathbf{w}^*) + \eta_t^2 \|\tilde{\mathbf{g}}_t\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}^*) \pm 2\eta_t \mathbf{1}\{t \in \mathcal{L}\}\mathbf{x}_t^\top (\mathbf{w}_t - \mathbf{w}^*) + \eta_t^2 \|\tilde{\mathbf{g}}_t\|^2 \\
&\le \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \mathbf{g}_t^\top (\mathbf{w}_t - \mathbf{w}^*) + 4\eta_t \mathbf{1}\{t \in \mathcal{L}\} + \eta_t^2 G^2 \qquad (12) \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \nabla F(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}^*) + 2\eta_t Z_t + 4\eta_t \mathbf{1}\{t \in \mathcal{L}\} + \eta_t^2 G^2 \\
&\le \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \lambda \|\mathbf{w}_t - \mathbf{w}^*\|^2 + 2\eta_t Z_t + 4\eta_t \mathbf{1}\{t \in \mathcal{L}\} + \eta_t^2 G^2 \qquad (13) \\
&= (1 - 2\lambda\eta_t) \|\mathbf{w}_t - \mathbf{w}^*\|^2 + 2\eta_t Z_t + 4\eta_t \mathbf{1}\{t \in \mathcal{L}\} + \eta_t^2 G^2
\end{aligned}$$

where in Eq. (12) we used $\mathbf{x}_t^\top (\mathbf{w}_t - \mathbf{w}^*) \le \|\mathbf{x}_t\| \|\mathbf{w}_t - \mathbf{w}^*\| \le 2$ and $\|\tilde{\mathbf{g}}_t\|^2 \le G^2$. In Eq. (13) we used Eq. (11). For any $T' \in \{2, \dots, T_\alpha\}$ let $Y_t(T') = \prod_{t'=t+1}^{T'} (1 - 2\lambda\eta_{t'})$. Unrolling the above recurrence till $t = 2$ yields

$$\|\mathbf{w}_{T'+1} - \mathbf{w}^*\|^2 \le Y_1(T') \|\mathbf{w}_2 - \mathbf{w}^*\|^2 + 2\sum_{t=2}^{T'} \eta_t Z_t Y_t(T') + 4\sum_{t=2}^{T'} \eta_t \mathbf{1}\{t \in \mathcal{L}\} Y_t(T') + G^2 \sum_{t=2}^{T'} \eta_t^2 Y_t(T').$$

Now substitute $\eta_t = \frac{1}{\lambda t}$, and note that since $(1 - 2\lambda\eta_2) = 0$ and $T' \ge 2$ we have $Y_1(T') = 0$, so the first term is zero. Also the second term is equal to $Z(T')$ by the definition in Eq. (8) in Appendix A.5. Simplifying leads to

$$\|\mathbf{w}_{T'+1} - \mathbf{w}^*\|^2 \le Z(T') + \frac{4}{\lambda} \sum_{t=2}^{T'} \mathbf{1}\{t \in \mathcal{L}\} \frac{Y_t(T')}{t} + \frac{G^2}{\lambda^2} \sum_{t=2}^{T'} \frac{Y_t(T')}{t^2}. \qquad (14)$$

Now observe that for $t \ge 2$

$$\log Y_t(T') = \sum_{t'=t+1}^{T'} \log\left(1 - \frac{2}{t'}\right) \le -2 \sum_{t'=t+1}^{T'} \frac{1}{t'} = -2\left(\sum_{t'=1}^{T'} \frac{1}{t'} - \sum_{t'=1}^{t} \frac{1}{t'}\right) \le -2(\log T' - \log t - 1),$$

where the last inequality uses a lower bound on the t-th harmonic number and upper bound on the T'-th harmonic number. Thus, $Y_t(T') \le \frac{e^2 t^2}{T'^2}$ and plugging back into Eq. (14) yields

$$\|\mathbf{w}_{T'+1} - \mathbf{w}^*\|^2 \le Z(T') + \frac{4e^2}{\lambda T'^2} \sum_{t=2}^{T'} \mathbf{1}\{t \in \mathcal{L}\} t + \frac{e^2 G^2}{\lambda^2 T'} \le Z(T') + \frac{4e^2 L}{\lambda T'} + \frac{e^2 G^2}{\lambda^2 T'}.$$

where the second inequality follows from $\sum_{t=2}^{T'} \mathbf{1}\{t \in \mathcal{L}\} t \le LT'$. Now, to bound the term $Z(T')$, we apply Lemma 9 from Appendix A.5 and conclude that for $\delta \in [0, 1/e]$, with probability at least $1 - \delta$, for all $T' \in \{2, \dots, T_\alpha\}$

$$Z(T') \le \frac{16G\sqrt{\log(\log(T')/\delta)}}{\lambda(T'-1)T'} \sqrt{\sum_{t=2}^{T'} (t-1)^2 \|\mathbf{w}_t - \mathbf{w}^*\|^2} + \frac{16G^2 \log(\log(T')/\delta)}{\lambda^2 T'}.$$

Plugging this back in and simplifying we get, with probability at least $1-\delta$, for all $T' \in \{2, \dots, T_\alpha\}$

$$\|\mathbf{w}_{T'+1} - \mathbf{w}^*\|^2 \le$$

$$\frac{16G\sqrt{\log(\log(T')/\delta)}}{\lambda(T'-1)T'} \sqrt{\sum_{t=2}^{T'} (t-1)^2 \|\mathbf{w}_t - \mathbf{w}^*\|^2} + \frac{1}{T'}\left(\frac{(16\log(\log(T')/\delta) + e^2)G^2}{\lambda^2} + \frac{4e^2 L}{\lambda}\right).$$

In order to apply Lemma 10 in Appendix A.5 let

$$\begin{aligned}
a &= \frac{(624\log(\log(T_\alpha)/\delta) + e^2)G^2}{\lambda^2} + \frac{4e^2 L}{\lambda}, \\
b &= \frac{16G\sqrt{\log(\log(T')/\delta)}}{\lambda}, \text{ and} \\
c &= \frac{(16\log(\log(T')/\delta) + e^2)G^2}{\lambda^2} + \frac{4e^2 L}{\lambda}.
\end{aligned}$$

It is a straightforward calculation to show that $a \geq \frac{9b^2}{4} + 3c$. Also for any T'

$$G \left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\| \geq \left\| \nabla F(\mathbf{w}_{T'}) \right\| \left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\| \geq \nabla F(\mathbf{w}_{T'})^\top (\mathbf{w}_{T'} - \mathbf{w}^*) \geq \lambda \left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\|^2$$

where the last inequality follows from Eq. (11). Dividing both sides by $\lambda \left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\|$ proves $\left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\| \leq \frac{G}{\lambda}$ for all T', which implies $\left\| \mathbf{w}_{T'} - \mathbf{w}^* \right\|^2 \leq a/T'$ for $T' \in \{1, 2\}$. Now we can apply Lemma 10 in Appendix A.5 to show

$$\left\| \mathbf{w}_{T_\alpha + 1} - \mathbf{w}^* \right\|^2 \leq \frac{1}{T_\alpha + 1} \Big(\frac{(624 \log(\log(T_\alpha)/\delta) + e^2) G^2}{\lambda^2} + \frac{4e^2 L}{\lambda} \Big),$$

which completes the proof. □

A.7 Kernelized LEAP algorithm

For what follows, we define the projection operation

$$\Pi_K \big(\boldsymbol{\beta}, (\mathbf{x}_1, \ldots, \mathbf{x}_t) \big) = \frac{\boldsymbol{\beta}}{\sqrt{\sum_{i,j=1}^t \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j)}}.$$

The kernelized LEAP algorithm is given below.

Algorithm 2 Kernelized LEAP algorithm

- Let $K(\cdot, \cdot)$ be a PDS function s.t. $\forall \mathbf{x} : |K(\mathbf{x}, \mathbf{x})| \leq 1$, $0 \leq \alpha \leq 1$, $T_\alpha = \lceil \alpha T \rceil$, $\boldsymbol{\beta} = \mathbf{0} \in \mathbb{R}^{T_\alpha}$, $\epsilon \geq 0$, $\lambda > 0$.
- For $t = 1, \ldots, T_\alpha$
  - Offer $p_t \sim U$
  - Observe a_t
  - $\beta_t = -\frac{2}{\lambda t} \big(\sum_{i=1}^{t-1} \beta_i K(\mathbf{x}_i, \mathbf{x}_t) - a_t \big)$
  - $\boldsymbol{\beta} = \Pi_K \big(\boldsymbol{\beta}, (\mathbf{x}_1, \ldots, \mathbf{x}_t) \big)$
- For $t = T_\alpha + 1, \ldots, T$
  - Offer $p_t = \sum_{i=1}^{T_\alpha} \beta_i K(\mathbf{x}_i, \mathbf{x}_t) - \epsilon$

Proposition 2. *Algorithm 2 is a kernelized implementation of the LEAP algorithm with* $\mathcal{W} = \{\mathbf{w} \colon \|\mathbf{w}\|_2 \leq 1\}$ *and* $\mathbf{w}_1 = \mathbf{0}$. *Furthermore, if we consider the feature space induced by the kernel* K *via an explicit mapping* $\phi(\cdot)$, *the learned linear hypothesis is represented as* $\mathbf{w}_t = \sum_{i=1}^{t-1} \beta_i \phi(\mathbf{x}_i)$ *which satisfies* $\|\mathbf{w}_t\| = \sum_{i,j=1}^{t-1} \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) \leq 1$. *The gradient is* $\mathbf{g}_t = 2 \Big(\sum_{i=1}^{t-1} \beta_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t) - a_t \Big) \phi(\mathbf{x}_t)$, *and* $\|\mathbf{g}_t\| \leq 4$.

Proof. We will use an inductive argument. Note that, before the projection step $\beta_1 = 2a_1/\lambda$ and after projection $\beta_1 = a_1 / \sqrt{K(\mathbf{x}_1, \mathbf{x}_1)}$. Thus, $\mathbf{w}_1 = \mathbf{0}$ and $\mathbf{w}_2 = \beta_1 \phi(\mathbf{x}_1) = \frac{a_1}{\sqrt{K(\mathbf{x}_1, \mathbf{x}_1)}} \phi(\mathbf{x}_1)$ match the hypotheses returned by the LEAP algorithm when operating in the feature space induced by $\phi(\cdot)$ and using the projection $\Pi_{\mathcal{W}}$ for $\mathcal{W} = \{\mathbf{w} \colon \|\mathbf{w}\|_2 \leq 1\}$. Now, assuming the inductive hypothesis, we have $\mathbf{w}_t = \sum_{i=1}^{t-1} \beta_i \phi(\mathbf{x}_i)$ and we have, before projection,

$$\sum_{i=1}^{t} \beta_i \phi(\mathbf{x}_i) = \mathbf{w}_t + \beta_t = \mathbf{w}_t - \frac{2}{\lambda t} \big(\sum_{i=1}^{t-1} \beta_i K(\mathbf{x}_i, \mathbf{x}_t) - a_t \big) \phi(\mathbf{x}_t) = \mathbf{w}_t - \frac{2}{\lambda t} (\mathbf{w}_t^\top \phi(\mathbf{x}_t) - a_t) \phi(\mathbf{x}_t)$$

and, after projection,

$$\frac{\sum_{i=1}^t \beta_i \phi(\mathbf{x}_i)}{\sqrt{\sum_{i,j=1}^t \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j)}} = \frac{\sum_{i=1}^t \beta_i \phi(\mathbf{x}_i)}{\| \sum_{i=1}^t \beta_i \phi(\mathbf{x}_i) \|} = \frac{\mathbf{w}_t - \frac{2}{\lambda t} (\mathbf{w}_t^\top \phi(\mathbf{x}_t) - a_t) \phi(\mathbf{x}_t)}{\| \mathbf{w}_t - \frac{2}{\lambda t} (\mathbf{w}_t^\top \phi(\mathbf{x}_t) - a_t) \phi(\mathbf{x}_t) \|}$$

$$= \Pi_{\mathcal{W}} \Big(\mathbf{w}_t - \frac{2}{\lambda t} (\mathbf{w}_t^\top \phi(\mathbf{x}_t) - a_t) \phi(\mathbf{x}_t) \Big) = \mathbf{w}_{t+1}$$

which proves the equivalence of the first phase of the two algorithms in the feature space induced by $\phi(\cdot)$. Note, in the second phase neither $\boldsymbol{\beta}$ or $\mathbf{w}_{T_\alpha+1}$ is updated, and from the preceding argument we have

$$p_t = \sum_{i=1}^{T_\alpha} \beta_i K(\mathbf{x}_i, \mathbf{x}_t) - \epsilon = \Big(\sum_{i=1}^{T_\alpha} \beta_i \phi(\mathbf{x}_i) \Big) \phi(\mathbf{x}_t) - \epsilon = \mathbf{w}_{T_\alpha+1}^\top \phi(\mathbf{x}_t) - \epsilon \,,$$

which shows the equivalence of the two algorithms in the second phase as well.

The bound $\|\mathbf{w}_t\| \leq 1$ follows directly from the definition of the projection Π_K. Using $\mathbf{w}_t = \sum_{i=1}^{t-1} \beta_i \phi(\mathbf{x}_i)$, we have that the gradient is

$$\mathbf{g}_t = 2(\mathbf{w}_t^\top \phi(\mathbf{x}_t) - a_t)\phi(\mathbf{x}_t) = 2\Big(\sum_{i=1}^{t} \beta_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t) - a_t \Big) \phi(\mathbf{x}_t) \,.$$

Finally, we can bound $\|\mathbf{g}_t\| \leq 2(|\mathbf{w}_t^\top \phi(\mathbf{x}_t)| + 1)\|\phi(\mathbf{x}_t)\| \leq 2(\|\mathbf{w}_t\|\|\phi(\mathbf{x}_t)\| + 1) \leq 4$, which follows from $\|\mathbf{w}_t\| \leq 1$ and $\|\phi(\mathbf{x}_t)\| = \sqrt{K(\mathbf{x}_t, \mathbf{x}_t)} \leq 1$. □

A.8 Doubling trick

Corollary 2. *Partition all T rounds into $\lceil \log_2 T \rceil$ consecutive phases, where each phase i has length $T_i = 2^i$. Run an independent instance of the LEAP algorithm in each phase, tuning ϵ and α as in Theorem 2, using horizon length T_i. Against a surplus-maximizing buyer, the seller's regret against a surplus-maximizing buyer is $R(T) \leq O\big(T^{2/3}\sqrt{\frac{\log(T)}{\log(1/\gamma)}}\big)$.*

Proof. Since an independent instance of the algorithm is run in each phase, the buyer will behave so as to maximize surplus in each phase independently, without regard to what occurs in other phases. Moreover, the discount factor for the sth round in any phase i is $\gamma^{t_i+s} = \gamma^{t_i}\gamma^s$, where t_i is the first round of phase i. It is easy to see that the behavior of a surplus-maximizing buyer is unchanged if we scale her surplus in every round by a constant. Therefore the analysis of Theorem 2 is directly applicable to every phase, and we can combine the analysis for all phases using the doubling trick, as follows.

Let R_i be the seller's strategic regret in phase i and $n = \lceil \log_2 T \rceil$. By Theorem 2 there exists a constant C depending only on λ such that

$$R(T) = \sum_{i=1}^{\lceil \log_2 T \rceil} R_i \leq \frac{C}{\sqrt{\log(1/\gamma)}} \sum_{i=1}^{\lceil \log_2 T \rceil} T_i^{2/3} \sqrt{\log_2 T_i} = \frac{C}{\sqrt{\log(1/\gamma)}} \sum_{i=1}^{\lceil \log_2 T \rceil} \Big(2^{2/3}\Big)^i \sqrt{i} \quad (15)$$

Let $S_{r,n} = \sum_{i=1}^{n} r^i \sqrt{i}$. Observe that

$$S_{r,n+1} = \sum_{i=1}^{n+1} r^i \sqrt{i} = r^{n+1}\sqrt{n+1} + \sum_{i=1}^{n} r^i \sqrt{i} = r^{n+1}\sqrt{n+1} + S_{r,n}$$

and

$$S_{r,n+1} = r \sum_{i=1}^{n+1} r^{i-1}\sqrt{i} \geq r \sum_{i=1}^{n+1} r^{i-1}\sqrt{i-1} = r \sum_{i=2}^{n+1} r^{i-1}\sqrt{i-1} = r \sum_{i=1}^{n} r^i \sqrt{i} = r S_{r,n}$$

Combining the previous two inequalities proves $r^{n+1}\sqrt{n+1} + S_{r,n} \geq r S_{r,n}$, which can be rearranged to show

$$\sum_{i=1}^{n} r^i \sqrt{i} \leq \frac{r^{n+1}\sqrt{n+1}}{r-1}.$$

Applying the above inequality for $n = \lceil \log_2 T \rceil$ and $r = 2^{2/3}$ proves

$$\begin{aligned}\sum_{i=1}^{\lceil \log_2 T \rceil} \left(2^{2/3}\right)^i \sqrt{i} &\leq \frac{(2^{2/3})^{\lceil \log_2 T \rceil + 1} \sqrt{\lceil \log_2 T \rceil + 1}}{2^{2/3} - 1} \\ &\leq \frac{(2^{2/3})^{\log_2 T + 2} \sqrt{\log_2 T + 2}}{2^{2/3} - 1} \\ &= \frac{2^{4/3}}{2^{2/3} - 1} T^{2/3} \sqrt{\log_2 T + 2}.\end{aligned}$$

Combining the above with Eq (15) proves the corollary. □