Learning against Non-credible Second-Price Auctions

Anonymous Author(s)*

Abstract

The standard framework of online bidding algorithm design assumes that the seller commits himself to faithfully implementing the rules of the adopted auction. However, the seller may attempt to cheat in execution to increase his revenue if the auction belongs to the class of non-credible auctions. For example, in a second-price auction, the seller could create a fake bid between the highest bid and the second highest bid. This paper focuses on one such case of online bidding in repeated second-price auctions. At each time t, the winner with bid b_t is charged not the highest competing bid d_t but a manipulated price $p_t = \alpha_0 d_t + (1 - \alpha_0) b_t$, where the parameter $\alpha_0 \in [0, 1]$ in essence measures the seller's credibility. Unlike classic repeated-auction settings where the bidder has access to samples $(d_s)_{s=1}^{t-1}$, she can only receive mixed signals of $(b_s)_{s=1}^{t-1}$, $(d_s)_{s=1}^{t-1}$ and α_0 in this problem. The task for the bidder is to learn not only the bid distributions of her competitors but also the seller's credibility. We establish regret lower bounds in various information models and provide corresponding online bidding algorithms that can achieve near-optimal performance. Specifically, we consider three cases of prior information based on whether the credibility α_0 and the distribution of the highest competing bids are known. Our goal is to characterize the landscape of online bidding in non-credible second-price auctions and understand the impact of the seller's credibility on online bidding algorithm design under different information structures.

Keywords

Non-credible Auctions, No-regret Learning, Bidding Algorithms

ACM Reference Format:
Anonymous Author(s). 2018. Learning against Non-credible Second-Price Auctions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Digital advertising has experienced significant expansion due to the rapid rise of online-activities, surpassing traditional advertising as the dominant marketing influence in various industries. Between 2021 and 2022, digital advertising revenues in U.S. grew 10.8% year-over-year totalling $209.7 billion dollars [19]. In practice, a huge amount of online ads are sold via real-time auctions implemented on advertising platforms and advertisers participate in such repeated online auctions to purchase advertising opportunities. This has motivated a flourishing line of work to focus on the problem of online bidding algorithm design. In particular, learning to bid in repeated second-price auctions—often with constraints or unknown own valuations—has been well studied due to the popularity of this auction format in practice [6, 7, 10, 11, 14, 15, 28]. However, these studies are in fact based on an implicit assumption that the seller commits himself to faithfully implementing the rules of the announced second-price auction.

The possibility that a seller can profitably cheat in a second-price auction was pointed out as early as the seminal paper [26] that introduced this auction format. After observing all the bids, the seller can strategically exaggerate the highest competing bid and overcharge the winner up to the amount of her own bid. Several following papers studied the issue of seller cheating in second-price auctions [22–24]. Recent theoretical work by [1] formally modelled *credibility* in an extensive-form game where the seller is allowed to deviate from the auction rules as long as the deviation cannot be detected by bidders. They defined an auction to be *credible* if it is incentive-compatible for the seller to follow the rules in the presence of cheating opportunities. In a second-price auction, the seller can even charge the winner the amount of her own bid to obtain higher utility, with an innocent explanation that the highest and second-highest bids are identical. Therefore, the prevalent second-price auction belongs to the class of non-credible auctions in this framework. Taking credibility into consideration, advertisers are confronted with a question of practical importance: how should an advertiser bid in repeated non-credible second-price auctions to maximize her cumulative utility?

In this work, we formulate the above problem as an online learning problem for a single bidder. We consider the scenario with a single seller who runs repeated non-credible second-price auctions. At each time t, the winner with bid b_t is charged not the highest competing bid d_t but a manipulated price $p_t = \alpha_0 d_t + (1 - \alpha_0) b_t$ for some $\alpha_0 \in [0, 1]$. The parameter α_0 in essence captures the seller's credibility, the extent to which the seller deviates from the second-price auction rules. Our linear model is equivalent to the classic bid-shilling model [22–24] in expectation. In the bid-shilling model, the seller cheats by inserting a shill bid after observing all of the bids with probability P^c. The seller is assumed to take full advantage of his power so the winner will pay her own bid if the seller does cheat. Then the winner's expected payment is $P^c b_t + (1 - P^c) d_t$. Moreover, no matter what charging rules the seller actually uses, as long as the estimated credibility within this linear model is away from 1, it can be confirmed that the seller is cheating. We believe our results and techniques have implications for the more complicated setting with $p_t = h(b_t, d_t; \alpha)$.

We assume the highest competing bids $(d_t)_{t=1}^T$ are *i.i.d.* sampled from a distribution G. The bidder aims to maximize her expected cumulative utility, which is given by the expected difference between the total value and the total payment. Moving to the information model, We investigate three cases of prior information: (1) known α_0 and unknown G; (2) unknown α_0 and known G; (3) unknown

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

α_0 and unknown G. For all three cases, we consider bandit feedback where the bidder can observe the realized allocation and cost at each round. For the last case where neither is unknown, we additionally consider full feedback where the price p_t is always observable regardless of the auction outcome. More discussions on modeling will be placed in Section 2.

The key challenge of this problem lies in the lack of credibility and its impact on the learning process. If assuming the seller has full commitment, it is well known that truthful bidding is the dominant strategy in second-price auctions, but this truthful property no longer holds in non-credible second-price auctions. Identifying optimal bidding strategies for utility maximization requires not only knowing the bidder's own values but also considering the strategies of her competitors and the seller. In classic repeated-auction settings (assuming a trustworthy seller), online bidding algorithms can collect historical samples $(d_s)_{s=1}^{t-1}$ to estimate distribution G. However, the bidder in non-credible auctions needs to cope with an additional dimension of uncertainty: the available observations under either bandit or full feedback are all manipulated prices, i.e., mixed signals of $(b_s)_{s=1}^{t-1}, (d_s)_{s=1}^{t-1}$ and α_0. As a result, difficulties arise in the estimation of the distribution of her competitors' bids and the seller's credibility.

1.1 Main Contributions

First, we characterize the optimal clairvoyant bidding strategy in non-credible second-price auctions when the bidder knows both credibility α_0 and distribution G. This optimal clairvoyant bidding strategy is also used as the benchmark strategy in the regret definition.

Next, we establish regret lower bounds in various information models and provide corresponding online bidding algorithms that are optimal up to log factors. Our results are summarized in Table 1.

- For the case where G is unknown and α_0 is known, we explore the landscape by discussing how the problem varies with different credibility parameter $\alpha_0 = 0, \alpha_0 = 1$ and $\alpha_0 \in (0, 1)$. We mainly contribute to the regret analysis for $\alpha_0 \in (0, 1)$, with a proven $\Omega(\sqrt{T})$ lower bound, and a concrete near-optimal $\widetilde{O}(\sqrt{T})$ algorithm.
- For the case where G is known and α_0 is unknown, we develop an $O(\log^2 T)$ algorithm, which adopts a dynamic estimation approach to approximate α_0.
- For the challenging case where both G and α_0 are unknown, we observe that under bandit feedback, an $\Omega(T^{2/3})$ lower bound and an $\widetilde{O}(T^{2/3})$ algorithm follow directly from existing algorithms. We then turn to the more interesting setting with full information feedback, for which we propose an episodic bidding algorithm that learns α_0 and G simultaneously in an efficient manner, while achieving a near-optimal regret of $\widetilde{O}(\sqrt{T})$.

Overall, this work provides a theoretical regret analysis for learning against non-credible auctions. We aim to characterize the landscape of online bidding in non-credible auctions and analyze how the seller' credibility influences the design of online bidding algorithms under different information structures.

1.2 Related Work

Credibility in auctions. The issue of seller cheating has been studied by the game-theoretic literature in a strategic framework [22–24]. Recently, [1] explored the setting where the seller deviates from the auction rules in a way that can be innocently explained. To this end, they defined credibility based on the detectability of seller deviations. They further established an impossibility result that no optimal auction simultaneously achieves staticity, credibility and strategy-proofness. They showed that the first-price auction is the unique static optimal auction that achieves credibility. [16] considered the general allocation problem and introduced the definition of *verifiability* (i.e., allowing participants to check the correctness of their assignments) and *transparency* (i.e., allowing participants to check whether the allocation rule is deviated). These are stronger security notions than the credibility concept investigated by [1]. [12] studied how the credibility of an auction format affects bidding behavior and final outcomes via laboratory experiments. Their empirical findings confirm the theory that sellers do have incentives to break the auction rules and overcharge the winning bidder. These pioneering works discussed how a participant can potentially detect and learn non-credible mechanisms as we do. In contrast, our work is based on the online learning framework, where information revelation is partial and sequential in nature.

Learning to bid. Our work is closely related with the line of literature on learning to bid in repeated auctions. [3, 4, 17, 18, 29] studied the problem of learning in repeated first-price auctions. [8, 27] studied no-regret learning in repeated first-price auctions with budget constraints. As for repeated second-price auctions, [6, 7, 11] considered the bidding problem with budget constraints and [14, 15] further considered return-on-spend (RoS) constraints. All these works assume that the seller has full commitment to the announced auction rules.

2 Problem Formulation

We consider the problem of online learning in repeated non-credible second-price auctions. We focus on a single bidder in a large population of bidders during a time horizon T. In each round $t = 1, \ldots, T$, there is an available item auctioned by a single seller. The bidder perceives a value $v_t \in [0, 1]$ for this item, and then submits a bid $b_t \in \mathbb{R}_+$ based on v_t and all historical observations available to her. We denote the maximum bid of all other bidders by $d_t \in \mathbb{R}_+$. As usual, we use the bold symbol $\boldsymbol{v}$ without subscript t to denote the vector $(v_1, \ldots, v_T)$; the same goes for other variables in the present paper.

We consider a *stochastic* setting where v_t is *i.i.d.* sampled from a distribution F and d_t is *i.i.d.* sampled from a distribution G. The latter assumption follows from the standard mean-field approximation [5, 20] and is a common practice in literature. The main rationale behind this assumption is that when the number of other bidders is large, on average their valuations and bidding strategies are static over time. Whether G is known to the bidder depends on the information structure while F is always unknown to the bidder.

The seller claims that he follows the rules of the second-price auction, but he actually uses a combination of the first-price auction and the second-price auction. For simplicity, we assume a linear model. The auction outcome in round t is then as follows: if $b_t \geq d_t$,

Table 1: Result Summary.

α_0	G	Feedback	Upper bound	Lower bound	Theorem
Known, $\alpha_0 = 1$	Unknown	Bandit	0	0	Theorem 3.1
Known, $\alpha_0 \in (0, 1)$			$\widetilde{O}(T^{1/2})$	$\Omega(T^{1/2})$	
Known, $\alpha_0 = 0$			$\widetilde{O}(T^{2/3})$	$\Omega(T^{2/3})$	
Unknown	Known	Bandit	$\widetilde{O}(1)$	$\Omega(1)$	Theorem 4.3[1]
Unknown	Unknown	Bandit	$\widetilde{O}(T^{2/3})$	$\Omega(T^{2/3})$	Corollary 5.1
		Full	$\widetilde{O}(T^{1/2})$	$\Omega(T^{1/2})$	Theorem 5.6[1]

[1] The regret bounds of these theorems rely on corresponding assumptions.

the bidder wins the item and pays $p_t \coloneqq \alpha_0 d_t + (1-\alpha_0) b_t$, where α_0 is assumed to be a fixed weight throughout the period; if $b_t < d_t$, the bidder loses the auction and pays nothing. Here we assume that ties are broken in favor of the bidder we concern to simplify exposition. We only consider $\alpha_0 \in [0, 1]$ since $\alpha_0 < 0$ will be immediately detected by the winner and $\alpha_0 > 1$ will lead to lower revenue than mere second-price auctions. One can observe the pricing rule follows the second-price auction when $\alpha_0 = 1$, and it follows the first-price auction when $\alpha_0 = 0$.

Let $x_t \coloneqq \mathbb{I}\{b_t \geq d_t\}$ be the binary variable indicating whether the bidder wins the item. Let $c_t \coloneqq x_t p_t$ be the bidder's cost and let $r_t \coloneqq x_t v_t - c_t$ be the corresponding reward.

Information structure. In this paper, we investigate three cases of prior information: 1) known credibility α_0 and unknown distribution G; 2) unknown credibility α_0 and known distribution G; 3) unknown credibility α_0 and unknown distribution G.

The first two cases not only serve as warm-up analysis, but also have practical significance in their own right. In reality, a bidder may receive some additional signals beyond the learning process to construct her belief over the seller's credibility or the strategies of other bidders, e.g. the seller's reputation heard from other bidders, bidding data collected through other credible channels. Then her bidding algorithm mainly aims to learn the other part in the competing environment.

We consider two cases of information feedback:

(1) *Bandit information feedback.* The bidder can observe the allocation x_t and the cost c_t at the end of each round t.
(2) *Full information feedback.* The bidder can observe the allocation x_t and the price p_t at the end of each round t.

The second information feedback makes sense in non-censored auctions where the seller-side platform (SSP) is supposed to provide the minimum winning price for every bidder regardless of the outcome. If the bidder wins, a dishonest seller will overstate the minimum winning price to overcharge the winner; if the bidder loses, a dishonest seller will understate the minimum winning price to trick the bidder into raising her bids in the following rounds. The full-feedback model for simplicity assumes that these two types of deceptions are symmetric, controlled by the same parameter α_0. Note that when $\alpha_0 = 0$, both feedback models are equivalent to the binary feedback model in first-prices auctions.

We denote the historical observations available to the bidder before submitting a bid in round t by $\mathcal{H}_t$. For the two cases of information feedback, we have, respectively,

$$\mathcal{H}_t^B \coloneqq (v_s, x_s, c_s)_{s=1}^{t-1}, \ \mathcal{H}_t^F \coloneqq (v_s, x_s, p_s)_{s=1}^{t-1}.$$

We will omit the superscript B or F in the remaining of this paper when the context is clear.

Bidding strategy and regret. A bidding strategy maps $(\mathcal{H}_t, v_t)$ to a (possibly random) bid b_t for each t. For a strategy π, we denote by $\mathcal{R}(\pi)$ the expected performance of π, defined as follows:

$$\mathcal{R}(\pi) = \mathbb{E}^{\pi}_{\boldsymbol{v},\boldsymbol{d}}\left[\sum_{t=1}^{T} r_t^{\pi}\right] = \mathbb{E}^{\pi}_{\boldsymbol{v},\boldsymbol{d}}\left[\sum_{t=1}^{T} \mathbb{I}\left\{b_t^{\pi} \geq d_t\right\}\left(v_t - p_t^{\pi}\right)\right],$$

where the expectation is taken with respect to the values $\boldsymbol{v}$, the highest competing bids $\boldsymbol{d}$ and any possible randomness embedded in the strategy π. The expect regret of the bidder is defined to be the difference in the expected cumulative rewards of the bidder's strategy and the optimal bidding strategy, which has the perfect knowledge of α_0, F and G to maximize the target:

$$\mathrm{Regret}(\pi) = \max_{\pi'} \mathcal{R}(\pi') - \mathcal{R}(\pi).$$

Now we look at the the optimal bidding strategy. Using the independence of d_t from v_t, one has

$$\mathcal{R}(\pi) = \sum_{t=1}^{T} \mathbb{E}^{\pi}_{\mathcal{H}_t, v_t}\left[(v_t - b_t^{\pi})G(b_t^{\pi}) + \alpha_0 \int_0^{b_t^{\pi}} G(y)dy\right].$$

Let $r(v, b, \alpha) = (v - b)\, G(b) + \alpha \int_0^b G(y)\, dy$ and let $b^*(v, \alpha) = \arg\max_b r(v, b, \alpha)$ (taking the largest in a case of a tie). Then with the perfect knowledge of α_0 and G, the optimal strategy in each round submits $b^*(v_t, \alpha_0)$, denoted by b_t^* for short. The expect regret of strategy π can written as

$$\mathrm{Regret}(\pi) = \mathbb{E}^{\pi}_{\boldsymbol{v},\boldsymbol{d}}\left[\sum_{t=1}^{T} r(v_t, b_t^*, \alpha_0) - \sum_{t=1}^{T} r(v_t, b_t^{\pi}, \alpha_0)\right].$$

We will omit the superscript π in the remaining of this paper when the context is clear .

3 Learning with Known α and Unknown G

We start with the scenario where the credibility parameter α is known but the distribution G of the highest competing bids is unknown.

Observe that when α reaches an endpoint of $[0, 1]$, this problem will degenerate into a bidding problem in repeated first-price or second-price auctions. Therefore, We will consider three cases separately: $\alpha_0 = 0$, $\alpha_0 = 1$, and $\alpha_0 \in (0, 1)$. The following Theorem 3.1 provides a comprehensive characterization of this setting. For each case, it establishes a regret lower bound and gives an algorithm with optimal performance up to log factors.

Theorem 3.1. *For repeated non-credible second-price auctions with known credibility α_0, unknown distribution G and bandit feedback:*

(1) when $\alpha_0 = 1$, truthful bidding achieves no regret;
(2) when $\alpha_0 = 0$, there exists a bidding algorithm (Algorithm A.1) that achieves an $\widetilde{O}(T^{2/3})$ regret, and the lower bound on regret for this case is $\Omega(T^{2/3})$;
(3) when $\alpha_0 \in (0, 1)$, there exists a bidding algorithm (Algorithm A.2) that achieves an $\widetilde{O}(T^{1/2})$ regret, and the lower bound on regret for this case is $\Omega(T^{1/2})$.

The lower bounds in Theorem 3.1 show that the hardness of this learning problem increases as α_0 decreases, which demonstrates the impact of reduced credibility on online bidding optimization. When α_0 deviates from 1, truthful bidding is no longer a dominant strategy. The bidder has to learn the competitors' bids to make decisions, and thus the distribution estimation error would introduce an inevitable regret of order $\Omega(T^{1/2})$. As long as $\alpha_0 > 0$, the bidder can infer the highest competing bid d_t from her payment once she wins an auction by measuring the difference between c_t and b_t. However, when α_0 becomes 0, $c_t \equiv b_t$ in winning rounds provides no additional information about d_t. The complete loss of credibility cripples the bidder's ability to observe the competitive environment and estimate the distribution G, so the regret lower bound leaps from $\Omega(T^{1/2})$ to $\Omega(T^{2/3})$.

The first statement of Theorem 3.1 is trivial due to the nature of the second-price auction.

The second case is equivalent to bidding in repeated first-price auctions with binary feedback (receiving only $x_t = 1$ or 0), which can be modeled as a contextual bandits problem with cross learning:

- *Cross learning over contexts.* The bidder in round t not only receives the reward r_t under (v_t, b_t), but also observes the rewards r' under (v', b_t) for every other v'.

For the contextual bandits problem with cross-learning over contexts in the stochastic setting, [4] proposed a UCB-based algorithm that can achieve an $O(\sqrt{KT})$ regret, where K is the number of actions. Applying this algorithm to the auction setting results in a regret bound of $\widetilde{O}(\sqrt{KT} + T/K)$, where the last term comes from the discretization error and the upper bound becomes $\widetilde{O}(T^{2/3})$ with $K \sim T^{1/3}$. They also proved the regret lower bound is $\Omega(T^{2/3})$ via a reduction to the problem of dynamic pricing.

Lemma 3.2 ([4]). *For repeated first-price auctions with binary feedback, Algorithm A.1 can achieve an $\widetilde{O}(T^{2/3})$ regret, and there exists a problem instance where any algorithm must incur a regret of at least $\Omega(T^{2/3})$ regret.*

The third case is similar to bidding in repeated first-price auctions with censored feedback, where the seller runs first-price auctions and always reveals the winner's bid to each bidder so the bidder can see the highest competing bid only if she loses the auction. [18] modeled that problem as a contextual bandits problem with cross learning, partial ordering:

- *Cross learning over contexts.*
- *Partial ordering over actions.* There exists a partial order $\preceq_{\mathcal{B}}$ over the action set $\mathcal{B}$. The bidder in round t not only receives the reward r_t under (v_t, b_t), but also observes the rewards r' under (v_t, b') for every other $b' \preceq_{\mathcal{B}} b_t$.
- *Partial ordering over contexts.* There exists a partial order $\preceq_{\mathcal{V}}$ over the context set $\mathcal{V}$ such that if $v_1 \preceq_{\mathcal{V}} v_2$, then $b^*(v_1) \preceq_{\mathcal{B}} b^*(v_2)$ where $b^*(v)$ is the optimal auction under context v.

Lemma 3.3. *[[18]] For the contextual bandits problem with cross-learning over contexts, partial ordering over auctions and contexts in the stochastic setting, there exists a bidding algorithm than achieves an $\widetilde{O}(T^{1/2})$ regret.*

We carefully adapt their algorithm to our setting (Algorithm A.2) and verify that the third case with $\alpha_0 \in (0, 1)$ satisfies the above three properties:

- *Cross learning over values.* The bidder can calculate r' under (v', b_t) by
$$r' = r_t + x_t(v' - v_t). \tag{1}$$
- *Partial ordering over bids.* If the bidder wins in round t, she can infer the highest competing bid d_t by
$$d_t = (c_t - (1 - \alpha_0)b_t) / \alpha_0. \tag{2}$$
Therefore the reward r' under (v_t, b') for any other $b < b_t$ can be calculated by using the corresponding allocation $\mathbb{I}\{b \geq d_t\}$ and price $\alpha_0 d_t + (1 - \alpha_0)b$. If the bidder loses the auction with b_t, she should also lose with $b' < b_t$ and the reward r' is 0.
- *Partial ordering over values.* We have shown in the previous section that the optimal bid under value v is $b^*(v, \alpha_0) = \arg\max_b r(v, b, \alpha_0)$. The following lemma shows it is a non-decreasing function in v.

Lemma 3.4. *$b^*(v, \alpha) = \arg\max_b r(v, b, \alpha)$ (taking the largest in the case of a tie) is a non-decreasing function in both v and α.*

Therefore, Algorithm A.2 can achieve an $\widetilde{O}(T^{1/2})$ regret when $\alpha_0 \in (0, 1)$. For the last piece of the puzzle, we prove the following regret lower bound. Remark that Lemma 3.5 holds for any α_0, though the bound is not tight when $\alpha_0 = 0$.

Lemma 3.5. *For repeated non-credible second-price auctions with known credibility α_0, unknown distribution G, there exists a constant $c > 0$ such that*

$$\inf_\pi \sup_G \text{Regret}(\pi) \geq c \cdot (1 - \alpha_0)\sqrt{T},$$

even in the special case with $v_t \equiv 1$ and full feedback.

4 Learning with Unknown α and Known G

We next consider the scenario where distribution G is known but credibility α_0 is unknown.

Algorithm 4.1: Learning with unknown α_0, known G and bandit feedback

---

Input: Time horizon T; distribution G.
Initialization: The bidder submits $b_1 = 1$.
for $t \leftarrow 2$ **to** T **do**
  The bidder receives the value $v_t \in [0, 1]$.
  The bidder estimates the seller's credibility by

$$\widetilde{\alpha}_t = \arg\min_{\alpha \in [0,1]} \left| \sum_{s=1}^{t-1} (r_s - r(v_s, b_s, \alpha)) \right|, \tag{3}$$

  The bidder submits $b_t = \arg\max_b r(v_t, b, \widetilde{\alpha}_t)$.
end

Our bidding algorithm for this setting is depicted in Algorithm 4.1. The bidder first conducts a one-round exploration to make an appropriate initialization. After receiving the value v_t in each round $t = 2, \dots, T$, the bidder computes $\widetilde{\alpha}_t$, which is the estimation of α_0 based on the historical observations in the past $t-1$ rounds. Recall that the optimal bid b_t^* shown in Section 2 maximizes $r(v_t, b, \alpha_0)$. Thus, by the choice of b_t, if the estimator $\widetilde{\alpha}_t$ is close to α_0, the expected reward $r(v_t, b_t, \alpha_0)$ will be close to the optimal reward $r(v_t, b_t^*, \alpha_0)$.

Equation (3) in Algorithm 4.1 aims to estimate α_0 by fitting the observed rewards $\{r_s\}_{s=1}^{t-1}$ to the expected rewards $\{r(v_s, b_s, \alpha)\}_{s=1}^{t-1}$, which can be computed given G is known. We apply the Azuma-Hoeffding inequality to obtain the following result.

Lemma 4.1. *Under Algorithm 4.1, we have with probability at least $1 - \delta$, $\forall t \in [T]$,*

$$|\widetilde{\alpha}_t - \alpha_0| \le w_t, \text{ where } w_t \text{ is given by}$$

where w_t is given by

$$w_t = \frac{2\sqrt{2(t-1)\log(2T/\delta)}}{\sum_{s=1}^{t-1} \int_0^{b_s} G(y)dy}.$$

For technical purpose, we make the following assumption.

Assumption 4.2. G is twice differentiable and log-concave with density function g. There exist positive constants B_1, B_2 such that $B_1 \le g(x) \le B_2$ for $x \in [0, 1]$.

The CDFs of many common distributions, such as gamma distributions, Gaussian distributions and uniform distributions, are all log-concave. The existence of positive bounds on the density function is also a standard and common assumption in various learning problems.

Theorem 4.3. *Suppose that Assumption 4.2 holds. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and bandit feedback, there exists a bidding algorithm (Algorithm 4.1) that achieves an $O(\log^2 T)$ regret, and any algorithm must incur at least a constant regret.*

Remark that due to the additional assumption, we cannot actually draw the conclusion that estimating α is generally easier than estimating G by comparing two lower bounds in Theorem 4.3 and Theorem 3.1. For example, the two-point method used in the proof of Lemma 3.5 constructs two discrete G distributions to show no bidding algorithm can obtain low regret simultaneously under both distributions, while Assumption 4.2 has ruled out such bad cases.

A key step in the proof of Theorem 4.3 involves showing that the reward per round obtained by the bidder is close to the reward under optimal bid with high probability. We have with probability at least $1 - \delta$, $\forall t$, $r(v_t, b_t^*, \alpha_0) - r(v_t, b_t, \alpha_0) \le w_t^2/B_1$. Then the regret upper bound can be established by showing $w_t \sim \sqrt{\log T/t}$ and summing up through 1 to T. It is also worth discussing the robustness of this result. Even without Assumption 4.2, we can get $r(v_t, b_t^*, \alpha_0) - r(v_t, b_t, \alpha_0) \le w_t$. And with a weaker continuity condition, it still holds that $w_t \sim \sqrt{\log T/t}$.

Corollary 4.4. *Suppose that G is continuous. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and bandit feedback, Algorithm 4.1 can achieve an $\widetilde{O}(\sqrt{T})$ regret.*

An intuition on why the proof of the regret upper bound may fail under some discontinuous distributions is given in Example B.3. In spite of this, we conjecture that Algorithm 4.1 can also guarantee a lower regret in the case with discontinuous G and we leave that as a future direction.

5 Learning with Unknown α and Unknown G

Bandit feedback. For the last scenario with both α and G unknown, we first consider bandit feedback. Although the seller's credibility is unknown, this case still satisfies the cross-learning property over values, i.e., Equation (1) holds. Thus, Algorithm A.1, which actually does not use the value of α_0, can still work in this case and achieve an $\widetilde{O}(T^{2/3})$ regret. The regret lower bound also directly follows the third statement of Theorem 3.1 since any bidding strategy cannot obtain a better regret guarantee than $O(T^{2/3})$ when $\alpha_0 = 0$. Hence, we have the following result.

Corollary 5.1. *For repeated non-credible second-price auctions with unknown credibility α_0, unknown distribution G and bandit feedback, there exists a bidding algorithm (Algorithm A.1) that achieves an $\widetilde{O}(T^{2/3})$ regret, and the lower bound on regret for this problem is $\Omega(T^{2/3})$.*

In spite of the sublinear regret, the result of Corollary 5.1 is not satisfactory. Algorithm A.1 only falls into the category of usual UCB policies with cross learning, but does not make full use of the properties of this auction setting. In fact, it treats the seller's mechanism as a black box without really estimating the seller's credibility or other bidder's strategies.

Full feedback. With the above in mind, we explore whether we can make any improvements with richer feedback and more meticulous estimation of α_0 and G. In the full feedback model, $p_t = \alpha_0 d_t + (1 - \alpha_0) b_t$ is always observable, which would intuitively help our estimation. Nevertheless, when $\alpha_0 = 0$, binary feedback in first-prices auctions still results in the $\Omega(T^{2/3})$ lower bound. Thus, in what follows we exclude the extreme case of $\alpha = 0$

when it is impossible to achieves a better bound than $O(T^{2/3})$. Note that sellers in reality often do not set $\alpha_0 = 0$ due to concerns about reputation or cheating costs [22]. This assumption is also consistent with the recent empirical findings by [12] that although sellers in non-credible second-price auctions often overcharge, they typically do not use the rules of the first-price auction to maximize revenue.

Assumption 5.2. There exists a constant $\underline{\alpha} > 0$ such that $\alpha_0 \in [\underline{\alpha}, 1]$.

We also make a slightly stronger assumption over G.

Assumption 5.3. G is twice differentiable with log-concave density function g. There exist positive constants B_1, B_2, B_3, W such that $B_1 < g(x) < B_2$ and $|g'(x)| \leq B_3$ for $x \in [0, 1 + W]$.

Our bidding algorithm for full feedback is presented in Algorithm 5.1. It runs in an episodic manner, similar to [9],[21], [3]. During a time horizon T, the bidding algorithm is divided into S episodes, each containing $T_s = T^{1-2^{-s}}$ time steps. Denote Γ_s be the time steps in stage s, s.t. $|\Gamma_s| = T_s$. For any time step t in the first episode, we set $b_t = 1$ for a proper initialization. For any time step t in episode $s (s \geq 2)$, we use the estimated parameter $\widetilde{\alpha}_{s-1}$ and distribution $\hat{G}_{s-1}$ in the $(s-1)$-th episode to set the bid

$$b_t = \arg\max_b \, (v_t - b)\, \hat{G}_{s-1}(b) + \widetilde{\alpha}_{s-1} \int_0^b \hat{G}_{s-1}(y)\, dy,$$

and only update these estimators at the end of episode s by using the data observed in episode s.

The main difficulty is how to update the estimators of G and α_0 in each episode. Recall that when G is known, the estimation step (Equation (3)) in Algorithm 4.1 is essentially similar to using the maximum likelihood estimation (MLE) method to find the most probable α. However, without the knowledge of G, the bidder cannot directly estimate α_0 by matching observed rewards to expected rewards. To handle this challenge, we combine the non-parametric log-concave density estimator and MLE method, to learn α_0 and G simultaneously.

We first introduce the non-parametric estimator of log-concave density function g, which is adopted from [13]. In each episode s, given realized $p_t, t \in \Gamma_s$, Algorithm 5.1 for any parameter α gives an estimator $\hat{g}_s(\cdot, \alpha)$,

$$\hat{g}_s(\cdot, \alpha) = \arg\max_{g \text{ is log-concave}} \frac{1}{T_s} \sum_{t \in \Gamma_s} \log g\left(\frac{p_t - (1-\alpha) b_t}{\alpha}\right) - \int g(y) dy. \qquad (11)$$

In this work, we restrict the function class of $\hat{g}_s(\cdot; \alpha)$ for any α and s as below,

$$\mathcal{P} = \left\{p : p(z) \leq B_2, \forall z \in [0, 1+W], \int p(z) dz = 1\right\}.$$

It is *w.l.o.g.* to re-parameterize $g(y) = \exp(\Psi(y))$, where $\Psi(y)$ is a concave function *w.r.t* y. Then it is equivalent to get an estimator $\hat{\Psi}_s(\cdot, \alpha)$,

$$\hat{\Psi}_s(\cdot, \alpha) = \arg\max_{\Psi \text{ is concave}} \frac{1}{T_s} \sum_{t \in \Gamma_s} \Psi\left(\frac{p_t - (1-\alpha) b_t}{\alpha}\right) - \int \exp(\Psi(y; \alpha)) dy. \qquad (12)$$

Algorithm 5.1: Learning with unknown α_0, unknown G and full feedback

---

Input: Time horizon T.

for $t \in \Gamma_1$ **do**

  The bidder receives the value $v_t \in [0, 1]$.

  The bidder submits a bid $b_t = 1$.

end

Estimate α_0 by using $\widetilde{\alpha}_1$, which is computed by,

$$\widetilde{\alpha}_1 = \arg \min_{\alpha \in [\underline{\alpha}, 1]} \mathcal{L}_1(\alpha), \qquad (4)$$

where $\mathcal{L}_1(\alpha)$ is defined in Equation (13).

The bidder computes $\hat{\Psi}(\cdot; \widetilde{\alpha}_1)$ by

$$\hat{\Psi}_1(\cdot, \widetilde{\alpha}_1) = \arg\max_{\Psi \text{ is concave}} \frac{1}{T_s} \sum_{t \in \Gamma_s} \Psi\left(\frac{p_t - (1-\widetilde{\alpha}_1) b_t}{\widetilde{\alpha}_1}\right) - \int \exp(\Psi(y; \widetilde{\alpha}_1)) dy. \qquad (5)$$

The bidder estimates G by

$$\hat{G}_1(\cdot; \widetilde{\alpha}_1) = \int \exp(\hat{\Psi}_1(y; \widetilde{\alpha}_1)) dy. \qquad (6)$$

for $s = 2, 3, \cdots, S$ **do**

  for $t \in \Gamma_s$ **do**

    The bidder receives the value $v_t \in [0, 1]$.

    The bidder submits a bid b_t, computed by,

$$b_t = \arg\max_b \left[(v_t - b)\, \hat{G}_{s-1}(b) + \widetilde{\alpha}_{s-1} \int_0^b \hat{G}_{s-1}(y)\, dy \right], \qquad (7)$$

  end

  The bidder updates the estimator for α_0 in episode s by using $\widetilde{\alpha}_s$, which is computed by,

$$\widetilde{\alpha}_s = \arg \min_{\alpha \in [\underline{\alpha}, 1]} \mathcal{L}_s(\alpha). \qquad (8)$$

  where $\mathcal{L}_s(\alpha)$ is defined in Equation (13). The bidder updates $\hat{\Psi}(\cdot; \widetilde{\alpha}_1)$ by

$$\hat{\Psi}_s(\cdot, \widetilde{\alpha}_s) = \arg\max_{\Psi \text{ is concave}} \frac{1}{T_s} \sum_{t \in \Gamma_s} \Psi\left(\frac{p_t - (1-\widetilde{\alpha}_s) b_t}{\widetilde{\alpha}_s}\right) - \int \exp(\Psi(y; \widetilde{\alpha}_s)) dy. \qquad (9)$$

  The bidder updates the estimation of G by

$$\hat{G}_s(\cdot; \widetilde{\alpha}_s) = \int \exp(\hat{\Psi}_1(y; \widetilde{\alpha}_s)) dy. \qquad (10)$$

end

Let $\hat{G}_s(y;\alpha) = \int_0^y \hat{g}_s(z,\alpha)dz = \int_0^y \exp(\hat{\Psi}_s(z;\alpha))dz$ be the estimated empirical distribution using the above estimator. Let $\mathbb{G}_s$ be the empirical distribution of samples $\{d_t\}_{t\in\Gamma_s}$ in episode s such that $\mathbb{G}_s(y) = \frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t \le y\}$. [13] proved the following result.

Lemma 5.4 ([13]). *The optimizer $\hat{\Psi}_s(\cdot,\alpha_0)$ exists and is unique. For any $d_t = (p_t - (1-\alpha_0)b_t)/\alpha_0, t \in \Gamma_s$, $\mathbb{G}_s(d_t) - \frac{1}{T_s} \le \hat{G}_s(d_t;\alpha_0) \le \mathbb{G}_s(d_t)$.*

Give the above characterization of $\hat{\Psi}_s(\cdot,\alpha_0)$ and $\hat{G}_s(\cdot;\alpha_0)$ we provide the uniform convergence bound for $|\hat{G}_s(d;\alpha_0) - G(z)|$ in the following lemma.

Lemma 5.5. *Suppose that $T_s \gg \log^2(2/\delta)$ for some $\delta > 0$, then with probability at least $1-\delta$, $\forall d \in [0,1], |\hat{G}_s(d;\alpha_0) - G(z)| \le O(\sqrt{\log(1/\delta)/T_s})$ holds .*

Given the non-parametric estimator $\hat{G}_s(\cdot;\alpha)$ introduced in the above, we minimize the following MLE loss function to compute $\widetilde{\alpha}_s$,

$$\mathcal{L}_s(\alpha) = -\frac{1}{T_s}\sum_{t\in\Gamma_s}\Big[\xi_t \log \hat{G}_s(\varepsilon_t(\alpha);\alpha) + (1-\xi_t)\log(1-\hat{G}_s(\varepsilon_t(\alpha);\alpha))\Big], \quad (13)$$

where $\xi_t = \mathbb{I}\{p_t \le b_t + \underline{\alpha}W/2\}$ and $\varepsilon_t(\alpha) = b_t + \underline{\alpha}W/(2\alpha)$. Here, the indicator ξ_t is carefully chosen so that $\mathbb{E}[\xi_t] = G(\varepsilon_t(\alpha))$, and $\forall\alpha \in [\underline{\alpha},1]$, $0 < \varepsilon_t(\alpha) < 1+W$.

Theorem 5.6. *Suppose that Assumption 5.2 holds. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and full feedback, there exists a bidding algorithm Algorithm 5.1 that achieves $\widetilde{O}(T^{1/2})$ if Assumption 5.3 holds. And the lower bound on regret for this problem is $\Omega(T^{1/2})$.*

We first give a bound on distance between $\widetilde{\alpha}_s$ and α_0 in Lemma C.3. Next we show that the distribution estimator $\hat{G}_s(\cdot;\widetilde{\alpha}_s)$ uniformly converges to the real distribution G in Lemma C.4. Our algorithm and the regret analysis follow the same spirit as in Theorem 4.3 in [3]. The main difference is that we use ξ_t rather than x_t as our indicator function.

6 Limitations

This is the first work on the bidding problem when the seller can potentially deviate from his announced auction rule, but the setting is a little bit specific and restrictive, which is the main limitation of this work. We have made several simplifications and assumptions to make the model tractable. In a more realistic setting, the seller's credibility α_0 might not be fixed or the seller might not even use a linear charging rule. The independence between values an highest competing bids is also a strong assumption, though standard in literature. We are working on how to generalize our techniques in this paper to other cases.

References

[1] Mohammad Akbarpour and Shengwu Li. 2020. Credible auctions: A trilemma. *Econometrica* 88, 2 (2020), 425–467.

[2] Mark An. 1996. *Log-concave Probability Distributions: Theory and Statistical Testing*. Game Theory and Information. University Library of Munich, Germany. https://EconPapers.repec.org/RePEc:wpa:wuwpga:9611002

[3] Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. 2023. Learning to Bid in Contextual First Price Auctions. In *Proceedings of the ACM Web Conference 2023*. 3489–3497.

[4] Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, and Jon Schneider. 2019. Contextual bandits with cross-learning. *Advances in Neural Information Processing Systems* 32 (2019).

[5] Santiago R Balseiro, Omar Besbes, and Gabriel Y Weintraub. 2015. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science* 61, 4 (2015), 864–884.

[6] Santiago R Balseiro and Yonatan Gur. 2019. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science* 65, 9 (2019), 3952–3968.

[7] Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. 2022. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research* (2022).

[8] Matteo Castiglioni, Andrea Celli, and Christian Kroer. 2022. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*. PMLR, 2767–2783.

[9] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2014. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory* 61, 1 (2014), 549–564.

[10] Yurong Chen, Qian Wang, Zhijian Duan, Haoran Sun, Zhaohua Chen, Xiang Yan, and Xiaotie Deng. 2023. Coordinated Dynamic Bidding in Repeated Second-Price Auctions with Budgets. *arXiv preprint arXiv:2306.07709* (2023).

[11] Zhaohua Chen, Chang Wang, Qian Wang, Yuqi Pan, Zhuming Shi, Chuyue Tang, Zheng Cai, Yukun Ren, Zhihua Zhu, and Xiaotie Deng. 2022. Dynamic Budget Throttling in Repeated Second-Price Auctions. *arXiv preprint arXiv:2207.04690* (2022).

[12] Ahrash Dianat and Mikhail Freer. 2021. Credibility in Second-Price Auctions: An Experimental Test. *arXiv preprint arXiv:2105.00204* (2021).

[13] Lutz Dümbgen and Kaspar Rufibach. 2009. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* 15, 1 (2009), 40–68.

[14] Zhe Feng, Swati Padmanabhan, and Di Wang. 2022. Online Bidding Algorithms for Return-on-Spend Constrained Advertisers. *arXiv preprint arXiv:2208.13713* (2022).

[15] Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. 2021. Bidding and pricing in budget and ROI constrained markets. *arXiv preprint arXiv:2107.07725* (2021).

[16] Rustamdjan Hakimov and Madhav Raghavan. 2023. Improving Transparency and Verifiability in School Admissions: Theory and Experiment. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-94449-1

[17] Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. 2020. Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568* (2020).

[18] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. 2020. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795* (2020).

[19] The Interactive Advertising Bureau IAB. 2023. Internet Advertising Revenue Report, Full year 2022 results. https://www.iab.com/wp-content/uploads/2023/04/IAB_PwC_Internet_Advertising_Revenue_Report_2022.pdf

[20] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. 2014. Mean field equilibria of dynamic auctions with learning. *Management Science* 60, 12 (2014), 2949–2970.

[21] Adel Javanmard and Hamid Nazerzadeh. 2019. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research* 20, 1 (2019), 315–363.

[22] David McAdams and Michael Schwarz. 2007. Who pays when auction rules are bent? *International Journal of Industrial Organization* 25, 5 (2007), 1144–1157.

[23] Ryan Porter and Yoav Shoham. 2005. On cheating in sealed-bid auctions. *Decision Support Systems* 39, 1 (2005), 41–54.

[24] Michael H Rothkopf and Ronald M Harstad. 1995. Two models of bid-taker cheating in Vickrey auctions. *Journal of Business* (1995), 257–267.

[25] Alexandre B. Tsybakov. 2009. *Introduction to Nonparametric Estimation*. Springer-Verlag, New York.

[26] William Vickrey. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance* 16, 1 (1961), 8–37.

[27] Qian Wang, Zongjun Yang, Xiaotie Deng, and Yuqing Kong. 2023. Learning to bid in repeated first-price auctions with budgets. *arXiv preprint arXiv:2304.13477* (2023).

[28] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. 2016. Online learning in repeated auctions. *Journal of Machine Learning Research* 49, June (2016), 1562–1583.

[29] Wei Zhang, Yanjun Han, Zhengyuan Zhou, Aaron Flores, and Tsachy Weissman. 2022. Leveraging the Hints: Adaptive Bidding in Repeated First-Price Auctions. *arXiv preprint arXiv:2211.06358* (2022).

A Missing Algorithms and Proofs of Section 3

Algorithm A.1: Learning with known $\alpha_0 = 0$ (or unknown α_0), unknown G and bandit feedback

Input: Time horizon T.

Initialization: Let $\mathcal{B} = \{b^1, \cdots, b^K\}$ be the bid set with $b^k = (k-1)/K$; for $t \in [K]$, the bidder submits a bid $b_t = b^t$.

for $t \leftarrow K+1$ **to** T **do**

The bidder receives the value $v_t \in [0, 1]$.

For every $k \in [K]$, the bidder counts the observation by

$$n_t^k \leftarrow \sum_{s=1}^{t-1} \mathbb{I}\left\{b_s = b^k\right\}. \tag{14}$$

For every $k \in [K]$, the bidder computes the confidence bound:

$$w_t^k = \sqrt{\frac{2 \ln T}{n_t^k}}, \tag{15}$$

For every $k \in [K]$, the bidder estimates the reward by

$$\widetilde{r}_t(v_t, b^k) = \frac{1}{n_t^k} \sum_{s=1}^{t-1} \mathbb{I}\left\{b_s = b^k\right\} (x_s v_t - c_s), \tag{16}$$

The bidder submits a bid $b_t = \arg\max_{b \in \mathcal{B}} \left(\widetilde{r}_t(v_t, b) + w_t^k\right)$.

end

Lemma 3.4. $b^*(v, \alpha) = \arg\max_b r(v, b, \alpha)$ *(taking the largest in the case of a tie) is a non-decreasing function in both* v *and* α.

Proof of Lemma 3.4. Fix any v_1, v_2 with $v_1 \le v_2$. For any $b \le b^*(v_1, \alpha)$, it holds that

$$\begin{aligned} r(v_2, b, \alpha) &= r(v_1, b, \alpha) + (v_2 - v_1) G(b) \\ &\le r(v_1, b^*(v_1, \alpha), \alpha) + (v_2 - v_1) G(b^*(v_1, \alpha)) \\ &= r(v_2, b^*(v_1, \alpha), \alpha), \end{aligned}$$

where the inequality follows from the definition of $b^*(v_1, \alpha)$ and the conditions $v_1 \le v_2, b \le b^*(v_1, \alpha)$. Thus, all bids no larger than $b^*(v_1, \alpha)$ cannot be the largest maximizer for v_2.

Fix any α_1, α_2 with $\alpha_1 \le \alpha_2$. For any $b \le b^*(v, \alpha_1)$, it holds that

$$\begin{aligned} r(v, b, \alpha_2) &= r(v, b, \alpha_1) + (\alpha_2 - \alpha_1) \int_0^b G(y)\, dy \\ &\le r(v, b^*(v, \alpha_1), \alpha_1) + (\alpha_2 - \alpha_1) \int_0^{b^*(v, \alpha_1)} G(y)\, dy \\ &= r(v, b^*(v, \alpha_1), \alpha_2), \end{aligned}$$

where the inequality follows from the definition of $b^*(v, \alpha)$ and the conditions $\alpha_1 \le \alpha_2, b \le b \le b^*(v, \alpha_1)$. Thus, all bids no larger than $b^*(v, \alpha_1)$ cannot be the largest maximizer for α_2. □

Lemma 3.5. *For repeated non-credible second-price auctions with known credibility* α_0*, unknown distribution* G*, there exists a constant* $c > 0$ *such that*

$$\inf_\pi \sup_G \mathrm{Regret}(\pi) \ge c \cdot (1 - \alpha_0)\sqrt{T},$$

even in the special case with $v_t \equiv 1$ *and full feedback.*

Proof of Lemma 3.5. The proof follows the Le Cam's two-point method [25]. For any $\alpha_0 \in (0, 1)$, consider the following two candidate distributions supported on $[0, 1]$:

$$G_1(x) = \begin{cases} 0 & \text{if } x \in \left[0, \frac{1-\alpha_0}{3-2\alpha_0}\right) \\ \frac{1}{2} + \Delta & \text{if } x \in \left[\frac{1-\alpha_0}{3-2\alpha_0}, \frac{2-\alpha_0}{3-2\alpha_0}\right) \\ 1 & \text{if } x \in \left[\frac{2-\alpha_0}{3-2\alpha_0}, 1\right] \end{cases}, \qquad G_2(x) = \begin{cases} 0 & \text{if } x \in \left[0, \frac{1-\alpha_0}{3-2\alpha_0}\right) \\ \frac{1}{2} - \Delta & \text{if } x \in \left[\frac{1-\alpha_0}{3-2\alpha_0}, \frac{2-\alpha_0}{3-2\alpha_0}\right) \\ 1 & \text{if } x \in \left[\frac{2-\alpha_0}{3-2\alpha_0}, 1\right] \end{cases},$$

Algorithm A.2: Learning with known $\alpha_0 \in (0,1)$, unknown G and bandit feedback

Input: Time horizon T; credibility parameter $\alpha_0 \in (0,1)$.

Initialization: Let $\mathcal{V} = [v^1, \cdots, v^M]$ be the value set with $v^m = (m-1)/M$; let $\mathcal{B} = \{b^1, \cdots, b^K\}$ be the bid set with $b^k = (k-1)/K$; set $\mathcal{B}_1^m \leftarrow \mathcal{B}$ for each $v_m \in \mathcal{V}$; select failure probability $\delta \in (0,1)$.

for $t \leftarrow 1$ **to** T **do**

  The bidder receives the value $v_t \in [0,1]$ and round it to $v^{m(t)}$ by

$$v^{m(t)} = \max\{u \in \mathcal{V} : u \leq v_t\}. \tag{17}$$

  The bidder submits a bid $b_t = \sup \mathcal{B}_t^{m(t)}$.

  For every $k \in [K]$, the bidder counts the observation by

$$n_t^k \leftarrow \sum_{s=1}^{t} \mathbb{I}\left\{b_s \geq b^k\right\}. \tag{18}$$

  For every $m \in [M], k \in [K]$, the bidder estimates the reward by

$$\widetilde{r}_t(v^m, b^k, \alpha_0) = \frac{1}{n_t^k} \sum_{s=1}^{t} \mathbb{I}\left\{b_s \geq b^k\right\} \mathbb{I}\left\{b^k \geq \widetilde{d}_s\right\} x_s \left(v^m - \alpha_0 \widetilde{d}_s - (1-\alpha_0) b^k\right), \tag{19}$$

  where $\widetilde{d}_s = (c_s - (1-\alpha_0) b_s)/\alpha_0$.

  for $m = M, M-1, \ldots, 1$ **do**

    The bidder eliminates bids by:

$$\mathcal{B}_t^m = \left\{b^k \in \mathcal{B}_t^m : b^k \leq \min_{s>m} \sup \mathcal{B}_{t+1}^s\right\}. \tag{20}$$

    The bidder computes the confidence bound:

$$w_t^m = \sqrt{\frac{4 \ln T \log(KT/\delta)}{N_t^m}}, \tag{21}$$

    where $N_t^m = \min_{b^k \in \mathcal{B}_t^m} n_t^k$.

    The bidder eliminates bids by:

$$\mathcal{B}_{t+1}^m \leftarrow \left\{b^k \in \mathcal{B}_t^m : \widetilde{r}_t(v^m, b^k, \alpha_0) \geq \max_{b \in \mathcal{B}_t^m} \widetilde{r}_t(v^m, b, \alpha_0) - 2 w_t^m\right\}. \tag{22}$$

  end

end

where Δ is some parameter to be chosen later. In other words, G_1 corresponds to a discrete random variable taking value in $\left\{\frac{1-\alpha_0}{3-2\alpha_0}, \frac{2-\alpha_0}{3-2\alpha_0}\right\}$ with probability $(1/2+\Delta, 1/2-\Delta)$, and G_2 corresponds to the probability $(1/2-\Delta, 1/2+\Delta)$. Fixing $v_t \equiv 1$, let $R_1(b)$ and $R_2(b)$ be the expected per-round reward when bidding b under G_1 and G_2, respectively. After some algebra, it is straightforward to check that

$$\max_{b \in [0,1]} R_1(b) = \max_{b \in [0,1]} \left[(1-b) G_1(b) + \alpha_0 \int_0^b G_1(x) \mathrm{d}x\right] = \frac{2-\alpha_0}{3-2\alpha_0}\left(\frac{1}{2} + \Delta\right),$$

$$\max_{b \in [0,1]} R_2(b) = \max_{b \in [0,1]} \left[(1-b) G_2(b) + \alpha_0 \int_0^b G_2(x) \mathrm{d}x\right] = \frac{1}{3-2\alpha_0}\left(1 - \frac{\alpha_0}{2} + \alpha_0 \Delta\right),$$

$$\max_{b \in [0,1]} (R_1(b) + R_2(b)) = \max_{b \in [0,1]} \left[(1-b)\left(G_1(b) + G_2(b)\right) + \alpha_0 \int_0^b \left(G_1(x) + G_2(x)\right) \mathrm{d}x\right] = \frac{2-\alpha_0}{3-2\alpha_0}.$$

Hence, for any $b_t \in [0,1]$, we have

$$\begin{aligned}
&\left(\max_{b\in[0,1]} R_1(b) - R_1(b_t)\right) + \left(\max_{b\in[0,1]} R_2(b) - R_2(b_t)\right) \\
&\geq \max_{b\in[0,1]} R_1(b) + \max_{b\in[0,1]} R_2(b) - \max_{b\in[0,1]} (R_1(b) + R_2(b)) \\
&= \frac{2-2\alpha_0}{3-2\alpha_0}\Delta. \qquad (23)
\end{aligned}$$

The inequality (23) is the separation condition required in the two-point method: there is no single bid b_t that can obtain a uniformly small instantaneous regret under both G_1 and G_2.

In the full information model, when $\alpha_0 \neq 0$, the bidder can always infer d_t by

$$d_t = (p_t - (1-\alpha_0)b_t)/\alpha_0.$$

Let $P_i^t, i \in \{1,2\}$ be the distribution of all historical samples $(d_1, \cdots, d_{t-1})$ at the beginning of time t. Then for any policy π,

$$\begin{aligned}
\sup_G \text{Regret}(\pi; G) &\overset{(a)}{\geq} \frac{1}{2}\text{Regret}(\pi; G_1) + \frac{1}{2}\text{Regret}(\pi; G_2) \\
&= \frac{1}{2}\sum_{t=1}^{T}\left(\mathbb{E}_{P_1^t}\left[\max_{b\in[0,1]} R_1(b) - R_1(b_t)\right] + \mathbb{E}_{P_2^t}\left[\max_{b\in[0,1]} R_2(b) - R_2(b_t)\right]\right) \\
&\overset{(b)}{\geq} \frac{1}{2}\sum_{t=1}^{T}\frac{2-2\alpha_0}{3-2\alpha_0}\Delta\int \min\{\mathrm{d}P_1^t, \mathrm{d}P_2^t\} \\
&\overset{(c)}{=} \frac{1}{2}\sum_{t=1}^{T}\frac{2-2\alpha_0}{3-2\alpha_0}\Delta\left(1 - \|P_1^t - P_2^t\|_{\text{TV}}\right) \\
&\geq \frac{1-\alpha_0}{3}\Delta\sum_{t=1}^{T}\left(1 - \|P_1^t - P_2^t\|_{\text{TV}}\right), \qquad (24)
\end{aligned}$$

where (a) is due to the fact that the maximum is no smaller than the average, (b) follows from (23), and (c) is due to the identity $\int \min\{\mathrm{d}P, \mathrm{d}Q\} = 1 - \|P - Q\|_{\text{TV}}$. Invoking Lemma A.1 and using the fact that for $\Delta \in (0, 1/4)$,

$$\begin{aligned}
D_{\text{KL}}(P_1^t || P_2^t) &= (t-1)D_{\text{KL}}(G_1||G_2) \\
&= (t-1)\left(\left(\frac{1}{2}+\Delta\right)\log\frac{1/2+\Delta}{1/2-\Delta} + \left(\frac{1}{2}-\Delta\right)\log\frac{1/2-\Delta}{1/2+\Delta}\right) \\
&\leq 16T\Delta^2, \qquad (25)
\end{aligned}$$

we have the following inequality on total variation distance:

$$1 - \|P_1^t - P_2^t\|_{\text{TV}} \geq \frac{1}{2}\exp\left(-16T\Delta^2\right). \qquad (26)$$

Finally, choosing $\Delta = \frac{1}{4\sqrt{T}}$, combining inequalities (24) and (26), we conclude that Lemma 3.5 holds with the constant $c = \frac{1}{24e}$. □

Lemma A.1 ([25]). *Let P, Q be two probability measures on the same space. It holds that*

$$1 - \|P - Q\|_{\text{TV}} \geq \frac{1}{2}\exp\left(-\frac{D_{\text{KL}}(P||Q) + D_{\text{KL}}(Q||P)}{2}\right).$$

B Missing Proofs of Section 4

Lemma 4.1. *Under Algorithm 4.1, we have with probability at least $1-\delta, \forall t \in [T]$,*

$$|\widetilde{\alpha}_t - \alpha_0| \leq w_t, \text{ where } w_t \text{ is given by}$$

where w_t is given by

$$w_t = \frac{2\sqrt{2(t-1)\log(2T/\delta)}}{\sum_{s=1}^{t-1}\int_0^{b_s} G(y)dy}.$$

Proof of Lemma 4.1. Taking expectation on r_s w.r.t. d_s, one has

$$\mathbb{E}\left[r_s\right] = r(v_s, b_s, \alpha_0).$$

Then by applying Azuma–Hoeffding inequality, with probability at least $1-\delta/T$,

$$\left|\sum_{s=1}^{t-1}(r_s - r(v_s, b_s, \alpha_0))\right| \le \sqrt{2(t-1)\log(2T/\delta)}.$$

Then by (3) in Algorithm 4.1,

$$\begin{aligned}\left|\sum_{s=1}^{t-1}(\widetilde{\alpha}_t-\alpha_0)\int_0^{b_s}G(y)\,dy\right| &\le \left|\sum_{s=1}^{t-1}(r_s - r(v_s,b_s,\widetilde{\alpha}_t))\right| + \left|\sum_{s=1}^{t-1}(r_s - r(v_s,b_s,\alpha_0))\right|\\ &\le 2\left|\sum_{s=1}^{t-1}(r_s - r(v_s,b_s,\alpha_0))\right|\\ &\le 2\sqrt{2(t-1)\log(2T/\delta)}.\end{aligned}$$

Then the proof can be concluded by applying a union bound. □

Theorem 4.3. *Suppose that Assumption 4.2 holds. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and bandit feedback, there exists a bidding algorithm (Algorithm 4.1) that achieves an $O(\log^2 T)$ regret, and any algorithm must incur at least a constant regret.*

Proof of Theorem 4.3. First, we show that the loss incurred by choosing such a bid b_t is small with probability at least $1-\delta$. For all $t \ge 2$,

$$\begin{aligned} r(v_t,b_t,\alpha_0) &= r(v_t,b_t,\widetilde{\alpha}_t) - (\widetilde{\alpha}_t-\alpha_0)\int_0^{b_t}G(y)\,dy,\\ &\overset{(a)}{\ge} r(v_t,b_t^*,\widetilde{\alpha}_t) - (\widetilde{\alpha}_t-\alpha_0)\int_0^{b_t}G(y)\,dy,\\ &= r(v_t,b_t^*,\alpha_0) + (\widetilde{\alpha}_t-\alpha_0)\int_{b_t}^{b_t^*}G(y)\,dy,\\ &\overset{(b)}{\ge} r(v_t,b_t^*,\alpha_0) - \frac{1}{B_1}(\widetilde{\alpha}_t-\alpha_0)^2,\\ &\overset{(c)}{\ge} r(v_t,b_t^*,\alpha_0) - \frac{1}{B_1}w_t^2, \end{aligned} \tag{27}$$

where (a) holds by the choice of b_t and (b) holds since $b_t^* = b^*(v_t,\alpha_0)$, $b_t = b^*(v_t,\widetilde{\alpha}_t)$ together with Lemma B.1; (c) follows from Lemma 4.1.

Next, we have

$$\begin{aligned}\sum_{s=1}^{t-1}\int_0^{b_s}G(y)dy &= \sum_{s=1}^{t-1}\int_0^{b^*(v_s,\widetilde{\alpha}_s)}G(y)\,dy.\\ &\overset{(a)}{\ge}\sum_{s=1}^{t-1}\int_0^{b^*(v_s,0)}G(y)\,dy.\\ &\overset{(b)}{\ge}(t-1)\cdot\underbrace{\mathbb{E}_v\left[\int_0^{b^*(v,0)}G(y)\,dy\right]}_{C_0} - \sqrt{2(t-1)\log(T/\delta)},\end{aligned}$$

where (a) holds by Lemma 3.4 or Lemma B.1 and (b) holds with probability at least $1-\delta/T$ by applying Azuma–Hoeffding inequality.

By Lemma B.2, If $C_0 = 0$, no algorithm can obtain positive expected reward under regardless of the seller's credibility α_0, which also implies any algorithm that guarantees $b_t \le v_t$ can get a zero regret. Therefore, we can presume C_0 is a positive constant depending on the known distribution G. Then for some constant $C_1 \in (0, C_0)$, when $t > t_0 = \frac{2\log(T/\delta)}{(C_0-C_1)^2}$, we have

$$\begin{aligned}
w_t &\leq \frac{2\sqrt{2(t-1)\log(2T/\delta)}}{C_0(t-1)-\sqrt{2(t-1)\log(T/\delta)}} \\
&\leq \frac{2\sqrt{2\log(2T/\delta)}}{C_0\sqrt{t-1}-\sqrt{2\log(T/\delta)}} \\
&\leq \frac{2\sqrt{2\log(2T/\delta)}}{C_1\sqrt{t-1}}. \qquad (28)
\end{aligned}$$

Combining (27) and (28), we have with probability at least $1-2\delta$,

$$\begin{aligned}
\sum_{t=1}^{T} r(v_t, b_t^*, \alpha_0) - \sum_{t=1}^{T} r(v_t, b_t, \alpha_0) \leq & t_0 + \frac{1}{B_1}\sum_{t=t_0+1}^{T} w_t^2 \\
\leq & t_0 + \frac{8\log(2T/\delta)}{B_1 C_1^2}\sum_{t=t_0+1}^{T}\frac{1}{t-1} \\
\leq & \frac{2\log(T/\delta)}{(C_0-C_1)^2} + \frac{8\log(2T/\delta)}{B_1 C_1^2}\log T,
\end{aligned}$$

which is $O(\log^2 T)$ by taking $\delta \sim T^{-1}$. The proof of the regret upper bound can be finished by taking expectation.

The lower bound part trivially holds. We have for some constant c,

$$\inf_{\pi}\sup_{\alpha_0} \text{Regret}(\pi) \geq c,$$

since at least in the first round, no single bid b_t that can obtain a uniformly small instantaneous regret for all α_0. □

Corollary 4.4. *Suppose that G is continuous. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and bandit feedback, Algorithm 4.1 can achieve an $\widetilde{O}(\sqrt{T})$ regret.*

Proof of Corollary 4.4. The proof is analogous to that of Theorem 4.3. We first have

$$\begin{aligned}
r(v_t, b_t, \alpha_0) &= r(v_t, b_t, \widetilde{\alpha}_t) - (\widetilde{\alpha}_t - \alpha_0)\int_0^{b_t} G(y)\,dy, \\
&\geq r(v_t, b_t^*, \widetilde{\alpha}_t) - (\widetilde{\alpha}_t - \alpha_0)\int_0^{b_t} G(y)\,dy, \\
&= r(v_t, b_t^*, \alpha_0) + (\widetilde{\alpha}_t - \alpha_0)\int_{b_t}^{b_t^*} G(y)\,dy, \\
&\geq r(v_t, b_t^*, \alpha_0) - w_t, \qquad (29)
\end{aligned}$$

where the second inequality uses the face $\int_{b_t}^{b_t^*} G(y)\,dy \leq 1$ rather than Lemma B.1. Together the inequality (28), which still holds given that G is continuous, the proof can be finished by bounding $\sum_{t=1}^{T} w_t$. □

Lemma B.1. *Suppose that Assumption 4.2 holds. Then $b^*(v, \alpha)$ is strictly increasing in α with derivative bounded by $1/B_1$.*

Proof of Lemma B.1. First, we have

$$\frac{\partial r(v,b,\alpha)}{\partial b} = 0 \Longrightarrow 1 - (v-b)\frac{g(b)}{G(b)} = \alpha.$$

Denote $\phi(b) = 1-(v-b)\frac{g(b)}{G(b)}$. Give that Assumption 4.2 holds, we have $\phi'(b) = \log' G(b) - (v-b)\log'' G(b) \geq B_1$. Thus, $\phi(b)$ is a strictly increasing function and we have

$$\frac{\partial b^*(v,\alpha)}{\partial \alpha} = (\phi^{-1}(\alpha))' = \frac{1}{\phi'(\phi^{-1}(\alpha))} \leq \frac{1}{B_1}.$$

□

Lemma B.2. *Suppose that G is continuous. If $\mathbb{E}_v\left[\int_0^{b^*(v,0)} G(y)\,dy\right] = 0$, then for any α_0, no algorithm can obtain positive expected reward.*

Proof of Lemma B.2. First, as $b^*(v,0)$ is increasing in v by Lemma 3.4, $\int_0^{b^*(v,0)} G(y)\,dy$ is also increasing in v. Thus, if $\mathbb{E}_v\left[\int_0^{b^*(v,0)} G(y)\,dy\right] = 0$, it holds that for nearly all v,

$$\int_0^{b^*(v,0)} G(y)\,dy = 0,$$

except a set contained in $[F^{-1}(1), 1]$ on which the probability distribution F has a zero measure.

Next, by the continuity and monotonicity of G, for all $v \in [0, F^{-1}(1))$, we have $G(y) \equiv 0$ on $[0, b^*(v,0)]$. As

$$r(v, b^*(v,0), 0) = (v - b^*(v,0))G(b^*(v,0)) = 0 = r(v, v, 0),$$

we have $b^*(v,0) \geq v$ by the tie-breaking rule in the definition of $b^*(v,0)$. Thus, $G(y)$ remains zero on $[0, v]$ for all $v < F^{-1}(1)$. Again by the continuity of G, we have $G(y) \equiv 0$ on $[0, F^{-1}(1)]$ where $F^{-1}(1) = \inf\{v : F(v) = 1\}$.

For nearly all v except a zero-measure set,

$$\begin{aligned} r(v, b, \alpha_0) &= (v - b)G(b) + \alpha \int_0^b G(y)\,dy \\ &= \alpha_0 \int_v^b \alpha_0 G(y) - G(v)\,dy \\ &\leq 0, \end{aligned}$$

where the inequality holds (1) by $G(y) \equiv 0$ when $b \leq v$ and (2) by $\alpha G(y) \leq G(v)$ for $y \in [v, b]$ when $b \geq v$. Therefore, no algorithm can obtain positive expected reward under such a distribution G regardless of the seller's credibility α_0. □

Example B.3. Let $v \equiv 1$. Consider the following distribution:

$$G(y) = \begin{cases} 0 & \text{if } 0 \leq y < \frac{1}{3} \\ \frac{3}{4}y + \frac{1}{4} & \text{if } \frac{1}{3} \leq y \leq 1 \end{cases}.$$

Then for $b \in [1/3, 1]$,

$$\begin{aligned} r(v, b, \alpha_0) &= (1 - b)G(b) + \alpha_0 \int_0^b G(y)\,dy \\ &= -\frac{6 - 3\alpha_0}{8} b^2 + \frac{2 + \alpha_0}{4} b + \frac{2 - \alpha_0}{8}. \end{aligned}$$

The optimal bid is calculated by

$$b^* = \arg\max_b r(v, b, \alpha_0) = \frac{1}{3}\left(1 + \frac{2}{2/\alpha_0 - 1}\right).$$

Taking $\alpha_0 = \frac{1}{\sqrt{T}+1/2}$, we have

$$\int_0^{b^*} G(y)dy = \int_{\frac{1}{3}}^{\frac{1}{3}\cdot\left(1+\frac{1}{\sqrt{T}}\right)} \left(\frac{3}{4}y + \frac{1}{4}\right) dy = \frac{1}{24}\left(\frac{1}{T} + \frac{4}{\sqrt{T}}\right) \longrightarrow 0.$$

Thus, C_0 in the proof of Theorem 4.3 is neither a positive constant nor 0, but an infinitesimal o(1). As a result, we are not able to give an $O(\log T/\sqrt{t-1})$ upper bound for w_t.

Intuitively, it requires $b_t \to b_t^*$ for an algorithm to achieve low regret. However, when b_t is approaching $\frac{1}{3}\left(1 + \frac{1}{\sqrt{T}}\right)$, although the bidder can still win with positive probability, it becomes harder and harder for the bidder to distinguish different α_0 under bandit feedback since in the winning rounds $b_t - d_t \leq b_t - \frac{1}{3} \longrightarrow 0$.

C The Algorithm and Missing Proofs of Section 5

Proposition C.1. *Suppose that Assumption 5.3 holds. There exists positive constants l_W and h_W, such that $\forall x \in [0, 1 + W]$,*

$$\max\{|\log' G(x)|, |\log'(1 - G(x))|\} \leq h_W, \tag{30}$$

$$\min\{-\log'' G(x), -\log''(1 - G(x))\} \geq l_W. \tag{31}$$

Note that if the density function g is log-concave, then the cumulative distribution function G and $1 - G$ are both log-concave [2]. Thus, Equation (30) and Equation (31) trivially holds for the bounded interval $[0, 1 + W]$.

Lemma 5.5. *Suppose that $T_s \gg \log^2(2/\delta)$ for some $\delta > 0$, then with probability at least $1-\delta$, $\forall d \in [0, 1]$, $|\hat{G}_s(d; \alpha_0) - G(z)| \leq O(\sqrt{\log(1/\delta)/T_s})$ holds .*

Proof of Lemma 5.5. Denote $r_s = \frac{1}{2B_1\sqrt{T_s}}$. For any $d \in [0, 1+W]$, the probability that there exists a point $y \in \{d_t\}_{t\in\Gamma_s}$ such that $|y-d| \le r_s$ is at least

$$1-(1-2B_1\cdot r_s)^{T_s} = 1-\left(1-\frac{1}{\sqrt{T_s}}\right)^{T_s} \approx 1-e^{-\sqrt{T_s}} \ge 1-\delta/2.$$

Then for any d and y such that $|y-d| \le r_s$, we can decompose $|\hat{G}_s(d;\alpha_0) - G(z)|$ in the following,

$$\begin{aligned}
&|\hat{G}_s(d;\alpha_0) - G(z)| \\
&\le |\hat{G}_s(d;\alpha_0) - \hat{G}_s(y;\alpha_0)| + |\hat{G}_s(y;\alpha_0) - \mathbb{G}_s(y)| + |\mathbb{G}_s(y) - G(y)| + |G(y)-G(d)| \\
&\le B_2\cdot r_s + \frac{1}{T_s} + \sqrt{\frac{log(4/\delta)}{2T_s}} + B_2\cdot r_s \\
&= \frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{log(4/\delta)}{2T_s}} + \frac{1}{T_s}.
\end{aligned}$$

The second inequality follows from $\hat{G}_s(\cdot;\alpha_0)$ is B_2-Lipschitz, Lemma 5.4, Dvoretzky-Kiefer-Wolfowitz (DKW) inequality and G is B_2-Lipschitz. It also holds with probability at least $1-\delta/2$ since $T_s \gg log^2(2/\delta)$. □

Given Lemma 5.5, $\hat{G}_s(\cdot;\alpha_0)$ is arbitrarily close to G when T_s is sufficiently large. In addition, [13] also show $\hat{g}_s(\cdot,\alpha_0)$ is arbitrarily close to g when T_s is sufficiently large. Therefore, we can show,

Proposition C.2. *Suppose that Assumption 5.3 holds. Under Algorithm 5.1, there exists positive constants $\tilde{l}_W$ and $\tilde{h}_W$, such that both*

$$max\{|\log'\hat{G}_s(x;\alpha_0)|, |\log'(1-\hat{G}_s(x;\alpha_0))|\} \le \tilde{h}_W, \forall x \in [0,1+W], \forall s \in [S]$$

and

$$min\{-\log''\hat{G}_s(x;\alpha), -\log''(1-\hat{G}_s(x;\alpha))\} \ge \tilde{l}_W, \forall x \in [0,1+W], \forall \alpha \in [\underline{\alpha},1],, \forall s \in [S]$$

hold almost surely.

Theorem 5.6. *Suppose that Assumption 5.2 holds. For repeated non-credible second-price auctions with unknown credibility α_0, known distribution G and full feedback, there exists a bidding algorithm Algorithm 5.1 that achieves $\widetilde{O}(T^{1/2})$ if Assumption 5.3 holds. And the lower bound on regret for this problem is $\Omega(T^{1/2})$.*

Proof of Theorem 5.6. The lower bound directly follows from Lemma 3.5 since no algorithm can avoid $\Omega(\sqrt{T})$ regret even with known α_0. Then we mainly focus on the upper bound. We first rewrite the regret per round in the following way,

$$\begin{aligned}
&r(v_t, b_t^*, \alpha_0) - r(v_t, b_t, \alpha_0) \\
=&\int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)g(x)dx - \int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)g(x)dx \\
=&\int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)g(x)dx - \int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx \\
&+\int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx - \int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx \\
&+\int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx - \int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)g(x)dx \\
\le&\int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)g(x)dx - \int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx \\
&+\int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx - \int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)g(x)dx \\
=&(v_t-b_t^*)\cdot[G(b_t^*)-\hat{G}_{s-1}(b_t^*;\widetilde{\alpha}_{s-1})] - (v_t-b_t)\cdot[G(b_t)-\hat{G}_{s-1}(b_t;\widetilde{\alpha}_{s-1})] + \alpha_0\int_{b_t}^{b_t^*}[G(x)-\hat{G}_{s-1}(x;\widetilde{\alpha}_{s-1})]dx \\
\le&|G(b_t^*)-\hat{G}_{s-1}(b_t^*;\widetilde{\alpha}_{s-1})| + |G(b_t)-\hat{G}_{s-1}(b_t;\widetilde{\alpha}_{s-1})| + \sup_x|G(x)-\hat{G}_{s-1}(x;\widetilde{\alpha}_{s-1})|,
\end{aligned}$$

where the first inequality holds by the choice of b_t, i.e.,

$$\int_0^{b_t^*}(v_t-\alpha_0 x-(1-\alpha_0)b_t^*)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx \le \int_0^{b_t}(v_t-\alpha_0 x-(1-\alpha_0)b_t)\hat{g}_{s-1}(x;\widetilde{\alpha}_{s-1})dx.$$

Let Regret_s be the regret achieved in episode s. We have

$$\begin{aligned}\text{Regret}_s &\le \sum_{t\in\Gamma_s} |G(b_t^*) - \hat{G}_{s-1}(b_t^*;\widetilde{\alpha}_{s-1})| + \sum_{t\in\Gamma_s} |G(b_t) - \hat{G}_{s-1}(b_t;\widetilde{\alpha}_{s-1})| + T_s \sup_x |G(x) - \hat{G}_{s-1}(x;\widetilde{\alpha}_{s-1})| \\ &\le \frac{12AB_2(2+W)T_s}{\underline{\alpha}^2 W^2 \tilde{l}_W} + \frac{3B_2\sqrt{T_s}}{B_1} + 3\sqrt{\frac{\log(16S/\delta)T_s}{2}} + 3,\end{aligned}$$

where the second inequality holds by Lemma C.3 and Lemma C.4 (setting $\mathcal{K}_s = \frac{4A}{\underline{\alpha}^2 W^2 \tilde{l}_W}$ and $\delta := \delta/2S$). Hence, the inequality holds with probability at least $1-\delta/S$. Finally, by a union bound over S episodes and the fact that $T_s/T_{s-1} = \sqrt{T}$ and $S \le \log\log T$, we complete our proof. □

Lemma C.3. *For each episode s, we have with probability at least $1-\delta/2S$,*

$$||\widetilde{\alpha}_s - \alpha_0|| \le \frac{4A}{\underline{\alpha}^2 W^2 \tilde{l}_W},$$

where

$$A = \frac{W\tilde{h}_W}{2\overline{\beta}T_s}\cdot\left(\frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{log(16S/\delta)}{2T_s}} + \frac{1}{T_s}\right) + \frac{W\tilde{h}_W}{2\overline{\beta}}\sqrt{\frac{\log(2S/\delta)}{T_s}}.$$

Proof. For notation simplicity, we re-parameterize α by denoting $\beta = 1/\alpha$. By Assumption 5.2, β has an upper bound $\overline{\beta} = 1/\underline{\alpha}$. To compute $\hat{\beta}_s$, we minimize the following MLE loss function,

$$\begin{aligned}\mathcal{L}_s(\beta) = -\frac{1}{T_s}\sum_{t\in\Gamma_s}\Bigg[&\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} \ge p_t\right\}\log\hat{G}_s\left(b_t + \frac{W}{2\overline{\beta}}\beta;\beta\right) \\ &+ \mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} < p_t\right\}\log\left(1-\hat{G}_s\left(b_t + \frac{W}{2\overline{\beta}}\beta;\beta\right)\right)\Bigg].\end{aligned}$$

We abuse notation and denote $\varepsilon_t(\beta) = b_t + \frac{W}{2\overline{\beta}}\beta$. Observe that

$$\mathbb{E}\left[\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} \ge p_t\right\}\right] = \mathbb{E}\left[\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} \ge \alpha d_t + (1-\alpha)b_t\right\}\right] = \mathbb{E}\left[\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}}\beta \ge d_t\right\}\right] = G(\varepsilon_t(\beta)).$$

When $\hat{G}_s$ equals to the real G, $\beta_0 = 1/\alpha_0$ should minimize the MLE loss function. By the second-order Taylor theorem, we have

$$\mathcal{L}_s(\hat{\beta}_s) - \mathcal{L}_s(\beta_0) = \mathcal{L}'(\beta_0)(\hat{\beta}_s - \beta_0) + \frac{1}{2}\mathcal{L}''(\tilde{\beta})(\hat{\beta}_s - \beta_0)^2$$

for some $\tilde{\beta}$ on the line segment between β and $\hat{\beta}_s$. Given the definition of $\mathcal{L}_s(\beta;\cdot)$, we have

$$\mathcal{L}_s'(\beta) = \frac{1}{T_s}\sum_{t\in\Gamma_s}\frac{W}{2\overline{\beta}}\eta_t(\beta), \qquad \mathcal{L}_s''(\beta) = \frac{1}{T_s}\sum_{t\in\Gamma_s}\frac{W^2}{4\overline{\beta}^2}\zeta_t(\beta)$$

where $\eta_t(\beta)$ and $\zeta_t(\beta)$ are defined as follows,

$$\eta_t(\beta) = -\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} \ge p_t\right\}\log'\hat{G}_s(\varepsilon_t(\beta);\beta) - \mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} < p_t\right\}\log'(1-\hat{G}_s(\varepsilon_t(\beta);\beta)),$$

$$\zeta_t(\beta) = -\mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} \ge p_t\right\}\log''\hat{G}_s(\varepsilon_t(\beta);\beta) - \mathbb{I}\left\{b_t + \frac{W}{2\overline{\beta}} < p_t\right\}\log''(1-\hat{G}_s(\varepsilon_t(\beta);\beta)).$$

Based on our construction of the algorithm, b_t is independent with d_t. Therefore, $\varepsilon_t(\beta_0) = b_t + \frac{W}{2\overline{\beta}}\beta_0$ are independent with d_t for any $t\in\Gamma_s$, we have

$$\begin{aligned}\mathbb{E}[\eta_t(\beta_0)] &= -\frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{\hat{G}_s(\varepsilon_t(\beta_0);\beta_0)}\cdot G(\varepsilon_t(\beta_0)) + \frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{\hat{G}_s(1-\varepsilon_t(\beta_0);\beta_0)}\cdot(1-G(\varepsilon_t(\beta_0))) \\ &= [\hat{G}_s(\varepsilon_t(\beta_0);\beta_0) - G(\varepsilon_t(\beta_0)))]\cdot[\frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{\hat{G}_s(\varepsilon_t(\beta_0);\beta_0)} + \frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{1-\hat{G}_s(\varepsilon_t(\beta_0);\beta_0)}].\end{aligned}$$

Thus, by Lemma 5.5 and Proposition C.2, we have

$$|\mathbb{E}[\eta_t(\beta_0)]| = [\hat{G}_s(\varepsilon_t(\beta_0);\beta_0) - G(\varepsilon_t(\beta_0)))] \cdot [\frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{\hat{G}_s(\varepsilon_t(\beta_0);\beta_0)} + \frac{\hat{g}_s(\varepsilon_t(\beta_0);\beta_0)}{1-\hat{G}_s(\varepsilon_t(\beta_0);\beta_0)}]$$
$$\le 2\tilde{h}_W \cdot \left(\frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{log(16S/\delta)}{2T_s}} + \frac{1}{T_s}\right)$$

holds with probability at least $1-\delta/4S$. Then by Hoeffding's inequality and union bound

$$|\mathcal{L}_s'(\beta_0)| \le \frac{W}{2\overline{\beta}T_s}\sum_{t\in\Gamma_s}|\mathbb{E}[\eta_t(\beta_0)]| + \frac{W\tilde{h}_W}{2\overline{\beta}}\sqrt{\frac{\log(8S/\delta)}{T_s}}$$
$$\le \frac{W\tilde{h}_W}{2\overline{\beta}T_s}\cdot\left(\frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{log(16S/\delta)}{2T_s}} + \frac{1}{T_s}\right) + \frac{W\tilde{h}_W}{2\overline{\beta}}\sqrt{\frac{\log(8S/\delta)}{T_s}} := A$$

holds with probability at least $1-\delta/2S$. By the optimality of $\hat{\beta}_s$,

$$\mathcal{L}_s(\hat{\beta}_s) - \mathcal{L}_s(\beta_0) \le 0.$$

Invoking into $\mathcal{L}_s(\beta)$, we have

$$\frac{1}{2}\mathcal{L}_s''(\beta_0)(\hat{\beta}_s-\beta_0)^2 \le -\mathcal{L}_s'(\beta_0)(\hat{\beta}_s-\beta_0) \le A|\hat{\beta}_s-\beta_0|,$$

$$|\hat{\beta}_s-\beta_0| \le \frac{2A}{2\tilde{l}_W\cdot W^2/(4\overline{\beta}^2)} = \frac{4\overline{\beta}^2A}{W^2\tilde{l}_W}.$$

holds with probability at least $1-\delta/2S$. So we have

$$||\widetilde{\alpha}_s-\alpha_0|| \le \frac{4\overline{\beta}^2A}{W^2\tilde{l}_W} = \frac{4A}{\underline{\alpha}^2W^2\tilde{l}_W}.$$

holds with probability at least $1-\delta/2S$. □

Lemma C.4. *For any fixed $\delta>0$, suppose $T_s \gg \log^2(2/\delta)$ and conditioned on $|\hat{\beta}_s-\beta_0|\le\mathcal{K}_s$, we have for all $d\in[0,1+W]$,*

$$|\hat{G}_s(d;\hat{\beta}_s) - G(d)| \le 3B_2(2+W)\mathcal{K}_s + \frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{\log(8/\delta)}{2T_s}} + \frac{1}{T_s}$$

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

Proof. Let $\hat{\mathbb{G}}_s$ be the empirical distribution of samples $\{\hat{\beta}_s(p_t-b_t)+b_t\}_{t\in\Gamma_s}$, i.e.

$$\hat{\mathbb{G}}_s(d) = \frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{\hat{\beta}_s(p_t-b_t)+b_t\le d\} = \frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t\le d+(\beta_0-\hat{\beta}_s)(p_t-b_t)\}.$$

First, we give a uniform convergence bound for $|\hat{\mathbb{G}}_s(d)-G(d)|$. The main challenge is that we cannot directly apply DKW inequality, since $\hat{\beta}_s$ depends on d_t, $t\in\Gamma_s$. To handle this challenge, we bound the lower bound and upper bound of $\hat{\mathbb{G}}_s(d)$ separately. Since $|\hat{\beta}_s-\beta_0|\le\mathcal{K}_s$, and $b_t\le 1$, $p_t\le 1+W$, we have

$$\frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t\le d-(2+W)\mathcal{K}_s\} \le \hat{\mathbb{G}}_s(d) \le \frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t\le d+(2+W)\mathcal{K}_s\}$$

Thus, conditioned on $|\hat{\beta}_s-\beta_0|\le\mathcal{K}_s$, for any $\gamma>0$, we have

$$\mathbb{P}(\hat{\mathbb{G}}_s(d) - G(d+(2+W)\mathcal{K}_s)\le\gamma)$$
$$\ge \mathbb{P}(\frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t\le d+(2+W)\mathcal{K}_s\} - G(d+(2+W)\mathcal{K}_s)\le\gamma)$$
$$\ge 1-\mathbb{P}(\sup_d|\frac{1}{T_s}\sum_{t\in\Gamma_s}\mathbb{I}\{d_t\ge d+(2+W)\mathcal{K}_s\} - G(d+(2+W)\mathcal{K}_s)|>\gamma)$$
$$\ge 1-2exp(-2T_s\gamma^2).$$

Similarly, we have $\mathbb{P}(G(d+(2+\hat{W})\mathcal{K}_s) - \mathbb{G}_s(d) \le \gamma) \ge 1 - 2exp(-2T_s\gamma^2)$ for any $\gamma > 0$, conditioned on $|\hat{\beta}_s - \beta_0| \le \mathcal{K}_s$.
Therefore, applying a union bound and Lipschitzness of G we have,

$$|\hat{\mathbb{G}}_s(d) - G(d)| \le \sqrt{\frac{\log(8/\delta)}{2T_s}} + B_2(2+W)\mathcal{K}_s \tag{32}$$

holds with probability at least $1-\delta/2$.
Second, we apply the similar technique used in Lemma 5.5 to bound $|\hat{G}_s(d;\hat{\beta}_s) - G(d)|$. Denote $\hat{d}_t = \hat{\beta}_s(p_t - b_t) + b_t, \forall t \in \Gamma_s$. Thus, for any $\hat{d}_t$, there must exist at least one d_t s.t. $|d_t - \hat{d}_t| = |(\beta_0 - \hat{\beta}_s)(p_t - b_t)| \le (2+W)\mathcal{K}_s$. Let $r_s = \frac{1}{2B_1\sqrt{T_s}}$, then for any $d \in [0, 1+W]$, the probability that there exists a point $y \in \{\hat{d}_t\}_{t\in\Gamma_s}$ s.t. $|y - z| \le r_s + (2+W)\mathcal{K}_s$, is at least,

$$1 - (1 - 2B_1 \cdot r_s)^{T_s} = 1 - (1 - \frac{1}{\sqrt{T_s}}) \approx 1 - e^{-\sqrt{T_s}} \ge 1 - \delta/2$$

Therefore, for any $d \in [0, 1+W]$, we can decompose $\hat{G}_s(d;\hat{\beta}_s) - G(d)$ in the following,

$$\begin{aligned}&|\hat{G}_s(d;\hat{\beta}_s) - G(d)|\\ \le&|\hat{G}_s(d;\hat{\beta}_s) - \hat{G}_s(y;\hat{\beta}_s)| + |\hat{G}_s(y;\hat{\beta}_s) - \hat{\mathbb{G}}_s(y)| + |\hat{\mathbb{G}}_s(y)| - G(y)| + |G(y) - G(d)|.\end{aligned}$$

Indeed, the characterization results by Lemma 5.4 applies to samples $\hat{d}_t$. Then we have $|\hat{G}_s(y;\hat{\beta}_s) - \hat{\mathbb{G}}_s(y)| \le \frac{1}{T_s}$. . By the Lipshitzness of $\hat{G}_s(\cdot;\hat{\beta}_s)$ and G, Equation (32) and union bound, we have

$$\begin{aligned}|\hat{G}_s(d;\hat{\beta}_s) - G(d)| &\le 2B_2[r_s + (2+W)\mathcal{K}_s] + \frac{1}{T_s} + \sqrt{\frac{\log(8/\delta)}{2T_s}} + B_2(2+W)\mathcal{K}_s\\ &= 3B_2(2+W)\mathcal{K}_s + \frac{B_2}{B_1\sqrt{T_s}} + \sqrt{\frac{\log(8/\delta)}{2T_s}} + \frac{1}{T_s}\end{aligned}$$

holds with probability at least $1-\delta$ when $T_s \gg \log^2(2/\delta)$. □

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009