NO-REGRET LEARNING IN REPEATED FIRST-PRICE AUCTIONS WITH BUDGET CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently the online advertising market has exhibited a gradual shift from second-price auctions to first-price auctions. Although there has been a line of works concerning online bidding strategies in first-price auctions, it still remains open how to handle budget constraints in the problem. In the present paper, we initiate the study for a buyer with a budget to learn her online bidding strategies in repeated first-price auctions. We propose an RL-based bidding algorithm against the optimal non-anticipating strategy under stationary competition. Our algorithm obtains $\widetilde{O}(\sqrt{T})$-regret if the bids are all revealed at the end of each round, where $\widetilde{O}(\cdot)$ is a variant of the big-O that hides logarithmic factors. With the restriction that the buyer only sees the winning bid after each round, our modified algorithm obtains $\widetilde{O}(T^{\frac{7}{12}})$-regret by techniques developed from survival analysis. Our analysis extends to the more general scenario where the buyer can have any bounded instantaneous utility function with regrets of the same order. Simulation experiments show that the constant factor inside the regret bound is rather small.

1 INTRODUCTION

There has been extensive growth in the online advertising market in recent years. It was estimated that the volume of online advertising worldwide would reach 500 billion dollars in 2022 (Statista, 2021). In such a market, advertising platforms use auctions to allocate ad opportunities. Typically, each advertiser has a limited amount of capital for an advertisement campaign. Therefore, consecutive rounds of competition are *interconnected by budgets of participating advertisers*. Furthermore, each advertiser has very limited knowledge of 1) her valuation of certain keywords and 2) the competitors she is facing. There are many works being devoted to studying algorithms for *learning* strategies for optimally spending the budget in repeated *second*-price auctions (see Section 1.1).

In practice, on the other hand, we have witnessed numerous switches from second-price auctions to first-price auctions in the online advertising market. A recent remarkable example is Google AdSenses' integrated move at the end of 2021 (LLC, 2021). Earlier examples also include AppNexus, Index Exchange, and OpenX (Sluis, 2017). This industry-wide shift is due to various factors including a fairer transactional process and increased transparency. Therefore, the shift to first-price auctions brings about major importance to the following open question which is barely considered in previous works:

How should budget-constrained advertisers learn to compete in repeated first-price *auctions?*

This paper thus initiates the study of learning to bid with budget constraints in repeated first-price auctions. It has been noted that the application of first-price auctions with budgets is not limited to online advertising mentioned above. Traditional competitive environments like mussel trade in Netherlands (van Schaik et al., 2001), modern price competition, and procurement auctions (e.g. U.S. Treasury Securities auction (Chari and Weber, 1992)) are examples as well.

Challenges and contributions The challenges in this setting are two-fold.

The first challenge relates to the specific information structure of first-price auctions. In practice, it is often the case that only the highest bid is revealed to all participants (Esponda, 2008). This is known as *censored*-feedback or an informational version of *winner's curse* in literature (Capen et al., 1971).

This affects the information structure of learning since the buyer learns less information when she wins. This makes the problem challenging compared to standard contextual bandits (c.f. Section 1.1).

The second challenge is more fundamental. It is known that the strategy in first-price auctions is notoriously complex to analyze, even in the static case (Lebrun, 1996). To get an intuitive feeling of this difficulty in our problem compared to repeated second-price auctions. Let us consider the offline case where the opponents' bids are all known. Given the budget, the problem for second-price auctions can be reduced to a pure knapsack problem, where the budget is regarded as weight capacity and prices as weights. This structure enables mature techniques including duality theory to be applied to study the benchmark strategy. Pitifully in first-price auctions, since the payment depends on the buyer's own bid, the previous approach/benchmark is not directly usable. We provide a concrete example to further illustrate such difficulties.

Example 1.1. *Consider a case where the buyer's value v follows a uniform distribution on $[0.4, 1]$ and the highest bid m of her opponents' follows a uniform distribution on $[0, 0.5]$. The time horizon is T and the buyer's budget $B = 0.5T$. The first-best benchmark (an anticipating[1] strategy that knows her values and her opponents' bids) can be viewed as a knapsack problem, which is*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{v}\sim F^T\\ \boldsymbol{m}\sim G^T}}\left[\max_{b_1,\ldots,b_T}\sum_{t=1}^{T}(v_t-b_t)\mathbf{1}_{\{b_t\geq m_t\}}\right] \quad \text{subject to} \quad \sum_{t=1}^{T}\mathbf{1}_{\{b_t\geq m_t\}}b_t \leq B \quad \forall (v_t)_{t=1}^T;(m_t)_{t=1}^T,$$

where v_t is her value and m_t is her opponents' highest bid at time t. The buyer wants to determine each b_t to maximize the revenue. In hindsight, we need to pay as close to m_t as possible. Using the theory of knapsack, the utility is $T\cdot\mathbb{E}[\mathbf{1}_{\{v\geq m\}}(v-m)]^+ = 0.45T$. On the contrary, the optimal non-anticipating bidding strategy in a first-price auction is to bid $\frac{v}{2}$ and the utility is $T\cdot\mathbb{E}[\mathbf{1}_{\{\frac{v}{2}\geq m\}}\frac{v}{2}] = 0.26T$. There is already an $\Omega(T)$ separation between the first-best benchmark and the ideal case with full information.

This example shows that simple characterization of the optimum in Balseiro and Gur (2019) is not applicable to our problem. Furthermore, it remains unclear what methodology can be applied in first-price auctions with budgets. The state-of-the-art adaptive pacing strategy downgrades to truthful bidding as the budget increases, so in first-price auctions, it may result in *near-zero* reward and thus cannot have any theoretical guarantee (see (Balseiro and Gur, 2019, §2.4) for further discussions).

The present paper takes the first step to combat the challenges mentioned above with a dynamic programming approach. Correspondingly, our contribution is also two-fold:

- We provide an RL-based learning algorithm. Through the characterization of the optimal strategy, we obtain $\widetilde{O}(\sqrt{T})$-regret guarantee for the algorithm in the full-feedback case[2].
- In the censored-feedback setting, by techniques developed from survival analysis, we modify our algorithm and obtain a regret of $\widetilde{O}(T^{\frac{7}{12}})$.

1.1 Related Work

Repeated second-price auctions with budgets There is a flourishing source of literature on bidding strategies in repeated auctions with budgets. Through the lens of online learning, Balseiro and Gur (2019) identify asymptotically optimal online bidding strategies known as *pacing* (a.k.a. bid-shading in literature) in repeated second-price auctions with budgets. Inspired by the pacing strategy, Flajolet and Jaillet (2017) develop no-regret non-anticipating algorithms for learning with contextual information in repeated second-price auctions. Another line of works that uses similar techniques in the present paper includes Amin et al. (2012); Tran-Thanh et al. (2014); Gummadi et al. (2012). Gummadi et al. (2012) and Amin et al. (2012) study bidding strategies in repeated second-price auctions with budget constraints, but the former does not involve any learning and the latter does not provide any regret analysis (their estimator is biased). Tran-Thanh et al. (2014) derive regret bounds for the same scenario but the optimization objective is the number of items won instead of value or surplus. Baltaoglu et al. (2017) also use dynamic programming to tackle repeated

[1] An algorithm is anticipating if bid selection depends on future observations, see Flajolet and Jaillet (2017).

[2] This is especially practical in public-sector auctions (Chari and Weber, 1992) as regulations mandate all bids to be revealed.

second-price auctions with budgets. However, they assume per-round budget constraints and their dynamic programming algorithm is for allocating bids among multiple items. Again, we emphasize that no prior work has been done in repeated first-price auctions with budgets since the structure of the problem (compared to second-price variants) is fundamentally different (recall Example 1.1).

Repeated first-price auctions without budgets Two notable works concerning repeated first-price auctions are Han et al. (2020b) and Han et al. (2020a). Han et al. (2020b) introduce a new problem called monotone group contextual bandits and obtain an $O(\sqrt{T}\ln^2 T)$-regret algorithm for repeated first-price auctions *without* budget constraints under stationary settings. This bound is improved to $O(T^{\frac{1}{3}+\epsilon})$ by Achddou et al. (2021) with additional assumptions on distributions. Han et al. (2020a) concentrate on an adversarial setting and develop a mini-max optimal online bidding algorithm with $O(\sqrt{T}\ln T)$ regret against all Lipschitz bidding strategies. Badanidiyuru et al. (2021) consider the problem in a contextual setting. A crucial difference is that in the present paper, budgets are involved thus the algorithms from previous works are not directly suitable for our needs.

Bandit with knapsack From the bandit side, Badanidiyuru et al. (2013) develop bandit algorithms under resource constraints. They show that their algorithm can be used in dynamic procurement, dynamic posted pricing with limited supply, etc. However, since the bidder observes her value *before* bidding in our problem, results by Badanidiyuru et al. (2013) cannot be directly applied to our setting. Our setting also relates to contextual bandit problems with resource constraints (Badanidiyuru et al., 2014; Agrawal and Devanur, 2016; Agrawal et al., 2016). Nevertheless, applying this contextual bandit approach requires discretizing the action space, which needs Lipschitz continuity of distributions. Our approach does not rely on any continuity assumption. Further, the performance guarantee (typically $\widetilde{O}(T^{\frac{2}{3}})$) is worse than ours. It also does not fit into our information structure when the feedback is censored.

2 PRELIMINARIES

Auction mechanism We consider a repeated first-price auction with budgets. Specifically, we suppose that the buyer has a limited budget B to spend in a time horizon of $T \leq +\infty$ (can be *unknown* to her) rounds. At the beginning of each round $t = 1, 2, \ldots, T$, the bidder privately observes a value v_t for a fresh copy of item and bids b_t according to her past observations $\boldsymbol{h}_t$ and value v_t. Denote the strategy she employs as $\pi\colon (v_t, B_t, \boldsymbol{h}_t) \to b_t$, which maps her current budget B_t, value v_t and past history $\boldsymbol{h}_t$ to a bid. Let m_t be the maximum bid of the other bidders. Since the auction is a *first* price auction, if b_t is larger than m_t, then the buyer wins the auction, is charged b_t, and obtains a utility of $v_t - b_t$; otherwise, she loses and the utility is 0. Therefore, the instantaneous utility of the buyer is

$$r_t = (v_t - b_t)\mathbf{1}_{\{b_t \geq m_t\}}.$$

The exact information structure of history the buyer observes is dictated by how the mechanism reveals m_t. In full generality, we assume that the feedback is censored, i.e. only the highest bid is revealed at the end of each round and the winner does not observe m_t exactly. This is considered to be an informational version of *winner's curse* (Capen et al., 1971) and is of practical interest (Esponda, 2008). For the purpose of modeling, we suppose that ties are broken in favor of the buyer but this choice is arbitrary and by no means a limitation of our approach.

Next, we state the assumptions on m_t and v_t. Without loss of generality, we assume that b_t, m_t, v_t are normalized to be in $[0, 1]$. In the present paper, we consider a stochastic setting where m_t and v_t are drawn from some distributions F, G *unknown* to the buyer, respectively, and independent from the past. We will also refer to the cumulative distribution functions of F, G with the same notations. No further assumptions will be made on F, G. Now, the expected instantaneous utility of the buyer at time t with strategy π is

$$R^\pi(v_t, b_t) = \mathop{\mathbb{E}}_{m_t \sim F}[r_t] = (v_t - b_t)F(b_t).$$

To argue for the reasonability of this assumption, note that although other buyers may also involve learning behavior, it is typical that in a real advertising market, there are a large number of buyers (Kahng et al., 2004). The specific buyer only faces a different small portion of them and is completely oblivious of whom she is facing in each round. Therefore, the buyer's sole objective is to maximize her utility (instead of fooling other buyers) and by the law of large numbers, the price m_t and value v_t the buyer observes are independent and identically distributed at least for a period of time[3].

Buyer's target and regret The buyer aims at maximizing her long-term accumulated utility subject to budget constraints. Recall that the instantaneous utility of the buyer is $r_t = (v_t - b_t)\mathbf{1}_{\{b_t \geq m_t\}}$. The payment is $c_t = b_t\mathbf{1}_{\{b_t \geq m_t\}}$ and the budget will then decrease accordingly as the payment incurs. She can continue to bid as long as the budget has not run out but must stop at

$$\tau^* = \min\left\{T+1, \min\left\{t \in \mathbb{N} : \sum_{\tau=1}^{t} c_\tau = B\right\} + 1\right\}.$$

The buyer's problem now becomes determining how much to bid in each round to maximize her accumulated utility. In line with works Gummadi et al. (2012); Golrezaei et al. (2019); Deng and Zhang (2021), the buyer adopts a discount factor $\lambda \in (0, 1)$. She takes discounts since she does not know T or τ^* — Discount factors can be interpreted to be the estimate of the probability that the repeated auction will last for at least t rounds (Devanur et al., 2014; Drutsa, 2018). It means that the process will terminate at each round with probability $1 - \lambda$ (Uehara et al., 2021). On the economic side, in important real-world markets like online advertising platforms, buyers are impatient for opportunities since companies of different sizes have different capabilities. Discount factors model how impatient[4] the buyer is in (Drutsa, 2017; Vanunts and Drutsa, 2019). Now the buyer's optimization problem is to determine a *non-anticipating* strategy π for the following:

$$\max_{\pi} \quad \mathop{\mathbb{E}}_{\substack{\boldsymbol{v}\sim F^T \\ \boldsymbol{m}\sim G^T}}\left[\sum_{t=1}^{T} \lambda^{t-1} r_t\right] \quad \text{subject to} \quad \sum_{t=1}^{T} \mathbf{1}_{\{b_t \geq m_t\}} b_t \leq B \quad \forall (v_t)_{t=1}^T; (m_t)_{t=1}^T,$$

where $b_t = \pi(v_t, B_t, \boldsymbol{h}_t)$. Here, $\boldsymbol{v} \coloneqq (v_1, \ldots, v_T)$ denotes the sequence of private values the buyer observes, and $\boldsymbol{m} \coloneqq (m_1, \ldots, m_T)$ is the sequence of the highest bids of the other bidders. $V^\pi(B, t)$ denotes the expected accumulated utility using strategy π with budget B and starting from time t. Let π^* denote the optimal bidding strategy with the knowledge of the underlying distributions F and G. The corresponding expected accumulated utility is $V^{\pi^*}(B, t)$. (We sometimes use $V(\cdot, \cdot)$ to represent $V^{\pi^*}(\cdot, \cdot)$ for convenience in the rest of the paper.)

We now come to define the regret. First, write the per-episode revenue suboptimality for each round t as

$$\text{SubOpt}_t(\pi_t) = V^{\pi^*}(B_t, t) - V^{\pi_t}(B_t, t),$$

where π_t is the strategy used in round t. Our evaluation metric is then the sum of suboptimality for $t = 1, \ldots, T$, namely

$$\text{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \text{SubOpt}_t(\pi_t)\right], \tag{1}$$

where the expectation is taken over the trajectories of the achievement of $\boldsymbol{v}$ and the realization of others' bids inexplicitly.

The definition of regret comes from traditional reinforcement learning (RL) literature of infinite-horizon discounted model (Kaelbling et al., 1996). The definition is also inspired by the recent advances in Yang et al. (2021); He et al. (2021); Liu and Su (2020); Zhou et al. (2021). Zhou et al. (2021) call it cumulative error. It reflects the suboptimality for π_t to learn the optimal valuation of attending the auction.

In the most common scenario, such as Balseiro and Gur (2019), the budget constraint is linear to time horizon T, i.e. $\frac{B}{T} \sim \Theta(1)$. Therefore, a bidder has an expectation that she will win for $O(T)$ rounds.

[3]This assumption has support from experimental evidence (Pin and Key, 2011). It can also be theoretically justified using mean field asymptotics. Please also see Han et al. (2020b) for another justification.

[4]As an additional explanation, in budget-constrained first-price auctions, the bidder always bids below or equal to her value. So she is very sensitive to the market price. However, by not winning the auction at a certain price, the bidder creates a future opportunity to win an equivalent auction at a lower price. The use of a bid discount factor adds flexibility to tune this behavior. As the bidder has a preference for present utility over future utility, the discount factor moderates the extent of underbidding that she finds to be optimal, which makes the model more general.

With a sub-optimal policy, it is easy to suffer $O(T)$ regret which is intolerable for bidders. It leads to the challenge to achieve a sublinear regret in first-price auctions and we design algorithms to answer the question.

3 Bidding Algorithm and Analysis

In this section, we present our bidding algorithm and the high-level ideas in the analysis of regret. We first consider the case where the feedback is not censored, i.e. the buyer is aware of m_t no matter whether she wins or not. Then we extend our algorithm to the case where the feedback is censored with techniques developed from survival analysis.

3.1 Full Feedback

When F and G are known, the buyer's problem can be viewed as offline. The technical challenge lies in the observation that even when the distributions are known, the buyer's problem cannot be directly analyzed as a knapsack problem. To tackle this challenge, we use a dynamic programming approach to solve the problem. In particular, the optimal strategy π^* satisfies the following Bellman equation:

$$b^*(B_\tau, v) \in \arg\max_b [(v-b) + \lambda V(B_\tau - b, \tau+1)]F(b) + \lambda V(B_\tau, \tau+1)(1-F(b)),$$
$$V(B_\tau, \tau) = \mathop{\mathbb{E}}_v [(v-b^*) + \lambda V(B_\tau - b^*, \tau+1)]F(b^*) + \lambda V(B_\tau, \tau+1)(1-F(b^*)),$$

for all $\tau \in \mathbb{N}$ and $0 \le B_\tau \le B$. Note that for any $B_\tau < 0$, $V(B_\tau, \tau) = -\infty$. By choosing appropriate initialization conditions, we can solve the equation recursively and obtain the optimal bidding strategy together with the function $V(\cdot,\cdot)$. The above recursion also characterizes the optimal solution, which will be used in the analysis later.

When the buyer does not have the information of F and G, she can learn the distributions from past observations. Therefore, it is natural to maintain estimations $\hat{F}$ and $\hat{G}$ of the true distributions. Our algorithm for the full-feedback case is now depicted in Algorithm 1. To ease technical loads, we first assume the knowledge of G and only estimate F in Algorithm 1. Later, we will add the estimation of G and its analysis is presented in Theorem 3.2.

Algorithm 1 Algorithm for the full-feedback case

1: **Input**: Initial budget B and constant C_1 $\triangleright$ C_1 is an arbitrary positive constant
2: Initialize the estimation $\hat{F}$ of F to a uniform distribution over $[0,1]$ and $B_1 \leftarrow B$
3: **for** $t = 1, 2, \ldots$ **do**
4: Observe the value v_t in round t
5: Let t_0 be the smallest integer that satisfies $\lambda^{t_0 - t}\frac{1}{1-\lambda} < \frac{C_1}{\sqrt{t}}$
6: Set $V_{\hat{F}}(B_{t_0}, t_0) = 0$ for any B_{t_0} $\triangleright$ $V_{\hat{F}}$ is algorithm's estimation of V
7: **for** $\tau = t_0, t_0 - 1, \ldots, t$ **do**
8: $Q_{v,\hat{F}}(B_\tau, \tau, b) \leftarrow [(v-b) + \lambda V_{\hat{F}}(B_\tau - b, \tau+1)]\hat{F}(b) + \lambda V_{\hat{F}}(B_\tau, \tau+1)(1-\hat{F}(b))$
9: Solve the optimization problem $\hat{b}^*_\tau \leftarrow \arg\max_b Q_{v,\hat{F}}(B_\tau, \tau, b)$
10: $V_{\hat{F}}(B_\tau, \tau) \leftarrow \mathbb{E}_{v\sim G}[Q_{v,\hat{F}}(B_\tau, \tau, \hat{b}^*_\tau)]$
11: **end for**
12: Place a bid $\hat{b}_t \leftarrow \arg\max_b Q_{v,\hat{F}}(B_t, t, b)$
13: Observe m_t, c_t and r_t from this round of auction and update $\hat{F}(x) = \frac{1}{t}\sum_{i=1}^t \mathbf{1}_{\{m_i \le x\}}$.
14: $B_{t+1} \leftarrow B_t - c_t$. If $B_{t+1} \le 0$ then halt.
15: **end for**

Similar to prior work (Amin et al., 2012), Algorithm 1 performs exploration and exploitation simultaneously. The buyer initializes the estimation of F to a uniform distribution (Line 2). At round t, the buyer observes her valuation. Then, she uses her estimation of F to solve the dynamic programming problem recursively[5] to obtain an estimation of the optimal bid (Line 7~Line 11). To

[5]For the non-trivial case $B \le T$, this can be solved in $O\left(\frac{T^{4.5}}{(1-\lambda)^6}\right)$ time with only $O(T^{-\frac{1}{2}})$ loss (see, e.g. Chow and Tsitsiklis, 1989).

provide a base case for recursion, note that for sufficiently large $t_0 \gg t$, $V_{\hat{F}}(\cdot, t_0)$'s impact to $V_{\hat{F}}(\cdot, t)$ is negligible due to the discount $\lambda^{t_0 - t}$. Therefore, the buyer can approximate $V_{\hat{F}}(\cdot, t_0)$ with zero for t_0 (Line 5). Finally, the auction proceeds with m_t, r_t, c_t revealed and the buyer updates her information accordingly (Line 13~Line 14).

Analysis of regret To analyze the algorithm, we first assume that Algorithm 1 knows the distribution G exactly and establishes the regret. Then we add the contribution of the estimation of F.

Theorem 3.1. *Under the circumstance that F is unknown, the worst-case regret of Algorithm 1 is $\widetilde{O}(\sqrt{T})$, where the regret is computed according to Equation* (1). *Explicitly, if we take $C_1 = 1$,*

$$\textit{Regret}(T) \leq \left(4\sqrt{\ln(\sqrt{2}T)}\frac{1+\lambda}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T} + 1.$$

To show an example of the application of the result, let us take the budget B to scale linearly with T as in Balseiro and Gur (2019); Flajolet and Jaillet (2017). Specifically, assume that $T < +\infty$ and there exists a constant β such that the budget $B = \beta T$, then we establish that the regret is $\widetilde{O}(\sqrt{T})$ in this special case. Indeed, under this condition, we can simply set $t_0 = T + 1$ and $V_{\hat{F}}(B_{T+1}, T+1) = 0$ for any B_{T+1} in Algorithm 1. Therefore, $C_1 = 0$ and the worst-case regret is bounded by $\left(4\sqrt{\ln(\sqrt{2}T)}\frac{1+\lambda}{(1-\lambda)^3}\right)\sqrt{T} + 1$.

Next, we deal with the case where G is also initially unknown. Based on Algorithm 1, we additionally maintain an estimation $\hat{G}$ of G based on past observations of valuations. $\hat{G}$ is initialized to be a uniform distribution and will be used to solve the dynamic programming problem (see Line 7 of Algorithm 2). Using similar techniques as before (with more work), we obtain the following theorem.

Theorem 3.2. *Under the circumstance that F, G are both unknown, it holds that the worst-case regret of Algorithm 1 using empirical distribution functions to estimate F and G is $\widetilde{O}(\sqrt{T})$. Explicitly, if we take $C_1 = 1$,*

$$\textit{Regret}(T) \leq \left(\sqrt{\ln(2T)}\frac{6(1+\lambda)}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T} + 1.$$

3.2 Censored Feedback

In this subsection, we deal with the case that the buyer can only see the winner's bid after each round. In other words, the feedback is left-censored. Concretely, the buyer's observation is

$$o_t = \max\{b_t, m_t\}.$$

When she wins, the exact value of m_t is not revealed. The buyer only knows that m_t is smaller than her bid in the current round. To estimate the distribution of m_t, there is a classical statistics (KM estimator) developed by Kaplan and Meier (1958) for the estimation of F in this scenario. However, the KM estimator assumes the sequence $(m_t)_{t=1}^{T}$ is deterministic, which does not fit our needs. Although Suzukawa (2004)'s extension allows random censorship, it requires independence between b_t and m_t, which is not realistic since we use the estimated distribution to place bids.

To tackle this problem, we first introduce an estimator proposed by Zeng (2004) denoted by $\hat{F}_n$ to take place of the previous empirical distribution used in Algorithm 1.

Estimation procedure We now present the procedure for estimating F under censored feedback. It suffices to estimate the distribution function of $1 - m_t$ which is right-censored by $1 - b_t$. Let $y_t = \min\{1 - m_t, 1 - b_t\}, r_t = \mathbf{1}_{\{m_t \geq b_t\}}$. The observations can now be described as $(y_t, r_t, \boldsymbol{h}_t)_{t=1}^{T}$.

Roughly speaking, to decouple the dependency between m_t, b_t, we use the fact that b_t and m_t are independent conditioning on $\boldsymbol{h}_t$. Intuitively, the history $\boldsymbol{h}_t$ provides information for getting enough effective samples for m_t. Next, we establish models to estimate the hazard rate functions[6] of $1 - m_t, 1 - b_t$ using $\boldsymbol{h}_t$. With the hazard rate functions, we use the maximum likelihood method with a kernel to compute the final estimation $\hat{F}_t$ and obtain Equation (3).

[6]The hazard rate function of a random variable X with p.d.f. f and c.d.f. F is $H_X(x) = \frac{f(x)}{1-F(x)}$.

Details follow. We initialize with a sequence (bandwidth) $(a_t)_{t=1}^T$ such that $\frac{\log^2 a_t}{ta_t^2} \to 0, ta_t^2 \to \infty, ta_t^4 \to 0$ as $t \to +\infty$ and a symmetric kernel function $K(\cdot,\cdot) \in C^2(\mathbb{R}^2)$ with bounded gradient. Now, at each time t, we compute two vectors $\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t$ which maximize each of the following likelihood functions (these can be regarded as loss functions of the estimation that we aim to optimize)

$$f(\boldsymbol{\beta}) = \sum_{\tau=1}^{t} \frac{r_\tau}{t} \left(\boldsymbol{\beta}^\top \boldsymbol{h}_\tau - \log \sum_{y_i \geq y_\tau} e^{\boldsymbol{\beta}^\top \boldsymbol{h}_\tau} \right), g(\boldsymbol{\gamma}) = \sum_{\tau=1}^{t} \frac{1-r_\tau}{t} \left(\boldsymbol{\gamma}^\top \boldsymbol{h}_\tau - \log \sum_{y_i \geq y_\tau} e^{\boldsymbol{\gamma}^\top \boldsymbol{h}_\tau} \right). \tag{2}$$

We arbitrarily pad $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_t$ with zeros to make their length the same (we will show that this is without loss of generality in a moment). Compute $\boldsymbol{Z}_t = (\boldsymbol{\beta}_t^\top \boldsymbol{h}_t, \boldsymbol{\gamma}_t^\top \boldsymbol{h}_t)^\top$. The survival function of $1 - m_t$, or equivalently the cumulative distribution function of m_t, is estimated based on Zeng (2004)'s estimator

$$\hat{F}_t(x) = \frac{1}{t} \sum_{i=1}^{t} \prod_{j=1}^{t} \left(1 - \frac{K((\boldsymbol{Z}_i - \boldsymbol{Z}_j)/a_n) \mathbf{1}_{\{y_j \leq x\}} r_j}{\sum_{m=1}^{n} K((\boldsymbol{Z}_i - \boldsymbol{Z}_m)/a_n) \mathbf{1}_{\{y_j \leq y_m\}}} \right). \tag{3}$$

Now, we are ready to apply the estimator to design the algorithm for the censored-feedback case. Note that the new estimator's convergence rate is slower than that for the full-feedback case. Therefore, compared to Algorithm 1, Algorithm 2 is now a multi-phase algorithm. The algorithm only updates the estimation of $\hat{F}$ at the end of each phase (see Figure 1 for an illustration). The other elements of each phase in Algorithm 2 are similar to Algorithm 1.

Algorithm 2 Algorithm for the censored-feedback case

Input: Initial budget B and constant C_1 ▷ C_1 is an arbitrary positive constant
Initialize the estimation $\hat{F}$ of F and the estimation $\hat{G}$ of G to uniform distributions over $[0,1]$
$B_1 \leftarrow B$
for Phase $i = 1, 2, \ldots$ **do** ▷ Phase i $(i > 1)$ lasts for 2^i rounds. Phase 1 lasts for 2 rounds
  for each t in the time interval of round i **do**
    Observe the value v_t in round t
    Update $\hat{G}(x) = \frac{1}{t}\sum_{i=1}^{t} \mathbf{1}_{\{v_i \leq x\}}$.
    Let t_0 be the smallest integer that satisfies $\lambda^{t_0 - t}\frac{1}{1-\lambda} < \frac{C_1}{\sqrt{t}}$
    Set $V_{\hat{F},\hat{G}}(B_{t_0}, t_0) = 0$ for any B_{t_0} ▷ $V_{\hat{F},\hat{G}}$ is algorithm's estimation of V
    for $\tau = t_0, t_0 - 1, \ldots, t$ **do** ▷ This loop can be moved to the end of each phase to reduce the invocation time from T to $\ln T$
      $Q_{\hat{F},\hat{G}}(B_\tau, \tau, b) \leftarrow [(v - b) + \lambda V_{\hat{F},\hat{G}}(B_\tau - b, \tau + 1)]\hat{F}(b) + \lambda V_{\hat{F},\hat{G}}(B_\tau, \tau + 1)(1 - \hat{F}(b))$
      Solve the optimization problem $\hat{b}_\tau^* \leftarrow \arg\max_b Q_{\hat{F},\hat{G}}(B_\tau, \tau, b)$
      $V_{\hat{F},\hat{G}}(B_\tau, \tau) \leftarrow \mathbb{E}_{v \sim G}[Q_{\hat{F},\hat{G}}(B_\tau, \tau, \hat{b}_\tau^*)]$
    end for
    Place a bid $\hat{b}_t \leftarrow \arg\max_b Q_{\hat{F},\hat{G}}(B_t, t, b)$
    Observe o_t, c_t and r_t from this round of auction
    $B_{t+1} \leftarrow B_t - c_t$. If $B_{t+1} \leq 0$ then halt.
  end for
  Update $\hat{F}$ using the estimator specified in Equation (3) with data observed before this phase
end for

Analysis of regret To analyze the performance of Algorithm 2, we will prove a series of lemmas on the estimation error of Equation (3). We concentrate on the performance of the new estimator since this is the major difference between Algorithm 1 and Algorithm 2. In particular, our proof relies on the following convergence result.

Lemma 3.3 (Zeng)**.** *Let $\hat{F}_n$ be the estimation of F after using n observations. We have*

$$\sqrt{n}(\hat{F}_n(1-x) - F(1-x)) \Longrightarrow \mathcal{B}(x) \quad \text{in} \quad \ell^\infty([0,1]),$$

where $\mathcal{B}(x)$ is a Gaussian process.

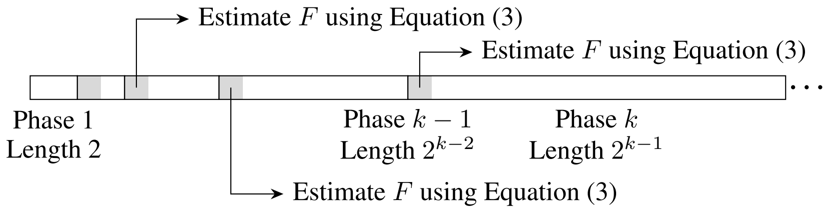

Figure 1: Schematic representation of the phases in Algorithm 2. Algorithm 2 updates its estimates of F at the end of each phase.

Before we proceed to apply the lemma, we verify a series of prerequisites mentioned in Zeng (2004) to make sure it holds. First, we make sure that conditioning on $\boldsymbol{h}_t$, the random variables $1-m_t$ and $1-b_t$ are independent. Indeed, b_t is completely decided by $\boldsymbol{h}_t$ and m_t is independent of $\boldsymbol{h}_t$. Second, we note that the maximizer shown in Equation (2) is essentially doing Cox's proportional hazards regression analysis. To establish consistency of the estimator, we show that at least one of $\widetilde{m}_t \coloneqq 1-m_t$ and $1-b_t$ follows Cox's proportional model. That is to say, there exists $\boldsymbol{\beta}$ and a function $f(y)$ such that the hazard rate function of $\widetilde{m}_t$ or $1-b_t$ conditioning on $\boldsymbol{h}_t$ exactly follows

$$H(y \mid \boldsymbol{h}_t) = f(y)e^{\boldsymbol{\beta}^\top \boldsymbol{h}_t}. \tag{4}$$

Equation (4) holds for $\widetilde{m}_t$. In fact, taking $\boldsymbol{\beta}=\mathbf{0}$ and $f(y)=\frac{F'(1-y)}{1-F(1-y)}$ suffices. Since we take $\boldsymbol{\beta}=\mathbf{0}$, consistency establishes regardless of the way we pad $\boldsymbol{h}_t$.

Next, consider some phase at time $2^n \le t \le 2^{n+1}-1$. The estimation $\hat{F}_n$ is computed using the first 2^n observed data points. Applying similar techniques for the rate of convergence of empirical process (Norvaiša and Paulauskas, 1991), we obtain the following lemma on the performance of $\hat{F}$ in Algorithm 2.

Lemma 3.4. *Under the update process in Algorithm 2, for any $2^n \le t \le 2^{n+1}-1$, we have the following bounds for the estimation $\hat{F}_n$:*

$$|\Pr(\sup_b |\sqrt{2^n}(\hat{F}_n(1-b)-F(1-b))| \ge r) - \Pr(\sup_b |\mathcal{B}(1-b)| \ge r)| \le M(1+r)^{-3}\ln^2(t)\cdot t^{-\frac{1}{6}},$$

where M is a constant depending only on F and Algorithm 2.

Finally, we now bound the difference between $\hat{F}_n$ and F with the help of Lemma 3.4.

Lemma 3.5. *Recall that we use the first 2^n data points to estimate $\hat{F}$. Under the update procedure of Algorithm 2, for any $2^n \le t \le 2^{n+1}-1$, with probability at least $1-T^{-\frac{5}{12}}/(2\ln T)$*

$$\sup_x |\hat{F}(x)-F(x)| \le \sqrt{2}C_5(4M\ln^3 T)^{\frac{1}{3}} t^{-\frac{5}{9}} T^{\frac{5}{36}},$$

where C_5 is an absolute constant.

With Lemma 3.5 in hand, we now have

Theorem 3.6. *Under the circumstance that F, G are both unknown and the feedback is censored, the worst-case regret of Algorithm 2 is $\widetilde{O}(T^{\frac{7}{12}})$. Explicitly, if we take $C_1=1$,*

$$\textit{Regret}(T) \le \left(\frac{9\sqrt{2}(1+\lambda)}{2(1-\lambda)^3}C_5(4M\ln^3 T)^{\frac{1}{3}}+1\right)T^{\frac{7}{12}} + \left(\sqrt{\frac{1}{2}\ln\left(4T^{\frac{17}{12}}\right)}\frac{2(1+\lambda)}{(1-\lambda)^3}+\frac{4}{1-\lambda}\right)\sqrt{T}.$$

Remark 3.7. *The setting in Han et al. (2020b) is a special case of ours, where there are no budget constraints and $\lambda=0$ (thus removing the $\frac{1}{1-\lambda}$ factor in our results). The buyer's aim is to maximize $(v-b)F(b)$ each round. This is equivalent to $V_{\hat{F}}=0$ in our setting with no need to estimate G, yielding regret $\widetilde{O}(\sqrt{T})$ in the full-feedback case and regret $\widetilde{O}(T^{\frac{7}{12}})$ in the censored-feedback case.*

Remark 3.8. *The regret-bound looks unusual at a first glance. The reason is that the convergence rate of the estimator is lower than that in the commonly used "Hoeffding-type" or "Bernstein-type" inequalities. However, due to the information structure, they are not suitable to be used in our environment to our best knowledge.*

4 LOWER BOUND

Here, we discuss the lower bound for the regret under such settings. This will shed light on the optimality gap of the proposed policies.

Full Feedback We have proposed a general solution framework that works for any β where the budget constraint is $B = \beta T$ and any discount rate $\lambda \in [0, 1)$. Note that the general lower bound is no less than the lower bound for a specific case. Consider the case when $\beta = 1$ and $\lambda = 0$. Our problem reduces to the case where the buyer essentially does not face budget constraints and is extremely myopic. Under this circumstance, the problem is a multi-armed bandit problem. Auer et al. (2002) shows that it suffers from a $\Theta(\sqrt{T})$-regret lower bound. This means that our algorithm is *optimal* up to logarithmic terms.

Censored Feedback The $\widetilde{O}(\sqrt{T})$ lower bound also applies here. However, the upper bound and the lower bound have not matched yet. We leave this as an intriguing open problem as there is a lack of relevant literature to show regret lower bounds under the censored-feedback case. However, we want to provide some evidence that our upper bound is sufficiently good. For example, parallel to our work, Gaitonde et al. (2022) extend the pacing techniques to a class of auction forms including first-price auctions. They obtain an $\widetilde{O}(T^{\frac{3}{4}})$-regret bounds against *the best linear policy* under the *value-maximization* objective. Under a censored-feedback information structure with contextual valuations, Cesa-Bianchi et al. (2017) show an $\widetilde{O}(T^{\frac{d-1/3}{d+2/3}})$-regret upper bound without budget constraints where d is the dimension of the context. And a similar information and payment structure in Bayesian persuation yield an $\widetilde{O}(T^{\frac{4}{5}})$ regret bound (Castiglioni et al., 2020).

5 DISCUSSION AND CONCLUSION

In this paper, we develop a learning algorithm to adaptively bid in repeated first-price auctions with budgets. On the theoretical side, our algorithm, together with its analysis of $\widetilde{O}(\sqrt{T})$-regret in the full-feedback case and $\widetilde{O}(T^{\frac{7}{12}})$-regret in the censored-feedback case, takes the first step in understanding the problem. On the practical side, our algorithm is simple and readily applicable to the digital world that has shifted to first-price auctions[7].

Questions raise themselves for future explorations. We observe here that in the limiting case $\lambda \to 1$, the optimal bidding strategy in Algorithm 2 is similar to a *pacing* strategy, which relates to the open question[8] raised in Balseiro and Gur (2019). In the limiting case of $\lambda \to 1$, the optimal bid of Algorithm 2 is of the form $\frac{v_t}{1+x_t}$, where x_t is a pacing multiplier that depends only on B_t and F and can be computed without solving the dynamic programming problem. This observation can be viewed as a corollary of (Theorem 3.1 Gummadi et al., 2012). This connection between Algorithm 2 and pacing suggests further investigations. Other immediate open questions include closing the gap between upper and lower bounds for the censored feedback case.

REFERENCES

Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Fast rate learning in stochastic first price bidding. In *Asian Conference on Machine Learning*, pages 1754–1769. PMLR, 2021.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18. PMLR, 2016.

[7]For simulation experiments, please refer to Appendix D.

[8]"A question of theoretical and practical interest is how to extend the adaptive pacing approach that is suggested in this paper to nontruthful mechanisms such as first-price auctions."

Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget optimization for sponsored search: Censored learning in MDPs. *arXiv preprint arXiv:1210.4847*, 2012.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. Learning to bid in contextual first price auctions. *arXiv preprint arXiv:2109.03173*, 2021.

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

M. Sevi Baltaoglu, Lang Tong, and Qing Zhao. Online learning of optimal bidding strategy in repeated multi-commodity auctions. *Advances in Neural Information Processing Systems*, 30, 2017.

Edward C. Capen, Robert V. Clapp, and William M. Campbell. Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, 23(06):641–653, 1971.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, pages 465–481. PMLR, 2017.

V. Chari and Robert Weber. How the us treasury should auction its debt. *Quarterly Review*, 16(4), 1992.

Chef-Seng Chow and John N Tsitsiklis. The complexity of dynamic programming. *Journal of complexity*, 5(4):466–488, 1989.

Yuan Deng and Hanrui Zhang. Prior-independent dynamic auctions for a value-maximizing buyer. *Advances in Neural Information Processing Systems*, 34, 2021.

Nikhil R Devanur, Yuval Peres, and Balasubramanian Sivan. Perfect bayesian equilibria in repeated sales. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 983–1002. SIAM, 2014.

Alexey Drutsa. Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In *Proceedings of the 26th International Conference on World Wide Web*, pages 33–42, 2017.

Alexey Drutsa. Weakly consistent optimal pricing algorithms in repeated posted-price auctions with strategic buyer. In *International Conference on Machine Learning*, pages 1319–1328. PMLR, 2018.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

Ignacio Esponda. Information feedback in first price auctions. *The RAND Journal of Economics*, 39 (2):491–508, 2008.

Gadi Fibich and Nir Gavish. Asymmetric first-price auctions—a dynamical-systems approach. *Mathematics of Operations Research*, 37(2):219–243, 2012.

Arthur Flajolet and Patrick Jaillet. Real-time bidding with side information. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jason Gaitonde, Yingkai Li, Bar Light, Brendan Lucier, and Aleksandrs Slivkins. Budget Pacing in Repeated Auctions: Regret and Efficiency without Convergence. *arXiv e-prints*, art. arXiv:2205.08674, May 2022.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Ramakrishna Gummadi, Peter Key, and Alexandre Proutiere. Repeated auctions under budget constraints: Optimal bidding strategies and equilibria. In *the Eighth Ad Auction Workshop*, volume 4, 2012.

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

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

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

B Kahng, I Yang, H Jeong, and A-L Barabási. Emergence of power-law behaviors in online auctions. In *The Application of Econophysics*, pages 204–209. Springer, 2004.

Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Bernard Lebrun. Existence of an equilibrium in first price auctions. *Economic Theory*, 7(3):421–443, 1996.

Shuang Liu and Hao Su. Regret bounds for discounted MDPs. *arXiv preprint arXiv:2002.05138*, 2020.

Google LLC. Faqs about adsense moving to a first-price auction - google adsense help. `https://support.google.com/adsense/answer/10858748?hl=en`, 2021. Accessed: 2022-05-06.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

R. Norvaiša and V. Paulauskas. Rate of convergence in the central limit theorem for empirical processes. *Journal of Theoretical Probability*, 4(3):511–534, 1991.

Furcy Pin and Peter Key. Stochastic variability in sponsored search auctions: observations and models. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 61–70, 2011.

Sarah Sluis. Big changes coming to auctions, as exchanges roll the dice on first-price. `https://www.adexchanger.com/platforms/big-changes-coming-auctions-exchanges-roll-dice-first-price/`, 2017.

Statista. Digital advertising spending worldwide from 2019 to 2024 (in billion u.s. dollars). `https://www.statista.com/statistics/237974/online-advertising-spending-worldwide/`, 2021. Accessed: 2022-05-05.

Akio Suzukawa. Unbiased estimation of functionals under random censorship. *Journal of the Japan Statistical Society*, 34(2):153–172, 2004.

Long Tran-Thanh, Lampros Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *30th Conference on Uncertainty in AI*, 2014.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Frans D. J. van Schaik, Jack P. C. Kleijnen, et al. *Sealed-bid auctions: case study*. Center for Economic Research, 2001.

Arsenii Vanunts and Alexey Drutsa. Optimal pricing in repeated posted-price auctions with different patience of the seller and the buyer. *Advances in Neural Information Processing Systems*, 32, 2019.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

Donglin Zeng. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics*, 32(4):1533–1555, 2004.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

A PROOF OF THEOREM 3.1

The establishment of the regret bounds will be given in two steps. First, we will show that the buyer's estimation $V_{\hat{F}}$ is approximately accurate with a sufficient number of samples. This relies on the estimation of F. Concentration inequalities are thus intrinsic to our analysis. Second, we bridge the regret defined in Equation (1) with V and $V_{\hat{F}}$. This is done by a series of auxiliary quantities measuring the regret. We obtain the desired result by combining the two steps.

We first present the following lemma (Dvoretzky et al., 1956; Massart, 1990), which states a uniform convergence result for the estimation of cumulative distribution functions.

Lemma A.1 (Dvoretzky–Kiefer–Wolfowitz). *Given $t \in \mathbb{N}$, let $m_1, m_2, \ldots, m_t$ be real-valued independent and identically distributed random variables with cumulative distribution function F. Let $\hat{F}_t$ denote the associated empirical distribution function defined by $\hat{F}_t(x) = \frac{1}{t}\sum_{i=1}^{t} \mathbf{1}_{\{m_i \le x\}}$ where $x \in \mathbb{R}$. Then with probability $1-\delta$, it holds*

$$\sup_x |\hat{F}_t(x) - F(x)| \le \sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\, t^{-\frac{1}{2}}.$$

With Lemma A.1 in hand, given any T, we show a bound for the difference between $V(B_t, t)$ and $V_{\hat{F}}(B_t, t)$ for any $1 \le t \le T$ (recall that $V(B_t, t)$ is the accumulated utility of the optimal strategy with the knowledge of F). We prove this using induction. Note that the induction basis is a little tricky due to the base case of recursion in Algorithm 1. Therefore, in the following lemma, we first deal with the induction step.

Lemma A.2. *For any round $t \le T$, budget B_t and with probability at least $1 - \frac{\delta}{T}$, we have the following bounds for the estimated $V_{\hat{F}}$ and the ground truth with F if $\sup_{B_{t_0}} |V(B_{t_0}, t_0) - V_{\hat{F}}(B_{t_0}, t_0)| \le \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2} t^{-\frac{1}{2}}$:*

$$|V(B_t, t) - V_{\hat{F}}(B_t, t)| \le C t^{-\frac{1}{2}} = \widetilde{O}\left(\frac{1}{\sqrt{t}}\right) \quad \textit{where} \quad C = \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2}.$$

Proof. We will use backward induction to show that

$$|V(B_t, t) - V_{\hat{F}}(B_t, t)| \le C t^{-\frac{1}{2}}.$$

The inequality holds trivially with the condition of the lemma for the basis ($t = t_0$). Suppose the bound holds for $t+1$. Now we write out the difference of the value functions

$$\begin{aligned} V(B_t, t) - V_{\hat{F}}(B_t, t) &= \mathop{\mathbb{E}}_{v \sim G}[Q_{v,F}(B_t, t, b_t^*) - Q_{v,\hat{F}}(B_t, t, b_t)] \\ &\le \mathop{\mathbb{E}}_{v \sim G}[Q_{v,F}(B_t, t, b_t^*) - Q_{v,\hat{F}}(B_t, t, b_t^*)], \end{aligned}$$

where b_t is Algorithm 1's estimated optimal bid and b_t^* is the bid of the benchmark. The inequality establishes by noting that b_t^* is sub-optimal under $\hat{F}$. Next consider the term inside the expectation which is rewritten as follows:

$$\begin{aligned} &Q_{v,F}(B_t, t, b_t^*) - Q_{v,\hat{F}}(B_t, t, b_t^*) \le \\ &\underbrace{|(v_t - b_t^*)(F(b_t^*) - \hat{F}(b_t^*))|}_{\Delta_1} \\ &+ \underbrace{\lambda F(b_t^*)|(V(B_t - b_t^*, t+1) - V_{\hat{F}}(B_t - b_t^*, t+1))| + \lambda|(F(b_t^*) - \hat{F}(b_t^*))V_{\hat{F}}(B_t - b_t^*, t+1)|}_{\Delta_2} \\ &+ \underbrace{\lambda|(1 - F(b_t^*))(V(B_t, t+1) - V_{\hat{F}}(B_t, t+1))| + \lambda|(\hat{F}(b_t^*) - F(b_t^*))V_{\hat{F}}(B_t, t+1)|}_{\Delta_3}. \end{aligned}$$

To bound the above equation, we deal with the three terms $\Delta_1, \Delta_2, \Delta_3$ separately. Using Lemma A.1 and union bound of T rounds, $\Delta_1 \le \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}} t^{-\frac{1}{2}}$ with probability at least $1-\delta$. By the induction hypothesis $|V(B_t,t) - V_{\hat{F}}(B_t,t)| \leq C(t+1)^{-\frac{1}{2}} \leq Ct^{-\frac{1}{2}}$ and that any value function is bounded by $1+\lambda+\lambda^2+\cdots = \frac{1}{1-\lambda}$, we have

$$\Delta_2 \leq \lambda C F(b_t^*) t^{-\frac{1}{2}} + \frac{\lambda}{1-\lambda}\sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\, t^{-\frac{1}{2}},$$
$$\Delta_3 \leq \lambda C(1-F(b_t^*)) t^{-\frac{1}{2}} + \frac{\lambda}{1-\lambda}\sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\, t^{-\frac{1}{2}}.$$

Therefore,

$$\begin{aligned} Q_{v,F}(B_t,t,b_t^*) - Q_{v,\hat{F}}(B_t,t,b_t^*) &\leq \left(\frac{2\lambda}{1-\lambda}\sqrt{\frac{1}{2}\ln\frac{2T}{\delta}} + \lambda C + \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\right) t^{-\frac{1}{2}} \\ &= \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\,\frac{1+\lambda}{(1-\lambda)^2} t^{-\frac{1}{2}}. \end{aligned}$$

This establishes that $V(B_t,t) - V_{\hat{F}}(B_t,t) \leq Ct^{-\frac{1}{2}}$. Finally, by symmetry—swap F and $\hat{F}$ and repeat the proof above, it holds that

$$|V(B_t,t) - V_{\hat{F}}(B_t,t)| \leq Ct^{-\frac{1}{2}}.$$

This finishes the induction step and the claim follows. □

Note that the lemma above assumes that $\sup_{B_{t_0}} |V(B_{t_0},t_0) - V_{\hat{F}}(B_{t_0},t_0)|$ is bounded above by $\sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2}t^{-\frac{1}{2}}$, which holds if the base case $V_{\hat{F}}(B_{t_0},t_0)$ is set accurately (i.e. $V(B_{t_0},t_0) = V_{\hat{F}}(B_{t_0},t_0)$). We use $\widetilde{V}_{\hat{F}}(B_t,t)$ to denote the estimated value function using $\hat{F}$ if the base is indeed accurate. In the following lemma, we show that using the alternative initialization method specified in Line 5 of Algorithm 1, $\sup_{B_t} |V_{\hat{F}}(B_t,t) - \widetilde{V}_{\hat{F}}(B_t,t)|$ actually satisfies the condition of Lemma A.2.

Lemma A.3. *Suppose $\widetilde{V}_{\hat{F}}(B_{t_0},t_0) = V(B_{t_0},t_0)$ and $\widetilde{V}_{\hat{F}}(B_t,t)$ is then computed by the recursive procedure in Algorithm 1. Then it holds that for any $\tau \leq t_0$ and B_τ:*

$$|V_{\hat{F}}(B_\tau,\tau) - \widetilde{V}_{\hat{F}}(B_\tau,\tau)| \leq \frac{1}{1-\lambda}\lambda^{t_0-\tau}.$$

In particular, when $\tau = t$, we have $\sup_{B_t} |V_{\hat{F}}(B_t,t) - \widetilde{V}_{\hat{F}}(B_t,t)| \leq \frac{C_1}{\sqrt{t}}$ (by construction of t_0).

To prove Lemma A.3, we will state a general form of it concerning the error in the initialization of the base case. This lemma will come in handy in the following sections.

Lemma A.4. *For any fixed distributions F, G, consider the value function $V_{F,G}(B_t,t)$. Suppose we use an arbitrary value in $\left[0, \frac{1}{1-\lambda}\right]$ to initialize the base case $V_{F,G}(B_{t_0},t_0)$ and recursively compute thereon to obtain $\widetilde{V}_{F,G}(B_t,t)$, then it holds that for any $t \leq t_0$:*

$$\sup_{B_t} |V_{F,G}(B_t,t) - \widetilde{V}_{F,G}(B_t,t)| \leq \frac{1}{1-\lambda}\lambda^{t_0-t}.$$

Proof. When $\tau = t_0$, the claim holds because $V_{F,G}$ and $\widetilde{V}_{F,G}(\cdot,\cdot)$ are both upper bounded by $\frac{1}{1-\lambda}$ and lower bounded by 0.

Supposing the claim holds when $\tau = t+1$, then for $\tau = t$, we have

$$\begin{aligned} \widetilde{V}_{F,G}(B_t,t) - V_{F,G}(B_t,t) \leq \mathop{\mathbb{E}}_{v\sim G}&[(v_t - b_t^*)F(b_t^*) + \lambda F(b_t^*)\widetilde{V}_{F,G}(B_t - b_t^*, t+1) \\ &+ \lambda(1-F(b_t^*))\widetilde{V}(B_t,t+1) - (v_t - b_t^*)F(b_t^*) - \lambda F(b_t^*)V_{F,G}(B_t - b_t^*, t+1) \\ &- \lambda(1-F(b_t^*))V_{F,G}(B_t,t+1)] \\ &\leq \frac{1}{1-\lambda}\lambda^{t_0-t-1}\lambda \end{aligned}$$

$$= \frac{1}{1-\lambda}\lambda^{t_0-t},$$

In the derivation above, b_t^* denotes the optimal bidding strategy obtained by computing $V(B_t, t)$. The first inequality holds since b_t^* is not be the optimal bidding strategy under $\hat{V}(\cdot, t)$'s view. The second inequality holds since $|\widetilde{V}(B_{t+1}, t+1) - V(B_{t+1}, t+1)| \leq \frac{1}{1-\lambda}\lambda^{t_0-t-1}$ for any B_{t+1}.

And by symmetry, we have $|\widetilde{V}(B_t, t) - V(B_t, t)| \leq \frac{1}{1-\lambda}\lambda^{t_0-t}$. This concludes the induction step and yields the lemma. $\square$

In particular, using Lemma A.4, under the condition of Lemma A.3, the initialization is taken to be 0. We have

$$\sup_{B_t} |V_{\hat{F}}(B_t, t) - \widetilde{V}_{\hat{F}}(B_t, t)| \leq \frac{1}{1-\lambda}\lambda^{t_0-t}.$$

Remark A.5. *For convenience, similar to the notations used in this lemma, for any value function ν, we will use $\widetilde{\nu}$ to denote its approximately-initialized counterpart. Furthermore, we will invoke the lemma many times for other value functions in the rest of the proofs.*

Synthesizing Lemma A.2 and Lemma A.3, we have $|V(B_t, t) - V_{\hat{F}}(B_t, t)| \leq (C + C_1)t^{\frac{1}{2}}$. A crucial next step is to relate this bound to the final regret. This is achieved by two transformations. Roughly speaking, the buyer's "regret" can be viewed in two parts: 1) she does not bid according to the optimal strategy; 2) her strategy is not optimally spending the budget which leads to future losses. These two transformations are done with this intuitive observation and summarized in the following lemma that bounds the performance of the buyer's strategy. Below we first condition on the good event that the estimation succeeds for every t. Then we add the contribution of the bad event to the regret in Theorem 3.1.

Lemma A.6. *For any given B_t and t, denote $V^\pi(B_t, t) = \mathbb{E}_{\boldsymbol{v}\sim G^T}\left[\sum_{\tau=t}^{T}\lambda^{\tau-t}R^\pi(v_\tau, b_\tau)\right]$, then*

$$|V(B_t, t) - V^\pi(B_t, t)| \leq \frac{4\left(\sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2} + C_1\right)}{(1-\lambda)\sqrt{t}}.$$

By Lemma A.6 and further transformations, we can now establish the regret bound of Algorithm 1. Below we first condition on the good event that the estimation (of Lemma A.1) succeeds for every t. Then we add the contribution of the bad event to the regret in Theorem 3.1 finally.

To proof this theorem, we will first bound the following auxiliary "regret" with Lemma A.2 and Lemma A.4.

Let us first make an intuitive and approximate description of the regret. The buyer's "regret" can be viewed in two parts: 1) she does not bid according to the optimal strategy; 2) her strategy is not optimally spending the budget which leads to future losses. Given remaining budget B_t at time t with strategy π, the above intuition guides us to first look at

$$\begin{aligned} R_1 = \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\Bigg[&\sum_{t=1}^{T}(R_t^{\pi^*}(v_t, b_t^*) - R_t^\pi(v_t, b_t)) \\ &+ [\lambda(F(b_t^*)V(B_t - b_t^*, t+1) + (1 - F(b_t^*))V(B_t, t+1)) \\ &- \lambda(F(b_t)V(B_t - b_t, t+1) + (1 - F(b_t))V(B_t, t+1))]\Bigg]. \end{aligned}$$

Lemma A.7. *Suppose Lemma A.2 and Lemma A.4 hold for some constants C and C_1, i.e. $|V(B_t, t) - V_{\hat{F}}(B_t, t)| \leq (C + C_1)t^{-\frac{1}{2}}$. Assume further that $\sup_x |F(x) - \hat{F}(x)| \leq Kt^{-\frac{1}{2}}$ for some constant K. We have*

$$R_1 \leq 2\left(\frac{K(1+\lambda)}{1-\lambda} + (1+\lambda)C + 2C_1\right)\sqrt{T}.$$

Proof. To ease description, we first let

$$\hat{H}_t := (v_t - b_t)\hat{F}(b_t) + \lambda(\hat{F}(b_t)V_{\hat{F}}(B_t - b_t, t+1) + (1 - \hat{F}(b_t))V_{\hat{F}}(B_t, t+1)),$$

$$H_t \coloneqq (v_t - b_t^*)F(b_t^*) + \lambda(F(b_t^*)V(B_t - b_t^*, t+1) + (1 - F(b_t^*))V(B_t, t+1)),$$
$$\widetilde{H}_t \coloneqq (v_t - b_t)F(b_t) + \lambda(F(b_t)V(B_t - b_t, t+1) + (1 - F(b_t))V(B_t, t+1)).$$

(Recall that b_t is Algorithm 1's estimated optimal bid and b_t^* is the bid of the benchmark.) Using the notations above, R_1 now becomes

$$\mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}(H_t - \widetilde{H}_t)\right] \leq \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}(|H_t - \hat{H}_t| + |\hat{H}_t - \widetilde{H}_t|)\right].$$

Use the induction part in the proof of Lemma A.2 and Lemma A.4, $|H_t - \hat{H}_t| \leq (C + C_1)t^{-\frac{1}{2}}$ follows from the condition. In order to bound $|\hat{H}_t - \widetilde{H}_t|$, we write

$$\begin{aligned}
|\hat{H}_t - \widetilde{H}_t| \leq& |(v_t - b_t)(F(b_t) - \hat{F}(b_t))| + \lambda|\hat{F}(b_t) - F(b_t)|V_{\hat{F}}(B_t - b_t, t+1) \\
&+ \lambda F(b_t)|V(B_t - b_t, t+1) - V_{\hat{F}}(B_t - b_t, t+1)| \\
&+ \lambda|F(b_t) - \hat{F}(b_t)|V_{\hat{F}}(B_t, t+1) + \lambda(1 - F(b_t))|V(B_t, t+1) - V_{\hat{F}}(B_t, t+1)| \\
\leq& Kt^{-\frac{1}{2}}\left(1 + \frac{2\lambda}{1-\lambda}\right) + \lambda\left(Ct^{-\frac{1}{2}} + \frac{C_1}{\lambda}t^{-\frac{1}{2}}\right) \\
=& \left(\frac{K(1+\lambda)}{1-\lambda} + \lambda C + C_1\right)t^{-\frac{1}{2}}.
\end{aligned}$$

The first inequality holds because of the triangle inequality and the second inequality establishes using the conditions specified (Lemma A.1, Lemma A.2 and Lemma A.4). Note that it holds that $|V(B_{t+1}, t+1) - V_{\hat{F}}(B_{t+1}, t+1)| \leq \frac{1}{1-\lambda}\lambda^{t_0 - t - 1} = \frac{C_1}{\lambda\sqrt{t}}$.

Finally, we sum up the above estimation and obtain

$$R_1 \leq \sum_{t=1}^{T}\left(\frac{K(1+\lambda)}{1-\lambda} + (1+\lambda)C + 2C_1\right)t^{-\frac{1}{2}},$$

as desired. □

In particular, taking $C = \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}\frac{1+\lambda}{(1-\lambda)^2}}$ and $K = \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}$ as it holds in the analysis of Algorithm 1, we have $R_1 \leq 4(C + C_1)\sqrt{T}$.

Next, we will build connections between R_1 and the accumulated differences in the values i.e. regret.

Let us recall the definition of regret that

$$\text{Regret}(T) = \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}(V^(B_t, t) - V^\pi(B_t, t))\right].$$

Lemma A.8. *For any given B_t and t, suppose that the conditions for Lemma A.7 holds. We have*

$$|V(B_t, t) - V^\pi(B_t, t)| \leq \frac{C'}{(1-\lambda)\sqrt{t}} \quad \textit{where} \quad C' = \frac{K(1+\lambda)}{1-\lambda} + (1+\lambda)C + 2C_1.$$

Proof. To start, we introduce the following notation:

$$H_t^\pi = (v_t - b_t)F(b_t) + \lambda(F(b_t)V^\pi(B_t - b_t, t+1) + (1 - F(b_t))V^\pi(B_t, t+1)).$$

Note that we have $\mathbb{E}_{\boldsymbol{v}\sim G^T}[H_t] = V(B_t, t)$ and $\mathbb{E}_{\boldsymbol{v}\sim G^T}[H_t^\pi] = V^\pi(B_t, t)$. We will use backward induction to bound the difference between $V(B_t, t)$ and $V^\pi(B_t, t)$ for any B_t. First choose a sufficient large t_0 and assume $V(B_{t_0}, t_0) = \widetilde{V}^\pi(B_{t_0}, t_0)$. When $\tau = t_0$, the induction basis holds since $V(B_{t_0}, t_0) = \widetilde{V}^\pi(B_{t_0}, t_0)$. Now consider $\tau = t + 1$. By the induction hypothesis, $|V(B_\tau, \tau) - \widetilde{V}^\pi(B_\tau, \tau)| \leq \frac{C'}{(1-\lambda)\sqrt{\tau}}$, and when $\tau = t$, it holds that

$$\widetilde{H}_t - H_t^\pi = \mathop{\mathbb{E}}_{v\sim G}[\lambda F(b_t)(V(B_t - b_t, t+1) - \widetilde{V}^\pi(B_t - b_t, t+1))$$

$$+ \lambda(1 - F(b_t))(V(B_t, t+1) - \widetilde{V}^\pi(B_t, t+1))]$$
$$\leq \frac{C'\lambda}{(1-\lambda)\sqrt{t+1}}.$$

It follows that

$$|V(B_t, t) - \widetilde{V}^\pi(B_t, t)| \leq \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G}[|H_t - \widetilde{H}_t| + |\widetilde{H}_t - H_t^\pi|] \leq \frac{C'}{\sqrt{t}} + \frac{C'\lambda}{(1-\lambda)\sqrt{t+1}} \leq \frac{C'}{(1-\lambda)\sqrt{t}}.$$

and this concludes the induction step.

Applying the proof techniques in Lemma A.4 with the condition for V^π, we have

$$|V^\pi(B_t, t) - \widetilde{V}^\pi(B_t, t)| \leq \frac{1}{1-\lambda}\lambda^{t_0 - t}.$$

Since t_0 is arbitrarily chosen (as long as it is sufficiently large), we can take $t_0 \to +\infty$. Note that we have $\lim_{t_0\to+\infty} \widetilde{V}^\pi(B_t, t) = V^\pi(B_t, t)$. Therefore, it holds that

$$|V(B_t, t) - V^\pi(B_t, t)| \leq \frac{C'}{(1-\lambda)\sqrt{t}},$$

which ends the proof. □

This establishes

$$\text{Regret}(T) \leq \frac{2C'}{1-\lambda}\sqrt{T},$$

and leads to the Theorem 3.1.

In particular, take $C = \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}\frac{1+\lambda}{(1-\lambda)^2}}$ and $K = \sqrt{\frac{1}{2}\ln\frac{2T}{\delta}}$ as it holds in the analysis of Algorithm 1, we have $\text{Regret}(T) \leq \frac{4(C+C_1)}{1-\lambda}\sqrt{T}$.

The bound above is conditional on the good event that the estimation (of Lemma A.1) succeeds for every $1 \leq t \leq T$. Now take $\delta = \frac{1}{T}$ and note that $\Pr[\text{bad event}] \leq \frac{1}{T}$. By using the trivial regret bound T for the failure event, we have

$$\text{Regret}(T) \leq \left(4\sqrt{\ln(\sqrt{2}T)}\frac{1+\lambda}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T} + 1.$$

and this concludes the proof of Theorem 3.1.

Corollary A.9. *We can establish the relationship between our regret and the loss of cash flow. Let's define the loss of cash flow as* $R_2 = \mathbb{E}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^T R^{\pi^*}(v_t, b_t^*) - R^\pi(v_t, b_t)\right]$, *then it holds that*

$$R_2 \leq \left(\left(4\sqrt{\ln(\sqrt{2}T)}\frac{1+\lambda}{(1-\lambda)^2} + 4C_1\right) + 1 - \lambda\right)\sqrt{T} + \frac{\lambda}{2}\log_{\frac{1}{\lambda}}\frac{T}{(1-\lambda)^2} + 1.$$

Proof. In order to prove Corollary A.9, we show the following lemma first. It relates the regret defined in Equation (1) with R_2 in Corollary A.9.

Lemma A.10. *Suppose that* $\text{Regret}(T) \leq \frac{2C'}{1-\lambda}\sqrt{T}$, *then* $R_2 \leq (2C' + 1 - \lambda)\sqrt{T} + \frac{1-\lambda}{2}\log_{\frac{1}{\lambda}}\frac{T}{1-\lambda}$.

Proof. In fact, note that

$$\text{Regret}(T) = \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^T (1 + \lambda + \cdots + \lambda^{t-1})(R^*(v_t, b_t^*) - R^\pi(v_t, b_t))\right],$$

and any expected instantaneous reward is no greater than 1. We have

$$\left|\frac{1}{1-\lambda}R_2 - \text{Regret}(T)\right| = \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\left|\sum_{t=1}^T \frac{\lambda^t}{1-\lambda}(R^*(v_t, b_t^*) - R^\pi(v_t, b_t))\right|\right]$$

$$\leq \frac{\lambda}{2(1-\lambda)} \log_{\frac{1}{\lambda}} \frac{T}{(1-\lambda)^2} + \frac{1}{\sqrt{T}} T.$$

The above bound holds because when $t \geq \log_{\frac{1}{\lambda}} \frac{\sqrt{T}}{1-\lambda}$, we have $\frac{\lambda^t}{1-\lambda} \leq \frac{1}{\sqrt{T}}$. The considered quantity is divided into two parts with this threshold $\log_{\frac{1}{\lambda}} \frac{\sqrt{T}}{1-\lambda}$. The first part is no greater than $\log_{\frac{1}{\lambda}} \frac{\sqrt{T}}{1-\lambda} \frac{\lambda}{1-\lambda}$ and the second part is no greater than $\frac{1}{\sqrt{T}} T = \sqrt{T}$. The result follows after simple rearrangements. $\square$

Therefore, take $C' = \frac{K(1+\lambda)}{1-\lambda} + (1+\lambda)C + 2C_1$ as it holds in the analysis of Algorithm 1, we have

$$R_2 \leq \left(\left(4\sqrt{\frac{1}{2} \ln \frac{2T}{\delta}} \frac{1+\lambda}{(1-\lambda)^2} + 4C_1 \right) + 1 - \lambda \right) \sqrt{T} + \frac{\lambda}{2} \log_{\frac{1}{\lambda}} \frac{T}{(1-\lambda)^2},$$

conditioning on the good event. Combining the bad event leads to our Corollary A.9 finally. $\square$

B Proof of Theorem 3.2

Note that the function Q depends on both F and G as V does. With the additional condition that G also needs to be estimated, we use $Q_{\cdot,\hat{G}}$ and $Q_{\cdot,G}$ to denote the corresponding version computed using $\hat{G}$ and G respectively. We also extend this notation on V in this scenario (c.f. Algorithm 2).

We start with two simple lemmas. These two lemmas are direct corollaries of Lemma A.1 and Lemma A.4.

Lemma B.1. *Let $\hat{G}_t$ be the estimated distribution at round t in Algorithm 1. With probability $1-\delta$,*

$$\sup_x |\hat{G}_t(x) - G(x)| \leq \sqrt{\frac{1}{2} \ln \frac{2}{\delta}} t^{-\frac{1}{2}}.$$

Lemma B.2. *For any round $t \leq T$, budget B_t, it holds that*

$$\sup_{B_t} |V_{\hat{F},\hat{G}}(B_t, t) - \widetilde{V}_{\hat{F},\hat{G}}(B_t, t)| \leq \frac{1}{1-\lambda} \lambda^{t_0 - t}.$$

In the following proof, we will bridge $V(B_t, t)$ and $V_{\hat{F},\hat{G}}(B_t, t)$—the estimated value function, gradually. For distributions A, B, the notation $V_{A,B}$ refers to the value function computed when $F = A$ and $G = B$.

Lemma B.3. *With probability at least $1 - \frac{\delta}{2T}$, for any given $t \leq T$ and budget B_t, we have*

$$|V_{\hat{F},\hat{G}}(B_t, t) - V_{\hat{F},G}(B_t, t)| \leq \sqrt{\frac{1}{2} \ln \frac{4T}{\delta}} \frac{1}{(1-\lambda)^2} t^{-\frac{1}{2}}.$$

Proof. First we note that Lemma B.1 states with probability at least $1 - \frac{\delta}{2T}$, $\sup_x |G_t(x) - G(x)| \leq \sqrt{\frac{1}{2} \ln \frac{4T}{\delta}} t^{-\frac{1}{2}}$. Now we apply backward induction. When $t = t_0$, the induction basis holds trivially since $|V_{\hat{F},\hat{G}}(B_t, t) - V_{\hat{F},G}(B_t, t)| = 0 \leq \sqrt{\frac{1}{2} \ln \frac{4T}{\delta}} \frac{1}{1-\lambda} t^{-\frac{1}{2}}$. Assume the induction hypothesis holds for $\tau = t + 1$. For any $t < t_0$, v_t and B_t, it holds that

$$\begin{aligned}
Q_{\hat{F},G}(B_t, t, b_t^*) - Q_{\hat{F},\hat{G}}(B_t, t, b_t) &\leq Q_{\hat{F},G}(B_t, t, b_t^*) - Q_{\hat{F},\hat{G}}(B_t, t, b_t^*) \\
&= \lambda \hat{F}(b_t^*) V_{\hat{F},G}(B_t - b_t^*, t+1) + \lambda(1 - \hat{F}(b_t^*)) V_{\hat{F},G}(B_t, t+1) \\
&- \lambda \hat{F}(b_t^*) V_{\hat{F},G}(B_t - b_t^*, t+1) - \lambda(1 - \hat{F}(b_t^*)) V_{\hat{F},\hat{G}}(B_t, t+1) \\
&\leq \lambda \hat{F}(b_t^*) |V_{\hat{F},G}(B_t - b_t^*, t+1) - V_{\hat{F},\hat{G}}(B_t - b_t^*, t+1)| \\
&+ \lambda(1 - \hat{F}(b_t^*)) |V_{\hat{F},G}(B_t, t+1) - V_{\hat{F},\hat{G}}(B_t, t+1)| \\
&\leq \sqrt{\frac{1}{2} \ln \frac{4T}{\delta}} \frac{\lambda}{(1-\lambda)^2} t^{-\frac{1}{2}}.
\end{aligned}$$

The first inequality holds since by construction b_t^* is the optimal bid under $\hat{F}$ and G's view while b_t is the optimal bid under $\hat{F}$ and $\hat{G}$'s view. The second inequality holds with the triangle inequality. And the third inequality is due to the induction hypothesis. Finally, by symmetry—swap G and $\hat{G}$ and repeat the proof above, it holds that $|Q_{\hat{F},G}(B_t,t,b_t) - Q_{\hat{F},\hat{G}}(B_t,t,b_t^*)| \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{\lambda}{(1-\lambda)^2}t^{-\frac{1}{2}}$.

Next, since $V_{\hat{F},\hat{G}}(B_t,t) = \mathbb{E}_{\hat{G}}[Q_{\hat{F},\hat{G}}(B_t,t,b_t)]$ and $V_{\hat{F},G}(B_t,t) = \mathbb{E}_G[Q_{\hat{F},v_t}(B_t,t,b_t^*)]$, we have

$$|V_{\hat{F},G}(B_t,t) - V_{\hat{F},\hat{G}}(B_t,t)| \leq \Delta_1 + \Delta_2,$$

where

$$\Delta_1 = |\mathop{\mathbb{E}}_{G}[Q_{\hat{F},G}(B_t,t,b_t^*)] - \mathop{\mathbb{E}}_{\hat{G}}[Q_{\hat{F},G}(B_t,t,b_t^*)]| \text{ and}$$
$$\Delta_2 = |\mathop{\mathbb{E}}_{\hat{G}}[Q_{\hat{F},\hat{G}}(B_t,t,b_t)] - \mathop{\mathbb{E}}_{\hat{G}}[Q_{\hat{F},G}(B_t,t,b_t^*)]|.$$

- To bound Δ_1: since $Q_{\hat{F},G}$ is supported on $\left[0,\frac{1}{1-\lambda}\right]$ and the difference between G and $\hat{G}$ is upper bounded by $\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}$, Δ_1 is bounded by $\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}\frac{1}{1-\lambda}$. Note that we use the fact that with integration by parts, $\int_0^1 Q\,\mathrm{d}(G-\hat{G}) = Q(G-\hat{G})|_0^1 - \int_0^1(G-\hat{G})\,\mathrm{d}Q$. Therefore, it holds that $|\Delta_1| \leq \int_0^1 |G-\hat{G}||\mathrm{d}Q|$. Since Q is monotone w.r.t. v_t and is two-sided bounded, it holds that $|\Delta_1| \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}\frac{1}{1-\lambda}$.
- To bound Δ_2, it is clear that $\Delta_2 \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{\lambda}{(1-\lambda)^2}t^{-\frac{1}{2}}$ by linearity of expectation.

Therefore, we obtain

$$|V_{\hat{F},G}(B_t,t) - V_{\hat{F},\hat{G}}(B_t,t)| \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{1}{(1-\lambda)^2}t^{-\frac{1}{2}},$$

which finishes induction step and concludes the proof. □

Lemma B.4. *For any $t \leq T$ and budget B_t, with probability at least $1-\frac{\delta}{T}$, it holds that:*

$$|V_{F,G}(B_t,t) - \widetilde{V}_{\hat{F},\hat{G}}(B_t,t)| \leq \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{2+\lambda}{(1-\lambda)^2} + C_1\right)t^{-\frac{1}{2}}.$$

Proof. In order to bound the difference between $\widetilde{V}_{\hat{F},\hat{G}}(B_t,t)$ and $V_{F,G}(B_t,t)$. We first rewrite

$$|V_{F,G}(B_t,t) - \widetilde{V}_{\hat{F},\hat{G}}(B_t,t)| \leq \Delta_1 + \Delta_2 + \Delta_3,$$

where

$$\Delta_1 = |\widetilde{V}_{\hat{F},\hat{G}}(B_t,t) - V_{\hat{F},\hat{G}}(B_t,t)|,$$
$$\Delta_2 = |V_{\hat{F},\hat{G}}(B_t,t) - V_{\hat{F},G}(B_t,t)| \text{ and}$$
$$\Delta_3 = |V_{\hat{F},G}(B_t,t) - V_{F,G}(B_t,t)|.$$

- To bound Δ_1: using Lemma B.2 and the definition of t_0, we conclude that $\Delta_1 \leq \frac{C_1}{\sqrt{t}}$.
- To bound Δ_2: using Lemma B.3, it holds that $\Delta_2 \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{1}{(1-\lambda)^2}t^{-\frac{1}{2}}$.
- To bound Δ_3: using Lemma A.2, we obtain that $\Delta_3 \leq \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2}t^{-\frac{1}{2}}$. Note that we use Bonferroni's method to divide δ into two parts for the union bound of error in the estimations of F and G.

Therefore, it holds that:

$$|V_{F,G}(B_t,t) - \widetilde{V}_{\hat{F},\hat{G}}(B_t,t)| \le \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{2+\lambda}{(1-\lambda)^2} + C_1\right) t^{-\frac{1}{2}}.$$

□

Now, we provide a bound similar to Lemma A.7 for Theorem 3.2. Similarly, we define

$$\begin{aligned}
\hat{H}_t &\coloneqq (v_t - b_t)\hat{F}(b_t) + \lambda\hat{F}(b_t)\widetilde{V}_{\hat{F},\hat{G}}(B_t - b_t, t+1) + \lambda(1-\hat{F}(b_t))\widetilde{V}_{\hat{F},\hat{G}}(B_t, t+1),\\
\widetilde{H}_t &\coloneqq (v_t - b_t)F(b_t) + \lambda F(b_t)V_{F,G}(B_t - b_t, t+1) + \lambda(1-F(b_t))V_{F,G}(B_t, t+1),\\
H_t &\coloneqq (v_t - b_t^*)F(b_t^*) + \lambda F(b_t^*)V_{F,G}(B_t - b_t, t+1) + \lambda(1-F(b_t^*))V_{F,G}(B_t, t+1).
\end{aligned}$$

Recall that $R_1 = \mathbb{E}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}(H_t - \widetilde{H}_t)\right]$. The following series of arguments holds. First, we have

$$R_1 \le \left|\mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}(H_t - \hat{H}_t)\right]\right| + \mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}|\widetilde{H}_t - \hat{H}_t|\right].$$

Followed from Lemma B.3, it holds that

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{v}\sim G^T}\left[\sum_{t=1}^{T}H_t - \hat{H}_t\right]\right| \le \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{2+\lambda}{(1-\lambda)^2} + C_1\right)\sum_{t=1}^{T}t^{-\frac{1}{2}}.$$

It suffices to bound $|\widetilde{H}_t - \hat{H}_t|$. To do so we rewrite

$$|\widetilde{H}_t - \hat{H}_t| \le \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5,$$

where

$$\begin{aligned}
\Delta_1 &= |(v_t - b_t)\hat{F}(b_t) - (v_t - b_t)F(b_t)|,\\
\Delta_2 &= \lambda\hat{F}(b_t)|\widetilde{V}_{\hat{F},\hat{G}}(B_t - b_t, t+1) - V_{F,G}(B_t - b_t, t+1)|,\\
\Delta_3 &= |\lambda\hat{F}(b_t)V_{F,G}(B_t - b_t, t+1) - \lambda F(b_t)V_{F,G}(B_t - b_t, t+1)|,\\
\Delta_4 &= \lambda(1-\hat{F}(b_t))|\widetilde{V}_{\hat{F},\hat{G}}(B_t, t+1) - V_{F,G}(B_t, t+1)| \text{ and}\\
\Delta_5 &= |\lambda(1-\hat{F}(b_t))V_{F,G}(B_t, t+1) - \lambda(1-F(b_t))V_{F,G}(B_t, t+1)|.
\end{aligned}$$

- To bound Δ_1: using Lemma A.1, it holds $\Delta_1 \le \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}$.
- To bound $\Delta_2 + \Delta_4$: using Lemma B.3 and setting $C_1 \leftarrow \frac{C_1}{\lambda}$, it holds that $\Delta_2 + \Delta_4 \le \lambda\left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{2+\lambda}{(1-\lambda)^2} + \frac{C_1}{\lambda}\right)t^{-\frac{1}{2}}$.
- To bound Δ_3: since $V_{F,G}(\cdot,\cdot) \le \frac{1}{1-\lambda}$, then it holds that $\Delta_3 \le \frac{\lambda}{1-\lambda}\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}$.
- To bound Δ_5: like the way we deal with Δ_3, it holds that $\Delta_5 \le \frac{\lambda}{1-\lambda}\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}t^{-\frac{1}{2}}$.

After summing them up, we have

$$|\widetilde{H}_t - \hat{H}_t| \le \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{1+2\lambda}{(1-\lambda)^2} + C_1\right)t^{-\frac{1}{2}}.$$

Therefore, conditioning on the good event that the estimation succeeds (of both F and G) for every t, R_1 is bounded by

$$R_1 \le \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{3+3\lambda}{(1-\lambda)^2} + 2C_1\right)\sum_{t=1}^{T}t^{-\frac{1}{2}} \le \left(\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{6(1+\lambda)}{(1-\lambda)^2} + 4C_1\right)\sqrt{T}.$$

We then apply the same techniques in Lemma A.8 to yield

$$\text{Regret}(T) \leq \left(\sqrt{\ln(2T)}\frac{6(1+\lambda)}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T},$$

conditioning on the good event.

Now take $\delta = \frac{1}{T}$ and note that $\Pr[\text{bad event}] \leq \frac{1}{T}$. By using the trivial regret bound T for the failure event, we obtain the desired bound

$$\text{Regret}(T) \leq \left(\sqrt{\ln(2T)}\frac{6(1+\lambda)}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T} + 1.$$

Similar to Corollary A.9, we have the following corollary.

Corollary B.5. *The loss of cash flow when distribution of G is unknown is upper bound by* $\left(\sqrt{\ln(2T)}\frac{6(1+\lambda)}{(1-\lambda)^2} + 5 - \lambda\right)\sqrt{T} + \frac{\lambda}{2}\log_{\frac{1}{\lambda}}\frac{T}{(1-\lambda)^2} + 1.$

C Omitted proof in Section 3.2

C.1 Proof of Lemma 3.4

We first use the techniques that appeared in (Norvaiša and Paulauskas, 1991) to obtain

Lemma C.1. *Under Algorithm 2, let $\hat{F}_n$ be an empirical process of updates and $\mathcal{B}$ be a general Gaussian process, respectively, indexed by a class $\mathscr{F}$ of real measurable functions. We have:*

$$|\Pr(\{\|\hat{F}_n\|_{\mathscr{F}} \geq r\}) - \Pr(\{\|\mathcal{B}\|_{\mathscr{F}} \geq r\})| \leq C_2(1+r)^{-3}\ln^2(t)\cdot t^{-\frac{1}{6}},$$

where C_2 is a constant depending only on $\hat{F}_n$ and t is the size of data used to update the estimation.

Using Lemma C.1, we have $|\Pr(\sup_b |\sqrt{2^n}(\hat{F}_n(1-b) - F(1-b))| \geq r) - \Pr(\sup_b |B(1-b)| \geq r)| \leq C_2(1+r)^{-3}\ln^2(2^n)(2^{-\frac{n}{6}})$. Since $2^n \leq t \leq 2^{n+1}-1$, it holds that:

$$\begin{aligned}&|\Pr(\sup_b |\sqrt{2^n}(\hat{F}_n(1-b) - F(1-b))| \geq r) - \Pr(\sup_b |B(1-b)| \geq r)|\\ &\leq C_2(1+r)^{-3}\ln^2(t)\left(\frac{t}{2}\right)^{-\frac{1}{6}}.\end{aligned}$$

Taking $M = C_2 2^{\frac{1}{6}}$ concludes the proof.

C.2 Proof of Lemma 3.5

Before we proceed to bound the difference between $\hat{F}_n$ and F, we also need the following lemma to to characterize a property of Gaussian processes.

Lemma C.2. *Let $\mathcal{B}(1-b)$ be a Gaussian process, we have*

$$\Pr(\sup_b |\mathcal{B}(1-b)| \geq r)| \leq C_3 e^{-C_4 r^2},$$

where C_3 and C_4 are constants.

For a Gaussian process, the tail satisfies Gaussian distribution. For a normal distribution, we have the following well-known inequality

$$\int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,\mathrm{d}t \leq C_6 e^{-C_7 x^2},$$

where $x \geq 0$ and C_6, C_7 are constants. For example, we can take $C_6 = \frac{1}{2}e^{\frac{1}{2}}$ and $C_7 = \frac{1}{2}$ in this inequality.

To bound for a certain Gaussian process, we rescale the random variable and the tail distribution also satisfies above property. So, there exists C_3 and C_4 to make Lemma C.2 hold.

Now we give a bound for the estimation of $\hat{F}$. First, we have

$$\Pr(\sup_x |\sqrt{2^n}(\hat{F}(x) - F(x))| \geq r) \leq \Pr(|\mathcal{B}| \geq r) + M(1+r)^{-3}\ln^2(t) \cdot t^{-\frac{1}{6}},$$

We show that Lemma 3.5 establishes if we take $r = \max\{r_1, r_2\}$ where $r_1 = \sqrt{\frac{1}{C_4}\ln\frac{4C_3\ln T}{\delta}}$ and $r_2 = (4M\ln^3 T)^{\frac{1}{3}} t^{-\frac{1}{18}}\delta^{-\frac{1}{3}}$. Indeed, if $r \geq r_1$ then the first part on the right side is not greater than $\frac{\delta}{4\ln T}$ by Lemma C.2 and if $r \geq r_2$ then the second part is not greater than $\frac{\delta}{4\ln T}$. Take $\delta = T^{-\frac{5}{12}}$. It follows from simple comparison (between the rate of growth of r_1 and r_2) that there exists a constant C_5 for any t and T such that $\max\{r_1, r_2\} \leq C_5 r_2$. Now,

$$\Pr\left(\sup_x \left|\sqrt{\frac{t}{2}}(\hat{F}(x) - F(x))\right| \geq C_5 r_2\right) \leq \Pr(\sup_x |\sqrt{2^n}(\hat{F}(x) - F(x))| \geq C_5 r_2) \leq \frac{1}{2T^{\frac{5}{12}}\ln T}$$

since $t \leq 2^{n+1}$. This concludes the proof of the lemma.

C.3 Proof of Theorem 3.6

Similar to the proof in the full-feedback case, let us first assume the knowledge of G. The regret of learning G will be considered later. The proof of Theorem 3.6 is short thanks to the tool-sets established by previous sections. As usual, we first condition on the good event that the estimation succeeds for every t. Then we add the contribution of the bad event to the regret in the final proof.

In order to bound the final regret, we need to bound the difference between F and $\hat{F}_n$ and then apply the methodology used in the proof of Theorem 3.1.

Dropping the first two rounds, for any $2^n \leq t \leq 2^{n+1} - 1$, the estimation of F is $\hat{F}_n$. Then using Lemma 3.5, we obtain that, with probability at least $1 - \frac{\delta}{2\ln T}$,

$$\sup_x |\hat{F}_n(x) - F(x)| \leq \sqrt{2}C_5(4M\ln^3 T)^{\frac{1}{3}} t^{-\frac{5}{9}} T^{\frac{5}{36}}., \tag{5}$$

where $\delta = T^{-\frac{5}{12}}$. Note that we use the fact that if $1 \leq t \leq T$, the algorithm updates for at most $\lfloor \ln T \rfloor$ times. Now, the new concentration bound (Lemma C.1) effective changes K, C and the convergence rate in Lemma A.7, Lemma A.8 and Lemma A.10. Therefore, by substituting $\frac{C}{\sqrt{t}}$ with $\sqrt{2}C_5(4M\ln^3 T)^{\frac{1}{3}} t^{-\frac{5}{9}} T^{\frac{5}{36}} \frac{1+\lambda}{(1-\lambda)^2}$ and $\frac{K}{\sqrt{t}}$ with $\sqrt{2}C_5(4M\ln^3 T)^{\frac{1}{3}} t^{-\frac{5}{9}} T^{\frac{5}{36}}$ in the lemmas we obtain that conditioning on the good event (w.p. $1 - \frac{\delta}{2}$) that the estimation of F succeeds every time,

$$R_1 \leq 4C_1\sqrt{T} + \frac{9\sqrt{2}(1+\lambda)}{2(1-\lambda)^2} C_5 (4M\ln^3 T)^{\frac{1}{3}} T^{\frac{7}{12}},$$

if Algorithm 2 has the knowledge of G.

Next we add the estimation of G. Note that Lemma B.3 decouples the regret of estimating F and G, hence we may apply the same approach in the proof of Theorem 3.2 by taking $C = \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{1+\lambda}{(1-\lambda)^2}$ and $K = \sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}$ and we have

$$R_1 \leq 4C_1\sqrt{T} + \frac{9\sqrt{2}(1+\lambda)}{2(1-\lambda)^2} C_5 (4M\ln^3 T)^{\frac{1}{3}} T^{\frac{7}{12}} + \left[\sqrt{\frac{1}{2}\ln\frac{4T}{\delta}}\frac{2(1+\lambda)}{(1-\lambda)^2}\right]\sqrt{T},$$

conditioning on the good event (w.p. $1 - \delta$) that the estimation of F succeeds every time.

Finally, we use Lemma A.8 to transform R_1 into the final regret, which yields

$$\text{Regret}(T) \leq \left(\frac{9\sqrt{2}(1+\lambda)}{2(1-\lambda)^3} C_5 (4M\ln^3 T)^{\frac{1}{3}}\right) T^{\frac{7}{12}} + \left(\sqrt{\frac{1}{2}\ln{(4T^{\frac{17}{12}})}}\frac{2(1+\lambda)}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T},$$

conditioning on the good event that the estimations of F and G succeeds for every $1 \leq t \leq T$. Note that $\Pr[\text{bad event}] \leq \frac{1}{T^{\frac{5}{12}}}$. By using the trivial regret bound T for the failure event, we have

$$\text{Regret}(T) \leq \left(\frac{9\sqrt{2}(1+\lambda)}{2(1-\lambda)^3} C_5 (4M\ln^3 T)^{\frac{1}{3}} + 1\right) T^{\frac{7}{12}} + \left(\sqrt{\frac{1}{2}\ln{(4T^{\frac{17}{12}})}}\frac{2(1+\lambda)}{(1-\lambda)^3} + \frac{4}{1-\lambda}\right)\sqrt{T}.$$

This concludes the proof.

D EMPIRICAL EVALUATION

Although this is mostly a theory paper, we briefly provide some experimental evidence to show that our algorithm is efficient and performs well to complement the theoretical result.

Setup The datasets of values and highest bids of the opponents are obtained by 1) sampling from uniform distributions over $[0, 1]$, and 2) sampling from right-truncated exponential distributions. Note that by domain knowledge (Fibich and Gavish, 2012; Han et al., 2020b), most valuations and bids in online auctions follow exponential distributions, so we believe that our simulations can correctly reflect real scenarios[9].

The discount factor is set to be an extreme $\lambda = 1 - 10^{-7}$ to test the robustness of our algorithm. On each realized trajectory of values and highest bids, we only reveal v_t and $\max(m_t, b_t)$ to the bidder (i.e. Algorithm 2). The dynamic programming part of the algorithm is computed using heuristics (sampling and gradient descent). For the benchmark, we use a standard baseline — fixed pacing parameter that has the knowledge of the realized values and bids. Although this is an anticipating strategy, we believe it serves as a strong and suitable baseline because pacing is widely applied in practice.

We fix $T = 2000000$ and vary $\rho = B/T$ from 0.005 to 1.0 — the target average expenditure rate that reflects how much the bidder is constrained to budgets. The utility of the online algorithm is compared to the baseline to evaluate how the performance varies w.r.t. budgets.

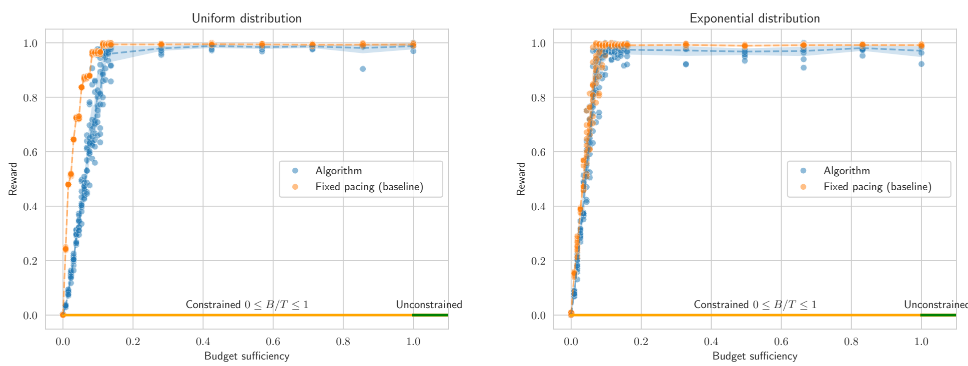

Figure 2: Utilities of the algorithm and the baseline with $T = 2000000$ and B/T from 0.005 to 1.0. Left: uniform distribution. Right: exponential distribution. The fixed-pacing baseline knows opponents' highest bids in advance and uses them to choose an optimal pacing parameter.

Results and discussion The results are shown in Figure 2, where we plot the rewards of the algorithm and the baseline as a function of $\rho = B/T$. We make the following observations regarding the results. First, although the baseline is anticipating, our algorithm achieves almost the same revenue. The loss incurred by our strategy is less than 5% in cases for almost all the ρ values. This suggests that the actual constant factor of our algorithm is rather small. Second, we observe that the strategy of our algorithm in the $\rho = 1$ region recovers that of the Bayesian optimal, providing additional evidence that the algorithm indeed learns the best non-anticipating strategy. Finally, we highlight that in our experiments, each step of decisions takes less than $10\,\text{ms}$ to perform, which is realistic for real-time deployment.

Additional analysis on λ We have also repeated our experiments with $\lambda = 0.9$ and $\lambda = 0.1$ to find out the impact of λ on our algorithm. The results for uniform distributions and exponential distributions are listed in Table 1 and Table 2, respectively. We can see that due to small λs, the baseline is much more unstable than our algorithm, and the performance of our algorithm is superior.

[9]We are unable to obtain real datasets on first-price auctions with budgets due to the fact that currently, most platforms are on the transition to first-price auctions so there are no public datasets to the best of our knowledge.

	$\lambda = 0.9$		$\lambda = 0.1$	
Budget sufficiency	Algorithm	Fixed pacing (baseline)	Algorithm	Fixed pacing (baseline)
0.005	0.052±0.028	**0.371**±0.187	0.082±0.068	**0.174**±0.258
0.01	0.09±0.038	**0.352**±0.151	0.109±0.076	**0.144**±0.2
0.02	0.128±0.038	**0.372**±0.138	0.124±0.08	**0.174**±0.222
0.03	0.179±0.045	**0.342**±0.173	0.143±0.093	**0.181**±0.219
0.04	0.22±0.059	**0.364**±0.15	**0.175**±0.09	0.137±0.198
0.05	0.262±0.071	**0.378**±0.174	**0.185**±0.09	0.131±0.209
0.06	0.319±0.067	**0.378**±0.151	0.192±0.111	**0.211**±0.244
0.07	0.342±0.075	**0.397**±0.193	**0.201**±0.126	0.184±0.24
0.08	**0.389**±0.065	0.374±0.165	**0.21**±0.106	0.158±0.198
0.09	**0.425**±0.085	0.38±0.156	**0.244**±0.113	0.197±0.256
0.10	**0.501**±0.086	0.429±0.18	**0.251**±0.124	0.186±0.233
0.11	**0.53**±0.102	0.373±0.183	**0.265**±0.127	0.141±0.2
0.12	**0.552**±0.081	0.353±0.154	**0.265**±0.116	0.141±0.198
0.13	**0.613**±0.099	0.363±0.157	**0.278**±0.13	0.155±0.216
0.14	**0.64**±0.108	0.356±0.151	**0.314**±0.144	0.196±0.244
0.15	**0.668**±0.093	0.408±0.178	**0.277**±0.118	0.161±0.221
0.16	**0.701**±0.113	0.359±0.183	**0.331**±0.146	0.193±0.265
0.17	**0.723**±0.087	0.366±0.142	**0.307**±0.131	0.168±0.222
0.18	**0.734**±0.104	0.402±0.171	**0.326**±0.138	0.17±0.238
0.19	**0.758**±0.098	0.411±0.16	**0.353**±0.167	0.173±0.24
0.4	**0.722**±0.09	0.336±0.165	**0.325**±0.153	0.217±0.26
0.6	**0.736**±0.091	0.391±0.181	**0.343**±0.16	0.137±0.196
0.8	**0.738**±0.087	0.357±0.147	**0.333**±0.144	0.235±0.266
1.0	**0.736**±0.098	0.359±0.159	**0.333**±0.145	0.148±0.202

Table 1: More results with $\lambda = 0.9$ and $\lambda = 0.1$ for uniform distributions. All the rewards are normalized and the standard deviations are shown. The fixed-pacing baseline knows opponents' highest bids in advance and uses them to choose an optimal pacing parameter.

	$\lambda = 0.9$		$\lambda = 0.1$	
Budget sufficiency	Algorithm	Fixed pacing (baseline)	Algorithm	Fixed pacing (baseline)
0.005	0.057±0.03	**0.299**±0.133	0.08±0.066	**0.111**±0.163
0.01	0.12±0.047	**0.263**±0.122	**0.124**±0.083	0.106±0.165
0.02	0.182±0.049	**0.269**±0.151	0.148±0.095	**0.149**±0.216
0.03	0.259±0.059	**0.32**±0.161	**0.191**±0.108	0.136±0.194
0.04	**0.328**±0.068	0.286±0.152	**0.204**±0.111	0.127±0.174
0.05	**0.403**±0.074	0.32±0.158	**0.251**±0.127	0.108±0.176
0.06	**0.434**±0.068	0.309±0.149	**0.252**±0.121	0.131±0.189
0.07	**0.52**±0.083	0.3±0.138	**0.287**±0.11	0.136±0.18
0.08	**0.584**±0.099	0.301±0.12	**0.288**±0.141	0.106±0.177
0.09	**0.628**±0.096	0.249±0.135	**0.348**±0.154	0.152±0.23
0.10	**0.688**±0.112	0.294±0.134	**0.317**±0.141	0.125±0.194
0.11	**0.665**±0.087	0.29±0.143	**0.346**±0.162	0.104±0.155
0.12	**0.672**±0.096	0.294±0.137	**0.337**±0.148	0.132±0.216
0.13	**0.668**±0.082	0.316±0.146	**0.353**±0.137	0.129±0.187
0.14	**0.674**±0.1	0.307±0.129	**0.344**±0.138	0.114±0.197
0.15	**0.667**±0.1	0.307±0.152	**0.334**±0.142	0.126±0.18
0.16	**0.661**±0.091	0.266±0.11	**0.341**±0.163	0.122±0.183
0.17	**0.698**±0.098	0.297±0.147	**0.343**±0.159	0.132±0.195
0.18	**0.676**±0.106	0.293±0.16	**0.336**±0.158	0.103±0.165
0.19	**0.689**±0.114	0.274±0.125	**0.351**±0.156	0.119±0.191
0.4	**0.645**±0.09	0.296±0.136	**0.337**±0.15	0.146±0.217
0.6	**0.677**±0.094	0.311±0.146	**0.354**±0.143	0.151±0.231
0.8	**0.674**±0.101	0.307±0.163	**0.345**±0.149	0.106±0.152
1.0	**0.679**±0.087	0.274±0.149	**0.355**±0.161	0.099±0.139

Table 2: More results with $\lambda = 0.9$ and $\lambda = 0.1$ for exponential distributions. All the rewards are normalized and the standard deviations are shown. The fixed-pacing baseline knows opponents' highest bids in advance and uses them to choose an optimal pacing parameter.