Pricing against a Budget and ROI Constrained Buyer

Negin Golrezaei
MIT Sloan

Patrick Jaillet
MIT EECS

Jason Cheuk Nam Liang
MIT Operations Research Center

Vahab Mirrokni
Google

Abstract

Internet advertisers (buyers) repeatedly procure ad impressions from ad platforms (sellers) with the aim to maximize total conversion (i.e. ad value) while respecting both budget and return-on-investment (ROI) constraints for efficient utilization of limited monetary resources. Facing such a constrained buyer who aims to learn her optimal strategy to acquire impressions, we study from a seller's perspective how to learn and price ad impressions through repeated posted price mechanisms to maximize revenue. For this two-sided learning setup, we propose a learning algorithm for the seller that utilizes an episodic binary-search procedure to identify a revenue-optimal selling price. We show that such a simple learning algorithm enjoys low seller regret when within each episode, the budget and ROI constrained buyer approximately best responds to the posted price. We present simple yet natural buyer's bidding algorithms under which the buyer approximately best responds while satisfying budget and ROI constraints, leading to a low regret for our proposed seller pricing algorithm. The design of our seller algorithm is motivated by the fact that the seller's revenue function admits a bell-shaped structure when the buyer best responds to prices under budget and ROI constraints, enabling our seller algorithm to identify revenue-optimal selling prices efficiently.

1 INTRODUCTION

In online advertising markets, advertisers (i.e. buyers) run ad campaigns by procuring ad impressions in selling mechanisms held by the platform (i.e. seller). To efficiently utilize limited monetary resources that are allocated to a certain campaign, advertisers' strategies in the procurement process are typically subject to financial constraints, which generally include budget and *return-on-investment* (ROI) constraints. Budget constraints primarily reflect advertisers' monetary limits due to organizational planning, whereas ROI constraints enforces the desired performance/return on the amount of capital spent Kireyev et al. (2016); Golrezaei et al. (2018); Balseiro et al. (2019b). The presence of such financial constraints, along with the increasing availability of real time data, motivates buyers' deployment of complex algorithms to procure impressions. Such financial constraint and algorithm driven buyer behavior introduces significant challenges to sellers' design of selling mechanisms, primarily due to the fact that buyer algorithms adapt quickly and constantly to data generated by buyer-seller interactions, and also sellers' lack of information on buyers' model primitives such as target ROI, budget, buyer algorithm, etc. In this work, we address the following question:

From the perspective of a seller (e.g. ad platform), what is an optimal selling strategy against a buyer who adopts value-maximizing algorithms under both budget and ROI constraints?

We study the setting where a seller repeatedly sells items to a single budget and ROI constrained buyer through a posted price mechanism. This single-buyer setup is primarily motivated by ad platforms' targeting practices that enable advertisers to target users who may be more interested in their ads, as such practices along with advertisers' heterogeneous targeting criteria lead to a very small number of advertisers/buyers per ad impression, justifying our single-buyer setup. Throughout the repeated mechanism, the seller posts a price for an impression during each period, and the buyer decides on whether to accept and pay the posted price for the sold impression. Our key focus lies in the practical two-sided learning setup where buyers adopt learning algorithms under both budget and ROI constraints, and the seller sets prices algorithmically based on past transactions. The key challenge for the seller's problem of interest is two-fold: the seller does not know the buyer parameters such as target ROI, budget or algorithm, and buyer actions constantly adapt to the past buyer-seller algorithmic interactions. The goal of this work is to design a revenue-maximizing seller pricing strategy against algorthmic and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

financially constrained buyers in such a limited information setting.

The main contribution of this work is that we propose a simple seller algorithm that does not require explicitly learning buyer's parameters nor reverse engineering the buyer's learning algorithms. We show that our algorithm is feasible in achieving high revenue under limited information by exploiting a salient property of the seller revenue function against financially-constrained buyers. In particular, we summarize our contributions as followed:

Main contributions. We first characterize the seller revenue function against a clairvoyant budget and ROI constrained buyer who always best responds to posted prices. To begin with, we show that the buyer's best response to a posted price is a "threshold strategy", i.e. the buyer accepts the sold item if her valuation exceeds a certain threshold that depends on the posted price. With this characterization of buyer best response, we show that the seller revenue function against a best-responding clairvoyant buyer admits a salient "bell-shaped" structure: as the seller increases prices, the corresponding per-period seller revenue first monotonically increases and decreases. We argue that such a structure is exploitable by the seller to extract revenue even without knowing buyer model primitives such as value distribution, budget rate, and target ROI.

We exploit this bell-shaped structure and design an episodic binary search seller pricing algorithm. In each episode, the algorithm sets a single price, and then moves on to the next episode with an updated price based on a binary search procedure w.r.t. the realized revenue of previous prices. We also characterize general buyer-algorithm adaptiveness properties that allow buyers to adapt quickly to prices in seller episodes, and present regret analyses against buyer algorithms that are adaptive to seller prices in the sense of our defined adaptiveness properties. Moreover, we argue that seller regret of our proposed algorithm is driven by the agent (i.e. seller or buyer) who incurs a larger loss in terms of learning error.

Finally, we analyze example buyer algorithms which satisfy the aforementioned adaptiveness properties and aim to maximize total value under both budget and ROI constraints. In particular, we consider clairvoyant buyers who best respond in each period, as well as buyers who make decisions based on machine-learned advice that take the form of value distribution estimates. For each of these buyers, we show that both buyer and seller regret are sublinear.

1.1 Literature review

Mechanism design for budget and ROI constrained buyers. One relevant line of research addresses the mechanism design problem for budget or ROI constrained buyers. As one of the pioneering works regarding mechanism for financially constrained buyers, Laffont and Robert (1996) derives the optimal mechanism for symmetric buyers and public budget information. On the contrary, a more recent paper Pai and Vohra (2014) studies the general multidimensional mechanism design setting against buyers with private budgets. Regarding ROI constrained buyers, Golrezaei et al. (2018) shows that the optimal mechanism for symmetric ROI-constrained buyers is either second-price auctions with reduced reserve prices or subsidized second-price auctions. The work also derives an optimal mechanism for asymmetric ROI buyers. There is also a wide range of work that study dynamic mechanism design for budget constrained buyers, and we refer the reader to the survey Bergemann and Said (2010) and references therein. There have also been recent developments for designing auctions in a setup called *autobidding*, where advertisers simultaneously participate in parallel auction to maximize total value while subject to a coupled ROI constraint across all auctions (see e.g. Aggarwal et al. (2019); Deng et al. (2021); Balseiro et al. (2021); Deng et al. (2022)). All aforementioned works focus on the static mechanism design problem, whereas in this paper we address the topic of designing repeated posted price mechanisms to sell to both budget and ROI constrained buyers.

Selling to strategic or learning buyers. Kleinberg and Leighton (2003) studies the scenario where the seller sells items through a repeated posted price mechanism to a single truthful buyer who accepts the price if her valuation is greater than the offered price. The work presents optimal algorithms in the settings where the buyer's valuations are fixed, stochastic and adversarial, respectively. Amin et al. (2013) also concerns selling through a posted price mechanism, but to a strategic buyer who may choose not to accept a price bellow her valuation (or accept a price above her valuation). The work presents learning algorithms in both the fixed valuation and stochastic valuation settings under the assumption that discount their utilities over time. Other related works include Golrezaei et al. (2020) which studies the dynamic pricing problem for repeated contextual second price auctions facing multiple strategic buyers. The work proposes learning algorithms that are robust to buyers' strategic behavior under various seller information structures and provides corresponding performance guarantees. Golrezaei et al. (2019) relaxes several assumptions for one of the settings in Golrezaei et al. (2020), and presents an algorithm with improved performance guarantees. Finally, Balseiro et al. (2019c) considers the dynamic mechanism design problem against strategic buyers, and further identifies a class of problems in which the optimal mechanism is to simply repeat some static mechanism over time. The closest previous work to this paper is Braverman et al. (2018), where it studies the pricing problem against a single unconstrained quasi-linear buyer who adopts a certain class of learning algorithms, which they refer to as "mean-based"

algorithms (e.g. Follow the Perturbed Leader algorithm and EXP3), the seller can extract the buyer's entire surplus; see Deng et al. (2019) for an extension of this work. We remark that all works discussed here do not consider constrained buyers, and therefore this paper distinguishes itself by studying the pricing problem against buyers with both budget and ROI constraints, which further allows us to characterize special structures of seller revenue (see Section 3).

We refer readers to Appendix A for an extended literature review.

2 PRELIMINARIES

Notation. Let $\mathbb{R}_+$ be all non-negative real numbers, and $\mathbb{R}_{++}$ be all strictly positive real numbers. For integer $N \in \mathbb{N}$, denote $[N] = \{1, 2, \ldots, N\}$ and $\Delta_N = \left\{\boldsymbol{p} \in [0,1]^N : \sum_{n\in[N]} p_n = 1\right\}$ be the N-dimensional probability simplex. Finally, denote $\|\cdot\|$ as the Euclidean norm.

Model setup: Consider a seller repeatedly selling items to a buyer over T periods through a posted price mechanism: in each period t, the seller posts a price d_t for the item to be sold, and the buyer makes a take it or leave it decision $z_t \in \{0,1\}$ based on her value v_t of the item, where $z_t = 1$ when the buyer takes the item at price d_t, and 0 otherwise.

We assume the seller commits to a finite price set $\mathcal{D} = \{D_m\}_{m\in[M]}$ where $1 \geq D_1 > \cdots > D_M > 0$ from which she chooses the posted prices $\{d_t\}_{t\in[T]}$, and we assume the the buyer's valuations are drawn independently each period from a distribution over $\boldsymbol{g} = (g_1 \ldots g_N) \in \Delta_N$ ($g_n \in \mathbb{R}_{++}$ for all $n \in [N]$) over a finite support $\mathcal{V} = \{V_n\}_{n\in[N]}$ where $1 \geq V_1 > \cdots > V_N > 0$ such that $\mathbb{P}(v_t = V_n) = g_n$ for any period $t \in [T]$.

ROI and budget constrained buyers: The buyer aims to maximize total acquired value over T periods, while subject to an ROI constraint with the target ROI of $\gamma \geq 1$ and a budget constraint with budget rate $\rho \in (0,1)$.[1] Mathematically, using the shorthand notation $\boldsymbol{d}_{1:T}$ for the sequence of prices $\{d_t\}_{t\in[T]}$, the buyer's hindsight optimization problem can be written as followed

$$\begin{aligned} \text{B-OPT}(\boldsymbol{d}_{1:T}) = \max_{\boldsymbol{z}\in[0,1]^T} \quad & \mathbb{E}\left[\sum_{t\in[T]} v_t z_t\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\sum_{t\in[T]} (v_t - \gamma d_t) z_t\right] \geq 0 \\ & \mathbb{E}\left[\sum_{t\in[T]} d_t z_t\right] \leq \rho T\,. \end{aligned} \tag{1}$$

We remark that both budget and ROI constraints are studied in expectation. Such "soft" constraints are useful in practice due to the fact that real-world advertisers typically engage in many different online advertising campaigns, so it is reasonable to maintain these financial constraints in an average sense. We note that such soft financial constraints are also studied in mechanism design and online learning literature such as Vaze (2018); Golrezaei et al. (2018).

We denote the optimal hindsight buyer decision sequence to Equation (1) as $\{z_t^*(\boldsymbol{d}_{1:T})\}_{t\in[T]}$. When all prices are equal, i.e. $d_t = d$ for all t, we use the shorthand notation B-OPT(d) and $\{z_t^*(d)\}_{t\in[T]}$. Note that optimal hindsight decisions $\{z_t^*(\boldsymbol{d}_{1:T})\}_{t\in[T]}$ may possibly be fractional, which can be implemented by randomization.

The buyer's target ROI γ and budget rate ρ are private to the buyer and unknown to the seller. Also, both the seller and the buyer do not know the valuation distribution $\boldsymbol{g}$.

Seller's benchmark revenue and regret.

The seller does not know the buyer's model primitives, namely the buyer's valuation distribution $\boldsymbol{g}$, target ROI γ and budget rate ρ. Furthermore, the seller only observes the buyer's decision $z_t \in \{0,1\}$, and does not observe buyer values. Under such information structure, we focus on non-anticipative seller pricing strategies that post prices based on historical data, i.e. in each period t, the decision z_t can only depend on $\{(d_\tau, z_\tau)\}_{\tau\in[t-1]}$. We evaluate the performance of any sequence of pricing decision $\{d_t\}_{t\in[T]} \in \mathcal{D}^T$ by benchmarking its realized revenue, namely $\sum_{t\in[T]} d_t z_t$, to the maximum revenue that could have been obtained if (i) the seller had set a fixed price over all T periods and (ii) the buyer makes optimal hindsight decisions given her ROI and budget constraints. Mathematically, assume the seller fixes price $d \in \mathcal{D}$ over all T periods, and the buyer's optimal decisions are $\{z_t^*(d)\}_{t\in[T]}$. Then, the seller's benchmark revenue is $\max_{d\in\mathcal{D}} \mathbb{E}[d \sum_{t\in[T]} z_t^*(d)]$ and her regret can be defined as follows

$$\text{Reg}_{\text{sell}} = \max_{d\in\mathcal{D}} \mathbb{E}\left[d \sum_{t\in[T]} z_t^*(d)\right] - \sum_{t\in[T]} \mathbb{E}\left[d_t z_t\right]\,, \tag{2}$$

where the expectation is taken w.r.t. $\{v_t\}_{t\in[T]}$ and randomness in the buyer's strategy (and thus randomness in $\{z_t^*(d)\}_{t\in[T]}$).

Remark 1. *The seller's regret resembles that of an M-arm multi-arm bandit (MAB) problem (see Lattimore and Szepesvári (2020) for a detailed introduction), where we can view each price $d \in \mathcal{D}$ as an arm and $d \cdot z_t$ as the reward by pulling arm m. Nevertheless, we point out that our setting is more complex than the vanilla MAB setting as the seller's reward $d \cdot z_t$ for setting price d during period t not only depends on the seller algorithm which determines prices based on historical observations , but also the buyer's algorithm to optimize Equation (1).*

We point out that the benchmark revenue in the seller's regret of Equation (2) is strong, as it represents the maximum

[1] Note that in the literature another common buyer objective is to optimize linear utility that takes the form $\sum_{t\in[T]}(v_t - \alpha d_t)z_t$ for some parameter $\alpha \geq 0$. We point out that all results in this paper can be extended easily to such linear objectives.

seller revenue when both the buyer and seller have complete information and act optimally, i.e. if the seller knows everything about the buyer, in each period she myopically posts a revenue-maximizing price under best buyer response.

Our goal is to develop a seller pricing algorithm to minimize regret when facing a buyer who optimizes Equation (1) via running some online learning algorithm (to be discussed in later sections).

3 SELLER'S REVENUE AND REGRET

In this section, we present a reformulation for the seller's benchmark revenue in the seller's regret (Equation (2)), and then further characterize special structures of this reformulation which will later motivate the design of our pricing algorithm.

3.1 Reformulating the seller's benchmark revenue

Recall the seller's benchmark revenue in Equation (2) which depends on the buyer's best response decision sequence over the entire horizon T under a fixed price. To present our reformulation of this benchmark, we first show that for any price d, although the buyer's hindsight optimal decisions $\{z_t^*(d)\}_{t\in[T]}$ may seemingly be interdependent across periods due to the coupling of budget and ROI constraints over the entire horizon, the optimal buyer decision in each period t simply requires the buyer to myopically make a decision z_t that maximizes single-period expected value under "single-period budget and ROI constraints", namely $\mathbb{E}\left[(v_t - \gamma d)\, z_t\right] \geq 0$ and $\mathbb{E}\left[d z_t\right] \leq \rho$.

Formally, consider the following myopic buyer optimization problem: for a given posted price d, let $\boldsymbol{x} \in [0,1]^N$ be some vector whose nth entry x_n denotes the probability of accepting the price when the buyer's realized value is V_n. Then, the myopic buyer optimization problem can be written as Equation (3) whose optimal solution is shown in the following Lemma 1 (see proof in Appendix C.1).

$$\begin{aligned} U(d) = \max_{\boldsymbol{x}\in[0,1]^N} & \sum_{n\in[N]} g_n V_n x_n \\ \text{s.t.} & \sum_{n\in[N]} g_n \left(V_n - \gamma d\right) x_n \geq 0 \\ & d \sum_{n\in[N]} g_n x_n \leq \rho\,. \end{aligned} \tag{3}$$

Lemma 1. *For any price d, the optimal solution to Equation (3) is unique, and takes the form $\boldsymbol{x}_d = (1,1,\dots q,0,0\dots 0) \in [0,1]^N$ for some $q \in (0,1]$.*

The special form of the optimal solution of Equation (3) suggests a buyer strategy that accepts all items when buyer value is beyond a certain threshold. We formalize such a strategy in the following definition.

Definition 1 (Threshold strategy)**.** *For a given vector $\boldsymbol{x}$ that takes the form $\boldsymbol{x} = (1,1,\dots q,0,0\dots 0) \in [0,1]^N$ where $q \in (0,1]$ is the nth entry, we say a buyer adopts a threshold strategy w.r.t. $\boldsymbol{x}$ if, regardless of the posted price, she accepts the item when her value is $V_1 \dots V_{n-1}$; accepts w.p. q when her value is V_n; and rejects the item otherwise.*

As an example, for $N = 4$ and some vector $\boldsymbol{x} = (1,1,0.3,0)$, the buyer adopts a threshold strategy w.r.t. $\boldsymbol{x}$ if she accepts the item when her value is V_1 or V_2; accepts w.p. 0.3 when her value is V_3, and rejects when her value is V_4.

With Lemma 1 and the notion of threshold strategies in Definition 1, we can formally define the buyer's best response to a given price d:

Definition 2 (Buyer best response)**.** *We say a buyer best responds to a posted price d if she adopts a threshold strategy w.r.t. $\boldsymbol{x}_d \in [0,1]^N$ which is the optimal solution to $U(d)$ (see Lemma 1).*

Note that in order for to best respond to a posted price, the buyer would need to know the value distribution $\boldsymbol{g}$.

Our main result for this subsection is illustrated in the following theorem, which states that buyer's hindsight optimal decision sequence $\{z_t^*(d)\}_{t\in[T]}$ for B-OPT(d) in Equation (1) simply requires the buyer to independently best respond to the posted price in each period.

Proposition 2. *Given a single price d posted across all periods, the optimal buyer decision in each period t is to best respond according to a threshold strategy w.r.t. $\boldsymbol{x}_d$ (Definition 2), where $\boldsymbol{x}_d \in [0,1]^N$ is the unique optimal threshold solution to $U(d)$ (Equation (3)). Further, the best response buyer decision induces a per-period expected revenue*

$$\pi(d) := d \sum_{n\in[N]} g_n x_{d,n}\,. \tag{4}$$

Then, $\max_{d\in\mathcal{D}} \mathbb{E}\left[d \sum_{t\in[T]} z_t^*(d)\right] = T \max_{d\in\mathcal{D}} \pi(d)$ and thus $\textit{Reg}_{\textit{sell}} = T \max_{d\in\mathcal{D}} \pi(d) - \sum_{t\in[T]} \mathbb{E}\left[d_t z_t\right]$.

We refer readers to the proof in Appendix C.2.

3.2 Structure of Benchmark Seller Revenue

Here, we present a special underlying structure of the seller revenue $\pi(d)$ defined in Equation (4) which will motivate our pricing algorithm in the next Section 4. The goal of this section is to develop efficient ways to identify $\arg\max_{d\in\mathcal{D}} \pi(d)$ by avoiding exploring each possible price in $\mathcal{D}$. In the rest of the paper, we make the following assumption to rule out trivial problem instances (e.g. cases when the optimal solution $\boldsymbol{x}_d$ corresponding to some $d \in \mathcal{D}$ has all 0 entries or when one of the constraints are redundant):

Assumption 1. *For any $d \in \mathcal{D}$, assume $V_N - \gamma d < 0 < V_1 - \gamma d$ and $\sum_{n\in[N]}(V_n - \gamma d)g_n \neq 0$. Furthermore, assume $D_M < \rho < D_1$.*

To begin with, we categorize all prices $d \in \mathcal{D}$ based on whether constraints are binding under the corresponding optimal solution $\boldsymbol{x}_d$.

Definition 3. *For price d let $\boldsymbol{x}_d$ be the optimal threshold-based solution to $U(d)$ in Equation (3). Then we call d*

- ***Non-binding**, if under $\boldsymbol{x}_d$, both constraints are non binding, i.e., $d\sum_{n\in[N]} g_n x_{d,n} < \rho$ and $\sum_{n\in[N]}(V_n - \gamma d)\, g_n x_{d,n} > 0$;*
- ***Budget binding** if under $\boldsymbol{x}_d$, the budget constraints is binding, i.e. $d\sum_{n\in[N]} g_n x_{d,n} = \rho$ and $\sum_{n\in[N]}(V_n - \gamma d)g_n x_{d,n} > 0$;*
- ***ROI binding** if under $\boldsymbol{x}_d$, the ROI constraint is binding, i.e. $\sum_{n\in[N]}(V_k - \gamma d)g_n x_{d,n} = 0$ and $d\sum_{n\in[N]} g_n x_{d,n} \leq \rho$.*

It is apparent that any price $d \in \mathcal{D}$ must belong to at least one of these categories. Also, if a price is non-binding, it cannot be budget binding or ROI binding.

Our main result of this subsection is the following Theorem 3, which states that as we traverse $\mathcal{D}$ in increasing price order, prices are first non-binding and the revenue $\pi(d)$ increases in d; then prices become budget binding, where revenue remains constant at $\pi(d) = \rho$; finally prices become ROI binding, where $\pi(d)$ decreases in d. The proof can be found in Appendix C.3.

Theorem 3 (Bell-shaped Structure of the Revenue Function)**.** *Suppose that Assumption* 1 *holds. Then, the following hold*

1. *For any non-binding prices $d, \widetilde{d}$, if $d < \widetilde{d}$ then $\pi(d) < \pi(\widetilde{d})$.*
2. *If d is budget binding, any price $\widetilde{d} > d$ cannot be non-binding, which means $\widetilde{d}$ is budget binding or ROI binding.*
3. *If d is ROI binding, then any $\widetilde{d} > d$ must also be ROI binding. Furthermore, $\pi(d) > \pi(\widetilde{d})$.*

We provide an illustration of Theorem 3 in Figure 1 that depicts the "non-binding → budget binding → ROI binding" transition phenomenon, as well as a corresponding revenue "increase → plateau → decrease", as we traverse prices in increasing order. We note that for specific model primitives $\boldsymbol{g}, \gamma, \rho$, there may exist no budget binding prices (as shown in right subfigure in Figure 1), meaning that there are scenarios in which it is impossible for the buyer to extract the entire buyer budget. Nevertheless, this transition phenomena suggests that we can efficiently identify the maximizing revenue $\arg\max_{d\in\mathcal{D}} \pi(d)$ by utilizing a simple binary search approach. Hence, we utilize this structure of $\pi(d)$ to motivate our pricing algorithm.

4 PRICING ALGORITHM AGAINST AN ROI AND BUDGET CONSTRAINED BUYER

The main challenge the seller faces is her lack of knowledge on the buyer's model primitives, namely the buyer's valuation distribution $\boldsymbol{g}$, target ROI γ and budget rate ρ. Furthermore, the seller has limited information feedback as she only observes whether the buyer accepted the price or not, i.e., the seller only observes the outcome $z_t \in \{0, 1\}$. This lack of information makes it very difficult for the seller to estimate the buyer's model primitives. Nevertheless, we propose a simple pricing algorithm that bypasses this lack of knowledge via exploiting the price transition phenomenon as characterized in Theorem 3 and Figure 1. We demonstrate later in subsection 4.1 that this algorithm achieves good performance when facing a general class of algorithms that is adaptive to nonstationary environments.

Our proposed pricing algorithm consists of an exploration phase and an exploitation phase. During the exploration phase, the algorithm searches for a revenue maximizing price $\arg\max_{d\in\mathcal{D}} \pi(d)$ through an episodic structure: the seller initiates the first episode $\mathcal{E}_1$, and fixes the price chosen in this episode D_1 for E consecutive periods. At the end of the episode (i.e. after E periods since the beginning of the episode), the seller records the average per-period revenue $\hat{\pi}(D_1) = \frac{D_1}{E}\sum_{t\in\mathcal{E}_1} z_t$, where $z_t \in \{0, 1\}$ indicates whether the buyer takes the price at time $t \in \mathcal{E}_1$. The process then repeats as the seller moves on to episodes $\mathcal{E}_2, \ldots$ This exploration phase eventually terminates when the seller has explored enough prices. The seller's pricing decision in each episode is governed by a binary search procedure over the price set $\mathcal{D}$, such that every price is chosen at most once across all episodes, and the exploration phase will have $\mathcal{O}(\log(M))$ episodes. Our pricing algorithm is detailed in Algorithm 1.

We note that our proposed algorithm does not try to learn the buyer's model primitives. We further point out that such a binary-search approach is a natural choice to identify revenue-optimal prices in the simplest monopolistic pricing setting under a typical unimodal assumption, [2] and one may wonder whether this approach can have good performances against a much more complex setting where the buyer is ROI and budget constrained and aims to learn her optimal bidding strategy. Surprisingly, in the next section we are in fact able to show this simple approach achieves

[2] In monopolistic pricing, the revenue-optimal price p^* is charachterized by $d^* = \arg\max_d dF(d)$, where F is the cdf of buyer valuations. A typical assumption is such that the function $dF(d)$ is unimodal.

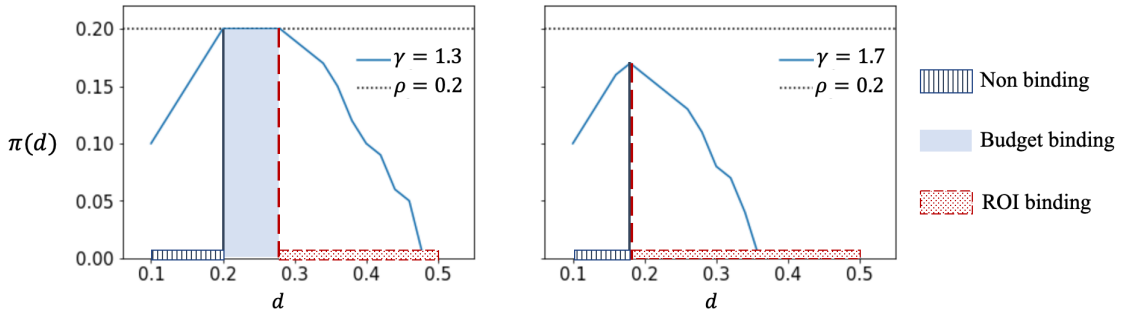

Figure 1: **Seller revenue function bell-shape structure.** Model primitives: number of unique buyer valuations $N = 6$, valuation set $\mathcal{V} = (0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$, valuation distribution $\boldsymbol{g} = (0.1, 0.1, 0.2, 0.1, 0.2, 0.3)$, seller price set $\mathcal{D} = (0.5, 0.48 \dots 0.1)$, buyer budget rate $\rho = 0.2$. The left and right subfigures correspond to target ROI $\gamma = 1.3$ and 1.7 respectively. In both cases, prices transition from non-binding to budget binding, and finally to ROI binnding. Revenue $\pi(d)$ increases as in d when prices are non-binding, decreases in d when prices are ROI binding, and remains at ρ when prices are budget binding. Note that when $\gamma = 1.7$, there are no budget binding prices.

good performances against buyers who are adaptive to price changes.

Algorithm 1 Episodic Binary Search

Input: Exploration episode length E.
Initialize iteration index iter $= 1$.
Exploration episodes:
Set D_1 for E consecutive periods, and record per-period revenue $\hat{\pi}(D_1)$. Then set D_M for E consecutive periods, and record average per-period revenue $\hat{\pi}(D_M)$.
Set $m^* \leftarrow \arg\max_{m \in \{1,M\}} \hat{\pi}(D_m)$ $\mathrm{L} = 1$, $\mathrm{R} = M$, $\text{med} = \lfloor \frac{\mathrm{L}+\mathrm{R}}{2} \rfloor$.
while $\mathrm{L} < \mathrm{R}$ **do**
  iter $\leftarrow$ iter $+ 1$.
  if per-period revenue $\hat{\pi}(D_k)$ is not recorded for $k = \text{med}, \text{med}+1$ **then**
    Set price D_k for E consecutive periods and record per-period revenue $\hat{\pi}(D_k)$ for $k = \text{med}, \text{med}+1$
  end if
  if $\hat{\pi}(D_{\text{med}}) < \hat{\pi}(D_{\text{med}+1})$ **then**
    Set $m^* \leftarrow \arg\max_{m \in \{m^*, \text{med}+1\}} \hat{\pi}(D_m)$, $\mathrm{L} \leftarrow \text{med}+1$, $\text{med} \leftarrow \lfloor \frac{\mathrm{L}+\mathrm{R}}{2} \rfloor$
  else
    Set $m^* \leftarrow \arg\max_{m \in \{m^*, \text{med}\}} \hat{\pi}(D_m)$, $\mathrm{R} \leftarrow \text{med}-1$, $\text{med} \leftarrow \lfloor \frac{\mathrm{L}+\mathrm{R}}{2} \rfloor$
  end if
end while
Exploitation episode:
Set price D_{m^*} for the remaining periods.

For notation convenience, we denote $\mathcal{E}_h$ as the collection of periods in episode h. Finally, we remark that the exploration episode length E is deterministic and depends on the total number of periods T.

4.1 Regret Analysis of Pricing Algorithm

In this section, we provide theoretical guarantees for our proposed pricing algorithm against buyer algorithms whose induced decisions approximate single-round best responses (see Definition 2) in the average sense. We formally define algorithms with such properties as follows:

Definition 4 (ξ-Adaptive Buyer Algorithms)**.** *We say a buyer algorithm is ξ-adaptive to seller algorithm 1 for some $\xi \in (0,1)$ if the induced decisions $\{z_t\}_{t \in [T]}$ in any exploration or exploitation episode $\mathcal{E}_h$ satisfies*

$$\left| \frac{D_h}{|\mathcal{E}_h|} \sum_{t \in \mathcal{E}_h} z_t - \pi(D_h) \right| \leq \frac{\phi(|\mathcal{E}_h|)}{|\mathcal{E}_h|} \tag{5}$$

with probability (w.p.) at least $1 - 1/T$ for some increasing error function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\phi(x) = \mathcal{O}(x^{1-\xi})$. Here D_h is the price set in episode h, and $\pi(\cdot)$ is the per-period revenue function under buyer best response defined in Equation (4).

The term $\left| \frac{D_h}{|\mathcal{E}_h|} \sum_{t \in \mathcal{E}_h} z_t - \pi(D_h) \right|$ is the seller's average revenue loss, relative to the revenue from optimal buyers, over a certain period with a fixed price D_h. Alternatively, the term can be viewed as the buyer's deviation from best responding since $\frac{\pi(D_h)}{D_h} = \sum_{n \in [N]} g_n x_{D_h, n}$ is the optimal probability with which the buyer should take price D_h.

The main result of this subsection is presented in Theorem 4, which characterizes the performance of our pricing algorithm against any ξ-adaptive buyer algorithm. The proof of Theorem 4 can be found in Appendix D.1.

Theorem 4 (Pricing against ξ-adaptive buyers)**.** *Consider the seller runs Algorithm 1 against an ξ-adaptive buyer algorithm (Definition 4). Fix $\epsilon \in (0, \xi)$ independent of T. Then by setting exploration episode length $E = T^{1-\xi+\epsilon}$ in seller algorithm 1, for large enough T under Assumption 1*

the seller's regret is bounded as

$$\begin{aligned} \textit{Reg}_{\textit{sell}} \leq & 2\left(\lfloor \log_2(M) \rfloor + 1\right) \cdot T^{1-\xi+\epsilon} + \phi\left(T\right) \\ & + \left(\lfloor \log_2(M) \rfloor + 1\right)^2 / 2\,, \end{aligned} \tag{6}$$

where ϕ is the error function defined Equation (5).

The first term $T^{1-\xi+\epsilon}$ in the seller's regret (see Equation (2)) characterizes the number of periods required for the buyer's algorithm to approximate the best-responding decisions in each episode facing a fixed price; the second term $\phi\left(T\right)$ represents the buyer's deviation from the best response. Finally, we point out that although in Theorem 6 we set the exploration episode length to be $E = T^{1-\xi+\epsilon}$, the seller does not need to know the exact value of ξ as a lower bound would be sufficient: if the seller knows some lower bound for ξ, say $\xi' < \xi$, she can set $E = T^{1-\xi'}$, and the final seller regret would become $\text{Reg}_{\text{sell}} \leq 2\left(\lfloor \log_2(M) \rfloor + 1\right) \cdot T^{1-\xi'} + \phi\left(T\right) + \left(\lfloor \log_2(M) \rfloor + 1\right)^2 / 2$ for large enough T.

Another interesting observation for the seller regret is that its dependence on the price set dimension M is logarithmic, meaning that our Algorithm 1 is robust w.r.t. the size of the seller's decision set. In fact, later in Section 6, we discuss that this nice logarithmic dependence on M allows us to easily handle continuous price sets without causing decay in seller performance by using a simple discretization approach.

5 EXAMPLE OF ADAPTIVE AND BUYER-REGRET MINIMIZING ALGORITHMS

In this section, we present simple examples of buyer algorithms that are adaptive in the sense of Definition 4, and also aim to satisfy budget and ROI constraints (Equation (1)) while attaining low buyer regret, where the regret of the buyer is defined as

$$\text{Reg}_{\text{buy}} = \text{B-OPT}(\boldsymbol{d}_{1:T}) - \sum_{t\in[T]} \mathbb{E}\left[v_t z_t\right]\,. \tag{7}$$

Here $\{z_t\}_{t\in[T]}$ is the sequence of buyer binary decisions produced by the buyer algorithm. Also recall B-OPT is the buyer's optimal hindsight total value described in Equation (1). In the following subsections, we consider a clairvoyant buyer who best responds in each period as well as a buyer who possess machine-learned (ML) advice with which she uses to make decisions. We then further characterize seller regret of our proposed Algorithm 1 against such buyers.

5.1 Best-responding buyer

As a warm-up buyer example, we first consider a clairvoyant buyer who knows her value distribution $\boldsymbol{g}$, which means the buyer has nothing to learn from the data and thus can best respond in the sense of Definition 2 during each period to maximize value under both budget and ROI constraints (Equation (1). We show in the following lemma that best responding is adaptive (see proof in Appendix E.1).

Lemma 5 (Best-responding is 1/2-adaptive)**.** *There exists some $T_0 \in \mathbb{N}$ such that for all $T > T_0$, best responding is $\frac{1}{2}$-adaptive (Definition 4).*

Combining Lemma 5 and Theorem 4, we present the regret of Algorithm 1 against a best responding buyer in the following theorem whose proof can be found in Appendix E.2

Theorem 6 (Seller's regret against best responding buyer)**.** *Assume the buyer always best responds, then for a fixed $\epsilon \in (0, \frac{1}{2})$ independent of T, if the seller sets prices with episode length $E = T^{\frac{1}{2}+\epsilon}$ using Algorithm 1, then for large enough T, the seller's regret is bounded as $\textit{Reg}_{\textit{sell}} \leq \mathcal{O}(T^{\frac{1}{2}+\epsilon})$. On the other hand, the buyer also incurs $\mathcal{O}(T^{\frac{1}{2}+\epsilon})$ regret, and both budget and ROI constraints are satisfied.*

In this clairvoyant buyer setting, since the buyer is not learning and always best responds, the $T^{\frac{1}{2}}$ constituent in the seller regret is due to learning error from the seller. In the next section, we introduce a buyer who is non-clairvoyant and also constantly learns how to respond, and further discuss how buyer and seller learning errors simultaneously impact seller regret.

5.2 Buyer with machine-learned (ML) advice

In a real world scenario, buyers typically do not know their value distribution; e.g. buyers may be unaware of the likelihood of conversion of their ad impressions. However, the emergence of data-driven tools for online advertising platforms have provided buyers with additional analytics, or so-called ML advice, to help buyers estimate ad conversion. In this subsection, we consider a buyer who possesses ML advice in the form of distribution estimates of $\boldsymbol{g}$ with which she uses to approximate best responses against posted prices. Formally, we characterize such ML-advice-driven buyer responses as followed:

Definition 5 (Approximate best response with ML advice.)**.** *Assume in each period t, the buyer obtains ML advice $\hat{\boldsymbol{g}}_t \in \Delta_N$ that only depends on historical data $\{v_\tau\}_{\tau\in[t]}$ s.t. $\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| < \ell_t$ where ℓ_t is some estimation error. Then, the buyer solves for the optimal solution $\hat{\boldsymbol{x}}_t$ in Equation (3) via replacing the true distribution $\boldsymbol{g}$ with $\hat{\boldsymbol{g}}_t$, and then adopts a threshold strategy w.r.t. $\hat{\boldsymbol{x}}_t$ (see Definition 1).*

We remark that ML advice in the form of distributional estimates is very common. For model-based approaches, ML algorithms assume distributions take a certain parametric form and then uses data to estimate unknown distribution parameters; see e.g. Eliason (1993) for an intro on maximum likelihood estimation. For more general non-parametric approaches, ML advice concerns using empirical estimates (or

so-called histogram estimates), which we will later discuss in Theorem 8.

The following lemma relates ML advice driven approximate responses to our notion of buyer adaptivity in Definition 4, with which we are able to quantify seller regret in light of Theorem 4. The detailed proof can be found in Appendix E.3

Theorem 7 (Seller regret against approximate best responding buyer with ML advice)**.** *Assume the buyer approximate best responds with ML advice (Definition 5) and there exists some $L \in (0,1)$ s.t. in each exploration or exploitation episode h of Algorithm 1 the estimation errors, denoted by ℓ_t's, satisfy $\lim_{t\to\infty} \ell_t = 0$ and $\sum_{t\in\mathcal{E}_h} \ell_t \le \widetilde{\phi}(|\mathcal{E}_h|)$ for some increasing function $\widetilde{\phi}: \mathbb{R}_+ \to \mathbb{R}^+$ and $\widetilde{\phi}(x) \le \mathcal{O}(x^{1-L})$. Then this buyer algorithm is ξ-adaptive for $\xi = \min\{\frac{1}{2}, L\}$. Further, by setting exploration episode length $E = T^{1-\xi+\epsilon}$ for some $\epsilon \in (0,\xi)$ independent of T, the seller regret is exactly that in Equation (6) of Theorem 4 for large enough T. On the buyer side, we have $\text{Reg}_{buy} \le \mathcal{O}(T^{1-\xi})$ and the induced buyer decisions $\{z_t\}_{t\in[T]}$ satisfy*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} (v_t - \gamma d_t)\, z_t\right] \ge -\Theta(T^{-L})$$

and

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} d_t z_t\right] \le \rho + \Theta(T^{-L})\,.$$

We remark that best responding buyers considered in Section 5.1 can be viewed as a special case of buyers with ML advice where the advice is perfect, i.e. $\ell_t = 0$ for all t so $\widetilde{\phi}(x) \equiv 0$ and consequently $L = 1$. This recovers our results in Theorem 6.

Here, we also quickly discuss the aggregate impact of buyer and seller learning error on the seller regret of our proposed Algorithm 1. In particular, the constituent $T^{1-\xi} = T^{1-\min\{\frac{1}{2},L\}}$ in the seller regret arises from learning errors of both the buyer and the seller. We can view the seller's learning rate to be in the order of $t^{-\frac{1}{2}}$, and the buyer learning rate to be of order t^{-L}, and thus we see that the seller regret is governed by the agent that learns at a slower rate: if the buyer is learning more slowly, i.e. $L < \frac{1}{2}$, then the seller regret is driven by the buyer learning loss; a similar argument applies for the case when the buyer learns more quickly.

To conclude this section, we present a concrete example for buyers with ML advice: consider the simple ML advice that is an empirical estimate of the buyer's value distribution:

$$\hat{\boldsymbol{g}}_t = \frac{1}{t}\cdot\left(\sum_{\tau\in[t]} \mathbb{I}\{v_\tau = V^1\}, \ldots, \sum_{\tau\in[t]} \mathbb{I}\{v_\tau = V^N\}\right). \quad (8)$$

Then, both the buyer and seller regret are characterized in the following theorem (see proof in Appendix E.4).

Theorem 8 (Seller regret against approximate best responding buyer with empirical distribution estimates)**.** *When the buyer approximate best responds with ML advice in the form of empirical estimates as defined in Equation (8), Theorem 7 holds for $L = \xi = \frac{1}{2}$ w.p. at least $1 - 1/T$.*

6 ADDITIONAL DISCUSSIONS

Continuous price set. We remark that our main results in this paper, specifically the analyses of Algorithm 1 and the corresponding seller regret, can be easily extended to handle continuous seller price sets, as the seller regret in Theorem 4 only depends logarithmically on M which we recall to be the size of a discrete price set. Assuming the price decision set is $[0,1]$, the approach that the seller can take is to discretize the decision set into $\mathcal{D} = \{\frac{1}{T}, \frac{2}{T} \ldots 1\}$ with size $|\mathcal{D}| = T$. Recall $\pi(d)$ defined in Equation (4) is the expected per-period seller revenue under buyer best response, and define $d^* = \arg\max_{d\in[0,1]} \pi(d)$ to be the optimal price w.r.t. the continuous set, such that the seller regret is now $\text{Reg}_{\text{sell}} = T\cdot\pi(d^*) - \sum_{t\in[T]} \mathbb{E}\left[d_t z_t\right]$ (see Proposition 2). Then, for a price $\widetilde{d} \in \mathcal{D}$ in the discretized set $\mathcal{D}$ that is close to d^* such that $|\widetilde{d} - d^*| < \frac{1}{T}$, similar to our proof in Theorem 7 we can show that the optimal solutions $\boldsymbol{x}_d$ and $\boldsymbol{x}_{\widetilde{d}}$ to the per-period buyer optimization problem $U(d)$ and $U(\widetilde{d})$ (see Equation (1)), respectively, are also close to one another. Further, we can show that $\pi(d^*) - \pi(\widetilde{d}) \le O(\frac{1}{T})$. Therefore, via running Algorithm 1 w.r.t. the discretized price set $\mathcal{D}$, our seller regret when facing a ξ-adaptive buyer (Definition 4) can be bounded as

$$\begin{aligned}
&\text{Reg}_{\text{sell}} = T\max_{d\in[0,1]}\pi(d) - \sum_{t\in[T]}\mathbb{E}\left[d_t z_t\right] \\
&= \underbrace{T\left(\pi(d^*) - \max_{d\in\mathcal{D}}\pi(d)\right)}_{\text{discretization error}} + T\max_{d\in\mathcal{D}}\pi(d) - \sum_{t\in[T]}\mathbb{E}\left[d_t z_t\right] \\
&\le T(\pi(d^*) - \pi(\widetilde{d})) + T\max_{d\in\mathcal{D}}\pi(d) - \sum_{t\in[T]}\mathbb{E}\left[d_t z_t\right] \\
&\le \mathcal{O}(1) + T\max_{d\in\mathcal{D}}\pi(d) - \sum_{t\in[T]}\mathbb{E}\left[d_t z_t\right] \\
&\le \mathcal{O}(1) + \mathcal{O}(\log(T)T^{1-\xi+\epsilon} + \phi(T))\,,
\end{aligned}$$

where the final inequality follows from the seller regret (Equation (6)) in Theorem 4 by setting the price set size $M = T$. That being said, the discretization error introduced to the seller regret is only in the order of $\mathcal{O}(1)$, and this is due to the the fact that the bell-shape structure of seller's revenue (Theorem 3) along with our seller algorithm yields a seller regret that is logarthmic in the discrete price set size.

Ethics of buyer-seller interactions. As modern online ad platforms run selling mechanisms to sell ad impressions,

they also offer services for buyers to help procure ad impressions on their behalf. This raises potential ethical concerns regarding the issue of platforms controlling both the buyer algorithms and auction/pricing protocol. For example, the platform can sets high prices and simultaneously run buyer algorithms that would accept such prices, leading to large platform revenue margins while enforcing high costs to buyers. Nevertheless, in reality, procurement (on behalf of buyers) and pricing are either conducted by different and independent entities, or two non-collusive parties of the same entity where collaboration and any type of information flow that encourages collusion is prohibited. Our main goal of this paper is to shed light on the possible behavior and dynamics for online advertising markets under real financial considerations, and we believe that preventing such collusive behavior between buyer-seller interactions is a future research direction of practical and ethical importance.

Other future directions. One natural future research direction that is of both theoretical and practical interest involves designing pricing algorithms when facing multiple financially constrained buyers. The multi-buyer analogue to our single-buyer posted price setup in this work is to set a single reserve price in each period over time where constrained buyers compete in a second-price auction (see e.g. setup in Golrezaei et al. (2019) for non-constrained buyers). The key challenge lies in the fact that in this multi-buyer setup we no longer have the salient bell-shape structure in the seller revenue function, and more importantly buyer algorithmic interactions introduce significant difficulties to the analyses of seller regret. Similar challenges that arise from selling to multiple learning buyers have also been discussed (but not resolved) in related works such as Braverman et al. (2018); Deng et al. (2019).

References

Aggarwal, G., Badanidiyuru, A., and Mehta, A. (2019). Autobidding with constraints. In *International Conference on Web and Internet Economics*, pages 17–30. Springer.

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

Agrawal, S., Wang, Z., and Ye, Y. (2014). A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890.

Alon, N., Awerbuch, B., and Azar, Y. (2003). The online set cover problem. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 100–105.

Amin, K., Rostamizadeh, A., and Syed, U. (2013). Learning prices for repeated auctions with strategic buyers. *arXiv preprint arXiv:1311.6838*.

Arlotto, A. and Gurvich, I. (2019). Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 9(3):231–260.

Azar, Y., Bhaskar, U., Fleischer, L., and Panigrahi, D. (2013). Online mixed packing and covering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 85–100. SIAM.

Azar, Y., Cohen, I. R., and Panigrahi, D. (2014). Online covering with convex objectives and applications. *arXiv preprint arXiv:1412.3507*.

Babaioff, M., Immorlica, N., Kempe, D., and Kleinberg, R. (2007). A knapsack secretary problem with applications. In *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*, pages 16–28. Springer.

Balseiro, S., Deng, Y., Mao, J., Mirrokni, V., and Zuo, S. (2021). Robust auction design in the auto-bidding world. *Advances in Neural Information Processing Systems*, 34:17777–17788.

Balseiro, S., Golrezaei, N., Mahdian, M., Mirrokni, V., and Schneider, J. (2019a). Contextual bandits with cross-learning. *Advances in Neural Information Processing Systems*, 32:9679–9688.

Balseiro, S., Golrezaei, N., Mirrokni, V., and Yazdanbod, S. (2019b). A black-box reduction in mechanism design with private cost of capital. *Available at SSRN 3341782*.

Balseiro, S., Kim, A., and Russo, D. J. (2019c). On the futility of dynamics in robust mechanism design. *Columbia Business School Research Paper Forthcoming*.

Bergemann, D. and Said, M. (2010). Dynamic auctions. *Wiley Encyclopedia of Operations Research and Management Science*.

Braverman, M., Mao, J., Schneider, J., and Weinberg, M. (2018). Selling to a no-regret buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 523–538.

Dantzig, G. B. (1957). Discrete-variable extremum problems. *Operations research*, 5(2):266–288.

Deng, Y., Golrezaei, N., Jaillet, P., Liang, J. C. N., and Mirrokni, V. (2022). Fairness in the autobidding world with machine-learned advice. *arXiv preprint arXiv:2209.04748*.

Deng, Y., Mao, J., Mirrokni, V., and Zuo, S. (2021). Towards efficient auctions in an auto-bidding world. In *Proceedings of the Web Conference 2021*, pages 3965–3973.

Deng, Y., Schneider, J., and Sivan, B. (2019). Prior-free dynamic auctions with low regret buyers. *Advances in Neural Information Processing Systems*, 32.

Devanur, N. R. and Hayes, T. P. (2009). The adwords problem: online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 71–78.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Number 96. Sage.

Golrezaei, N., Jaillet, P., and Liang, J. C. N. (2019). Incentive-aware contextual pricing with non-parametric market noise. *arXiv preprint arXiv:1911.03508*.

Golrezaei, N., Javanmard, A., and Mirrokni, V. (2020). Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Operations Research*.

Golrezaei, N., Lobel, I., and Paes Leme, R. (2018). Auction design for roi-constrained buyers. *Available at SSRN 3124929*.

Han, Y., Zhou, Z., Flores, A., Ordentlich, E., and Weissman, T. (2020a). Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*.

Han, Y., Zhou, Z., and Weissman, T. (2020b). Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*.

Hoffman, A. J. (2003). On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 174–176. World Scientific.

Kireyev, P., Pauwels, K., and Gupta, S. (2016). Do display ads influence search? attribution and dynamics in online advertising. *International Journal of Research in Marketing*, 33(3):475–490.

Kleinberg, R. (2005). A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 630–631. Citeseer.

Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE.

Laffont, J.-J. and Robert, J. (1996). Optimal auction with financially constrained buyers. *Economics Letters*, 52(2):181–186.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Pai, M. M. and Vohra, R. (2014). Optimal auctions with financially constrained buyers. *Journal of Economic Theory*, 150:383–425.

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. (2020). Concentration inequalities for multinoulli random variables. *arXiv preprint arXiv:2001.11595*.

Vaze, R. (2018). Online knapsack problem under expected capacity constraint. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 2159–2167. IEEE.

Weed, J., Perchet, V., and Rigollet, P. (2016). Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

Zhou, Y., Chakrabarty, D., and Lukose, R. (2008). Budget constrained bidding in keyword auctions and online knapsack problems. In *International Workshop on Internet and Network Economics*, pages 566–576. Springer.

Appendices for
Pricing against a Budget and ROI Constrained Buyer

A EXTENDED LITERATURE REVIEW

As the most closely related works have been discussed in the introduction section, here we only further discuss broader related works.

Other related work in online resource allocation There has been extensive research on online resource allocation with budget/capacity constraints (see e.g. Kleinberg (2005); Devanur and Hayes (2009); Agrawal et al. (2016)) and here we briefly discuss those that are the most relevant.[3] Zhou et al. (2008) studies the budget-constrained bidding problem for sponsored search in an adversarial setting and present an algorithm with competitive ratio that depends on upper and lower bounds on the value-to-cost ratios; Babaioff et al. (2007); Arlotto and Gurvich (2019) study variants of the knapsack and secretary problems under the random order arrival model and stochastic arrival model, respectively, both presenting near optimal algorithms in their respective settings. Our work differs from this line of research as we incorporate an ROI constraint while also considering the problem of how to price against budget and ROI constrained buyers. Finally, Agrawal et al. (2014) utilizes a primal-dual framework to study online linear programming (LP) with packing constraints, where the positive-valued constraint matrix is revealed column by column (each column corresponds to a highest competing bid d_t) along with the corresponding objective coefficient (corresponding to utility $v_t - \alpha d_t$). Their algorithm determines the decision variable corresponding to the arriving column based on the dual variables of past revealed columns.

Online bidding in repeated auctions under feedback constraints Other than budget capacities and ROI targets, buyers are also typically constrained in terms of the amount information available as they participate in auctions. For example, Balseiro et al. (2019a) studies bidding problem in first price auctions under different feedback structures where an unconstrained quasi-linear buyer only observes whether or not she wins the auction, and Han et al. (2020b,a) study a similar problem where the buyer also gets to observe the highest competing bid if she did not win the auction. As another related work, Weed et al. (2016) studies the bidding problem where the buyer does not know her valuation before submitting her bid, and only observes her valuation if she wins the auction. The work considers the stochastic and adversarial highest competing bid settings, and presents algorithms that build on the UCB and EXP3 algorithms, respectively.

Online optimization with covering constraints The buyer's ROI constraint takes the form of a long-term covering constraint. The related problem of optimization under online covering constraints have been studied in Alon et al. (2003); Azar et al. (2013, 2014). However, the setting in these works differ from ours: Instead of making irrevocable online decisions, these works focus on updating a decision vector upon the arrival of a covering constraint each period such that this constraint is satisfied. In other words, they consider the decision problem where covering constraints are satisfied in each period, while our buyers of interest only need to satisfy the covering (ROI) constraint in the long run. Another key difference is that in these works the covering constraints are all positive, which means these constraints can be easily satisfied (per period) by increasing each entry of the decision vector. On the contrary, in our problem the ROI balance per period $(v_t - \gamma d_t)z_t$ may be negative, and hence makes constraint satisfaction more difficult.

Autobidding. This paper is also related to a recent line of work that studies so-called "autobidders" who simultaneously participate in parallel auctions with the aim to maximize total value subject to a global ROI and budget constraint, which says that total value accrued across auctions is no less than total spend times some multiple (i.e. the target ROI), and the total spend is less than a global budget. Aggarwal et al. (2019) has first formulated the optimization problem for autobidders, and presented optimal bidding strategies for such bidders when all parallel auctions are truthful. Deng et al. (2021); Balseiro et al. (2021) study the price of anarchy when multiple autobidders bid in parallel auctions of classic formats such as VCG, GSP and GFP. Deng et al. (2022) show auctioneers can set personalized reserve prices using predictions on bidder values (i.e. machine-learned advice) to improve welfare guarantees for individual bidders.

[3]The buyer's online bidding problem can be viewed as an online resource allocation problem. However, a key difference is that in bidding, the buyer does not observe the highest competing bid d_t (equivalently the amount of resource depleted) before making a decision; as in the resource allocation problem, both the reward and resource depletion are revealed before decision making. Therefore, to apply a resource allocation algorithm in the bidding problem, one must additionally impose some bidding mechanic that indirectly achieves the desired allocation through constructing appropriate bid values.

B ADDITIONAL DEFINITIONS

In this section, we introduce some additional definitions that will be used throughout the appendices.

Definition 6 (Threshold vectors)**.** *We say that an N-dimensional vector $\boldsymbol{x} \in \mathbb{R}^N$ is a threshold vector if it takes the form of $\boldsymbol{x} = (1 \dots 1, q, 0 \dots 0)$, where the first $J \in \{0, \dots N\}$ entries are 1's, followed by some number $q \in [0, 1)$, and trailing with $(N - J - 1)_+$ 0's.*[4] *Any threshold vector is uniquely characterized by its dimension N, as well as, a tuple $(J, q) \in \{0, \dots N\} \times [0, 1)$, so we denote the vector as $\psi(J, q)$. In the special case when $J = N$, take $q = 0$.*

For any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, let $\min\{\boldsymbol{a}, \boldsymbol{b}\} = (\min\{a_i, b_i\})_{i \in [n]}$ be the element-wise minimum. We write $\boldsymbol{a} \preceq \boldsymbol{b}$ if and only if $a_i \leq b_i$ and $\boldsymbol{a} \succeq \boldsymbol{b}$ if and only if $a_i \geq b_i$ for all $i \in [n]$.

C PROOFS FOR SECTION 3

C.1 Proof of Lemma 1

Here, we show a more detailed version of the lemma stated as followed:

Theorem 9 (Detailed version of Lemma 1)**.** *For a fixed price d, define*

$$\begin{aligned} R &= \max\left\{n \in [N] : \sum_{\ell \in [n]} g_\ell \left(V_\ell - \gamma d\right) \geq 0\right\}, \quad q_{\scriptscriptstyle R} = \frac{\sum_{k \in [R]} g_n \left(V_n - \gamma d\right)}{g_{R+1} \cdot \left|V_{R+1} - \gamma d\right|}, \\ B &= \max\left\{n \in [N] : d \sum_{\ell \in [n]} g_\ell \leq \rho\right\}, \quad \textit{and} \quad q_{\scriptscriptstyle B} = \frac{\rho - d \sum_{k \in [B]} g_n}{g_{B+1} \cdot d}, \end{aligned} \tag{9}$$

If we let $\boldsymbol{x}_{\scriptscriptstyle R} = \psi(R, q_{\scriptscriptstyle R})$ and $\boldsymbol{x}_{\scriptscriptstyle B} = \psi(B, q_{\scriptscriptstyle B})$ be two threshold vectors (see Definition 6), then $\boldsymbol{x}_d = \min\left\{\boldsymbol{x}_{\scriptscriptstyle R}, \boldsymbol{x}_{\scriptscriptstyle B}\right\}$ is the unique optimal solution to $U(d)$ in Equation (3)*. Furthermore, $\boldsymbol{x}_d$ is also a threshold vector characterized by tuple (J, q) where*

$$J = \min\{R, B\}, \quad q = x_{d,J+1} = \min\left\{x_{\scriptscriptstyle B,J+1}, x_{\scriptscriptstyle R,J+1}\right\}. \tag{10}$$

Proof.

Our proof for Theorem 9 consists of 3 steps:

- **Step 1.** We show that $\boldsymbol{x}_{\scriptscriptstyle \mathrm{B}}$ is the unique optimal solution to the "budget constraint only" problem:

$$\texttt{P-Budget} = \max_{\boldsymbol{x} \in [0,1]^N} \sum_{n \in [N]} g_n V_n x_n \text{ s.t. } d \sum_{n \in [N]} g_n x_n \leq \rho\,, \tag{11}$$

- **Step 2.** We show that $\boldsymbol{x}_{\scriptscriptstyle \mathrm{R}}$ is the unique optimal solution the "ROI constraint only" problem:

$$\texttt{P-ROI} = \max_{\boldsymbol{x} \in [0,1]^N} \sum_{n \in [N]} g_n V_n x_n \text{ s.t. } \sum_{n \in [N]} g_n \left(V_n - \gamma d\right) x_n \geq 0\,, \tag{12}$$

- **Step 3.** We show that $\boldsymbol{x}_d = \min\{\boldsymbol{x}_{\scriptscriptstyle \mathrm{B}}, \boldsymbol{x}_{\scriptscriptstyle \mathrm{R}}\}$ is feasible to $U(d)$. In other words, we show $\boldsymbol{x}_d$ is feasible to both `P-Budget` and `P-ROI`.

Step 1. We recognize that `P-Budget` is the linear program (LP) relaxation of a 0-1 knapsack problem, in which the items' "value-to-cost ratio", namely $\frac{g_n V_n}{d g_n} = \frac{V_n}{d}$ are ordered: $\frac{V_1}{d} > \dots \frac{V_N}{d}$ since $V_1 > \dots V_N > 0$. Therefore, it is a well known result that the unique optimal solution to `P-Budget` is exactly $\boldsymbol{x}_{\scriptscriptstyle \mathrm{B}}$ (a threshold vector) defined in the statement of Theorem 9; see e.g. Dantzig (1957) for the optimal solution to the 0-1 knapsack LP relaxation.

[4] For the edge case of $(1, \dots 1) \in \mathbb{R}^N$, $J = N$ and hence the number of trailing 0's is $(N - J - 1)_+ = 0$.

Step 2. Let $\widetilde{\boldsymbol{x}} \in [0,1]^N$ be any optimal solution to P-ROI. Define $\kappa = \max\{n \in [N] : V_n \geq \gamma d\}$ so that $V_n \geq \gamma d$ for all $n \leq \kappa$. Then it is easy to see for any $n = 1 \dots \kappa$, we have $\widetilde{x}_n = 1$. This is because if there exists some $j \leq \kappa$ such that $\widetilde{x}_j < 1$, then the solution $\boldsymbol{x} = (\widetilde{x}_1 \dots \widetilde{x}_{j-1}, 1, \widetilde{x}_{j+1}, \dots \widetilde{x}_N)$ is feasible and yields a strictly larger objective than $\widetilde{\boldsymbol{x}}$:

$$\sum_{n\in[N]} g_n V_n x_n - \sum_{n\in[N]} g_n V_n \widetilde{x}_n = V_j(1-\widetilde{x}_j) > 0\,. \tag{13}$$

Hence, the optimal solution to P-ROI takes the form of $\widetilde{\boldsymbol{x}} = (\underbrace{1 \dots 1}_{\kappa \text{ 1's}}, y_{\kappa+1}, \dots y_N) \in [0,1]^N$. Hence, we know that $\widetilde{\boldsymbol{y}} := (y_\kappa, \dots y_N)$ must satisfy

$$\widetilde{\boldsymbol{y}} \;\in\; \arg\max_{\boldsymbol{x}\in[0,1]^{N-\kappa}} \sum_{k=\kappa+1}^{N} g_n V_n x_n \text{ s.t. } \sum_{k=\kappa+1}^{N} g_n\left(\gamma d - V_n\right) x_n \leq \widetilde{c}\,, \tag{14}$$

where we defined $\widetilde{c} = \sum_{n\in[\kappa]} g_n\left(V_n - \gamma d\right) > 0$. Note that we have $\gamma d - V_n > 0$ for all $k = \kappa+1 \dots N$, and hence the optimization problem in Equation (14) is again an LP relaxation of the 0-1 knapsack problem. Thus similar to Step 1, we again consider the "value-to-cost-ratios": for any $i, j \in \{\kappa+1 \dots N\}$, we have

$$V_i > V_j \iff \frac{g_i V_i}{g_i\left(\gamma d - V_i\right)} > \frac{g_j V_j}{g_j\left(\gamma d - V_j\right)}\,.$$

Hence the "value-to-cost-ratios" $\frac{g_n V_n}{g_n(\gamma d - V_n)}$ decreases in n for $n \in \{\kappa+1 \dots N\}$. Therefore, the optimal solution $\widetilde{\boldsymbol{y}}$ to the 0-1 knapsack LP relaxation in Equation (14) is again unique, and is a threshold vector (again see Dantzig (1957)). Hence, the unique optimal solution to P-ROIis a threshold vector, and following Step 1., it is easy to see this unique optimal solution is $\boldsymbol{x}_{\text{R}}$ defined in the statement of Theorem 9.

Step 3. Since $g_n d > 0$ for all $n \in [N]$ and $\boldsymbol{x}_d = \min\{\boldsymbol{x}_{\text{B}}, \boldsymbol{x}_{\text{R}}\} \preceq \boldsymbol{x}_{\text{B}}$, we can apply Lemma 10 (i) with $a_n = g_n d$, $\boldsymbol{Z} = \boldsymbol{x}_{\text{B}}$ and $\boldsymbol{Y} = \boldsymbol{x}_d$, which yields

$$d\sum_{n\in[N]} g_n x_{d,n} \leq d\sum_{n\in[N]} g_n x_{\text{B},n} \leq \rho\,,$$

where the last inequality is due to the fact that $\boldsymbol{x}_{\text{B}}$ is feasible to P-Budget. This implies $\boldsymbol{x}_d$ is also feasible to P-Budget.

On the other hand, again define $\kappa = \max\{n \in [N] : V_n \geq \gamma d\}$ so that $V_n \geq \gamma d$ for all $n \leq \kappa$. Then since $\boldsymbol{x}_d = \min\{\boldsymbol{x}_{\text{B}}, \boldsymbol{x}_{\text{R}}\} \preceq \boldsymbol{x}_{\text{R}}$, and since $g_n\left(V_n - \gamma d\right) > 0$ for $n = 1 \dots \kappa$ and $g_n\left(V_n - \gamma d\right) < 0$ for $n = \kappa+1 \dots N$, we can apply Lemma 10 (ii) with $b_n = g_n\left(V_n - \gamma d\right)$, $\boldsymbol{Z} = \boldsymbol{x}_{\text{R}}$ and $\boldsymbol{Y} = \boldsymbol{x}_d$, which shows

$$\sum_{n\in[N]} g_n\left(V_n - \gamma d\right) x_{\text{R},n} \overset{(i)}{\geq} 0 \overset{(ii)}{\Longrightarrow} \sum_{n\in[N]} g_n\left(V_n - \gamma d\right) x_{d,k} \geq 0\,,$$

where (i) follows from the fact that $\boldsymbol{x}_{\text{R}}$ is feasible to P-ROI and (ii) follows from the first half of Lemma 10 (ii). Hence $\boldsymbol{x}_d$ is also feasible to P-ROI.

The rest of the proof is straightforward: P-Budget, P-ROI and $U(d)$ have the same objectives, while each of P-Budget and P-ROI has one less constraint than $U(d)$, respectively. So P-Budget $\geq U(d)$ and P-ROI $\geq U(d)$. If $\boldsymbol{x}_d = \boldsymbol{x}_{\text{B}}$, because from Step 3. we know $\boldsymbol{x}_d$ is feasible to $U(d)$, then P-Budget $= U(d)$ and $\boldsymbol{x}_d$ is the optimal solution to both P-Budget and $U(d)$. Similarly, when $\boldsymbol{x}_d = \boldsymbol{x}_{\text{R}}$, $\boldsymbol{x}_d$ is the optimal solution to both P-ROI and $U(d)$.

Finally, we argue $\boldsymbol{x}_d$ is the unique optimal solution to $U(d)$. Assume by contradiction there exists some other vector $\boldsymbol{x} \in [0,1]^N$ that is an optimal solution to $U(d)$ and $\boldsymbol{x}_d \neq \boldsymbol{x}$. Then, again if $\boldsymbol{x}_d = \boldsymbol{x}_{\text{B}}$, we know that P-Budget $= U(d)$, and because both $\boldsymbol{x}_d, \boldsymbol{x}$ achieve total value $U(d)$, then both $\boldsymbol{x}_d, \boldsymbol{x}$ are optimal solutions to P-Budget, which contradicts uniqueness of the optimal solution to P-Budget as argued in Step 1. Similarly, we can again arrive at a contradiction for the case when $\boldsymbol{x}_d = \boldsymbol{x}_{\text{R}}$. Hence, the optimal solution to $U(d)$ is unique.

□

C.2 Proof of Proposition 2

The proof for this proposition consists of two steps. First, we show that the buyer's optimal hindsight problem w.r.t. a single price d, namely B-OPT(d) in Equation (1) is upper bounded by $T \cdot U(d)$, which is the single-period myopic optimization problem denoted in Equation (3). Next, we show playing the threshold strategy w.r.t. $\boldsymbol{x}_d \in [0,1]^N$ (i.e. the optimal solution to $U(d)$) every period, gives the buyer a total value exactly $T \cdot U(d)$ while simultaneously satisfying both budget and ROI constraints. Therefore playing the threshold strategy w.r.t. $\boldsymbol{x}_d$ is the optimal value maximizing strategy to the buyer under a fixed price across all periods.

Step 1. Recall the linear program (LP) in Equation (3) that denotes the buyer's single-period myopic optimization problem. It is easy to see the optimal value is bounded and the LP is feasible (consider the solution with all entries set to be 0). Then, strong duality holds, and therefore for any d, there exists corresponding optimal dual variables $(\lambda, \mu) \in \mathbb{R}^2_+$ s.t.

$$\begin{aligned} U(d) &= \max_{\boldsymbol{x} \in [0,1]^N} \sum_{n \in [N]} \left(g_n(1+\lambda)V_n - (\gamma\lambda + \mu)d\right) x_n + \rho\mu \\ &= \sum_{n \in [N]} \left(g_n(1+\lambda)V_n - (\gamma\lambda + \mu)d\right)_+ + \rho\mu \end{aligned} \tag{15}$$

On the other hand, when the sequence of posted prices stays constant at d, we have

$$\begin{aligned} \text{B-OPT}(d) \;&\leq\; \max_{\boldsymbol{z} \in [0,1]^T} \sum_{t \in [T]} \mathbb{E}\left[\left((1+\lambda)v_t - (\gamma\lambda + \mu)d\right) z_t\right] + T\rho\mu \\ &\leq\; \sum_{t \in [T]} \mathbb{E}\left[\left((1+\lambda)v_t - (\gamma\lambda + \mu)d\right)_+\right] + T\rho\mu \\ &=\; T\left(\sum_{n \in [N]} g_n \left((1+\lambda)V_n - (\gamma\lambda + \mu)d\right)_+ + \rho\mu\right) \\ &=\; T \cdot U(d) \end{aligned} \tag{16}$$

Step 2. Let $\boldsymbol{x}_d \in [0,1]^N$ be the optimal solution to $U(d)$ in Equation (3). Then, the threshold strategy w.r.t $\boldsymbol{x}_d$ (see Definition 1) can be represented as

$$z_t^* = \sum_{n \in [N]} x_{d,n} \mathbb{I}\{v_t = V_n\} \tag{17}$$

It is easy to see $\{z_t^*\}_{t \in [T]}$ is feasible to the buyer's optimal hindsight problem B-OPT(d) because:

$$\begin{aligned} \mathbb{E}\left[\sum_{t \in [T]} (v_t - \gamma d)\, z_t^*\right] \;&=\; \sum_{t \in [T]} \mathbb{E}\left[(v_t - \gamma d) \sum_{n \in [N]} x_{d,n} \mathbb{I}\{v_t = V_n\}\right] \\ &=\; \sum_{t \in [T]} \sum_{n \in [N]} g_n (V_n - \gamma d) x_{d,n} \overset{(i)}{\geq} 0 \end{aligned} \tag{18}$$

and

$$\begin{aligned} \mathbb{E}\left[\sum_{t \in [T]} d z_t^*\right] \;&=\; d \sum_{t \in [T]} E\left[\sum_{n \in [N]} x_{d,n} \mathbb{I}\{v_t = V_n\}\right] \\ &=\; T \cdot d \sum_{t \in [T]} \sum_{n \in [N]} g_n x_{d,n} \\ &\overset{(ii)}{\leq} \rho T \end{aligned} \tag{19}$$

where both (i) and (ii) hold because $\boldsymbol{x}_d$ is feasible to $U(d)$. Finally, the threshold strategy yields a total value exactly $TU(d)$ because

$$\sum_{t\in[T]} \mathbb{E}\left[v_t z_t^*\right] = \sum_{t\in[T]} \mathbb{E}\left[v_t \sum_{n\in[N]} x_{d,n}\mathbb{I}\{v_t = V_n\}\right] = T\cdot \sum_{n\in[N]} g_n V_n x_{d,n} = T\cdot U(d)\,, \tag{20}$$

where the final equality follows from the fact that $\boldsymbol{x}_d$ is optimal to $U(d)$.

Therefore, in light of the upper bound shown in Equation (16), the threshold strategy in Equation (17) is optimal to the buyer's hindsight problem B-OPT(d).

Finally, the seller's revenue under the buyer's optimal threshold strategy is $d\sum_{t\in[T]} z_t^* = T\cdot\sum_{n\in[N]} g_n x_{d,n} = T\cdot\pi(d)$ where $\pi(d)$ is the per-period revenue defined in Equation (4). □

C.3 Proof of Theorem 3

Our proof relies on the following fact

Fact 1. *If price d is nonbinding, then the corresponding optimal solution $\boldsymbol{x}_d$ to $U(d)$ is $\boldsymbol{x}_d = (1\dots 1)\in\mathbb{R}_+^n$.*

Proof. We prove the claim via contradiction. Assume there is some index $k\in[N]$ such that $\boldsymbol{x}_{d,k} < 1$. Then consider the solution $\boldsymbol{x} = (x_{d,1}\dots x_{d,k-1}, y, x_{d,k+1},\dots x_{d,n})$ where we replaced the k'th entry of $\boldsymbol{x}_d$ with

$$y = x_{d,k}+\epsilon, \quad \text{where } \epsilon := \min\left\{1-x_{d,k}, \frac{\rho-\sum_{n\in[N]} g_n x_{d,n}}{dg_k}, \frac{\sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n}}{|V_k-\gamma d|\, g_k}\right\} \overset{(i)}{>} 0\,,$$

where (i) follows from the fact that $\boldsymbol{x}_d$ is nonbinding, i.e. $\rho > \sum_{n\in[N]} g_n x_{d,n}$ and $\sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} > 0$. Then

$$d\sum_{n\in[N]} g_n x_n \;=\; d\sum_{n\in[N]} g_n x_{d,n} + dg_k\epsilon \;\le\; d\sum_{n\in[N]} g_n x_{d,n} + \left(\rho - \sum_{n\in[N]} g_n x_{d,n}\right) = \rho\,.$$

On the other hand, if $V_k-\gamma d > 0$, then

$$\sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} = \sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} + (V_k-\gamma d)\, g_k\epsilon > \sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} > 0\,.$$

If $V_k-\gamma d < 0$, then

$$\begin{aligned}\sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} \;&=\; \sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} + (V_k-\gamma d)\, g_k\epsilon \\ &\ge\; \sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n} + (V_k-\gamma d)\cdot\frac{\sum_{n\in[N]}(V_n-\gamma d)\, g_n x_{d,n}}{|V_k-\gamma d|} = 0\end{aligned}$$

where in the last equality we used $|V_n-\gamma d| = -(V_n-\gamma d)$ since $V_n-\gamma d < 0$.

The above shows $\boldsymbol{x}$ is feasible to $U(d)$. On the other hand, $\sum_{n\in[N]} V_n g_n x_{d,n} < \sum_{n\in[N]} V_n g_n x_n$, so $\boldsymbol{x}$ yields a strictly larger objective than $\boldsymbol{x}_d$, contradicting the optimality of $\boldsymbol{x}_d$. □

We now return to our proof for Theorem 3.

(1). When both $d, \widetilde{d}$ are non-binding, Fact 1 implies $\boldsymbol{x}_d = \boldsymbol{x}_{\widetilde{d}} = (1\dots 1)$.

$$\pi(d) = d\sum_{n\in[N]} g_n x_{d,n} = d\sum_{n\in[N]} g_n < \widetilde{d}\sum_{n\in[N]} g_n = \widetilde{d}\sum_{n\in[N]} g_n x_{\widetilde{d},n} = \pi(\widetilde{d})\,.$$

(2). We prove this claim by contradiction. Assume $\widetilde{d}$ is non-binding and $\widetilde{d} > d$ where d is budget binding. Fact 1 states that $\boldsymbol{x}_{\widetilde{d}} = (1 \dots 1)$. Hence

$$\rho = \pi(d) = d \sum_{n\in[N]} g_n x_{d,n} \le d \sum_{n\in[N]} g_n x_{\widetilde{d},n} < \widetilde{d} \sum_{n\in[N]} g_n x_{\widetilde{d},n} \overset{(i)}{<} \rho\,,$$

where (i) follows from the definition that $\widetilde{d}$ is non-binding. Hence we obtain a contradiction, and $\widetilde{d}$ cannot be non-binding. This means $\widetilde{d}$ must be budget or ROI binding.

(3). Here we show that if some price $d \in \mathcal{D}$ is ROI binding so that $\sum_{n\in[N]}(V_n - \gamma d) g_n x_{d,n} = 0$, any price $\widetilde{d} > d$ must also be ROI binding. We first claim that $\boldsymbol{x}_{\widetilde{d}} \preceq \boldsymbol{x}_d$. To show this, we use a contradiction argument by assuming $\boldsymbol{x}_{\widetilde{d}} \succeq \boldsymbol{x}_d$.

Let the threshold vector $\boldsymbol{x}_d$ be characterized by $\boldsymbol{x}_d = \psi(J, q)$ (see definition of threshold vectors in Definition 6). Under Assumption 1, we note that $\boldsymbol{x}_d$ cannot have all 0 entries and hence $x_{d,1} > 0$. However, since $\sum_{n\in[N]}(V_n - \gamma d) g_n x_{d,n} = 0$, it must be the case that $V_{J+1} - \gamma d < 0$. Now, applying the ordering property for threshold vectors in the second half of Lemma 10 (ii) by taking $\boldsymbol{Z} = \boldsymbol{x}_{\widetilde{d}}$, $\boldsymbol{Y} = \boldsymbol{x}_d$, and $b_i = V_i - \gamma d$ we have

$$0 = \sum_{n\in[N]} (V_n - \gamma d) g_n x_{d,n} \ge \sum_{n\in[N]} (V_n - \gamma d) g_n x_{\widetilde{d},n} > \sum_{n\in[N]} (V_n - \gamma \widetilde{d}) g_n x_{\widetilde{d},n}\,.$$

In the last inequality we used the fact that $\widetilde{d} > d$. Hence, this contradicts the feasibility of $\boldsymbol{x}_{\widetilde{d}}$, so we conclude that $\boldsymbol{x}_{\widetilde{d}} \preceq \boldsymbol{x}_d$. This further implies

$$\rho \ge \underbrace{d \sum_{n\in[N]} g_n x_{d,n}}_{=\pi(d)} \overset{(i)}{=} \frac{1}{\gamma} \sum_{n\in[N]} V_n g_n x_{d,n} \overset{(ii)}{>} \frac{1}{\gamma} \sum_{n\in[N]} V_n g_n x_{\widetilde{d},n} \overset{(iii)}{\ge} \underbrace{\widetilde{d} \sum_{n\in[N]} g_n x_{\widetilde{d},n}}_{=\pi(\widetilde{d})}\,,$$

where (i) follows from d being ROI binding, i.e. $\sum_{n\in[N]}(V_n - \gamma d) g_n x_{d,n} = 0$; (ii) follows from $\boldsymbol{x}_{\widetilde{d}} \preceq \boldsymbol{x}_d$; (iii) follows from feasibility of $\widetilde{d}$ so that $\sum_{n\in[N]}(V_n - \gamma \widetilde{d}) g_n x_{\widetilde{d},n} \ge 0$. Therefore, $\rho \ge \pi(d) > \pi(\widetilde{d})$.

Finally, $\rho > \pi(\widetilde{d})$ implies that $\widetilde{d}$ is either non-binding or ROI binding. We note that it is not possible for $\widetilde{d}$ to be non-binding, because $\widetilde{d}$ non-binding implies $\boldsymbol{x}_{\widetilde{d}} = (1 \dots 1)$ according Fact 1, contradicting $\boldsymbol{x}_{\widetilde{d}} \preceq \boldsymbol{x}_d$ which we showed earlier. Here we used the fact that $\boldsymbol{x}_d \ne (1 \dots 1)$ because $\boldsymbol{x}_d$ is ROI binding and Assumption 1 states for any $d \in \mathcal{D}$, $\sum_{n\in[N]}(V_n - \gamma d) g_n \ne 0$. □

C.4 Additional lemmas for Section 3

Lemma 10 (Ordering property for threshold vectors)**.** *Consider $\{a_i\}_{i\in[N]} \subseteq \mathbb{R}^N_+$ and $\{b_i\}_{i\in[N]} \subseteq \mathbb{R}^N$ where there exists some $j \in [N]$ such that $b_i > 0$ for all $i = 1 \dots j$ and $b_i < 0$ for all $i = j+1, \dots m$. Let $\boldsymbol{Z}, \boldsymbol{Y} \in [0,1]^N$ be two threshold vectors (see Definition 6) such that $\boldsymbol{Y} = \psi(J_Y, q_Y)$, $\boldsymbol{Z} = \psi(J_Z, q_Z)$, and $\boldsymbol{Z} \succeq \boldsymbol{Y}$. Then the following hold:*

(i) $\sum_{i\in[N]} a_i Z_i \ge \sum_{i\in[N]} a_i Y_i$.

(ii) If $\sum_{i\in[N]} b_i Z_i \ge 0$ *then* $\sum_{i\in[N]} b_i Y_i \ge 0$. *Furthermore, if* $b_{J_Y+1} < 0$, *then* $\sum_{i\in[N]} b_i Y_i \ge \sum_{i\in[N]} b_i Z_i \ge 0$.

(iii) If $\sum_{i\in[N]} b_i Y_i < 0$ *then* $\sum_{i\in[N]} b_i Z_i < 0$.

Proof.

(i) Since $a_i > 0$ for all $i \in [N]$, and $\boldsymbol{Z} \succeq \boldsymbol{Y}$ (i.e. $Z_i \ge Y_i$ for all $i \in [N]$), it is easy to see $\sum_{i\in[N]} a_i Z_i \ge \sum_{i\in[N]} a_i Y_i$.

(ii) By the definition of threshold vectors, we have $Y_{J_Y+1} = q_Y$ while $Y_i = 0$ for all $i > J_Y + 1$. We prove the claim by contradiction by assuming $\sum_{i\in[N]} b_i Y_i < 0$.

First, it is easy to see $b_{J_Y+1} < 0$. This is because if $b_{J_Y+1} > 0$, then $b_i > 0$ for all $i = 1 \dots J_Y + 1$ by the definition of $\{b_i\}_{i\in[N]}$, and hence $\sum_{i\in[N]} b_i Y_i = \sum_{i\in[J_Y+1]} b_i Y_i \ge 0$ contradicting our assumption that $\sum_{i\in[N]} b_i Y_i < 0$.

Next, since $\sum_{i\in[N]} b_i Y_i < 0 \leq \sum_{i\in[N]} b_i Z_i$, we have $\sum_{i\in[N]} b_i(Z_i - Y_i) \geq 0$. On the other hand,

$$\sum_{i\in[N]} b_i(Z_i - Y_i) \overset{(i)}{=} \sum_{i=J_Y+1}^{N} b_i(Z_i - Y_i) \overset{(ii)}{<} 0\,.$$

Here, (i) follows from the definition of a threshold vector so that $Y_i = 1$ for all $i = 1 \dots J_Y$ and also $Z_i = 1$ for all $i = 1 \dots J_Y$ due to $\boldsymbol{Z} \succeq \boldsymbol{Y}$. (ii) follows from the fact that $b_{J_Y+1} < 0$ so $b_i < 0$ for all $i \geq J_Y + 1$ due to the definition of $\{b_i\}_{i\in[N]}$. Hence, we arrive at a contradiction, which allows us to conclude the first half of the claim, i.e. $\sum_{i\in[N]} b_i Z_i \geq 0$ implies $\sum_{i\in[N]} b_i Y_i \geq 0$.

We now show the second half of the claim i.e. $b_{J_Y+1} < 0$ implies $\sum_{i\in[N]} b_i Y_i \geq \sum_{i\in[N]} b_i Z_i \geq 0$. If $b_{J_Y+1} < 0$, then $b_i < 0$ for all $i = J_Y + 1 + \dots J_Z + 1$, and hence

$$\sum_{i\in[N]} b_i(Z_i - Y_i) = b_{J_Y+1}(Z_{J_Y+1} - Y_{J_Y+1}) + \sum_{i=J_Y+2}^{J_Z+1} b_i Z_i \overset{(i)}{<} 0\,.$$

Note that in the above inequality the summand $\sum_{i=J_Y+2}^{J_Z+1} b_i Z_i$ does not exist if $J_Y = J_Z$, and in (i) we also used the fact that $Y_i = 0$ for all $i > J_Y + 1$ using the definition of a threshold veector.

(iii) We again use a contradiction argument by assuming $\sum_{i\in[N]} b_i Z_i \geq 0$, and the rest of the proof is almost identical to that of (ii) so we will omit it here. □

D PROOFS FOR SECTION 4

D.1 Proof of Theorem 4

Define $G := \min_{d,\widetilde{d}\in\mathcal{D}:\pi(d)\neq\pi(\widetilde{d})} \left|\pi(d) - \pi(\widetilde{d})\right|$ to be the minimum revenue gap for all price pairs that do not yield the same revenue, where $\pi(d) := d\sum_{n\in[N]} g_n x_{d,n}$ for any $d \in \mathcal{D}$ is the per-period average seller revenue defined in Equation (4). Recall $\hat{\pi}(D_h) = \frac{D_h}{|\mathcal{E}_h|}\sum_{t\in\mathcal{E}_h} z_t$ is the estimate of $\pi(D_h)$ for episode h with fixed price D_h (see Algorithm 1).

For any exploration episode $\mathcal{E}_h$ whose length is $|\mathcal{E}_h| = T^{1-\xi+\epsilon}$, we have w.p. at least $1 - 1/T$

$$\begin{aligned} &\left|\frac{\hat{\pi}(D_h)}{D_h} - \frac{\pi(D_h)}{D_h}\right| = \left|\frac{1}{|\mathcal{E}_h|}\sum_{t\in\mathcal{E}_h} z_t - \frac{\pi(D_h)}{D_h}\right| \overset{(i)}{\leq} \frac{\phi(|\mathcal{E}_h|)}{|\mathcal{E}_h|} \overset{(ii)}{\leq} \frac{\phi(T)}{T^{1-\xi+\epsilon}} \\ &\overset{(iii)}{\Rightarrow} |\hat{\pi}(D_h) - \pi(D_h)| \leq \frac{\phi(T)}{T^{1-\xi+\epsilon}} \end{aligned} \tag{21}$$

where (i) is due to the definition of ξ-adaptive buyers in Definition 4; (ii) is due to the fact that ϕ is an increasing function and the exploration episode lengths are $|\mathcal{E}_h| = T^{1-\xi+\epsilon}$; (iii) is due to the fact that all prices are less than 1.

Since $\phi(T) = \mathcal{O}(T^{1-\xi})$, there exists some $T_\epsilon \in \mathbb{N}$ such that when $T > T_\epsilon$ we have

$$\frac{\phi(T)}{T^{1-\xi+\epsilon}} < \frac{G}{2} \tag{22}$$

The rest of the proof relies on the following lemma:

Lemma 11. *Assume $T > T_\epsilon$ s.t. Equation* (22) *holds. If $\hat{\pi}(D_i) \geq \hat{\pi}(D_j)$ for some exploration episodes i, j s.t. $i \neq j$, then w.p. at least $1 - \frac{1}{T}$, $\pi(D_i) \geq \pi(D_j)$. Furthermore, the following event $\mathcal{G}$*

$$\mathcal{G} = \{\hat{\pi}(D_i) \geq \hat{\pi}(D_j) \Longrightarrow \pi(D_i) \geq \pi(D_j) \text{ for all exploration episodes } i \neq j\} \tag{23}$$

holds with probability at least $1 - \frac{H(H-1)}{2T}$, where $H = \lfloor \log_2(M) \rfloor + 1$ is the maximum number of binary search iterations (i.e. number of episodes in the exploration phase).

Proof of Lemma 11. Because $\hat{\pi}(D_i) \geq \hat{\pi}(D_j)$, applying Equation (21) for episodes i, j yields

$$\pi(D_i) + \frac{\phi(T)}{T^{1-\xi+\epsilon}} \geq \hat{\pi}(D_i) \geq \hat{\pi}(D_j) \geq \pi(D_j) - \frac{\phi(T)}{T^{1-\xi+\epsilon}} \implies \frac{2\phi(T)}{T^{1-\xi+\epsilon}} \geq \pi(D_j) - \pi(D_i)\,,$$

Now, contrary to our claim, suppose that $\pi(D_i) < \pi(D_j)$. We then have

$$\frac{2\phi(T)}{T^{1-\xi+\epsilon}} \geq \pi(D_j) - \pi(D_i) \geq \min_{d,\widetilde{d}\in\mathcal{D}:\pi(d)\neq\pi(\widetilde{d})} \left|\pi(d) - \pi(\widetilde{d})\right| := G\,,$$

which contradicts Equation (22) for $T > T_\epsilon$. As there are $H(H-1)/2$ pairs (i, j) such that $i \neq j$, a simple union bound shows event $\mathcal{G}$ holds with probability at least $1 - \frac{H(H-1)}{2T}$. □

We now return to our proof of Theorem 4. We first show that under event $\mathcal{G}$ (see Equation (23)),the final price in the exploitation phase D_{m^*} is revenue-optimal, i.e. $\max_{d\in\mathcal{D}} \pi(d) = \pi\left(D_{m^*}\right)$

We use an induction argument that shows after each iteration of the binary search procedure in the exploration phase of Algorithm 1, $\pi(D_m) \leq \pi(D_{m^*})$ for all $m \leq \mathrm{L}$ and $m \geq \mathrm{R}$. The base case is the first iteration, where we have $\mathrm{L} = 1$, $\mathrm{R} = M$. If $m^* = \mathrm{L} = 1$, then under event $\mathcal{G}$ we get

$$\hat{\pi}(D_1) \geq \hat{\pi}(D_M) \overset{(i)}{\implies} \pi(D_1) \geq \pi(D_M)\,.$$

Hence after the first iteration $\pi(D_m) \leq \pi(D_{m^*})$ for any $m \leq \mathrm{L}$ and $m \geq \mathrm{R}$. The case for $m^* = \mathrm{R}$ follows from the same argument.

Now assume that the induction hypothesis holds, i.e. at the beginning of some iteration with the tuple $(\mathrm{L}, \mathrm{R}, m^*)$, we have $\pi(D_m) \leq \pi(D_{m^*})$ $m \leq \mathrm{L}$ and $m \geq \mathrm{R}$. According to Algorithm 1, we only need to show two cases in order to validate the induction procedure.

- **Case 1.** If $\hat{\pi}(D_{\text{med}}) < \hat{\pi}(D_{\text{med}+1})$, then we show $\pi(D_m) \leq \pi(D_{\text{med}+1})$ for all $m = 1 \ldots \text{med} + 1$
- **Case 2.** If $\hat{\pi}(D_{\text{med}}) \geq \hat{\pi}(D_{\text{med}+1})$, then we show $\pi(D_m) \geq \pi(D_{\text{med}})$ for all $m = \text{med} + 1 \ldots M$

Note that under Case 1., med $+ 1$ will be the new value of m^* in the next iteration (i.e. the next induction step). So by showing $\pi(D_m) \leq \pi(D_{\text{med}+1})$ for all $m = 1 \ldots \text{med} + 1$, we validate the induction hypothesis for the next induction step. A similar argument holds for Case 2.

Case 1. When $\hat{\pi}(D_{\text{med}}) < \hat{\pi}(D_{\text{med}+1})$, under event $\mathcal{G}$ (see Equation (23)) we have $\pi(D_{\text{med}}) \leq \pi(D_{\text{med}+1})$. We claim that D_{med} cannot be an ROI binding price. Assume the contrary that D_{med} is ROI binding. Then, part (3) of Theorem 3 states $\pi(D_{\text{med}+1}) < \pi(D_{\text{med}})$, leading to a contradiction. Hence D_{med} must be either a nonbinding price or a budget binding price. Applying part (1) of Theorem 3, we can then conclude that for any $m \leq \text{med}$, $\pi(D_m) \leq \pi(D_{\text{med}})$, so

$$\pi(D_m) \leq \pi(D_{\text{med}}) \leq \pi(D_{\text{med}+1}) \quad \forall m = 1 \ldots \text{med}\,.$$

At the end of the iteration, as we update $m^{*+} = \text{med} + 1$ (here we denote m^{*+} as the updated value to distinguish from its initial value at the start of the iteration), we have $\pi(D_{m^{*+}}) \geq \pi(D_{\text{med}+1}) \geq \pi(D_{\text{med}}) \ldots \pi(D_1)$. On the other hand, since $\hat{\pi}(D_{m^{*+}}) = \max_{m\in\{m^*,\text{med}+1\}} \hat{\pi}(D_m) \geq \hat{\pi}(D_{m^*})$, event $\mathcal{G}$ implies

$$\pi(D_{m^{*+}}) \geq \pi(D_{m^*}) \overset{(i)}{\geq} \pi(D_m) \quad \forall m = \mathrm{R} \ldots M\,,$$

where (i) follows from the induction hypothesis. Therefore, we have

$$\pi(D_{m^{*+}}) \geq \pi(D_m) \quad \forall m = \mathrm{R} \ldots M \text{ and } m = 1 \ldots \text{med} + 1\,,$$

and by realizing the tuple $(\text{med} + 1, \mathrm{R}, m^{*+})$ is the initial tuple for the next iteration concludes the induction step.

Case 2. The case when $\hat{\pi}(D_{\text{med}}) \geq \hat{\pi}(D_{\text{med}+1})$ follows from an identical argument, and we will omit the details. This concludes the induction proof.

The above implies that when the event $\mathcal{G} = \{\hat{\pi}(D_i) \geq \hat{\pi}(D_j) \Longrightarrow \pi(D_i) \geq \pi(D_j) \text{ for all } i,j \in [H]\}$ holds throughout the exploration phase, the above induction argument implies we have $\pi(D_{m^*}) \geq \pi(D_m)$for all $m \in [M]$. Hence $\pi(D_{m^*}) = \max_{d\in\mathcal{D}} \pi(d)$ w.p. at least $1 - \frac{H(H-1)}{2T}$ according to Lemma 11 where $H = \lfloor \log_2(M) \rfloor + 1$.

Furthermore, we point out that in each iteration of the binary search procedure the seller explores at most two prices. Hence the entire exploration phase, which consists of all periods in exploration episodes and we denote as $\mathcal{E}$, has length at most $2E\left(\lfloor \log_2(M) \rfloor + 1\right) = 2T^{1-\xi+\epsilon}\left(\lfloor \log_2(M) \rfloor + 1\right)$ periods. Therefore, the seller's regret can be upper bounded as

$$
\begin{aligned}
\text{Reg}_{\text{sell}} &= T \max_{d\in\mathcal{D}} \pi(d) - \sum_{t\in[T]} \mathbb{E}\left[d_t z_t\right] \\
&\leq |\mathcal{E}| + \sum_{t=|\mathcal{E}|+1}^{T} \max_{d\in\mathcal{D}} \pi(d) - \mathbb{E}\left[d_t z_t\right] \\
&\overset{(i)}{\leq} |\mathcal{E}| + \sum_{t\in[T]/\mathcal{E}} \mathbb{E}\left[\left(\pi(D_{m^*}) - D_{m^*} z_t\right) \mathbb{I}\left\{\mathcal{G}\right\}\right] + (T - |\mathcal{E}|)\mathbb{P}\left(\mathcal{G}^c\right) \\
&\leq |\mathcal{E}| + D_{m^*}(T - |\mathcal{E}|) \cdot \mathbb{E}\left[\frac{\pi(D_{m^*})}{D_{m^*}} - \frac{1}{T - |\mathcal{E}|} \sum_{t\in[T]/\mathcal{E}} z_t\right] + (T - |\mathcal{E}|)\mathbb{P}\left(\mathcal{G}^c\right) \\
&\overset{(ii)}{\leq} |\mathcal{E}| + \phi\left(T - |\mathcal{E}|\right) + (T - |\mathcal{E}|) \cdot \mathbb{P}\left(\left|\frac{\pi(D_{m^*})}{D_{m^*}} - \frac{1}{T - |\mathcal{E}|} \sum_{t\in[T]/\mathcal{E}} z_t > \frac{\phi(T - |\mathcal{E}|)}{T - |\mathcal{E}|}\right|\right) + T\mathbb{P}\left(\mathcal{G}^c\right) \\
&\overset{(iii)}{\leq} |\mathcal{E}| + \phi\left(T - |\mathcal{E}|\right) + 1 + T\mathbb{P}\left(\mathcal{G}^c\right) \\
&\overset{(iv)}{\leq} 2\left(\lfloor \log_2(M) \rfloor + 1\right) \cdot T^{1-\xi+\epsilon} + \phi\left(T\right) + \left(\lfloor \log_2(M) \rfloor + 1\right)^2 / 2 \,.
\end{aligned}
$$

In (i) we used the fact that $\max_{d\in\mathcal{D}} \pi(d) = \pi(D_{m^*})$ under event $\mathcal{G}$ and $d_t = D_{m^*}$ for all exploitation periods $t \in [T]/\mathcal{E}$; in (ii) and (iii) we used the definition of ξ-adaptive buyer algorithm (see Definition 4) so that for the exploitation phase $[T]/\mathcal{E}$, the event $\left|\frac{\pi(D_{m^*})}{D_{m^*}} - \frac{1}{T-|\mathcal{E}|}\sum_{t\in[T]/\mathcal{E}} z_t\right| \leq \frac{\phi(T-|\mathcal{E}|)}{T-|\mathcal{E}|}$ holds with probability at least $1 - 1/T$, and also ϕ is an increasing function; In (iv), we used the fact that all periods in exploration episodes $\mathcal{E}$, has length at most $2E\left(\lfloor \log_2(M) \rfloor + 1\right) = 2T^{1-\xi+\epsilon}\left(\lfloor \log_2(M) \rfloor + 1\right)$ periods, and the fact that $\mathbb{P}\left(\mathcal{G}^c\right) \leq \frac{\left(\lfloor \log_2(M) \rfloor + 1\right) \cdot \lfloor \log_2(M) \rfloor}{2T}$ according to Lemma 11, so $1 + T\mathbb{P}\left(\mathcal{G}^c\right) \leq \left(\lfloor \log_2(M) \rfloor + 1\right)^2 / 2$ given $M \geq 2$. □

E PROOFS FOR SECTION 5

E.1 Proof of Lemma 5

Recall that when the buyer best responds, she adopts the threshold strategy w.r.t $\boldsymbol{x}_d$ where $\boldsymbol{x}_d \in [0,1]^N$ is the optimal solution to $U(d)$ in Equation (3); see Definition 2 for best response. Further, the threshold strategy can be represented as decision

$$
z_t^* = \sum_{n\in[N]} x_{d,n} \mathbb{I}\{v_t = V_n\} \,.
$$

Then, for any exploration or exploitation episode $\mathcal{E}$ (whose posted price we denote as d), for the best response decisions $\{z_t\}_{t\in\mathcal{E}}$ defined above, we have for any $t \in \mathcal{E}$

$$
\mathbb{E}[d z_t^*] = d \sum_{n\in[N]} g_n x_{d,n} = \pi(d)
$$

where $\pi(d)$ is the per-period expected revenue defined in Equation (4). Hence, by defining

$$
Y_t = d z_t^* - \pi(d)
$$

we know that the sequence $\{Y_t\}_{t\in\mathcal{E}}$ is a martingale difference sequence such that that $|Y_t| \le d \le 1$ for all $t \in \mathcal{E}$. By Azuma Hoeffding's inequality (see Lemma 13) we have for any $\delta \in (0,1)$

$$\mathbb{P}\left(\left|d\sum_{t\in\mathcal{E}} z_t^* - |\mathcal{E}|\cdot\pi(d)\right| > \sqrt{2|\mathcal{E}|\log(2/\delta)}\right) \le \delta\,.$$

Hence, by taking $\delta = 1/T$ and considering the increasing function $\phi(x) = \sqrt{2x\log(2T)} = \mathcal{O}(x^{1/2})$, for any exploration/exploitation episode $\mathcal{E}$ (whose price we denote as d) we have

$$\left|\frac{d}{|\mathcal{E}|}\sum_{t\in\mathcal{E}} z_t^* - \pi(d)\right| \le \frac{\phi(|\mathcal{E}|)}{|\mathcal{E}|}$$

with probability (w.p.) at least $1 - 1/T$. Therefore best responding is $\frac{1}{2}$-adaptive. □

E.2 Proof of Theorem 6

For the seller, the regret upper bound is a direct result of Lemma 5 and Theorem 4.

On the other hand for the buyer, following the exact proof Step 2. in the proof of Proposition 2 (see Appendix C.2), in particular Equations (17), (18) and (19), we know that by best responding, the buyer's budget and ROI constraints are satisfied. Finally, we bound the buyer's regret (see Definition in (7)) as followed:

Let d be the posted price in the final exploitation episode (see Algorithm 1). Using an argument similar to Step 1. in the proof of Proposition 2, for the linear program (LP) $U(d)$ in Equation (3) that denotes the buyer's single-period myopic optimization problem, it is easy to see the optimal value is bounded and the LP is feasible (consider the solution with all entries set to be 0). Then, strong duality holds, and there exists corresponding optimal dual variables $(\lambda,\mu) \in \mathbb{R}_+^2$ w.r.t. the exploitation price d s.t.

$$\begin{aligned} U(d) &= \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} g_n\left((1+\lambda)V_n - (\gamma\lambda+\mu)d\right)x_n + \rho\mu \\ &= \sum_{n\in[N]} g_n\left((1+\lambda)V_n - (\gamma\lambda+\mu)d\right)_+ + \rho\mu \end{aligned} \tag{24}$$

Similar to Equation (16), by denoting $\mathcal{E}$ to be all periods within exploration episodes, the buyer's hindsight objective can be bounded as

$$\begin{aligned} \text{B-OPT}(\boldsymbol{d}_{1:T}) &\le \max_{\boldsymbol{z}\in[0,1]^T} \sum_{t\in[T]} \mathbb{E}\left[\left((1+\lambda)v_t - (\gamma\lambda+\mu)d\right)z_t\right] + T\rho\mu \\ &\le \sum_{t\in[T]} \mathbb{E}\left[\left((1+\lambda)v_t - (\gamma\lambda+\mu)d\right)_+\right] + T\rho\mu \\ &= \sum_{t\in[T]}\left(\sum_{n\in[N]} g_n\left((1+\lambda)V_n - (\gamma\lambda+\mu)d\right)_+ + \rho\mu\right) \\ &\le (1+\lambda+\rho\mu)\cdot|\mathcal{E}| + \sum_{t\in[T]/\mathcal{E}}\left(\sum_{n\in[N]} g_n\left((1+\lambda)V_n - (\gamma\lambda+\mu)d\right)_+ + \rho\mu\right) \\ &\overset{(i)}{=} \Theta(T^{\frac{1}{2}+\epsilon}) + (T-|\mathcal{E}|)U(d)\,. \end{aligned} \tag{25}$$

Here (i) follows from Equation (24) and the fact that there are at most $2\left(\lfloor\log_2(M)\rfloor+1\right)$ exploration episodes, which implies in $\mathcal{E}$ there are at most $2T^{1-\xi+\epsilon}\left(\lfloor\log_2(M)\rfloor+1\right) = \Theta(T^{\frac{1}{2}+\epsilon})$ periods. The buyer's regret can be thus bounded as followed

$$\text{Reg}_{\text{buy}} = \text{B-OPT}(\boldsymbol{d}_{1:T}) - \sum_{t\in[T]}\mathbb{E}\left[v_t z_t\right] \le \Theta(T^{\frac{1}{2}+\epsilon}) + (T-|\mathcal{E}|)U(d) - \sum_{t\in[T]/\mathcal{E}}\mathbb{E}\left[v_t z_t\right] = \Theta(T^{\frac{1}{2}+\epsilon})$$

where in the final equality, we used the fact that the buyer's expected utility is exactly $U(d)$ for each exploitation period when best responding as shown in Equation (20). □

E.3 Proof of Theorem 7

This proof consists of two parts, namely bounding seller's regret, and bounding buyer's regret as well the "balance" of buyer's budget and ROI constraints.

Part 1. Bounding seller's regret. Here, we only need to show that the buyer's strategy is ξ-adaptive (see Definition in 4), and the rest of the proof follows from Theorem 4.

For notation convenience, fix some exploration or exploitation episode $\mathcal{E}$, and denote the corresponding price in the episode as d. In light of Lemma 1, we let $\boldsymbol{x}_d \in [0,1]^N$ be the unique optimal threshold vector (see Definition 6) solution to $U(d)$. According to the definition of the per-period seller expected revenue $\pi(d)$ under buyer best response in Equation (4), we can further write the seller's per-period expected revenue for episode h as

$$\pi(d) = d \sum_{n\in[N]} x_{d,n} g_n \,. \tag{26}$$

Let $\mathcal{F}_t$ be the sigma algebra generated by $\{(v_\tau, d_\tau, z_\tau)\}_{\tau\in[t]}$, which characterizes all randomness in the buyer and seller's behavior up to period t. Recall $\hat{\boldsymbol{x}}_t$ is the optimal solution to $U(d_t)$ of Equation (3) via replacing the true distribution $\boldsymbol{g} \in \Delta_N$ with the estimate $\hat{\boldsymbol{g}}_t \in \Delta_N$. The buyer adopting a threshold strategy w.r.t. $\hat{\boldsymbol{x}}_t$ implies the buyer's decision to be

$$z_t = \sum_{n\in[N]} \hat{x}_{t,n} \mathbb{I}\{v_t = V_n\} \tag{27}$$

Since $\hat{\boldsymbol{x}}_t$ is $\mathcal{F}_{t-1}$-measurable, for $t \in \mathcal{E}$ we have

$$\mathbb{E}\left[z_t \middle| \mathcal{F}_{t-1}\right] \;=\; \sum_{n\in[N]} g_n \hat{x}_{t,n}$$

Thus, the by defining

$$Y_t = \sum_{n\in[N]} g_n \hat{x}_{t,n} - z_t \,, \tag{28}$$

we know that the sequence $\{Y_t\}_{t\in\mathcal{E}}$ is a martingale difference sequence such that that $|Y_t| \le 1$ for all t. By Azuma Hoeffding's inequality (see Lemma 13) we have for any $\delta \in (0,1)$

$$\mathbb{P}\left(\widetilde{\mathcal{G}}\right) \ge 1-\delta \text{ where } \widetilde{\mathcal{G}} := \left\{ \left| \sum_{t\in\mathcal{E}} \left(\sum_{n\in[N]} g_n \hat{x}_{t,n} - z_t \right) \right| \le \sqrt{2|\mathcal{E}| \log(2/\delta)} \right\} . \tag{29}$$

The remaining proof relies on the following lemma whose proof can be found in Appendix E.5

Lemma 12. *Fix some price d and define the following problem which is solved by the approximate best response buyer with ML advice to obtain $\hat{\boldsymbol{x}}_t$ (see Definition 5):*

$$\hat{U}_t(d) = \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} \hat{g}_{t,n} V_n x_n \quad \textit{s.t.} \sum_{n\in[N]} \hat{g}_{t,n} (V_n - \gamma d) x_n \ge 0 \quad \textit{and} \quad d \sum_{n\in[N]} \hat{g}_{t,n} x_n \le \rho \,. \tag{30}$$

Here, recall $\hat{\boldsymbol{g}}_t \in \Delta_N$ is the ML advice obtained in period t which is an estimate for the true value distribution $\boldsymbol{g} \in \Delta_N$. Further, define the following values

$$(A) = \left(U(d) - \sum_{n\in[N]} g_n V_n \hat{x}_{t,n} \right)_+ , \quad (B) = \left(- \sum_{n\in[N]} g_n (V_n - \gamma d) \hat{x}_{t,n} \right)_+ , \quad (C) = \left(d \sum_{n\in[N]} g_n \hat{x}_{t,n} - \rho \right)_+ , \tag{31}$$

where we recall $U(d)$ is defined in Equation (3). Then, the values $(B), (C)$ are upper bounded by $\sqrt{N}\|\boldsymbol{g} - \hat{\boldsymbol{g}}_t\|$ for all t. Further, because the estimation error $\lim_{t\to\infty} \ell_t = 0$ there exists some $T_0 \in \mathbb{N}$ s.t. $\|\boldsymbol{g} - \hat{\boldsymbol{g}}_t\| \le \ell_t < \frac{g_1}{2}$ for all $t > T_0$. Then, there exists an absolute constant C that only depends on buyer model primitives $(\boldsymbol{g}, \boldsymbol{V}, \rho, \gamma)$ s.t. the values (A) and $\|\boldsymbol{x}_d - \hat{\boldsymbol{x}}_t\|$ are upper bounded by $C\sqrt{N}\|\boldsymbol{g} - \hat{\boldsymbol{g}}_t\|$ for $t > T_0$, where $\boldsymbol{x}_d$ is the optimal solution to $U(d)$.

We now show a high probability bound for $\frac{\pi(d)}{d} - \sum_{t\in\mathcal{E}} z_t$. Assume event $\widetilde{\mathcal{G}}$ (Equation (29)) holds, then

$$\begin{aligned}
\left|\sum_{t\in\mathcal{E}}\left(\frac{\pi(d)}{d} - z_t\right)\right| &= \left|\sum_{t\in\mathcal{E}}\left(\sum_{n\in[N]} x_{d,n}g_n - z_t\right)\right| \\
&\le \left|\sum_{t\in\mathcal{E}}\left(\sum_{n\in[N]} \hat{x}_{t,n}g_n - z_t\right)\right| + \sum_{t\in\mathcal{E}}\left|\sum_{n\in[N]}\hat{x}_{t,n}g_n - \sum_{n\in[N]} x_{d,n}g_n\right| \\
&\overset{(i)}{\le} \sqrt{2|\mathcal{E}|\log(2T)} + \sum_{t\in\mathcal{E}}\|\boldsymbol{x}_d - \hat{\boldsymbol{x}}_t\|\cdot\|\boldsymbol{g}\| \\
&\overset{(ii)}{\le} \sqrt{2|\mathcal{E}|\log(2T)} + T_0 + C\sqrt{N}\sum_{t\in\mathcal{E}:t>T_0}\ell_t \\
&\le \sqrt{2|\mathcal{E}|\log(2T)} + T_0 + C\sqrt{N}\sum_{t\in\mathcal{E}}\ell_t \\
&\overset{(iii)}{\le} \sqrt{2|\mathcal{E}|\log(2T)} + T_0 + C\sqrt{N}\widetilde{\phi}(|\mathcal{E}|) \\
&:= \phi(|\mathcal{E}|)
\end{aligned}$$

where in (i) we plugged in the Azuma-Hoeffding inequality result showed in Equation (29) with $\delta = \frac{1}{T}$; in (ii) we applied Lemma 12 and some constant absolute constant C for $t > T_0$ (defined in statement of Lemma 12), and the fact that $\|\boldsymbol{g}\| \le 1$ since $\boldsymbol{g}$ is a probability simplex; in (iii) we used the assumption that there exists some increasing function $\widetilde{\phi}$ s.t. $\sum_{t\in\mathcal{E}}\ell_t \le \widetilde{\phi}(|\mathcal{E}|)$. Therefore w.p. at least $1 - 1/T$ (since $\widetilde{\mathcal{G}}$ holds w.p. at least $1 - 1/T$ when $\delta = 1/T$), we have

$$\left|\frac{d}{|\mathcal{E}|}\sum_{t\in\mathcal{E}} z_t - \pi(d)\right| \le \frac{\phi(|\mathcal{E}|)}{|\mathcal{E}|}$$

Since $\widetilde{\phi}(x) \le \mathcal{O}(x^{1-L})$, we know that $\phi(x) = \mathcal{O}(x^{1-\xi})$ for $\xi = \min\{\frac{1}{2}, L\}$. Hence, for large enough T s.t. the exploration episode length $E = T^{1-\xi+\epsilon} > T_0$, the buyer's approximate best responding with ML advice is $1-\xi$-adaptive for $\xi = \min\{\frac{1}{2}, L\}$.

Part 2. Bounds for the buyer. We first follow a similar approach as the proof of Theorem 6 to upper bound the buyer regret.

Let d be the posted price in the final exploitation episode (see Algorithm 1), and denote $\mathcal{E} = \Theta(T^{1-\xi+\epsilon})$ as all periods within exploration episodes. Then using the same arguments as in Equations (24) and (25), we can show the buyer's hindsight objective can be bounded as

$$\text{B-OPT}(\boldsymbol{d}_{1:T}) \;\le\; \Theta(T^{\xi+\epsilon}) + (T - |\mathcal{E}|)U(d)\,.$$

Since the buyer approximately best responds w.r.t. $\hat{\boldsymbol{x}}_t$ which is the optimal solution to the problem $\hat{U}_t(d)$ Equation (30), recall the buyer's decision z_t can be written as in Equation (27):

$$z_t = \sum_{n\in[N]}\hat{x}_{t,n}\mathbb{I}\{v_t = V_n\}$$

Hence, $\mathbb{E}[v_t z_t|\mathcal{F}_{t-1}] = \sum_{n\in[N]} g_n V_n \hat{x}_{t,n}$. Let C and T_0 be defined as in Lemma 12, and thus the buyer's regret can be

thus bounded as followed

$$\begin{aligned}
\text{Reg}_{\text{buy}} &= \text{B-OPT}(\boldsymbol{d}_{1:T}) - \sum_{t\in[T]} \mathbb{E}\left[v_t z_t\right] \\
&\leq \Theta(T^{\xi+\epsilon}) + \sum_{t\in[T]/\mathcal{E}} \mathbb{E}\left[(U(d) - g_n \hat{x}_{t,n})\right] \\
&\leq \Theta(T^{\xi+\epsilon}) + T_0 + \sum_{t\in[T]/\mathcal{E}:t>T_0} \mathbb{E}\left[(U(d) - g_n \hat{x}_{t,n})\right] \\
&\overset{(i)}{\leq} \Theta(T^{\xi+\epsilon}) + T_0 + C\sqrt{N} \sum_{t\in[T]/\mathcal{E}:t>T_0} \|\boldsymbol{g} - \hat{\boldsymbol{g}}_t\| \\
&\overset{(ii)}{\leq} \Theta(T^{\xi+\epsilon}) + T_0 + C\sqrt{N} \sum_{t\in[T]/\mathcal{E}} \ell_t \\
&\overset{(iii)}{\leq} \Theta(T^{\xi+\epsilon}) + T_0 + C\sqrt{N}\widetilde{\phi}(T - |\mathcal{E}|) \\
&\overset{(iv)}{=} \Theta(T^{\xi+\epsilon})\,.
\end{aligned}$$

In (i), we applied Lemma 12 for the value (A) defined in Equation (31); (ii) follows from the definition of the estimation errors $\ell_t \geq \|\boldsymbol{g} - \hat{\boldsymbol{g}}_t\|$; (iii) follows from the assumption that for any exploration or exploitation episode $\mathcal{E}_h$, the total error $\sum_{t\in\mathcal{E}_h} \ell_t$ is upper bounded by $\widetilde{\phi}(T - |\mathcal{E}|$ where $\widetilde{\phi}$ is an increasing function; (iv) follows from the fact that $\widetilde{\phi}(x) \leq \mathcal{O}(T^{1-L}) \leq \mathcal{O}(T^{1-\xi})$.

Now we show the buyer constraint violation is small, namely

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} \left(v_t - \gamma d_t\right) z_t\right] \geq -\Theta(T^{-L}) \quad \text{and} \quad \frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} d_t z_t\right] \leq \rho + \Theta(T^{-L})\,.$$

The proofs for both inequalities are very similar, so here we just show $\frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} \left(v_t - \gamma d_t\right) z_t\right] \geq -\Theta(T^{-L})$. Similar to the above where we bounded buyer's regret, we have $\mathbb{E}[\left(v_t - \gamma d_t\right) z_t | \mathcal{F}_{t-1}] = \sum_{n\in[N]} g_n \left(V_n - \gamma d\right) \hat{x}_{t,n}$, and thus for all exploration and exploitation episodes $\mathcal{E}_1 \ldots \mathcal{E}_H$ (assuming there are H episodes), we have

$$\begin{aligned}
-\mathbb{E}\left[\sum_{t\in[T]} \left(v_t - \gamma d_t\right) z_t\right] &= \sum_{t\in[T]} \mathbb{E}\left[-\left(\sum_{n\in[N]} g_n \left(V_n - \gamma d\right) \hat{x}_{t,n}\right)\right] \\
&\leq \sum_{t\in[T]} \mathbb{E}\left[\left(-\sum_{n\in[N]} g_n \left(V_n - \gamma d\right) \hat{x}_{t,n}\right)_+\right] \\
&\overset{(i)}{\leq} \sqrt{N} \sum_{t\in[T]} \ell_t \\
&= \sqrt{N} \sum_{h\in[H]} \sum_{t\in\mathcal{E}_h} \ell_t \\
&\overset{(ii)}{\leq} \sqrt{N} \sum_{h\in[H]} \mathcal{O}(|\mathcal{E}_h|^{1-L}) \\
&= \Theta(T^{1-L})
\end{aligned}$$

where (i) follows from the upper bound of (B) (Equation (31)) in Lemma 12; (ii) follows from the assumption that for any exploration and exploitation episode $\mathcal{E}_h$ the errors $\{\ell_t\}_t$ satisfy $\sum_{t\in\mathcal{E}_h} \ell_t \leq \widetilde{\phi}(|\mathcal{E}_h|)$ for some increasing function $\widetilde{\phi} : \mathbb{R}_+ \to \mathbb{R}^+$ and $\widetilde{\phi}(x) \leq \mathcal{O}(x^{1-L})$.

Finally, dividing both sides by T yields the desired bound $\frac{1}{T}\mathbb{E}\left[\sum_{t\in[T]} \left(v_t - \gamma d_t\right) z_t\right] \geq -\Theta(T^{-L})$. □

E.4 Proof of Theorem 8

We know that the empirical estimates $\hat{\boldsymbol{g}}_t \in \Delta_N$ for the buyer's value distribution $\boldsymbol{g} \in \Delta_N$ defined in Equation (8) follow a multinomial distribution, i.e. $\hat{\boldsymbol{g}}_t \sim \frac{1}{t}\text{Multinomial}(t, \boldsymbol{g})$. Therefore, applying Lemma 15 by taking $\delta = 1/T^2$, we have w.p. at least $1 - 1/T^2$ the following event holds

$$\mathcal{G}_t := \left\{ \|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \le \ell_t := \sqrt{\frac{2N\log(2T^2)}{t}} \right\} \tag{32}$$

Here we used the fact that $\|\boldsymbol{x}\| \le \|\boldsymbol{x}\|_1$ for any vector $\boldsymbol{x}$. Hence, using a simple union bound, the event $\cup_{t\in[T]}\mathcal{G}_t$ holds w.p. at least $1 - 1/T$. Further, for any exploration or exploitation episode $\mathcal{E}$, we have

$$\sum_{t\in\mathcal{E}} \ell_t \le \sum_{\tau\in[|\mathcal{E}|]} \sqrt{\frac{2N\log(2T^2)}{\tau}} \le \widetilde{\phi}(|\mathcal{E}|) \tag{33}$$

for some increasing function $\widetilde{\phi}$ s.t. $\widetilde{\phi}(x) \le \mathcal{O}(x^{\frac{1}{2}})$. Hence, w.p. at least $1 - 1/T$, the estimation errors $\{\ell_t\}_t$ defined above satisfy the conditions in Theorem 7 for large enough T, i.e. $\lim_{t\to\infty} \ell_t = 0$ and $\sum_{t\in\mathcal{E}} \ell_t \le \widetilde{\phi}(|\mathcal{E}|)$ for any any exploration or exploitation episode $\mathcal{E}$ where increasing function $\widetilde{\phi} : \mathbb{R}_+ \to \mathbb{R}^+$ and $\widetilde{\phi}(x) \le \mathcal{O}(x^{1-L})$. The rest of the proof directly follows from Theorem 7.

□

E.5 Proof of Lemma 12

Consider the region

$$\mathcal{C} = \left\{ \boldsymbol{x} \in [0,1]^N : -\sum_{n\in[N]} g_n V_n x_n \le -U(d), -\sum_{n\in[N]} g_n \left(V_n - \gamma d\right) x_n \le 0, d\sum_{n\in[N]} g_n x_n \le \rho \right\} \tag{34}$$

By Lemma 1, we know that $\boldsymbol{x}_d$ is the unique optimal solution to $U(d)$, and hence $\mathcal{C}$ consists of the single point $\boldsymbol{x}_d$, namely $\mathcal{C} = \{\boldsymbol{x}_d\}$. Now consider the optimal solution $\hat{\boldsymbol{x}}_t \in [0,1]^N$ to $\hat{U}_t(d)$ in Equation (30), by the Hoffman bound (Lemma 14), there exists some constant $H > 0$ that only depends on $(\boldsymbol{g}, \boldsymbol{V})$ s.t.

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_d\| \le H \left(\underbrace{\left(U(d) - \sum_{n\in[N]} g_n V_n \hat{x}_{t,n} \right)_+}_{(A)} + \underbrace{\left(-\sum_{n\in[N]} g_n \left(V_n - \gamma d\right) \hat{x}_{t,n} \right)_+}_{(B)} + \underbrace{\left(d\sum_{n\in[N]} g_n \hat{x}_{t,n} - \rho \right)_+}_{C} \right) \tag{35}$$

where we used the inequality $\|(\boldsymbol{y})_+\| \le \sum_{n\in[N]} (y_n)_+$ for any vector $y \in \mathbb{R}^N$. We now bound (A), (B) and (C) respectively.

Bounding (A). Similar to the proof of Theorem 6, strong duality holds for the LP $\hat{U}_t(d)$, and hence there exists optimal dual variables $\hat{\lambda}, \hat{\mu} \in \mathbb{R}_+$ s.t.

$$\hat{U}_t(d) = \sum_{n\in[N]} g_n V_n \hat{x}_{t,n} = \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} \hat{g}_{t,n} \left((1+\hat{\lambda})V_n - (\gamma\hat{\lambda} + \hat{\mu})d \right) x_n + \rho\hat{\mu} \tag{36}$$

Since $\hat{U}_t(d) \le 1$, it is easy to see $\hat{\mu} \in [0, 1/\rho]$, and further by considering $\boldsymbol{x} = (1, 0 \dots 0) \in \mathbb{R}^N$, we have

$$\begin{aligned} 1 \ge \hat{U}_t(d) \ &\ge\ \hat{g}_{t,1} \left((1+\hat{\lambda})V_1 - (\gamma\hat{\lambda} + \hat{\mu})d \right) \ \overset{(i)}{\ge}\ \hat{g}_{t,1}\hat{\lambda}(V_1 - \gamma d) - \hat{\mu}d \ \overset{(ii)}{\ge}\ \frac{g_1}{2} \cdot \hat{\lambda}(V_1 - \gamma d) - \frac{1}{\rho} \\ \overset{(iii)}{\Rightarrow}\ \hat{\lambda} &\le 2\left(1 + \frac{1}{\rho}\right) \frac{V_1 - \gamma D_1}{g_1}, \end{aligned} \tag{37}$$

where in (i) we used the fact that $\hat{g}_{t,1} \in [0,1]$; in (ii) we used the fact that $|\hat{g}_{t,1} - g_1| \leq \|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \leq \ell_t < \frac{g_1}{2}$ for all $t > T_0$, and also $d \in [0,1]$ as well as $\hat{\mu} \in [0, 1/\rho]$; in (iii), we used Assumption 1 s.t. $V_1 - \gamma d > 0$ for all $d \in \mathcal{D}$, and $g_1 > 0$.

On the other hand, we have

$$\begin{aligned} U(d) \;&\leq\; \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} g_n \Big((1+\hat{\lambda})V_n - (\gamma\hat{\lambda}+\hat{\mu})d\Big) x_n + \rho\hat{\mu} \\ &\overset{(i)}{\leq} \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} \hat{g}_{t,n} \Big((1+\hat{\lambda})V_n - (\gamma\hat{\lambda}+\hat{\mu})d\Big) x_n + (1+\hat{\lambda}) \sum_{n\in[N]} |\hat{g}_{t,n} - g_n| + \rho\hat{\mu} \\ &\overset{(ii)}{\leq} \max_{\boldsymbol{x}\in[0,1]^N} \sum_{n\in[N]} \hat{g}_{t,n} \Big((1+\hat{\lambda})V_n - (\gamma\hat{\lambda}+\hat{\mu})d\Big) x_n + \rho\hat{\mu} + (1+\hat{\lambda})\sqrt{N}\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \\ &\overset{(iii)}{=} \hat{U}_t(d) + 2\left(1+\frac{1}{\rho}\right)\frac{V_1 - \gamma D_1}{g_1} \cdot \sqrt{N}\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \end{aligned} \tag{38}$$

In (i), we used the fact that for all $n \in [N]$, $x_n \in [0,1]$ and $(1+\hat{\lambda})V_n - (\gamma\hat{\lambda}+\hat{\mu})d \leq (1+\hat{\lambda})V_n \leq 1+\hat{\lambda}$ since all possible values $V_n \in [0,1]$; (ii) applies Cauchy–Schwarz inequality; (iii) plugs in Equation (36) and (37).

Therefore, if $\hat{U}_t(d) = \sum_{n\in[N]} g_n V_n \hat{x}_{t,n} \geq U(d)$, then (A) = 0, whereas if $\hat{U}_t(d) = \sum_{n\in[N]} g_n V_n \hat{x}_{t,n} < U(d)$, Equation (38) implies

$$(A) \leq 2\left(1+\frac{1}{\rho}\right)\frac{V_1 - \gamma D_1}{g_1}\sqrt{N}\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \tag{39}$$

Bounding (B) and (C). The bounds for (B) and (C) are similar, and therefore we only show that for (B).

$$\begin{aligned} (B) = \left(-\sum_{n\in[N]} g_n\left(V_n - \gamma d\right)\hat{x}_{t,n}\right)_+ &\overset{(i)}{\leq} \left(-\sum_{n\in[N]} \hat{g}_{t,n}\left(V_n - \gamma d\right)\hat{x}_{t,n}\right)_+ + \left|\sum_{n\in[N]} \left(\hat{g}_{t,n} - g_n\right)\left(V_n - \gamma d\right)\hat{x}_{t,n}\right| \\ &\overset{(ii)}{\leq} \sum_{n\in[N]} |\hat{g}_{t,n} - g_n| \\ &\overset{(iii)}{\leq} \sqrt{N}\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \end{aligned} \tag{40}$$

Here, (i) follows from the basic inequality sequence $(a+b)_+ \leq (a)_+ + (b)_+ \leq (a)_+ + |b|$; (ii) follows from the fact that $\hat{\boldsymbol{x}}_t$ is feasible to $\hat{U}_t(d)$ so that $\sum_{n\in[N]} \hat{g}_{t,n}\left(V_n - \gamma d\right)\hat{x}_{t,n} \geq 0$, and also $|V_n - \gamma d| \leq V_n \leq 1$ and $\hat{x}_{t,n} \in [0,1]$; (iii) follows from the Cauchy–Schwarz inequality.

We can similarly show

$$(C) \leq \sqrt{N}\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\| \tag{41}$$

Finally, combining Equations (35), (39), (40), and (41) yields the desired result. □

F SUPPLEMENTARY LEMMAS

Lemma 13 (Azuma–Hoeffding inequality)**.** *beLet $Y_1 \ldots Y_n$ be a martingale difference sequence with a uniform bound $|Y_j| \leq 1$ for all $j \in [n]$. Then for any $\delta \in (0, 1/e)$,*

$$\mathbb{P}\left(\left|\sum_{j\in[n]} Y_j\right| > \sqrt{2n\log(2/\delta)}\right) \leq \delta\,.$$

Lemma 14 (Hoffman bound Hoffman (2003))**.** *Consider a non-empty linear region $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ for some $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then, there exists some constant $H > 0$ that only depends on $\boldsymbol{A}$ s.t. for any $\boldsymbol{y} \in \mathbb{R}^n$ we have $\inf_{\boldsymbol{z}\in\mathcal{C}}\|\boldsymbol{z} - \boldsymbol{y}\| \leq H\|(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b})_+\|$. Here $(\boldsymbol{y})_+$ is the vector that takes the positive parts for each entry in $\boldsymbol{y}$, i.e. $(\boldsymbol{y})_+ = ((y_1)_+ \ldots (y_n)_+)$.*

Lemma 15 (Empirical distribution concentration inequality Weissman et al. (2003))**.** *Let $\boldsymbol{g} \in \Delta_N$ be a N-dimensional probability simplex ($N \geq 2$), and $\hat{\boldsymbol{g}}_t \sim \frac{1}{t}$Multinomial$(t, \boldsymbol{g})$. Then for any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\|\hat{\boldsymbol{g}}_t - \boldsymbol{g}\|_1 > \sqrt{\frac{2N\log(2/\delta)}{t}}\right) \leq \delta$$

See also Qian et al. (2020) Proposition 2. for a similar statement to that of Lemma 15.