OPTIMAL ONLINE AUTOBIDDING IN FIRST-PRICE AUCTIONS WITH ROI CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a value-maximizing autobidder in T repeated first-price auctions under an expected return-on-investment (ROI) constraint. Values and competing bids are drawn i.i.d. from two unknown, mutually independent distributions, and the bidder observes only whether it wins. The benchmark knows the competing-bid distribution and the entire value sequence and may be randomized. We show that an optimal benchmark randomizes in at most one round. Our algorithm, UCB-DMD, combines action-adaptive confidence estimates with a consistent optimistic Lagrangian and dual update. It achieves $O(T^{2/3}(\log T)^{1/3})$ expected regret and expected ROI violation, improving the previous $\widetilde{O}(T^{3/4})$ expected-regret bound without requiring continuity of the competing-bid distribution. Under a fixed positive-surplus condition, a surplus-target modification of of UCB-DMD guarantees the ROI constraint is satisfied in expectation for all sufficiently large horizons with the same regret rate without any additional input for the algorithm. A joint regret and ROI violation lower bound establishes optimality of the common rate up to logarithms. The lower-bound instance also satisfies the fixed positive-surplus condition and establishes near-optimal regret among algorithms that satisfy the expected ROI constraint.

1 INTRODUCTION

Autobidding plays a central role in online advertising: advertisers specify a value-maximization objective and set constraints such as a budget or target return on ad spend, while the platform selects bids on their behalf (Aggarwal et al., 2024). The platform uses machine learning algorithms to predict values for each auction; the bidding algorithm then uses these estimates to optimize bids that balance the obtained value and payment.

For a value-maximizing bidder, ROI couples the value obtained with the corresponding payment. Thus, the optimal bidding strategy must trade off the probability of winning against the payment made upon winning (Li et al., 2026; Aggarwal et al., 2025). The objective is total value. Wins with positive surplus can compensate for losses on other wins. Consequently, requiring every bid to be below its value can be unnecessarily restrictive. The learner must identify both valuable allocations and the surplus that can support them. Although randomized bidding can be better, as illustrated by Li et al. (2026), we show that an optimal bidding strategy randomizes in at most one round if the algorithm knows the values in advance. This result clarifies the limited role of randomization in the offline benchmark.

Learning which bids to select remains a separate challenge. Binary feedback does not reveal the competing bid, and nearby prices need not have similar winning probabilities: an atom of the competing-bid distribution creates a jump. Learning on a finite bid grid therefore requires care both in estimating outcomes and in relating grid bids to the unrestricted benchmark. Under bandit feedback, Li et al. (2026) achieve $\widetilde{O}(T^{3/4})$ regret through an explore-then-optimize approach using an optimistic distribution estimate and upward bid adjustments. We also separate virtual and executed prices. The distinction is that UCB-DMD updates action-specific confidence estimates throughout the horizon and uses the same optimistic surplus in bid selection and dual mirror descent. Its analysis compares directly against the original randomized benchmark and does not require a concave-envelope computation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The key observation is that the observed allocation, together with the known value and submitted bid, determines both the value obtained and the corresponding payment. Learning the probability of winning therefore informs both the reward objective and the ROI constraint. UCB-DMD uses this shared information consistently when selecting bids and adjusting its emphasis on ROI, rather than treating the two tasks independently. This consistency is central to the improved guarantee.

A vanishing surplus target in the dual update additionally gives terminal expected ROI feasibility under a fixed positive-surplus condition. We also prove a joint lower bound on regret and ROI violation, rather than importing the budget-only lower bound of Aggarwal et al. (2025). These results close the rate gap left by Li et al. (2026) for the ROI-only model, up to logarithms.

To summarize, our contributions are as follows.

- **Structure of the benchmark.** An optimal offline policy randomizes in at most one round and uses at most two bids there. The best deterministic bidding plan is within one unit of the expected value of the randomized benchmark (Theorem 1).
- **Two online learning algorithms** UCB-DMD achieves $O(T^{2/3}(\log T)^{1/3})$ expected regret and expected ROI violation (Theorem 2 and Proposition 1). Under Assumption 1, a modified algorithm retains this regret rate without ROI violation in expectation (Theorem 3).
- **Matching lower bounds.** Any algorithm with $o(T^{2/3})$ expected ROI violation has $\Omega(T^{2/3})$ worst-case regret (Theorem 4). The common rate for regret and ROI violation is therefore optimal up to a logarithmic factor.

2 Related Work

Online bidding in repeated auctions. The recent work studies online bidding algorithms in repeated auctions. Balseiro & Gur (2019) and Balseiro et al. (2020) study budget-constrained bidding in second-price auctions via a primal-dual approach, and Wang et al. (2023) consider the same setting in first-price auctions. Kumar et al. (2024) design strategically robust learning algorithms for first-price auctions, and Susan et al. (2023) consider the setting where bidders can manage budgets across several platforms. Earlier work by Weed et al. (2016) treats repeated auctions as bandit problems, and Feng et al. (2021) analyze the convergence of no-regret bidders. The most related work to ours considers no-regret bidding algorithms for ROI-constrained value maximizers. Feng et al. (2023) give an algorithm with nearly optimal $\widetilde{O}(\sqrt{T})$ regret under truthful auctions, building on the dual mirror descent framework of Balseiro et al. (2020; 2023). Castiglioni et al. (2024) obtain no-regret guarantees under budget and ROI constraints when the value and bid spaces are finite, and Vijayan et al. (2025) consider the harder setting in which the value must be learned. For non-truthful auctions, Aggarwal et al. (2025) obtain no regret against the best Lipschitz-continuous bidding strategy and provide an $\Omega(T^{2/3})$ lower bound under bandit feedback. Li et al. (2026) study the same setting with ours and they replace the Lipschitz benchmark by the unconditionally optimal randomized policy and obtain $\widetilde{O}(\sqrt{T})$ regret under full feedback and $\widetilde{O}(T^{3/4})$ under bandit feedback. Our work adopts the same benchmark and improve the regret to $\widetilde{O}(T^{2/3})$ under bandit feedback.

Bandits and online allocation with constraints. Our algorithm is related to the stochastic bandits in online allocation with constraints. The upper confidence bounds for stochastic bandits can go back to Auer et al. (2002). Discretizing a continuous action space into a grid is the basic approach to Lipschitz bandits (Kleinberg et al., 2008). Dual updates for long-run constraints stem from online linear programming and resource allocation (Mehta et al., 2007; Agrawal et al., 2014; Devanur et al., 2019; Balseiro et al., 2015) and in bandits with knapsacks (Badanidiyuru et al., 2018; Agrawal & Devanur, 2014; Immorlica et al., 2022; Kesselheim & Singla, 2020; Castiglioni et al., 2022; Fikioris & Tardos, 2023; Slivkins et al., 2023). Contextual bandits (Balseiro et al., 2019) is another way to exploit the structure of auction feedback in online bidding. The ROI constraint studied in our work differs from a resource allocation constraint (i.e., budget constraint in auction) in that it couples payment with the value obtained, so its slack can be negative and positive in different rounds, which is why we bound the violation separately from the regret.

Auctions with autobidders. Another related literature studies the market-level consequences of autobidding, such as the efficiency of standard auction mechanisms with value maximizers (Aggarwal et al., 2019; Deng et al., 2021; Liaw et al., 2023; Deng et al., 2024a;b; Feng et al., 2024) and

the design of optimal or approximately optimal mechanisms for such bidders (Balseiro et al., 2021; Deng et al., 2022; Lu et al., 2023; Lv et al., 2023; Balseiro et al., 2024; Bei et al., 2025). Aggarwal et al. (2024) survey this area in detail. Our paper considers the bidding problem of a single value maximizer with ROI constraint in the autobidding world, who faces a stable environment.

3 PRELIMINARIES

A single value-maximizing autobidder takes part in a sequence of T single-item first-price auctions. In round $t \in \{1, \dots, T\}$ the bidder observes its value $v_t \in [0, 1]$ and then submits a bid $b_t \in [0, 1]$. Let $d_t \in [0, 1]$ denote the competing bid, defined as the highest bid among the other participants. The bidder wins if $b_t \geq d_t$, with ties broken in favor of the bidder, and pays its bid when it wins. The realized allocation and payment are

$$x_t(b_t, d_t) = \mathbf{1}\{b_t \geq d_t\}, \qquad p_t(b_t, d_t) = b_t \, \mathbf{1}\{b_t \geq d_t\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The values $v_1, \dots, v_T$ are drawn independently from an unknown distribution F on $[0, 1]$, and the competing bids $d_1, \dots, d_T$ are drawn independently from an unknown distribution G on $[0, 1]$, independently of the values. We identify G with its cumulative distribution function, so $G(b) = \Pr(d_t \leq b)$ is nondecreasing and right-continuous, and for $c \in (0, 1]$ we write $G(c_-) = \lim_{b \uparrow c} G(b)$ for its left limit. For any bid b, the expected allocation and payment in a round are $G(b)$ and $bG(b)$.

The bidder aims to maximize its total expected value subject to a return-on-investment constraint with target $\gamma \geq 1$: the expected total value must be at least γ times the expected total payment. Dividing every value by γ leaves the set of feasible policies and the optimal policies unchanged and scales the objective by $1/\gamma$, so we assume $\gamma = 1$ without loss of generality; values remain in $[0, 1]$ because $\gamma \geq 1$. Write $V^T = (v_1, \dots, v_T)$ for the value sequence. A bidding policy $\pi = (\pi_1, \cdots, \pi_T)$ is a rule that selects a bid $b_t \sim \pi_t$ in each round, possibly at random, using the current value, the past values, and the past auction outcomes. Because d_t is independent of everything that determines b_t, the expected reward and payment of π on a fixed value sequence are

$$\mathrm{Reward}(\pi \mid G, V^T) = \mathbb{E}_\pi\Big[\sum_{t=1}^{T} v_t G(b_t)\Big], \qquad \mathrm{Payment}(\pi \mid G, V^T) = \mathbb{E}_\pi\Big[\sum_{t=1}^{T} b_t G(b_t)\Big],$$

where the expectation is over the policy's internal randomization and over the past auction outcomes that influence its bids. The offline benchmark is the value of the program

$$\begin{aligned} \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) = \quad & \sup_{\pi} \quad \mathrm{Reward}(\pi \mid G, V^T), \\ & \text{s.t.} \quad \mathrm{Reward}(\pi \mid G, V^T) \geq \mathrm{Payment}(\pi \mid G, V^T) \end{aligned} \tag{1}$$

where π ranges over all policies that know G and the entire sequence V^T in advance but do not observe d_t before bidding in round t. We show in Section 4 that the supremum is attained, and we call a policy attaining it OPT. Note that the ROI constraint in (1) is imposed in expectation.

The online learner knows neither F nor G. We study bandit feedback: at the end of round t the learner observes only the allocation x_t (and hence the payment $b_t x_t$), not the competing bid d_t. An online algorithm ALG maps the current value and the history of past values, bids, and allocations to a bid. Its signed expected ROI violation and its regret on an instance (G, V^T) are

$$\Delta(\mathrm{ALG} \mid G, V^T) = \mathrm{Payment}(\mathrm{ALG} \mid G, V^T) - \mathrm{Reward}(\mathrm{ALG} \mid G, V^T),$$
$$\mathrm{Regret}(\mathrm{ALG} \mid G, V^T) = \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \mathrm{Reward}(\mathrm{ALG} \mid G, V^T).$$

We evaluate both quantities in expectation over the value sequence:

$$\mathrm{Regret}(\mathrm{ALG} \mid G, F) = \mathbb{E}_{V^T \sim F^T}\big[\mathrm{Regret}(\mathrm{ALG} \mid G, V^T)\big],$$
$$\Delta(\mathrm{ALG} \mid G, F) = \mathbb{E}_{V^T \sim F^T}\big[\Delta(\mathrm{ALG} \mid G, V^T)\big],$$

where F^T denotes the product of T copies of F. Our goal is a sublinear upper bound on regret and on $[\Delta]_+$ for every (G, F), where $[z]_+ = \max\{z, 0\}$.

For the signed realized ROI violation $D_T^{\mathrm{real}} = \sum_{t=1}^{T} (b_t - v_t) x_t$, we have $\Delta = \mathbb{E}[D_T^{\mathrm{real}}]$. Expected ROI feasibility means $\Delta \leq 0$, averaging over values, feedback, and internal randomization. The offline constraint is conditional on each fixed V^T; the online guarantees average over V^T.

4 STRUCTURE OF THE OFFLINE BENCHMARK

Li et al. (2026) give an example in which randomized bidding strictly improves the expected value attainable under an ROI constraint in a first-price auction, so the benchmark in (1) cannot be restricted to deterministic bids without loss. We show that the loss is small: an optimal policy needs to randomize in at most one round, and the best deterministic plan is within one unit of value of the benchmark. Fix G and V^T. A deterministic bidding plan is a vector $(b_1, \dots, b_T) \in [0,1]^T$ chosen with knowledge of G and V^T, and the deterministic benchmark is

$$\mathrm{Reward}(\mathrm{det} \mid G, V^T) = \sup_{b_1,\dots,b_T \in [0,1]} \Big\{ \sum_{t=1}^{T} v_t G(b_t) \ : \ \sum_{t=1}^{T} (v_t - b_t) G(b_t) \geq 0 \Big\}. \tag{2}$$

Theorem 1. *Fix a competing-bid distribution G. For every value sequence $V^T \in [0,1]^T$, the supremum in (1) is attained by a policy that uses a deterministic bid in all but at most one round and, in that round, randomizes between two bids. Moreover,*

$$0 \leq \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \mathrm{Reward}(\mathrm{det} \mid G, V^T) \leq \max_{1 \leq t \leq T} v_t \leq 1. \tag{3}$$

Proof sketch. When G has atoms, the set of allocation–payment pairs induced by deterministic bids need not be closed, so attainment of the benchmark is not immediate. For every point in its closure, we construct a mixture of at most two actual bids that preserves its expected payment and achieves at least its winning probability. Using a compact relaxation, this construction establishes attainment and reduces the benchmark to finitely supported bid distributions. The resulting linear program has T probability normalization constraints and one aggregate ROI constraint, so an optimal basic solution has at most $T+1$ positive bid probabilities. Since every round needs at least one, only one round can mix two bids. Replacing that mixture by the bid with the larger expected ROI surplus preserves feasibility and loses at most that round's value. Appendix A.1 gives the full proof.

Here, a deterministic plan may use different bids in different rounds, even when their values are the same. This is an offline structural statement about the benchmark. Separately, UCB-DMD selects bids by directly maximizing an optimistic Lagrangian score, without computing a concave envelope, and its analysis compares against the original randomized benchmark.

5 ALGORITHM WITH SUBLINEAR ROI VIOLATION

We first present UCB-DMD and prove bounds on its regret and expected ROI violation. Section 6 then modifies the algorithm to satisfy the original ROI constraint in expectation, while still retaining the same regret rate under Assumption 1. The matching lower bounds are given in Section 7.

5.1 THE UCB-DMD ALGORITHM

Partition the bid space into K intervals of width $1/K$. For $k \in \{0, 1, \dots, K\}$ define the grid points and their winning probabilities

$$a_k = \frac{k}{K}, \qquad q_k = G(a_k).$$

For positive values, the algorithm submits bids from $\{a_1, \dots, a_K\}$; bid a_k is action k, with virtual price $c_k = a_{k-1}$. When $v_t = 0$, it instead sets $k_t = 0$ and bids $a_0 = 0$. Define $c_0 = 0$ and the bookkeeping value $U_{t,0} = 1$; no estimate of $G(0)$ is needed, since action 0 is used only when its reward and surplus are exactly zero. The zero-bid feedback is ignored. For $k \in \{1, \dots, K\}$, before round t let $N_{t,k}$ and $W_{t,k}$ denote the number of times action k was played and the number of wins it produced,

$$N_{t,k} = \sum_{s<t} \mathbf{1}\{b_s = a_k\}, \qquad W_{t,k} = \sum_{s<t} \mathbf{1}\{b_s = a_k\}\, x_s.$$

Fix a failure probability $\delta \in (0,1)$ and a confidence parameter $L \geq \log(2KT/\delta)$. The upper confidence bound on q_k is

$$U_{t,k} = \begin{cases} 1, & N_{t,k} = 0 \\ \min\left\{1, \ \dfrac{W_{t,k}}{N_{t,k}} + \sqrt{\dfrac{L}{2N_{t,k}}}\right\}, & N_{t,k} > 0. \end{cases} \tag{4}$$

The algorithm maintains a dual variable $\lambda_t \geq 0$ for the ROI constraint, initialized at $\lambda_1 = 1$. In round t, after observing v_t, it selects

$$k_t = \begin{cases} 0, & v_t = 0, \\ \min \arg\max_{1 \leq k \leq K} \big((1+\lambda_t)v_t - \lambda_t c_k\big) U_{t,k}, & v_t > 0. \end{cases} \tag{5}$$

using the smallest maximizing index when $v_t > 0$, and bids $b_t = a_{k_t}$. For $v_t > 0$, the score in (5) is an optimistic Lagrangian: it uses the upper confidence bound $U_{t,k}$ in place of the winning probability and the virtual price c_k in place of the price. The algorithm forms the surrogate ROI surplus and updates the dual variable using dual mirror descent with step size $\eta > 0$:

$$\widetilde{g}_t = (v_t - c_{k_t})\, U_{t,k_t}, \qquad \lambda_{t+1} = \lambda_t \exp(-\eta\, \widetilde{g}_t). \tag{6}$$

This is dual mirror descent with the entropy regularizer (Balseiro et al., 2020): a positive surplus lowers λ and makes the algorithm bid more aggressively, and a negative surplus raises it. Algorithm 1 summarizes the procedure.

Remark 1. *Downward rounding can lose a constant amount of expected value when G has jumps; upward rounding introduces a price error multiplied by λ_t. Using the winning probability at a_k and price c_k avoids this coupling. On $\mathcal{E}$, the maximum grid score dominates the continuous Lagrangian supremum without additive error (Lemma 4). The score and dual update share the same surplus surrogate, while the $1/K$ difference between the surrogate price and the submitted bid enters the ROI violation bound.*

Algorithm 1 UCB with Dual Mirror Descent (UCB-DMD)

Require: Horizon T, grid size K, step size η, failure probability δ, confidence parameter $L \geq \log(2KT/\delta)$

1: Set $N_{1,k} = W_{1,k} = 0$ for every $k \in \{1, \dots, K\}$ and $\lambda_1 = 1$
2: **for** $t = 1, \dots, T$ **do**
3: Compute $U_{t,k}$ for every k from $N_{t,k}$ and $W_{t,k}$ using (4)
4: Observe the value v_t and select k_t by (5)
5: Submit the bid $b_t = a_{k_t}$ and observe the allocation $x_t = \mathbf{1}\{d_t \leq b_t\}$
6: Compute $\widetilde{g}_t$ and λ_{t+1} by (6)
7: If $k_t > 0$, set $N_{t+1,k_t} \leftarrow N_{t,k_t} + 1$, $W_{t+1,k_t} \leftarrow W_{t,k_t} + x_t$; other counters unchanged
8: **end for**

5.2 Regret and ROI Guarantees

Theorem 2. *Let $T \geq 4$ and run Algorithm 1 with*

$$\eta = T^{-1/2}, \quad \delta = T^{-2}, \quad L = \log(2T^4), \quad K = \lceil (T/L)^{1/3} \rceil.$$

Then for every G and F,

$$\mathrm{Regret}(\mathrm{ALG} \mid G, F) \leq 6\sqrt{T} + K + 2\sqrt{2KTL} + \frac{1}{T} = O\big(T^{2/3}(\log T)^{1/3}\big).$$

Proposition 1. *Under the parameter choices of Theorem 2, for every G and F,*

$$\Delta(\mathrm{ALG} \mid G, F) \leq \frac{T}{K} + \sqrt{T}\log\big(1 + 2\sqrt{T}\big) + K + 2\sqrt{2KTL} + \frac{1}{T} = O\big(T^{2/3}(\log T)^{1/3}\big).$$

The grid size balances learning cost $\sqrt{KTL}$ against the price difference T/K. Appendix E.1 details the comparison and the role of consistency between the optimistic score and the dual update. We give the proof sketches below; all complete proofs are in Appendix B.

For $v, b \in [0,1]$ and $\lambda \geq 0$, define

$$\ell_G(v, b, \lambda) = G(b)\big((1+\lambda)v - \lambda b\big), \qquad h_G(v, \lambda) = \sup_{b \in [0,1]} \ell_G(v, b, \lambda).$$

For the analysis, we generate statistically equivalent feedback as follows. For each action k let $Z_{k,1}, \dots, Z_{k,T}$ be independent Bernoulli random variables with mean q_k, independent across k and

independent of V^T. When action k is played for the s-th time, the learner observes $x_t = Z_{k,s}$. In the original model the allocation in round t is, conditionally on the history, a Bernoulli variable with mean q_{k_t}, and the same is true in the coupled model. Hence the joint law of $(v_t, k_t, x_t)_{t\le T}$ is identical in the two models, and so are all quantities defined through it, including the regret and the violation. For $1 \le n \le T$ define the empirical mean $\hat{q}_{k,n} = \frac{1}{n}\sum_{s=1}^{n} Z_{k,s}$, and let

$$\mathcal{E} = \bigcap_{k=1}^{K}\bigcap_{n=1}^{T}\Big\{\big|\hat{q}_{k,n} - q_k\big| \le \sqrt{\tfrac{L}{2n}}\Big\}$$

be the event that all empirical means stay within their confidence intervals. Zero-value rounds use a separate independent Bernoulli stream of mean q_0, whose outcomes are ignored. The event $\mathcal{E}$ depends only on the positive-grid variables $Z_{k,s}$ and is therefore independent of V^T. Note also that k_t is determined by $(Z_{k,s})_{k,s}$ and $v_1, \dots, v_t$, and that λ_t is determined by $(Z_{k,s})_{k,s}$ and $v_1, \dots, v_{t-1}$.

Define the trajectory-dependent regret as $R_T = \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \sum_{t=1}^{T} v_t q_{k_t}$. Since q_{k_t} is the conditional winning probability, $\mathrm{Regret}(\mathrm{ALG} \mid G, F) = \mathbb{E}[R_T]$. We decompose R_T into four terms and control each separately.

Lemma 1. *On every sample path, $R_T = \mathrm{C}_T + \mathrm{O}_T + \mathrm{D}_T + \mathrm{S}_T$, where*

$$\mathrm{C}_T = \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \sum_{t=1}^{T} h_G(v_t, \lambda_t), \quad \mathrm{O}_T = \sum_{t=1}^{T}\big[h_G(v_t, \lambda_t) - v_t U_{t,k_t} - \lambda_t \widetilde{g}_t\big],$$

$$\mathrm{D}_T = \sum_{t=1}^{T} \lambda_t \widetilde{g}_t, \qquad \mathrm{S}_T = \sum_{t=1}^{T} v_t\big(U_{t,k_t} - q_{k_t}\big).$$

C_T compares the benchmark with the Lagrangian relaxation, O_T is the optimism residual, D_T is controlled by the dual update, and S_T is the estimation error of the winning probabilities.

Lemma 2. *On the event $\mathcal{E}$, $\mathrm{S}_T \le K + 2\sqrt{2KTL}$.*

Proof sketch. The first play of an action contributes at most one; subsequent confidence widths are bounded by $\sqrt{2L/n}$ after n observations. Summing within each action and applying Cauchy–Schwarz gives the bound. See Appendix B.3.

The next lemma is the step in which the i.i.d. structure of the values is used. Let

$$H(\lambda) = \int_0^1 h_G(v, \lambda)\, F(\mathrm{d}v)$$

denote the expected Lagrangian supremum. The next lemma bounds C_T in conditional expectation.

Lemma 3. $\mathbb{E}[\mathrm{C}_T \big| \mathcal{E}] \le 0$.

Proof sketch. For every fixed $\lambda \ge 0$, weak duality gives $\mathbb{E}[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)] \le TH(\lambda)$. Independence of the current value from the sample arrays and past values gives $\mathbb{E}[h_G(v_t, \lambda_t) \mid \mathcal{E}] = \mathbb{E}[H(\lambda_t) \mid \mathcal{E}]$. Combining these facts with independence of $\mathcal{E}$ and V^T proves the claim; no independence between λ_t and $\mathcal{E}$ is needed. See Appendix B.4.

Next we bound the term O_T.

Lemma 4. *On the event $\mathcal{E}$, for every round t, value $v \in [0,1]$, and multiplier $\lambda \ge 0$,*

$$\max_{1\le k\le K} U_{t,k}\big((1+\lambda)v - \lambda c_k\big) \ge h_G(v, \lambda). \tag{7}$$

In particular, $v_t U_{t,k_t} \ge h_G(v_t, \lambda_t) - \lambda_t \widetilde{g}_t$ for every t, and hence $\mathrm{O}_T \le 0$.

Proof sketch. The maximum grid score is nonnegative. For a bid b with nonnegative Lagrangian coefficient, choose a cell with $c_k \le b \le a_k$. On $\mathcal{E}$, $U_{t,k} \ge G(a_k) \ge G(b)$ and the virtual price is at most b, proving the comparison. Negative-coefficient bids cannot improve it. See Appendix B.5.

Then, we bound D_T.

Lemma 5. *For $0 < \eta \le 1/2$, the following hold on every sample path:*

$$\mathrm{D}_T = \sum_{t=1}^{T} \lambda_t \widetilde{g}_t \le \frac{2}{\eta} + 4\eta T, \tag{8}$$

$$-\sum_{t=1}^{T} \widetilde{g}_t \le \frac{\log(1+2\eta T)}{\eta}. \tag{9}$$

Proof sketch. The selected score is nonnegative, so $\lambda_t \widetilde{g}_t \ge -v_t U_{t,k_t} \ge -1$. Together with $|\widetilde{g}_t| \le 1$, this bounds the second-order term of the multiplicative update. Telescoping controls D_T and gives $\lambda_{t+1} \le 1 + 2\eta t$; taking logarithms yields the second bound. See Appendix B.6.

Proof sketch of Theorem 2 and Proposition 1. On $\mathcal{E}$, Lemma 4 gives

$$v_t q_{k_t} \ge h_G(v_t, \lambda_t) - \lambda_t \widetilde{g}_t - v_t(U_{t,k_t} - q_{k_t}). \tag{10}$$

The statistical error is not multiplied by λ_t, and the dual term uses exactly the surplus in the update. Summing confidence widths along selected actions gives

$$\sum_{t=1}^{T} v_t(U_{t,k_t} - q_{k_t}) \le W_T := K + 2\sqrt{2KTL}.$$

This uses each bid's observation count rather than a separate exploration phase. Lemmas 5 and 3 then control the dual and benchmark terms. For ROI, the executed price exceeds the virtual price by at most $1/K$, so on $\mathcal{E}$, so on $\mathcal{E}$,

$$\sum_{t=1}^{T}(a_{k_t} - v_t)q_{k_t} \le \frac{T}{K} - \sum_{t=1}^{T} \widetilde{g}_t + \sum_{t=1}^{T} v_t(U_{t,k_t} - q_{k_t}).$$

Combining the same width bound with (9), and accounting for failure of $\mathcal{E}$ by δT, yields, for $0 < \eta \le 1/2$ and $L \ge \log(2KT/\delta)$,

$$\mathrm{Regret}(\mathrm{ALG} \mid G, F) \le \frac{2}{\eta} + 4\eta T + W_T + \delta T, \tag{11}$$

$$\Delta(\mathrm{ALG} \mid G, F) \le \frac{T}{K} + \frac{\log(1+2\eta T)}{\eta} + W_T + \delta T. \tag{12}$$

Thus T/K is an execution cost in the ROI bound, not a reward-approximation error. Taking $\eta = T^{-1/2}$ and balancing T/K with $\sqrt{KTL}$ gives the stated rates. Appendices B.7 and B.8 give the complete proofs.

6 Terminal ROI Feasibility in Expectation

We further introduce a positive surplus target into the dual update to compensate for the violations in Proposition 1. Under the following condition, the resulting algorithm achieves terminal ROI feasibility in expectation for all sufficiently large horizons while retaining the same regret rate against the original benchmark.

Assumption 1. *There is a constant $\kappa_{\min} \in (0, 1]$, independent of T, such that*

$$\kappa(G, F) := \mathbb{E}_{v\sim F, d\sim G}[(v - d)_+] \ge \kappa_{\min}. \tag{13}$$

For analysis, define the maximum attainable expected per-round surplus

$$s(v) = \sup_{0 \le b \le v} (v - b)G(b), \qquad \sigma(G, F) = \mathbb{E}_{v\sim F}[s(v)]. \tag{14}$$

Appendix C.1 proves

$$\frac{\kappa}{2(1 + \log(1/\kappa))} \le \sigma \le \kappa \qquad (\kappa > 0). \tag{15}$$

Thus Assumption 1 supplies a fixed positive σ without an additional assumption; the sharper implicit bound is $\kappa \le \sigma(1 + \log(1/\sigma))$.

The modified algorithm. Use the parameters of Theorem 2. With $W_T = K + 2\sqrt{2KTL}$, define the surplus target ε_T by

$$A_T = \frac{T}{K} + W_T + \frac{\log(1+2\eta T)}{\eta}, \qquad \varepsilon_T = \frac{A_T}{T} + 2\delta. \tag{16}$$

The three terms in A_T account for the price difference, cumulative estimation error, and dual contribution in the ROI bound. The additional 2δ accounts for confidence failure. Thus $\varepsilon_T = O(T^{-1/3}(\log T)^{1/3})$. Retain the confidence estimates, bid selection rule, and observation updates of Algorithm 1, and replace the dual update by

$$\lambda_{t+1} = \lambda_t \exp\big[-\eta(\widetilde{g}_t - \varepsilon_T)\big], \qquad \lambda_1 = 1. \tag{17}$$

The multiplier now increases when $\widetilde{g}_t < \varepsilon_T$ and decreases when $\widetilde{g}_t > \varepsilon_T$. Thus even a positive surrogate surplus can increase the emphasis on ROI if it falls below the target. Equivalently, the score is $v_t U_{t,k} + \lambda_t((v_t - c_k)U_{t,k} - \varepsilon_T)$. The term $-\lambda_t \varepsilon_T$ is common to all actions, so the maximization rule is unchanged. Denote the policy by $\mathrm{ALG}^{\mathrm{tar}}$.

Algorithm 2 UCB-DMD with a Surplus Target

Require: Parameters of Theorem 2, surplus target ε_T from (16)
1: Set $N_{1,k} = W_{1,k} = 0$ for all k and $\lambda_1 = 1$
2: **for** $t = 1, \dots, T$ **do**
3: Compute $U_{t,k}$ by (4); observe v_t and select k_t by (5)
4: Submit $b_t = a_{k_t}$ and observe x_t
5: Set $\widetilde{g}_t = (v_t - c_{k_t})U_{t,k_t}$ and update λ_{t+1} by (17)
6: If $k_t > 0$, set $N_{t+1,k_t} \leftarrow N_{t,k_t} + 1$, $W_{t+1,k_t} \leftarrow W_{t,k_t} + x_t$; others unchanged
7: **end for**

Theorem 3. *Let $T \geq 16$ and write $\sigma = \sigma(G, F)$. If $2\varepsilon_T \leq \sigma$, then Algorithm 2 satisfies*

$$\Delta(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) \leq -\frac{1}{T} + \frac{2}{T^3} < 0, \tag{18}$$

$$\mathrm{Regret}(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) \leq 6\sqrt{T} + W_T + \frac{16\varepsilon_T T}{\sigma} + \frac{1}{T}. \tag{19}$$

Consequently, under Assumption 1, there exist constants C and T_{ROI}, depending only on $\kappa_{\min}$, such that for every instance in this class and every $T \geq T_{\mathrm{ROI}}$,

$$\Delta(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) \leq 0, \qquad \mathrm{Regret}(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) \leq CT^{2/3}(\log T)^{1/3}. \tag{20}$$

Proof sketch. On $\mathcal{E}$, the cumulative difference between the surrogate surplus and the expected surplus of the submitted bids is at most $T/K + W_T$. Iterating the dual update gives

$$-\sum_{t=1}^{T} \widetilde{g}_t = \frac{\log \lambda_{T+1}}{\eta} - \varepsilon_T T.$$

To control the terminal multiplier, optimism implies $\widetilde{g}_t \geq s(v_t) - 1/\lambda_t$. Since v_t is independent of the past, the available surplus σ limits the growth of the expected multiplier. Specifically, when $2\varepsilon_T \leq \sigma$ and $\eta \leq 1/4$, $m_t := \mathbb{E}[\lambda_t \mid \mathcal{E}]$ satisfies

$$m_{t+1} \leq (1 - \eta\sigma/4)m_t + 2\eta, \qquad m_t \leq \min\{1 + 2\eta(t-1), 8/\sigma\}.$$

Jensen's inequality bounds the expected logarithmic term by $\eta^{-1}\log(1+2\eta T)$. The definition of ε_T then compensates for this term and the surplus error, leaving $2\delta T$ to account for confidence failure. The bound $m_t \leq 8/\sigma$ also limits the additional regret contribution to $2\varepsilon_T \sum_t m_t \leq 16\varepsilon_T T/\sigma$. The benchmark comparison remains unchanged. Appendix C.2 gives the full proof.

Since ε_T converges to 0 as T increases and $\sigma \geq \kappa_{\min}/[2(1 + \log(1/\kappa_{\min}))]$, the condition holds for all sufficiently large T. Neither σ nor $\kappa_{\min}$ is an algorithm input. The surplus target guides the dual update; the guarantee concerns nonnegative expected total surplus, averaged over values and feedback.

Remark 2. *Assumption 1 concerns surplus averaged over values and competing bids. For example, $\Pr(v - d \geq m_0) \geq p_0$ for fixed $p_0, m_0 > 0$ implies $\kappa \geq p_0 m_0$. Related positive-surplus conditions also appear in Feng et al. (2023, Section 4.1) and Castiglioni et al. (2024, Section 3). The assumption permits individual rounds with no surplus and an optimal policy whose ROI constraint binds. It is used only for terminal ROI feasibility in expectation; Theorem 2 and Proposition 1 require no such condition.*

7 LOWER BOUNDS

We first establish a joint lower bound on regret and ROI violation, and then derive regret lower bounds under a uniform bound on ROI violation.

Theorem 4. *There exist a positive constant T_0, such that for every integer $T \geq T_0$ and every online algorithm with bandit feedback, possibly randomized, there exists an instance (G, F) satisfying*

$$\mathrm{Regret}(\mathrm{ALG} \mid G, F) + \frac{1}{12} \max\{\Delta(\mathrm{ALG} \mid G, F), 0\} \geq \Omega(T^{2/3}). \tag{21}$$

Proof sketch. The construction uses a fixed value and competing-bid distributions that differ only on one short interval. We use $m = \Theta(T^{1/3})$ disjoint intervals and perturb the competing-bid distribution on one interval by order $\varepsilon = \Theta(T^{-1/3})$. The corresponding feasible benchmark gains $\Theta(\varepsilon)$ per round. Under the base distribution, some interval receives at most T/m bids in expectation. Since each observation there contributes only $O(\varepsilon^2)$ relative entropy, a change-of-measure argument shows that the learner also samples this interval infrequently under the perturbed distribution. A payment–winning-probability inequality converts this sampling bound into a joint lower bound on regret and expected ROI violation of order $\varepsilon T = \Theta(T^{2/3})$. Accounting for violation is essential: an infeasible learner may obtain high reward by overbidding, so regret alone does not capture the difficulty of satisfying ROI. The joint bound incorporates the expected ROI violation associated with such reward gains. See Appendix D.1. The joint lower bound yields the following consequence when the violation of the original ROI constraint is $o(T^{2/3})$.

Proposition 2. *Let $(\mathrm{ALG}_T)_{T\geq 1}$ be a family of online algorithms with bandit feedback, where ALG_T runs for T rounds. Suppose that there is a nonnegative sequence $r_T = o(T^{2/3})$ such that, for every T and every pair (G, F),*

$$\max\{\Delta(\mathrm{ALG}_T \mid G, F), 0\} \leq r_T. \tag{22}$$

Then, for every $T \geq T_0$,

$$\sup_{G,F} \mathrm{Regret}(\mathrm{ALG}_T \mid G, F) \geq \Omega(T^{2/3}) - \frac{r_T}{12}. \tag{23}$$

Since $r_T = o(T^{2/3})$, there is a threshold $T_1 \geq T_0$, possibly depending on (r_T), for which the right-hand side is at least $\Omega(T^{2/3})$ for every $T \geq T_1$.

Setting $r_T = 0$ gives a regret lower bound of $\Omega(T^{2/3})$ for algorithms that satisfy the original ROI constraint in expectation on every instance. Conversely, Theorem 4 implies that any algorithm with $o(T^{2/3})$ regret must have $\Omega(T^{2/3})$ violation on some instance. Since UCB-DMD achieves $O(T^{2/3}(\log T)^{1/3})$ for both regret and ROI violation, their common rate is optimal up to the logarithmic factor. By Theorem 1, the lower bound (23) also holds against the deterministic benchmark (2) with an additional subtraction of 1 in the regret, so its $\Omega(T^{2/3})$ consequence is unchanged.

Remark 3. *The hard instances in Appendix D.1 have $v_t = 4/13$ and $\Pr(d_t = 0) = 3/4$, while all positive competing bids are at least $1/3$. Hence, $\kappa(G, F) = \sigma(G, F) = \frac{3}{13}$. Bidding zero generates positive surplus but earns only $3T/13$, whereas the feasible benchmark achieves at least $T/4$. Sublinear regret therefore requires learning which higher bids use this surplus effectively, a task made difficult by binary feedback.*

Thus, for every fixed $0 < \kappa_{\min} \leq 3/13$, the joint lower bound holds within the corresponding instance class of Assumption 1. Any algorithm satisfying the expected ROI constraint throughout this class must incur $\Omega(T^{2/3})$ worst-case regret. Together with Theorem 3, this establishes regret optimality up to logarithmic factors for sufficiently large horizons, even when a policy generating positive surplus is available.

8 CONCLUSION AND SCOPE

UCB-DMD closes the bandit-feedback rate gap for regret and ROI violation up to logarithms against the optimal randomized offline benchmark. Under a fixed positive surplus assumption, the target modification also gives terminal ROI feasibility in expectation. The lower bound establishes joint rate optimality and regret optimality for feasible algorithms on positive surplus classes containing the hard instances.

AI USE STATEMENT

Generative AI tools assisted with editing, checking notation and proof consistency. The authors are responsible for verifying all claims, proofs.

REFERENCES

Gagan Aggarwal, Ashwinkumar Badanidiyuru, and Aranyak Mehta. Autobidding with constraints. In *Web and Internet Economics: 15th International Conference, WINE 2019*, pp. 17–30. Springer, 2019.

Gagan Aggarwal, Ashwinkumar Badanidiyuru, Santiago R Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Gagan Goel, Christopher Liaw, Haihao Lu, Mohammad Mahdian, et al. Auto-bidding and auctions in online advertising: A survey. *ACM SIGecom Exchanges*, 22(1):159–183, 2024.

Gagan Aggarwal, Giannis Fikioris, and Mingfei Zhao. No-regret algorithms in non-truthful auctions with budget and ROI constraints. In *Proceedings of the ACM on Web Conference 2025*, pp. 1398–1415, 2025.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the 15th ACM Conference on Economics and Computation*, pp. 989–1006, 2014.

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

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3):1–55, 2018.

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

Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pp. 613–628. PMLR, 2020.

Santiago Balseiro, Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Optimal mechanisms for a value maximizer: The futility of screening targets. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 1101–1101, 2024.

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

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

Santiago R Balseiro, Yuan Deng, Jieming Mao, Vahab S Mirrokni, and Song Zuo. The landscape of auto-bidding auctions: Value versus utility maximization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 132–133, 2021.

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023.

Xiaohui Bei, Pinyan Lu, Zhiqi Wang, Tao Xiao, and Xiang Yan. Optimal auction design for mixed bidders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13614–13621, 2025.

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

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning under budget and ROI constraints via weak adaptivity. In *International Conference on Machine Learning*, pp. 5792–5816. PMLR, 2024.

Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Towards efficient auctions in an autobidding world. In *Proceedings of the Web Conference 2021*, pp. 3965–3973, 2021.

Yuan Deng, Vahab Mirrokni, and Hanrui Zhang. Posted pricing and dynamic prior-independent mechanisms with value maximizers. *Advances in Neural Information Processing Systems*, 35: 24158–24169, 2022.

Yuan Deng, Mohammad Mahdian, Jieming Mao, Vahab Mirrokni, Hanrui Zhang, and Song Zuo. Efficiency of the generalized second-price auction for value maximizers. In *Proceedings of the ACM Web Conference 2024*, pp. 46–56, 2024a.

Yuan Deng, Jieming Mao, Vahab Mirrokni, Hanrui Zhang, and Song Zuo. Efficiency of the first-price auction in the autobidding world. *Advances in Neural Information Processing Systems*, 37: 139270–139293, 2024b.

Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM*, 66(1):1–41, 2019.

Yiding Feng, Brendan Lucier, and Aleksandrs Slivkins. Strategic budget selection in a competitive autobidding world. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 213–224, 2024.

Zhe Feng, Guru Guruganesh, Christopher Liaw, Aranyak Mehta, and Abhishek Sethi. Convergence analysis of no-regret bidding algorithms in repeated auctions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5399–5406, 2021.

Zhe Feng, Swati Padmanabhan, and Di Wang. Online bidding algorithms for return-on-spend constrained advertisers. In *Proceedings of the ACM Web Conference 2023*, pp. 3550–3560, 2023.

Giannis Fikioris and Éva Tardos. Approximately stationary bandits with knapsacks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3758–3782. PMLR, 2023.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pp. 2286–2305. PMLR, 2020.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Rachitesh Kumar, Jon Schneider, and Balasubramanian Sivan. Strategically-robust learning algorithms for bidding in first-price auctions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 893–893, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Yilin Li, Yuan Deng, Wei Tang, and Hanrui Zhang. No-regret online autobidding algorithms in first-price auctions. *Advances in Neural Information Processing Systems*, 38:62201–62219, 2026.

Christopher Liaw, Aranyak Mehta, and Andres Perlroth. Efficiency of non-truthful auctions in autobidding: The power of randomization. In *Proceedings of the ACM Web Conference 2023*, pp. 3561–3571, 2023.

Pinyan Lu, Chenyang Xu, and Ruilong Zhang. Auction design for value maximizers with budget and return-on-spend constraints. In *International Conference on Web and Internet Economics*, pp. 474–491. Springer, 2023.

Hongtao Lv, Xiaohui Bei, Zhenzhe Zheng, and Fan Wu. Auction design for bidders with ex post ROI constraints. In *International Conference on Web and Internet Economics*, pp. 492–508. Springer, 2023.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. AdWords and generalized online matching. *Journal of the ACM*, 54(5):22, 2007.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular Lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023.

Fransisca Susan, Negin Golrezaei, and Okke Schrijvers. Multi-platform budget management in ad markets with non-IC auctions. *arXiv preprint arXiv:2306.07352*, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Sushant Vijayan, Zhe Feng, Swati Padmanabhan, Karthikeyan Shanmugam, Arun Suggala, and Di Wang. Online bidding under RoS constraints without knowing the value. In *Proceedings of the ACM on Web Conference 2025*, pp. 3096–3107, 2025.

Qian Wang, Zongjun Yang, Xiaotie Deng, and Yuqing Kong. Learning to bid in repeated first-price auctions with budgets. In *International Conference on Machine Learning*, pp. 36494–36513. PMLR, 2023.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pp. 1562–1583. PMLR, 2016.

APPENDIX

A OMITTED PROOFS OF SECTION 4

A.1 PROOF OF THEOREM 1

Proof of Theorem 1. Step 1: reduction to per-round bid distributions. Let π be any policy feasible for (1) and let π_t be the marginal distribution of its bid b_t. Because d_t is independent of b_t, the expected reward and payment of π are $\sum_t \int v_t G(b)\,\pi_t(\mathrm{d}b)$ and $\sum_t \int bG(b)\,\pi_t(\mathrm{d}b)$, which depend on π only through $(\pi_1, \dots, \pi_T)$. The policy that draws $b_t \sim \pi_t$ independently in each round, without looking at past outcomes, has the same marginals and hence the same reward and payment. Thus the benchmark is unchanged if we restrict to policies that fix a bid distribution π_t for each round in advance.

Step 2: the set of achievable pairs. For a deterministic bid b, the pair $(bG(b), G(b))$ records its expected payment and winning probability. Let

$$\mathcal{A} = \{(bG(b), G(b)) : b \in [0,1]\} \subseteq [0,1]^2,$$

let $\overline{\mathcal{A}}$ be its closure, and let $\mathcal{C} = \mathrm{conv}(\overline{\mathcal{A}})$ be the convex hull of the closure, that is, the set of all finite weighted averages of points of $\overline{\mathcal{A}}$ with nonnegative weights summing to one. The set $\overline{\mathcal{A}}$ is closed and bounded, hence compact. By Carathéodory's theorem in the plane, every point of $\mathcal{C}$ is a weighted average of at most three points of $\overline{\mathcal{A}}$, so $\mathcal{C}$ is the image of the compact set $\overline{\mathcal{A}}^3 \times \{(w_1, w_2, w_3) \geq 0 : \sum_i w_i = 1\}$ under a continuous map and is therefore compact. For any bid distribution π_t, the pair $(\int bG(b)\,\pi_t(\mathrm{d}b), \int G(b)\,\pi_t(\mathrm{d}b))$ is the expectation of a random point of $\mathcal{A} \subseteq \mathcal{C}$; since $\mathcal{C}$ is compact and convex, this expectation lies in $\mathcal{C}$.

Step 3: an auxiliary problem. For each round t write $(p_t, q_t) \in \mathcal{C}$ for a candidate pair of expected payment and winning probability, and consider

$$\max_{(p_t,q_t)\in\mathcal{C},\ t=1,\dots,T} \Big\{ \sum_{t=1}^{T} v_t q_t \ :\ \sum_{t=1}^{T} (v_t q_t - p_t) \geq 0 \Big\}. \tag{24}$$

Here $v_t q_t$ is the expected reward and $v_t q_t - p_t$ the expected ROI surplus of round t. By Steps 1 and 2, the pairs induced by any feasible policy are feasible for (24) with the same objective, so the optimal value of (24) is at least $\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$. The feasible set of (24) is nonempty (the pair $(0, G(0))$ of bidding zero is feasible in every round), closed, and bounded, and the objective is continuous, so the maximum is attained at some $(p_t^*, q_t^*)_{t\leq T}$.

Step 4: realizing points of $\mathcal{C}$ by actual bids. We show that every point of $\overline{\mathcal{A}}$ can be realized by a mixture of at most two bids with the same expected payment and at least the same winning probability. Points of $\mathcal{A}$ are realized by their own bid. Every point of $\overline{\mathcal{A}} \setminus \mathcal{A}$ is a limit of points $(b_n G(b_n), G(b_n))$ with $b_n \to c$ for some $c \in [0,1]$; since G is nondecreasing and right-continuous, the only possible limits other than $(cG(c), G(c))$ arise from $b_n \uparrow c$ with $c \in (0,1]$ and equal $(cG(c_-), G(c_-))$. If $G(c) = 0$ this point is $(0,0)$, which is realized by bidding zero, since then $G(0) = 0$. Otherwise, bid c with probability $G(c_-)/G(c)$ and zero with the remaining probability. The expected payment is $c\,G(c) \cdot G(c_-)/G(c) = cG(c_-)$, and the winning probability is $G(c_-) + (1 - G(c_-)/G(c))G(0) \geq G(c_-)$. By Carathéodory's theorem, each (p_t^*, q_t^*) is a weighted average of at most three points of $\overline{\mathcal{A}}$, and each of these is realized as above; mixing the realizations with the same weights gives a bid distribution supported on at most six bids with expected payment p_t^* and winning probability at least q_t^*. Increasing the winning probability at a fixed payment does not decrease $v_t q_t$ or $v_t q_t - p_t$ because $v_t \geq 0$. The resulting policy is therefore feasible for (1) and its reward is at least the optimal value of (24), which is at least $\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$. Hence the supremum in (1) is attained by a policy using at most six bids per round.

Step 5: reducing to one randomized round. For each round t let $\mathcal{B}_t$ be the set of bids to which the policy of Step 4 assigns positive probability, so $|\mathcal{B}_t| \leq 6$, and let $w_{t,b}$ be the probability of bid $b \in \mathcal{B}_t$. With the bid sets fixed, consider the linear program

$$\max_{w_{t,b}\geq 0} \sum_{t=1}^{T} \sum_{b\in\mathcal{B}_t} w_{t,b}\, v_t G(b) \quad \text{s.t.} \quad \sum_{b\in\mathcal{B}_t} w_{t,b} = 1 \ \ (t = 1, \dots, T), \quad \sum_{t=1}^{T} \sum_{b\in\mathcal{B}_t} w_{t,b}(v_t - b)G(b) \geq 0. \tag{25}$$

The weights of the policy from Step 4 are feasible, and every feasible solution defines a feasible policy for (1) with objective equal to its reward, so (25) has optimal value $\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$. Adding a nonnegative slack variable turns the ROI inequality into an equality; the resulting standard-form program has $T+1$ equality constraints, is feasible, and is bounded (its objective is at most T), so it has an optimal basic feasible solution. A basic feasible solution has at most as many positive variables as there are linearly independent equality constraints, hence at most $T+1$ positive variables including the slack. Since each of the T probability constraints forces at least one positive $w_{t,b}$ in its round, at most one round can have a second positive weight. Thus the optimal solution uses one bid in every round except at most one, in which it mixes two bids.

Step 6: comparison with deterministic plans. If the policy of Step 5 is deterministic, then $\mathrm{Reward}(\mathrm{det} \mid G, V^T) \geq \mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$. Otherwise let s be the unique randomized round and let b', b'' be its two bids. Replace the mixture in round s by whichever of b', b'' has the larger ROI surplus $(v_s - b)G(b)$. The larger surplus is at least the weighted average of the two, so the total surplus does not decrease and the resulting deterministic plan is feasible for (2). Only round s changes, and its expected reward before and after lies in $[0, v_s]$, so the total reward decreases by at most v_s. Hence $\mathrm{Reward}(\mathrm{det} \mid G, V^T) \geq \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - v_s \geq \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \max_t v_t$. Conversely, every deterministic plan is a feasible policy for (1), so $\mathrm{Reward}(\mathrm{det} \mid G, V^T) \leq \mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$. Together these prove (3). □

B Omitted Proofs from Section 5.2

B.1 Confidence event and Estimation Bounds

We verify the properties of the confidence event defined in Section 5.2.

Lemma 6. *If $L \geq \log(2KT/\delta)$, then $\Pr(\mathcal{E}) \geq 1 - \delta$. On $\mathcal{E}$, for all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$,*

$$0 \leq U_{t,k} - q_k \leq \begin{cases} 1, & N_{t,k} = 0, \\ \min\{1, \sqrt{2L/N_{t,k}}\}, & N_{t,k} > 0. \end{cases} \tag{26}$$

Proof. Probability of $\mathcal{E}$. Fix k and n. The variables $Z_{k,1}, \ldots, Z_{k,n}$ are independent, take values in $[0,1]$, and have mean q_k, so Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr\Big(|\hat{q}_{k,n} - q_k| > \sqrt{\tfrac{L}{2n}}\Big) \leq 2\exp\Big(-2n \cdot \tfrac{L}{2n}\Big) = 2e^{-L}.$$

A union bound over the KT pairs (k, n) gives $\Pr(\mathcal{E}^c) \leq 2KTe^{-L} \leq \delta$, where the last inequality is $L \geq \log(2KT/\delta)$.

Bounds on $\mathcal{E}$. Fix a sample path in $\mathcal{E}$, a round t, and an action k. If $N_{t,k} = 0$ then $U_{t,k} = 1$ and $0 \leq U_{t,k} - q_k = 1 - q_k \leq 1$. If $N_{t,k} = n \geq 1$, then in the coupled model the learner has observed exactly $Z_{k,1}, \ldots, Z_{k,n}$ from action k, so its empirical mean is $W_{t,k}/N_{t,k} = \hat{q}_{k,n}$. Let $r_n = \sqrt{L/(2n)}$. On $\mathcal{E}$ we have $q_k - r_n \leq \hat{q}_{k,n} \leq q_k + r_n$, hence $q_k \leq \hat{q}_{k,n} + r_n \leq q_k + 2r_n$. Since $q_k \leq 1$, the definition (4) gives

$$q_k \leq \min\{1, \hat{q}_{k,n} + r_n\} = U_{t,k} \leq \min\{1, q_k + 2r_n\}.$$

Subtracting q_k yields $0 \leq U_{t,k} - q_k \leq \min\{1 - q_k, 2r_n\} \leq \min\{1, \sqrt{2L/n}\}$, which is (26). □

B.2 Proof of Lemma 1

Proof. Adding and subtracting $\sum_t h_G(v_t, \lambda_t)$ gives

$$R_T = \mathrm{C}_T + \sum_t \big(h_G(v_t, \lambda_t) - v_t q_{k_t}\big).$$

For each t, adding and subtracting $v_t U_{t,k_t}$ and $\lambda_t \widetilde{g}_t$ yields

$$h_G(v_t, \lambda_t) - v_t q_{k_t} = \big(h_G(v_t, \lambda_t) - v_t U_{t,k_t} - \lambda_t \widetilde{g}_t\big) + \lambda_t \widetilde{g}_t + v_t(U_{t,k_t} - q_{k_t}).$$

Summing over t gives the claim. □

B.3 Proof of Lemma 2

Proof. Fix a sample path in $\mathcal{E}$. For positive-grid actions, Lemma 6 gives $U_{t,k_t} - q_{k_t} \geq 0$; zero-value rounds contribute zero. Since $v_t \leq 1$,

$$\mathrm{S}_T \leq \sum_{t:k_t>0} (U_{t,k_t} - q_{k_t}).$$

Zero-value rounds contribute zero to S_T. For each positive-grid action k let $N_k = N_{T+1,k}$ be its total number of plays. An action with $N_k = 0$ contributes nothing. If $N_k \geq 1$, then at its first play $N_{t,k} = 0$ and the width is at most 1, and at its i-th play, $2 \leq i \leq N_k$, we have $N_{t,k} = i - 1$ and the width is at most $\sqrt{2L/(i-1)}$ by (26). Therefore

$$\sum_{t:\, k_t=k} (U_{t,k} - q_k) \leq 1 + \sqrt{2L} \sum_{n=1}^{N_k-1} \frac{1}{\sqrt{n}} \leq 1 + 2\sqrt{2LN_k},$$

where we used

$$\sum_{n=1}^{M} n^{-1/2} \leq \int_0^M x^{-1/2}\,\mathrm{d}x = 2\sqrt{M}$$

for every integer $M \geq 0$ (an empty sum is zero). Summing over k and applying the Cauchy–Schwarz inequality,

$$\mathrm{S}_T \leq K + 2\sqrt{2L} \sum_{k=1}^{K} \sqrt{N_k} \leq K + 2\sqrt{2L}\sqrt{K \sum_{k=1}^{K} N_k} \leq K + 2\sqrt{2KTL},$$

where the last step uses $\sum_{k=1}^{K} N_k \leq T$. □

B.4 Proof of Lemma 3

Proof. For a value $v \in [0,1]$, a bid $b \in [0,1]$, and a multiplier $\lambda \geq 0$, define the per-round Lagrangian and its supremum over bids by

$$\ell_G(v, b, \lambda) = G(b)\big((1+\lambda)v - \lambda b\big), \qquad h_G(v, \lambda) = \sup_{b\in[0,1]} \ell_G(v, b, \lambda).$$

Since $\ell_G(v, 0, \lambda) = G(0)(1+\lambda)v \geq 0$ and $\ell_G \leq 1 + \lambda$, we have $0 \leq h_G(v, \lambda) \leq 1 + \lambda$. For fixed λ, $h_G(\cdot, \lambda)$ is a supremum of affine functions of v, hence convex and measurable. For a policy π, let π_t denote the marginal distribution of its bid in round t. Then

$$\mathrm{Reward}(\pi \mid G, V^T) = \sum_t \int_0^1 v_t G(b)\, \pi_t(\mathrm{d}b) \qquad \mathrm{Payment}(\pi \mid G, V^T) = \sum_t \int_0^1 bG(b)\, \pi_t(\mathrm{d}b).$$

Thus, for $\lambda \geq 0$,

$$\mathrm{Reward}(\pi \mid G, V^T) + \lambda(\mathrm{Reward}(\pi \mid G, V^T) - \mathrm{Payment}(\pi \mid G, V^T)) = \sum_{t=1}^{T} \int \ell_G(v_t, b, \lambda)\pi_t(\mathrm{d}b), \tag{27}$$

where all integrals over b are over $[0,1]$.

Step 1: Fix V^T and $\lambda \geq 0$, and let π be any policy feasible for (1), so that

$$\mathrm{Reward}(\pi \mid G, V^T) - \mathrm{Payment}(\pi \mid G, V^T) \geq 0.$$

Multiplying this nonnegative quantity by $\lambda \geq 0$ and using (27),

$$\mathrm{Reward}(\pi \mid G, V^T) \leq \sum_{t=1}^{T} \int_0^1 \ell_G(v_t, b, \lambda)\, \pi_t(\mathrm{d}b) \leq \sum_{t=1}^{T} h_G(v_t, \lambda),$$

where the last step bounds the integrand by its supremum over b. Taking the supremum over feasible π gives

$$\mathrm{Reward}(\mathrm{OPT} \mid G, V^T) \le \sum_{t=1}^{T} h_G(v_t, \lambda).$$

for every V^T and every $\lambda \ge 0$. Taking expectations over $V^T \sim F^T$ and using that each v_t has distribution F,

$$\mathbb{E}\big[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)\big] \le T\, H(\lambda) \qquad \text{for every } \lambda \ge 0. \tag{28}$$

Step 2: Fix t. The multiplier λ_t is a function of $(Z_{k,s})_{k,s}$ and $v_1, \dots, v_{t-1}$, while v_t is independent of these variables and has distribution F. Hence,

$$\mathbb{E}[h_G(v_t, \lambda_t) \mid (Z_{k,s})_{k,s}, v_1, \dots, v_{t-1}] = H(\lambda_t).$$

Since $\mathbf{1}\{\mathcal{E}\}$ is a function of $(Z_{k,s})_{k,s}$ alone, the tower property gives

$$\mathbb{E}[h_G(v_t, \lambda_t)\mathbf{1}\{\mathcal{E}\}] = \mathbb{E}[H(\lambda_t)\mathbf{1}\{\mathcal{E}\}].$$

All quantities are integrable because $0 \le h_G(v_t, \lambda_t) \le 1 + \lambda_t \le 1 + e^{\eta(t-1)}$, where the last bound uses $|\widetilde{g}_s| \le 1$ in (6).

Step 3: Applying (28) with the realized value of λ_t shows that

$$H(\lambda_t) \ge \frac{1}{T}\mathbb{E}[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)]$$

on every sample path. Therefore

$$\mathbb{E}\Big[\sum_{t=1}^{T} h_G(v_t, \lambda_t)\mathbf{1}\{\mathcal{E}\}\Big] = \mathbb{E}\Big[\sum_{t=1}^{T} H(\lambda_t)\mathbf{1}\{\mathcal{E}\}\Big] \ge \Pr(\mathcal{E})\,\mathbb{E}\big[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)\big].$$

On the other hand, $\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)$ is a function of V^T, which is independent of $\mathcal{E}$, so $\mathbb{E}[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)\mathbf{1}\{\mathcal{E}\}] = \Pr(\mathcal{E})\,\mathbb{E}[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)]$. Subtracting the two displays gives $\mathbb{E}[\mathrm{C}_T\mathbf{1}\{\mathcal{E}\}] \le 0$. □

B.5 Proof of Lemma 4

Proof. Fix t, v, and λ. Since $c_1 = 0$, action 1 has score $U_{t,1}(1+\lambda)v \ge 0$, so the left-hand side of (7) is nonnegative. It therefore suffices to show that it is at least $\ell_G(v, b, \lambda)$ for every bid b with $(1+\lambda)v - \lambda b \ge 0$, because for all other bids $\ell_G(v, b, \lambda) \le 0$. For such a b, choose $k \in \{1, \dots, K\}$ with $c_k \le b \le a_k$. On $\mathcal{E}$, Lemma 6 and the monotonicity of G give $U_{t,k} \ge G(a_k) \ge G(b)$. Using $c_k \le b$, $\lambda \ge 0$, and then $(1+\lambda)v - \lambda b \ge 0$,

$$U_{t,k}\big((1+\lambda)v - \lambda c_k\big) \ge U_{t,k}\big((1+\lambda)v - \lambda b\big) \ge G(b)\big((1+\lambda)v - \lambda b\big) = \ell_G(v, b, \lambda).$$

Taking the supremum over b proves (7). For the second claim, the definition of $\widetilde{g}_t$ gives

$$U_{t,k_t}\big((1+\lambda_t)v_t - \lambda_t c_{k_t}\big) = v_t U_{t,k_t} + \lambda_t \widetilde{g}_t,$$

where k_t maximizes the score for $v_t > 0$ (and both sides are zero when $v_t = 0$), so (7) with $v = v_t$ and $\lambda = \lambda_t$ yields $v_t U_{t,k_t} + \lambda_t \widetilde{g}_t \ge h_G(v_t, \lambda_t)$. Every summand of O_T is thus nonpositive. □

B.6 Proof of Lemma 5

Proof of Lemma 5. Two elementary facts. Since $v_t, c_{k_t}, U_{t,k_t} \in [0, 1]$, the definition (6) gives $|\widetilde{g}_t| \le 1$. Moreover, since $c_1 = 0$, action 1 has score $U_{t,1}(1+\lambda_t)v_t \ge 0$, and the selected score is at least zero (also when $v_t = 0$), so $v_t U_{t,k_t} + \lambda_t \widetilde{g}_t \ge 0$, that is,

$$\lambda_t \widetilde{g}_t \ge -v_t U_{t,k_t} \ge -1. \tag{29}$$

From these two facts, $\lambda_t \widetilde{g}_t^2 \le \lambda_t |\widetilde{g}_t| \le \lambda_t \widetilde{g}_t + 2$: if $\widetilde{g}_t \ge 0$ the middle and right terms differ by 2, and if $\widetilde{g}_t < 0$ then $\lambda_t |\widetilde{g}_t| \le 1 \le \lambda_t \widetilde{g}_t + 2$ by (29).

One-step inequality. We use $e^{-x} \le 1 - x + x^2$ for $|x| \le 1/2$. For $x \ge 0$ this follows from $e^{-x} \le 1 - x + x^2/2$, and for $x = -y$ with $y \in (0, 1/2]$ from $e^y = 1 + y + y^2 \sum_{n\ge2} y^{n-2}/n! \le 1 + y + y^2 \sum_{n\ge2}(1/2)^{n-2}/2 = 1 + y + y^2$. Applying it with $x = \eta\widetilde{g}_t$, for which $|x| \le \eta \le 1/2$,

$$\begin{aligned}\lambda_{t+1} = \lambda_t e^{-\eta\widetilde{g}_t} &\le \lambda_t\big(1 - \eta\widetilde{g}_t + \eta^2\widetilde{g}_t^2\big)\\ &\le \lambda_t - \eta\lambda_t\widetilde{g}_t + \eta^2(\lambda_t\widetilde{g}_t + 2) = \lambda_t - \eta(1-\eta)\lambda_t\widetilde{g}_t + 2\eta^2.\end{aligned}$$

Proof of (8). Summing the one-step inequality over $t = 1, \dots, T$ and using $\lambda_1 = 1$ and $\lambda_{T+1} > 0$,

$$\eta(1-\eta)\mathrm{D}_T \le \lambda_1 - \lambda_{T+1} + 2\eta^2 T \le 1 + 2\eta^2 T.$$

Since $1 - \eta \ge 1/2$, dividing by $\eta(1-\eta)$ gives $\mathrm{D}_T \le 2/\eta + 4\eta T$.

Growth of λ_t. The one-step inequality and (29) give $\lambda_{t+1} - \lambda_t \le \eta(1-\eta) + 2\eta^2 = \eta(1+\eta) \le 2\eta$, so $\lambda_t \le 1 + 2\eta(t-1)$ for $t = 1, \dots, T+1$.

Proof of (9). Iterating the multiplicative update gives $\lambda_{T+1} = \exp(-\eta\sum_{t=1}^T \widetilde{g}_t)$, so $-\sum_t \widetilde{g}_t = \log(\lambda_{T+1})/\eta \le \log(1 + 2\eta T)/\eta$ by the growth bound. □

B.7 Proof of Theorem 2

Proof. Since $T \ge 4$, $\eta \le 1/2$ as required by Lemma 5. Also $L \ge \log 512 > 1$, so $K \le \lceil T^{1/3} \rceil \le T$ and hence $2KT/\delta \le 2T^4$, which gives $L \ge \log(2KT/\delta)$; Lemma 6 then yields $\Pr(\mathcal{E}) \ge 1 - \delta$. By Lemma 1, we have

$$\mathrm{Regret}(\mathrm{ALG} \mid G, F) = \mathbb{E}[R_T] = \mathbb{E}[R_T\mathbf{1}\{\mathcal{E}\}] + \mathbb{E}[R_T\mathbf{1}\{\mathcal{E}^c\}].$$

For the first term, Lemma 3 gives $\mathbb{E}[\mathrm{C}_T\mathbf{1}\{\mathcal{E}\}] \le 0$, and on $\mathcal{E}$ Lemmas 4, 5, and 2 give $\mathrm{O}_T \le 0$, $\mathrm{D}_T \le 2/\eta + 4\eta T$, and $\mathrm{S}_T \le K + 2\sqrt{2KTL}$. Hence

$$\mathbb{E}[R_T\mathbf{1}\{\mathcal{E}\}] \le \Big(\frac{2}{\eta} + 4\eta T + K + 2\sqrt{2KTL}\Big)\Pr(\mathcal{E}) \le \frac{2}{\eta} + 4\eta T + K + 2\sqrt{2KTL}.$$

For the second term, $R_T \le \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) \le T$ on every sample path because $v_t q_{k_t} \ge 0$ and each round contributes at most $v_t \le 1$ to the benchmark, so it is clear that

$$\mathbb{E}[R_T\mathbf{1}\{\mathcal{E}^c\}] \le T\Pr(\mathcal{E}^c) \le \delta T = 1/T.$$

Substituting $\eta = T^{-1/2}$ gives the stated bound. Finally, $K \le (T/L)^{1/3} + 1$ and $L = \Theta(\log T)$ give $\sqrt{KTL} = O(T^{2/3}L^{1/3})$, which is the claimed rate. □

B.8 Proof of Proposition 1

Proof. Since the algorithm bids a_{k_t} in round t, $\mathrm{Payment}(\mathrm{ALG} \mid G, V^T) - \mathrm{Reward}(\mathrm{ALG} \mid G, V^T) = \mathbb{E}[\sum_t (a_{k_t} - v_t)q_{k_t} \mid V^T]$, so $\Delta(\mathrm{ALG} \mid G, F) = \mathbb{E}[\sum_{t=1}^T (a_{k_t} - v_t)q_{k_t}]$. Using $a_{k_t} - c_{k_t} = \mathbf{1}\{k_t > 0\}/K$ and the definition of $\widetilde{g}_t$ in (6),

$$(a_{k_t} - v_t)q_{k_t} = \frac{\mathbf{1}\{k_t > 0\}q_{k_t}}{K} - (v_t - c_{k_t})q_{k_t} = \frac{\mathbf{1}\{k_t > 0\}q_{k_t}}{K} - \widetilde{g}_t + (v_t - c_{k_t})(U_{t,k_t} - q_{k_t}).$$

On $\mathcal{E}$, $U_{t,k_t} - q_{k_t} \ge 0$ by Lemma 6 for $k_t > 0$, and by $U_{t,0} = 1$ otherwise; together with $q_{k_t} \le 1$ and $c_{k_t} \ge 0$ this gives $(a_{k_t} - v_t)q_{k_t} \le 1/K - \widetilde{g}_t + v_t(U_{t,k_t} - q_{k_t})$. Summing over t and applying (9) and Lemma 2, on $\mathcal{E}$

$$\sum_{t=1}^T (a_{k_t} - v_t)q_{k_t} \le \frac{T}{K} - \sum_{t=1}^T \widetilde{g}_t + \mathrm{S}_T \le \frac{T}{K} + \frac{\log(1 + 2\eta T)}{\eta} + K + 2\sqrt{2KTL}.$$

On every sample path $\sum_t (a_{k_t} - v_t)q_{k_t} \le T$ because each summand is at most 1. Splitting the expectation according to $\mathcal{E}$ and using $\Pr(\mathcal{E}^c) \le \delta$ from Lemma 6,

$$\Delta(\mathrm{ALG} \mid G, F) \le \frac{T}{K} + \frac{\log(1 + 2\eta T)}{\eta} + K + 2\sqrt{2KTL} + \delta T.$$

Substituting $\eta = T^{-1/2}$ and $\delta = T^{-2}$ gives the stated inequality. For the rate, $T/K \le T^{2/3}L^{1/3}$ by the choice of K, $\sqrt{T}\log(1 + 2\sqrt{T}) = O(\sqrt{T}\log T)$, and $\sqrt{KTL} = O(T^{2/3}L^{1/3})$ as in Theorem 2. □

B.9 Realized terminal ROI violation and the zero-value safeguard

Proposition 3 (Terminal ROI violation bounds)**.** *Run Algorithm 1 with* $0 < \eta \le 1/2$, $L \ge \log(2KT/\delta)$, *and* $0 < \delta < 1$. *Write*

$$B_T = \frac{T}{K} + \frac{\log(1+2\eta T)}{\eta} + W_T, \qquad W_T = K + 2\sqrt{2KTL}.$$

For every $0 < \xi < 1$, *with probability at least* $1 - \delta - \xi$,

$$D_T^{\mathrm{real}} \le B_T + \sqrt{\frac{T}{2}\log\frac{1}{\xi}}. \tag{30}$$

Moreover,

$$\mathbb{E}[(D_T^{\mathrm{real}})_+] \le B_T + \delta T + \frac{\sqrt{T}}{4}. \tag{31}$$

Thus the parameter choices of Theorem 2 give the same $O(T^{2/3}(\log T)^{1/3})$ *rate for* $\mathbb{E}[(D_T^{\mathrm{real}})_+]$. *This proposition concerns Algorithm 1; it does not upgrade the terminal ROI feasibility in expectation of Algorithm 2 to almost-sure feasibility.*

Proof. Let $Q_T = \sum_t (b_t - v_t)G(b_t)$ and $M_T = \sum_t (b_t - v_t)(x_t - G(b_t))$, so $D_T^{\mathrm{real}} = Q_T + M_T$. The proof of Proposition 1 gives $Q_T \le B_T$ pointwise on $\mathcal{E}$. Under the natural interaction history, after v_t, b_t are determined but before x_t is observed, each summand of M_T has conditional mean zero and a conditional range of length $|b_t - v_t| \le 1$. The conditional Hoeffding bound therefore gives $\Pr(M_T > z) \le \exp(-2z^2/T)$ for $z \ge 0$. A union bound with $\Pr(\mathcal{E}^c) \le \delta$ proves (30). This argument never conditions the martingale on $\mathcal{E}$, which contains future feedback samples.

For the expectation bound, $Q_T \le T$ on every path and $B_T \ge 0$, hence $\mathbb{E}[(Q_T)_+] \le B_T + \delta T$. Orthogonality of martingale differences and the Bernoulli variance bound yield

$$\mathbb{E}[M_T^2] = \sum_t \mathbb{E}[(b_t - v_t)^2 G(b_t)(1 - G(b_t))] \le T/4.$$

Since $\mathbb{E}[M_T] = 0$, we have $\mathbb{E}[(M_T)_+] = \mathbb{E}|M_T|/2 \le \sqrt{T}/4$. Finally, $(Q_T + M_T)_+ \le (Q_T)_+ + (M_T)_+$ proves (31). □

Why the safeguard preserves all previous bounds. When $v_t = 0$, bidding zero gives zero reward, payment, and surplus, and $h_G(0, \lambda) = 0$. With $c_0 = 0$ and $U_{t,0} = 1$, the estimated reward and surplus and the estimation-error term also vanish. Hence the optimism comparison and both dual inequalities hold on these rounds. The selected positive-grid counts now satisfy $\sum_{k=1}^K N_{T+1,k} \le T$, which only improves the confidence-width bound. The executed-price error becomes $\mathbf{1}\{k_t > 0\}/K \le 1/K$. For the target update, the zero-value round still uses $\lambda_{t+1} = \lambda_t \exp(\eta\varepsilon_T)$; it must not be skipped. The surplus comparison holds because $s(0) = 0$. No extra unknown arm is introduced. In the instance $v_t = d_t = 0$ for all t, the unguarded algorithm pays exactly T/K, whereas the guarded algorithm pays zero on every trajectory. This is an edge-case correction, not a general pathwise-safety transformation.

C Omitted Proofs of Section 6

C.1 Comparing the two surplus margins

Lemma 7 (A sharper surplus-margin comparison)**.** *Let* $\kappa = \mathbb{E}[(v-d)_+]$ *and* $\sigma = \mathbb{E}_v[\sup_{0 \le b \le v}(v - b)G(b)]$. *Then* $\kappa^2/2 \le \sigma \le \kappa$. *In addition, with* $\phi(0) = 0$ *and* $\phi(z) = z(1+\log(1/z))$ *for* $z \in (0, 1]$,

$$\kappa \le \phi(\sigma). \tag{32}$$

In particular, for $\kappa > 0$,

$$\sigma \ge \frac{\kappa}{2(1 + \log(1/\kappa))} > 0. \tag{33}$$

The function on the right of (33) is increasing in κ. *These statements require no continuity of* G.

Proof. Write $s(v) = \sup_{0\le b\le v}(v-b)G(b)$. Equivalently, $s(v) = \sup_{b\in[0,1]}(v-b)_+G(b)$, which is bounded and 1-Lipschitz in v and hence measurable. For $b \le v$, $(v-b)\mathbf{1}\{d \le b\} \le (v-d)_+$; taking expectations and the supremum gives $\sigma \le \kappa$. For $v > 0$, averaging over $b \in [0, v]$ gives

$$s(v) \ge \frac{1}{v}\int_0^v (v-b)G(b)\,\mathrm{d}b = \mathbb{E}_d\left[\frac{(v-d)_+^2}{2v}\right].$$

Define the fraction to be zero at $v = 0$. Since $v \le 1$, averaging and applying Jensen's inequality proves $\sigma \ge \kappa^2/2$.

For the sharper comparison, fix v and let $Y_v = (v-d)_+$. For $0 < y \le v$,

$$y\Pr(Y_v \ge y) = yG(v-y) \le s(v).$$

The same tail probability is zero when $y > v$, and is at most one everywhere. Therefore, if $s(v) > 0$,

$$\mathbb{E}_d[Y_v] = \int_0^1 \Pr(Y_v \ge y)\,\mathrm{d}y \le \int_0^1 \min\{1, s(v)/y\}\,\mathrm{d}y = \phi(s(v)).$$

When $s(v) = 0$, the same conclusion follows since every strictly positive tail is zero. The function ϕ is continuous and concave on $[0, 1]$, with $\phi''(z) = -1/z < 0$ on $(0, 1]$. Thus Jensen's inequality gives $\kappa \le \mathbb{E}_v[\phi(s(v))] \le \phi(\sigma)$.

To obtain the explicit lower bound, put $a = 1 + \log(1/\kappa) \ge 1$ and $z_0 = \kappa/(2a)$. Since $\log(2a) \le a$ for $a \ge 1$,

$$\phi(z_0) = \frac{\kappa}{2a}\big(a + \log(2a)\big) \le \kappa.$$

The function ϕ is strictly increasing on $[0, 1]$, so $\phi(\sigma) \ge \kappa \ge \phi(z_0)$ implies $\sigma \ge z_0$. Direct differentiation proves monotonicity of $\kappa/[2(1 + \log(1/\kappa))]$. □

C.2 Proof of Theorem 3

Proof. Step 1: the original assumption supplies a fixed margin. Write $\kappa = \kappa(G, F)$ and $\sigma = \sigma(G, F)$. Lemma 7 gives $0 < \sigma \le 1$ and the uniform lower bound $\sigma \ge \sigma_0 := \kappa_{\min}/[2(1 + \log(1/\kappa_{\min}))]$ under Assumption 1. The lemma is only analytical; the algorithm does not execute an auxiliary conservative phase or require the margin as input.

Step 2: confidence coupling and surplus error. Generate the independent Bernoulli feedback arrays $(Z_{k,n})_{k\le K, n\le T}$ from Section 5.2, independently of the values, and use the next sample of the selected arm as feedback. This construction has the same transcript law as the actual auctions also for the modified update: the selected action is determined by the current value and the observed past, and the next allocation has mean q_k for that action. Let $\mathcal{E}$ be the simultaneous confidence event from that section. A union bound gives $\Pr(\mathcal{E}^c) \le 2KTe^{-L} \le \delta$, and $\mathcal{E}$ is independent of the entire value sequence. Zero-value rounds have zero estimation error. The count-based proof of Lemma 2 for positive-grid actions is otherwise unchanged by the dual update. Thus, on $\mathcal{E}$,

$$\sum_{t=1}^{T} v_t(U_{t,k_t} - q_{k_t}) \le W_T. \tag{34}$$

Define the conditional-mean surplus and two trajectory quantities by

$$\bar{g}_t = (v_t - a_{k_t})q_{k_t},$$
$$Q_T = -\sum_{t=1}^{T}\bar{g}_t, \qquad R_T = \mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \sum_{t=1}^{T} v_t q_{k_t}.$$

Under the natural interaction history, $\mathbb{E}[x_t \mid \text{past}, v_t, k_t] = q_{k_t}$, so unconditionally

$$\Delta(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) = \mathbb{E}[Q_T], \qquad \mathrm{Regret}(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) = \mathbb{E}[R_T]. \tag{35}$$

We split the expectations of these quantities according to $\mathcal{E}$; we do not assume that x_t remains unbiased after conditioning on this event, which includes future feedback samples.

On $\mathcal{E}$, since $U_{t,k_t} \geq q_{k_t}$,

$$\begin{aligned}\widetilde{g}_t - \bar{g}_t &= (v_t - c_{k_t})(U_{t,k_t} - q_{k_t}) + \mathbf{1}\{k_t > 0\} q_{k_t}/K \\ &\leq v_t(U_{t,k_t} - q_{k_t}) + 1/K.\end{aligned}$$

Summing and using (34) gives $\sum_t (\widetilde{g}_t - \bar{g}_t) \leq T/K + W_T$. The new update, with $\lambda_1 = 1$, gives the exact identity

$$-\sum_{t=1}^{T} \widetilde{g}_t = \frac{\log \lambda_{T+1}}{\eta} - \varepsilon_T T. \tag{36}$$

Consequently, pointwise on $\mathcal{E}$,

$$Q_T \leq \frac{T}{K} + W_T + \frac{\log \lambda_{T+1}}{\eta} - \varepsilon_T T. \tag{37}$$

It remains to control the terminal dual residual for the modified, rather than the original, update.

Step 3: a one-step bound for the modified dual update. Write $\varepsilon = \varepsilon_T$ and $\alpha = \eta(1-\eta)$. The unchanged score maximization implies, on every path,

$$|\widetilde{g}_t| \leq 1, \qquad v_t U_{t,k_t} + \lambda_t \widetilde{g}_t \geq 0, \qquad \lambda_t \widetilde{g}_t \geq -1.$$

The middle inequality follows from the nonnegative first-grid score when $v_t > 0$ and from $\widetilde{g}_t = 0$ when $v_t = 0$. These facts imply $\lambda_t \widetilde{g}_t^2 \leq \lambda_t \widetilde{g}_t + 2$: for $\widetilde{g}_t \geq 0$ use $\widetilde{g}_t^2 \leq \widetilde{g}_t$, and otherwise use $\lambda_t \widetilde{g}_t^2 \leq -\lambda_t \widetilde{g}_t \leq \lambda_t \widetilde{g}_t + 2$. Applying $e^{-x} \leq 1 - x + x^2$ for $|x| \leq 1/2$ with $x = \eta \widetilde{g}_t$ yields

$$\begin{aligned}\lambda_{t+1} &= e^{\eta\varepsilon} \lambda_t e^{-\eta \widetilde{g}_t} \\ &\leq e^{\eta\varepsilon} \left[\lambda_t - \alpha \lambda_t \widetilde{g}_t + 2\eta^2\right].\end{aligned} \tag{38}$$

Unlike the original dual bound, this inequality explicitly retains the factor $e^{\eta\varepsilon}$.

Step 4: bounding the expected dual variable. On $\mathcal{E}$, the pointwise optimism of Lemma 4 applies to every nonnegative multiplier, including those produced by (17). Hence

$$v_t U_{t,k_t} + \lambda_t \widetilde{g}_t \geq h_G(v_t, \lambda_t) \geq \lambda_t s(v_t),$$

since for $b \leq v$ its true Lagrangian score is at least $\lambda(v-b)G(b)$. Because $v_t U_{t,k_t} \leq 1$ and $\lambda_t > 0$,

$$\widetilde{g}_t \geq s(v_t) - 1/\lambda_t \qquad \text{on } \mathcal{E}. \tag{39}$$

Let $\mathcal{H}_{t-1}$ be generated by all feedback arrays and $v_1, \ldots, v_{t-1}$. Both λ_t and $\mathcal{E}$ are measurable with respect to $\mathcal{H}_{t-1}$, whereas v_t is independent of it and has law F. Thus, on $\mathcal{E}$,

$$\mathbb{E}[\widetilde{g}_t \mid \mathcal{H}_{t-1}] \geq \sigma - 1/\lambda_t.$$

Taking conditional expectations in (38) gives, on $\mathcal{E}$,

$$\mathbb{E}[\lambda_{t+1} \mid \mathcal{H}_{t-1}] \leq e^{\eta\varepsilon} \left[(1 - \eta(1-\eta)\sigma)\lambda_t + \eta + \eta^2\right]. \tag{40}$$

Now $T \geq 16$ implies $\eta \leq 1/4$. Set $z = \eta\sigma \leq 1/4$. The condition $2\varepsilon \leq \sigma$ gives $\eta\varepsilon \leq z/2$. Using $e^u \leq 1/(1-u)$ for $0 \leq u < 1$,

$$e^{\eta\varepsilon}(1 - \eta(1-\eta)\sigma) \leq \frac{1 - 3z/4}{1 - z/2} \leq 1 - z/4,$$

$$e^{\eta\varepsilon}(\eta + \eta^2) \leq \frac{(5/4)\eta}{1 - 1/8} = \frac{10}{7}\eta \leq 2\eta.$$

For $m_t = \mathbb{E}[\lambda_t \mid \mathcal{E}]$, (40) therefore implies

$$m_{t+1} \leq (1 - \eta\sigma/4) m_t + 2\eta, \qquad m_1 = 1.$$

Induction gives

$$\mathbb{E}[\lambda_t \mid \mathcal{E}] \leq \min\{1 + 2\eta(t-1),\, 8/\sigma\}. \tag{41}$$

The first bound drops the nonpositive term; the second follows because $1 \leq 8/\sigma$ and substituting $8/\sigma$ into the right-hand side leaves it unchanged. All expectations are finite, since $|\log \lambda_t| \leq \eta(t-1)(1+\varepsilon)$ on every path. Jensen's inequality now gives

$$\mathbb{E}[\log \lambda_{T+1} \mid \mathcal{E}] \leq \log(1 + 2\eta T). \tag{42}$$

Step 5: terminal expected ROI, including confidence failure. Combining (37) with (42) and the definition of ε,

$$\mathbb{E}[Q_T \mid \mathcal{E}] \le \frac{T}{K} + W_T + \frac{\log(1+2\eta T)}{\eta} - \varepsilon T = A_T - \varepsilon T = -2\delta T.$$

On every path, $Q_T \le T$. With $p = \Pr(\mathcal{E}^c) \le \delta$, (35) implies

$$\begin{aligned}\Delta(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) &\le (1-p)(-2\delta T) + pT \\ &\le -2\delta T + \delta T(1+2\delta) = -\delta T(1-2\delta) = -1/T + 2/T^3 < 0.\end{aligned}$$

This proves (18). Equivalently, expected actual surplus is at least $1/T - 2/T^3$. The surplus target compensates for the cumulative estimation and price errors, the terminal dual term, and confidence failure, yielding nonnegative expected total surplus.

Step 6: regret against the original benchmark. Rearranging (38) and summing gives

$$\begin{aligned}\alpha \sum_{t=1}^{T} \lambda_t \widetilde{g}_t &\le \sum_{t=1}^{T} (\lambda_t - e^{-\eta\varepsilon}\lambda_{t+1}) + 2\eta^2 T \\ &= 1 - e^{-\eta\varepsilon}\lambda_{T+1} + (1-e^{-\eta\varepsilon}) \sum_{t=2}^{T} \lambda_t + 2\eta^2 T \\ &\le 1 + \eta\varepsilon \sum_{t=1}^{T} \lambda_t + 2\eta^2 T.\end{aligned}$$

Since $\alpha \ge \eta/2$, it follows that

$$\begin{aligned}\sum_{t=1}^{T} \lambda_t \widetilde{g}_t &\le 2/\eta + 4\eta T + 2\varepsilon \sum_{t=1}^{T} \lambda_t, \\ \mathbb{E}\left[\sum_{t=1}^{T} \lambda_t \widetilde{g}_t \,\middle|\, \mathcal{E}\right] &\le 2/\eta + 4\eta T + 16\varepsilon T/\sigma, \end{aligned} \tag{43}$$

where the second line uses (41).

For completeness, let $\mathsf{B}_T = \mathbb{E}_{V^T}[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T)]$ and recall $H(\lambda) = \mathbb{E}_{v\sim F}[h_G(v,\lambda)]$. Weak duality for the original ROI benchmark gives $\mathsf{B}_T \le TH(\lambda)$ for every fixed $\lambda \ge 0$. Since v_t is independent of $\mathcal{H}_{t-1}$,

$$\mathbb{E}[h_G(v_t, \lambda_t) \mid \mathcal{H}_{t-1}] = H(\lambda_t) \ge \mathsf{B}_T/T.$$

Using measurability of $\mathcal{E}$ with respect to $\mathcal{H}_{t-1}$ and independence of $\mathcal{E}$ from V^T,

$$\mathbb{E}\left[\mathrm{Reward}(\mathrm{OPT} \mid G, V^T) - \sum_{t=1}^{T} h_G(v_t, \lambda_t) \,\middle|\, \mathcal{E}\right] \le 0. \tag{44}$$

This comparison does not restrict the benchmark to policies satisfying a positive surplus target. On $\mathcal{E}$, optimism yields

$$v_t q_{k_t} \ge h_G(v_t, \lambda_t) - \lambda_t \widetilde{g}_t - v_t(U_{t,k_t} - q_{k_t}).$$

Combining (34), (43), and (44),

$$\mathbb{E}[R_T \mid \mathcal{E}] \le 2/\eta + 4\eta T + W_T + 16\varepsilon T/\sigma.$$

Since $R_T \le T$ on every path, splitting the expectation according to $\mathcal{E}$ and using (35) gives

$$\mathrm{Regret}(\mathrm{ALG}^{\mathrm{tar}} \mid G, F) \le 2/\eta + 4\eta T + W_T + 16\varepsilon T/\sigma + \delta T.$$

Substituting $\eta = T^{-1/2}$ and $\delta = T^{-2}$ proves (19).

Step 7: the uniform asymptotic conclusion under the original assumption. The parameter choices give $A_T, W_T = O(T^{2/3}(\log T)^{1/3})$ and $\varepsilon_T \to 0$, with universal constants. For a fixed $\kappa_{\min} > 0$,

Step 1 gives the uniform lower bound $\sigma \geq \sigma_0 := \kappa_{\min}/[2(1+\log(1/\kappa_{\min}))]$. Choose an integer $T_{\mathrm{ROI}} \geq 16$ such that $2\varepsilon_T \leq \sigma_0$ for every integer $T \geq T_{\mathrm{ROI}}$. Such a threshold exists by convergence and depends only on $\kappa_{\min}$. The finite-horizon result applies on every instance in the class beyond this threshold, and

$$16\varepsilon_T T/\sigma \leq \frac{32(1+\log(1/\kappa_{\min}))}{\kappa_{\min}}\varepsilon_T T = O_{\kappa_{\min}}(T^{2/3}(\log T)^{1/3}).$$

This proves (20), with C and T_{ROI} independent of T and of the particular instance in the fixed-margin class. Neither constant is an input to the algorithm. The theorem asserts terminal expected feasibility only for the stated horizons; it does not assert sample-path feasibility or a uniform guarantee over arbitrarily small margins. □

C.3 A SHARPER FINITE-HORIZON FEASIBILITY CERTIFICATE

The condition $2\varepsilon_T \leq \sigma$ in Theorem 3 is a convenient sufficient condition, not a characterization of feasibility. The next proposition retains the exact coefficients of (40) instead of replacing them by $1-\eta\sigma/4$ and 2η. It analyzes the same target update and does not change its inputs.

Proposition 4 (Finite-horizon certificate)**.** *Run the target update with $0 < \eta \leq 1/2$, a deterministic target $\varepsilon \geq 0$, $L \geq \log(2KT/\delta)$, and $0 < \delta < 1$. Write $\sigma = \sigma(G,F)$ and define*

$$\rho = e^{\eta\varepsilon}\big(1-\eta(1-\eta)\sigma\big), \qquad c = e^{\eta\varepsilon}(\eta+\eta^2).$$

If $\rho < 1$, set

$$M = \max\{1, c/(1-\rho)\}, \qquad B = \frac{T}{K} + W_T + \frac{\log M}{\eta} - \varepsilon T.$$

Then

$$\Delta(\mathrm{ALG}^{\mathrm{tar}} \mid G,F) \leq B + \delta(T-B)_+, \tag{45}$$

$$\mathrm{Regret}(\mathrm{ALG}^{\mathrm{tar}} \mid G,F) \leq \frac{2}{\eta} + 4\eta T + W_T + 2\varepsilon T M + \delta T. \tag{46}$$

In particular, terminal expected ROI holds whenever $B \leq -\delta T/(1-\delta)$. For the target ε_T in (16), the simpler condition $\rho < 1$ and $M \leq 1 + 2\eta T$ recovers $\Delta \leq -\delta T(1-2\delta)$.

Proof. The conditional recurrence (40) was derived before imposing $2\varepsilon \leq \sigma$, so it remains valid here. For $m_t = \mathbb{E}[\lambda_t \mid \mathcal{E}]$ it gives $m_{t+1} \leq \rho m_t + c$. We have $0 \leq \rho < 1$ and $m_1 = 1$. Since $M \geq 1$ and $\rho M + c \leq M$, induction gives $m_t \leq M$ for all t. Jensen's inequality yields $\mathbb{E}[\log \lambda_{T+1} \mid \mathcal{E}] \leq \log M$. The exact target identity and surplus error bound in (37) then give $\mathbb{E}[Q_T \mid \mathcal{E}] \leq B$. On $\mathcal{E}^c$, $Q_T \leq T$. For $p = \Pr(\mathcal{E}^c) \leq \delta$,

$$\Delta \leq (1-p)B + pT = B + p(T-B) \leq B + \delta(T-B)_+.$$

If $B \leq -\delta T/(1-\delta)$, the right-hand side is nonpositive. For the stated target, $M \leq 1 + 2\eta T$ gives $B \leq -2\delta T$ and hence the last conclusion.

The rearrangement of (38) gives $\sum_t \lambda_t \widetilde{g}_t \leq 2/\eta + 4\eta T + 2\varepsilon \sum_t \lambda_t$ for $\eta \leq 1/2$. Combining $m_t \leq M$ with the unchanged benchmark comparison and estimation bound gives the regret result, including δT for confidence failure. □

Conservativeness and unavailable margins. A failed certificate is not evidence of algorithmic infeasibility. Conversely, a passed certificate uses a true margin or a valid analytical lower bound on it; substituting an uncertified empirical estimate is not justified by this proposition. The algorithm still uses only $T, K, \eta, \delta, L, \varepsilon_T$, not σ.

D OMITTED PROOFS OF SECTION 7

D.1 PROOF OF THEOREM 4

Proof. The construction uses a fixed value and a family of competing-bid distributions that differ only on one short interval each. For $T \geq 1728$ set

$$m = \lceil 8T^{1/3} \rceil, \qquad \varepsilon = \frac{1}{3m}, \qquad \beta_i = \frac{1}{3} + i\varepsilon, \quad i = 0, \ldots, m,$$

so that $\beta_0 = 1/3$ and $\beta_m = 2/3$. The value distribution $F^\star$ puts all its mass on $v^\star = 4/13$, so $v_t = v^\star$ in every round. The base competing-bid distribution is

$$G_\circ(b) = \begin{cases} 3/4, & 0 \le b < \beta_0, \\ 3/(4-\beta_i), & \beta_i \le b < \beta_{i+1}, \quad i = 0, \dots, m-1, \\ 9/10, & \beta_m \le b < 1, \\ 1, & b = 1. \end{cases}$$

Since $3/(4-\beta_i)$ increases with i from $9/11$ to $3/(4-\beta_m) = 9/10$, the function $G_\circ$ is nondecreasing and right-continuous, so it is a valid distribution function. For $j \in \{0, \dots, m-1\}$ let $I_j = [\beta_j, \beta_{j+1})$, $\theta_j = 3/(4-\beta_{j+1}) = G_\circ(\beta_{j+1})$, and

$$G_j(b) = G_\circ(b) + \big(\theta_j - G_\circ(\beta_j)\big)\mathbf{1}\{b \in I_j\}.$$

Thus G_j moves the probability mass that $G_\circ$ places at β_{j+1} to β_j; it equals θ_j on I_j and agrees with $G_\circ$ elsewhere, and it is again a valid distribution function. Instance j has values drawn from $F^\star$ and competing bids drawn from G_j.

The base distribution is chosen so that $bG_\circ(b) = 4G_\circ(b) - 3$ at every grid point $b \in \{0, \beta_0, \dots, \beta_m, 1\}$: the ROI surplus of every grid bid is the same affine function of its winning probability, so under $G_\circ$ a bidder gains nothing by locating the grid precisely. Under G_j, the bid β_j wins with the higher probability θ_j at the lower price β_j, which is worth about ε per round, but detecting this requires about ε^{-2} bids in I_j.

For two numbers $p, q \in (0,1)$ let $\mathrm{kl}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ denote the relative entropy between Bernoulli distributions with means p and q, and for two probability distributions P, Q on the same space let $\mathrm{KL}(P\|Q)$ denote their relative entropy.

Step 1: payment versus winning probability. We claim that for every j and every bid $b \in [0,1]$,

$$bG_j(b) \ge 4G_j(b) - 3 - \varepsilon\theta_j\mathbf{1}\{b \in I_j\}, \tag{47}$$

with equality at $b = 0$ and at $b = \beta_j$. First consider $G_\circ$. At $b = 0$ both sides of $bG_\circ(b) = 4G_\circ(b) - 3$ equal 0. At $b = \beta_i$ with $0 \le i \le m$, $G_\circ(\beta_i) = 3/(4-\beta_i)$ (for $i = m$ this is $9/10$), and $4 \cdot \frac{3}{4-\beta_i} - 3 = \frac{3\beta_i}{4-\beta_i} = \beta_i G_\circ(\beta_i)$. At $b = 1$ both sides equal 1. Between consecutive points of $\{0, \beta_0, \dots, \beta_m, 1\}$, $G_\circ$ is constant while $bG_\circ(b)$ is nondecreasing in b, so $bG_\circ(b) \ge 4G_\circ(b) - 3$ for all $b \in [0,1]$. Now G_j agrees with $G_\circ$ outside I_j, and on I_j we have $G_j(b) = \theta_j$ and, using $\beta_{j+1}\theta_j = 4\theta_j - 3$ from the identity at β_{j+1},

$$bG_j(b) = b\theta_j \ge \beta_j\theta_j = (\beta_{j+1} - \varepsilon)\theta_j = 4\theta_j - 3 - \varepsilon\theta_j,$$

with equality at $b = \beta_j$. This proves (47).

Step 2: a lower bound on the benchmark. Consider the policy that, in every round, bids β_j with probability

$$\alpha_j = \frac{(3/4)v^\star}{(3/4)v^\star + (\beta_j - v^\star)\theta_j}$$

and bids 0 otherwise. Since $\beta_j \ge 1/3 > 4/13 = v^\star$, we have $0 < \alpha_j < 1$. Moreover, $(3/4)v^\star = 3/13$, $\beta_j - v^\star \le 2/3 - 4/13 = 14/39$, and $\theta_j \le 9/10$, so the denominator is at most $3/13 + (14/39)(9/10) = 36/65$ and $\alpha_j \ge (3/13)(65/36) = 5/12$. Bidding 0 wins with probability $G_j(0) = 3/4$ at zero price, and bidding β_j wins with probability θ_j at price β_j. The per-round expected value is $v^\star\big((1-\alpha_j)\frac{3}{4} + \alpha_j\theta_j\big)$ and the per-round expected payment is $\alpha_j\beta_j\theta_j$; by the choice of α_j these are equal, since $v^\star\frac{3}{4} = \alpha_j\big(v^\star\frac{3}{4} + (\beta_j - v^\star)\theta_j\big)$ rearranges to $v^\star\big((1-\alpha_j)\frac{3}{4} + \alpha_j\theta_j\big) = \alpha_j\beta_j\theta_j$. The policy is therefore feasible for (1). To evaluate its reward r_j, use the equality case of (47) at β_j and $v^\star = 4/13$: from $r_j = \alpha_j\beta_j\theta_j = \alpha_j(4\theta_j - 3 - \varepsilon\theta_j)$ and $r_j = \frac{3}{13} + \frac{4}{13}\alpha_j(\theta_j - \frac{3}{4})$, multiplying the second identity by 13 and substituting the first gives $3 = 12\alpha_j(4\theta_j - 3) - 13\alpha_j\varepsilon\theta_j$, hence $\alpha_j(4\theta_j - 3) = \frac{3 + 13\alpha_j\varepsilon\theta_j}{12}$ and

$$r_j = \alpha_j(4\theta_j - 3) - \alpha_j\varepsilon\theta_j = \frac{1}{4} + \frac{\alpha_j\varepsilon\theta_j}{12} \ge \frac{1}{4} + \frac{5}{176}\varepsilon,$$

where the inequality uses $\alpha_j \geq 5/12$ and $\theta_j \geq 9/11$. Since the policy is feasible, summing over T rounds gives

$$\mathrm{Reward}(\mathrm{OPT} \mid G_j, V^T) \geq \frac{T}{4} + \frac{5}{176}\varepsilon T. \tag{48}$$

Step 3: the algorithm rarely bids in the perturbed interval. Let $P_\circ$ and P_j denote the joint distributions of the algorithm's bids and observed allocations over the T rounds under $G_\circ$ and under G_j, respectively, and write $\mathbb{E}_\circ$ and $\mathbb{E}_j$ for the corresponding expectations. Define $M_j = \sum_{t=1}^{T} \mathbf{1}\{b_t \in I_j\}$. The intervals $I_0, \ldots, I_{m-1}$ are disjoint, so $\sum_j M_j \leq T$ and there is an index j with $\mathbb{E}_\circ[M_j] \leq T/m$; fix this j. The winning probabilities of a bid b under the two distributions differ only for $b \in I_j$, where

$$\begin{aligned} G_j(b) - G_\circ(b) = \theta_j - G_\circ(\beta_j) &= \frac{3}{4-\beta_{j+1}} - \frac{3}{4-\beta_j} \\ &= \frac{3\varepsilon}{(4-\beta_{j+1})(4-\beta_j)} \leq \frac{3\varepsilon}{(10/3)^2} = \frac{27}{100}\varepsilon, \end{aligned}$$

using $\beta_i \leq 2/3$. Also $\theta_j \in [9/11, 9/10]$, so $\theta_j(1-\theta_j) \geq \min\{\frac{9}{11}\cdot\frac{2}{11}, \frac{9}{10}\cdot\frac{1}{10}\} = \frac{9}{100}$. The bound $\mathrm{kl}(p\|q) \leq (p-q)^2/(q(1-q))$, which follows from $\log x \leq x - 1$, gives for $b \in I_j$

$$\mathrm{kl}\big(G_\circ(b) \,\|\, G_j(b)\big) \leq \frac{(27\varepsilon/100)^2}{9/100} = \frac{81}{100}\varepsilon^2 \leq \varepsilon^2,$$

and the relative entropy is zero for $b \notin I_j$. The observation in round t is the allocation x_t, whose conditional law given the past and the bid b_t is Bernoulli with mean $G_\circ(b_t)$ or $G_j(b_t)$; the values are the constant $v^\star$; and the conditional law of b_t given the past is the same under both instances because it is determined by the algorithm. By the chain rule for relative entropy (see, e.g., Lattimore & Szepesvári, 2020),

$$\mathrm{KL}(P_\circ \| P_j) = \sum_{t=1}^{T} \mathbb{E}_\circ\big[\mathrm{kl}\big(G_\circ(b_t) \,\|\, G_j(b_t)\big)\big] \leq \varepsilon^2\, \mathbb{E}_\circ[M_j] \leq \frac{\varepsilon^2 T}{m}.$$

Since M_j takes values in $[0, T]$, Pinsker's inequality (Tsybakov, 2009) gives $\mathbb{E}_j[M_j] - \mathbb{E}_\circ[M_j] \leq T\sqrt{\mathrm{KL}(P_\circ\|P_j)/2}$, hence

$$n_j := \mathbb{E}_j[M_j] \leq \frac{T}{m} + T\varepsilon\sqrt{\frac{T}{2m}} \quad\Longrightarrow\quad \frac{n_j}{T} \leq \frac{1}{m} + \frac{\sqrt{T}}{3\sqrt{2}\,m^{3/2}}.$$

Using $m \geq 8T^{1/3}$ and $T \geq 1728 = 12^3$: $1/m \leq 1/(8\cdot 12) = 1/96$, and $\sqrt{T}/(3\sqrt{2}\,m^{3/2}) \leq \sqrt{T}/(3\sqrt{2}\cdot 8^{3/2}\sqrt{T}) = 1/96$. Therefore

$$n_j \leq \frac{T}{48}. \tag{49}$$

Step 4: combining. By (47), $(b-4)G_j(b) = bG_j(b) - 4G_j(b) \geq -3 - \varepsilon\theta_j \mathbf{1}\{b \in I_j\}$ for every bid b. Using the definitions of regret and violation with $v_t = v^\star$ and the identity $13v^\star = 4$,

$$\begin{aligned} &\mathrm{Regret}(\mathrm{ALG} \mid G_j, F^\star) + \frac{1}{12}\Delta(\mathrm{ALG} \mid G_j, F^\star) \\ &\quad= \mathrm{Reward}(\mathrm{OPT} \mid G_j, V^T) - \mathbb{E}_j\Big[\sum_{t=1}^{T} v^\star G_j(b_t)\Big] + \frac{1}{12}\mathbb{E}_j\Big[\sum_{t=1}^{T}(b_t - v^\star)G_j(b_t)\Big] \\ &\quad= \mathrm{Reward}(\mathrm{OPT} \mid G_j, V^T) + \frac{1}{12}\mathbb{E}_j\Big[\sum_{t=1}^{T}(b_t - 13v^\star)G_j(b_t)\Big] \\ &\quad= \mathrm{Reward}(\mathrm{OPT} \mid G_j, V^T) + \frac{1}{12}\mathbb{E}_j\Big[\sum_{t=1}^{T}(b_t - 4)G_j(b_t)\Big] \\ &\quad\geq \frac{T}{4} + \frac{5}{176}\varepsilon T - \frac{3T}{12} - \frac{\varepsilon\theta_j n_j}{12} \geq \varepsilon T\Big(\frac{5}{176} - \frac{1}{576}\Big) \geq \frac{\varepsilon T}{39}, \end{aligned}$$

where the first inequality uses (48), the second uses (49) and $\theta_j \leq 1$, and the last is a numerical check ($5/176 - 1/576 > 0.0266 > 1/39$). Replacing Δ by $\max\{\Delta, 0\}$ can only increase the left-hand side. Finally, $m \leq 8T^{1/3} + 1 \leq 9T^{1/3}$, so $\varepsilon T/39 = T/(117m) \geq T^{2/3}/1053$, which proves (21). □

D.2 Proof of Proposition 2

Proof. For each $T \geq 1728$, choose the instance $(G_j, F^\star)$ supplied by Theorem 4. Applying (22) on this instance to (21) gives

$$\mathrm{Regret}(\mathrm{ALG}_T \mid G_j, F^\star) \geq \frac{T^{2/3}}{1053} - \frac{1}{12}\max\{\Delta(\mathrm{ALG}_T \mid G_j, F^\star), 0\} \geq \frac{T^{2/3}}{1053} - \frac{r_T}{12}.$$

Taking the supremum over instances proves (23) with $c = 1/1053$. Since $r_T = o(T^{2/3})$, there is a threshold $T_1 \geq T_0$, which may depend on (r_T), such that the right-hand side is at least $T^{2/3}/2106$ for every $T \geq T_1$. □

E Technical Comparison with Fixed Exploration

E.1 What changes relative to fixed exploration

The bandit algorithm of Li et al. (2026, Section 5) uses a fixed exploration phase on a bid grid, followed by optimization under an optimistic distribution estimate and an upward adjustment of the selected bid. Their method already separates the prices used for optimization and execution. UCB-DMD combines this separation with action-specific confidence bounds updated throughout the horizon, direct maximization of an optimistic linear Lagrangian, and a common surplus estimate for bid selection and the dual update.

With n observations per grid action, the fixed exploration analysis has schematic costs $Kn + T/\sqrt{n} + T/K$. Balancing these terms gives $K \asymp T^{1/4}$, $n \asymp T^{1/2}$, and a $\widetilde{O}(T^{3/4})$ bound. UCB-DMD instead accumulates confidence widths along the selected actions, yielding leading learning and execution costs $K + O(\sqrt{KT\log T}) + T/K$. Balancing them gives the improved $O(T^{2/3}(\log T)^{1/3})$ rate.

Two design choices prevent the dual multiplier from amplifying approximation errors in the regret bound. At a fixed probability estimate, replacing the virtual price by the executed price changes the score by at most λ_t/K; using a different surplus estimate $\widehat{g}_t$ in the dual update introduces the mismatch $\sum_t \lambda_t(\widetilde{g}_t - \widehat{g}_t)$. Virtual-price optimism avoids the first error, while using the same surplus estimate in the score and update eliminates the second. The optimism argument in Lemma 4 uses pointwise upper confidence bounds on winning probabilities and does not require them to form a valid CDF. The nonnegative coefficient condition is essential: increasing the estimated winning probability lowers the Lagrangian score when its coefficient is negative.