ONLINE BILATERAL TRADE UNDER STOCHASTIC DELAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Online bilateral trade has recently attracted significant attention in the learning and economics literature. It studies sequential pricing between arriving seller–buyer pairs, where a learner posts prices to facilitate trades and maximize the gain from trade. Existing works typically assume immediate feedback, whereas practical platforms often experience delayed and incomplete responses due to factors such as network latency, payment verification, and slow user interactions. In this paper, we study online bilateral trade under stochastic delayed and censored acceptance-only feedback, where acceptance signals arrive after random delays and missing signals may correspond either to rejection or delayed feedback. We first establish a regret lower bound showing that stochastic delays introduce an unavoidable additional cost proportional to the expected delay. For exogenous stochastic delays, we exploit boundary prices that guarantee one-sided acceptance to develop a boundary-calibrated scouting bandit algorithm achieving near-optimal regret up to logarithmic factors. We further extend our results to valuation-dependent delays, where independence-based correction fails, and propose a boundary-calibrated adaptive window elimination algorithm achieving the same regret rate.

1 INTRODUCTION

Many online platforms, such as Uber, Upwork, and Lyft Bacchiocchi et al. (2025), operate by posting prices to arriving seller–buyer pairs and learning pricing strategies to facilitate trades. This interaction is captured by the online bilateral trade framework Cesa-Bianchi et al. (2024a), where a learner posts a price P_t for a seller–buyer pair with private valuations (S_t, B_t) and receives gain from trade. The learner aims to maximize cumulative gain from trade relative to the best fixed price.

Despite its practical relevance, existing online bilateral trade models typically assume that feedback from both sides of the market is observed immediately after each pricing decision. In practice, acceptance responses may be delayed due to network latency, payment verification, or slow user interactions. For example, transaction confirmations in online marketplaces such as eBay may arrive substantially later while the platform continues to make pricing decisions Fong & Waisman (2025); Backus et al. (2020).

Beyond delayed feedback, practical platforms face another challenge: missing acceptance signals are inherently ambiguous. A missing signal may correspond either to a true rejection or to an acceptance response that has not yet arrived. For example, in ride-sharing platforms, a user's failure to respond to a posted price may result from either delayed response or rejection, while the platform cannot distinguish between these cases.

Motivated by these considerations, we study online bilateral trade under stochastic delayed and censored acceptance-only feedback, where a missing signal may correspond either to rejection or delayed feedback. The learner must make pricing decisions based on delayed and incomplete information, a challenge not captured by existing models with instantaneous feedback.

Our contributions are summarized as follows:

- We introduce online bilateral trade with stochastic delayed and censored acceptance-only feedback. We establish a regret lower bound showing that stochastic delays introduce an unavoidable additional cost proportional to the expected delay.

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.

- We develop a boundary-calibrated learning algorithm for exogenous stochastic delays. By exploiting boundary prices that guarantee one-sided acceptance, the learner estimates delay information and achieves near-optimal regret up to logarithmic factors.
- We extend the framework to valuation-dependent delays, where delays may correlate with private valuations and independence-based correction is no longer applicable. We propose a boundary-calibrated adaptive window elimination algorithm that controls the induced bias through adaptive boundary calibration and achieves the same regret order up to logarithmic factors.

1.1 Related Work

Online bilateral trade. Cesa-Bianchi et al. (2024a) initiated the study of online bilateral trade from a regret-minimization perspective under stochastic seller and buyer valuations. Subsequent works have extended this framework to adversarial and smoothed markets Azar et al. (2022); Cesa-Bianchi et al. (2024b); Lunghi et al. (2026b), relaxed budget requirements and global budget balance Bernasconi et al. (2024); Chen et al. (2025); Lunghi et al. (2026a;c); Jin (2026), and other settings including asynchronous feedback, fairness, contextual valuations, brokerage, multi-buyer markets, and profit maximization Bacchiocchi et al. (2025); Bachoc et al. (2024a; 2026); Gaucher et al. (2025); Cosson et al. (2026); Bachoc et al. (2025); Bolić et al. (2024); Bachoc et al. (2024b); Cesari & Colomboni (2025); Lunghi et al. (2025); Di Gregorio et al. (2025). However, existing works do not consider delayed feedback in online bilateral trade, which is the main focus of our work.

Bandits with delayed feedback. Our work is also related to the literature on multi-armed bandits with delayed feedback. Since the underlying framework differs from ours, we provide only a brief overview. Early works studied fixed or bounded delays Dudik et al. (2011); Mandel et al. (2015), while subsequent works considered more general stochastic delayed-feedback settings Zhou et al. (2019); Vernade et al. (2020); Howson et al. (2023). Another line of work studies adversarial, nonstochastic, arm-dependent, or reward-dependent delays Thune et al. (2019); Lancewicki et al. (2021); Gyorgy & Joulani (2021); Van Der Hoeven & Cesa-Bianchi (2022); Zimmert & Seldin (2020); Tang et al. (2024). Related models also consider delay-dependent payoffs, where the delay itself affects the reward Cella & Cesa-Bianchi (2020); Schlisselberg et al. (2025). Our setting differs from standard delayed bandits in both the reward and feedback models. Rewards arise from bilateral trade and depend on the posted price and two latent valuations, rather than fixed arm distributions. Moreover, feedback is censored and acceptance-only: missing signals may indicate either rejection or delayed acceptance, rather than simply pending rewards.

2 Model

We consider online bilateral trade over a time horizon of T rounds. At each round t, a new seller–buyer pair arrives with private valuations $S_t, B_t \in [0,1]$, respectively. The seller's valuation S_t represents the minimum price she is willing to accept, whereas the buyer's valuation B_t represents the maximum price he is willing to pay. Upon the arrival of the pair, the learner posts a price $P_t \in [0,1]$. A trade occurs if and only if both agents accept the posted price, i.e., $S_t \le P_t \le B_t$.

Delayed Acceptance-Only Feedback. The learner does not observe the private valuations S_t and B_t. Instead, the seller and the buyer only send a signal when they accept the posted price, and such acceptance signals may be delayed.

We first define the underlying acceptance decisions at round t as

$$A_t^b = \mathbb{I}\{P_t \le B_t\}, \qquad A_t^s = \mathbb{I}\{P_t \ge S_t\},$$

where A_t^b and A_t^s denote the buyer-side and seller-side acceptance indicators, respectively.

Let D_t^b and D_t^s denote the corresponding stochastic feedback delays. The seller-side and buyer-side delays may follow different distributions. We denote their marginal delay distributions by $\mathcal{D}^s$ and $\mathcal{D}^b$, respectively. For any $s \le t$, the acceptance feedback available to the learner at time t for the pair arriving at round s is

$$A_{s,t}^b = \mathbb{I}\{P_s \le B_s\}\mathbb{I}\{D_s^b \le t-s\}, \qquad A_{s,t}^s = \mathbb{I}\{P_s \ge S_s\}\mathbb{I}\{D_s^s \le t-s\}.$$

Hence, $A^b_{s,t} = 1$ (respectively, $A^s_{s,t} = 1$) certifies that the buyer (respectively, seller) accepted the posted price and that the corresponding acceptance signal has already arrived at time step t. In contrast, an observation equal to zero is ambiguous: it may indicate either rejection or an acceptance signal that has not yet arrived.

Gain from Trade. The learner's performance at round t is measured by the gain from trade (GFT),

$$\mathit{GFT}_t(P_t) = \mathit{GFT}(P_t, S_t, B_t) = (B_t - S_t)\mathbb{I}\{S_t \le P_t \le B_t\}.$$

Thus, a successful trade generates the total surplus $B_t - S_t$, whereas the gain from trade is zero when no trade occurs. Importantly, the GFT itself is a performance criterion and is not directly observed by the learner; the learner only receives the delayed acceptance-only feedback defined above.

Regret. The learner chooses prices $P_1, \dots, P_T$ sequentially based only on the feedback available at the time of each decision. We compare the learner with the best fixed price in expectation. The regret is defined as

$$R_T = \max_{p\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathit{GFT}(p, S_t, B_t) - \sum_{t=1}^{T} \mathit{GFT}(P_t, S_t, B_t)\right].$$

Common Assumptions. The valuation pairs $\{(S_t, B_t)\}_{t=1}^T$ are i.i.d. across rounds, and S_t and B_t are independent within each round. We assume that the valuation distributions satisfy the standard M-Lipschitz condition:

$$|F_S(x) - F_S(y)| \le M|x - y|, \qquad |F_B(x) - F_B(y)| \le M|x - y|,$$

for all $x, y \in [0, 1]$. This condition is standard in online bilateral trade and is necessary for sublinear regret even without delays Cesa-Bianchi et al. (2024a); Bacchiocchi et al. (2025).

We assume that feedback delays are random variables with bounded support. Without loss of generality, delays are truncated at T, since feedback arriving after the horizon cannot affect the learner's decisions or regret.

3 Lower Bound

In this section, we establish a minimax regret lower bound for online bilateral trade under stochastic delays. Let $\mathcal{P}(M, \mu)$ denote the class of online bilateral trade instances satisfying the valuation assumptions with Lipschitz constant M, where the seller and buyer delays satisfy $\mathbb{E}[\mathcal{D}^s] \le \mu$ and $E[\mathcal{D}^b] \le \mu$. The minimax regret over this class is defined as $\mathfrak{R}_T(\mathcal{P}) = \inf_\pi \sup_{P\in\mathcal{P}(M,\mu)} \mathbb{E}_P[R_T(\pi)]$.

Even without delayed feedback, stochastic online bilateral trade requires regret of order $\Omega(T^{2/3})$ Cesa-Bianchi et al. (2024a). We next show that stochastic delays introduce an additional unavoidable cost depending on the expected delay.

For the delay-dependent lower bound, we construct a hard instance in the class $\mathcal{P}(M, \mu)$. Consider deterministic delays $D^b_t = D^s_t = d$ for all rounds, where $d = \lfloor \mu \rfloor$. This instance satisfies the delay constraint since $\mathbb{E}[\mathcal{D}^b] = \mathbb{E}[\mathcal{D}^s] = d \le \mu$. Under this construction, no acceptance feedback is revealed during the first d rounds, and therefore the learner receives no informative observation that can distinguish between different valuation instances.

We construct two valuation instances with separated supports such that every posted price is suboptimal for at least one of the two instances. Consequently, during the first d rounds, the learner must choose prices without knowing which instance is realized. By the separation of the valuation supports, there exists one instance under which the learner's posted prices yield zero gain from trade for at least a constant fraction of these rounds. Hence, the learner incurs regret of order $\Omega(d) = \Omega(\mu)$.

Theorem 1. *For any fixed Lipschitz constant $M > 0$ and delay budgets μ, the minimax regret of online bilateral trade under stochastic delays satisfies*

$$\mathfrak{R}_T(\mathcal{P}(M, \mu)) \ge c \max\left\{T^{2/3}, \mu\right\},$$

where $c > 0$ is a universal constant.

Algorithm 1 BOUNDARY-CALIBRATED SCOUTING BANDITS

Require: Horizon T, scouting length T_0, grid size K, calibration sizes m_0, m

Initialize $q_k \leftarrow k/K$ for $k \in [K]$, $J \leftarrow \lceil \log_2 T \rceil$, and $L_j \leftarrow 2^j$

◁ **Stage I: Adaptive window selection**

Collect m_0 buyer-side boundary samples by posting price 0, and m_0 seller-side boundary samples by posting price 1.

for $j = 0, 1, \ldots, J$ **do**

  Set $L_j \leftarrow 2^j$.

  Wait until all boundary samples reach age L_j and compute $\widehat{p}_j^b$ and $\widehat{p}_j^s$.

  if $\widehat{p}_j^b \geq 2/3$ and $\widehat{p}_j^s \geq 2/3$ **then**

    Set $L \leftarrow L_j$ and stop.

  end if

end for

◁ **Stage II: Boundary-price delay calibration**

Post price 0 and 1 for m rounds each and compute $\widetilde{p}_L^b$ and $\widetilde{p}_L^s$.

◁ **Stage III: Fixed-window scouting exploration**

for the next T_0 rounds **do**

  Draw $U_t \sim \mathrm{Unif}[0, 1]$ and post price U_t.

  Observe fixed-window feedback $A_{t,t+L}^b, A_{t,t+L}^s$.

end for

Wait L additional rounds and compute $\widehat{F}_k$ and $\widehat{G}_k$ for all $k \in [K]$.

◁ **Stage IV: Fixed-window bandit learning**

Run Bandit Elimination Algorithm 4 (see Appendix) using arms $\{q_k\}_{k=1}^K$, estimates $(\widehat{F}_k, \widehat{G}_k)$, and calibration estimates $(\widetilde{p}_L^b, \widetilde{p}_L^s)$.

4 ONLINE BILATERAL TRADE WITH EXOGENOUS STOCHASTIC DELAYS

We first consider *exogenous stochastic delays*, where the timing of an acceptance signal is independent of the underlying private valuations. Specifically, we assume $(D_t^s, D_t^b) \perp (S_t, B_t)$, and that the delay variables are independent across rounds. The learner does not know the delay distribution in advance.

Even under exogenous delays, the acceptance-only feedback creates a fundamental ambiguity: when no signal has been observed, the learner cannot distinguish between a rejection and an acceptance whose signal has not yet arrived. Consequently, directly treating missing signals as rejections leads to systematically biased estimates.

The exogeneity assumption, however, provides a useful factorization. For any price p and waiting time L, $\Pr(B \geq p, \mathcal{D}^b \leq L) = \Pr(B \geq p) \Pr(\mathcal{D}^b \leq L)$, with an analogous identity on the seller side. Therefore, once the arrival probabilities $p_L^b = \Pr(\mathcal{D}^b \leq L)$ and $p_L^s = \Pr(\mathcal{D}^s \leq L)$ are estimated, the bias caused by delayed acceptance signals can be corrected.

A key structural feature of bilateral trade allows us to estimate these quantities without prior knowledge of the delay distribution. Posting price 0 guarantees buyer acceptance, while posting price 1 guarantees seller acceptance. Hence, boundary-price rounds provide direct observations of whether the corresponding acceptance signals arrive within a prescribed window. We exploit these observations to calibrate the delay process and combine them with a scouting-bandit procedure for learning the optimal trading price.

The algorithm proceeds in four stages. It discretizes the price interval $[0, 1]$ into K grid points. First, it performs a doubling search over candidate windows and selects a suitable fixed observation window L. Second, it uses fresh boundary-price samples to estimate p_L^b and p_L^s. Third, it performs fixed-window scouting exploration and constructs estimates of the gain-from-trade components. Finally, it runs an elimination-based bandit algorithm over the discretized prices using fixed-window observations.

Stage I: Adaptive window selection The learner first performs a doubling search over the candidate windows $\{L_j = 2^j\}_{j=0}^{\lceil \log_2 T \rceil}$ to select an appropriate fixed observation window. The chosen window

should be large enough so that a constant fraction of delayed feedback signals arrive within the window, but not so large that the learner incurs unnecessary waiting time.

For the adaptive window selection, the learner first collects m_0 buyer-side and m_0 seller-side boundary samples by repeatedly posting prices 0 and 1, respectively. These samples are reused to evaluate all candidate windows $L_j = 2^j$. Specifically, for each candidate window L_j, the corresponding observations are $X^b_{\ell,j} = \mathbb{1}\{D^b_\ell \le L_j\}$ and $X^s_{\ell,j} = \mathbb{1}\{D^s_\ell \le L_j\}$ for $\ell = 1, \dots, m_0$.

The corresponding empirical arrival probabilities are

$$\widehat{p}^b_j = \frac{1}{m_0}\sum_{\ell=1}^{m_0} X^b_{\ell,j}, \qquad \widehat{p}^s_j = \frac{1}{m_0}\sum_{\ell=1}^{m_0} X^s_{\ell,j}.$$

The learner selects the first window satisfying $\widehat{p}^b_j \ge 2/3$ and $\widehat{p}^s_j \ge 2/3$.

The following lemma shows that the selected window is sufficiently large to control censoring while remaining comparable to the expected delays.

Lemma 1. *Suppose that $m_0 \ge C_0 \log((\lceil \log_2 T \rceil + 1)T^2)$ for a sufficiently large universal constant C_0. Then, with probability at least $1 - T^{-2}$, the window L selected in Algorithm 1 satisfies*

$$p^b_L \ge \frac{1}{2}, \qquad p^s_L \ge \frac{1}{2},$$

and

$$L \le C \max\{1, \mathbb{E}[\mathcal{D}^b], \mathbb{E}[\mathcal{D}^s]\},$$

where $C > 0$ is a universal constant.

Stage II: Boundary-price delay calibration. After selecting the fixed observation window L, the learner collects $2m$ fresh boundary samples to estimate the corresponding arrival probabilities:

$$\widehat{p}^b_L = \frac{1}{m}\sum_{i=1}^{m} X^b_i, \qquad \widehat{p}^s_L = \frac{1}{m}\sum_{i=1}^{m} X^s_i.$$

For stability, define $\widetilde{p}^b_L = \max\{\widehat{p}^b_L, 1/2\}$ and $\widetilde{p}^s_L = \max\{\widehat{p}^s_L, 1/2\}$.

The following lemma shows that this calibration step provides a nearly accurate estimate of the fixed-window arrival probability.

Lemma 2. *Conditioned on the event in Lemma 1, the calibration estimators satisfy*

$$\mathbb{E}[|\widetilde{p}^b_L - p^b_L|] \le O(m^{-1/2}), \qquad \mathbb{E}[|\widetilde{p}^s_L - p^s_L|] \le O(m^{-1/2}).$$

Stage III: Fixed-window scouting exploration. Given the estimate of the fixed-window arrival probability obtained in Stage II, the learner next performs scouting exploration. In this phase, the learner uses fixed-window observations, together with the calibrated arrival probability, to estimate the gain-from-trade function at the discretized prices.

For grid prices $q_k = k/K$, define

$$F_k = \Pr(q_k \le U \le B), \qquad G_k = \Pr(q_k \ge U \ge S),$$

where $U \sim \mathrm{Unif}[0,1]$.

Using the decomposition in Cesa-Bianchi et al. (2024a), the expected gain from trade associated with price q_k can be expressed in terms of F_k and G_k as

$$\mathbb{E}[\mathit{GFT}(q_k, S, B)] = F_k \Pr(q_k \ge S) + G_k \Pr(q_k \le B). \tag{1}$$

The learner samples random prices U_t and observes delayed acceptance signals within window L:

$$A^b_{t,t+L} = \mathbb{1}\{U_t \le B_t\}\mathbb{1}\{D^b_t \le L\}, \qquad A^s_{t,t+L} = \mathbb{1}\{U_t \ge S_t\}\mathbb{1}\{D^s_t \le L\}.$$

The scouting estimators are

$$\widehat{F}_k = \Pi_{[0,1]}\left(\frac{1}{T_0\widetilde{p}_L^b}\sum_t \mathbb{1}\{q_k \le U_t\}A_{t,t+L}^b\right), \qquad \widehat{G}_k = \Pi_{[0,1]}\left(\frac{1}{T_0\widetilde{p}_L^s}\sum_t \mathbb{1}\{q_k \ge U_t\}A_{t,t+L}^s\right),$$

where $\Pi_{[0,1]}$ denotes projection onto $[0,1]$.

The following lemma establishes the accuracy of the scouting estimators.

Lemma 3. *Conditioned on the event in Lemma 1, the scouting estimators satisfy*

$$\mathbb{E}\left[\max_k |\widehat{F}_k - F_k|\right] \le O\left(\sqrt{\frac{\log K}{T_0}} + \frac{1}{\sqrt{m}}\right),$$

and the same bound holds for $\widehat{G}_k$.

Stage IV: Fixed-window elimination bandit. The learner treats the discretized prices as arms and runs the elimination bandit procedure using the fixed-window observations. The surrogate utility of price q_k is constructed from $\widehat{F}_k$ and $\widehat{G}_k$, while the delayed feedback is corrected using $\widetilde{p}_L^b$ and $\widetilde{p}_L^s$ (details are in Appendix).

The following lemma gives the regret contribution of this bandit phase.

Lemma 4. *Conditioned on the event in Lemma 1 and the selected window L, the fixed-window bandit procedure satisfies*

$$R_{\text{bandit}} = O\left(\sqrt{KT\log T} + L\log K\right).$$

Combining the above lemmas gives the following result.

Theorem 2. *For exogenous stochastic delays with unknown delay distributions, Algorithm 1 satisfies*

$$R_T = O\left(T^{2/3}\sqrt{\log T} + (\mathbb{E}[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s])\log T\right).$$

5 Online Bilateral Trade with Valuation-Dependent Stochastic Delays

We now relax the exogeneity assumption and allow the feedback delays to depend on the agents' private valuations. Specifically, the tuples (S_t, B_t, D_t^s, D_t^b) are independent and identically distributed across rounds, with $D_t^s \sim Q_s(\cdot \mid S_t)$ and $D_t^b \sim Q_b(\cdot \mid B_t)$. Thus, the seller-side delay may depend on the seller's private valuation, while the buyer-side delay may depend on the buyer's private valuation. In particular, we no longer require $D_t^s \perp S_t$ or $D_t^b \perp B_t$.

This seemingly mild generalization fundamentally changes the learning problem. Under exogenous delays, the key correction relies on the factorization $\Pr(B \ge p, \mathcal{D}^b \le L) = \Pr(B \ge p)\Pr(\mathcal{D}^b \le L)$, and analogously on the seller side. Under valuation-dependent delays, this identity no longer holds in general.

Consequently, even perfect knowledge of the marginal delay distribution is insufficient to recover the true acceptance probability through a scalar correction. The observations that arrive within a short window may be systematically biased toward particular valuation regions.

Our algorithm avoids this difficulty by taking a different approach. Rather than attempting to debias every delayed observation, we choose, at each statistical resolution, a sufficiently long observation window so that the probability mass remaining beyond the window is negligible relative to the accuracy required at that stage.

We use a phased-elimination procedure. Let $\varepsilon_\ell = 2^{-\ell}$ denote the target accuracy in phase ℓ, and let $\mathcal{K}_\ell$ denote the set of active candidate prices at the beginning of the phase. Each phase consists of four main steps:

1. **Calibrate the delay window.** Use boundary prices to estimate delay tails and select the smallest geometric window L_ℓ with tail probability $O(\varepsilon_\ell)$.

Algorithm 2 BOUNDARY-CALIBRATED ADAPTIVE WINDOW ELIMINATION

Require: Horizon T, grid size K, confidence level δ

Initialize $q_k \leftarrow k/K$ for $k \in [K]$, $\mathcal{A}_0 \leftarrow [K]$, $\varepsilon_0 \leftarrow 1$, and $\ell \leftarrow 0$.
while $\varepsilon_\ell \geq \varepsilon_\star$ and time remains **do**
  $L_\ell \leftarrow$ BOUNDARYTAILWINDOW$(\varepsilon_\ell, \mathcal{A}_\ell)$.
  Set $m_\ell^{\mathrm{sc}} \leftarrow \lceil C_{\mathrm{sc}}\Lambda/\varepsilon_\ell^2 \rceil$ and $n_\ell \leftarrow \lceil C_{\mathrm{arm}}\Lambda/\varepsilon_\ell^2 \rceil$.
  ◁ **Scouting**
  for $r = 1, \ldots, m_\ell^{\mathrm{sc}}$ **do**
    Draw $U_r \sim \mathrm{Unif}[0,1]$ and post price $P_r = U_r$.
  end for
  ◁ **Active-price sampling**
  for $r = 1, \ldots, n_\ell$ **do**
    for each $k \in \mathcal{A}_\ell$ **do**
      Post price q_k once.
    end for
  end for
  Wait at most L_ℓ rounds, posting any price in $\mathcal{A}_\ell$, until all phase samples reach age L_ℓ.
  Compute $\widehat{F}_{k,\ell}$, $\widehat{G}_{k,\ell}$, $\widehat{\alpha}^b_{k,\ell}$, and $\widehat{\alpha}^s_{k,\ell}$ for all $k \in \mathcal{A}_\ell$, and set $\widehat{g}_{k,\ell} = \widehat{F}_{k,\ell}\widehat{\alpha}^s_{k,\ell} + \widehat{G}_{k,\ell}\widehat{\alpha}^b_{k,\ell}$.
  ◁ **Elimination**
  Set $\widehat{g}_\ell^{\max} \leftarrow \max_{j \in \mathcal{A}_\ell} \widehat{g}_{j,\ell}$.
  Update $\mathcal{A}_{\ell+1} \leftarrow \{k \in \mathcal{A}_\ell : \widehat{g}_{k,\ell} \geq \widehat{g}_\ell^{\max} - \varepsilon_\ell/2\}$.
  if $|\mathcal{A}_{\ell+1}| = 1$ **then**
    Play the remaining price for all remaining rounds and **return** .
  end if
  Set $\varepsilon_{\ell+1} \leftarrow \varepsilon_\ell/2$ and $\ell \leftarrow \ell + 1$.
end while
Play any price in the final active set for all remaining rounds.

2. **Estimate GFT components.** Use randomized scouting prices to estimate the truncated components F_{k,L_ℓ} and G_{k,L_ℓ} for all active prices.
3. **Estimate acceptance probabilities.** Sample each active price and estimate its truncated buyer- and seller-side acceptance probabilities.
4. **Eliminate suboptimal prices.** Combine the estimates into empirical GFT values and remove prices far below the empirical optimum.

As the algorithm progresses, ε_ℓ decreases geometrically. Consequently, the required censoring window becomes longer and the statistical estimates become more accurate. Early phases tolerate shorter windows and coarser estimates, whereas later phases use longer windows and more samples to distinguish prices with smaller GFT gaps. This multi-scale design avoids waiting for unnecessarily long delays before such accuracy is needed.

For convenience, let $\Lambda = \log\left(\frac{16KT^2}{\delta}\right)$, and $\varepsilon_\star = c_\star\sqrt{\frac{K\Lambda}{T}}$,, where $c_\star > 0$ is a sufficiently large universal constant.

Step 1: Adaptive boundary-tail calibration. We first describe how to select the phase-dependent observation window. For a window length L, define the buyer-side and seller-side delay tails as $\rho_b(L) = \Pr(\mathcal{D}^b > L)$ and $\rho_s(L) = \Pr(\mathcal{D}^s > L)$.

The key idea is to use boundary prices to directly estimate these tails. When posting price 0, the buyer always accepts, and thus $1 - A^b_{r,r+L} = \mathbb{I}\{D^b_r > L\}$ is a Bernoulli sample with mean $\rho_b(L)$. Similarly, posting price 1 gives $1 - A^s_{r,r+L} = \mathbb{I}\{D^s_r > L\}$ with mean $\rho_s(L)$.

At phase ℓ, the learner sets $\eta_\ell = \varepsilon_\ell/256$ and collects $m_\ell^{\mathrm{cal}} = \lceil C_{\mathrm{cal}}\Lambda/\eta_\ell^2 \rceil$ boundary samples from each side, denoted by $\mathcal{C}^b_\ell$ and $\mathcal{C}^s_\ell$. For each candidate window $L_j = 2^j$, the empirical tails are

$$\widehat{\rho}_b(L_j) = \frac{1}{m_\ell^{\mathrm{cal}}} \sum_{r \in \mathcal{C}^b_\ell} (1 - A^b_{r,r+L_j}), \qquad \widehat{\rho}_s(L_j) = \frac{1}{m_\ell^{\mathrm{cal}}} \sum_{r \in \mathcal{C}^s_\ell} (1 - A^s_{r,r+L_j}).$$

Algorithm 3 BOUNDARYTAILWINDOW$(\varepsilon, \mathcal{A})$

Require: target accuracy ε, active set $\mathcal{A}$

Set $\eta \leftarrow \varepsilon/256$ and $m^{\mathrm{cal}} \leftarrow \lceil C_{\mathrm{cal}}\Lambda/\eta^2 \rceil$.
Collect m^{cal} buyer-side boundary samples by posting price 0, and m^{cal} seller-side boundary samples by posting price 1.
for $j = 0, 1, \ldots, \lceil \log_2 T \rceil$ **do**
    Set $L_j \leftarrow \min\{2^j, T\}$.
    Wait until all calibration samples reach age L_j, while playing any price in $\mathcal{A}$.
    Compute $\widehat{\rho}_b(L_j)$ and $\widehat{\rho}_s(L_j)$.
    if $\widehat{\rho}_b(L_j) \le 2\eta$ and $\widehat{\rho}_s(L_j) \le 2\eta$ **then**
        return L_j.
    end if
end for

The learner selects the smallest window satisfying $\widehat{\rho}_b(L_j) \le 2\eta_\ell$ and $\widehat{\rho}_s(L_j) \le 2\eta_\ell$.

The following lemma shows that the selected window controls the remaining delayed-feedback probability while remaining bounded by the expected delays.

Lemma 5 (Boundary calibration)**.** *For a sufficiently large universal constant C_{cal}, with probability at least $1 - \delta/T$, simultaneously over all completed phases, every window L_ℓ returned by Algorithm 3 satisfies*

$$\rho_b(L_\ell) \le \frac{\varepsilon_\ell}{64}, \qquad \rho_s(L_\ell) \le \frac{\varepsilon_\ell}{64}.$$

Moreover,

$$L_\ell \le C\left(1 + \frac{E[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s]}{\varepsilon_\ell}\right)$$

for a universal constant C.

Step 2: Fixed-window scouting. Once L_ℓ is selected, the learner estimates the components of the GFT decomposition. For a grid price q_k, let g_k denote its expected gain from trade. Recall that

$$g_k = \mathbb{E}[\mathit{GFT}(q_k, S, B)] = F_k\alpha_k^s + G_k\alpha_k^b,$$

where $F_k = \Pr(q_k \le U \le B)$, $G_k = \Pr(q_k \ge U \ge S)$, $\alpha_k^b = \Pr(B \ge q_k)$, and $\alpha_k^s = \Pr(S \le q_k)$.

Under window L_ℓ, define the truncated scouting quantities

$$F_{k,L_\ell} = \Pr(q_k \le U \le B, \mathcal{D}^b \le L_\ell), \qquad G_{k,L_\ell} = \Pr(q_k \ge U \ge S, \mathcal{D}^s \le L_\ell).$$

The learner collects $m_\ell^{\mathrm{sc}} = \lceil C_{\mathrm{sc}}\Lambda/\varepsilon_\ell^2 \rceil$ fresh scouting samples over rounds $\mathcal{S}_\ell$. For each $r \in \mathcal{S}_\ell$, it draws $U_r \sim \mathrm{Unif}[0,1]$ and posts $P_r = U_r$. The estimators are

$$\widehat{F}_{k,\ell} = \frac{1}{m_\ell^{\mathrm{sc}}} \sum_{r \in \mathcal{S}_\ell} \mathbb{I}\{q_k \le U_r\} A^b_{r,r+L_\ell}, \qquad \widehat{G}_{k,\ell} = \frac{1}{m_\ell^{\mathrm{sc}}} \sum_{r \in \mathcal{S}_\ell} \mathbb{I}\{q_k \ge U_r\} A^s_{r,r+L_\ell}.$$

Unlike the exogenous-delay algorithm, these estimators are not rescaled by an estimated arrival probability. Instead, the bias is directly controlled through the calibrated tail bound from Lemma 5.

Lemma 6 (Scouting accuracy)**.** *Condition on the event in Lemma 5. For sufficiently large C_{sc}, with probability at least $1 - \delta/T$, simultaneously for all completed phases and all grid prices,*

$$|\widehat{F}_{k,\ell} - F_k| \le \frac{\varepsilon_\ell}{32}, \qquad |\widehat{G}_{k,\ell} - G_k| \le \frac{\varepsilon_\ell}{32}.$$

Step 3: Price-specific acceptance estimation. The scouting step provides the common GFT components, but evaluating a candidate price q_k additionally requires its seller-side and buyer-side acceptance probabilities.

Under the selected window, define $\alpha^b_{k,L_\ell} = \Pr(B \geq q_k, \mathcal{D}^b \leq L_\ell)$ and $\alpha^s_{k,L_\ell} = \Pr(S \leq q_k, \mathcal{D}^s \leq L_\ell)$. The learner pulls every active price q_k, $k \in \mathcal{A}_\ell$, for $n_\ell = \lceil C_{\mathrm{arm}} \Lambda / \varepsilon_\ell^2 \rceil$ fresh rounds. Let $\mathcal{T}_{k,\ell}$ denote the corresponding set of rounds. The empirical acceptance probabilities are

$$\widehat{\alpha}^b_{k,\ell} = \frac{1}{n_\ell} \sum_{r \in \mathcal{T}_{k,\ell}} A^b_{r,r+L_\ell}, \qquad \widehat{\alpha}^s_{k,\ell} = \frac{1}{n_\ell} \sum_{r \in \mathcal{T}_{k,\ell}} A^s_{r,r+L_\ell}.$$

The following lemma establishes the accuracy of the acceptance probability estimates.

Lemma 7 (Acceptance-probability accuracy)**.** *Condition on the event in Lemma 5. For sufficiently large C_{arm}, with probability at least $1 - \delta/T$, simultaneously for every completed phase and every active price,*

$$|\widehat{\alpha}^b_{k,\ell} - \alpha^b_k| \leq \frac{\varepsilon_\ell}{32}, \qquad |\widehat{\alpha}^s_{k,\ell} - \alpha^s_k| \leq \frac{\varepsilon_\ell}{32}.$$

Step 4: GFT estimation and elimination. The learner combines the scouting and acceptance estimates to obtain an empirical estimate of the GFT at each active price:

$$\widehat{g}_{k,\ell} = \widehat{F}_{k,\ell} \widehat{\alpha}^s_{k,\ell} + \widehat{G}_{k,\ell} \widehat{\alpha}^b_{k,\ell}.$$

By Lemmas 6 and 7, all components involved in the above decomposition can be estimated accurately up to the stage accuracy ε_ℓ. Since these quantities lie in $[0, 1]$, the component-wise estimation guarantees directly imply a uniform error bound for the estimated GFT.

Let $\widehat{g}^{\max}_\ell = \max_{j \in \mathcal{A}_\ell} \widehat{g}_{j,\ell}$. The learner then eliminates prices whose empirical GFT is sufficiently below the current best estimate and keeps the active set

$$\mathcal{A}_{\ell+1} = \left\{ k \in \mathcal{A}_\ell : \widehat{g}_{k,\ell} \geq \widehat{g}^{\max}_\ell - \frac{\varepsilon_\ell}{2} \right\}.$$

The following lemma establishes that this elimination rule preserves the optimal price while removing all prices whose GFT is separated from the optimum by more than the current accuracy level.

Lemma 8 (Safe elimination)**.** *On the intersection of the events in Lemmas 6 and 7,the optimal discretized price is never eliminated. Moreover, every price k surviving phase ℓ satisfies*

$$g^\star - g_k \leq \frac{3}{4} \varepsilon_\ell,$$

where $g^\star = \max_{k \in [K]} g_k$ is the optimal GFT among the discretized prices.

We are now ready to state the regret guarantee of Algorithm 2.

Theorem 3. *For sufficiently large universal constants C_{cal}, C_{sc}, C_{arm}, and $c_\star$, Algorithm 2 satisfies*

$$R_T = O\Big(T^{2/3} \sqrt{\log T} + (\mathbb{E}[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s]) \log T \Big).$$

6 Conclusion and Discussion

We study online bilateral trade under stochastic delayed and censored acceptance-only feedback. Our results show that delayed feedback introduces an additional unavoidable cost proportional to the expected delay beyond the intrinsic $T^{2/3}$ learning difficulty of stochastic bilateral trade. We first consider exogenous stochastic delays, where delays are independent of private valuations, and develop boundary-calibrated algorithms achieving near-optimal regret guarantees up to logarithmic factors. We further extend our results to valuation-dependent delays, where independence-based correction fails, and show that adaptive window elimination achieves the same regret order without explicitly learning the valuation–delay dependence.

Future Directions. Several directions remain open. First, extending our framework to non-stationary or adversarial valuation sequences would provide a more realistic model for dynamic marketplaces. Second, it remains interesting to study more general delay processes, such as adversarial or history-dependent delays, and understand whether similar boundary-based calibration techniques can achieve near-optimal regret.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret research results. The tasks related to dataset generation, experimental design, experimental implementation, and experimental analysis are not applicable to this work, as this paper does not contain an experimental component.

We have reviewed all AI-assisted work and verified the correctness and accuracy of the final manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yossi Azar, Amos Fiat, and Federico Fusco. An α-regret analysis of adversarial bilateral trade. *Advances in Neural Information Processing Systems*, 35:1685–1697, 2022.

Francesco Bacchiocchi, Matteo Castiglioni, Roberto Colomboni, and Alberto Marchesi. Online bilateral trade with minimal feedback: Don't waste seller's time. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

François Bachoc, Nicolò Cesa-Bianchi, Tom Cesari, and Roberto Colomboni. Fair online bilateral trade. *Advances in Neural Information Processing Systems*, 37:37241–37263, 2024a.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A contextual online learning theory of brokerage. *arXiv preprint arXiv:2407.01566*, 2024b.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A tight regret analysis of non-parametric repeated contextual brokerage. In *International Conference on Artificial Intelligence and Statistics*, pp. 2836–2844. PMLR, 2025.

François Bachoc, Roberto Colomboni, and Emilie Kaufmann. Repeated bilateral trade: The quest for fairness. *arXiv preprint arXiv:2606.15369*, 2026.

Matthew Backus, Thomas Blake, Brad Larsen, and Steven Tadelis. Sequential bargaining in the field: Evidence from millions of online bargaining interactions. *The Quarterly Journal of Economics*, 135(3):1319–1361, 2020.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. No-regret learning in bilateral trade via global budget balance. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 247–258, 2024.

Natasa Bolić, Tommaso Cesari, and Roberto Colomboni. An online learning theory of brokerage. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 216–224, 2024.

Leonardo Cella and Nicolo Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1168–1177. PMLR, 2020.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Bilateral trade: A regret minimization perspective. *Mathematics of Operations Research*, 49(1): 171–203, 2024a.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Regret analysis of bilateral trade with a smoothed adversary. *Journal of Machine Learning Research*, 25(234):1–36, 2024b.

Tommaso Cesari and Roberto Colomboni. An online learning theory of trading-volume maximization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Houshuang Chen, Yaonan Jin, Pinyan Lu, and Chihao Zhang. Tight regret bounds for fixed-price bilateral trade. *arXiv preprint arXiv:2504.04349*, 2025.

Romain Cosson, Federico Fusco, Anupam Gupta, Stefano Leonardi, Renato Paes Leme, and Matteo Russo. Contextual online bilateral trade. *arXiv preprint arXiv:2602.12903*, 2026.

Simone Di Gregorio, Paul Dütting, Federico Fusco, and Chris Schwiegelshohn. Nearly tight regret bounds for profit maximization in bilateral trade. *arXiv preprint arXiv:2509.22563*, 2025.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.

Jessica Fong and Caio Waisman. The effects of delay in bargaining: Evidence from ebay. *Management Science*, 2025.

Solenne Gaucher, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Vianney Perchet. Feature-based online bilateral trade. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pp. 3988–3997. PMLR, 2021.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *International Conference on Artificial Intelligence and Statistics*, pp. 6095–6119. PMLR, 2023.

Yaonan Jin. Tight regret bounds for bilateral trade under semi feedback. *arXiv preprint arXiv:2601.16412*, 2026.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pp. 5969–5978. PMLR, 2021.

Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. Online two-sided markets: Many buyers enhance learning. *arXiv preprint arXiv:2503.01529*, 2025.

Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. Better regret rates in bilateral trade via sublinear budget violation. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 6494–6536. SIAM, 2026a.

Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. Regret minimization in bilateral trade with perturbed markets. *arXiv preprint arXiv:2605.10475*, 2026b.

Anna Lunghi, Mattia Piccinato, Matteo Castiglioni, and Alberto Marchesi. A stronger benchmark for online bilateral trade: From fixed prices to distributions. *arXiv preprint arXiv:2602.05681*, 2026c.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Ofir Schlisselberg, Ido Cohen, Tal Lancewicki, and Yishay Mansour. Delay as payoff in mab. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20310–20317, 2025.

Yifu Tang, Yingfei Wang, and Zeyu Zheng. Stochastic multi-armed bandits with strongly reward-dependent delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3043–3051. PMLR, 2024.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Dirk Van Der Hoeven and Nicolo Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pp. 9712–9721. PMLR, 2020.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3285–3294. PMLR, 2020.

Algorithm 4 FIXED-WINDOW ELIMINATION BANDIT

Require: Horizon T, window length L, active set $\mathcal{K} \leftarrow [K]$, estimates $\{(\widehat{F}_k, \widehat{G}_k)\}_{k=1}^{K}$, calibrated arrival probabilities $\widetilde{p}_L^b, \widetilde{p}_L^s$

for epoch $i = 0, 1, 2, \ldots$ **do**
  ◁ **Round-robin exploration**
  for $r = 1, \ldots, |\mathcal{K}|2^i$ **do**
    Select the $(r \bmod |\mathcal{K}|)$-th arm I_t in $\mathcal{K}$ and post price q_{I_t}.
    Observe fixed-window reward $Y_t = \widehat{F}_{I_t} S_t^{(L)} / \widetilde{p}_L^s + \widehat{G}_{I_t} B_t^{(L)} / \widetilde{p}_L^b$.
  end for
  ◁ **Confidence estimation**
  for each $k \in \mathcal{K}$ **do**
    Compute arrived observations $n_k(t) = \sum_{s=1}^{t-L} \mathbb{1}\{I_s = k\}$.
    Compute empirical mean $\widehat{v}_k(t)$ and confidence bounds $LCB_k(t)$ and $UCB_k(t)$.
  end for
  ◁ **Elimination**
  Let $L^\star(t) = \max_{j \in \mathcal{K}} LCB_j(t)$.
  Remove all arms $k \in \mathcal{K}$ satisfying $UCB_k(t) \le L^\star(t)$.
end for

A ELIMINATION BANDIT

After the exploration phase, each grid price q_k is treated as an arm in a K-armed bandit problem. The learner uses the scouting estimates $\widehat{F}_k$ and $\widehat{G}_k$ to construct a surrogate utility for each arm:

$$u_k = \widehat{F}_k \Pr(S \le q_k) + \widehat{G}_k \Pr(B \ge q_k).$$

The goal of this stage is to identify an arm with the largest surrogate utility.

When the learner selects arm k at round t, it posts price q_k. Similar to the exploration phase, the learner uses fixed-window feedback to handle delayed acceptance signals. The two feedback components are corrected separately using the corresponding calibrated arrival probabilities. The fixed-window reward observation is defined as

$$Y_t = \widehat{F}_k \frac{S_t^{(L)}}{\widetilde{p}_L^s} + \widehat{G}_k \frac{B_t^{(L)}}{\widetilde{p}_L^b}.$$

The learner then runs the elimination-based bandit procedure described in Algorithm 4. The algorithm proceeds in epochs with geometrically increasing lengths. In each epoch, the active arms are explored in a round-robin manner using fixed-window observations. Arms whose upper confidence bounds are no larger than the largest lower confidence bound among active arms are eliminated.

Let $n_k(t)$ denote the number of fixed-window observations of arm k that have arrived by time t, i.e.,

$$n_k(t) = \sum_{s=1}^{t-L} \mathbb{1}\{I_s = k\}.$$

The empirical mean is given by

$$\widehat{v}_k(t) = \frac{1}{n_k(t)} \sum_{s=1}^{t-L} \mathbb{1}\{I_s = k\} Y_s,$$

and the confidence bounds are

$$LCB_k(t) = \widehat{v}_k(t) - 8\sqrt{\frac{\log T}{n_k(t)}}, \qquad UCB_k(t) = \widehat{v}_k(t) + 8\sqrt{\frac{\log T}{n_k(t)}}. \tag{2}$$

B Omitted Proofs

B.1 Proof of Theorem 1

Proof. First, by Theorem 4 of Cesa-Bianchi et al. (2024a), the minimax regret satisfies

$$R \geq cT^{2/3}, \tag{3}$$

where $c \geq 11/672$.

We next establish the delay-dependent lower bound. It suffices to consider the case $\mu \geq 1$, since otherwise the $\Omega(T^{2/3})$ lower bound already implies the desired result. Let $d = \lfloor \mu \rfloor$. We construct an instance in the class $\mathcal{P}(M, \mu)$ with deterministic delays

$$D_t^b = D_t^s = d$$

for all rounds t. This instance satisfies the delay constraint since

$$\mathbb{E}[\mathcal{D}^b] = \mathbb{E}[\mathcal{D}^s] = d \leq \mu.$$

We construct two valuation instances with separated supports.

In Case 1, the seller and buyer valuation densities are

$$f_S^{(1)}(x) = 8\mathbb{1}\left\{x \in [0, \tfrac{1}{8}]\right\}, \tag{4}$$

and

$$f_B^{(1)}(y) = 8\mathbb{1}\left\{y \in [\tfrac{1}{8}, \tfrac{1}{4}]\right\}. \tag{5}$$

The optimal price is $1/8$, which achieves

$$\mathbb{E}[\mathit{GFT}] = \frac{1}{8}.$$

Moreover, any price larger than $1/4$ yields zero gain from trade.

In Case 2, the seller and buyer valuation densities are

$$f_S^{(2)}(x) = 8\mathbb{1}\left\{x \in [\tfrac{3}{4}, \tfrac{7}{8}]\right\}, \tag{6}$$

and

$$f_B^{(2)}(y) = 8\mathbb{1}\left\{y \in [\tfrac{7}{8}, 1]\right\}. \tag{7}$$

The optimal price is $7/8$, which achieves

$$\mathbb{E}[\mathit{GFT}] = \frac{1}{8}.$$

Moreover, any price smaller than $3/4$ yields zero gain from trade.

During the first d rounds, all buyer-side and seller-side acceptance feedback is delayed. Therefore, the learner receives no informative observation that can distinguish between the two valuation instances. Consequently, any adaptive pricing strategy must choose prices without knowing which instance is realized.

The two zero-gain regions, $[0, 3/4)$ in Case 2 and $(1/4, 1]$ in Case 1, cover the entire price space. Hence, during the first d rounds, the learner must post prices belonging to one of these two regions for at least $d/2$ rounds. In the corresponding instance, these prices yield zero gain from trade, while the optimal fixed price achieves expected gain from trade $1/8$.

Therefore, the regret incurred during the first d rounds is at least

$$\frac{d}{2} \cdot \frac{1}{8} = \frac{d}{16}. \tag{8}$$

Since $d = \lfloor \mu \rfloor \geq \mu/2$ for $\mu \geq 1$, we obtain

$$R \geq \frac{\mu}{32}. \tag{9}$$

Thus, the minimax regret is lower bounded by

$$\Omega(\mu).$$

Combining this delay-dependent lower bound with the $\Omega(T^{2/3})$ lower bound proves

$$R = \Omega\left(\max\{T^{2/3}, \mu\}\right).$$

□

B.2 Proof of Lemma 1

Proof. For each candidate window $L_j = 2^j$, define $p_j^b = \Pr(\mathcal{D}^b \le L_j)$ and $p_j^s = \Pr(\mathcal{D}^s \le L_j)$. The empirical estimates $\widehat{p}_j^b$ and $\widehat{p}_j^s$ are averages of independent Bernoulli random variables with means p_j^b and p_j^s.

Consider the event

$$\mathcal{E} = \left\{ |\widehat{p}_j^b - p_j^b| \le \frac{1}{12}, |\widehat{p}_j^s - p_j^s| \le \frac{1}{12},\ \forall j \right\}.$$

By Hoeffding's inequality, for any fixed candidate window,

$$\Pr\left(|\widehat{p}_j^b - p_j^b| > \frac{1}{12} \right) \le 2\exp(-m_0/72),$$

and the same bound holds for the seller side. Taking a union bound over the two sides and all $O(\log T)$ candidate windows, and choosing $m_0 \ge C_0 \log((\lceil \log_2 T \rceil + 1)T^2)$ for a sufficiently large constant C_0, gives

$$\Pr(\mathcal{E}^c) \le T^{-2}.$$

We next condition on $\mathcal{E}$. If Algorithm 1 selects $L = L_j$, then by the stopping criterion, $\widehat{p}_j^b \ge 2/3$ and $\widehat{p}_j^s \ge 2/3$. Therefore, $p_L^b \ge 2/3 - 1/12 = 7/12 \ge 1/2$, and similarly $p_L^s \ge 1/2$.

It remains to bound the size of the selected window. Let $\mu_b = \mathbb{E}[\mathcal{D}^b]$ and $\mu_s = \mathbb{E}[\mathcal{D}^s]$. By Markov's inequality, $\Pr(\mathcal{D}^b > L) \le \mu_b/L$ and $\Pr(\mathcal{D}^s > L) \le \mu_s/L$. Choose a dyadic window $L^\star$ satisfying

$$L^\star \ge 4\max\{1, \mu_b, \mu_s\}.$$

Then $p_{L^\star}^b \ge 3/4$ and $p_{L^\star}^s \ge 3/4$. Hence, on $\mathcal{E}$,

$$\widehat{p}_{L^\star}^b \ge 2/3, \qquad \widehat{p}_{L^\star}^s \ge 2/3.$$

Therefore, the doubling search stops no later than $L^\star$. Since $L^\star$ is the smallest dyadic number satisfying the above condition,

$$L \le 2L^\star \le 8\max\{1, \mu_b, \mu_s\}.$$

□

B.3 Proof of Lemma 2

Proof. We prove the buyer-side bound; the seller-side bound follows analogously.

Condition on the event in Lemma 1, under which $p_L^b \ge 1/2$. The estimator $\widehat{p}_L^b$ is the empirical mean of m independent Bernoulli random variables with mean p_L^b. The calibrated estimator is defined as

$$\widetilde{p}_L^b = \max\{\widehat{p}_L^b, 1/2\}.$$

Since $p_L^b \ge 1/2$, the clipping operation cannot increase the estimation error, and therefore

$$|\widetilde{p}_L^b - p_L^b| \le |\widehat{p}_L^b - p_L^b|.$$

Moreover,

$$\mathbb{E}\left[(\widehat{p}_L^b - p_L^b)^2\right] = \frac{p_L^b(1-p_L^b)}{m} \le \frac{1}{4m}.$$

By Jensen's inequality,

$$\mathbb{E}\left[|\widehat{p}_L^b - p_L^b|\right] \le \sqrt{\mathbb{E}\left[(\widehat{p}_L^b - p_L^b)^2\right]} \le \frac{1}{2\sqrt{m}}.$$

Hence,

$$\mathbb{E}\left[|\widehat{p}_L^b - p_L^b|\right] \le O(m^{-1/2}).$$

The same argument applies to the seller-side estimator, yielding

$$\mathbb{E}\left[|\widehat{p}_L^s - p_L^s|\right] \le O(m^{-1/2}).$$

□

B.4 Proof of Lemma 3

Proof. We prove the buyer-side bound; the seller-side bound follows analogously.

Condition on the event in Lemma 1, under which $p_L^b \ge 1/2$. Define the oracle estimator using the true arrival probability:

$$\widetilde{F}_k = \frac{1}{T_0 p_L^b} \sum_t \mathbb{I}\{q_k \le U_t\} A_{t,t+L}^b.$$

By the independence between valuations and delays,

$$\mathbb{E}[\mathbb{I}\{q_k \le U_t\} A_{t,t+L}^b] = p_L^b F_k.$$

Hence, $\widetilde{F}_k$ is an unbiased estimator of F_k.

Moreover, since $p_L^b \ge 1/2$, the summands are bounded by a universal constant. By Hoeffding's inequality and a union bound over all K grid prices,

$$\mathbb{E}\left[\max_k |\widetilde{F}_k - F_k|\right] \le O\left(\sqrt{\frac{\log K}{T_0}}\right).$$

It remains to bound the error caused by replacing p_L^b with $\widetilde{p}_L^b$. Let

$$\widehat{F}_k = \Pi_{[0,1]}\left(\frac{1}{T_0 \widetilde{p}_L^b} \sum_t \mathbb{I}\{q_k \le U_t\} A_{t,t+L}^b\right).$$

Since projection onto $[0,1]$ cannot increase the distance to $F_k \in [0,1]$, it suffices to consider the unprojected estimator. We have

$$\left|\frac{p_L^b}{\widetilde{p}_L^b}\widetilde{F}_k - F_k\right| \le |\widetilde{F}_k - F_k| + \left|\frac{p_L^b}{\widetilde{p}_L^b} - 1\right| |F_k|.$$

By Lemma 2,

$$\mathbb{E}\left[|\widetilde{p}_L^b - p_L^b|\right] \le O(m^{-1/2}).$$

Since $\widetilde{p}_L^b \ge 1/2$ and $F_k \in [0,1]$, the second term is bounded by $O(m^{-1/2})$. Therefore,

$$\mathbb{E}\left[\max_k |\widehat{F}_k - F_k|\right] \le O\left(\sqrt{\frac{\log K}{T_0}} + \frac{1}{\sqrt{m}}\right).$$

The seller-side bound follows from the same argument using p_L^s, $\widetilde{p}_L^s$, and $\widehat{G}_k$. □

B.5 Proof of Lemma 4

Proof. We condition on the event in Lemma 1 and the calibration guarantee in Lemma 2. We first characterize the expected fixed-window reward.

Under the exogeneity assumption,

$$\mathbb{E}[S_t^{(L)} \mid I_t = k] = p_L^s \Pr(S \le q_k),$$

and

$$\mathbb{E}[B_t^{(L)} \mid I_t = k] = p_L^b \Pr(B \ge q_k).$$

Therefore,

$$\mathbb{E}[Y_t \mid I_t = k] = \frac{p_L^s}{\widetilde{p}_L^s} \widehat{F}_k \Pr(S \le q_k) + \frac{p_L^b}{\widetilde{p}_L^b} \widehat{G}_k \Pr(B \ge q_k).$$

By Lemma 2, the calibration errors satisfy

$$\mathbb{E}[|\widetilde{p}_L^b - p_L^b|] = O(m^{-1/2}), \qquad \mathbb{E}[|\widetilde{p}_L^s - p_L^s|] = O(m^{-1/2}).$$

Since Lemma 1 guarantees that $p_L^b, p_L^s \ge 1/2$, the correction ratios $p_L^b/\widetilde{p}_L^b$ and $p_L^s/\widetilde{p}_L^s$ are bounded by universal constants. Hence the fixed-window rewards are bounded random variables, and their expectation differs from the ideal surrogate utility only by the calibration error.

We next analyze the elimination procedure. Let

$$v_k = \mathbb{E}[Y_t \mid I_t = k]$$

denote the expected fixed-window reward of arm k. Since the calibration error only introduces a lower-order perturbation, it suffices to analyze the stochastic bandit problem with arm values $\{v_k\}_{k=1}^K$.

For every arm k, recall that $n_k(t)$ is the number of arrived fixed-window observations available at time t. Since the observations are bounded, Hoeffding's inequality gives

$$\Pr\left(|\widehat{v}_k(t) - v_k| > 8\sqrt{\frac{\log T}{n_k(t)}}\right) \le 2T^{-4}.$$

Taking a union bound over all arms and all time steps, with probability at least $1 - T^{-2}$, all confidence intervals are simultaneously valid.

On this event, the optimal arm is never eliminated. Indeed, if $k^\star$ denotes an optimal arm, then

$$LCB_{k^\star}(t) \le v_{k^\star} \le UCB_{k^\star}(t),$$

and a suboptimal arm cannot have an upper confidence bound smaller than the lower confidence bound of the optimal arm due to the validity of the confidence intervals.

Now consider a suboptimal arm k with gap

$$\Delta_k = v_{k^\star} - v_k.$$

For a sufficiently large constant C, once

$$n_k(t) \ge C\frac{\log T}{\Delta_k^2},$$

the confidence radius of arm k is at most $\Delta_k/4$. Therefore, its upper confidence bound is smaller than the lower confidence bound of the optimal arm, and the arm is eliminated. Hence, the total number of arrived observations of a suboptimal arm satisfies

$$n_k(T) = O\left(\frac{\log T}{\Delta_k^2}\right).$$

The resulting statistical regret is bounded by the standard elimination argument:

$$\sum_{k \ne k^\star} \Delta_k \min\left\{T, C\frac{\log T}{\Delta_k^2}\right\} = O(\sqrt{KT\log T}).$$

It remains to account for delayed observations. Because every fixed-window reward is revealed after L rounds, the learner may continue sampling an arm before its previous observations arrive. During an epoch with active set size $|\mathcal{K}_i|$, the number of pending observations for each arm is at most $O(L/|\mathcal{K}_i| + 1)$ under round-robin exploration. Summing this additional oversampling cost over all elimination epochs gives

$$O\left(L \sum_i \frac{n_i}{|\mathcal{K}_i|}\right) = O(L \log K),$$

where n_i denotes the number of arms eliminated in epoch i.

Combining all these above yields

$$R_{\text{bandit}} = O\left(\sqrt{KT \log T} + L \log K\right).$$

□

B.6 Proof of Theorem 2

Proof. Let

$$\mu_b = \mathbb{E}[\mathcal{D}^b], \qquad \mu_s = \mathbb{E}[\mathcal{D}^s].$$

We condition on the event where Lemmas 1, 2, 3, and 4 hold. We decompose the regret into the costs before and after the bandit phase.

Let $p^\star \in [0, 1]$ denote the optimal continuous price and let $q^\star$ be the closest grid price. The regret can be decomposed as

$$R_T \le R_{\text{pre}} + R_{\text{disc}} + R_{\text{scout}} + R_{\text{bandit}},$$

where R_{pre} denotes the regret incurred during window selection, calibration, and waiting periods, R_{disc} is the discretization error, R_{scout} is the error from estimating the GFT components, and R_{bandit} is the regret of the fixed-window elimination procedure.

We first bound the cost before the bandit phase. By Lemma 1, the selected window satisfies

$$L = O(1 + \max\{\mu_b, \mu_s\}).$$

The doubling search tests at most $O(\log T)$ candidate windows, and the waiting time incurred during the window selection and fixed-window exploration is proportional to the selected window. Therefore,

$$R_{\text{pre}} = O((\mu_b + \mu_s) \log T).$$

Next, consider the discretization error. By the Lipschitz assumption on the valuation distributions, the expected gain-from-trade function is Lipschitz. Hence, for a grid with spacing $1/K$,

$$R_{\text{disc}} \le T\left(g(p^\star) - g(q^\star)\right) = O(T/K).$$

We next bound the estimation error introduced by the scouting phase. By Lemma 3,

$$\mathbb{E}\left[\max_k |\widehat{F}_k - F_k|\right] = O\left(\sqrt{\frac{\log K}{T_0}} + \frac{1}{\sqrt{m}}\right),$$

and the same bound holds for $\widehat{G}_k$. Since the GFT decomposition is linear in these quantities and all coefficients are bounded by one, the resulting regret contribution is

$$R_{\text{scout}} = O\left(T\sqrt{\frac{\log K}{T_0}} + \frac{T}{\sqrt{m}}\right).$$

Finally, by Lemma 4, the fixed-window elimination procedure satisfies

$$R_{\text{bandit}} = O\left(\sqrt{KT \log T} + L \log K\right).$$

Using the bound on L from Lemma 1,

$$L \log K = O((\mu_b + \mu_s) \log T).$$

Combining the above bounds gives

$$R_T = O\left(\frac{T}{K} + T\sqrt{\frac{\log K}{T_0}} + \frac{T}{\sqrt{m}} + \sqrt{KT\log T} + (\mu_b + \mu_s)\log T\right).$$

Choosing

$$K = \Theta(T^{1/3}), \qquad T_0 = \Theta(T^{2/3}), \qquad m = \Theta(T^{2/3}),$$

we obtain

$$\frac{T}{K} = O(T^{2/3}),$$
$$T\sqrt{\frac{\log K}{T_0}} = O(T^{2/3}\sqrt{\log T}),$$
$$\frac{T}{\sqrt{m}} = O(T^{2/3}),$$

and

$$\sqrt{KT\log T} = O(T^{2/3}\sqrt{\log T}).$$

Therefore,

$$R_T = O\left(T^{2/3}\sqrt{\log T} + (\mu_b + \mu_s)\log T\right),$$

which proves the theorem. □

B.7 Proof of Lemma 5

Proof. We prove the two claims separately.

For phase ℓ, recall that

$$\eta_\ell = \frac{\varepsilon_\ell}{256}.$$

For any candidate window L_j, define the empirical tail estimators $\widehat{\rho}_b(L_j)$ and $\widehat{\rho}_s(L_j)$ as in Algorithm 3. Since the boundary prices guarantee acceptance from one side, the samples used in these estimators are independent Bernoulli random variables with means $\rho_b(L_j)$ and $\rho_s(L_j)$, respectively.

Consider the event

$$\mathcal{E}_\ell = \{|\widehat{\rho}_b(L_j) - \rho_b(L_j)| \le \eta_\ell,\ |\widehat{\rho}_s(L_j) - \rho_s(L_j)| \le \eta_\ell,\ \forall j\}.$$

By Hoeffding's inequality, for any fixed j,

$$\Pr\left(|\widehat{\rho}_b(L_j) - \rho_b(L_j)| > \eta_\ell\right) \le 2\exp(-2m_\ell^{\mathrm{cal}}\eta_\ell^2),$$

and the same bound holds for the seller side.

Choosing

$$m_\ell^{\mathrm{cal}} = \left\lceil C_{\mathrm{cal}}\frac{\Lambda}{\eta_\ell^2}\right\rceil$$

for a sufficiently large universal constant C_{cal}, a union bound over both sides and all candidate windows gives

$$\Pr(\mathcal{E}_\ell^c) \le \frac{\delta}{T^2}.$$

Taking another union bound over all completed phases, we obtain that with probability at least $1 - \delta/T$, the event $\mathcal{E}_\ell$ holds simultaneously for all phases.

We now condition on this event. Suppose that Algorithm 3 returns window L_ℓ. By the stopping rule,

$$\widehat{\rho}_b(L_\ell) \le 2\eta_\ell, \qquad \widehat{\rho}_s(L_\ell) \le 2\eta_\ell.$$

Therefore,

$$\rho_b(L_\ell) \le \widehat{\rho}_b(L_\ell) + \eta_\ell \le 3\eta_\ell = \frac{3}{256}\varepsilon_\ell \le \frac{\varepsilon_\ell}{64},$$

and similarly,

$$\rho_s(L_\ell) \le \frac{\varepsilon_\ell}{64}.$$

This proves the first claim.

It remains to bound the size of the selected window. By Markov's inequality, for any $L > 0$,

$$\rho_b(L) = \Pr(\mathcal{D}^b > L) \le \frac{\mu_b}{L},$$

and

$$\rho_s(L) = \Pr(\mathcal{D}^s > L) \le \frac{\mu_s}{L}.$$

Let

$$L^\star = \frac{128\mu_\Sigma}{\varepsilon_\ell},$$

where $\mu_\Sigma = \mu_b + \mu_s$. Then

$$\rho_b(L^\star) \le \frac{\mu_b}{L^\star} \le \frac{\varepsilon_\ell}{128},$$

and

$$\rho_s(L^\star) \le \frac{\mu_s}{L^\star} \le \frac{\varepsilon_\ell}{128}.$$

Hence, for sufficiently large constants, $L^\star$ satisfies the population tail requirement of the window search. Since Algorithm 3 searches over geometrically increasing windows and returns the first successful one, the selected window is at most a constant factor larger than $L^\star$. Therefore, there exists a universal constant C such that

$$L_\ell \le C\left(1 + \frac{\mu_\Sigma}{\varepsilon_\ell}\right).$$

Combining the two parts completes the proof. □

B.8 Proof of Lemma 6

Proof. We prove the buyer-side bound for $\widehat{F}_{k,\ell}$; the seller-side bound follows analogously.

Condition on the event in Lemma 5. Define the truncated scouting quantity

$$F_{k,L_\ell} = \Pr(q_k \le U \le B, \mathcal{D}^b \le L_\ell).$$

For each scouting sample r, the random variable

$$\mathbb{I}\{q_k \le U_r\} A^b_{r,r+L_\ell}$$

is a Bernoulli random variable with mean F_{k,L_ℓ}. Therefore, $\widehat{F}_{k,\ell}$ is an empirical average satisfying

$$\mathbb{E}[\widehat{F}_{k,\ell}] = F_{k,L_\ell}.$$

By Hoeffding's inequality, for any fixed grid price q_k,

$$\Pr\left(\left|\widehat{F}_{k,\ell} - F_{k,L_\ell}\right| > \frac{\varepsilon_\ell}{64}\right) \le 2\exp\left(-\frac{m^{\mathrm{sc}}_\ell \varepsilon_\ell^2}{2048}\right).$$

Choosing

$$m^{\mathrm{sc}}_\ell \ge C_{\mathrm{sc}} \frac{\Lambda}{\varepsilon_\ell^2}$$

for a sufficiently large universal constant C_{sc}, and applying a union bound over all grid prices and completed phases, gives

$$\left|\widehat{F}_{k,\ell} - F_{k,L_\ell}\right| \le \frac{\varepsilon_\ell}{64}$$

simultaneously with probability at least $1 - \delta/T$.

It remains to control the bias caused by truncating delayed feedback. Since

$$F_k = \Pr(q_k \le U \le B),$$

we have

$$0 \le F_k - F_{k,L_\ell} = \Pr(q_k \le U \le B, \mathcal{D}^b > L_\ell).$$

Therefore,

$$F_k - F_{k,L_\ell} \le \Pr(\mathcal{D}^b > L_\ell) = \rho_b(L_\ell).$$

By Lemma 5,

$$\rho_b(L_\ell) \le \frac{\varepsilon_\ell}{64}.$$

Combining the estimation error and truncation bias gives

$$\left|\widehat{F}_{k,\ell} - F_k\right| \le \left|\widehat{F}_{k,\ell} - F_{k,L_\ell}\right| + |F_{k,L_\ell} - F_k| \le \frac{\varepsilon_\ell}{32}.$$

The proof for the seller-side quantity is identical. □

B.9 Proof of Lemma 7

Proof. We prove the buyer-side bound; the seller-side bound follows analogously.

Condition on the event in Lemma 5. For an active price q_k, define the truncated buyer-side acceptance probability

$$\alpha^b_{k,L_\ell} = \Pr(B \ge q_k, \mathcal{D}^b \le L_\ell).$$

The estimator

$$\widehat{\alpha}^b_{k,\ell} = \frac{1}{n_\ell} \sum_{r \in \mathcal{T}_{k,\ell}} A^b_{r,r+L_\ell}$$

is an empirical average of independent Bernoulli random variables with mean α^b_{k,L_ℓ}.

By Hoeffding's inequality, for any fixed active price q_k,

$$\Pr\left(\left|\widehat{\alpha}^b_{k,\ell} - \alpha^b_{k,L_\ell}\right| > \frac{\varepsilon_\ell}{64}\right) \le 2\exp\left(-\frac{n_\ell \varepsilon_\ell^2}{2048}\right).$$

Choosing

$$n_\ell \ge C_{\mathrm{arm}} \frac{\Lambda}{\varepsilon_\ell^2}$$

for a sufficiently large universal constant C_{arm}, and applying a union bound over all active prices and completed phases, gives

$$\left|\widehat{\alpha}^b_{k,\ell} - \alpha^b_{k,L_\ell}\right| \le \frac{\varepsilon_\ell}{64}$$

simultaneously with probability at least $1 - \delta/T$.

It remains to control the bias introduced by the finite observation window. Since the only missing buyer-side acceptance signals are those with delay larger than L_ℓ, we have

$$0 \le \alpha^b_k - \alpha^b_{k,L_\ell} = \Pr(B \ge q_k, \mathcal{D}^b > L_\ell) \le \Pr(\mathcal{D}^b > L_\ell) = \rho_b(L_\ell).$$

By Lemma 5,

$$\rho_b(L_\ell) \le \frac{\varepsilon_\ell}{64}.$$

Therefore,

$$\left|\widehat{\alpha}^b_{k,\ell} - \alpha^b_k\right| \le \left|\widehat{\alpha}^b_{k,\ell} - \alpha^b_{k,L_\ell}\right| + \left|\alpha^b_{k,L_\ell} - \alpha^b_k\right| \le \frac{\varepsilon_\ell}{32}.$$

The seller-side bound follows from the same argument. □

B.10 Proof of Lemma 8

Proof. Let $k^\star$ denote an optimal discretized price, i.e.,

$$g^\star = g_{k^\star} = \max_{k\in[K]} g_k.$$

We first establish a uniform confidence bound for the empirical GFT estimates.

By Lemmas 6 and 7, on the considered event, for every active price $k \in \mathcal{A}_\ell$,

$$|\widehat{F}_{k,\ell} - F_k| \le \frac{\varepsilon_\ell}{32}, \qquad |\widehat{G}_{k,\ell} - G_k| \le \frac{\varepsilon_\ell}{32},$$

and

$$|\widehat{\alpha}^b_{k,\ell} - \alpha^b_k| \le \frac{\varepsilon_\ell}{32}, \qquad |\widehat{\alpha}^s_{k,\ell} - \alpha^s_k| \le \frac{\varepsilon_\ell}{32}.$$

Since all quantities are bounded in $[0,1]$, for any $x, y, \widehat{x}, \widehat{y} \in [0,1]$, we have

$$|xy - \widehat{x}\widehat{y}| \le |x - \widehat{x}| + |y - \widehat{y}|.$$

Applying this inequality to the two terms in the GFT decomposition gives

$$\begin{aligned}|\widehat{g}_{k,\ell} - g_k| &\le \left|\widehat{F}_{k,\ell}\widehat{\alpha}^s_{k,\ell} - F_k\alpha^s_k\right| + \left|\widehat{G}_{k,\ell}\widehat{\alpha}^b_{k,\ell} - G_k\alpha^b_k\right| \\ &\le \frac{\varepsilon_\ell}{16} + \frac{\varepsilon_\ell}{16} = \frac{\varepsilon_\ell}{8}.\end{aligned}$$

Therefore, the empirical GFT estimates are uniformly accurate.

We first show that the optimal discretized price is never eliminated. By the uniform estimation bound, for every active price k,

$$\widehat{g}_{k,\ell} \le g_k + \frac{\varepsilon_\ell}{8} \le g^\star + \frac{\varepsilon_\ell}{8},$$

and therefore

$$\widehat{g}^{\max}_\ell \le g^\star + \frac{\varepsilon_\ell}{8}.$$

Moreover,

$$\widehat{g}_{k^\star,\ell} \ge g^\star - \frac{\varepsilon_\ell}{8}.$$

Hence,

$$\widehat{g}_{k^\star,\ell} \ge \widehat{g}^{\max}_\ell - \frac{\varepsilon_\ell}{4} \ge \widehat{g}^{\max}_\ell - \frac{\varepsilon_\ell}{2}.$$

Therefore, $k^\star$ satisfies the elimination criterion and is never removed.

Then, consider any price k that survives phase ℓ. By the elimination condition,

$$\widehat{g}_{k,\ell} \ge \widehat{g}^{\max}_\ell - \frac{\varepsilon_\ell}{2} \ge \widehat{g}_{k^\star,\ell} - \frac{\varepsilon_\ell}{2}.$$

Using the uniform estimation bound,

$$g_k \ge \widehat{g}_{k,\ell} - \frac{\varepsilon_\ell}{8},$$

and

$$\widehat{g}_{k^\star,\ell} \ge g^\star - \frac{\varepsilon_\ell}{8}.$$

Therefore,

$$g_k \ge g^\star - \frac{\varepsilon_\ell}{8} - \frac{\varepsilon_\ell}{2} - \frac{\varepsilon_\ell}{8} = g^\star - \frac{3}{4}\varepsilon_\ell.$$

Equivalently,

$$g^\star - g_k \le \frac{3}{4}\varepsilon_\ell.$$

This proves both claims. □

B.11 Proof of Theorem 3

Proof. Let

$$\mu_\Sigma = \mathbb{E}[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s],$$

and set $\delta = T^{-3}$. Let $\mathcal{E}$ denote the event that the guarantees in Lemmas 5–8 hold simultaneously. By a union bound, $\Pr(\mathcal{E}) \geq 1 - \delta$. On the failure event, the regret is at most T, and therefore its contribution is bounded by

$$T \Pr(\mathcal{E}^c) \leq T\delta = O(1).$$

Hence, it suffices to analyze the regret conditioned on $\mathcal{E}$.

Let $g(p) = \mathbb{E}[\mathrm{GFT}(p, S, B)]$ and let $g_K^\star = \max_{k\in[K]} g_k$ denote the optimal discretized GFT. The regret is decomposed as

$$R_T \leq R_{\mathrm{disc}} + R_{\mathrm{cal}} + R_{\mathrm{stat}} + R_{\mathrm{wait}},$$

where R_{disc} is the discretization error, R_{cal} is the regret from boundary-tail calibration, R_{stat} is the statistical regret from scouting and elimination, and R_{wait} is the regret caused by waiting for delayed observations.

First, let $p^\star$ be the optimal continuous price and let $q^\star$ be the closest grid price. Since the grid spacing is $1/K$, the Lipschitz assumption implies

$$g(p^\star) - g(q^\star) \leq \frac{C}{K}.$$

Therefore,

$$R_{\mathrm{disc}} \leq T(g(p^\star) - g(q^\star)) = O(T/K).$$

Next, we bound the regret incurred during boundary-tail calibration. In phase ℓ, the learner collects

$$m_\ell^{\mathrm{cal}} = \left\lceil C_{\mathrm{cal}} \frac{\Lambda}{\eta_\ell^2} \right\rceil$$

boundary samples from each side, where $\eta_\ell = \varepsilon_\ell/256$. Each round incurs regret at most one, and hence

$$R_{\mathrm{cal}}^{(\ell)} = O(m_\ell^{\mathrm{cal}}).$$

Since

$$\eta_\ell = \Theta(\varepsilon_\ell),$$

we have

$$R_{\mathrm{cal}}^{(\ell)} = O\left(\frac{\Lambda}{\varepsilon_\ell^2}\right).$$

The phases are executed with geometrically decreasing accuracy $\varepsilon_\ell = 2^{-\ell}$ until

$$\varepsilon_\ell \leq \varepsilon_\star = c_\star \sqrt{\frac{K\Lambda}{T}}.$$

Hence the total calibration cost is dominated by the final phase:

$$R_{\mathrm{cal}} \leq O\left(\frac{\Lambda}{\varepsilon_\star^2}\right) = O\left(\frac{T}{K} \log KT\right).$$

We next consider the statistical regret. For phase ℓ, Lemmas 6 and 7 imply that the empirical GFT estimates satisfy

$$|\widehat{g}_{k,\ell} - g_k| \leq \frac{\varepsilon_\ell}{8}$$

for every active price. By Lemma 8, the optimal discretized price is never eliminated, and every surviving price after phase ℓ satisfies

$$g_K^\star - g_k \leq \frac{3}{4}\varepsilon_\ell.$$

For a suboptimal price k, define

$$\Delta_k = g_K^\star - g_k.$$

Such a price can remain active only while $\varepsilon_\ell = \Omega(\Delta_k)$. In phase ℓ, each active price is sampled

$$n_\ell = O\left(\frac{\Lambda}{\varepsilon_\ell^2}\right)$$

times. Therefore, the total number of samples allocated to price k satisfies

$$N_k \le C \sum_{\ell:\varepsilon_\ell \ge c\Delta_k} \frac{\Lambda}{\varepsilon_\ell^2} = O\left(\frac{\Lambda}{\Delta_k^2}\right).$$

Thus, the regret contribution of price k is

$$R_k \le \Delta_k \min\left\{T, C\frac{\Lambda}{\Delta_k^2}\right\}.$$

Summing over all suboptimal prices gives

$$R_{\text{stat}} \le \sum_{k\neq k^\star} \min\left\{T\Delta_k, C\frac{\Lambda}{\Delta_k}\right\} = O(\sqrt{KT\log T}).$$

Finally, we bound the waiting cost. In phase ℓ, the learner waits for at most $O(L_\ell)$ rounds until all observations reach the selected window. By Lemma 8, every active price in this phase has gap at most $O(\varepsilon_\ell)$. Hence,

$$R_{\text{wait}}^{(\ell)} \le C\varepsilon_\ell L_\ell.$$

By Lemma 5,

$$L_\ell \le C\left(1 + \frac{\mu_\Sigma}{\varepsilon_\ell}\right).$$

Therefore,

$$R_{\text{wait}}^{(\ell)} \le C\varepsilon_\ell + C\mu_\Sigma.$$

Since the number of phases is $O(\log T)$,

$$R_{\text{wait}} = O(\mu_\Sigma \log T).$$

Combining all components,

$$R_T = O\left(\frac{T}{K}\log KT + \sqrt{KT\log T} + \frac{T}{K} + \mu_\Sigma \log T\right).$$

Choosing

$$K = \lceil T^{1/3} \rceil$$

gives

$$\frac{T}{K} = O(T^{2/3}),$$

and

$$\sqrt{KT\log T} = O(T^{2/3}\sqrt{\log T}).$$

Finally, substituting $\mu_\Sigma = \mathbb{E}[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s]$ gives

$$R_T = O\left(T^{2/3}\log T + (\mathbb{E}[\mathcal{D}^b] + \mathbb{E}[\mathcal{D}^s])\log T\right).$$

□