OPTIMAL REGRET FOR ONLINE MARKET MAKING WITH LIMIT ORDER BOOK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online learning in market making, where, at each round, a market maker posts bid and ask prices before observing the market price and the private valuation of an incoming trader. In this setting, Maran & Restelli (2026) introduce a feedback model motivated by limit order books, in which the trader's valuation is revealed only if no transaction occurs. Assuming that trader valuations are drawn i.i.d. from an unknown distribution while market prices are chosen adversarially, they establish an expected regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$. In this work, we improve upon this guarantee by establishing a high-probability regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. As a warm-up, we first consider the full-feedback setting. We introduce a discretization of the bid–ask space based on two coupled grids and combine it with Hedge to achieve the desired regret rate. Building on these ideas, we then address the substantially weaker feedback induced by a limit order book and develop an algorithm that achieves the same guarantee. Finally, we investigate the limits of learnability in fully adversarial environments, where the valuations may vary arbitrarily as well. Perhaps surprisingly, we show that when both market prices and trader valuations are chosen adversarially, sublinear regret is impossible even under full feedback, thereby motivating our stochastic assumption on the valuations.

1 INTRODUCTION

Market makers provide liquidity to financial markets by posting prices at which they are willing to buy and sell an asset (Amihud & Mendelson, 1986). They aim to profit from the bid–ask spread while managing the risks of price changes and trading with better-informed participants, a central theme in the economics literature on market making. Today, market making is a core activity of quantitative trading firms such as Jane Street and Citadel Securities, which use mathematical models and automated trading systems to provide liquidity across financial markets (Biais et al., 2005). Designing effective strategies for choosing and updating these quotes is therefore an important problem in algorithmic trading (Glosten & Harris, 1988; Madhavan, 2000). A central challenge in designing market-making strategies is that quotes must be chosen under uncertainty about both the market price of the asset and the willingness of incoming traders to transact. Moreover, the available information depends on the quotes themselves, linking current trading decisions to future learning opportunities. Online learning provides a framework for studying this interaction through a sequential decision-making problem. At each round, a market maker posts a bid and an ask before observing the market price and the incoming trader's private valuation. Thus, the performance is measured by regret against the best fixed bid–ask pair in hindsight.

Cesa-Bianchi et al. (2024) study this problem under two-bit feedback, where the learner observes the market price and whether the trader buys, sells, or does not trade, while the trader's valuation remains hidden. They establish tight regret guarantees under several assumptions on market prices and valuations. In particular, the minimax regret is of order $T^{2/3}$ when market prices and trader valuations are i.i.d. and independent of each other. Their lower bound holds even under Lipschitz assumptions on the underlying distributions, showing that the difficulty persists in regular stochastic environments. They also establish impossibility results when suitable assumptions on regularity and independence are absent. These results highlight the role of the information available to the market maker in determining learnability.

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.

More recently, Maran & Restelli (2026) introduce a richer feedback model motivated by limit order books. In this model, the trader's valuation remains hidden when a transaction occurs but is revealed when no transaction takes place. Thus, the posted quotes determine both the reward and the information available for future decisions. Under i.i.d. trader valuations, they obtain high-probability regret bounds of $\widetilde{\mathcal{O}}(\sqrt{T})$ for independent stochastic market prices and extend these guarantees to certain mean-reverting price processes, without imposing smoothness on the valuation distribution. Differently, when the market prices are chosen by an oblivious adversary, while the valuations remain stochastic and stationary, they establish an expected regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$.

This leaves a gap between the guarantees available for stochastic and adversarial market prices under limit order book feedback. Is the slower rate an unavoidable consequence of arbitrary market prices, or can the information revealed by the order book still support a $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantee? Moreover, both guarantees rely on the assumption that trader valuations are drawn from a fixed distribution. It is therefore natural to ask whether learning remains possible when this assumption is removed. Specifically, in this work we address the following two research questions:

(i) *Can we achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret under limit order book feedback with adversarial market prices?*
(ii) *Can sublinear regret also be achieved when both market prices and trader valuations are adversarial?*

We answer the first question affirmatively and the second negatively. When trader valuations are drawn i.i.d. from an unknown distribution, we establish a high-probability regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ while allowing market prices to remain arbitrary. Finally, we show that sublinear regret is impossible in fully adversarial environments, even under full feedback.

1.1 Original Contributions

We study online market making under different assumptions on the environment and the feedback available to the learner. Our main focus is the setting in which market prices are chosen by an adaptive adversary, while private valuations are drawn from an arbitrary distribution. We establish optimal regret rates under both full feedback and limit order book feedback, and show that learning becomes impossible when private valuations are also chosen adversarially. Our contributions can be summarized as follows:

- **Optimal regret under full feedback.** As a warm-up, we establish a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound under full feedback. Our algorithm combines two Hedge instances, one for bid prices and one for ask prices. A key observation is that swapping the sampled quotes whenever they violate the feasibility constraint never decreases utility, allowing us to analyze the two instances separately and transfer their guarantees to the actions actually played.
- **Optimal high-probability regret under limit order book.** For the feedback model introduced by Maran & Restelli (2026), we establish a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound holding with high probability, even against an adaptive adversary. This improves upon the previous $\widetilde{\mathcal{O}}(T^{2/3})$ bound, which held in expectation and assumed an oblivious adversary. Our analysis exploits the feedback structure. Specifically, each round allows the learner to reconstruct losses for the interval between the sampled quotes. Using this additional information, we achieve the same regret rate as under full feedback.
- **Impossibility in a fully adversarial environment.** When both market prices and private valuations are chosen adversarially, we prove a linear regret lower bound. This impossibility result holds even under full feedback, showing that observing all parameters at the end of each round is insufficient to guarantee sublinear regret.

The $\tilde{\mathcal{O}}(\sqrt{T})$ upper bounds are optimal up to logarithmic factors, matching the standard $\Omega(\sqrt{T})$ lower bound for the multi-armed bandit problem (Lattimore & Szepesvári, 2020). Together, our results clarify how the assumptions on the environment and the available feedback affect the learnability of market making.

Table 1 summarizes our guarantees and compares them with previous work.

Table 1: Comparison between our results and the state-of-the-art.

Feedback	(M_t, V_t)	(Maran & Restelli, 2026)	**Our Work**
Full	**(adv)+(stoc)**	✗	$\widetilde{\mathcal{O}}(\sqrt{T})$
	(adv)+(adv)	✗	$\Omega(T)$
LOB	**(adv)+(stoc)**	$\widetilde{\mathcal{O}}(T^{2/3})$	$\widetilde{\mathcal{O}}(\sqrt{T})$

1.2 Related Works

Market making has long been studied in market microstructure (Bouchaud et al., 2008), with traditional approaches often formulating the choice of bid and ask prices as a stochastic control problem (Guéant, 2017). An alternative approach treats market making as an online learning problem, following the work of Abernethy & Kale (2013) and its more recent developments by Cesa-Bianchi et al. (2024) and Maran & Restelli (2026).

Within this line of research, Cesa-Bianchi et al. (2024) study two-bit feedback about private valuation, where the learner observes only whether a purchase and a sale occur, and derive regret guarantees under regularity assumptions on the distributions describing the environment. Although these assumptions support the analysis, they may exclude behavior concentrated around particular prices. For example, the tendency of bidders to bid round numbers documented by van Oldeniel et al. (2025) motivates considering distributions that need not be smooth.

Alongside the assumptions on the environment, the information available to the learner is another key aspect of the problem. Since two-bit feedback may omit information available in a limit order book, Maran & Restelli (2026) introduce a richer feedback model, for which they establish an expected regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$ against an oblivious adversary. Their guarantees are limited to expected regret in the adversarial-price case and do not cover the fully adversarial setting in which both market prices and private valuations are chosen adversarially.

More broadly, market making is related to other online learning problems in economics, such as dynamic pricing (Kleinberg & Leighton, 2003), first-price auctions (Cesa-Bianchi et al., 2026; Myerson, 1981) and bilateral trade (Cesa-Bianchi et al., 2023; Myerson & Satterthwaite, 1983), where the feedback makes it difficult to infer how rewards vary across actions.

2 Preliminaries

We study online market making over $T \in \mathbb{N}_{\geq 0}$ rounds. At each round $t \in \{1, \dots, T\}$, the learner (*i.e.*, the market maker) posts a bid quote B_t and an ask quote A_t, where $A_t \geq B_t$.[1] We normalize prices to the interval $[0, 1]$, which can always be done through an affine transformation of the original price space. Accordingly, the feasible decision set is:

$$\mathcal{U} = \{(a, b) \in [0, 1]^2 : a \geq b\}.$$

At the beginning of round $t \in \{1, \dots, T\}$, a trader with private valuation $V_t \in [0, 1]$ arrives. The trader buys the asset from the market maker whenever $V_t > A_t$, and sells the asset to the market maker whenever $V_t \leq B_t$. If $V_t \in (B_t, A_t]$, no trade occurs. We assume that, whenever a trade takes place, the market maker immediately offloads its position on a perfectly liquid secondary market at price $M_t \in [0, 1]$, which is observed at the end of the round only, incurring no additional transaction costs or inventory constraints. The resulting utility function of the market maker is:

$$u(a, b, v, m) = (a - m)\mathbb{I}\{v > a\} + (m - b)\mathbb{I}\{v \leq b\}.$$

Throughout the paper, we will use the shorthand notation of $u_t(a, b) = u(a, b, V_t, M_t)$.

The online market making interaction is summarized in Protocol 1. We consider two possible types of feedback: *full* feedback (Line 10) and *limit order book* feedback (Lines 4–9). In both cases, the

[1] This is a standard no-arbitrage condition. Otherwise, one could buy at the ask and immediately sell at the bid for a guaranteed profit.

learner observes the market price M_t at the end of each round. The realized trader valuation V_t is always observed at the end of each round under full feedback, while it is observed only if no trade occurs in the limit order book case.

Protocol 1 Online Market Making

for $t = 1, \dots, T$ **do**
  A new trader with valuation V_t arrives
  The learner posts ask and bid prices $(A_t, B_t) \in \mathcal{U}$

Limit order book

  $M_t \in [0,1]$ is observed
  if $B_t < V_t \leq A_t$ **then**
    $V_t \in [0,1]$ is observed
  else
    The learner observes $\mathbb{1}\{B_t \geq V_t\}$ and $\mathbb{1}\{A_t < V_t\}$
  end if

Full feedback

  $(V_t, M_t) \in [0,1]^2$ is observed

  The learner gains utility $u_t(A_t, B_t)$
end for

We consider two different models for the sequence $\{(V_t, M_t)\}_{t=1}^T$:

(adv) + (stoc) $\{V_t\}_{t=1}^T$ is an i.i.d. sequence, while $\{M_t\}_{t=1}^T$ is chosen by an adaptive adversary.

(adv) + (adv) Fully adversarial setting, in which the sequence $\{(V_t, M_t)\}_{t=1}^T$ is generated by an oblivious adversary.

The first case **(adv) + (stoc)** is the main focus of this work, while we provide a lower bound, which naturally extends to the case of an adaptive adversary, for the second case **(adv) + (adv)**.

Let us remark that we make no assumption on the regularity of the sequence of $\{V_t\}_{t=1}^T$. Indeed, differently from Cesa-Bianchi et al. (2024), we do *not* assume that the c.d.f. of the trader valuations distribution is Lipschitz–continuous.

2.1 Performance Metric

We use as performance metric the notion of *cumulative regret*, which is defined as follows:

$$R_T = \sup_{(a,b)\in\mathcal{U}} \mathbb{E}\left[\sum_{t=1}^T (u_t(a,b) - u_t(A_t, B_t))\right],$$

where the expectation is taken over the randomization of the environment. We define the c.d.f. for the random variable V underlying the trader valuations in the **(adv)+(stoc)** as $F(x) = \mathbb{P}(V \leq x)$ and, accordingly, we define the expected utility as:

$$U_t(a,b) \coloneqq \mathbb{E}_V[u_t(a,b)] = (M_t - b)F(b) + (a - M_t)(1 - F(a)).$$

We notice that the utility can be decomposed into two components, one for the bid and one for the ask price. Thus, we denote:

$$u_t^A(a) \coloneqq (a - M_t)\mathbb{1}\{V_t > a\}, \qquad u_t^B(b) \coloneqq (M_t - b)\mathbb{1}\{V_t \leq b\},$$

so that $u_t(a,b) = u_t^B(b) + u_t^A(a)$. Defining $\mu_T = \frac{1}{T}\sum_{t=1}^T M_t$, we also notice that:

$$\sup_{(a,b)\in\mathcal{U}} \frac{1}{T}\sum_{t=1}^T U_t(a,b) = \sup_{(a,b)\in\mathcal{U}} \Big[(\mu_T - b)F(b) + (a - \mu_T)(1 - F(a))\Big].$$

Finally, defining $U(a,b) \coloneqq (\mu_T - b)F(b) + (a - \mu_T)(1 - F(a))$, we can rewrite the regret as:

$$R_T = \sup_{(a,b)\in\mathcal{U}} \sum_{t=1}^T \Big[U_t(a,b) - U_t(A_t, B_t)\Big] = T \cdot \sup_{(a,b)\in\mathcal{U}} U(a,b) - \sum_{t=1}^T U_t(A_t, B_t). \tag{1}$$

Notice that, in case of adversarial trader valuations, where no statistical assumption is made on the sequence $\{(M_t, V_t)\}_{t=1}^T$, no expectation is taken in the regret definition.

In the next section, we start with assuming full feedback and stochastic trader valuations.

3 Warm up: The Full Feedback Case

In this section, we start by considering the **(adv)+(stoc)** scenario.

This setting is the one already studied in (Maran & Restelli, 2026), with the difference that we assume, throughout this section, to have full feedback, while we let the adversary to be adaptive. Our approach will employ two instances of the Hedge algorithm (Freund & Schapire, 1997), thus significantly departing from the existing technique of Maran & Restelli (2026), where the algorithm is mainly based on Follow the Perturbed Leader (FTPL) (Hutter & Poland, 2005).

3.1 Algorithm

Algorithm 2 Market Making with Experts

Input: Time horizon T, K, η

1: Initialize $A_1 \leftarrow 1$, $B_1 \leftarrow 0$, and grid $\mathcal{G}$
2: Initialize W_1^A, W_1^B to 1 on $\mathcal{G}$
3: **for** $t = 1, \dots, T$ **do**
4: Form w_t^A, w_t^B by normalizing W_t^A, W_t^B
5: Draw $Y_t \sim w_t^A$ and $X_t \sim w_t^B$
6: Post prices $(A_t, B_t) \leftarrow (\max\{X_t, Y_t\}, \min\{X_t, Y_t\})$
7: Gain utility $u_t(A_t, B_t)$
8: Observe (M_t, V_t) and reconstruct $u_t^A(a), u_t^B(b)$ for all $a, b \in \mathcal{G}$
9: Update weights for all $a, b \in \mathcal{G}$:

$$W_{t+1}^A(a) \leftarrow W_t^A(a) \exp\left(\eta\, u_t^A(a)\right), \quad W_{t+1}^B(b) \leftarrow W_t^B(b) \exp\left(\eta\, u_t^B(b)\right),$$

10: **end for**

Our algorithm, whose pseudocode is provided in Algorithm 2, discretizes the bid and ask prices using the uniform grid $\mathcal{G} = \left\{\frac{i}{K} : i = 0, \dots, K\right\}$. The resulting Cartesian product $\mathcal{G} \times \mathcal{G}$ consists of $(K+1)^2$ points in $[0,1]^2$ and serves as the discrete action set. We then run two independent instances of Hedge (Orabona, 2019) with the same learning rate $\eta > 0$, one for bid prices and one for ask prices. Each instance maintains weights $\{W_t^B\}_{t=1}^T$ and $\{W_t^A\}_{t=1}^T$, respectively, and the corresponding probability distributions $\{w_t^B\}_{t=1}^T$ and $\{w_t^A\}_{t=1}^T$ over the one-dimensional grid $\mathcal{G}$ (Line 2). At each round, the algorithm samples a bid and an ask price from the corresponding distributions and swaps them if the resulting pair violates the constraint defining $\mathcal{U}$ (Line 6). It then posts the resulting pair of prices and updates the weights of both Hedge instances (Line 9). We emphasize that the swap only affects the prices played, not the weight updates. Since feedback is fully observable, at the end of each round we can evaluate the utility associated with every grid price in its respective role as a bid or an ask. Each Hedge instance is therefore updated using these utilities, regardless of whether the sampled prices were swapped (Lines 8–9).

3.2 Theoretical Results

In this section, we analyze Algorithm 2 and establish the associated $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound.

We start by stating the following lemma, which shows how enforcing feasibility by swapping the prices predicted by the two Hedge instances never decreases the utility. More precisely, the utility remains unchanged when the sampled prices already satisfy the constraint, and increases by $X_t - Y_t$ otherwise. Thus, the swap in Line 6 of Algorithm 2 guarantees a pair of prices whose utility is at

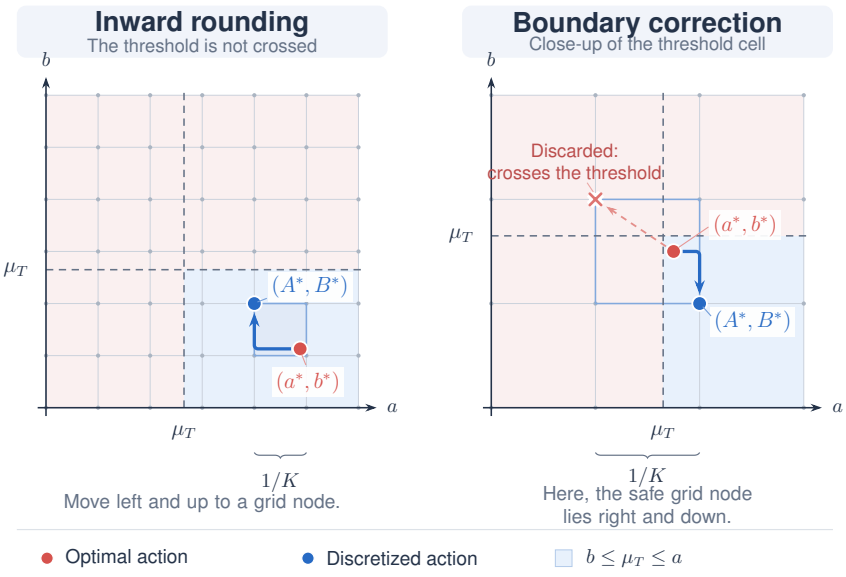

Figure 1: Discretization around the optimal action. *Left:* inward rounding. *Right:* correction when both coordinates would cross μ_T.

least as large as that of the original predictions. This property will allow us to transfer the utility guarantees of the two Hedge instances to the actions actually played by the algorithm.[2]

Lemma 3.1. *Consider $X_t \sim w_t^B$ and $Y_t \sim w_t^A$. Then it holds that:*

$$u_t(\max\{X_t, Y_t\}, \min\{X_t, Y_t\}) = u_t(Y_t, X_t) + (X_t - Y_t)_+,$$

where we denote $(x)_+ = \max\{0, x\}$. As a consequence, for any $t \in \{1, \dots, T\}$ it holds that $U_t(A_t, B_t) \ge U_t(Y_t, X_t)$ and $U(A_t, B_t) \ge U(Y_t, X_t)$.

Next, we show that restricting the action set to the grid incurs only a bounded loss in expected utility. Specifically, there exists a grid point whose expected utility is within $5/2K$ of the supremum over the continuous feasible set $\mathcal{U}$. This allows us to relate performance on the grid to the continuous optimum.

Lemma 3.2. *There exists a point $(A^*, B^*) \in \mathcal{G} \times \mathcal{G}$ such that:*

$$0 \le \sup_{(a,b)\in\mathcal{U}} U(a, b) - U(A^*, B^*) \le \frac{5}{2K}.$$

Figure 1 illustrates the idea behind the proof of Lemma 3.2. We start from a nearly optimal point and choose a nearby point on the grid, taking care not to cross the boundary defined by μ_T, as doing so could reduce the expected utility. This choice keeps the discretization loss small. We provide the full argument in Appendix A.

We are now ready to provide the final regret bound attained by Algorithm 2.

Theorem 3.3. *Let $\delta \in (0, 1)$ and $T \ge 2$. Algorithm 2 with $K = \lceil \sqrt{T} \rceil, \eta = \sqrt{\frac{\log(T)}{T}}$ attains regret:*

$$R_T \le \mathcal{O}\left(\sqrt{T \log(T)} + \sqrt{T \log(1/\delta)} + \log(1/\delta)\right),$$

with probability at least $1 - \delta$.

In Theorem 3.3, we derive a regret bound by adding the standard Hedge guarantees for the two one-dimensional instances. We analyze each instance using its original sampled quote, before any swap

[2]When full feedback is available, this lemma is not necessary. Indeed, it is possible to run a single Hedge instance on the two-dimensional grid to recover the same regret guarantees. We employ two Hedge instances invoking Lemma 3.1 as a warm-up to the next section.

Algorithm 3 Market Making with Limit Order Book

Input: Time horizon T, parameters η, β, γ, K

1: Initialize $\Delta \sim \mathcal{U}(0, \frac{1}{2K})$, $\mathcal{G}_A, \mathcal{G}_B$
2: Set W_1^A, W_1^B to 1
3: **for** $t = 1, \dots, T$ **do**
4: Form w_t^A, w_t^B normalizing W_t^A, W_t^B and probabilities p_t, q_t as in Equation (3)
5: Compute Q_t as in Equation (4) for all internal grid points
6: Draw $X_t \sim p_t$ and $Y_t \sim q_t$
7: Play prices $(A_t, B_t) = (\max\{X_t, Y_t\}, \min\{X_t, Y_t\})$
8: Observe M_t, observe V_t if no trade occurred
9: $\forall z \in (\mathcal{G}_A \cup \mathcal{G}_B) \cap [B_t, A_t]$, compute $\ell_t^s(z)$, for $s \in \{A, B\}$ respectively
10: Form the loss estimator $\hat{\ell}_t^s$, with $s \in \{A, B\}$, as in Equation (2)
11: Update weights W_{t+1}^A, W_{t+1}^B
12: **end for**

takes place. Although the swap couples the bid and ask prices actually posted, Lemma 3.1 shows that it never decreases utility. We can therefore analyze the two Hedge instances separately and carry their combined guarantee over to the actions played by the algorithm. The regret bound in Theorem 3.3 depends only logarithmically on the grid size K, as a consequence of the Hedge guarantee. This dependence is optimal in the standard expert setting. We would expect a different dependence on K under limit order book feedback, which provides less information than full feedback.

4 Optimal Bounds For The Limit Order Book Case

In this section, we study the more interesting setting of the limit order book feedback, where the information available to the learner depends on whether a trade occurs. If no exchange takes place, the learner receives full feedback; otherwise, only the utility is observed with the type of trade.

4.1 Algorithm

We now present Algorithm 3 for the limit order book feedback and analyze its regret.

Similarly to the previous approach, Algorithm 3 treats the bid and ask prices separately, maintaining disjoint grids $\mathcal{G}_B$ and $\mathcal{G}_A$ (Line 1), weights W_t^B and W_t^A (Line 2), and sampling distributions p_t and q_t (Lines 4–6). Unlike in the full-feedback setting, we must account for the possibility that grid points coincide with atoms of F[3]. We address this issue through a random shift $\Delta \sim \mathcal{U}(0, \frac{1}{2K})$ (Line 1), which almost surely prevents internal grid points from coincident with atoms of F, as shown in Lemma B.1 in Appendix B. This property is needed to reconstruct feedback over the whole interval $[B_t, A_t]$ in Equation (2), as it rules out boundary events such as $V_t = B_t$ at internal grid points.

For an integer $K \geq 1$, we define the two grids as:

$$\mathcal{G}_B = \left\{\Delta + \frac{i}{K} : i = 0, \dots, K-1\right\} \cup \{1\}, \; \mathcal{G}_A = \left\{\Delta + \frac{2i+1}{2K} : i = 0, \dots, K-1\right\} \cup \{0\}.$$

Each grid contains K internal points and one endpoint. We run two parallel instances of an EXP3-based algorithm (Auer et al., 2002). For convenience, we express the utility contributions as losses in $[0, 1]$ by defining $\ell_t^A(a) := {}^{(1-u_t^A(a))}/_2$, $\ell_t^B(b) := {}^{(1-u_t^B(b))}/_2$. Let $\beta > 0$ be an implicit exploration parameter (Neu, 2015). For each internal grid point $x \in \mathcal{G}_s$, with $s \in \{A, B\}$, let $Q_t(x) := \mathbb{P}\big(x \in [B_t, A_t] \cup \{0, 1\}\big)$ denote the probability that its loss is observed at round t. We set $Q_t(1) = Q_t(0) = 1$, since losses at the endpoints can always be computed from M_t, which is observed at every round. We further denote $O_t(x) := \mathbb{1}\{x \in [B_t, A_t] \cup \{0, 1\}\}$ as the event that a point $x \in \mathcal{G}_s$, $s \in \{A, B\}$ is observed. The limit order book feedback, together with the random shift Δ, allows us to recover

[3]We assume throughout the paper that F has no atom at zero. This reflects the fact that traders in financial markets typically assign strictly positive valuations to the assets they trade.

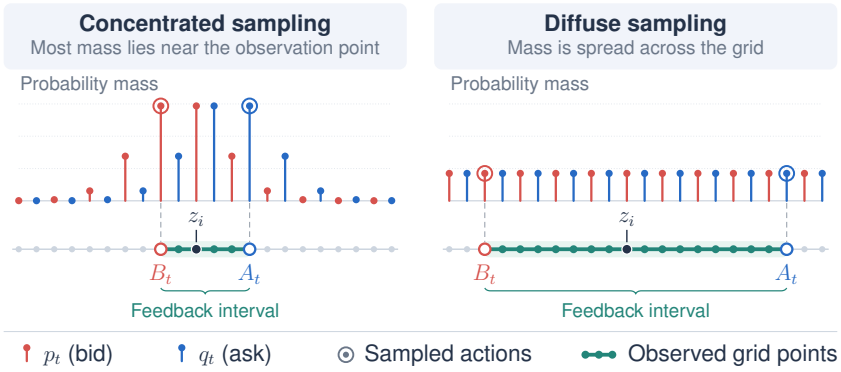

Figure 2: The two panels illustrate sampling under a narrower (left) and wider (right) bid–ask spread. The bars p_t and q_t denote the sampling probabilities over prices, while $Q_t(z_i)$ is the probability of observing an internal point $z_i \in [B_t, A_t]$. Concentrating the sampling mass induces greater utility, but weakens observability of internal points. Consequently, although $Q_t(z_i)$ may be small, the aggregate importance-weight ratios $p_t(z_i)/Q_t(z_i)$ and $q_t(z_i)/Q_t(z_i)$ remain bounded, yielding the $\mathcal{O}(\log K)$ dependence in Lemma 4.2.

the exact losses at all grid points in $[B_t, A_t]$, as we will show in Lemma 4.1. We use these observed losses to construct the estimates:

$$\hat{\ell}_t^A(x) = \frac{O_t(x)\ell_t^A(x)}{Q_t(x)+\beta}, \qquad \hat{\ell}_t^B(x) = \frac{O_t(x)\ell_t^B(x)}{Q_t(x)+\beta}. \tag{2}$$

The implicit exploration parameter β prevents the estimates from becoming too large when a grid point is rarely observed, helping to control their variance.

For each component $s \in \{A, B\}$, we update the weights according to:

$$W_{t+1}^s(x) = W_t^s(x)\exp\big(-\eta\hat{\ell}_t^s(x)\big),$$

and denote the corresponding normalized weights by w_t^s. Given an explicit exploration parameter $\gamma \in (0, 1)$, we define the sampling distributions for the buying and selling components (Line 4) as:

$$p_t(x) = (1-\gamma)w_t^B(x) + \frac{\gamma}{K+1}, \qquad q_t(x) = (1-\gamma)w_t^A(x) + \frac{\gamma}{K+1}. \tag{3}$$

As in EXP3.P (Auer et al., 2002), mixing the normalized weights with the uniform distribution ensures that every grid point is sampled with probability at least $\gamma/(K+1)$. This exploration helps maintain sufficient observation of the losses across the grids.

Finally, we notice that for any internal grid point $x \in \mathcal{G}_s$, with $s \in \{A, B\}$, we can compute the observation probability directly from the sampling distributions:

$$Q_t(x) = 1 - \left(\sum_{y\in\mathcal{G}_A:y<x} q_t(y)\right)\left(\sum_{y\in\mathcal{G}_B:y<x} p_t(y)\right) - \left(\sum_{y\in\mathcal{G}_A:y>x} q_t(y)\right)\left(\sum_{y\in\mathcal{G}_B:y>x} p_t(y)\right). \tag{4}$$

Indeed, x lies outside the played interval $[B_t, A_t]$ precisely when both sampled prices are strictly below x or both are strictly above it. Since the prices are sampled independently, the probabilities of these two events are given by the products above.

4.2 Theoretical Results

In this section, we establish the regret bound for Algorithm 3.

We begin by showing that the available feedback is sufficient to compute the loss estimates used by the algorithm, and then use these estimates to analyze its performance. Specifically, the first lemma shows that, despite the limited feedback, playing a price pair (A_t, B_t) allows us to recover the exact losses at every grid point between the two quotes. Thus, a single round provides information about an entire interval of grid points, which we use to construct the estimates in Equation (2).

Lemma 4.1. *Let (A_t, B_t) be a price pair with $B_t \leq A_t$, played at time step t. Then, for any point $x \in (\mathcal{G}_A \cup \mathcal{G}_B) \cap [B_t, A_t]$, it is possible to reconstruct $\ell_t^s(x),\ s \in \{A, B\}$, after playing (A_t, B_t).*

Lemma 4.1 follows from the structure of the limit order book feedback. If no trade occurs, the learner receives full feedback. Otherwise, observing whether a purchase or a sale occurred reveals on which side of the bid–ask spread the private value lies. In either case, the learner can determine how the private value compares with each grid point within the spread and, since M_t is always observed, reconstruct the corresponding utility.

Next, Lemma 4.2 bounds the ratios between the probability of sampling each grid point and the probability of observing its loss. These ratios measure how much information the feedback provides relative to how often each point is selected. Since each round reveals losses throughout the interval between the sampled quotes, the sum of these ratios grows only logarithmically with the grid size, with explicit exploration ensuring a minimum sampling probability at every point.

Lemma 4.2. *For any time step $t \in \{1, \dots, T\}$, with $T \geq 2$, it holds:*

$$\sum_{x \in \mathcal{G}_A} \left(\frac{q_t(x)}{Q_t(x)} \right) + \sum_{x \in \mathcal{G}_B} \left(\frac{p_t(x)}{Q_t(x)} \right) \leq 4 \left(1 + \log \left(\frac{2(K+1)}{\gamma} \right) \right).$$

Consequently, it holds:

$$\sum_{x \in \mathcal{G}_s,\ s \in \{A,B\}} \frac{w_t^s(x)}{Q_t(x) + \beta} \leq \frac{4}{1-\gamma} \left(1 + \log \left(\frac{2(K+1)}{\gamma} \right) \right). \tag{5}$$

Figure 2 illustrates the result of Lemma 4.2. Specifically, reconstructing the losses over the interval $[B_t, A_t]$ allows us to bound the term measuring the cost of the limit order book feedback by $\mathcal{O}(\log K)$. This lies between the full-feedback setting, where every loss is observed and the corresponding term is $\mathcal{O}(1)$, and the bandit setting, where only the loss of the sampled action is observed and the term scales as $\mathcal{O}(K)$.

We are now ready to show the regret guarantees of Algorithm 3.

Theorem 4.3. *Let $\delta \in (0, 1)$ and $T \geq 2$. There exist $\eta = \beta = \Theta(\sqrt{1/T})$ such that Algorithm 3 instantiated with $K = \lceil \sqrt{T} \rceil$, and $\gamma = 1/T$ attains:*

$$R_T = \mathcal{O} \left(\sqrt{T} \log \left(\frac{T}{\delta} \right) \right),$$

with probability at least $1 - \delta$.

Theorem 4.3 improves upon the bound of Maran & Restelli (2026) by establishing a high-probability regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$, which is optimal up to logarithmic factors and holds even against an adaptive adversary.

5 An Impossibility Result

We finally consider the full-feedback setting under the **(adv)+(adv)** case. We establish a linear regret lower bound, showing that market making is unlearnable in a fully adversarial environment, even when the entire valuations are observed at the end of each round.

This is done in the following theorem.

Theorem 5.1. *Consider the market making problem with adversarial sequence of $\{M_t\}_t$ and $\{V_t\}_t$ and full feedback. Then, for any algorithm $\mathcal{A}$, there exists a sequence $\{(M_t, V_t)\}_{t=1}^T$ where expected regret $R_T \geq \frac{T}{6}$.*

The proof of Theorem 5.1 employs the well-known Yao's minimax principle (Yao, 1977). Specifically, we build a distribution over problem instances in which the relevant price interval shrinks from either the left or the right at random in each round. The learner cannot anticipate which part of the interval will remain, whereas a single fixed action chosen in hindsight can be selected using the final interval. This uncertainty creates a constant gap between the learner's expected utility and that of the hindsight benchmark in each round, yielding a linear regret lower bound. The full proof is given in Appendix C.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing for clarity and readability, generating figures, assist with identifying relevant literature, and, in a limited number of cases, help turn author-developed proof sketches into clearer lemma statements. All AI-assisted text and mathematical statements were subsequently reviewed and verified by the authors, and the cited references were checked against the original sources.

REFERENCES

Jacob Abernethy and Satyen Kale. Adaptive market making via online learning. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/995e1fda4a2b5f55ef0df50868bf2a8f-Paper.pdf.

Yakov Amihud and Haim Mendelson. Asset pricing and the bid-ask spread. *Journal of financial Economics*, 17(2):223–249, 1986.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Bruno Biais, Larry Glosten, and Chester Spatt. Market microstructure: A survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets*, 8(2):217–264, 2005. ISSN 1386-4181. doi: https://doi.org/10.1016/j.finmar.2004.11.001. URL https://www.sciencedirect.com/science/article/pii/S1386418104000382.

Jean-Philippe Bouchaud, J. Doyne Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand, 2008. URL https://arxiv.org/abs/0809.0822.

Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Repeated bilateral trade against a smoothed adversary. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1095–1130. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/cesa-bianchi23a.html.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Luigi Foscari, and Vinayak Pathak. Market making without regret. *ArXiv*, abs/2411.13993, 2024. URL https://api.semanticscholar.org/CorpusID:274165544.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. *SIAM Journal on Computing*, 55(5):903–938, 2026. doi: 10.1137/24M1712308. URL https://doi.org/10.1137/24M1712308.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

Lawrence R Glosten and Lawrence E Harris. Estimating the components of the bid/ask spread. *Journal of financial Economics*, 21(1):123–142, 1988.

Olivier Guéant. Optimal market making, 2017. URL https://arxiv.org/abs/1605.01862.

Marcus Hutter and Jan Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6(22):639–660, 2005. URL http://jmlr.org/papers/v6/hutter05a.html.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '03, pp. 594, USA, 2003. IEEE Computer Society. ISBN 0769520405.

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

Haipeng Luo and Robert Schapire. A drifting-games analysis for online learning and applications to boosting. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/24402144990624b417229a96ad7fa7bc-Paper.pdf.

Ananth Madhavan. Market microstructure: A survey. *Journal of financial markets*, 3(3):205–258, 2000.

Davide Maran and Marcello Restelli. Online market making and the value of observing the order book. In Steve Hanneke and Tor Lattimore (eds.), *The Thirty Ninth Annual Conference on Learning Theory, June 29 - July 3, 2026, San Diego, California, USA*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4969–4998. PMLR, 2026. URL https://proceedings.mlr.press/v336/maran26a.html.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981. doi: 10.1287/moor.6.1.58. URL https://doi.org/10.1287/moor.6.1.58.

Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(83)90048-0. URL https://www.sciencedirect.com/science/article/pii/0022053183900480.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e5a4d6bf330f23a8707bb0d6001dfbe8-Paper.pdf.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

Mark van Oldeniel, Christopher M Snyder, and Adriaan R Soetevent. Round bidding in auctions. Working Paper 34564, National Bureau of Economic Research, December 2025. URL http://www.nber.org/papers/w34564.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227, 1977. doi: 10.1109/SFCS.1977.24.

APPENDIX

The Appendix is organized as follows:

- In Appendix A, we provide the omitted proofs and lemmas of Section 3.
- In Appendix B, we provide the omitted proofs and lemmas of Section 4.
- In Appendix C, we prove the lower bound stated in Section 5.

A WARM UP

We start by proving that, by ordering the bid and ask prices, the mean utility cannot reduce.

Lemma 3.1. *Consider $X_t \sim w_t^B$ and $Y_t \sim w_t^A$. Then it holds that:*

$$u_t(\max\{X_t, Y_t\}, \min\{X_t, Y_t\}) = u_t(Y_t, X_t) + (X_t - Y_t)_+,$$

where we denote $(x)_+ = \max\{0, x\}$. As a consequence, for any $t \in \{1, \dots, T\}$ it holds that $U_t(A_t, B_t) \geq U_t(Y_t, X_t)$ and $U(A_t, B_t) \geq U(Y_t, X_t)$.

Proof. If $X_t = Y_t$ then the statement is trivially proved, so we assume $X_t \neq Y_t$. If $X_t < Y_t$, then $u_t(Y_t, X_t) = u_t^B(B_t) + u_t^A(A_t)$ and $(X_t - Y_t)_+ = 0$ the claim is true.

If instead $X_t > Y_t$ and $V_t \neq \{Y_t, X_t\}$, then we have that for each case $V_t < Y_t < X_t$, $V_t \in (Y_t, X_t)$ and $Y_t < X_t < V_t$ we have an improvement of $(X_t - Y_t)_+$ in the utility by swapping X_t with Y_t. For example, if $V_t < Y_t < X_t$, then $u_t(X_t, Y_t) = u_t^A(X_t) + u_t^B(Y_t) = (M_t - Y_t) = u_t(Y_t, X_t) + (X_t - Y_t)_+$ and a similar reasoning can be applied to the other cases.

In the cases where $V_t \in \{A_t, B_t\}$, if $V_t = X_t$ and $X_t > Y_t$, then $u_t(Y_t, X_t) = (Y_t - X_t)$ and $u_t(X_t, Y_t) = 0$, so the statement still holds. Lastly, if $V_t = Y_t$ and $X_t > Y_t$, $u_t(Y_t, X_t) = (M_t - X_t)$ and $u_t(X_t, Y_t) = (M_t - Y_t)$, by which we conclude the first part of the proof.

By observing that $(X_t - Y_t)_+ \geq 0$, we take the expected value of the previously proved equality, yielding $U_t(A_t, B_t) \geq U_t(Y_t, X_t)$. Finally, by taking the empirical mean over the sequence $\{M_t\}_{t=1}^T$, we get $U(A_t, B_t) \geq U(Y_t, X_t)$. □

We now turn to the intermediate Lemma 3.2, which justifies the reduction to a finite set of actions in Algorithm 2, by showing explicitly that there exists an almost optimal point on the grid, even though the utility is not regular.

Lemma 3.2. *There exists a point $(A^*, B^*) \in \mathcal{G} \times \mathcal{G}$ such that:*

$$0 \leq \sup_{(a,b) \in \mathcal{U}} U(a, b) - U(A^*, B^*) \leq \frac{5}{2K}.$$

Proof. We define a point on $(A^*, B^*) \in \mathcal{G} \times \mathcal{G}$ and then prove that such point has utility $2/K$ close to the optimal action $(a^*, b^*) \in \mathcal{U}$. While the supremum is not necessarily attained for the utility, we select $0 < \varepsilon < \frac{1}{2K}$ and a point such that $U(a^*, b^*) \geq \sup_{(a,b) \in \mathcal{U}} U(a, b) - \varepsilon$.

Let us first notice that we can reduce to the case where $b^* \leq \mu_T \leq a^*$. This is because of the structure of our utility function: if $a^* < \mu_T$, we have $(a^* - \mu_T)(1 - F(a^*)) \leq 0$, so we would weakly improve our revenue by substituting a^* with μ_T: analogously, we can argue that $b^* \leq \mu_T$ We select the point (A^*, B^*) following the same order, using the same logic.

Let $(A^*, B^*) \in \mathcal{G} \times \mathcal{G}$ be defined as:

$$B^* = \begin{cases} \dfrac{\lceil Kb^* \rceil}{K}, & \text{if } \frac{\lceil Kb^* \rceil}{K} \leq \mu_T, \\ \dfrac{\lfloor K\mu_T \rfloor}{K}, & \text{otherwise.} \end{cases} \qquad A^* = \begin{cases} \dfrac{\lfloor Ka^* \rfloor}{K}, & \text{if } \frac{\lfloor Ka^* \rfloor}{K} \geq \mu_T, \\ \dfrac{\lceil K\mu_T \rceil}{K}, & \text{otherwise.} \end{cases} \tag{6}$$

We decouple the contributions from the ask and the bid components, only studying the bid case, as the other follows from a similar argument: we then we sum up the additional contribution of ε to measure the difference with the supremum. Suppose first that the first case holds and $B^* = \frac{\lceil Kb^* \rceil}{K} \leq \mu_T$. Since F is non decreasing and bounded in $[0, 1]$, $F(B^*) \geq F(b^*)$. Therefore, $(\mu_T - B^*)F(B^*) \geq (\mu_T - B^*)F(b^*) = (\mu_T - b^*)F(b^*) - (B^* - b^*)F(b^*) \geq (\mu_T - b^*)F(b^*) - \frac{1}{K}$.

Suppose instead that $B^* = \frac{\lfloor K\mu_T \rfloor}{K}$. In this case the first rounding point lies above μ_T, which implies $0 \leq \mu_T - b^* < \frac{1}{K}$. Since $F(b^*) \leq 1$, $0 \leq (\mu_T - b^*)F(b^*) < \frac{1}{K}$. On the other hand, $B^* \leq \mu_T$, so $(\mu_T - B^*)F(B^*) \geq 0$. Thus again $(\mu_T - b^*)F(b^*) - (\mu_T - B^*)F(B^*) \leq \frac{1}{K}$.

Hence, in all cases, $(\mu_T - b^*)F(b^*) - (\mu_T - B^*)F(B^*) \leq \frac{1}{K}$, which concludes the argument, by summing up the two contributions from bid and ask prices. We conclude the argument by observing that:

$$\sup_{(a,b)\in\mathcal{U}} U(a,b) - U(A^*,B^*) = \sup_{(a,b)\in\mathcal{U}} U(a,b) - U(a^*,b^*) + U(a^*,b^*) - U(A^*,B^*) \leq \frac{5}{2K}.$$

This concludes the proof. □

After proving that there exists an almost-optimal point with respect to K on the grid $\mathcal{G}\times\mathcal{G}$, we finally turn to proving our regret bound.

Theorem 3.3. *Let $\delta \in (0,1)$ and $T \geq 2$. Algorithm 2 with $K = \lceil\sqrt{T}\rceil, \eta = \sqrt{\frac{\log(T)}{T}}$ attains regret:*

$$R_T \leq \mathcal{O}\left(\sqrt{T\log(T)} + \sqrt{T\log(1/\delta)} + \log(1/\delta)\right),$$

with probability at least $1-\delta$.

Proof. Using Lemma 3.2 and the regret decomposition in Equation (1), we have $R_T \leq \frac{5T}{2K} + \max_{(A^*,B^*)\in\mathcal{G}\times\mathcal{G}} \sum_{t=1}^{T} \left(U_t(A^*,B^*) - U_t(A_t,B_t)\right)$. By applying the decomposition into two separate instances of two problems, one on the bid prices and one in the ask prices, we can rewrite:

$$\begin{aligned}
&\max_{(A^*,B^*)\in\mathcal{G}\times\mathcal{G}} \sum_{t=1}^{T}\Big(U_t(A^*,B^*) - U_t(A_t,B_t)\Big) \\
&\qquad \leq \max_{(A^*,B^*)\in\mathcal{G}\times\mathcal{G}} \sum_{t=1}^{T}\Big(U_t(A^*,B^*) - U_t(Y_t,X_t)\Big) \\
&\qquad = \max_{A^*\in\mathcal{G}} \sum_{t=1}^{T} \mathbb{E}\Big[u_t^A(A^*) - u_t^A(Y_t)\Big] + \max_{B^*\in\mathcal{G}} \sum_{t=1}^{T} \mathbb{E}\Big[u_t^B(B^*) - u_t^B(X_t)\Big],
\end{aligned}$$

where the inequality holds by applying Lemma 3.1, while the equality holds because the utility can be decomposed into independent components. We notice here that we use in the proof the points $\{(X_t,Y_t)\}_{t=1}^T$ selected in the original instances of Hedge for the bid and ask, so that the standard concentration argument can be applied, for example using Luo & Schapire (2014). Thus, observing that both instances of Hedge have the same learning rate $\eta > 0$, we have that:

$$\begin{aligned}
&\max_{A^*\in\mathcal{G}} \sum_{t=1}^{T} \mathbb{E}\Big[u_t^A(A^*) - u_t^A(Y_t)\Big] + \max_{B^*\in\mathcal{G}} \sum_{t=1}^{T} \mathbb{E}\Big[u_t^B(B^*) - u_t^B(X_t)\Big] \\
&\qquad \leq 2\frac{\log(K+1)}{\eta} + \eta T + 2\sqrt{2T\log\left(\frac{2}{\delta}\right)} + \frac{4}{3}\log\left(\frac{2}{\delta}\right),
\end{aligned}$$

Building on that, by choosing $K = \lceil\sqrt{T}\rceil$ and $\eta = \sqrt{\frac{\log(T)}{T}}$, we get the second result $R_T = \mathcal{O}(\sqrt{T\log(T)} + \sqrt{T\log(1/\delta)} + \log(1/\delta))$, which concludes the argument. □

B Optimal bounds for LOB

We start by proving that almost surely no grid point corresponds to an atom for the c.d.f. F.

Lemma B.1. *Let $\Delta \sim \mathcal{U}(0, \frac{1}{2K})$ and grids $\mathcal{G}_A, \mathcal{G}_B$ be defined as in Algorithm 3. Then with probability 1 no internal grid point coincides with an atom for F.*

Proof. Atoms for the c.d.f. F can be at most countably many and are defined as points $y \in [0,1]$ such that $\mathbb{P}(V_t = y) > 0$. We fix such an atom y for F. Take a point $x \in \mathcal{G}_s$, $s \in \{A,B\}$: by definition the point is $x = \Delta + c_i(K)$, where $c_i(K)$ denotes a constant in K with i the index on the corresponding grid. We have that $x = y$ is equivalent to having $\Delta + c_i(K) = y$, which has probability 0 as Δ is uniformly distributed. Taking a union bound both on the grid points and the atoms for F, we conclude that for any grid point the probability of having a grid point corresponding to an atom for F is 0. □

Lemma 4.1. *Let (A_t, B_t) be a price pair with $B_t \leq A_t$, played at time step t. Then, for any point $x \in (\mathcal{G}_A \cup \mathcal{G}_B) \cap [B_t, A_t]$, it is possible to reconstruct $\ell_t^s(x),\ s \in \{A, B\}$, after playing (A_t, B_t).*

Proof. If $V_t \in (B_t, A_t]$ the statement follows trivially by the fact that we observe (M_t, V_t) entirely.

If a purchase occurred and $x \in \mathcal{G}_B$, then $V_t \leq B_t \leq x$: therefore $u_t^B(x_B) = M_t - x_B$. If instead we consider a point on the other grid $y \in \mathcal{G}_A \cap [B_t, A_t]$, then $V_t \leq B_t < y$ and $u_t^A(y) = 0$, therefore we can again reconstruct the feedback. The same reasoning can be mirrored in case of a sale, thus we conclude the proof. □

After proving the consistency of our estimators, we now move on to the proof of Lemma 4.2. This lemma presents the first step for bounding in the regret analysis.

Lemma 4.2. *For any time step $t \in \{1, \dots, T\}$, with $T \geq 2$, it holds:*

$$\sum_{x \in \mathcal{G}_A} \left(\frac{q_t(x)}{Q_t(x)} \right) + \sum_{x \in \mathcal{G}_B} \left(\frac{p_t(x)}{Q_t(x)} \right) \leq 4 \left(1 + \log \left(\frac{2(K+1)}{\gamma} \right) \right).$$

Consequently, it holds:

$$\sum_{x \in \mathcal{G}_s,\, s \in \{A,B\}} \frac{w_t^s(x)}{Q_t(x) + \beta} \leq \frac{4}{1-\gamma} \left(1 + \log \left(\frac{2(K+1)}{\gamma} \right) \right). \tag{5}$$

Proof. For the sake of this proof, we fix a unique indexing on the union of the two grids $\mathcal{G}_A \cup \mathcal{G}_B$, with a total of $2(K+1)$ points, denoted as $z_1 \leq \cdots \leq z_N$. We impose that the vectors p_t, q_t on this common grid have value 0 when outside their specific domain: that is, $p_t(z) = 0$ if $z \in \mathcal{G}_A$ and $q_t(z) = 0$ if $z \in \mathcal{G}_B$. Let $S_t(z_i) = \sum_{j \leq i} \frac{p_t(z_j) + q_t(z_j)}{2}$ be the sum of half the probabilities in each point on the grid until point $z_i \in \mathcal{G}_A \cup \mathcal{G}_B$. Recall Equation (4): we observe that, selecting a point $z_i \in \mathcal{G}_A \cup \mathcal{G}_B$, it holds that $\sum_{j<i} p_t(z_j) + \sum_{j<i} q_t(z_j) = 2S_t(z_{i-1})$ and $\sum_{j>i} p_t(z_j) + \sum_{j>i} q_t(z_j) = 2\left(1 - S_t(z_i)\right)$ with our new notation. Thus we have that, from Identity (4):

$$\begin{aligned} Q_t(z_i) &\geq 1 - S_t^2(z_{i-1}) - (1 - S_t(z_i))^2 \\ &= 2S_t(z_i)(1 - S_t(z_{i-1})) - \left(\frac{p_t(z_i) + q_t(z_i)}{2} \right)^2 \\ &\geq S_t(z_i)(1 - S_t(z_{i-1})). \end{aligned}$$

By using the standard analysis inequality $ab \leq (\frac{a+b}{2})^2$ and the fact that:

$$S_t(z_i)\left(1 - S_t(z_{i-1})\right) \geq \left(\frac{p_t(z_i) + q_t(z_i)}{2} \right)^2,$$

since $\left(\frac{p_t(z_i) + q_t(z_i)}{2} \right) = S_t(z_i) - S_t(z_{i-1})$ is no larger than both $S_t(z_i)$ and $1 - S_t(z_{i-1})$.

Substituting into the desired terms, we deduce:

$$\begin{aligned} \sum_{i=1}^{N} \frac{p_t(z_i) + q_t(z_i)}{Q_t(z_i)} &\leq \sum_{i=1}^{N} \frac{p_t(z_i) + q_t(z_i)}{S_t(z_i)(1 - S_t(z_{i-1}))} \\ &\leq \sum_{i=1}^{N} \frac{p_t(z_i) + q_t(z_i)}{S_t(z_i)} + \sum_{i=1}^{N} \frac{p_t(z_i) + q_t(z_i)}{1 - S_t(z_{i-1})} \\ &= 2\sum_{i=1}^{N} \frac{S_t(z_i) - S_t(z_{i-1})}{S_t(z_i)} + 2\sum_{i=1}^{N} \frac{S_t(z_i) - S_t(z_{i-1})}{1 - S_t(z_{i-1})}. \end{aligned}$$

Which holds by applying definitions and using that $1/(ab) \leq 1/a + 1/b$ with $a + b \geq 1$. We now conclude the argument by using on the deduced two sums that $1 - 1/a \leq \log(a)$ and telescoping the two series from above:

$$\sum_{i=1}^{N} \frac{p_t(z_i) + q_t(z_i)}{Q_t(z_i)} \leq 2 \left(1 + \log \left(\frac{2}{p_t(z_1) + q_t(z_1)} \right) + 1 + \log \left(\frac{2}{p_t(z_N) + q_t(z_N)} \right) \right).$$

Observing that by definition $\frac{p_t(z_i)+q_t(z_i)}{2} \geq \frac{\gamma}{N} = \frac{\gamma}{2(K+1)}$ for every point z_i in the union of the two grids, we have proved the first part of the thesis. Inequality (5) follows by considering that $w_t^B(z_i) \leq \frac{p_t(z_i)}{1-\gamma}$ and $w_t^A(z_i) \leq \frac{q_t(z_i)}{1-\gamma}$ and substituting this inequality in the first claim of the lemma.

We define for the following proofs the factor in Equation (5) as:

$$D := \frac{4}{1-\gamma}\left(1+\log\left(\frac{2(K+1)}{\gamma}\right)\right),$$

to simplify later derivations. □

Before proving Theorem 4.3, let us state some intermediate steps Lemma B.2, Lemma B.3 and Lemma B.4, providing high probability bounds on the conditional mean and variance of the estimated losses at each time step, given the data from the previous rounds $\mathcal{H}_{t-1} = \{\Delta, (A_1, B_1), (V_1, M_1), \ldots, (V_{t-1}, M_{t-1})\}$.

In the following Lemma B.2, we bound the cumulative difference between the estimated loss of the two current algorithms, denoted as Z_t, and that of any fixed bid–ask pair $(a, b) \in \mathcal{G}_A \times \mathcal{G}_B$, up to the usual logarithmic and second-order terms, denoted as Y_t.

Lemma B.2. *Let Z_t, Y_t be defined as:*

$$Z_t = \sum_{x\in\mathcal{G}_s,\, s\in\{A,B\}} w_t^s(x)\hat{\ell}_t^s(x), \qquad Y_t = \sum_{x\in\mathcal{G}_s,\, s\in\{A,B\}} w_t^s(x)(\hat{\ell}_t^s(x))^2.$$

Then we have that, for any $b \in \mathcal{G}_B$, $a \in \mathcal{G}_A$:

$$\sum_{t=1}^{T}\left(Z_t - \left(\hat{\ell}_t^B(b) + \hat{\ell}_t^A(a)\right)\right) \leq \frac{2\log(K+1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} Y_t.$$

Proof. The proof follows by manipulation of the weights W_t^s. Fix a time step $t \leq T$ and let $s \in \{A, B\}$. First, consider that:

$$\frac{W_{t+1}^s(x)}{\sum_{y\in\mathcal{G}_s} W_t^s(y)} = \frac{W_t^s(x)}{\sum_{y\in\mathcal{G}_s} W_t^s(y)}\exp\left(-\eta\hat{\ell}_t^s(x)\right) = w_t^s(x)\exp\left(-\eta\hat{\ell}_t^s(x)\right).$$

Therefore, if we sum the equality above over the grid points $x \in \mathcal{G}_s$, we have:

$$\begin{aligned}\frac{\sum_{x\in\mathcal{G}_s} W_{t+1}^s(x)}{\sum_{y\in\mathcal{G}_s} W_t^s(y)} &= \sum_{x\in\mathcal{G}_s} w_t^s(x)\exp\left(-\eta\hat{\ell}_t^s(x)\right)\\ &\leq \sum_{x\in\mathcal{G}_s} w_t^s(x)\left(1-\eta\hat{\ell}_t^s(x)+\frac{\eta^2}{2}(\hat{\ell}_t^s(x))^2\right)\\ &= 1-\eta\sum_{x\in\mathcal{G}_s} w_t^s(x)\hat{\ell}_t^s(x)+\frac{\eta^2}{2}\sum_{x\in\mathcal{G}_s} w_t^s(x)(\hat{\ell}_t^s(x))^2,\end{aligned}$$

where we used that $\exp(-x) \leq 1 - x + \frac{x^2}{2}$. After taking the logarithm on both sides and employing that $\log(1+x) \leq x$:

$$\log\left(\frac{\sum_{x\in\mathcal{G}_s} W_{t+1}^s(x)}{\sum_{y\in\mathcal{G}_s} W_t^s(y)}\right) \leq -\eta\sum_{x\in\mathcal{G}_s} w_t^s(x)\hat{\ell}_t^s(x)+\frac{\eta^2}{2}\sum_{x\in\mathcal{G}_s} w_t^s(x)(\hat{\ell}_t^s(x))^2,$$

and by summing over the time steps $t = 1, \ldots, T$ and $s \in \{A, B\}$, and telescoping, we deduce:

$$\log\left(\sum_{y\in\mathcal{G}_A} W_{T+1}^A(y)\right)+\log\left(\sum_{y\in\mathcal{G}_B} W_{T+1}^B(y)\right)-2\log(K+1) \leq \sum_{t=1}^{T}\left(-\eta Z_t+\frac{\eta^2}{2}Y_t\right).$$

where we used $\sum_{y\in\mathcal{G}_s} W_1^s(y) = K+1$, $s \in \{A, B\}$ and the definitions of Z_t and Y_t. We conclude the argument observing that, fixing an $s \in \{A, B\}$ and $x \in \mathcal{G}_s$, $\sum_{y\in\mathcal{G}_s} W_{T+1}^s(y) \geq W_{T+1}^s(x) = \exp\left(-\eta\sum_{t=1}^{T}\hat{\ell}_t^s(x)\right)$, since it represents the last update, and substituting this inequality into the last step of the previous chain of inequalities. □

We now provide a high probability bound on the difference between the estimated losses $\hat{\ell}$ and the real ℓ.

Lemma B.3. *Consider the realized mean loss a time t and point $x \in \mathcal{G}_B$, defined as $\bar{\ell}_t^B(x) := \frac{1-(M_t-x)F(x)}{2}$ and for $y \in \mathcal{G}_A$ as $\bar{\ell}_t^A(y) := \frac{1-(y-M_t)(1-F(y))}{2}$.*

For any $x \in \mathcal{G}_s$, with $s \in \{A, B\}$ and for all $u > 0$ it holds that:

$$\mathbb{P}\left(\sum_{t=1}^{T}\left(\hat{\ell}_t^s(x) - \bar{\ell}_t^s(x)\right) \geq \frac{u}{\beta}\right) \leq \exp(-u).$$

Hence, setting $\delta \in (0,1)$, with probability $1-\delta/3$ we have that for any $z \in \mathcal{G}_s$:

$$\sum_{t=1}^{T}\left(\hat{\ell}_t^s(x) - \bar{\ell}_t^s(x)\right) \leq \frac{1}{\beta}\log\left(\frac{6(K+1)}{\delta}\right).$$

Proof. Let $x \in \mathcal{G}_s$ and fix a time step t. We observe that $O_t(x)$ is distributed as a Bernoulli with parameter $Q_t(x)$, given the history $\mathcal{H}_{t-1}$ and loss $\ell_t^s(x)$. By taking the exponential of the estimator $\hat{\ell}_t^s(x)$, we can therefore bound:

$$\begin{aligned}\mathbb{E}_{O_t(x)}\left[\exp\left(\frac{\beta O_t(x)\ell_t^s(x)}{Q_t(x)+\beta}\right) \mid \mathcal{H}_{t-1}, \ell_t^s(x)\right] &= 1 - Q_t(x) + Q_t(x)\exp\left(\frac{\beta\,\ell_t^s(x)}{Q_t(x)+\beta}\right)\\ &\leq 1 + \frac{\beta\,Q_t(x)\,\ell_t^s(x)}{Q_t(x)+\beta-\beta\ell_t^s(x)}\\ &\leq 1 + \beta\,\ell_t^s(x).\end{aligned}$$

Where we used that, for $x \in [0,1)$, $\exp(x) \leq 1 + \frac{x}{1-x}$, then that $\frac{,Q_t(x)}{Q_t(x)+\beta-\beta\ell_t^s(x)} \leq 1$.

By the tower rule on $\ell_t^s(z)$, we can rewrite the expected value $\mathbb{E}\left[\exp\left(\hat{\ell}_t^s(x)\right) \mid \mathcal{H}_{t-1}\right]$:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{O_t(x)}\left[\exp\left(\beta\hat{\ell}_t^s(x)\right) \mid \mathcal{H}_{t-1}, \ell_t^s(x)\right] \mid \mathcal{H}_{t-1}\right] &\leq 1 + \beta\,\bar{\ell}_t^s(x)\\ &\leq \exp\left(\beta\,\bar{\ell}_t^s(x)\right),\end{aligned}$$

where the last inequality follows by $1 + x \leq \exp(x)$.

Hence, the process $\exp\left(\beta\sum_{\tau=1}^{t}\left(\hat{\ell}_\tau^s(x) - \bar{\ell}_\tau^s(x)\right)\right)$ is a nonnegative supermartingale that has an initial value of 1.

The first part of the thesis is deduced by applying Markov's inequality to the difference of the sums $\sum_{t=1}^{T}\left(\hat{\ell}_t^s(x) - \bar{\ell}_t^s(x)\right)$ and applying the deduced inequality in terms of the exponential. That is,

$$\begin{aligned}&\mathbb{P}\left(\sum_{t=1}^{T}\left(\hat{\ell}_t^s(x) - \bar{\ell}_t^s(x)\right) \geq \frac{u}{\beta}\right) =\\ &= \mathbb{P}\left(\exp\left(\beta\sum_{t=1}^{T}\left(\hat{\ell}_t^s(x) - \bar{\ell}_t^s(x)\right)\right) \geq \exp(u)\right) \leq \frac{1}{\exp(u)}.\end{aligned}$$

The second part of the claim is obtained building on the first, by a union bound on the $2(K+1)$ total grid points and where we select $u = \log\left(\frac{6(K+1)}{\delta}\right)$. □

Lemma B.4. *Consider $\eta \in (0, \beta]$ and recall the definition of Z_t and Y_t given in Lemma B.2. Then, it holds that $0 \leq Z_t - \frac{\eta}{2}Y_t \leq D$, and $\mathbb{E}[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}] \leq 2$. Moreover:*

$$\sum_{x\in\mathcal{G}_s, s\in\{A,B\}} w_t^s(x)\,\bar{\ell}_t^s(x) - \mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] \leq \left(\beta + \frac{\eta}{2}\right)D.$$

Furthermore, for any $u > 0$:

$$\mathbb{P}\left(\sum_{t=1}^{T}\left(\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] - \left(Z_t - \frac{\eta}{2}Y_t\right)\right) > 2\sqrt{DTu}\right) \leq \exp(-u),$$

and for every $x \in \mathcal{G}_B$ and $y \in \mathcal{G}_A$ it holds that:

$$\sum_{t=1}^{T}\left(Z_t - \frac{\eta}{2}Y_t\right) - \sum_{t=1}^{T}\left(\hat{\ell}_t^B(x) + \hat{\ell}_t^A(y)\right) \leq \frac{2\log(K+1)}{\eta}.$$

Proof. Let $H_t = \sum_{x\in\mathcal{G}_B} p_t(x)\bar{\ell}_t^B(x) + \sum_{x\in\mathcal{G}_A} q_t(x)\bar{\ell}_t^A(x)$ for readability.

We first notice that:

$$\left|\sum_{x\in\mathcal{G}_B}\left(w_t^B(x) - p_t(x)\right)\bar{\ell}_t^B(x) + \sum_{x\in\mathcal{G}_A}\left(w_t^A(x) - q_t(x)\right)\bar{\ell}_t^A(x)\right| \leq 2\gamma,$$

because of the definition of the probabilities in Equation (3) and summing over $K+1$ points.

It also holds that:

$$H_t \geq \mathbb{E}[\bar{\ell}_t^A(A_t) + \bar{\ell}_t^B(B_t) \mid \mathcal{H}_{t-1}], \qquad H_t = \mathbb{E}[\bar{\ell}_t^A(Y_t) + \bar{\ell}_t^B(X_t) \mid \mathcal{H}_{t-1}].$$

The inequality in the first expression derives from the fact that we might swap the prices (Y_t, X_t) initially selected by Algorithm 3 using Lemma 3.1,for which we weakly decrease the loss even if in the step t we had swapped the prices.

We consider the expected value for the losses estimates. It holds, by definition:

$$\mathbb{E}[\hat{\ell}_t^s(x) \mid \mathcal{H}_{t-1}] = \frac{Q_t(x)}{Q_t(x)+\beta}\bar{\ell}_t^s(x),$$
$$\mathbb{E}[(\hat{\ell}_t^s(x))^2 \mid \mathcal{H}_{t-1}] = \frac{Q_t(x)}{(Q_t(x)+\beta)^2}\mathbb{E}[\ell_t^s(x)^2 \mid \mathcal{H}_{t-1}] \leq \frac{1}{Q_t(x)+\beta}.$$

Thus, follows that $\sum_{x\in\mathcal{G}_s, s\in\{A,B\}} w_t^s(x)\bar{\ell}_t^s(x) - \mathbb{E}[Z_t \mid \mathcal{H}_{t-1}] \leq \beta D$ and $\mathbb{E}[Y_t \mid \mathcal{H}_{t-1}] \leq D$, from Lemma 4.2.

Notice that, for any time step t and $x \in \mathcal{G}_s$, $\hat{\ell}_t^s(x) \in [0, \frac{1}{\beta}]$: so $0 \leq \hat{\ell}_t^s(x)\left(1 - \frac{\eta}{2}\hat{\ell}_t^s(x)\right) \leq \hat{\ell}_t^s(x)$. Therefore, by recalling Lemma B.2, we have that $0 \leq Z_t - \frac{\eta}{2}Y_t \leq Z_t \leq D$. By taking the expectation:

$$\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] \leq \mathbb{E}\left[Z_t \mid \mathcal{H}_{t-1}\right] \leq \sum_{x\in\mathcal{G}_s, s\in\{A,B\}} w_t^s(x)\bar{\ell}_t^s(x) \leq 2.$$

We deduce the second claim from Lemma B.2. We also notice that $\mathbb{E}[\left(Z_t - \frac{\eta}{2}Y_t\right)^2 \mid \mathcal{H}_{t-1}] \leq D\mathbb{E}[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}] \leq 2D$.

Therefore, $\forall\lambda \geq 0$ and using $\exp(-x) \leq 1 - x + \frac{x^2}{2}$ and $1 + x \leq \exp(x)$, we have that:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\left(\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] - \left(Z_t - \frac{\eta}{2}Y_t\right)\right)\right) \mid \mathcal{H}_{t-1}\right] \\
&\leq \exp\left(\lambda\,\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right]\right)\left(1 - \lambda\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] + \frac{\lambda^2}{2}\mathbb{E}\left[\left(Z_t - \frac{\eta}{2}Y_t\right)^2 \mid \mathcal{H}_{t-1}\right]\right) \\
&\leq \exp\left(\frac{\lambda^2}{2}\mathbb{E}\left[\left(Z_t - \frac{\eta}{2}Y_t\right)^2 \mid \mathcal{H}_{t-1}\right]\right) \\
&\leq \exp\left(\lambda^2 D\right).
\end{aligned}$$

By summing over the t and applying Markov's inequality with $\lambda = \sqrt{u/TD}$ we deduce the third claim, while the last claim is a reordering of Lemma B.2. □

We finally prove the full Theorem 4.3 statement, using the results from the previous Lemma.

Theorem 4.3. *Let $\delta \in (0,1)$ and $T \geq 2$. There exist $\eta = \beta = \Theta(\sqrt{1/T})$ such that Algorithm 3 instantiated with $K = \lceil \sqrt{T} \rceil$, and $\gamma = 1/T$ attains:*

$$R_T = \mathcal{O}\left(\sqrt{T}\log\left(\frac{T}{\delta}\right)\right),$$

with probability at least $1-\delta$.

Proof. Recall H_t as defined in Lemma B.4 and Z_t, Y_t as in Lemma B.2. We first observe that by conditional Hoeffding's bound and the observations in Lemma B.4 for all time steps t:

$$\mathbb{E}\left[\exp\left(\lambda(\bar{\ell}_t^A(Y_t) + \bar{\ell}_t^B(X_t) - H_t\right) \mid \mathcal{H}_{t-1}\right] \leq \exp\left(\lambda^2/2\right).$$

Then with a union bound and Markov inequality:

$$\mathbb{P}\left(\sum_{t=1}^{T}\left(\bar{\ell}_t^A(Y_t) + \bar{\ell}_t^B(X_t) - H_t\right) \geq \sqrt{2Tu}\right) \leq \exp(-u).$$

To get the bound, we use the results in the previous Lemma B.2, Lemma B.3 and Lemma B.4 by adding and subtracting the terms defined there.

We consider as a comparator action $(a,b) \in \mathcal{G}_A \times \mathcal{G}_B$. This is justified as we notice that an identical construction to the one in Lemma 3.2 repeated in the case of $\mathcal{G}_A \times \mathcal{G}_B$ establishes that we can reduce the problem to our discrete set at the total cost of $\frac{5T}{K}$, accounting for the additional constants in the loss.

In particular fix any pair $(a,b) \in \mathcal{G}_A \times \mathcal{G}_B$: with probability at least $1-\delta$, which derives from summing up the failure probabilities from the previous lemmas:

$$\begin{aligned}
&\left[\sum_{t=1}^{T}\left(\bar{\ell}_t^A(A_t) + \bar{\ell}_t^B(B_t) - \bar{\ell}_t^A(a) - \bar{\ell}_t^B(b)\right)\right] \\
&\quad\leq \sum_{t=1}^{T}\left(\bar{\ell}_t^A(Y_t) + \bar{\ell}_t^B(X_t) - H_t\right) + \sum_{t=1}^{T}\left(H_t - \sum_{x\in\mathcal{G}_s, s\in\{A,B\}} w_t^s(x)\bar{\ell}_t^s(x)\right) \\
&\qquad+ \sum_{t=1}^{T}\left(\sum_{x\in\mathcal{G}_s, s\in\{A,B\}} w_t^s(x)\bar{\ell}_t^s(x) - \mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right]\right) \\
&\qquad+ \sum_{t=1}^{T}\left(\mathbb{E}\left[Z_t - \frac{\eta}{2}Y_t \mid \mathcal{H}_{t-1}\right] - Z_t + \frac{\eta}{2}Y_t\right) \\
&\qquad+ \sum_{t=1}^{T}\left(Z_t - \frac{\eta}{2}Y_t - \hat{\ell}_t^A(a) - \hat{\ell}_t^B(b)\right) + \sum_{t=1}^{T}\left(\hat{\ell}_t^A(a) + \hat{\ell}_t^B(b) - \bar{\ell}_t^A(a) - \bar{\ell}_t^B(b)\right) \\
&\quad\leq \sqrt{2T\log(3/\delta)} + 2\gamma T + \left(\beta + \frac{\eta}{2}\right)DT + 2\sqrt{DT\log(3/\delta)} \\
&\qquad+ \frac{2\log(K+1)}{\eta} + \frac{2\log(6(K+1)/\delta)}{\beta}.
\end{aligned}$$

By applying this bound to the regret decomposition, we obtain the general claim. We also used Lemma 3.1 in the first inequality.

Consequently, by selecting $K = \lceil\sqrt{T}\rceil$, $\gamma = 1/T$, $\eta = \beta = \sqrt{\frac{\log(6(K+1)/\delta)}{DT}}$, we have $D = \frac{4(1+\log(2(K+1)T))}{1-1/T}$ and deduce the optimal rate $R_T \leq \tilde{\mathcal{O}}(\sqrt{T})$. In particular, with probability $1-\delta$ it holds that $R_T = \mathcal{O}\left(\sqrt{T}\log\left(\frac{T}{\delta}\right)\right)$, which concludes the proof. □

Table 2: Upper bound on $u_t(A_t, B_t)$ for each case of the random sequence

(A_t, B_t) \ (V_t, M_t)	$(L_{t-1}, 1)$	$(y, 1)$	$(x, 0)$	$(U_{t-1}, 0)$	**Upper bound**
$A_t \in [L_{t-1}, x), B_t \in [L_{t-1}, x)$,	$1 - B_t$	$A_t - 1$	A_t	A_t	$\frac{1}{3}$
$A_t \in [x, y), B_t \in [L_{t-1}, x)$	$1 - B_t$	$A_t - 1$	0	A_t	$\frac{1}{4}$
$A_t \in [y, U_{t-1}], B_t \in [L_{t-1}, x)$	$1 - B_t$	0	0	A_t	$\frac{1}{3}$
$A_t \in [x, y), B_t \in [x, y)$	$1 - B_t$	$A_t - 1$	$-B_t$	A_t	$\frac{1}{18}$
$A_t \in [y, U_{t-1}], B_t \in [x, y)$	$1 - B_t$	0	$-B_t$	A_t	$\frac{1}{3}$
$A_t \in [y, U_{t-1}], B_t \in [y, U_{t-1}]$	$1 - B_t$	$1 - B_t$	$-B_t$	A_t	$\frac{1}{3}$

C PROOF FOR THE IMPOSSIBILITY RESULT

In this final section, we give a proof of the unlearnability result for market making, under the favorable condition of full feedback and oblivious adversary.

Theorem 5.1. *Consider the market making problem with adversarial sequence of $\{M_t\}_t$ and $\{V_t\}_t$ and full feedback. Then, for any algorithm $\mathcal{A}$, there exists a sequence $\{(M_t, V_t)\}_{t=1}^T$ where expected regret $R_T \geq \frac{T}{6}$.*

Proof. To prove the statement, we rely on Yao's minimax principle (Yao, 1977): we construct a random instance of the problem and prove that any deterministic bid-ask sequence suffers linear regret with such sequence.

Consider the following construction for a sequence bounded in $[0, 1]$ $(L_t, U_t)_t$. Let $D_t = U_t - L_t$ at any given time step t. We initialize $(L_0, U_0) = (1/3, 2/3)$, then for any $t = 1, \dots, T$:

$$\begin{cases} (L_t, U_t) = (L_{t-1}, U_{t-1} - \frac{2D_{t-1}}{3}) & \text{with probability } \frac{1}{2} \\ (L_t, U_t) = (L_{t-1} + \frac{2D_{t-1}}{3}, U_{t-1}) & \text{with probability } \frac{1}{2} \end{cases}$$

Given this sequence, we construct the instance of prices $\{(V_t, M_t)\}_t$ as follows:

$$\begin{cases} (V_t, M_t) = (L_t, 1), & \text{with probability } \frac{1}{2}, \\ (V_t, M_t) = (U_t, 0), & \text{with probability } \frac{1}{2}, \end{cases}$$

where the sequence is drawn independently of the previous one.

We now consider an optimal in hindsight action for such instance of the problem. We notice that there exists a (a^*, b^*) where $a^*, b^* \in [L_T, U_T]$ such that the learner always wins. Defining a constant $\varepsilon \leq \frac{D_T}{2}$, we can take such action as $(a^*, b^*) = (U_T - \varepsilon, L_T)$. The total expected revenue with respect to the randomization of the environment is:

$$\mathbb{E}\left[\sum_{t=1}^{T} u_t(a^*, b^*)\right] = T\frac{1 - L_T + U_T - \varepsilon}{2} = T\frac{1 + D_T - \varepsilon}{2} \geq \frac{T}{2} + \frac{T\,D_T}{4}$$

where we imposed $\varepsilon = \frac{D_T}{2}$ and $D_T = \frac{1}{3^{T+1}}$ and vanishes as T grows.

We now turn to the expected revenue for any deterministic sequence. For any time step $A_t > U_t$ or $B_t < L_t$ gains 0 revenue for the corresponding utility, while having $A_t \leq L_t$ or $B_t \geq U_t$ gives negative revenue in the corresponding problem. With this observation, it only makes sense to consider the cases of bid-ask prices $A_t \in (L_{t-1}, U_{t-1}]$ and $B_t \in [L_{t-1}, U_{t-1})$. We have six cases to consider, depending on where the bid ask prices lie in the three intervals $(L_{t-1}, L_{t-1} + \frac{D_{t-1}}{3})$, $(L_{t-1} + \frac{D_{t-1}}{3}, L_{t-1} + \frac{2D_{t-1}}{3})$ and $(L_{t-1} + \frac{2D_{t-1}}{3}, U_{t-1})$ and given our general constraint $A_t \geq B_t$. For simplicity, we call $x = L_{t-1} + \frac{D_{t-1}}{3}$ and $y = L_{t-1} + \frac{2D_{t-1}}{3}$ in Table 2. We proceed by computing the expected revenue given each event in the original sequence and summing up the

resulting revenue for each case, thus obtaining the general expected revenue for a deterministic sequence of prices. We summarize the results for each case with Table 2.

Overall, the expected utility for a deterministic algorithm playing $(A_t, B_t) \in \mathcal{U}$ is $\mathbb{E}\left[u_t(A_t, B_t)\right] \leq \frac{1}{3}$, given our construction.

Therefore by Yao's principle that for any probabilistic algorithm $\{(A_t, B_t)\}_{t=1}^{T}$:

$$R_T \geq \frac{T}{2} - \frac{T}{3} = \frac{T}{6},$$

which proves the statement, by Yao's principle. □