Online Market Making and the Value of Observing the Order Book

Davide Maran DAVIDE.MARAN@POLIMI.IT
Politecnico di Milano, Piazza Leonardo da Vinci, 32-36 - Città Studi, Milano (MI)

Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Politecnico di Milano, Piazza Leonardo da Vinci, 32-36 - Città Studi, Milano (MI)

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study an online market-making problem in which a learner sequentially posts bid and ask prices for a single asset while interacting with traders holding private valuations. Unlike existing online learning formulations that assume fully censored feedback, we introduce an action-dependent feedback model inspired by real limit order books: when a trade occurs, the trader's valuation remains hidden, whereas when no trade occurs, informative feedback about supply and demand is revealed. We show that this additional information fundamentally changes the learnability of the problem. In the stochastic setting with i.i.d. market prices, we propose an elimination-based algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with high probability, without requiring any smoothness assumptions on the distribution of trader valuations. We then extend this result to a broad class of mean-reverting price processes by considering both local, autoregressive dynamics and a weaker global drift condition based on cumulative deviations from the mean. Under either assumption, we establish high-probability $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bounds, relying on a new concentration inequality of independent interest. Finally, in the adversarial setting with oblivious prices, we design an explore-then-perturb algorithm that guarantees $\tilde{\mathcal{O}}(T^{2/3})$ regret in expectation. Our results quantify the value of observing the order book in online market making and demonstrate that even limited, action-dependent feedback can substantially improve regret guarantees compared to standard bandit feedback models.

Keywords: Online learning; Bandits with partial feedback; Action-dependent feedback; Elimination algorithms; Mean-reverting processes

1. Introduction

Market making is the activity performed by intermediaries who provide liquidity to an asset by simultaneously quoting buying prices (bid) and selling prices (ask). This function is essential for the efficiency of financial markets, as it reduces transaction costs and facilitates the immediate matching of supply and demand (Amihud and Mendelson, 1986). Without market makers, investors would face wider spreads and increased price volatility (Glosten and Harris, 1988; Madhavan, 2000).

In the context of online learning, market making is framed as an iterative game between the agent, "maker", and a taker (which models the rest of the market). The agent decides sequentially one bid/ask pair B_t, A_t, without knowing the taker's private valuation V_t or the future market value of the asset M_t, aiming to minimize regret relative to the best fixed strategy in hindsight. This perspective introduces a fundamental exploration-exploitation trade-off, where the agent must balance learning the latent distribution of trader valuations with the goal of maximizing immediate profit. Previous work in this domain (Cesa-Bianchi et al., 2024) focused on a restricted feedback model where the agent only observes the market value M_t and a binary indicator of whether a transaction

© 2026 D. Maran & M. Restelli.

occurred (i.e., $V_t \leq B_t$ or $V_t \geq A_t$). Under such limited information, only weak regret guarantees, typically sublinear only under restrictive assumptions, can be achieved.

This feedback accurately reflects the fact that, whenever the agent makes a deal, it is not possible to understand which was the true valuation V_t of the taker. Nonetheless, we observe that, in the opposite case, when no deal happens, it is fair to assume the V_t is revealed to the learner. In fact, this assumption aligns with the operational reality of modern electronic exchanges, where the Limit Order Book (LOB) serves as a public ledger of intent. While a completed transaction only reveals that a price was met, the absence of a trade at the maker's spread allows the agent to observe the surrounding "resting" limit orders, which explicitly represent the buy and sell valuations of other market participants. Observing the book of limit orders is vital for every modern market maker. This information is so valuable that industry leaders such as Jane Street, Citadel Securities, and Hudson River Trading invest hundreds of millions of dollars annually in high-speed, "Level 3" market data feeds to gain full visibility into these unexecuted orders. Global spending on financial market data reached a record $44 billion in 2024 (Burton-Taylor International Consulting, 2025), reflecting the industry's consensus that observing the order book is not a luxury, but a fundamental requirement for effective market making.

1.1. Original contribution

The main contribution of this paper is the introduction of a novel action-dependent feedback model for online market making that closely reflects the information structure of real limit order books. Unlike classical formulations, where feedback is either fully censored (bandit feedback) or fully revealed (full feedback), we propose a feedback mechanism in which the information received by the learner depends continuously on the chosen bid–ask spread. This feedback structure captures a fundamental feature of real-world market making: information about demand and supply is revealed precisely when liquidity is *not* consumed. From an Online Learning perspective, this setting gives rise to a previously unexplored regime in which: (i) the feedback is partial and censored, (ii) the type of feedback depends on the learner's action, and (iii) informative feedback is obtained exactly when the instantaneous reward is zero. To the best of our knowledge, this is the first formalization of such a feedback structure in the field of Online Learning. Building on this model, we establish regret guarantees in three progressively more challenging environments:

1. Stochastic prices. When the market price process is i.i.d. with unknown mean, we design an algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with high probability.

2. Mean-reverting prices. We extend the stochastic analysis to a broad class of mean-reverting price processes, significantly relaxing the i.i.d. assumption. Under a mild martingale-type condition, we show that the same $\widetilde{\mathcal{O}}(\sqrt{T})$ regret rate can still be achieved. This result is based on a simple yet novel concentration inequality of independent interest.

3. Adversarial prices. When market prices are allowed to be an arbitrary oblivious sequence, we propose an explore-then-perturb algorithm that guarantees $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$ regret in expectation.

In contrast, lower bounds for the setting with bandit feedback (Cesa-Bianchi et al., 2024) show that a regret of $\Omega(T^{\frac{2}{3}})$ in the stochastic case, and of $\Omega(T)$ in the adversarial unless a Lipschitz condition on the c.d.f. of V_t is met. Our results quantify the value of observing the order book in online market making and show that even a limited, action-dependent form of feedback can dramatically improve theoretical guarantees.

2. Setting

In this paper, we model the market-making problem as a discrete-time online learning interaction between an agent (the "maker") and a sequence of "takers". Formally, the Online Market Making setting is defined by two main components: first, a sequence of takers who, at each round $t = 1, \dots, T$, hold a private valuation V_t for a single unit of the underlying asset; second, a sequence of market prices M_t representing the asset's objective value at the end of each round. At each round, the market maker selects a pair of bid and ask prices (B_t, A_t) from the available action space, without prior knowledge of either V_t or M_t. Consequently, three scenarios may arise:

- **Maker Buys:** If $V_t \leq B_t$, the taker sells the asset to the maker at price B_t, resulting in a reward of $M_t - B_t$ for the agent.
- **No Trade:** If $B_t < V_t < A_t$, no transaction occurs, and the agent receives a reward of 0.
- **Maker Sells:** If $V_t \geq A_t$, the taker purchases the asset from the maker at price A_t, yielding a reward of $A_t - M_t$ for the agent.

Importantly, in real markets, the agent does not interact with a single taker at a time, and no taker is willing to buy or sell at the same V_t. Focusing on a single private valuation V_t per round is a necessary simplification, somewhat close to the Glosten-Milgrom model (Das*, 2005; Touzo et al., 2021), which serves as a powerful abstraction that captures the idea of "best opportunity of the market". In mathematical terms, the reward function of the agent writes as

$$r(b, a; v, m) := \mathbb{1}(v \leq b)(m - b) + \mathbb{1}(a \leq v)(a - m) \tag{1}$$

for $v = V_t$ and $m = M_t$. The following assumptions, introduced by Cesa-Bianchi et al. (2024), will be always considered:

Assumption 1 (Bounded price) *At any time step $V_t, M_t \in [0, 1]$.*

Assumption 2 (Stochastic evaluation) *The evaluations V_t are sampled from a probability distribution with c.d.f. $F : [0, 1] \to [0, 1]$ independently at any time step. By convenience, we always call $S(x) := 1 - F(x)$ the corresponding survival function.*

From Assumption 1, it follows that the agent has no need to choose any pair (b, a) outside of the following set $\mathcal{U} := \{b, a : 0 \leq b < a \leq 1\}$. The former thus defines the agent's action space. We do not yet need explicit assumptions on M_t. Different results apply when M_t is stochastic, i.i.d., and independent of V_t (stochastic case) versus when it is an arbitrary sequence, as shown by Cesa-Bianchi et al. (2024). Our main novelty relative to that paper is the following "order book access" assumption.

Assumption 3 (Order book access) *At the end of the round, if $\boxed{B_t < V_t < A_t}$ the agent observes V_t, M_t, otherwise only M_t.*

In other words, *information arrives exactly when reward does not*. The former assumption reflects a real problem of market makers: if V_t falls outside of (B_t, A_t), the deal occurs at one of the extrema, and the agent has no possibility to infer the value at which the taker was willing to buy or sell; if V_t falls within the spread, its value becomes observable through the clearing of limit orders,

allowing the agent to capture the transaction price resulting from the interaction between other market participants. From an Online Learning perspective, this feedback structure is particularly noteworthy. Specifically, it yields full feedback when $V_t \in (B_t, A_t)$ and bandit feedback otherwise. Given that the agent receives zero reward whenever V_t falls within this interval, such a hybrid feedback mechanism significantly complicates the exploration-exploitation trade-off. The agent's goal is to minimize regret, measured w.r.t. the optimal bid-ask choice in hindsight. Calling

$$J_t(a, b) := F(b)(M_t - b) + S(a)(a - M_t), \tag{2}$$

Assumption 2 allows us to write the regret as follows

$$R_T = \sup_{b,a\in[0,1]} \mathop{\mathbb{E}}_{V_t} \left[\sum_{t=1}^{T} r(b, a, V_t, M_t) - r(B_t, A_t, V_t, M_t) \right] \tag{3}$$

$$= \sup_{b,a\in[0,1]} \sum_{t=1}^{T} J_t(b, a) - J_t(B_t, A_t). \tag{4}$$

Given the particular structure of J_t in Equation (2), the supremum of the sum which appears in Equation (3) may be written in closed form as

$$\sup_{b,a\in[0,1]} \frac{1}{T} \sum_{t=1}^{T} J_t(b, a) = \sup_{b,a\in[0,1]} (\widehat{\mu}_T - b)F(b) + (a - \widehat{\mu}_T)S(a), \qquad \widehat{\mu}_T = \frac{1}{T} \sum_{t=1}^{T} M_t. \tag{5}$$

As already noted, the former definition says nothing about the sequence M_t. In the following section, we are going to explore the theoretical guarantees that correspond to some assumptions on the market price. The following sections consider different scenarios in order of generality. First, the case of a sequence of independent random variables (section 3), then a generalization which introduces different forms of temporal correlation which formalize the concept of *mean reversion* (section 4) and, finally, the one when M_t is an arbitrary sequence which cannot adapt to the choices of the agent (section 5).

3. Stochastic Independent Market Prices

We begin by examining the case where M_t is an independent stochastic process. In this section, we assume the market price follows an independent process with unknown mean μ.

Assumption 4 (Stochastic independent market price) *The evaluations M_t are an independent sequence with mean μ.*

Under Assumption 4, the environment is fully stochastic. In this setting, the most natural approach is to estimate μ with a sample mean estimator $\widehat{\mu}_t$ and plan the bid-ask spreads in order to be a near-maximizer of a surrogate reward function which takes the form $\widetilde{F}(b)(\widehat{\mu}_t - b) + \widetilde{S}(a)(a - \widehat{\mu}_t)$.

Dealing with the estimator for the mean is the easy part. Indeed, by a relatively standard version of Azuma-Hoeffding's inequality that we show in appendix B, with probability at least $1 - \delta$,

$$\forall 1 \le t \le T \qquad |\widehat{\mu}_t - \mu| \le \sqrt{\frac{\log(3\log(T)/\delta)}{t}}. \tag{6}$$

Since the sample mean converges quickly to the true mean, in this setting there is essentially no difference between measuring the regret Equation (5) with respect to μ or $\widehat{\mu}_T$. In similar stochastic settings, one therefore often uses the following notion of regret (pseudo-regret) as the objective.

$$\mathcal{R}_T = T \cdot \sup_{b,a \in [0,1]} J(b,a) - \sum_{t=1}^{T} J(B_t, A_t) \qquad J(b,a) = F(b)(\mu - b) + S(a)(a - \mu). \tag{7}$$

In the stochastic setting, pseudo-regret is close to the regret with high probability, as formalized in Proposition 6. The other ingredient to perform this step is an estimator $\widehat{F}$ for the c.d.f. of V_t. Define the empirical c.d.f. as follows:

$$\widehat{F}_t(x) = \frac{\sum_{\tau=1}^{t} \mathbb{1}(V_\tau \leq x)}{t}. \tag{8}$$

By the Glivenko-Cantelli theorem (see Chapter 19 in (Van der Vaart, 2000)), this estimator is known to converge to F *uniformly*. Unfortunately, in our interaction $\widehat{F}_t$ cannot always be built, as the agent is only aware of $\text{clip}(V_t; B_t, A_t)$ at any step t, by Assumption 3. When playing a spread $A_t - B_t$ that is too narrow, the agent gathers no information to improve this estimate. On the other hand, when the agent receives information, i.e., $B_t < V_t < A_t$, no trade occurs; therefore, the agent receives no reward. To address these two problems, we will propose an algorithm that is both elimination-based, to ensure that the c.d.f. of all active prices is estimated at each time step, and optimistic, to ensure that the optimal pair a, b is not discarded with high probability.

3.1. Algorithm

The learner faces a tension between narrowing the spread to gain reward and widening it to gain information. OPSR (algorithm 1) resolves this by shrinking the action set only when confidence intervals certify suboptimality, while always playing the most conservative surviving quotes. This algorithm keeps in memory the value of the e.c.d.f. for all "interesting" prices, which need to be in a discrete set. Therefore, define the following uniform quantization

$$\mathcal{A} := \left\{ \frac{n}{\lceil \sqrt{T} \rceil} : n = 0, 1, \ldots, \lceil \sqrt{T} \rceil \right\}.$$

At any time-step, our algorithm is going to choose a pair $B_t, A_t \in \mathcal{A}$ and we will call $\mathcal{A}_t := \mathcal{A} \cup [B_t, A_t]$. While the agent cannot always access the value V_t, which is formally required to build eq. (8), the observation is sufficient to build this surrogate

$$V_t^{\text{clip}} := \begin{cases} B_t & V_t < B_t \\ V_t & B_t \leq V_t \leq A_t \\ A_t + 0.01 & V_t > A_t \end{cases} \qquad \chi_t(x) := \mathbb{1}(V_t^{\text{clip}} \leq x).$$

If $x \in [B_t, A_t]$, then $\{V_t^{\text{clip}} \leq x\}$ corresponds to $\{V_t \leq x\}$, whichever the value of V_t (in fact, the choice of 0.01 is arbitrary, any positive number would work for for purposes). Taking any price x that belongs to the whole sequence of spreads, the following proposition holds.

Proposition 1 *For any $t \in [T], x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$, $\widehat{F}_t(x) = \frac{1}{t}\sum_{\tau=1}^{t} \chi_\tau(x)$ and, for any $\delta > 0$,*

$$\mathbb{P}\left(\exists t \in [T],\ \exists x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau, \quad |\widehat{F}_t(x) - F(x)| > \sqrt{\frac{3\log(3T/\delta)}{4t}}\right) \leq \delta.$$

The former proposition shows that the estimated c.d.f. concentrates nicely, almost as fast as $\widehat{\mu}_t$ for $x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$. Outside of this set, the estimation of $\widehat{F}_t$ gets more difficult, if not impossible. Therefore, the idea of our algorithm is to let the sets $\mathcal{A}_t$ be monotonically decreasing, so that $\bigcap_{\tau=1}^{t} \mathcal{A}_\tau = \mathcal{A}_t$. In this way, the agent is guaranteed to have all the information available for all "interesting" prices at any time step. Once $x \notin \mathcal{A}_t$ at one time step, meaning $x < B_t$ or $x > A_t$, there is no chance this price is played again. This idea is the same as elimination-based algorithms (Even-Dar et al., 2006). Eliminating an arm is risky and may lead the agent to pay linear regret. Therefore, we define the following optimistic/pessimistic estimators for the c.d.f.

$$\widehat{F}_t^{\text{up}}(x) = \min\{\widehat{F}_{t-1}^{\text{up}}(x), \widehat{F}_t(x) + \psi(t,\delta)\}, \qquad \widehat{F}_t^{\text{low}}(x) = \max\{\widehat{F}_{t-1}^{\text{low}}(x), \widehat{F}_t(x) - \psi(t,\delta)\} \tag{9}$$

$$\psi(t,\delta) := \sqrt{\frac{3\log(3T/\delta)}{4t}}, \tag{10}$$

where $\widehat{F}_0^{\text{up}}(x) := 1$ and $\widehat{F}_0^{\text{low}}(x) := 0$. When dealing with asks instead of bids, it is more natural to talk about the survival function than the cumulative distribution. Indeed, to get its optimistic/pessimistic estimators, one just needs to define $\widehat{S}_t^{\text{up}}(x) = 1 - \widehat{F}_t^{\text{low}}(x)$ and $\widehat{S}_t^{\text{low}}(x) = 1 - \widehat{F}_t^{\text{up}}(x)$. The same can be done for μ, exploiting eq. (6),

$$\widehat{\mu}_t^{\text{up}} := \min\{\widehat{\mu}_{t-1}^{\text{up}}, \widehat{\mu}_t + \phi(t,\delta)\} \qquad \widehat{\mu}_t^{\text{low}} := \max\{\widehat{\mu}_{t-1}^{\text{low}}, \widehat{\mu}_t - \phi(t,\delta)\} \tag{11}$$

$$\phi(t,\delta) := \sqrt{\frac{\log(3\log(T)/\delta)}{t}}, \tag{12}$$

where $\widehat{\mu}_0^{\text{up}} := 1, \widehat{\mu}_0^{\text{low}} := 0$. The algorithm presented in algorithm 1 leverages all available estimators to eliminate sub-optimal bid-ask pairs until it converges to a quasi-optimal solution. Its scheme alternates between optimistic and pessimistic estimates for each element in $\mathcal{A}_{t-1}$. At the beginning of each round, optimistic estimates for $(x-\mu)F(x)$ and $(\mu - x)S(x)$ are computed in lines 6 and 7, respectively. Subsequently, pessimistic estimates are calculated for every element in $\mathcal{A}_{t-1}$ (line 8 and line 9), from which the respective maxima Γ^*_{bid} and Γ^*_{ask} are derived. Finally, the spread is narrowed in lines 13 and 14 by increasing B_t until an element satisfying $\Theta_{\text{bid}}(x) \geq \Gamma^*_{\text{bid}}$ is encountered, with A_t being decreased analogously. This last step is crucial: even if heavily sub-optimal bid-ask pairs remain within $\mathcal{A}_t$, they do not increase the regret, as the played values are always the extreme points of the interval.

This structure allows OPSR to have a regret guarantee under our assumptions. First, we show that these guarantees hold outside the following failure event. Define

$$E := \bigcup_{t=1}^{T} \left\{ \bigcup_{a \in \mathcal{A}_t} |\widehat{F}_t(x) - F(x)| > \psi(t,\delta) \right\} \cup \left\{ |\widehat{\mu}_t - \mu| > \phi(t,\delta) \right\}. \tag{13}$$

This event corresponds to the fact that at least one of our confidence bounds in Equation (9) and Equation (12) fails. As already proved in Equation (6) and Proposition 1, $\mathbb{P}(E)$ does not exceed 2δ. The following theorem shows that, as long as E does not hold, the regret is under control.

Algorithm 1: Optimistic/Pessimistic Successive Rejects (OPSR)

Data: Time horizon T.

Set $B_0 \leftarrow 0, A_0 \leftarrow 1$
for $t = 1, \dots T$ **do**
  Play (B_{t-1}, A_{t-1})
  Update Equation (9) and Equation (12)
  for $x \in \mathcal{A}_{t-1}$ **do**
    $\Theta_{\text{bid}}(x) \leftarrow (\widehat{\mu}_t^{\text{low}} - x)\widehat{F}_t^{\text{up}}(x)$
    $\Theta_{\text{ask}}(x) \leftarrow (x - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(x)$
    $\Gamma_{\text{bid}}(x) \leftarrow (\widehat{\mu}_t^{\text{low}} - x)\widehat{F}_t^{\text{low}}(x)$
    $\Gamma_{\text{ask}}(x) \leftarrow (x - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(x)$
  end
  $\Gamma^*_{\text{bid}} \leftarrow \max_{x \in \mathcal{A}_{t-1}} \Gamma_{\text{bid}}(x)$
  $\Gamma^*_{\text{ask}} \leftarrow \max_{x \in \mathcal{A}_{t-1}} \Gamma_{\text{ask}}(x)$
  $B_t \leftarrow \min\{a \in \mathcal{A}_{t-1} : \Theta_{\text{bid}}(a) \geq \Gamma^*_{\text{bid}}\}$
  $A_t \leftarrow \max\{a \in \mathcal{A}_{t-1} : \Theta_{\text{ask}}(a) \geq \Gamma^*_{\text{ask}}\}$
end

Theorem 2 *Under Assumptions 1 and 3. Under E^c (the failure event in Equation (13)), the pseudo-regret suffered by Algorithm 1 is bounded by $\mathcal{R}_T \leq \sqrt{T} + 2 + 4\sum_{t=1}^{T-1} \psi(t,\delta) + \phi(t,\delta)$ which, for ϕ, ψ as in Equations (12) and (9), writes as*

$$\mathcal{R}_T \leq \sqrt{48T\log(3T/\delta)} + \mathcal{O}\left(\sqrt{T\log(\log(T))}\right).$$

Building on the previous result, we derive an upper bound on the regret by bounding the failure probability and the gap between the regret and the pseudo-regret. By Theorem 2, Proposition 1, and Proposition 6, for Algorithm 1 we have, with probability at least $1-\delta$,

$$R_T \leq \sqrt{48T\log(9T/\delta)} + \mathcal{O}\left(\sqrt{T\log(\log(T))}\right).$$

This result mirrors several bounds for stochastic bandits Lattimore and Szepesvári (2020). Interestingly, while continuous bandits with an action space in the interval $[0,1]^2$ suffer a regret of $\widetilde{\mathcal{O}}(T^{3/4})$ under Lipschitz reward assumptions Kleinberg (2004), or $\widetilde{\mathcal{O}}(\sqrt{T})$ under much stronger assumptions Bubeck et al. (2011); Liu et al. (2021), our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret without requiring any smoothness assumptions on the cumulative distribution function.

Computational complexity In the domain of market making, latency is a critical bottleneck for any algorithmic framework. The high-frequency nature of modern markets necessitates near-instantaneous decision-making, as liquidity opportunities often vanish within a few milliseconds. Consequently, there is significant interest in designing algorithms that simultaneously provide robust no-regret guarantees and maintain low per-step computational complexity. Our proposed algorithm 1 involves updating the estimates $\widehat{F}_t(x)$ for all $x \in \mathcal{A}_t$, followed by extracting the maxima Γ^*_{bid} and Γ^*_{ask}. These operations entail a per-step complexity of $\mathcal{O}(|\mathcal{A}|) = \mathcal{O}(\sqrt{T})$. Notably, despite the bivariate nature of the optimization problem, we avoid $\mathcal{O}(|\mathcal{A}|^2)$ complexity due to the additive separability of the reward function with respect to bid and ask prices.

To improve computational efficiency without affecting the regret bound, we introduce the LAZY-OPSR algorithm (Algorithm 3, detailed in the Appendix). This variant stores V_t and updates $\widehat{F}_t$ only at time steps t that are powers of two. For most iterations, the per-step complexity is $\mathcal{O}(1)$. In the remaining $\mathcal{O}(\log T)$ iterations, $\widehat{F}_t$ is reconstructed from all past data in $\mathcal{O}(T + |\mathcal{A}|) = \mathcal{O}(T)$ using prefix sums. Thus, the amortized per-step complexity is $\mathcal{O}(\log T)$.

4. Stochastic Prices with Temporal Correlation

In the former section, we covered the stochastic independent case. The proof of the regret bound relies on the definition of a failure event and ensures small regret whenever that event is not verified. By its definition (see Equation (13)), whenever one can ensure that the event

$$\bigcup_{t=1}^{T} \{|\widehat{\mu}_t - \mu| > \phi(t, \delta)\} \qquad \phi(t, \delta) \approx \sqrt{\frac{\log(t/\delta)}{t}} \tag{14}$$

does not happen, Theorem 2 ensures small (pseudo) regret. Assuming that M_t forms an independent sequence is a good way to say that the event in Equation (14) has low probability, but there do exist many other processes with this quality. For example, let

$$M_1 = \text{Unif}(0, 1), \qquad M_{t+1} = 1 - M_t. \tag{15}$$

This sequence is not independent. Nonetheless, their sample mean $\widehat{\mu}_t$ is always either $1/2$, for $t = 2n$ or $\frac{M_1+n}{2n+1}$, for $t = 2n + 1$. In both cases, $|\widehat{\mu}_t - \mu| \leq 1/t$, a rate that is even faster than the i.i.d. case. Stochastic processes of this nature, characterized by a negative correlation between the current value and its preceding deviations from the long-term equilibrium, informally defined as mean-reverting processes. These models are of paramount importance in the field of quantitative finance, as a vast array of algorithmic trading strategies is predicated on mean-reverting assumptions about asset price dynamics. We formalize this phenomenon with two assumptions: one based on the Ornstein-Uhlenbeck process, capturing short-term mean reversion, and another with a martingale viewpoint, capturing mean reversion relative to the process's entire history.

Assumption 5 (Local mean reversion) *For some arbitrary initial conditions $\{M_t\}_{t=-\tau}^{0} \subset [0, 1]$, the market price evolves according to the following $AR(k)$ process:*

$$\forall t \in \mathbb{N} \qquad M_{t+1} = \sum_{\tau=0}^{k-1} \gamma_\tau M_{t-\tau} + (1 - \gamma)\eta_{t+1},$$

where η_{t+1} is a random variable with support in $[0, 1]$ and such that $\mathbb{E}[\eta_{t+1}|\mathcal{F}_t] = \mu$. $\{\gamma_\tau\}_{\tau=0}^{k-1}$ are non-negative real numbers such that $\sum_{\tau=0}^{k-1} \gamma_\tau = \gamma < 1$.

For $k = 1$, the former assumption covers the discrete Ornstein-Uhlenbeck process (Grimmett and Stirzaker, 2020), which is often used to model mean reversion. In that case, assumption 5 may be written as $\Delta M_{t+1} = (1 - \gamma)(\eta_{t+1} - M_t)$, with $\mathbb{E}[\eta_{t+1}|\mathcal{F}_t] = \mu$.

Assumption 6 (Global mean reversion) *There is some constant μ such that, given $X_t := M_t - \mu$ and $S_t = \sum_{\tau=1}^{t} X_\tau$, one has, for every $t \leq T$*

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] \cdot S_t \leq 0.$$

Assumptions 5 and 6 represent two possible formalizations of the mean-reverting mechanism; indeed, analogous definitions have been established in the literature (see, e.g., (Vasicek, 1977; Bouchaud and Potters, 2003)). While the scope of the two assumptions differs slightly, they both capture interesting scenarios. If the market prices are independent as in the previous section, Assumption 5 holds for $k = 1, \gamma = 0$ and Assumption 6 is satisfied with $\mu = \mathbb{E}[M_1]$ as $\mathbb{E}[X_{t+1}|\mathcal{F}_t] = 0$. On the other side, only Assumption 6 covers the sequence of random variables in Equation (15). In both cases, we can prove concentration inequalities that allow to generalize the independent case.

Theorem 3 *Under Assumption 1 and one between Assumptions 5 and 6.*
The event $\bigcup_{t=1}^{T} \left\{|\widehat{\mu}_t - \mu| > \overline{\phi}(t,\delta)\right\}$ *has a probability not larger then* δ*, where the specific form of* $\overline{\phi}$ *depends on the assumption.*

1. *Under Assumption 5,* $\overline{\phi}(t,\delta) := \sqrt{\frac{\log(2T/\delta)}{4t}} + \frac{\gamma(k+1)}{(1-\gamma)t}$.
2. *Under Assumption 6,* $\overline{\phi}(t,\delta) := \sqrt{\frac{4\log(2T/\delta)}{t}}$.

In the second case, the proof follows by a novel martingale argument of independent interest, Theorem 7. To our knowledge, this is the first concentration result that exploits global mean-reversion expressed through cumulative deviations rather than mixing or spectral assumptions. The formed value of $\overline{\phi}$ is worse than ϕ that we used in section 3, by only logarithmic factors ($\log\log T$ becomes $\log T$). After Theorem 3, nothing more needs to be proved to show the bound on the pseudo-regret. Indeed, Theorem 2, which does not assume any specific form for ϕ, can be applied for the two values of $\phi \leftarrow \overline{\phi}$ in Theorem 3 giving, under E^c,

$$\begin{aligned}\mathcal{R}_T &\leq \sqrt{T} + 2 + 4\sum_{t=1}^{T-1} \psi(t,\delta) + \overline{\phi}(t,\delta) \\ &\leq \begin{cases} \sqrt{36 \cdot T\log(3T/\delta)} + \frac{\gamma(k+1)\log(T)}{(1-\gamma)} + \mathcal{O}(\sqrt{T}) & \text{under Assumption 5} \\ \sqrt{526 \cdot T\log(3T/\delta)} + \mathcal{O}(\sqrt{T}) & \text{under Assumption 6} \end{cases}\end{aligned}$$

At this point, Theorem 3 shows that $\mathbb{P}(E) < 2\delta$, so the former result is a valid high-probability pseudo-regret bound. Interestingly, with respect to the analogous result in the stochastic case, there is only some constant change, not the regret order, even when comparing the logarithmic terms.

To pass from the former result to an upper bound for the regret R_T is not trivial. In fact, Proposition 6, which bounds the discrepancy between $\mathcal{R}_T$ and R_T is only valid in the stochastic setting. For this reason, we have to slightly modify Algorithm 1, by performing action eliminations only at times-steps t such that $t = 2^p$ for some $p \in \mathbb{N}$. After employing this technique, sometimes called the *doubling trick* (Besson and Kaufmann, 2018), we refer to the algorithm as LAZYOPSR. For a detailed implementation of this algorithm, see the Appendix A.

Theorem 4 *Under Assumptions 1, 2, 3 and one between 5 and 6, the regret suffered by* LAZYOPSR *is bounded, with probability at least* $1-\delta$*, by*

$$R_T \leq \begin{cases} \sqrt{2007 \cdot T\log(6T/\delta)} + \frac{8\gamma(k+1)\log(T)}{(1-\gamma)} + \mathcal{O}(\sqrt{T}) & \textit{under Assumption 5} \\ \sqrt{20783 \cdot T\log(6T/\delta)} + \mathcal{O}(\sqrt{T}) & \textit{under Assumption 6} \end{cases}$$

Algorithm 2: Explore Then Perturb (ETP)

Data: Time horizon T, number of exploratory rounds κ.

$B_0 \leftarrow 0, A_0 \leftarrow 1$
for $t = 1, \dots \kappa$ **do**
  Play (B_{t-1}, A_{t-1}) $B_t \leftarrow 0, A_t \leftarrow 1$
end
Compute $\widehat{F}$ with Equation (8)
$\widehat{S}(\cdot) \leftarrow 1 - \widehat{F}(\cdot)$
$\widehat{\mu}_t \leftarrow 0$
Sample $\varepsilon \sim \text{Unif}(0, T^{-1/2})$
for $t = \kappa + 1, \dots T$ **do**
  Play (B_{t-1}, A_{t-1})
  $\widehat{\mu}_t \leftarrow \frac{(t-\kappa-1)\widehat{\mu}_{t-1} + M_t}{t-\kappa}$
  $B_t \leftarrow \text{argmax}_{b \in \mathcal{A}} \widehat{F}(b)(\widehat{\mu}_t + \varepsilon - b)$
  $A_t \leftarrow \text{argmax}_{a \in \mathcal{A}} \widehat{S}(a)(a + \varepsilon - \widehat{\mu}_t)$
end

The motivation for employing a scheme such as LAZYOPSR in this setting stems from the partially adversarial nature of the environment. Given that Assumption 6 is considerably weaker than the standard i.i.d. assumption on the sequence $\{M_t\}_{t=1}^T$, it is prudent to restrict the algorithm to a limited number of action switches, specifically $\mathcal{O}(\log T)$ in our case. This "lazy" update schedule serves as a regularization mechanism, preventing the agent's choice from correlating with M_t in an unpredictable way. Despite this doubling mechanism, which slows down learning, the regret is only a constant worse than the stochastic case (note $\sqrt{20783} \approx 144$ while $\sqrt{48} \approx 7$).

5. Adversarial Market Values

After discussing the stochastic case and its generalization, in this section, we focus on the case where no assumption is put on M_t, which is allowed to be any sequence of values in $[0, 1]$. This regime introduces significantly greater challenges: under such agnostic conditions, the historical realization of the sequence $\{M_t\}_{t \geq 1}$ provides no predictive information regarding its future evolution. Consequently, the empirical mean price $\widehat{\mu}_t$, which exhibited relative stationarity in the stochastic setting, may now undergo substantial fluctuations throughout the learning process. This volatility necessitates continuous exploration across the entire bid-ask action space, precluding any strategy aimed at a monotonic narrowing of the spread.

Our strategy, Algorithm 2, is to split the two problems, the first being the partial feedback, which allows to estimate $\widehat{F}(x)$ only for $x \in [B_t, A_t]$, and the other being that $\widehat{\mu}_t$ is not guaranteed to converge to any specific value. To begin with, Algorithm 2 utilizes the first κ iterations with the environment to estimate the function $\widehat{F}(x)$ across the entire interval, maintaining $B_t = 0$ and $A_t = 1$ during this phase. Selecting bid and ask prices at the boundaries of the interval causes the regret to grow rapidly; however, it is the only way to ensure a uniformly good estimate of F. This initial phase, which continues until line 6, is essentially based on the principles of the EXPLORE-THEN-COMMIT algorithm (Garivier et al., 2016). The subsequent part of the algorithm no longer

attempts to improve the estimate $\widehat{F}$; instead, it focuses entirely on minimizing the regret. The framework we draw inspiration from is the FOLLOW-THE-PERTURBED-LEADER algorithm (Kalai and Vempala, 2005), the core idea of which is to inject random noise to perturb the cumulative reward function, thereby stabilizing the optimizer's decisions. This is implemented in line algorithm 2, where the noise ε is sampled from the distribution $\text{Unif}(0, T^{-1/2})$. The utility of this seemingly counterintuitive choice lies in the need to prevent the sequence of chosen $\{B_t, A_t\}$ from being excessively volatile. In non-convex contexts such as ours, a minimal variation in the cumulative reward function (which, in our case, is essentially $\widehat{F}(b)(\widehat{\mu}_t - b) + \widehat{S}(a)(a - \widehat{\mu}_t)$) can cause abrupt shifts in its maximizer. The inclusion of noise allows us to prove that, on average, this instability remains bounded for any sequence of market prices $\{M_t\}_t$ that cannot adapt to ε.

Interestingly, the noise applied here differs from the usual implementations of FTPL. Here, instead of adding the noise to $\sum_{\tau=1}^{t} r_\tau(b, a)$, where r_τ is the reward relative to any pair b, a at step t, the noise is added to $\widehat{\mu}_t$. This choice is suited to the particular structure of our problem, and we do not know if a more standard implementation of FTPL could achieve a similar regret bound.

Theorem 5 *Under Assumptions 1, 2 and 3 and let $\{M_t\}_{t=1}^{T}$ be any unkown oblivious sequence. For $\kappa = \left\lceil \log(3T)^{\frac{1}{3}} T^{\frac{2}{3}} \right\rceil$, the regret suffered by Algorithm 2 is bounded, in expectation, by*

$$\mathbb{E}[R_T] \leq 5 \log(T)^{\frac{1}{3}} T^{\frac{2}{3}} + \mathcal{O}\left(\sqrt{T} \log(T)\right).$$

Theorem 5 shows that ETP is capable of doing a sublinear regret in the adversarial case. With respect to the previous results for the stochastic case, this performance guarantee has two drawbacks. First, the bound scales roughly with $T^{2/3}$, which is worse than $\sqrt{T}$. Second, the guarantees only hold in expectation, not with high probability. This fact is a consequence of the use of ε: for unfortunate choices, the regret distribution could in principle have heavy tails. Lastly, note that the regret bound holds for any sequence $\{M_t\}_t$ that is oblivious, meaning that its values must not depend on the sampled ε. An assumption of this form is often called *oblivious adversary*, and is standard in the literature. Generalizing to a sequence that can adapt to the agent's actions may be even more difficult.

6. Related Work

Online market making has been extensively studied in financial economics and market microstructure, with classical models such as Glosten–Milgrom capturing the interaction between market makers and traders with private information (Das*, 2005; Touzo et al., 2021). Cesa-Bianchi et al. (2024) recently introduced a regret-minimization framework for market making under fully censored (bandit) feedback, where the learner only observes whether a trade occurred. They show that under such feedback, sublinear regret is achievable only under restrictive smoothness assumptions on the distribution of trader valuations, and that regret lower bounds are significantly worse in more general settings. Notably, the extension of their framework to richer feedback models is explicitly identified as an important direction for future work. A summary of the best-known regret guarantees for online market making under different feedback models and assumptions on the price process is reported in Table 1.

The feedback structure studied in this paper concerns bandit problems with partial or censored observations. In most existing work, however, the censoring mechanism is exogenous and independent of the learner's action. Action-dependent feedback has been studied in other online learning

Table 1: Comparison of regret guarantees for online market making under different feedback models. The mixed feedback model captures order-book observability when no trade occurs.

Price dynamics	**Stochastic**	**Mean-Reverting**	**Adversarial**
Bandit feedback (Cesa-Bianchi et al., 2024)	$T^{\frac{2}{3}}$	N/A	$T^{\frac{2}{3},*}$
Mixed (Assumption 3, This paper)	$T^{\frac{1}{2}}$	$T^{\frac{1}{2}}$	$T^{\frac{2}{3}}$

* assumes that the c.d.f. in Assumption 2 is Lipschitz continuous

settings, such as feedback graphs, in which selecting an action reveals losses for neighboring actions. In contrast, our model exhibits a continuous and asymmetric form of action-dependent feedback: the informativeness of the observation depends on the chosen bid–ask spread, and informative feedback is obtained precisely in rounds where the instantaneous reward is zero. To the best of our knowledge, this specific feedback structure has not been previously analyzed.

From a bandit perspective, our problem is a continuous-action bandit with a two-dimensional action space. Classical continuum-armed bandits typically need Lipschitz or smoothness assumptions on the reward to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret; weaker assumptions give slower rates (Kleinberg, 2004; Bubeck et al., 2011; Liu et al., 2021). In contrast, we obtain $\tilde{\mathcal{O}}(\sqrt{T})$ regret in stochastic and mean-reverting settings without smoothness assumptions, instead leveraging richer, action-dependent feedback. Mean-reverting price dynamics are central in quantitative finance and are commonly modeled by autoregressive or Ornstein–Uhlenbeck processes (Vasicek, 1977; Bouchaud and Potters, 2003). We adopt a learning-oriented perspective, imposing both local autoregressive dynamics and a weaker global mean-reverting drift condition based on cumulative deviations from the mean. This abstraction relaxes the classical i.i.d. assumption while remaining compatible with standard mean-reverting models, and yields high-probability regret guarantees in settings not covered by existing online market-making analyses.

Relation with Partial Monitoring An interesting connection of this work lies within the framework of online learning with partial monitoring (Cesa-Bianchi et al., 2006; Bartók et al., 2011; Lattimore and Szepesvári, 2019). In standard partial monitoring games, the feedback received by the agent is decoupled from the reward, creating a delicate trade-off between exploitation and information acquisition. In our market-making formulation, this structure manifests in a specific manner: our feedback allows the learner to see the exact private valuation of the trader whenever an order remains unexecuted. While the setting could resemble the "easy" class presented in the classic partial monitoring literature (Lattimore and Szepesvári, 2019), where a regret of $\tilde{\mathcal{O}}(\sqrt{T})$ is achievable, the key distinction is that here we consider a continuous set of arms without any global smoothness assumption. In the finite outcome and action settings characterized by Lattimore and Szepesvári (2019), "easy" games rely on strict geometric observability conditions to reconstruct unobserved losses. In contrast, by moving to an infinite action space standard partial monitoring techniques would need to employ discretization techniques, which typically lead to a $\Theta(T^{2/3})$ regret bound. From this perspective, our main result for the stochastic case may be seen as a way to show that the agent can play good arms while ensuring to always have full feedback for the candidate optimal arm, effectively bypassing the exploration-exploitation dilemma.

7. Conclusions

Motivated by a key feature of real limit order books, private valuations are revealed precisely when no trade occurs, we design a novel action-dependent feedback structure for Online Learning. Under this feedback, we prove substantially improved regret guarantees over standard bandit assumptions. In the stochastic independent setting, OPSR Algorithm 1 attains regret $\sqrt{T\log(T)}$ (Theorem 2). A minor modification that slows the update (LAZYOPSR) yields the same bound in a more general mean-reverting price setting. Finally, (ETP) Algorithm 2, combining an explore-then-commit scheme with the FTPL approach, achieves regret $T^{\frac{2}{3}}\log(T)^{\frac{1}{3}}$ in the adversarial case.

Future works While this work focuses on a stylized market-making model, we view it as a foundational step toward understanding how information structure affects learnability in online market-making. In particular, our analysis isolates the role of action-dependent feedback by deliberately abstracting away additional sources of complexity. A natural next direction is to incorporate inventory constraints, which play a central role in practical market making. Extending the proposed framework to settings in which the learner must control bid and ask quotes while maintaining a balanced inventory would allow one to study the interaction between risk management and action-dependent feedback, and to quantify the additional learning cost induced by inventory control. Another key extension concerns the symmetry of the trading mechanism. The current model summarizes interaction with the market through a single private valuation and a reference price, enabling a clean analysis of feedback. A more realistic approach would model buy and sell pressure separately, matching bids and asks to distinct order streams and rewarding the market maker through the realized spread.

References

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

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 133–154. JMLR Workshop and Conference Proceedings, 2011.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge university press, 2003.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Burton-Taylor International Consulting. Exchange global share and segment sizing 2025. Technical report, Burton-Taylor International Consulting, 2025. URL https://tpicap.com/burtontaylor/reports/2025exchangebenchmark#.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Luigi Foscari, and Vinayak Pathak. Market making without regret. *arXiv preprint arXiv:2411.13993*, 2024.

Sanmay Das*. A learning market-maker in the glosten–milgrom model. *Quantitative Finance*, 5 (2):169–180, 2005.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 2004.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory*, pages 529–556. PMLR, 2019.

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

Yusha Liu, Yining Wang, and Aarti Singh. Smooth bandit optimization: generalization to holder space. In *International Conference on Artificial Intelligence and Statistics*, pages 2206–2214. PMLR, 2021.

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

Léo Touzo, Matteo Marsili, and Don Zagier. Information thermodynamics of financial markets: The glosten–milgrom model. *Journal of Statistical Mechanics: Theory and Experiment*, 2021 (3):033407, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188, 1977.

Appendix A. LAZYOPSR

In this section, we present LAZYOPSR, which is a modification of our main algorithm OPSR (see Algorithm 1) with a lazy update and a doubling-trick schedule. The rest of the structure follows exactly the same logic, so it does not need any particular comment.

Algorithm 3: LAZYOPSR

Data: Time horizon T.

Set $B_0 \leftarrow 0, A_0 \leftarrow 1$
$p \leftarrow 0$
for $t = 1, \dots T$ **do**
  Play (B_{t-1}, A_{t-1})
  Update Equation (9) and Equation (12)
  if $t = 2^p$ **then**
    for $x \in \mathcal{A}_{t-1}$ **do**
      $\Theta_{\text{bid}}(x) \leftarrow (\widehat{\mu}_t^{\text{low}} - x)\widehat{F}_t^{\text{up}}(x)$
      $\Theta_{\text{ask}}(x) \leftarrow (x - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(x)$
      $\Gamma_{\text{bid}}(x) \leftarrow (\widehat{\mu}_t^{\text{low}} - x)\widehat{F}_t^{\text{low}}(x)$
      $\Gamma_{\text{ask}}(x) \leftarrow (x - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(x)$
    end
    $\Gamma^*_{\text{bid}} \leftarrow \max_{x \in \mathcal{A}_{t-1}} \Gamma_{\text{bid}}(x)$
    $\Gamma^*_{\text{ask}} \leftarrow \max_{x \in \mathcal{A}_{t-1}} \Gamma_{\text{ask}}(x)$
    $B_t \leftarrow \min\{a \in \mathcal{A}_{t-1} : \Theta_{\text{bid}}(a) \geq \Gamma^*_{\text{bid}}\}$
    $A_t \leftarrow \max\{a \in \mathcal{A}_{t-1} : \Theta_{\text{ask}}(a) \geq \Gamma^*_{\text{ask}}\}$
    $p \leftarrow p + 1$
  end
end

Appendix B. Proofs from Section 3

We start proving Equation (6).

Proof M_t is an independent process bounded in $[0, 1]$ by Assumption 1. Therefore, Theorem 9.2 by Lattimore and Szepesvári (2020), for any $\varepsilon > 0$ ensures that, for any n,

$$\mathbb{P}\left(\exists t \leq n, \quad \sum_{\tau=1}^{t} M_\tau - t\mu \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{n}\right).$$

Fixing $\delta > 0$ and calling $\varepsilon = \sqrt{n \log(1/\delta)/2}$ gives

$$\mathbb{P}\left(\exists t \leq n, \quad \sum_{\tau=1}^{t} M_\tau - t\mu \geq \sqrt{\frac{n \log(1/\delta)}{2}}\right) \leq \delta.$$

Making a union bound for $n = 2, 4, 8 \dots 2^{\lceil \log(T) \rceil}$ gives

$$\mathbb{P}\left(\exists t \leq T, \quad \sum_{\tau=1}^{t} M_\tau - t\mu \geq \sqrt{\frac{2^{\lceil \log(t) \rceil} \log(1/\delta)}{2}}\right) \leq \delta \lceil \log(T) \rceil.$$

At this point, we note that $2^{\lceil \log(t) \rceil} \leq 2t$, so the former writes as

$$\mathbb{P}\left(\exists t \leq T, \quad \sum_{\tau=1}^{t} M_\tau - t\mu \geq \sqrt{t \log(1/\delta)}\right) \leq \delta \lceil \log(T) \rceil,$$

which, by definition of the sample mean, corresponds to

$$\mathbb{P}\left(\exists t \leq T, \quad \widehat{\mu}_t - \mu \geq \sqrt{\frac{\log(1/\delta)}{t}}\right) \leq \delta\lceil\log(T)\rceil.$$

Flipping the sign proves that also

$$\mathbb{P}\left(\exists t \leq T, \quad \mu - \widehat{\mu}_t \geq \sqrt{\frac{\log(1/\delta)}{t}}\right) \leq \delta\lceil\log(T)\rceil.$$

and then $\delta \leftarrow \delta/(2\lceil\log(T)\rceil)$ ends the proof. ■

Proposition 1 *For any $t \in [T], x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$, $\widehat{F}_t(x) = \frac{1}{t}\sum_{\tau=1}^{t} \chi_\tau(x)$ and, for any $\delta > 0$,*

$$\mathbb{P}\left(\exists t \in [T],\ \exists x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau, \quad |\widehat{F}_t(x) - F(x)| > \sqrt{\frac{3\log(3T/\delta)}{4t}}\right) \leq \delta.$$

Proof Let us fix $x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$. Then, by Equation (8)

$$\begin{aligned}\widehat{F}_t(x) &= \frac{\sum_{\tau=1}^{t} \mathbb{1}(V_\tau \leq x)}{t} \\ &= \frac{\sum_{\tau=1}^{t} \mathbb{1}(V_\tau^{\text{clip}} \leq x)}{t} \\ &= \frac{1}{t}\sum_{\tau=1}^{t} \chi_\tau(x).\end{aligned}$$

The second equality for the follows because $x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$, so

1. If $V_\tau < B_\tau$ then $V_\tau^{\text{clip}} = B_\tau$, so $V_\tau^{\text{clip}} \leq x$ always holds.
2. If $B_\tau \leq V_\tau \leq A_\tau$ then $V_\tau^{\text{clip}} = V_\tau$.
3. If $V_\tau > A_\tau$ then $V_\tau^{\text{clip}} > A_\tau$ and $V_\tau^{\text{clip}} \leq x$ never holds.

At this point, by Assumptions 1 and 2, $\widehat{F}_t(x)$ is the sample mean of a sequence of i.i.d. random variables bounded in $[0, 1]$. Thus, Hoeffding's inequality ensures that, with probability at most $1 - \delta$

$$|\widehat{F}_t(x) - F(x)| \leq \sqrt{\frac{\log(2/\delta)}{2t}}.$$

To obtain a uniform bound, we use a union bound. Indeed, $\bigcap_{\tau=1}^{t} \mathcal{A}_\tau \subset \mathcal{A}$, with $|\mathcal{A}| \leq \sqrt{T} + 1$. Therefore, at the same time for every $t, x \in \bigcap_{\tau=1}^{t} \mathcal{A}_\tau$, with probability at least $1 - \delta$,

$$|\widehat{F}_t(x) - F(x)| \leq \sqrt{\frac{\log(2T(\sqrt{T}+1)/\delta)}{2t}} \leq \sqrt{\frac{3\log(3T/\delta)}{4t}},$$

which completes the proof. ■

Proposition 6 *Under Assumptions 1, 2 and 4, for any choice B_t, A_t as a $\mathcal{F}_{t-1}$ measurable sequence, with probability at least $1-\delta$*

$$|R_T - \mathcal{R}_T| \leq \sqrt{8T\log(4/\delta)}.$$

Proof By definition of regret and Equations (3) and (5),

$$\begin{aligned} R_T &= \sup_{b,a\in[0,1]} \sum_{t=1}^{T} J_t(b,a) - J_t(B_t,A_t) \\ &= T\left[\sup_{b,a\in[0,1]} (\widehat{\mu}_T - b)F(b) + (a-\widehat{\mu}_T)S(a)\right] - \sum_{t=1}^{T} J_t(B_t,A_t). \end{aligned}$$

To bound the discrepancy with $\mathcal{R}_t$, we are going to compare the two parts with the corresponding parts of Equation (7). First, by Hoeffding's inequality, with probability at least $1-\delta$,

$$\begin{aligned} &\left|T\left[\sup_{b,a\in[0,1]} (\widehat{\mu}_T - b)F(b) + (a-\widehat{\mu}_T)S(a)\right] - T\cdot\sup_{b,a\in[0,1]} J(b,a)\right| \\ &\qquad \leq T\left|\sup_{b,a\in[0,1]} (\widehat{\mu}_T - \mu)F(b) + (\mu-\widehat{\mu}_T)S(a)\right| \\ &\qquad \leq 2T|\widehat{\mu}_T - \mu| \leq 2T\sqrt{\frac{\log(2/\delta)}{2T}} = \sqrt{2T\log(2/\delta)}. \end{aligned}$$

Let us focus on the second one.

$$\sum_{t=1}^{T} J_t(B_t,A_t) - \sum_{t=1}^{T} J(B_t,A_t) = \sum_{t=1}^{T} F(B_t)(M_t-\mu) + S(A_t)(\mu - M_t)$$

Thanks to Assumption 4, M_t is sampled at each time step independently from the past. Therefore, as B_t, A_t form a $\mathcal{F}_{t-1}$ measurable sequence, the previous sum is a martingale, whose increments are bounded by

$$\max_{b,a,m\in[0,1]} |F(b)(m-\mu) + S(a)(\mu - m)| \leq 2.$$

Therefore, another application of the Azuma-Hoeffding's inequality gives, with probability at least $1-\delta$

$$\sum_{t=1}^{T} J_t(B_t,A_t) - \sum_{t=1}^{T} J(B_t,A_t) \leq \sqrt{2T\log(2/\delta)}.$$

The union of the former two events completes the proof. ■

Theorem 2 *Under Assumptions 1 and 3. Under E^c (the failure event in Equation (13)), the pseudo-regret suffered by Algorithm 1 is bounded by $\mathcal{R}_T \leq \sqrt{T} + 2 + 4\sum_{t=1}^{T-1} \psi(t,\delta) + \phi(t,\delta)$ which, for ϕ, ψ as in Equations (12) and (9), writes as*

$$\mathcal{R}_T \leq \sqrt{48T\log(3T/\delta)} + \mathcal{O}\left(\sqrt{T\log(\log(T))}\right).$$

Proof As our algorithm works on the discrete set $\mathcal{A}$, it is necessary to reduce the supremum over the continuous interval to a maximum over $\mathcal{A}$. Let $\tilde{b}, \tilde{a}$ such that

$$J(\tilde{b}, \tilde{a}) \geq \sup_{b,a \in [0,1]} J(b, a) - T^{-1}.$$

The existence of such a pair follows by definition of supremum. Now, let

$$b^\star := \min\{x \in \mathcal{A} : x \geq \tilde{b}\}, \qquad a^\star := \max\{x \in \mathcal{A} : x \leq \tilde{a}\}.$$

By definition of $\mathcal{A}$, $b^\star - \tilde{b} \leq T^{-1/2}$ and $\tilde{a} - a^\star \leq T^{-1/2}$. Then,

$$\begin{aligned}(\mu - b^\star)F(b^\star) + (a^\star - \mu)S(a^\star) &\geq (\mu - b^\star)F(\tilde{b}) + (a^\star - \mu)S(\tilde{a}) \\ &\geq (\mu - \tilde{b})F(\tilde{b}) + (\tilde{a} - \mu)S(\tilde{a}) - 2T^{-1/2},\end{aligned}$$

where the first inequality comes from the fact that $F(x)$ is non-decreasing and $S(x)$ is non-increasing, and the second by the bound on the differences $b^\star - \tilde{b}, \tilde{a} - a^\star$. This entails that

$$T \cdot \sup_{b,a \in \mathcal{A}} J(b, a) \geq T \cdot J(b^\star, a^\star) \geq T \cdot \sup_{b,a \in [0,1]} J(b, a) - 2\sqrt{T} - 1.$$

This fact allows us to focus on the regret w.r.t. the optimum within $\mathcal{A}$. As $\mathcal{A}$ is a finite set, there exists

$$b^\star, a^\star \in \arg\max_{b,a \in \mathcal{A}} J(b, a).$$

We still name them in this way, with a small overload of notation. Under our assumptions, Lemma 2 ensures that $b^\star, a^\star$ are in $\mathcal{A}_t$ at any step t. The following inequality holds by design of the algorithm

$$\begin{aligned}(A_{t+1} - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(A_{t+1}) &= \Theta_{\text{ask,t}}(A_{t+1}) \\ &\geq \Gamma_{\text{ask,t}}(a^\star) \\ &= (a^\star - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(a^\star). \end{aligned} \tag{16}$$

Where the inequality comes from the fact that $a^\star \in \mathcal{A}_t$. In a symmetric way, one also has

$$(\widehat{\mu}_t^{\text{low}} - B_{t+1})\widehat{F}_t^{\text{up}}(B_{t+1}) \geq (\widehat{\mu}_t^{\text{low}} - b^\star)\widehat{F}_t^{\text{low}}(b^\star) \tag{17}$$

At the same time, by definition of E, one also has

$$(A_{t+1} - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(A_{t+1}) - (A_{t+1} - \mu)S(A_{t+1}) \leq \psi(t, \delta) + \phi(t, \delta)$$

$$(a^\star - \mu)S(a^\star) - (a^\star - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(a^\star) \leq \psi(t, \delta) + \phi(t, \delta),$$

and analogous equation holding for the bids. Replacing these results in equations (17) and (16) gives

$$(\mu - b^\star)F(b^\star) - (\mu - B_{t+1})F(B_{t+1}) \leq 2\psi(t, \delta) + 2\phi(t, \delta), \tag{18}$$

$$(a^\star - \mu)S(a^\star) - (A_{t+1} - \mu)S(A_{t+1}) \leq 2\psi(t, \delta) + 2\phi(t, \delta). \tag{19}$$

Putting together all previous passages gives

$$\begin{aligned}\mathcal{R}_T &= T \cdot \sup_{b,a\in[0,1]} J(b,a) - \sum_{t=1}^{T} J(B_t, A_t) \\ &\leq \sum_{t=1}^{T} J(b^\star, a^\star) - J(B_t, A_t) + \sqrt{T} + 1 \\ &\leq \sqrt{T} + 2 + \sum_{t=1}^{T-1} J(b^\star, a^\star) - (\mu - B_{t+1})F(B_{t+1}) - (A_{t+1} - \mu)S(A_{t+1}) \\ &\leq \sqrt{T} + 2 + \sum_{t=1}^{T-1} 4\psi(t,\delta) + 4\phi(t,\delta).\end{aligned}$$

The last sums can be upper-bounded in an explicit way:

$$\sum_{t=1}^{T-1} \psi(t,\delta) = \sum_{t=1}^{T-1} \sqrt{\frac{3\log(3T/\delta)}{4t}} \leq \sqrt{3T\log(3T/\delta)},$$

and

$$\sum_{t=1}^{T-1} \phi(t,\delta) = \sum_{t=1}^{T-1} \sqrt{\frac{\log(3\log(T)/\delta)}{t}} \leq \sqrt{4T\log(3\log(T)/\delta)}.$$

This completes the proof. ■

Lemma 1 (Elimination scheme monotonicity) *For every $t = 1, \dots T-1$ we have $\Gamma^*_{t,bid} \leq \Gamma^*_{t+1,bid}$ and $\Gamma^*_{t,ask} \leq \Gamma^*_{t+1,ask}$.*

Proof We prove $\Gamma^*_{\text{t,bid}} \leq \Gamma^*_{\text{t+1,bid}}$, as the other part is analogous. By definition,

$$\Gamma^*_{\text{t,bid}} = \max_{x\in\mathcal{A}_{t-1}} (\widehat{\mu}_t^{\text{low}} - x)\widehat{F}_t^{\text{low}}(x).$$

Call x^* the arm realizing this maximum. As x^* had the highest lower bound at t, it is still active at $t+1$,

$$\begin{aligned}\Gamma^*_{\text{t+1,bid}} &\geq (\widehat{\mu}_{t+1}^{\text{low}} - x^*)\widehat{F}_{t+1}^{\text{low}}(x^*) \\ &\geq (\widehat{\mu}_t^{\text{low}} - x^*)\widehat{F}_t^{\text{low}}(x^*) = \Gamma^*_{\text{t,bid}},\end{aligned}$$

where the second inequality comes from the monotonic structure of eq. (9) and eq. (12). ■

Lemma 2 *Under Assumptions 1, 2, 3, and 4. Let*

$$b^\star, a^\star \in \arg\max_{b,a\in\mathcal{A}} J(b,a).$$

Under E^c, the failure event in Equation (13), the optimal arms $b^\star, a^\star \in \bigcap_{t=1}^{T} \mathcal{A}_t$ while running Algorithm 1.

Proof We prove that for any time step t, $A_t \geq a^\star$. The proof for the optimal bid is equivalent to reversing the signs. This completes the statement: as $b^\star \leq \mu \leq a^\star$, proving that, at any time-step t, $A_t \geq a^\star$ and $b^\star \geq B_t$ implies that both are always in $\mathcal{A}_t$.

By design of algorithm 1, a necessary condition to eliminate $a^\star$ is that at some t there is $a < a^\star$ such that

$$\Theta_{\text{ask}}(a^\star) < \Gamma_{\text{t,ask}} = \Gamma_{\text{ask}}(a).$$

We first note that for this condition to occur, it is necessary that $a \geq \widehat{\mu}_t^{\text{up}}$. Indeed, if this condition is not satisfied, applying lemma 1 gives

$$0 = \Gamma_{\text{0,ask}} \leq \Gamma_{\text{t,ask}} = \Gamma_{\text{ask}}(a) < 0.$$

We can therefore continue the proof assuming $a - \widehat{\mu}_t^{\text{up}}$ to be positive.

The condition $\Theta_{\text{ask}}(a^\star) < \Gamma_{\text{ask}}(a)$ wites as

$$(a^\star - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(a^\star) < (a - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(a).$$

By definition of event E and $\widehat{S}_t^{\text{up}}, \widehat{S}_t^{\text{low}}$ we have $\widehat{S}_t^{\text{low}}(a) \leq S(a) \leq \widehat{S}_t^{\text{up}}(a)$ and $\mu \leq \mu_t^{\text{up}}$. It follows that

$$\begin{aligned}
(a - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{low}}(a) &\leq (a - \widehat{\mu}_t^{\text{up}})S(a) && (20)\\
&= (a - \mu)S(a) + (\mu - \widehat{\mu}_t^{\text{up}})S(a) && (21)\\
&\leq (a^\star - \mu)S(a^\star) + (\mu - \widehat{\mu}_t^{\text{up}})S(a) && (22)\\
&\leq (a^\star - \mu)S(a^\star) + (\mu - \widehat{\mu}_t^{\text{up}})S(a^\star) && (23)\\
&= (a^\star - \widehat{\mu}_t^{\text{up}})S(a^\star) && (24)\\
&\leq (a^\star - \widehat{\mu}_t^{\text{up}})\widehat{S}_t^{\text{up}}(a^\star). && (25)
\end{aligned}$$

Passage (22) comes from the optimality of $a^\star$, while (23) from the fact that $a < a^\star$, so $S(a^\star) < S(a)$ and $\mu - \widehat{\mu}_t^{\text{up}}$ is negative. ■

Appendix C. Proofs from Section 4

Theorem 7 *Let X_n for $n \in \mathbb{N}$ be a stochastic process bounded in $[-\sigma/2, \sigma/2]$ (for some $\sigma > 0$), and $S_n := \sum_{i=1}^n X_i$ such that for any n*

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \cdot S_n \leq 0. \tag{26}$$

If $n\sigma^2 > 2$, with probability at least $1 - \delta$,

$$|S_n| \leq \sqrt{2n\sigma^2 \log(2n/\delta)}.$$

Proof For any $\lambda \in \mathbb{R}$, the following inequalities hold

$$\begin{aligned}
\mathbb{E}[\exp(\lambda S_{i+1})|\mathcal{F}_i] &= \mathbb{E}[\exp(\lambda S_i + \lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i] + \lambda(X_{i+1} - \mathbb{E}[X_{i+1}|\mathcal{F}_i]))|\mathcal{F}_i]\\
&= \mathbb{E}[\exp(\lambda S_i + \lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i])|\mathcal{F}_i]\\
&\quad \cdot \mathbb{E}[\exp(\lambda(X_{i+1} - \mathbb{E}[X_{i+1}|\mathcal{F}_i]))|\mathcal{F}_i]\\
&\leq \exp(\lambda S_i)\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i])e^{\frac{\lambda^2\sigma^2}{2}}.
\end{aligned}$$

The last steps follow from the fact that $X_{i+1} - \mathbb{E}[X_{i+1}|\mathcal{F}_i]$ is zero-mean, independent from the past (by definition of conditional expectation), and bounded in $[-\sigma, \sigma]$, so also $\sigma-$sub-Gaussian.

By definition,

$$\exp(\lambda S_i)\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i]) = (\mathbb{1}\{S_i \le 0\} + \mathbb{1}\{S_i > 0\})\exp(\lambda S_i)\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i]).$$

Using Equation (26) it follows that only one between S_i and X_{i+1} can be positive. Therefore, the first term is

$$\mathbb{1}\{S_i \le 0\}\exp(\lambda S_i)\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i]) \le \mathbb{1}\{S_i \le 0\}\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i]).$$

On the other side,

$$\mathbb{1}\{S_i > 0\}\exp(\lambda S_i)\exp(\lambda\,\mathbb{E}[X_{i+1}|\mathcal{F}_i]) \le \mathbb{1}\{S_i > 0\}\exp(\lambda S_i).$$

Together, the two inequalities imply

$$\begin{aligned}\mathbb{E}[\exp(\lambda S_{i+1})|\mathcal{F}_i] &\le e^{\frac{\lambda^2\sigma^2}{2}}\max\{\exp(\lambda S_i), \exp(\lambda)\}\\ &\le e^{\frac{\lambda^2\sigma^2}{2}}\left(\exp(\lambda S_n) + \exp(\lambda)\right).\end{aligned}$$

Which, by induction, means

$$\mathbb{E}[\exp(\lambda S_n)] \le e^{\frac{n\lambda^2\sigma^2}{2}} + e^{\lambda}\sum_{m=0}^{n-1} e^{\frac{m\lambda^2\sigma^2}{2}} \le ne^{\frac{n\lambda^2\sigma^2}{2}} + ne^{\lambda}.$$

Let us $\lambda = t/n\sigma^2$, so that

$$e^{\frac{n\lambda^2\sigma^2 - 2\lambda t}{2}} \to e^{-\frac{t^2}{2n\sigma^2}} \qquad e^{\lambda(1-t)} \to e^{\frac{t-t^2}{n\sigma^2}}$$

and Markov's inequality:

$$\begin{aligned}\mathbb{P}(S_n > t) \le \exp(-\lambda t)\,\mathbb{E}[\exp(\lambda S_n)] &\le e^{\frac{n\lambda^2\sigma^2-2\lambda t}{2}} + \sum_{m=0}^{n-1} e^{\frac{m\lambda^2\sigma^2}{2}} e^{\lambda(1-t)}\\ &\le e^{\frac{n\lambda^2\sigma^2-2\lambda t}{2}} + n\max\{, e^{\lambda(1-t)}\}\\ &\le ne^{-\frac{t^2}{2n\sigma^2}} + ne^{\frac{t-t^2}{n\sigma^2}}\\ &\overset{t\ge 2}{\le} 2ne^{-\frac{t^2}{2n\sigma^2}}.\end{aligned}$$

Fixing $t = \sqrt{2n\sigma^2\log(2n/\delta)}$, the former passages show that

$$\mathbb{P}\left(S_n > \sqrt{2n\sigma^2\log(n/\delta)}\right) \le 2ne^{-\frac{2n\sigma^2\log(n/\delta)}{2n\sigma^2}} = \delta.$$

Changing the sign $X_i \to -X_i$ in the whole proof shows the bound the other way round and completes the proof (note that eq. (26) does not change) ■

Theorem 3 *Under Assumption 1 and one between Assumptions 5 and 6.*
The event $\bigcup_{t=1}^{T} \left\{ |\widehat{\mu}_t - \mu| > \overline{\phi}(t, \delta) \right\}$ *has a probability not larger then* δ*, where the specific form of* $\overline{\phi}$ *depends on the assumption.*

1. *Under Assumption 5,* $\overline{\phi}(t, \delta) := \sqrt{\frac{\log(2T/\delta)}{4t}} + \frac{\gamma(k+1)}{(1-\gamma)t}$.
2. *Under Assumption 6,* $\overline{\phi}(t, \delta) := \sqrt{\frac{4\log(2T/\delta)}{t}}$.

Proof We prove the two parts separately.

(Part 1) By superposition of the effects, we can write $M_t = \overline{M}_t + \widetilde{M}_t$, where, calling $\epsilon_t := \eta_t - \mu$,

$$\widetilde{M}_{t+1} = \sum_{\tau=0}^{k-1} \gamma_\tau \widetilde{M}_{t-\tau} + (1-\gamma)\epsilon_{t+1} \qquad \overline{M}_{t+1} = \sum_{\tau=0}^{k-1} \gamma_\tau \overline{M}_{t-\tau} + (1-\gamma)\mu.$$

By convention, we assume $\widetilde{M}_t = 0$ for $t \leq 0$, moving the initial conditions to the deterministic part.

The stochastic part satisfies the following equations

$$\begin{aligned}
\widetilde{S}_n := \sum_{t=1}^{n} \widetilde{M}_t &= \sum_{t=0}^{n-1} \sum_{\tau=0}^{k-1} \gamma_\tau \widetilde{M}_{t-\tau} + (1-\gamma)\epsilon_{t+1} \\
&= (1-\gamma)\xi_n + \sum_{\tau=0}^{k-1} \sum_{t=0}^{n-1} \gamma_\tau \widetilde{M}_{t-\tau} \\
&= (1-\gamma)\xi_n + \sum_{\tau=0}^{k-1} \gamma_\tau \widetilde{S}_{n-1-\tau},
\end{aligned}$$

where

$$\xi_n := \sum_{t=1}^{n} \epsilon_t.$$

We can prove by induction that, for any $t \leq n$, $\widetilde{S}_t$ is $\sqrt{n}/2-$subgaussian. For S_1 the result holds trivially as $\epsilon_1 \in [0, 1]$. For general $t+1 \leq n$, denoting $\|\cdot\|_{\psi_2}$ the Orclidz norm, we have

$$\begin{aligned}
\|\widetilde{S}_{t+1}\|_{\psi_2} &\leq (1-\gamma)\|\xi_{t+1}\|_{\psi_2} + \sum_{\tau=0}^{k-1} \gamma_\tau \|\widetilde{S}_{t-\tau}\|_{\psi_2} \\
&\leq (1-\gamma)(\sqrt{t+1}/2) + \gamma \max_{\tau \leq t} \|\widetilde{S}_\tau\|_{\psi_2} \\
&\leq (1-\gamma)(\sqrt{t+1}/2) + \gamma(\sqrt{t+1}/2) = \sqrt{t+1}/2.
\end{aligned}$$

As the Orclidz norm corresponds to the sug-Gaussian parameter, this shows that, at any time-step t,

$$\forall t \in \mathbb{N} \qquad \widetilde{S}_t \text{ is } \frac{\sqrt{t}}{2}\text{-sub-Gaussian.} \tag{27}$$

A similar decomposition holds for the deterministic part

$$\begin{aligned}\overline{S}_n := \sum_{t=1}^{n} \overline{M}_t &= \sum_{t=0}^{n-1}\sum_{\tau=0}^{k-1} \gamma_\tau \overline{M}_{t-\tau} + (1-\gamma)\mu \\ &= (1-\gamma)n\mu + \sum_{\tau=0}^{k-1}\sum_{t=0}^{n-1} \gamma_\tau \overline{M}_{t-\tau} \\ &= (1-\gamma)n\mu + \sum_{\tau=0}^{k-1} \gamma_\tau \left(\overline{S}_{n-1-\tau} + \sum_{q=1}^{\tau} M_{-q} \right).\end{aligned}$$

This time, we want to prove by induction that

$$|\overline{S}_t - t\mu| \le \frac{\gamma(\mu+k)}{1-\gamma} =: C.$$

Indeed,

$$\begin{aligned}|\overline{S}_{t+1} - (t+1)\mu| &\le \left| (1-\gamma)(t+1)\mu + \sum_{\tau=0}^{k-1} \gamma_\tau \left(\overline{S}_{t-\tau} + \sum_{q=1}^{\tau} M_{-q} \right) - (t+1)\mu \right| \\ &= \left| (1-\gamma)(t+1)\mu + \sum_{\tau=0}^{k-1} \gamma_\tau Z_\tau - (t+1)\mu \right|,\end{aligned}$$

where

$$Z_\tau = S_{t-\tau} + \sum_{q=1}^{\tau} M_{-q}, \qquad Z_\tau \in [(t-\tau)\mu - C, (t-\tau)\mu + C + \tau],$$

since, by induction, $|S_{t-\tau} - (t-\tau)\mu| \le C$ and $0 \le \sum_{q=1}^{\tau} M_{-q} \le \tau$. For this reason,

$$\gamma((t-k)\mu - C) \le \sum_{\tau=0}^{k-1} \gamma_\tau Z_\tau \le \gamma(t\mu + k + C).$$

Taking the two extrema, we have

$$\begin{aligned}(1-\gamma)(t+1)\mu + \sum_{\tau=0}^{k-1} \gamma_\tau Z_\tau - (t+1)\mu &\le (1-\gamma)(t+1)\mu + \gamma(t\mu + k + C) \\ &\quad - (t+1)\mu \\ &= \gamma(k + C - \mu).\end{aligned}$$

and

$$\begin{aligned}(1-\gamma)(t+1)\mu + \sum_{\tau=0}^{k-1} \gamma_\tau Z_\tau - (t+1)\mu &\ge (1-\gamma)(t+1)\mu + \gamma((t-k)\mu - C) \\ &\quad - (t+1)\mu \\ &= -\gamma(\mu + k + C).\end{aligned}$$

Replacing $C = \frac{\gamma(\mu+k)}{1-\gamma}$, gives the same quantity. As $\mu \leq 1$, this proves that

$$\forall t \in \mathbb{N} \qquad |\overline{S}_t - t\mu| \leq \frac{\gamma(k+1)}{1-\gamma}. \tag{28}$$

It is now possible to recollect the previous formulas to give an upper bound to the discrepancy between μ and $\widehat{\mu}_t$. From eq. (27), it follows that, w.p. at least $1-\delta$,

$$\forall 1 \leq t \leq T \qquad \left|\frac{\widetilde{S}}{t}\right| \leq \sqrt{\frac{\log(2T/\delta)}{4t}}.$$

From eq. (28), we get that, with the same probability

$$\begin{aligned}|\mu - \widehat{\mu}_t| = \left|\mu - \frac{\sum_{\tau=1}^{t} M_\tau}{t}\right| &= \left|\mu - \frac{\widetilde{S}_t + \overline{S}_t}{t}\right| \\ &\leq \frac{\gamma(k+1)}{(1-\gamma)t} + \left|\frac{\widetilde{S}_t}{t}\right| \leq \frac{\gamma(k+1)}{(1-\gamma)t} + \sqrt{\frac{\log(2T/\delta)}{4t}},\end{aligned}$$

which completes the proof.

(Part 2) The former theorem 7 applies with $\sigma = 1$. At any time-step $t > 2$, with probability at least $1-\delta$,

$$\begin{aligned}|\widehat{\mu}_t - \mu| &= \left|\frac{1}{t}\sum_{\tau=1}^{t}(M_t - \mu)\right| \\ &= \left|\frac{S_t}{t}\right| \leq \sqrt{\frac{2\log(2t/\delta)}{t}}.\end{aligned}$$

For $t \leq 2$, we note that the inequality is immediately verified by the fact that $S_t/t \in [-1,1]$ a.s.. Making a union bound over $t = 1, \ldots T$ ends the proof. ■

Theorem 4 *Under Assumptions 1, 2, 3 and one between 5 and 6, the regret suffered by* LAZYOPSR *is bounded, with probability at least* $1-\delta$*, by*

$$R_T \leq \begin{cases} \sqrt{2007 \cdot T\log(6T/\delta)} + \frac{8\gamma(k+1)\log(T)}{(1-\gamma)} + \mathcal{O}(\sqrt{T}) & \textit{under Assumption 5} \\ \sqrt{20783 \cdot T\log(6T/\delta)} + \mathcal{O}(\sqrt{T}) & \textit{under Assumption 6} \end{cases}$$

Proof The proof is done under E^c, for E defined as in eq. (13) in $\phi \leftarrow \overline{\phi}$ and μ coming from assumption 6. Thanks to assumptions 1 2 and 6, proposition 1 and theorem 3 show that $\mathbb{P}(E) < 2\delta$.

Let us call $p = 0, \ldots \lfloor \log T \rfloor$ the current phase of the algorithm. By design, the sequences of bids and asks can be written as sequences $B_{(p)}, A_{(p)}$ each repeated 2^p times.

By equation (3), the regret writes as follows

$$\begin{aligned}
R_T &\leq \sqrt{T} + \sup_{b,a\in\mathcal{A}} \sum_{t=1}^{T} J_t(b,a) - J_t(B_t, A_t) \\
&\leq \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} \sup_{b,a\in\mathcal{A}} \sum_{\tau=0}^{2^p-1} J_{2^p+\tau}(b,a) - J_{2^p+\tau}(B_{2^p+\tau}, A_{2^p+\tau}) \\
&\overset{(*)}{\leq} \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} \sup_{b,a\in\mathcal{A}} \sum_{\tau=0}^{2^p-1} J_{2^p+\tau}(b,a) - J_{2^p+\tau}(B_{(p)}, A_{(p)}) \\
&=: \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} R_{T,\ell}.
\end{aligned}$$

Where $(*)$ follows from the structure of LAZYOPSR, which only switches arms when p changes. When this happens, that is, for $t = 2^p$, we can apply Equations (18) and (19):

$$\sup_{b,a\in\mathcal{A}} J(b,a) - J(B_{(p)}, A_{(p)}) \leq 4\psi(2^p,\delta) + 4\overline{\phi}(2^p,\delta) =: \kappa_1(p,\delta) \tag{29}$$

Coming back to the regret, the following upper bound holds for any p:

$$\begin{aligned}
R_{T,p} \leq & \sup_{b,a\in\mathcal{A}} \sum_{\tau=0}^{2^p-1} F(b)(M_{2^p+\tau} - b) + S(a)(a - M_{2^p+\tau}) \\
& - \sum_{\tau=0}^{2^p-1} F(B_{(p)})(M_{2^p+\tau} - B_{(p)}) + S(A_{(p)})(A_{(p)} - M_{2^p+\tau}) \\
= & \sup_{b,a\in\mathcal{A}} F(b)\left(\sum_{\tau=0}^{2^p-1} M_{2^p+\tau} - 2^p b\right) + S(a)\left(2^p a - \sum_{\tau=0}^{2^p-1} M_{2^p+\tau}\right) \\
& - F(B_{(p)})\left(\sum_{\tau=0}^{2^p-1} M_{2^p+\tau} - 2^p B_{(p)}\right) + S(A_{(p)})\left(2^p A_{(p)} - \sum_{\tau=0}^{2^p-1} M_{2^p+\tau}\right) \\
= & \, 2^p \left[\sup_{b,a\in\mathcal{A}} F(b)\left(\widehat{\mu}_{(p)} - b\right) + S(a)\left(a - \widehat{\mu}_{(p)}\right)\right] \\
& - 2^p \left[F(B_{(p)})\left(\widehat{\mu}_{(p)} - B_{(p)}\right) + S(A_{(p)})\left(A_{(p)} - \widehat{\mu}_{(p)}\right)\right] \qquad \widehat{\mu}_{(p)} := 2^{-p}\sum_{\tau=0}^{2^p-1} M_{2^p+\tau}.
\end{aligned}$$

In the previous result, $\widehat{\mu}_{(p)}$ corresponds to the sample mean of the prices M_t during phase p. Crucially, when compared to the global sample mean at the beginning of the phase, $\widehat{\mu}_{2^p}$ and the one at the end, $\widehat{\mu}_{2^{p+1}}$, the following equation holds true.

$$\widehat{\mu}_{2^{p+1}} = \frac{\widehat{\mu}_{2^p} + \widehat{\mu}_{(p)}}{2} \implies \widehat{\mu}_{(p)} = 2\widehat{\mu}_{2^{p+1}} - \widehat{\mu}_{2^p}.$$

By definition of event E^c, both $\widehat{\mu}_{2^p}$ and $\widehat{\mu}_{2^{p+1}}$ are at most $\overline{\phi}(t,\delta)$-far from the true mean, with $t = 2^p$ / 2^{p+1} respectively. Therefore,

$$|\mu_{(p)} - \mu| \leq 2\overline{\phi}(2^{p+1},\delta) + \overline{\phi}(2^p,\delta) := \kappa_2(p,\delta) \tag{30}$$

We can use Equations (29) and (30) to complete the regret bound on $R_{T,p}$. In fact

$$\begin{aligned}
R_{T,p} &\leq 2^p \left[\sup_{b,a \in \mathcal{A}} F(b) \left(\widehat{\mu}_{(p)} - b \right) + S(a) \left(a - \widehat{\mu}_{(p)} \right) \right] \\
&\quad - 2^p \left[F(B_{(p)}) \left(\widehat{\mu}_{(p)} - B_{(p)} \right) + S(A_{(p)}) \left(A_{(p)} - \widehat{\mu}_{(p)} \right) \right] \\
&\overset{eq.\ (30)}{\leq} 2^p \left[\sup_{b,a \in \mathcal{A}} F(b) \left(\mu - b \right) + S(a) \left(a - \mu \right) + 2\kappa_2(p,\delta) \right] \\
&\quad - 2^p \left[F(B_{(p)}) \left(\mu - B_{(p)} \right) + S(A_{(p)}) \left(A_{(p)} - \mu \right) - 2\kappa_2(p,\delta) \right] \\
&= 2^{p+2} \kappa_2(p,\delta) + 2^p \left[\sup_{b,a \in \mathcal{A}} J(b,a) - J(B_{(p)}, A_{(p)}) \right] \\
&\overset{eq.\ (29)}{\leq} 2^{p+2} \kappa_2(p,\delta) + 2^p \kappa_1(p,\delta).
\end{aligned}$$

The whole term $R_{T,p}$ is thus bounded by

$$2^{p+2} \kappa_2(p,\delta) + 2^p \kappa_1(p,\delta) = 2^{p+2} \psi(2^p,\delta) + 2^{p+3} \overline{\phi}(2^p,\delta) + 2^{p+3} \overline{\phi}(2^{p+1},\delta).$$

From this point on, the proof depends on the exact definition of ψ and $\overline{\phi}$. In both cases of assumption 6 and assumption 5, the former writes as eq. (9). Therefore,

$$2^{p+2} \psi(2^p,\delta) = \sqrt{\frac{3 \log(3T/\delta)}{2^{p+2}}} = \sqrt{12 \cdot 2^p \log(3T/\delta)}.$$

The latter instead corresponds, by theorem 3, to

$$2^{p+3} \overline{\phi}(2^p,\delta) = \begin{cases} \sqrt{16 \cdot 2^p \log(2T/\delta)} + \frac{8\gamma(k+1)}{(1-\gamma)} & \text{Assumption 5} \\ \sqrt{256 \cdot 2^p \log(2T/\delta)} & \text{Assumption 6,} \end{cases}$$

and

$$2^{p+3} \overline{\phi}(2^{p+1},\delta) = \begin{cases} \sqrt{16 \cdot 2^{p+1} \log(2T/\delta)} + \frac{8\gamma(k+1)}{(1-\gamma)} & \text{Assumption 5} \\ \sqrt{256 \cdot 2^{p+1} \log(2T/\delta)} & \text{Assumption 6.} \end{cases}$$

Replacing this values in the total regret achieves, for

$$C_1 = \sqrt{12 \log(3T/\delta)} + \sqrt{16 \log(2T/\delta)} + \sqrt{32 \log(2T/\delta)}, \qquad C_2 = \frac{8\gamma(k+1)}{(1-\gamma)}$$

in assumption 5 and

$$C_1 = \sqrt{12 \log(3T/\delta)} + \sqrt{256 \log(2T/\delta)} + \sqrt{518 \log(2T/\delta)}, \qquad C_2 = 0,$$

the following expression

$$\begin{aligned} R_T &\leq \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} R_{T,\ell} \\ &\leq \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} 2^{p+2}\psi(2^p,\delta) + 2^{p+3}\overline{\phi}(2^p,\delta) + 2^{p+3}\overline{\phi}(2^{p+1},\delta) \\ &\leq \sqrt{T} + \sum_{p=0}^{\lfloor \log T \rfloor} 2^{p/2}C_1 + C_2 \leq \sqrt{T} + \frac{C_1\sqrt{2T}}{\sqrt{2}-1} + C_2\log(T). \end{aligned}$$

Replacing the constants ends the proof. ■

Appendix D. Proofs from Section 5

Theorem 5 *Under Assumptions 1, 2 and 3 and let $\{M_t\}_{t=1}^T$ be any unkown oblivious sequence. For $\kappa = \left\lceil \log(3T)^{\frac{1}{3}} T^{\frac{2}{3}} \right\rceil$, the regret suffered by Algorithm 2 is bounded, in expectation, by*

$$\mathbb{E}[R_T] \leq 5\log(T)^{\frac{1}{3}} T^{\frac{2}{3}} + \mathcal{O}\left(\sqrt{T}\log(T)\right).$$

Proof By definition and Equation (2), the regret can be written as follows, for $J_t(a,b) := F(b)(M_t - b) + S(a)(a - M_t)$

$$\begin{aligned} R_T &= \sup_{b,a\in[0,1]} \sum_{t=1}^T J_t(b,a) - J_t(B_{t-1},A_{t-1}) \\ &= \sqrt{T} + \underbrace{\sup_{b\in\mathcal{A}} \sum_{t=1}^T F(b)(M_t - b) - F(B_{t-1})(M_t - B_{t-1})}_{R_T^{\text{bid}}} \\ &\quad + \underbrace{\sup_{a\in\mathcal{A}} \sum_{t=1}^T S(a)(a - M_t) - S(A_{t-1})(A_{t-1} - M_t)}_{R_T^{\text{ask}}}. \end{aligned}$$

Below, we show how to bound the R_T^{bid} part, relative to bidding. The result for R_T^{ask} follows analogously by just flipping the signs. As Algorithm 2 starts choosing $B_t = 0, A_t = 1$ in the first κ rounds, Proposition 1 ensures that for $\mathcal{A}_t = \mathcal{A}$, with probability at least $1-\delta$

$$\forall x \in \mathcal{A}, \qquad |\widehat{F}(x) - F(x)| \leq \frac{\sqrt{3\log(3T/\delta)}}{2\sqrt{\kappa}}.$$

Fixing $\delta = T^{-1}$, since the bids belong to $[0,1]$,

$$\mathbb{E}\left[R_T^{\text{bid}}\right] \leq \mathbb{E}\Bigg[\underbrace{\sup_{b\in\mathcal{A}}\sum_{t=1}^{T}\widehat{F}(b)(M_t-b)-\widehat{F}(B_{t-1})(M_t-B_{t-1})}_{=:\widetilde{R}_T^{\text{bid}}}\Bigg]+\frac{\sqrt{6\log(3T)}T}{\sqrt{\kappa}}+1.$$

We now need to bound $\widetilde{R}_T^{\text{bid}}$.

$$\begin{aligned}\widetilde{R}_T^{\text{bid}} &= \sup_{b\in\mathcal{A}}\sum_{t=1}^{T}\widehat{F}(b)(M_t-b)-\sum_{t=1}^{T}\widehat{F}(B_{t-1})(M_t-B_{t-1})\\ &\leq \kappa+\underbrace{\sup_{b\in\mathcal{A}}\sum_{t=\kappa+1}^{T}\widehat{F}(b)(M_t-b)-\sum_{t=\kappa+1}^{T}\widehat{F}(B_t)(M_t-B_t)}_{\text{P2}}\\ &\quad+\underbrace{\sum_{t=\kappa+1}^{T}\widehat{F}(B_t)(M_t-B_t)-\sum_{t=\kappa+1}^{T}\widehat{F}(B_{t-1})(M_t-B_{t-1})}_{\text{P2}}.\end{aligned}$$

In the former equations, we have split between the first κ rounds, which are devoted to pure exploration, and the following $T-\kappa$, where the interesting parts of the algorithm take place. After this, the regret is split into two terms, $T1$ taking into account the difference between the clairvoyant and B_t, and $T2$ measuring the difference between B_t and B_{t-1}. This decomposition, is relatively standard in the analysis of follow-the-leader algorithms.

Let us examine the first part. As $0\leq\varepsilon\leq 1/\sqrt{T}$ almost surely, one has

$$\left|\text{P1}-\sup_{b\in\mathcal{A}}\sum_{t=\kappa+1}^{T}\widehat{F}(b)(M_t+\varepsilon-b)+\sum_{t=\kappa+1}^{T}\widehat{F}(B_t)(M_t+\varepsilon-B_t)\right|\leq 2\sqrt{T}.$$

We are going to prove by induction that for any $\kappa\leq K\leq T$,

$$\sup_{b\in\mathcal{A}}\sum_{t=\kappa+1}^{T}\widehat{F}(b)(M_t+\varepsilon-b)-\sum_{t=\kappa+1}^{T}\widehat{F}(B_t)(M_t+\varepsilon-B_t)\leq 0.$$

Now, we perform the inductive step, assuming the thesis holds for K

$$\begin{aligned}\sup_{b\in\mathcal{A}}\sum_{t=\kappa+1}^{K+1}\widehat{F}(b)(M_t+\varepsilon-b) &= \sum_{t=\kappa+1}^{K+1}\widehat{F}(B_{K+1})(M_t+\varepsilon-B_{K+1})\\ &= \sum_{t=\kappa+1}^{K}\widehat{F}(B_{K+1})(M_t+\varepsilon-B_{K+1})+\widehat{F}(B_{K+1})(M_{K+1}+\varepsilon-B_{K+1})\\ &\leq \sup_{b\in\mathcal{A}}\sum_{t=\kappa+1}^{K}\widehat{F}(b)(M_t+\varepsilon-b)+\widehat{F}(B_{K+1})(M_{K+1}+\varepsilon-B_{K+1})\\ &\leq \sum_{t=\kappa+1}^{K}\widehat{F}(B_t)(M_t+\varepsilon-B_t)+\widehat{F}(B_{K+1})(M_{K+1}+\varepsilon-B_{K+1}).\end{aligned}$$

which completes the inductive part. This proves that

$$\text{P1} \leq 2\sqrt{T}. \tag{31}$$

Then, we work on P2. This part is a random variable, and we start proving the following expectation upper bound. Let $f : [0,1] \to [0,1]$ be any fixed function and $\Delta_t := \widehat{\mu}_{t+1} - \widehat{\mu}_t$. Since ε has uniform law on $(0, 1/\sqrt{T})$,

$$\begin{aligned}
\mathbb{E}[f(B_t)] &= \sqrt{T}\int_0^{T^{-1/2}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_{t+1} + x - b))\, dx \\
&= \sqrt{T}\int_0^{T^{-1/2}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_t + \Delta_{t+1} + x - b))\, dx \\
&= \sqrt{T}\int_{\Delta_{t+1}}^{T^{-1/2}+\Delta_{t+1}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_t + x - b))\, dx \\
&= \underbrace{\sqrt{T}\int_0^{T^{-1/2}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_t + x - b))\, dx}_{\mathbb{E}[f(b_t)]} \\
&\quad + \sqrt{T}\int_{T^{-1/2}}^{T^{-1/2}+\Delta_{t+1}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_t + x - b))\, dx \\
&\quad - \sqrt{T}\int_0^{\Delta_{t+1}} f(\text{argmax}_b \widehat{F}(b)(\widehat{\mu}_t + x - b))\, dx \\
&\leq \mathbb{E}[f(B_t)] + \sqrt{T}\Delta_{t+1} \leq \mathbb{E}[f(B_t)] + \sqrt{\frac{T}{(t+1)^2}},
\end{aligned}$$

where the key step is the fact that the sample mean change Δ_t does not exceed $1/(t+1)$ from step t to $t+1$, a consequence of assumption 1. If one is interested in the sum over $\sum_{t=\kappa+1}^{T}$, we get, for any sequence of functions f_t that are independent on the noise,

$$\mathbb{E}\left[\sum_{t=\kappa+1}^{T} f_t(B_t) - f_t(B_{t-1})\right] \leq \sum_{t=\kappa+1}^{T} \frac{\sqrt{T}}{t-\kappa} \leq \sqrt{T}\log(T).$$

Taking $f_t(b) := \widehat{F}(b)(M_t - b)$, that is bounded in $[0,1]$, this shows that

$$\mathbb{E}\left[\sum_{t=\kappa+1}^{T} \widehat{F}(B_t)(M_t - B_t) - \sum_{t=\kappa+1}^{T} \widehat{F}(B_{t-1})(M_t - B_{t-1})\right] \leq \sqrt{T}\log(T),$$

i.e. $\mathbb{E}[\text{P2}] \leq \sqrt{T}\log(T)$. Toghether with Equation (31), this proves that

$$\mathbb{E}[\widetilde{R}_T^{\text{bid}}] \leq \kappa + 2\sqrt{T} + \sqrt{T}\log(T).$$

As anticipated, the same passages show that $\mathbb{E}[\widetilde{R}_T^{\text{ask}}] \leq \kappa + 2\sqrt{T} + \sqrt{T}\log(T)$. Putting everything together,

$$\begin{aligned}
\mathbb{E}[R_T] &\leq \sqrt{T} + \mathbb{E}\left[R_T^{\text{bid}}\right] + \mathbb{E}\left[R_T^{\text{ask}}\right] \\
&\leq \sqrt{T} + \frac{\sqrt{6\log(3T)}T}{\sqrt{\kappa}} + 2 + \mathbb{E}\left[\widetilde{R}_T^{\text{bid}}\right] + \mathbb{E}\left[\widetilde{R}_T^{\text{ask}}\right] \\
&\leq \sqrt{T} + \frac{\sqrt{6\log(3T)}T}{\sqrt{\kappa}} + 2 + 2\kappa + 4\sqrt{T} + 2\sqrt{T}\log(T).
\end{aligned}$$

Replacing the value of $\kappa = \left\lceil \log(3T)^{\frac{1}{3}} T^{\frac{2}{3}} \right\rceil$, one gets,

$$R_T \leq 5\log(3T)^{\frac{1}{3}} T^{\frac{2}{3}} + \mathcal{O}\left(\sqrt{T}\log(T)\right).$$

■