Market Making without Regret

Nicolò Cesa-Bianchi   NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano and Politecnico di Milano, Milan, Italy

Tommaso Cesari   TCESARI@UOTTOWA.CA
University of Ottawa, Ottawa, Canada

Roberto Colomboni   ROBERTO.COLOMBONI@POLIMI.IT
Politecnico di Milano and Università degli Studi di Milano, Milan, Italy

Luigi Foscari   LUIGI.FOSCARI@UNIMI.IT
Università degli Studi di Milano, Milan, Italy

Vinayak Pathak   PATH.VINAYAK@GMAIL.COM
Independent researcher

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider a sequential decision-making setting where, at every round t, the learner (a *market maker*) posts a *bid* price B_t and an *ask* price A_t to an incoming trader (the *taker*) with a private valuation for some asset. If the trader's valuation is lower than the bid price, or higher than the ask price, then a trade (sell or buy) occurs. Letting M_t be the market price (observed only at the end of round t), the maker's utility is $M_t - B_t$ if the maker bought the asset, it is $A_t - M_t$ if they sold it, and it is 0 if no trade occurred. We characterize the maker's regret with respect to the best fixed choice of bid and ask pairs under a variety of assumptions (adversarial, i.i.d., and their variants) on the sequence of market prices and valuations. Our upper bound analysis unveils an intriguing connection relating market making to first-price auctions and dynamic pricing. Our main technical contribution is a lower bound for the i.i.d. case with Lipschitz distributions and independence between market prices and takers' valuations. The difficulty in the analysis stems from a unique relationship between the reward and feedback functions that allows learning algorithms to trade off reward for information in a continuous way.

Keywords: Regret minimization, online learning, market making, first-price auctions, dynamic pricing.

1. Introduction

Trading in financial markets is a crucial activity that helps keep the world's economy running, and several players, including hedge funds, prop trading firms, investment banks, central banks, and retail traders participate in it daily. While every actor has their own objective function (for example, a hedge fund wants to maximize profit whereas a central bank wants to keep inflation in check), at a fundamental level, trading can be viewed as a stochastic control problem where agents want the state to evolve so as to maximize their objective function. In this work we focus on market makers, i.e., traders whose job is to facilitate other trades to happen. One way to do so is by broadcasting, at all times, a price (*bid*) at which they are willing to buy, and a price (*ask*) at which they are willing to sell the asset being traded. This way when a buyer (resp., seller) arrives, they do not have to wait for a seller (resp., buyer) to be able to perform a transaction. A market where it is easy to make trades is called *liquid*, and liquidity is a desirable property for any kind of market. A

market maker thus provides an essential service by increasing the liquidity of the market. Market making is challenging, and a lot of thought goes into making it profitable, see Harris (2003) for an overview. A major risk that a market maker has to deal with, called adverse selection, is the risk that your counterpart is an *informed* trader who knows something about the future direction of the price movement. For example, an informed trader who knows that an asset is soon about to become cheaper will sell it to you thus forcing you to buy something whose price will crash. One way to mitigate this risk is to immediately offload your positions elsewhere. More generally, a market maker does not specialize in predicting the future movements of prices, and thus dislikes holding assets for too long. Market making desks often have limits on how much inventory they are allowed to hold at a time (Fender and Lewrick, 2015).

The strategy of immediately offloading your position, though simple, is profitable if one does it strategically. Consider, for example, the case where one can trade the same asset on two exchanges, where its bid-ask spread on the first is smaller than its bid-ask spread on the second. In this case, one can make a market on the second while offloading their positions on the first. This situation arises frequently for two closely related assets. For example, a *futures* contract on an asset is often more liquid than the asset itself and their price movements are highly correlated. For these reasons, we focus on a setting where the market maker offloads their positions shortly after each transaction.

1.1. Related works

Online market making. Our work is closely related to that of Abernethy and Kale (2013), which analyzes competition against N constant-spread dynamic strategies. While we also consider constant-spread (static) strategies, key differences make their results non-comparable to ours. Their model assumes full information feedback, with an adversarial yet bounded market price change and a fully known utility function at each step, leading to a regret of $\sqrt{T \ln N}$. In contrast, we compete against a continuum of ask-bid strategies while managing both estimation and approximation errors. Crucially, our model lacks full information feedback, as rewards depend on private valuations, requiring structural exploitation of the reward function to mitigate uncertainty.

Our work is also related to the classic Glosten-Milgrom model (Glosten and Milgrom, 1985), which introduced the concept of a market *specialist* (maker) providing liquidity to the market within an exchange. The specialist interacts with both informed traders (whose valuations are influenced by future market prices) and uninformed ones. The specialist's objective is to distinguish between informed and uninformed traders via interactions in an online fashion and optimize the bid-ask spread accordingly. In 2005, Das considers a learning algorithm in an extended version of the Glosten-Milgrom model for the market maker with a third type of "noisy-informed" trader, whose current valuation is a noisy version of the future market price. However, to the best of our knowledge, they do not provide any regret guarantees.

Smoothed adversary. Popularized by Spielman and Teng (2004); Rakhlin et al. (2011); Haghtalab et al. (2020), smoothness analysis provides a framework for analyzing algorithms in problems parameterized by distributions that are not "too" concentrated. Recent advancements in the smoothed analysis of online learning algorithms include contributions by Kannan et al. (2018); Haghtalab et al. (2022, 2021); Block et al. (2022); Durvasula et al. (2023); Cesa-Bianchi et al. (2021, 2023); Bolić et al. (2024); Cesa-Bianchi et al. (2024a). In this work, we exploit the connection between the smoothness of the takers' valuations distributions (i.e., the Lipschitzness of the corresponding cumulative distribution function, or, equivalently, the boundedness of the corre-

sponding probability density function) and Lipschitzness of the expected utility. This property is crucial to achieving sublinear regret guarantees in adversarial settings and affects learning depending on different assumptions on the level of information shared with the learner while bidding.

Please see Appendix B for further related works.

Online learning in digital markets. Our work unveils a connection between market making and first-price auctions/dynamic pricing, thus contributing to the existing body of research on the multi-armed bandit framework, which boasts a substantial literature base across statistics, operations research, computer science, and economics (Auer et al., 2002a; Lattimore and Szepesvári, 2020; Badanidiyuru et al., 2018; Kleinberg, 2004). The multi-armed bandit framework has proven to be an effective tool for addressing key problems in finance under uncertainty, including repeated first-price auctions (Cesa-Bianchi et al., 2024c; Han et al., 2024; Achddou et al., 2021; Flajolet and Jaillet, 2017) and dynamic pricing (Cesa-Bianchi et al., 2019; Babaioff et al., 2015; Kleinberg and Leighton, 2003). These settings are closely related to online market making (see Appendix B for further discussion).

Closely related to dynamic pricing is the problem of optimal taxation (Cesa-Bianchi et al., 2025), where the interaction between the learner and the environment follows a similar structure, but the objective shifts from revenue maximization to the design of pricing mechanisms that maximize social welfare.

Beyond first-price auction settings, related auction formats include repeated Vickrey auctions, studied from both the bidder's perspective (Weed et al., 2015) and the seller's perspective (Cesa-Bianchi et al., 2013).

Another important problem at the intersection of online learning and digital markets is repeated bilateral trade. In this setting, the learner acts as a broker, proposing price pairs to a sequence of traders with unknown private valuations. The foundational works of Cesa-Bianchi et al. (2021, 2024b) established optimal regret rates under i.i.d. valuations, assuming smoothness and independence between buyer and seller values, using the gain from trade as reward function, and mechanisms that are strongly budget balanced. This line of work has been extended in multiple directions: Azar et al. (2022) proposed weakly budget-balanced mechanisms achieving optimal 2-regret in adversarial settings; Cesa-Bianchi et al. (2023, 2024a) relaxed the i.i.d. and the independence between sellers and buyers assumptions while maintaining optimal regret guarantees under smoothness alone; and Bernasconi et al. (2024) introduced globally budget-balanced mechanisms with sublinear regret even in adversarial settings.

Contextual versions of bilateral trade have also been studied by Gaucher et al. (2025), while Bachoc et al. (2024) took a different perspective by proposing a fairness-oriented reward function establishing optimal regret bounds under the i.i.d. assumption with independent buyer and seller valuations, using strongly budget-balanced mechanisms.

Recent work has also considered scenarios in which agents are not assigned fixed roles as buyers or sellers. In this more general setting, Bolić et al. (2024) achieved faster regret rates than in standard bilateral trade using strongly budget-balanced mechanisms under the same assumptions as Cesa-Bianchi et al. (2021). In the contextual setting, Bachoc et al. (2025b) obtained optimal fast rates under a noisy-linear model, and Bachoc et al. (2025a) extended these results to nonparametric contexts. Finally, Cesari and Colomboni (2025) studied the objective of maximizing the volume of trade, achieving accelerated regret rates in this setting.

Setting	(adv)	(adv+lip)	(iid)	(iid+lip)	(iid+iv)	(iid+lip+iv)
Regret	T	$T^{2/3}$	T	$T^{2/3}$	$T^{2/3}$	$T^{2/3}$

Table 1: A summary of the regret guarantees we prove for each variant of the online market-making problem. All the rates are optimal up to constant factors.

1.2. Our contributions

We study the question of market making in an online learning setup. Here, trading happens in discrete time steps and an unknown stochastic process governs the prices of the asset being traded as well as the private valuations assigned to the asset by market participants (see details in Section 2). The market maker is an online learning algorithm that posts a bid and an ask at the beginning of each time step and receives some feedback at the end of the time step. We are interested in controlling the regret suffered by this learner at the end of T time steps. At the end of each round, the learner observes the market price, i.e., the price at which they are able to offload their position. Additionally, they see whether their bid or their ask (or neither) got successfully traded. We consider settings where the unknown stochastic process is independent and identically distributed (iid), as well as the adversarial setting (adv), where no such assumption is made on the process. We also consider settings where the cumulative distribution functions of takers' valuations are Lipschitz (lip), as well as settings where at each time step t the corresponding market price and taker's valuation are independent of each other (iv). Our results are summarized in Table 1.

We make the following contributions.

1. We design the M3 (meta-)algorithm (Algorithm 2) and prove a $\mathcal{O}(T^{2/3})$ upper bound on its regret under the assumption that either the cumulative distribution functions of the takers' valuations are Lipschitz (lip) or the sequence of market prices and takers' valuations are independent and identically distributed (iid) with market prices being independent of takers' valuations (iv) (Theorem 1).

2. Our main technical contribution is a lower bound of order $\Omega(T^{2/3})$ that matches M3's upper bound, and holds even under the simultaneous assumptions that the sequence of market prices and takers' valuations is independent and identically distributed (iid), market prices are independent of takers' valuations (iv), the cumulative distribution function of the takers' valuations is Lipschitz (lip), and the cumulative distribution function of the market prices is Lipschitz (Theorem 2).

3. We then investigate the necessity of assuming either Lipschitzness of the cumulative distribution function of the takers' valuations (lip) or the independence of market prices and takers' valuations (iv). We prove that, if both assumptions are dropped, learning is impossible in general, even when market prices and takers' valuations are independent and identically distributed (iid) (Theorem 3).

We can interpret the independence between takers' valuation and market prices (iv) as the taker being an uninformed trader (in the same spirit as the Glosten-Milgrom model). If we consider the (iid) setting, and contrast the impossibility of learning when neither Lipschitzness (lip) nor independence (iv) is assumed with the $\mathcal{O}(T^{2/3})$ upper bound when the market prices are independent of takers' valuations (iv), we get evidence for the practitioner's intuition that the existence of uninformed traders is good for the market maker.

Trading protocol 1: Online market making

---

for *time* $t = 1, 2, \ldots$ **do**

    Taker arrives with a private valuation $V_t \in [0, 1]$;

    Maker posts bid/ask prices $(B_t, A_t) \in \mathcal{U}$;

    Feedback $\mathbb{I}\{B_t \geq V_t\}$ and $\mathbb{I}\{A_t < V_t\}$ is revealed;

    if $B_t \geq V_t$ **then** Maker buys and pays B_t to Taker;

    else if $A_t < V_t$ **then** Maker sells (short)[1] and is paid A_t by Taker;

    else no trade happens;

    The market price $M_t \in [0, 1]$ is revealed;

    if *Maker bought from Taker* **then** Maker offloads to Market at price M_t;

    else if *Maker sold (short) to Taker* **then** Maker buys from Market at price M_t;

    Maker's utility is $U_t(B_t, A_t) \coloneqq u(B_t, A_t, M_t, V_t)$;

end

---

2. Setting

In online market making, the action space of the maker is the upper triangle $\mathcal{U} \coloneqq \big\{(b, a) \in [0, 1]^2 : b \leq a\big\}$, enforcing the constraint that a bid price b is never larger than the corresponding ask price a. For an overview of the notation, see Table 2, in Appendix A. The utility of the market maker, for all $(b, a) \in \mathcal{U}$ and $m, v \in [0, 1]$, is

$$u(b, a, m, v) \coloneqq \underbrace{(m - b)}_{\text{market-bid spread}} \times \underbrace{\mathbb{I}\{b \geq v\}}_{\text{Maker buys}} + \underbrace{(a - m)}_{\text{ask-market spread}} \times \underbrace{\mathbb{I}\{a < v\}}_{\text{Maker sells}}, \tag{1}$$

where v is the taker's private valuation and m is the market price.[2] Our online market making protocol is specified in Trading Protocol 1. At every round t, a taker arrives with a private valuation $V_t \in [0, 1]$ and the maker posts bid/ask prices $(B_t, A_t) \in \mathcal{U}$. If $B_t < V_t \leq A_t$ (i.e., if the taker is not willing to sell nor to buy at the proposed bid/ask prices), then no trade happens. If $B_t \geq V_t$ (buy) or $A_t < V_t$ (sell), a trade happens. At the end of each round, the maker observes the market price M_t and the type of trade (buy, sell, none) that took place in that round. For each $(b, a) \in \mathcal{U}$ and each round t, defining $U_t(b, a) \coloneqq u(b, a, M_t, V_t)$, the maker's utility at round t is $U_t(B_t, A_t) = u(B_t, A_t, M_t, V_t)$. Hence, any (possibly randomized) learning strategy for the maker at time t is a map that takes as input past feedback $\big\{\big(\mathbb{I}\{B_\tau \geq V_\tau\}, \mathbb{I}\{A_\tau < V_\tau\}, M_\tau\big)\big\}_{\tau < t}$, plus some optional random seeds, and returns as output a (random) pair (B_t, A_t) of bid/ask prices in $\mathcal{U}$.

---

1. We remark that in most markets, traders can sell assets they do not currently own (*short-selling*; see, e.g., the classic Black and Scholes 1973). For this reason, we do not assume that the market maker owns a unit of the asset before selling it. If the market maker sells short, they will buy the asset at the market price at the end of the interaction.

2. The choice of buying when $b \geq v$ and selling when $a < v$ models a taker that is slightly more inclined to sell rather than buy: in this case, the taker is willing to sell even when their valuation is exactly equal to the bid. This choice is completely immaterial for the results that follow and is merely done for the sake of simplicity. All the results we present still hold if trades happen according to a similar rule but with a taker that is slightly more inclined to buy rather than sell (i.e., if a trade occurs whenever $b > v$ or $a \leq v$) or if this inclination changes arbitrarily whenever a new taker comes at any new time step.

The maker's goal is to minimize, for any time horizon $T \in \mathbb{N}$, the regret after T time steps:

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

where the expectations are with respect to the (possible) randomness of the market prices, takers' valuations, and (possibly) the internal randomization of the algorithm.

We study this problem under a mix of several standard assumptions on the underlying process $(V_t, M_t)_{t \in \mathbb{N}}$ modeling private valuations and market prices.

(adv) $(V_t, M_t)_{t \in \mathbb{N}}$ is a stochastic process.

(iid) $(V_t, M_t)_{t \in \mathbb{N}}$ is an i.i.d. stochastic process.[3]

(iv) $(V_t, M_t)_{t \in \mathbb{N}}$ is a stochastic process such that, for all $t \in \mathbb{N}$, V_t is independent of M_t.[4]

(lip) $(V_t, M_t)_{t \in \mathbb{N}}$ is a stochastic process such that, there exists $L > 0$ such that, for all $t \in \mathbb{N}$, the cumulative distribution function of V_t is L-Lipschitz.

The variables A_t, B_t, V_t, M_t can be thought of as relative to a global reference value revealed at the beginning of each time step, aligning with the convention of quoting in "points" (or "ticks") above or below a benchmark. With this interpretation, in the regret definition, the benchmark strategy maintains a constant distance from a reference throughout the trading period (as opposed to a constant absolute level). These observations greatly broaden the scope of application of our setting.

3. $T^{2/3}$ Upper Bound (adv+lip) or (iid+iv)

Since the learner's action space contains a continuum of actions, standard bandit approaches require Lipschitzness (or the weaker one-side Lipschitzness) of the reward function, a property that is unfortunately missing in our setting, even when the environment is stable (iid) and the takers are uninformed (iv). One way to recover Lipschitzness (although only in expectation) is by assuming the takers' valuations are smoooth, i.e., by assuming that their cumulative distribution functions are Lipschitz (lip). Under the (lip) assumption, recalling that our action space $\mathcal{U}$ is the set of bid/ask pairs $\{(b, a) \in [0, 1]^2 \mid b \leq a\}$, and noticing that the feedback the learner receives at the end of each time step is sufficient to reconstruct bandit feedback, the market making problem can be seen as an instance of 2-dimensional Lipschitz bandits. However, a black-box application of existing d-dimensional Lipschitz bandits techniques (Slivkins, 2019, with $d = 2$) would give only a suboptimal regret rate of $\mathcal{O}(T^{(d+1)/(d+2)}) = \mathcal{O}(T^{3/4})$.

Using a different approach that exploits the structure of our feedback and reward functions, we manage not only to solve the non-Lipschitz (iid+iv) setting, but also to effectively reduce the dimensionality of the problem by 1 under the smoothness assumption, obtaining an optimal $\mathcal{O}(T^{2/3})$ regret rate in all cases. To do so, we first note that the utility of our market-making problem can be viewed as the sum of the utilities of two related sub-problems: the first addend in (1) corresponds to the utility in a repeated first-price auction problem with unknown valuations, while the second addend corresponds to the utility in a dynamic pricing problem with unknown costs (see below). This

---

3. Note that this assumption alone does not require that, for any $t \in \mathbb{N}$, V_t is independent of M_t.
4. Note that this assumption alone does not require that $(V_t, M_t)_{t \in \mathbb{N}}$ is an independent stochastic process.

suggests trying to use two algorithms for the two problems to solve our market-making problem. However, in our problem we have the further constraint that the bid we propose in the first-price auction is no greater than the price we propose in the dynamic pricing problem. This constraint prevents us from running directly two independent no-regret algorithms for the two different problems because the bid/ask pair produced this way may violate the constraint. To get around this roadblock, we present a meta-algorithm (Algorithm 2) that takes as input two algorithms for the two sub-problems and combines them to post pairs in $\mathcal{U}$ while preserving the respective guarantees. We now formally introduce the two related problems and build the explicit reduction.

Repeated first-price auctions with unknown valuations (FPA). Consider the following online problem (denoted by FPA) of repeated first-price auctions where the learner only learns the valuation of the good they are bidding on *after* the auction is cleared. At any time t, two $[0, 1]$-valued random variables Z_t and H_t are generated and hidden from the learner: Z_t is the current valuation of the good and H_t is the highest competing bid for the good, both unknown to the learner at the time of bidding. Then, the learner bids an amount $X_t \in [0, 1]$ and wins the bid if and only if their bid is higher than or equal to the highest competing bid. The utility of the learner is $Z_t - X_t$ if they win the auction and zero otherwise. Finally, the valuation of the good Z_t and the indicator function $\mathbb{I}\{X_t \geq H_t\}$ (representing whether or not the learner won the auction) are revealed to the learner.

The goal is to minimize, for any time horizon $T \in \mathbb{N}$, the regret after T time steps

$$R_T^{\mathrm{fpa}} := \sup_{x \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} (Z_t - x)\,\mathbb{I}\{x \geq H_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T} (Z_t - X_t)\,\mathbb{I}\{X_t \geq H_t\}\right]\,,$$

where the expectations are taken with respect to the (possible) randomness of $(Z_1, H_1), \ldots, (Z_T, H_T)$ and the (possible) internal randomization of the algorithm that outputs the bids $X_1, \ldots, X_T$.

The same setting was studied by Cesa-Bianchi et al. (2024c), but with different feedback models.

Dynamic pricing with unknown costs (DP). Consider the following online problem (denoted by DP) of dynamic pricing where the learner only learns the current cost of the item they are selling *after* interacting with the buyer. At any time t, two $[0, 1]$-valued random variables C_t and W_t are generated and hidden from the learner: C_t is the current cost of the item, and W_t is the buyer's valuation for the item. Then, the learner posts a price $P_t \in [0, 1]$ and the buyer buys the item if and only if their valuation is higher than the posted price. The utility of the learner is $P_t - C_t$ if the buyer bought the item and zero otherwise. Finally, the cost of the item C_t and the indicator function $\mathbb{I}\{P_t < W_t\}$ (representing whether or not the buyer bought the item) are revealed to the learner.

The goal is to minimize, for any time horizon $T \in \mathbb{N}$, the regret after T time steps

$$R_T^{\mathrm{dp}} := \sup_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} (p - C_t)\,\mathbb{I}\{p < W_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T} (P_t - C_t)\,\mathbb{I}\{P_t < W_t\}\right]\,,$$

where expectations are with respect to the (possible) randomness of $(C_1, W_1), \ldots, (C_T, W_T)$ and the (possible) internal randomization of the algorithm that outputs the posted prices $P_1, \ldots, P_T$.

Dynamic pricing is generally studied under the assumption that costs are known and all equal to zero, and assuming that the buyer buys whenever their evaluation is greater than *or equal to* the posted price—see, e.g., (Kleinberg and Leighton, 2003; Cesa-Bianchi et al., 2019; Leme et al., 2021; Luo et al., 2024).

Algorithm 2: M3 (Meta Market Making)

---

input: Algorithm $\mathcal{A}^{\text{fpa}}$ for FPA and algorithm $\mathcal{A}^{\text{dp}}$ for DP

for *time* $t = 1, 2, \ldots$ **do**

> Let X_t be the output of $\mathcal{A}^{\text{fpa}}$ at time t and let P_t be the output of $\mathcal{A}^{\text{dp}}$ at time t;
> **if** $X_t \leq P_t$ **then** let $B_t := X_t$ and $A_t := P_t$;
> **else if** $X_t > P_t$ **then** let $B_t := P_t$ and $A_t := X_t$;
> Post bid/ask prices (B_t, A_t) and observe feedback $\mathbb{I}\{B_t \geq V_t\}, \mathbb{I}\{A_t < V_t\}$, and M_t;
> **if** $X_t \leq P_t$ **then** feed $(\mathbb{I}\{B_t \geq V_t\}, M_t)$ to $\mathcal{A}^{\text{fpa}}$ and $(\mathbb{I}\{A_t < V_t\}, M_t)$ to $\mathcal{A}^{\text{dp}}$;
> **else if** $X_t > P_t$ **then** feed $(1 - \mathbb{I}\{A_t < V_t\}, M_t)$ to $\mathcal{A}^{\text{fpa}}$ and $(1 - \mathbb{I}\{B_t \geq V_t\}, M_t)$ to $\mathcal{A}^{\text{dp}}$;

end

Solving FPA and DP solves market making. In this section, we introduce a meta-algorithm that combines two sub-algorithms, one for repeated first-price auctions with unknown valuations and one for dynamic pricing with unknown costs. We thus obtain an algorithm for market-making and we show in Theorem 1 that its regret can be upper bounded by the sum of the regrets of the two sub-algorithms for the two corresponding sub-problems. The idea of our meta-algorithm M3 (Algorithm 2) begins with the observation that the utility of the market maker is a sum of two terms that correspond to the utilities of the learner in first-price auctions and dynamic pricing, respectively. This suggests maintaining two algorithms in parallel, one to determine bid prices and one to determine ask prices. M3 then enforces the constraint that bid prices should be no larger than ask prices by swapping the recommendations of the two sub-algorithms whenever they generate corresponding bid/ask prices that would violate the constraint. Finally, M3 leverages the available feedback $(\mathbb{I}\{B_t \geq V_t\}, \mathbb{I}\{A_t < V_t\}, M_t)$ at time t to reconstruct and relay back to the two sub-algorithms the counterfactual feedback they would have observed if the swap didn't happen (i.e., if the learner always posted the prices determined by the sub-algorithms).

The next result shows that the regret of M3 is upper bounded by the sum of the regrets of the two sub-algorithms, which in turn leads to optimal regret guarantees for the market making problem.

Theorem 1 *In the online market-making problem, the regret of M3 run with parameters $\mathcal{A}^{\text{fpa}}$ and $\mathcal{A}^{\text{dp}}$ over the sequence $(M_t, V_t)_{t \in [T]}$ of market prices and takers' valuations satisfies, for all T, $R_T \leq R_T^{\text{fpa}} + R_T^{\text{dp}}$, where R_T^{fpa} (resp., R_T^{dp}) is the regret of $\mathcal{A}^{\text{fpa}}$ (resp., $\mathcal{A}^{\text{dp}}$) in the FPA (resp., DP) problem over the sequence $(Z_t, H_t)_{t \in [T]} := (M_t, V_t)_{t \in [T]}$ (resp., $(C_t, W_t)_{t \in [T]} := (M_t, V_t)_{t \in [T]}$). Furthermore, for any horizon T, there exist $\mathcal{A}^{\text{fpa}}$ and $\mathcal{A}^{\text{dp}}$ such that, for any sequence $(M_t, V_t)_{t \in [T]}$ of market prices and takers' valuations, if one of the two following conditions holds:*

1. *For each $t \in [T]$, the cumulative distribution function of V_t is L-Lipschitz, for some $L > 0$.*

2. *The process $(M_t, V_t)_{t \in [T]}$ is i.i.d. and, for each $t \in [T]$, M_t is independent of V_t.*

Then, the regret of M3 run with parameters $\mathcal{A}^{\text{fpa}}$ and $\mathcal{A}^{\text{dp}}$ satisfies $R_T \leq cT^{2/3}$, with $c \leq 2L + 100$ (resp., $c \leq 102$) if Item 1 (resp., Item 2) holds.

Proof sketch First, notice that the feedback received by the two algorithms is equal to the feedback they would have retrieved in the corresponding instances regardless of whether or not the prices were

swapped. Hence, the regret guarantees of the prices posted by the two algorithm on the original sequence still hold true. Notice that, for each $t \in \mathbb{N}$, we have that

$$(\mathrm{I}_t) := U_t(B_t, A_t) \geq (M_t - X_t)\mathbb{I}\{X_t \geq V_t\} + (P_t - M_t)\mathbb{I}\{V_t > P_t\} =: (\mathrm{II}_t) .$$

In fact, if $X_t \leq P_t$ the two quantities are equal because $B_t = X_t$ and $A_t = P_t$. Instead, if $P_t < X_t$, then $B_t = P_t$ and $A_t = X_t$, and a direct computation shows that if $V_t \leq P_t$ then $(\mathrm{I}_t) = M_t - P_t > M_t - X_t = (\mathrm{II}_t)$; if $P_t < V_t \leq X_t$ then $(\mathrm{I}_t) = 0 > P_t - X_t = (\mathrm{II}_t)$; and if $X_t < V_t$ then $(\mathrm{I}_t) = X_t - M_t > P_t - M_t = (\mathrm{II}_t)$. It follows that

$$
\begin{aligned}
R_T = \ & \sup_{(b,a)\in\mathcal{U}} \mathbb{E}\left[\sum_{t=1}^{T} U_t(b,a)\right] - \mathbb{E}\left[\sum_{t=1}^{T}(\mathrm{I}_t)\right] \leq \sup_{(b,a)\in\mathcal{U}} \mathbb{E}\left[\sum_{t=1}^{T} U_t(b,a)\right] - \mathbb{E}\left[\sum_{t=1}^{T}(\mathrm{II}_t)\right] \\
\leq \ & \sup_{b\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T}(M_t - b)\mathbb{I}\{b \geq V_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T}(M_t - X_t)\mathbb{I}\{X_t \geq V_t\}\right] \\
& + \sup_{a\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T}(a - M_t)\mathbb{I}\{a < V_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T}(P_t - M_t)\mathbb{I}\{P_t < V_t\}\right] = R_T^{\mathrm{fpa}} + R_T^{\mathrm{dp}} .
\end{aligned}
$$

For the second part of the result, $\mathcal{A}^{\mathrm{fpa}}$ and $\mathcal{A}^{\mathrm{dp}}$ can be chosen by designing discretized versions of appropriately defined bandit algorithms. In the Lipschitz case, the proof follows from standard arguments showing that the expected reward of any price can be controlled by one of the closest points in the discretization grid. The (iid+iv) case is more subtle and it is based on the fact that, although we cannot accurately approximate the expected reward of all prices, under these assumptions, it is possible to approximate the expected rewards of near-optimal prices, which turns out to be sufficient to obtain the result. For all missing details, see Appendix C. ∎

4. $T^{2/3}$ Lower Bound (iid+lip+iv)

In this section, we present the main result of this work, a lower bound showing that no algorithm can achieve a worst-case regret better than $\Omega(T^{2/3})$ even when the sequence of market prices and takers' valuations is i.i.d. with a smooth distribution, and market prices and takers' valuations are independent of each other.

To give the reader some intuition on the nuances and peculiarities of this lower bound, we begin by comparing the trade-offs between regret and the amount of received feedback in hard instances of our setting and those of other (better-known) related settings. The three main ingredients in the standard lower bound approach for online learning with partial feedback are:

- *Hard instances*, constructed as "ε-perturbations" of a base instance obtained by slightly altering the base distributions of some of the random variables drawn by the environment.

- The *amount of information* acquired when playing an action, quantified by the Kullback-Leibler (KL) divergence between the distribution of the feedback obtained when selecting the action in the perturbed instance and the corresponding feedback distribution when selecting the action in the base instance—the higher, the better for the learner.

- The *regret* of an action, simply measured by the expected instantaneous regret of the action in the perturbed instance—the smaller, the better for the learner.
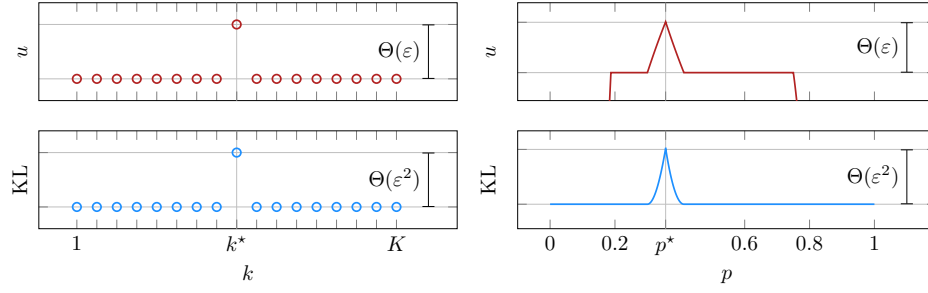
Figure 1: Utility (in red) and amount of acquired information (in blue) in a hard instance of the classic multi-armed bandit problem with a finite set of arms (on the left) and in a hard instance of the first-price auctions problem with a continuous set of bids (on the right)—a similar picture can be drawn for the dynamic pricing problem. Both quantities are maximized at the same point. In particular where there is no perturbation, there is no information and some amount of reward, otherwise the KL can grow up to a quantity of order ε^2 and the utility can grow by a quantity of order at least ε.

In hard instances of K-armed bandits, an arm is drawn uniformly at random from a K-sized set and assigned an expected reward ε-higher than the other arms in the set. Selecting an optimal arm costs zero regret and provides $\Theta(\varepsilon^2)$ bits of information. Selecting a non-optimal arm, instead, provides no information and costs ε regret (Auer et al., 2002b). This implies that any algorithm has only one way to acquire information about an arm: playing that arm. A similar situation occurs in other settings, including those with continuous decision spaces. For example, to acquire information about a bid in a standard hard instance of first-price auctions (or dynamic pricing), the only reasonable option is to bid that amount, incurring regret $\Theta(\varepsilon)$ if the bid is non-optimal and acquiring $\Theta(\varepsilon^2)$ bits of information if the arm is optimal (Cesa-Bianchi et al., 2024c). See Figure 1 for a pictorial representation of the expected reward and KL divergence in said hard instances.

Analogously, to build hard instances of market-making, we begin by drawing a price $\widetilde{b}$ uniformly at random from a known finite set of bid prices and the bid/ask pair $(\widetilde{b}, 1)$ is given a reward ε-higher than the other (potentially optimal) bid/ask pairs of the form $(b', 1)$. However, in the market-making problem, there is a trade-off between reward and feedback that is completely absent in the aforementioned classical problems. The amount of information and corresponding regret of each pair of prices (b, a) can be visualized by looking at the qualitative plots in Figure 2. Here, there are uncountably many ways to determine if a pair of prices $(b, 1)$ is optimal: playing any pair in the non-blue area of the left plot in Figure 2 allows to modulate the amount of exploration, with the points that yield the highest quality feedback being the ones with low reward and high regret (for a plot representing the growth of the amount of feedback and the corresponding reward as a function of the pair (b, a) taken along a representative path, see Figure 3). Consequently, the exploration-exploitation trade-off consists in *first* choosing whether to exploit with some pair $(b, 1)$ or instead to explore another potentially optimal pair $(b', 1)$, and *then*, in the second case, choosing from an uncountable set of options (b'', a'') how much regret one is willing to suffer in order to acquire more information about the (potential) optimality of $(b', 1)$.[5] To build a lower bound, we then need

---

5. The situation is even more complex because pairs (b, a) allow to test for the optimality not just of the pair $(b, 1)$, but also of the pair $(a, 1)$.
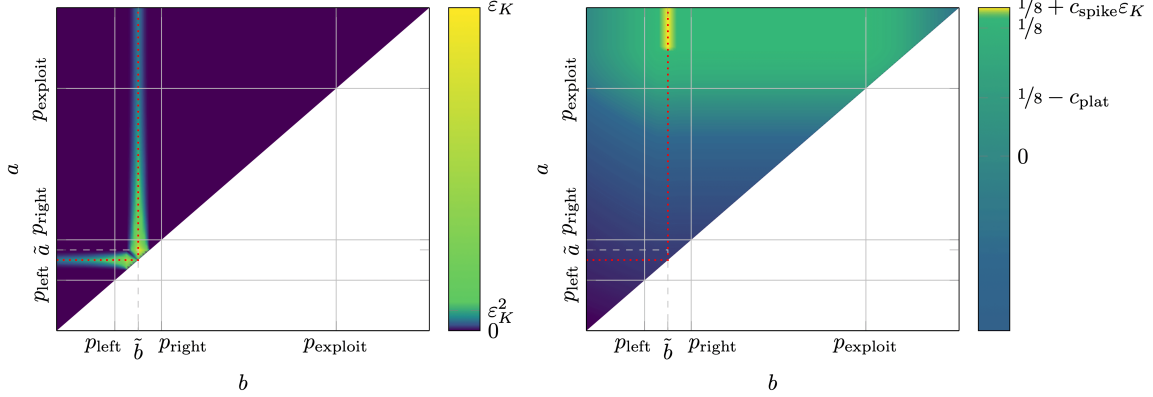
Figure 2: The heat maps of the amount of information (on the left) and the expected utilities (on the right) obtained by selecting pairs of prices in a hard instance we constructed in our $\Omega(T^{2/3})$ lower bound for i.i.d. smooth instances with market prices independent of takers' valuations (Theorem 2). To distinguish the optimal region with reward of (at least) $1/8 + c_{\text{spike}}\varepsilon_K$ (the yellow region at the top of the right plot) from the suboptimal region with reward $1/8$, one could play any point in the non-blue horizontal and vertical regions on the left plot. Note that the pairs that yield the highest amount of feedback (the two small yellow neighborhoods close to the diagonal on the left plot) give highly suboptimal rewards, which are at least $c_{\text{plat}} = \Omega(1)$ below $1/8$. For explicit definitions of $\varepsilon_K, c_{\text{spike}}, c_{\text{plat}}$, see Appendix D.

to consider *all* possible ways a learner can decide to leverage these multiple trade-offs between information and regret that arise in the market making problem.

Theorem 2 *In the online market-making problem, for each $L \geq 8$, each time horizon $T \geq 42$, and each (possibly randomized) algorithm for online market making, there exists a $[0,1]^2$-valued i.i.d. sequence $(M_t, V_t)_{t \in [T]}$ of market prices and takers' valuations such that for each $t \in [T]$ the two random variables M_t and V_t are independent of each other, they admit L-Lipschitz cumulative distribution functions, and the regret of the algorithm over the sequence $(M_t, V_t)_{t \in [T]}$ is lower bounded by $R_T \geq cT^{2/3}$ where $c \geq 10^{-6}$ is a constant.*

We present here a proof sketch and defer the full proof of this result to Appendix D.

Proof sketch The first step is to build a base joint distribution over market prices and takers' valuations such that the expected utility function of the learner is maximized over an entire segment of pairs of prices $(b, 1)$, for b that belongs to some interval—note that these are instances where it is never optimal to sell. Then, we partition this set of maximizers into K segments with the same size and build K perturbations of the base distribution such that, in each one of these perturbations, the corresponding expected utilities have a small $\varepsilon = \Theta(1/K)$-spike inside one of these subsegments.

We then make a simple observation: If we content ourselves with proving a lower bound for algorithms that play exclusively bid/ask pairs of the form $(b, 1)$, our problem reduces to repeated first-price auctions with unknown valuations. In this simplified problem, the learner has to essentially locate an ε-spike among one of $1/\varepsilon$ locations. Therefore, the learner can either refuse to
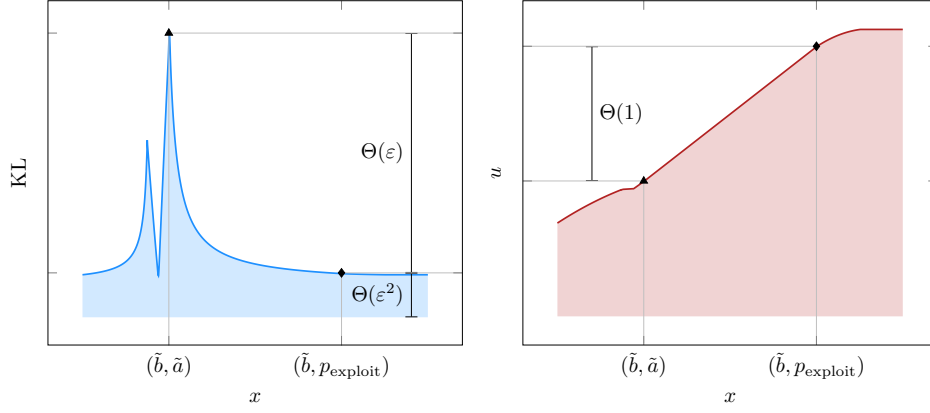
Figure 3: The amount of information (on the left) and the corresponding expected utility (on the right) as a function of pairs of prices (b, a), with (b, a) following the red dotted line in Figure 2 starting at the left boundary of the horizontal non-blue region at the point $(0, \widetilde{b})$, moving horizontally up to the diagonal, then vertically up to the optimal pair $(\widetilde{b}, 1)$, parameterized by a real number x. The points $(\widetilde{b}, \widetilde{a})$ and $(\widetilde{b}, p_{\text{exploit}})$ illustrate how regions of high reward and regions of high information do not coincide, in fact the cost of gathering more information (e.g., play $(\widetilde{b}, \widetilde{a})$ instead of $(\widetilde{b}, p_{\text{exploit}})$) can be a constant $\Theta(1)$ in the reward. On the other hand, a high-reward play gathers only $\Theta(\varepsilon^2)$ amount of information, hence costing $\Theta(\varepsilon)$ with respect to the maximum amount of information that can be gathered.

locate the spike, and pay an overall $\Theta(\varepsilon T)$ regret, or invest $\Omega(1/\varepsilon^2)$ rounds[6] in trying to locate each one of the $1/\varepsilon$ potential spikes, paying $\Omega(\varepsilon)$ each time a point in the wrong region is sampled, for an overall regret of $\Omega\left(\frac{1}{\varepsilon^2} \cdot \frac{1}{\varepsilon} \cdot \varepsilon\right) = \Omega(1/\varepsilon^2)$. The hardest instance is when $\varepsilon T = \Theta(1/\varepsilon^2)$, i.e., when $\varepsilon = \Theta(T^{1/3})$. In this case, no matter what the learner decides, they will always pay at least $\Omega(\varepsilon T) = \Omega(1/\varepsilon^2) = \Omega(T^{2/3})$ regret.

However, the problem is substantially more delicate in the general case where, to explore a potentially optimal action $(b, 1)$, the learner has access to uncountably many options (b', a') (see Figure 2). This also leads to another complication. In standard online learning lower bounds, the worst-case regret of an algorithm is analyzed by counting how many times the algorithm plays exploring vs. exploiting arms, then quantifying how much information was gathered from the exploration, and finally summing over all exploring arms. This strategy is hardly implementable in our setting because each exploiting pair $(b, 1)$ can be explored with uncountably many other arms (b', a'), where each arm trades off reward with feedback in a different way (see Figure 3)—for example, one could post two prices $(b, b+\theta)$, for some small $\theta > 0$ paying $\Omega(1)$ regret to obtain much higher quality feedback than if they played $(b, 1)$ (ε vs. ε^2 bits of information).

We circumvent this problem by clustering the action set $\mathcal{U}$ into a growing (with the time horizon T) number of disjoint subsets of non-uniform sizes that we further group in three macro-categories: "suffer high regret but (potentially) obtain a large amount of information", "suffer little (or potentially no) regret but also obtain little information", and clearly suboptimal actions where we "suffer

---

6. Given that the KL divergence between the base distribution of the feedback of points in a perturbed region and the perturbed distribution of the feedback of the same points is $O(\varepsilon^2)$, an information-theoretic argument shows that a sample of size $\Omega(1/\varepsilon^2)$ in any region is needed to determine whether the spike occurs in that region.

some regret and obtain no information". We then analyze the regret of any algorithm in terms of how many times it selects an action from each cluster, using similar techniques for all clusters belonging to the same macro-category. By doing so, we can prove that no matter how an algorithm decides to play, there will always be instances where it has to pay a regret of order at least $\Omega(T^{2/3})$.
∎

We conclude this section by remarking that, to the best of our knowledge, this is the first lower bound for a problem exhibiting such a continuum of trade-offs between reward and feedback. As our setting is arguably simple, we expect this phenomenon to occur in other applications of online learning to digital markets. Therefore, we believe the lower bound techniques developed in this work could be of independent interest.

5. Unlearnability without (lip) or (iv) Assumptions

In this section, we show that, even under the (iid) assumption, without smoothness (lip) or independence between market prices and takers' valuations (iv), learning is impossible in general.

Theorem 3 *In the online market-making problem, for each time horizon T and each $\delta \in (0, 1/2)$ and each (possibly randomized) algorithm for online market making, there exists a $[0,1]^2$-valued i.i.d. sequence $(M_t, V_t)_{t \in [T]}$ of market prices and takers' valuations such that the regret of the algorithm over the sequence $(M_t, V_t)_{t \in [T]}$ is lower bounded by $R_T \geq \left(\frac{1}{2} - \delta\right) \cdot T$.*

We present here a proof sketch and defer the full proof of this result to Appendix E.

Proof sketch   In an i.i.d. setting, without any loss of generality, we can assume that the learning algorithm is deterministic. Notice that, if market prices are always equal to 0 or 1, a deterministic algorithm can only generate finitely many points over a finite time horizon. Therefore, for any fixed deterministic algorithm and any small $0 < \varepsilon \approx 0$, there exists a small interval $[c, d]$ included in $\left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$ where the algorithm never plays over a finite time horizon. By building an i.i.d. family of market prices and takers' valuations such that $(M_t, V_t) = (0, d)$ or $(M_t, V_t) = (1, c)$ with probability $1/2$ each, one can prove that (i) the best fixed bid/ask pair belongs to the open interval (c, d), (ii) that this pair always wins (roughly) at least $1/2$ reward in expectation, while (iii) the learning algorithm (roughly) gains 0 in expectation. ∎

6. Conclusions and Future Directions

We initiated an investigation of market making under an online learning framework and provided tight bounds on the regret under various natural assumptions. While the regret of various problems related to financial trading was previously investigated, most of these results are from the viewpoint of a liquidity seeker. Liquidity providers are crucial for the functioning of markets, giving practical impact to our theoretical findings.

Extending our analysis to full feedback, in which the learner observes the private valuations at the end of each round, would shed light on the cost of the more realistic feedback studied in this paper. In a different direction, we could consider market makers that could clear their position at a later time, which would make our setting more closely related to *bandits-with-knapsacks* (Badanidiyuru et al., 2018), which is known to be unsolvable in the adversarial setting (Immorlica et al., 2023).

Acknowledgments

References

Jacob Abernethy and Satyen Kale. Adaptive market making via online learning. *Advances in Neural Information Processing Systems*, 26, 2013.

Jacob Abernethy, Rafael M Frongillo, and Andre Wibisono. Minimax option pricing meets Black-Scholes in the limit. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, pages 1029–1040, 2012.

Jacob Abernethy, Peter L Bartlett, Rafael Frongillo, and Andre Wibisono. How to hedge an option against an adversary: Black-Scholes pricing is minimax optimal. *Advances in Neural Information Processing Systems*, 26, 2013.

Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Fast rate learning in stochastic first price bidding. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of the 13th asian conference on machine learning*, volume 157 of *Proceedings of Machine Learning Research*, 2021.

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3: 5–40, 2001.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 2002a. doi: 10.1023/A:1013689704352.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

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

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic Pricing with Limited Supply. *ACM Trans. Econ. Comput.*, 3, 2015. ISSN 2167-8375. doi: 10.1145/2559152.

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

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A tight regret analysis of non-parametric repeated contextual brokerage. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2025a.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A parametric contextual online learning theory of brokerage. In *Forty-Second International Conference on Machine Learning*, 2025b.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with Knapsacks. *Journal of the ACM*, 2018. doi: 10.1145/3164539.

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

Dimitris Bertsimas and Andrew W Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 1716–1786. PMLR, 2022.

Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 309–313, 1997.

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

Allan Borodin, Ran El-Yaniv, and Vincent Gogan. On the competitive theory and practice of portfolio selection. In *LATIN 2000: Theoretical Informatics: 4th Latin American Symposium, Punta del Este, Uruguay, April 10-14, 2000 Proceedings 4*, pages 173–196. Springer, 2000.

Allan Borodin, Ran El-Yaniv, and Vincent Gogan. Can we learn to beat the best stock. *Advances in Neural Information Processing Systems*, 16, 2003.

Álvaro Cartea, Sebastian Jaimungal, and José Penalva. *Algorithmic and high-frequency trading*. Cambridge University Press, 2015.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, page 1190–1204, USA, 2013. Society for Industrial and Applied Mathematics. ISBN 9781611972511.

Nicolo Cesa-Bianchi, Tommaso Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In *Algorithmic Learning Theory*, pages 247–273. PMLR, 2019.

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 289–309, 2021.

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Repeated bilateral trade against a smoothed adversary. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1095–1130. PMLR, 2023.

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

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

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 225–236, 2024c.

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

Nicolò Cesa-Bianchi, Roberto Colomboni, and Maximilian Kasy. Adaptive maximization of social welfare. *Econometrica*, 93(3):1073–1104, 2025. doi: https://doi.org/10.3982/ECTA22351.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1991.

Thomas M Cover and Erik Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996.

Sanmay Das. A learning market-maker in the glosten–milgrom model. *Quantitative Finance*, 5: 169–180, 04 2005. doi: 10.1080/14697680500148067.

Peter DeMarzo, Ilan Kremer, and Yishay Mansour. Online trading algorithms and robust option pricing. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of computing*, pages 477–486, 2006.

Naveen Durvasula, Nika Haghtalab, and Manolis Zampetakis. Smoothed analysis of online non-parametric auctions. In *EC*, pages 540–560. ACM, 2023.

Ingo Fender and Ulf Lewrick. Shifting tides — market liquidity and market-making in fixed income instruments. *BIS Quarterly Review, March*, 2015.

Arthur Flajolet and Patrick Jaillet. Real-time bidding with side information. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in neural information processing systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/0bed45bd5774ffddc95ffe500024f628-Paper.pdf.

J Gatheral. *The volatility surface: A Practitioner's Guide*. John Wiley and Sons, Inc, 2011.

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

Lawrence R. Glosten and Paul R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14, 1985. ISSN 0304405X. doi: 10.1016/0304-405X(85)90044-3.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *NeurIPS*, 2020.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *FOCS*, pages 942–953. IEEE, 2021.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for smoothed adversaries. In *NeurIPS*, 2022.

Nils H Hakansson. Optimal investment and consumption strategies under risk for a class of utility functions. In *Stochastic Optimization Models in Finance*, pages 525–545. Elsevier, 1975.

Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. Optimal no-regret learning in repeated first-price auctions, 2024.

Larry Harris. *Trading and exchanges: Market microstructure for practitioners*. OUP USA, 2003.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

David P Helmbold, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. On-line portfolio selection using multiplicative updates. *Mathematical Finance*, 8(4):325–347, 1998.

J.C. Hull. *Options, Futures, and Other Derivatives*. Pearson Education, 2017. ISBN 9780134631493.

Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial Bandits with Knapsacks, 2023.

Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.

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

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

Hizuru Konishi. Optimal slice of a vwap trade. *Journal of Financial Markets*, 5(2):197–221, 2002.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828.

Renato Paes Leme, Balasubramanian Sivan, Yifeng Teng, and Pratik Worah. Learning to price against a moving target. In *International Conference on Machine Learning*, pages 6223–6232. PMLR, 2021.

Yiyun Luo, Will Wei Sun, and Yufeng Liu. Distribution-free contextual dynamic pricing. *Mathematics of Operations Research*, 49(1):599–618, 2024.

L.C. MacLean, E.O. Thorp, and W.T. Ziemba. *The Kelly Capital Growth Investment Criterion: Theory and Practice*. Handbook in Financial Economics. World Scientific, 2012. ISBN 9789814293495.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082, 15406261.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Stanislav Shalunov, Alexei Kitaev, Yakov Shalunov, and Arseniy Akopyan. Calculated boldness. *arXiv preprint arXiv:2012.13830*, 2020.

Steven Shreve. *Stochastic calculus for finance II: Continuous-time models*, volume 11. Springer, 2004.

Steven Shreve. *Stochastic calculus for finance I: the binomial asset pricing model*. Springer Science & Business Media, 2005.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004. doi: 10.1145/990308.990310.

E. O. Thorp. Optimal gambling systems for favorable games. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 1969. ISSN 03731138.

Edward O Thorp. Portfolio choice and the kelly criterion. In *Stochastic optimization models in finance*, pages 599–619. Elsevier, 1975.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions, 2015.

Julian Zimmert, Naman Agarwal, and Satyen Kale. Pushing the efficiency-regret pareto frontier for online learning of portfolios and quantum states. In *Conference on Learning Theory*, pages 182–226. PMLR, 2022.

Appendix A. Notation

In this section, we collect the main pieces of notation used in this paper.

T	Time horizon
K	Step of the grid over $[0,1]$
$\mathcal{U}$	Upper triangle over $[0,1]^2$

Market making	
M_t	Market price
V_t	Taker's valuation
B_t	Bid presented by the learner
A_t	Ask presented by the learner

First-price auction with unknown valuations	
Z_t	Unknown valuation of the auctioned item
H_t	Highest competing bid in the auction
X_t	Bid presented by the learner

Dynamic pricing with unknown costs	
C_t	Unknown cost of the item for sale
W_t	Buyer's valuation of the item
P_t	Price presented by the learner

Table 2: Notation

Appendix B. Additional Related Works

There exists a vast literature that treats various trading-related tasks as specific stochastic control problems and solves them using techniques from stochastic control theory (Shreve, 2004, 2005; Cartea et al., 2015; Gatheral, 2011). However, most of that work assumes that the parameters of the underlying stochastic process are known or have been fitted to historical data in a previous "calibration" step. Fitting a distribution to data is challenging, especially in a market with hundreds of thousands of assets. Moreover, the distribution-fitting process is typically unaware of the downstream optimization problem the distribution is going to be fed into, and thus is unlikely to minimize generalization error on the downstream task. The field of online learning (Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012; Hazan, 2016; Orabona, 2019) provides adversarial or distribution-free approaches for solving exactly the kinds of sequential decision making problems that are common in financial trading, and even though some recent research does apply these approaches to trading, they are still far from being widely adopted. An early contribution to this area is Cover's model of "universal portfolios" (Cover, 1991), where a problem of portfolio construction is solved in the case where asset returns are generated by an adversary. Cover showed that one can still achieve logarithmic regret with respect to the best "constantly rebalanced portfolio", that ensures the fraction of wealth allocated to each asset remains constant through time. A long line of work has built up

on his model to handle transaction costs (Blum and Kalai, 1997) and side information (Cover and Ordentlich, 1996), as well as optimizing the trade-off between regret and computation (Helmbold et al., 1998; Borodin et al., 2000, 2003; Zimmert et al., 2022). The problem of pricing derivatives such as options has also been tackled using online learning, where the prices are generated by an adversary as opposed to a geometric Brownian motion (DeMarzo et al., 2006; Abernethy et al., 2012, 2013). However, most of the problems considered in the online learning framework are from the point of view of a liquidity seeker as opposed to a liquidity provider, such as a market maker.

First-price auctions with unknown evaluations. When restricted solely to the option of buying stock, our problem can be viewed as a series of repeated first-price auctions with unknown valuations. This scenario has received prior attention within the context of regret minimization (Cesa-Bianchi et al., 2024c; Achddou et al., 2021). The present work leverages these existing results but applies them within a more intricate setting. Cesa-Bianchi et al. (2024c) explored this problem with varying degrees of transparency, defined as the information revealed by the auctioneer, and provide a comprehensive characterization. In our framework, the level of transparency is higher than that encountered in the *bandit* case, orthogonal to the *semi-transparent* case, and less informative that the *transparent* case; this positioning offers a novel perspective on the problem. Achddou et al. (2021) investigated the repeated first-price auction problem within a fixed stochastic environment with independence assumptions and provide instance-dependent bounds, we tackle a similar problem in Lemma 4 and present distribution-free guarantees. Other types of feedback have been considered, e.g., the case where the maximum bids $(V_t)_{t \in [T]}$ form an i.i.d. process and are observed only when the auction is lost (Han et al., 2024), for the case where the private evaluations $(M_t)_{t \in [T]}$ are i.i.d. it is possible to achieve regret $\widetilde{O}(\sqrt{T})$.

Dynamic pricing with unknown costs. When considering solely the option of selling stock, our problem aligns with the well-established field of dynamic pricing with unknown costs. This area boasts a rich body of research, prior research analyzed the setting in which the learner has I items to sell to N independent buyers and, with regularity assumptions of the underlying distributions, can achieve a $O((I \log T)^{2/3})$ regret bound against an offline benchmark with knowledge of the buyer's distribution (Babaioff et al., 2015). It is known that in the stochastic setting and, under light assumptions on the reward function, achieve regret $O(\sqrt{T \log T})$ (Kleinberg and Leighton, 2003); this result has been expanded upon by considering discrete price distributions supported on a set of prices of unknown size K and it has been shown to be possible to achieve regret of order $O(\sqrt{KT})$ (Cesa-Bianchi et al., 2019). Our work on dynamic pricing focuses on a scenario where the learner maintains an unlimited inventory and must incur an unknown cost per trade before realizing any profit.

Traditional finance. Traditionally, the finance literature first fits the parameters of a stochastic process to the market and then optimizes trading based on those parameters. For example, the Nobel prize-winning Black-Scholes-Merton formula (Black and Scholes, 1973) assumes that the price of a stock follows a geometric Brownian motion with known volatility and prices an option on the stock by solving a Hamilton-Jacobi-Bellman equation to compute a dynamic trading strategy whose value at any time matches the value of the option. Many similar formulae have since been derived for more exotic derivatives and for more complex underlying stochastic processes—see (Hull, 2017) for an overview. A stochastic control approach has also been taken to solve the problem of "optimal trade execution," which involves trading a large quantity of an asset over a specified period of time while

---

Algorithm 3: `Discretized Bandits for First Price Auctions`

---

input: Number of arms $K \in \{2, 3, \dots\}$, Algorithm $\mathcal{A}$ for K-armed bandits
initialization: Let $q_k := \frac{k-1}{K-1}$, for all $k \in [K]$
for *time* $t = 1, 2, \dots$ **do**
$\quad$ Let $I_t \in [K]$ be the output of $\mathcal{A}$ at time t;
$\quad$ Post bid $X_t := q_{I_t}$ and observe feedback Z_t and $\mathbb{I}\{X_t \geq H_t\}$;
$\quad$ Compute utility $(Z_t - X_t) \cdot \mathbb{I}\{X_t \geq H_t\}$ and feed it back to $\mathcal{A}$;
end

---

minimizing market impact (Almgren and Chriss, 2001; Bertsimas and Lo, 1998), to compute a sequence of trades such that the total cost matches a pre-specified benchmark (Konishi, 2002), and also for the problem of market making (Cartea et al., 2015). Once again, the underlying stochastic process is assumed to be known. The two-step approach of first fitting a distribution and then optimizing a function over the fitted distribution is very popular even for simpler questions. For example, the problem of "portfolio optimization/construction" deals with computing the optimal allocation of your wealth into various assets in order to maximize some notion of future utility. The celebrated Kelly criterion (Shalunov et al., 2020; Thorp, 1975, 1969; Hakansson, 1975; MacLean et al., 2012) recommends one to use the allocation that maximizes the expected log return, and the mean-variance theory (Markowitz, 1952) (another Nobel prize-winning work) says one should maximize a sum of expected return and the variance of returns (scaled by some measure of your risk tolerance). However both assume that one knows the underlying distribution over returns.

Appendix C. Missing details in the Proof of Theorem 1

To guide the reader through the proof of Theorem 1, we split it in several steps.

First, consider the problem of repeated first-price auctions with unknown valuations. Since the feedback received at the end of each round allows to compute the utility of the learner, a natural strategy is to simply discretize the interval $[0, 1]$ and run a bandit algorithm on the discretization. The pseudo-code of this simple meta-algorithm can be found in Algorithm 3. The following result shows its guarantees.

Lemma 4 *In the repeated first-price auctions with unknown valuations problem, let $T \in \mathbb{N}$ be the time horizon and let $(Z_t, H_t)_{t \in [T]}$ be the $[0, 1]^2$-valued stochastic process representing the sequence of valuations and highest competing bids. Assume that one of the two following conditions is satisfied:*

1. For each $t \in [T]$, the cumulative distribution function of H_t is L-Lipschitz, for some $L > 0$.

2. The process $(Z_t, H_t)_{t \in [T]}$ is i.i.d. and, for each $t \in [T]$, the two random variables Z_t and H_t are independent of each other.

Then, for any $K \geq 2$ and any K-armed bandit algorithm $\mathcal{A}$, letting R_T^K be the regret of $\mathcal{A}$ when the reward at any time $t \in [T]$ of any arm $k \in [K]$ is $(Z_t - q_k)\mathbb{I}\{q_k \geq H_t\}$, the regret of Algorithm 3 run with parameters K and $\mathcal{A}$ satisfies $R_T \leq R_T^K + \frac{\widetilde{L}+1}{2(K-1)}T$, with $\widetilde{L} = L$ (resp., $\widetilde{L} = 1$) if Item 1 (resp., Item 2) holds. In particular, if $T \geq 2$, by choosing $K := \lceil T^{1/3} \rceil + 1$ and, as the

underlying learning procedure $\mathcal{A}$, an adapted version of Poly INF (Audibert and Bubeck, 2010), the regret of Algorithm 3 run with parameters K and $\mathcal{A}$ satisfies $R_T \leq c \cdot T^{2/3}$, where $c \leq L + 50$ (resp. $c \leq 51$) if Item 1 (resp., Item 2) holds.

Proof The regret R_T^K of $\mathcal{A}$, when the reward at any time $t \in [T]$ of an arm $k \in [K]$ is $(Z_t - q_k)\mathbb{I}\{q_k \geq H_t\}$, is defined by

$$R_T^K = \max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_k) \cdot \mathbb{I}\{q_k \geq H_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_{I_t}) \cdot \mathbb{I}\{q_{I_t} \geq H_t\}\right].$$

The regret R_T of Algorithm 3 when the unknown valuations are $Z_1, \ldots, Z_T$ and the highest competing bids are $H_1, \ldots, H_T$ is

$$R_T = \sup_{x \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - x) \cdot \mathbb{I}\{x \geq H_t\}\right] - \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_{I_t}) \cdot \mathbb{I}\{q_{I_t} \geq H_t\}\right].$$

Hence $R_T = R_T^K + \delta_K$, where δ_K is the discretization error

$$\delta_K := \sup_{x \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - x) \cdot \mathbb{I}\{x \geq H_t\}\right] - \max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_k) \cdot \mathbb{I}\{q_k \geq H_t\}\right].$$

We now proceed to bound the discretization error δ_K in the two cases:

1. If the cumulative distribution function of H_t is L-Lipschitz, for some $L > 0$, then for all $t \in [T]$ the expected utility is $\phi(q) = \mathbb{E}\left[(Z_t - q) \cdot \mathbb{I}\{q \geq H_t\}\right]$ is $(L+1)$-Lipschitz; indeed, for all $q, q' \in [0,1]$, if (without loss of generality) $q > q'$, then

$$\left|\phi(q) - \phi(q')\right| = \left|\mathbb{E}\left[(Z_t - q)\mathbb{I}\{H_t \leq q\} - (Z_t - q')\mathbb{I}\{H_t \leq q'\}\right]\right|$$
$$= \left|\mathbb{E}\left[(Z_t - q)\mathbb{I}\{q' < H_t \leq q\} + (q' - q)\mathbb{I}\{H_t \leq q'\}\right]\right|$$
$$\leq \left(\mathbb{P}[H_t \leq q] - \mathbb{P}[H_t \leq q']\right) + (q - q') \leq (L+1)(q - q').$$

Define the supremum $x^\star \in [0,1]$ in the definition of δ_K, which exists because ϕ is continuous and $[0,1]$ is compact. Let $k^\star = \arg\min_{k \in [K]} |q_k - x^\star|$ be the point in the grid closest to $x^\star$. Since

$$\max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_k) \cdot \mathbb{I}\{q_k \geq H_t\}\right] \geq \mathbb{E}\left[\sum_{t=1}^{T}(Z_t - q_{k^\star}) \cdot \mathbb{I}\{q_{k^\star} \geq H_t\}\right],$$

the function ϕ is $(L+1)$-Lipschitz, and $|x^\star - q_{k^\star}| \leq 1/2(K-1)$ we obtain

$$\delta_K \leq \sum_{t=1}^{T}\left(\mathbb{E}\left[(Z_t - x^\star) \cdot \mathbb{I}\{x^\star \geq H_t\}\right] - \mathbb{E}\left[(Z_t - q_{k^\star}) \cdot \mathbb{I}\{q_{k^\star} \geq H_t\}\right]\right) \leq \frac{L+1}{2(K-1)}T .$$

2. If the process $(Z_t, H_t)_{t \in [T]}$ is i.i.d. and for each $t \in [T]$ the two random variables Z_t and H_t are independent of each other, then for all $t \in [T]$ the expected utility is $\psi(x) := (\mu - x)F(x)$ where $\mu := \mathbb{E}[Z_t]$ and $F(x) := \mathbb{P}[H_t \leq x]$. Fix $\eta > 0$. If $\sup_{x \in [0,1]} \psi(x) \leq \eta$, set $k^\star = K$ and note that for each $x \in [0, 1]$ we have that $\psi(x) \leq \eta \leq \psi(1) + \eta = \psi(q_{k^\star}) + \eta$, which means $\delta_K \leq \eta T$. Otherwise, let $x_\eta^\star$ be such that $\sup_{x \in [0,1]} \psi(x) - \psi(x_\eta^\star) \leq \eta$ and notice that we can assume that $x_\eta^\star \leq \mu$ because otherwise $\psi(x_\eta^\star) \leq 0$, hence we would have been in the first case when $\sup_{x \in [0,1]} \psi(x) \leq \eta$. The expected reward achieved by playing bid $x_\eta^\star$ can be controlled with any point q_k on the grid such that $q_k \geq x_\eta^\star$ as

$$\psi(x_\eta^\star) - \psi(q_k) = (\mu - x_\eta^\star)F(x_\eta^\star) - (\mu - q_k)F(q_k) \leq (\mu - x_\eta^\star)F(q_k) - (\mu - q_k)F(q_k) \leq q_k - x_\eta^\star .$$

Call $k^\star \in [K]$ the index of the arm closest to $x_\eta^\star$ such that $q_{k^\star} \geq x_\eta^\star$, note that $q_{k^\star} - x_\eta^\star \leq 1/(K-1)$. Thus

$$\delta_K \leq \eta T + \sum_{t=1}^{T} \left(\psi(x_\eta^\star) - \psi(q_{k^\star}) \right) \leq \eta T + \sum_{t=1}^{T} |x_\eta^\star - q_{k^\star}| \leq \eta T + \frac{T}{K-1} .$$

Given that η was chosen arbitrarily, taking the limit as $\eta \to 0$, we have that

$$\delta_K \leq \frac{T}{K-1} .$$

Define $\widetilde{L} := L$ in the first case and $\widetilde{L} := 1$ in the second case. Next, pick the Poly INF algorithm (Audibert and Bubeck, 2010) as the underlying learning procedure $\mathcal{A}$ and apply the appropriate rescaling of the utilities $x \mapsto \frac{x+1}{2}$, which is necessary because the utility yields values in $[-1, 1]$, while Poly INF was designed for rewards in $[0, 1]$, this costs a multiplicative factor of 2 in the regret guarantee. The regret R_T of Algorithm 3 can be upper bounded by

$$R_T \leq R_T^K + \frac{\widetilde{L}}{K-1} \cdot T \leq 50 \cdot T^{2/3} + \frac{\widetilde{L}}{K-1} \cdot T \leq (50 + \widetilde{L}) \cdot T^{2/3}$$

whenever $K = \lceil T^{1/3} \rceil + 1$ and the second inequality comes from the regret guarantees of the rescaled version of Poly INF (Audibert and Bubeck, 2010, Theorem 11). ∎

A completely analogous result can be proved for Algorithm 4 for the problem of dynamic pricing with unknown costs.

Lemma 5 *In the repeated dynamic pricing with unknown costs problem, let $T \in \mathbb{N}$ be the time horizon and let $(C_t, W_t)_{t \in [T]}$ be the $[0,1]^2$-valued stochastic process representing the sequence of costs and buyer's valuations. Assume that one of the two following conditions is satisfied:*

1. *For each $t \in [T]$, the cumulative distribution function of W_t is L-Lipschitz, for some $L > 0$.*

2. *The process $(C_t, W_t)_{t \in [T]}$ is i.i.d. and, for each $t \in [T]$, the two random variables C_t and W_t are independent of each other.*

Algorithm 4: `Discretized Bandits for Dynamic Pricing`

---

input: Number of arms $K \in \{2, 3, \dots\}$, Algorithm $\mathcal{A}$ for K-armed bandits
initialization: Let $q_k := \frac{k-1}{K-1}$, for all $k \in [K]$
for *time* $t = 1, 2, \dots$ **do**
    Let $I_t \in [K]$ be the output of $\mathcal{A}$ at time t;
    Post price $P_t := q_{I_t}$ and observe feedback C_t and $\mathbb{I}\{P_t < W_t\}$;
    Compute utility $(P_t - C_t) \cdot \mathbb{I}\{P_t < W_t\}$ and feed it back to $\mathcal{A}$;
end

Then, for any $K \geq 2$ and any K-armed bandit algorithm $\mathcal{A}$, letting R_T^K be the regret of $\mathcal{A}$ when the reward at any time $t \in [T]$ of any arm $k \in [K]$ is $(q_k - C_t)\mathbb{I}\{q_k < W_t\}$, the regret of Algorithm 4 run with parameters K and $\mathcal{A}$ satisfies

$$R_T \leq R_T^K + \frac{\widetilde{L} + 1}{2(K-1)} T \,,$$

with $\widetilde{L} = L$ (resp., $\widetilde{L} = 1$) if Item 1 (resp., Item 2) holds. In particular, if $T \geq 2$, by choosing $K := \lceil T^{1/3} \rceil + 1$ and, as the underlying learning procedure $\mathcal{A}$, an adapted version of Poly INF (Audibert and Bubeck, 2010), the regret of Algorithm 4 run with parameters K and $\mathcal{A}$ satisfies

$$R_T \leq c \cdot T^{2/3} \,,$$

where $c \leq L + 50$ (resp. $c \leq 51$) if Item 1 (resp., Item 2) holds.

Proof As in Lemma 4, we can bound R_T for the two cases by controlling the discretization error δ_K on the grid via Lipschitzness (Item 1) or leveraging the structure of the expected utility around the maximum (Item 2). The bound on the regret $\mathcal{R}_T$ of $\mathcal{A}$ follows from the same guarantees on Poly INF (Audibert and Bubeck, 2010). ∎

Next, we use Lemma 4 and Lemma 5 to show that the regret of an appropriately tuned `M3` is bounded by $\mathcal{O}(T^{2/3})$. This implies is the second part of Theorem 1, fully concluding its proof.

Lemma 6 *In the online market-making problem, let $T \in \mathbb{N}$ be the time horizon and let $(M_t, V_t)_{t \in [T]}$ be the $[0, 1]^2$-valued stochastic process representing the sequence of market prices and takers' valuations. Assume that one of the two following conditions is satisfied:*

1. *For each $t \in [T]$, the cumulative distribution function of V_t is L-Lipschitz, for some $L > 0$.*

2. *The process $(M_t, V_t)_{t \in [T]}$ is i.i.d. and, for each $t \in [T]$, the two random variables M_t and V_t are independent of each other.*

Then, let $K := \lceil T^{1/3} \rceil + 1$, let $\mathcal{A}^{\mathrm{fpa}}$ be the instance of Algorithm 3 that uses Poly INF as described in Lemma 4, and $\mathcal{A}^{\mathrm{dp}}$ be the instance of Algorithm 4 that uses Poly INF as described in Lemma 5. Then the regret of `M3` (Algorithm 2) run with parameters $\mathcal{A}^{\mathrm{fpa}}$ and $\mathcal{A}^{\mathrm{dp}}$ satisfies $R_T \leq cT^{2/3}$ with $c \leq 2L + 100$ (resp., $c \leq 102$) if Item 1 (resp., Item 2) holds.

Proof By Algorithm 2, the regret of `M3` is upper bounded by the regret of a repeated first-price auctions problem with unknown valuations plus a dynamic pricing problem with unknown costs. Plugging in the bounds from Lemma 4 and Lemma 5, we get the required result. ∎

Appendix D. Proof of Theorem 2

The following is the proof of our $T^{2/3}$ lower bound (Theorem 2).

Proof Fix $T \geq 42$. We define the following constants that will be used in the proof.

$$K := \lceil T^{1/3} \rceil \qquad \varepsilon_K := 1/16K \qquad \forall k \in [K], \ r_K^k := 3/16 + (k - 1/2)\,\varepsilon_K$$

$$p_{\textbf{left}} := 3/16 \qquad p_{\textbf{right}} := 1/4 \qquad p_{\textbf{exploit}} := 3/4$$

$$c_{\textbf{plat}} := 1/32 \qquad c_{\textbf{spike}} := 1/72$$

Define the density

$$f\colon [0,1] \to [0,\infty)\,, \qquad x \mapsto \frac{8}{9}\cdot \mathbb{I}_{[0,\frac{3}{16}]}(x) + \frac{1}{8}\frac{1}{\left(\frac{15}{16}-x\right)^2}\cdot \mathbb{I}_{\left(\frac{3}{16},\frac{3}{4}\right]}(x) + \frac{8}{3}\cdot \mathbb{I}_{\left(\frac{3}{4},\frac{7}{8}\right]}(x)\,,$$

so that the corresponding cumulative distribution function F satisfies, for each $x \in [0,1]$,

$$F(x) = \frac{8}{9}x \cdot \mathbb{I}_{[0,\frac{3}{16}]}(x) + \frac{1}{8}\frac{1}{\frac{15}{16}-x}\cdot \mathbb{I}_{\left(\frac{3}{16},\frac{3}{4}\right]}(x) + \frac{8}{3}\left(x-\frac{1}{2}\right)\cdot \mathbb{I}_{\left(\frac{3}{4},\frac{7}{8}\right]}(x) + \mathbb{I}_{\left(\frac{7}{8},1\right]}(x)\,. \qquad (2)$$

We define a family of perturbations parameterized by the set $\Xi := \{(r,\varepsilon) \in [\frac{3}{16},\frac{11}{16}] \times [0,1] \mid \frac{3}{16} \leq r - \frac{\varepsilon}{2} \leq r + \frac{\varepsilon}{2} \leq \frac{11}{16}\}$; for each $(r,\varepsilon) \in \Xi$, define $g_{r,\varepsilon} := \frac{1}{9}\cdot\mathbb{I}_{[r-\frac{\varepsilon}{2},r]} - \frac{1}{9}\cdot\mathbb{I}_{(r,r+\frac{\varepsilon}{2}]}$ and $f_{r,\varepsilon} := f + g_{r,\varepsilon}$ (see Figure 4). Notice that for each $(r,\varepsilon) \in \Xi$ the function $f_{r,\varepsilon}$ is still a density function whose corresponding cumulative distribution function $F_{r,\varepsilon}$ satisfies, for each $x \in [0,1]$,

$$F_{r,\varepsilon}(x) = F(x) + \frac{\varepsilon}{18}\Lambda_{r,\varepsilon}(x)\,,$$

where $\Lambda_{r,\varepsilon}$ is the tent function of height 1 and width ε centered in r, i.e., the function defined, for each $x \in \mathbb{R}$, by

$$\Lambda_{r,\varepsilon}(x) := \left(1 - \frac{2}{\varepsilon}(r-x)\right)\cdot \mathbb{I}_{[r-\frac{\varepsilon}{2},r]}(x) + \left(1 - \frac{2}{\varepsilon}(x-r)\right)\cdot \mathbb{I}_{(r,r+\frac{\varepsilon}{2}]}(x)\,.$$

Note that $F_{r,\varepsilon}$ is 4-Lipschitz; indeed, for each $(b,a) \in \mathcal{U}$,

$$|F_{r,\varepsilon}(a) - F_{r,\varepsilon}(b)| = \int_b^a f_{r,\varepsilon}(x)\ \mathrm{d}x = \max_{c\in[b,a]} f_{r,\varepsilon}(c)(a-b)$$

$$\leq \left(\max_{c'\in[0,1]} f(c') + \frac{1}{9}\right)(a-b) = 4(a-b)\,,$$

where f is maximized in $3/4$. Consider an independent family $\{M_t, V_t, V_{r,\varepsilon,t}\}_{t\in\mathbb{N},(r,\varepsilon)\in\Xi}$ such that for each $t \in \mathbb{N}$ the distribution μ of M_t is a uniform on $\left[\frac{7}{8},1\right]$ (therefore, M_t admits an 8-Lipschitz cumulative distribution function), for each $t \in \mathbb{N}$ the distribution ν of V_t has f as density, while for each $(r,\varepsilon) \in \Xi$ and each $t \in \mathbb{N}$ the distribution $\nu_{r,\varepsilon}$ of $V_{r,\varepsilon,t}$ has $f_{r,\varepsilon}$ as density. Notice that for each $k \in [K]$ we have that $(r_K^k, \varepsilon_K) \in \Xi$. Now, partition $\mathcal{U}$ in the following regions (see Figure 5 for a
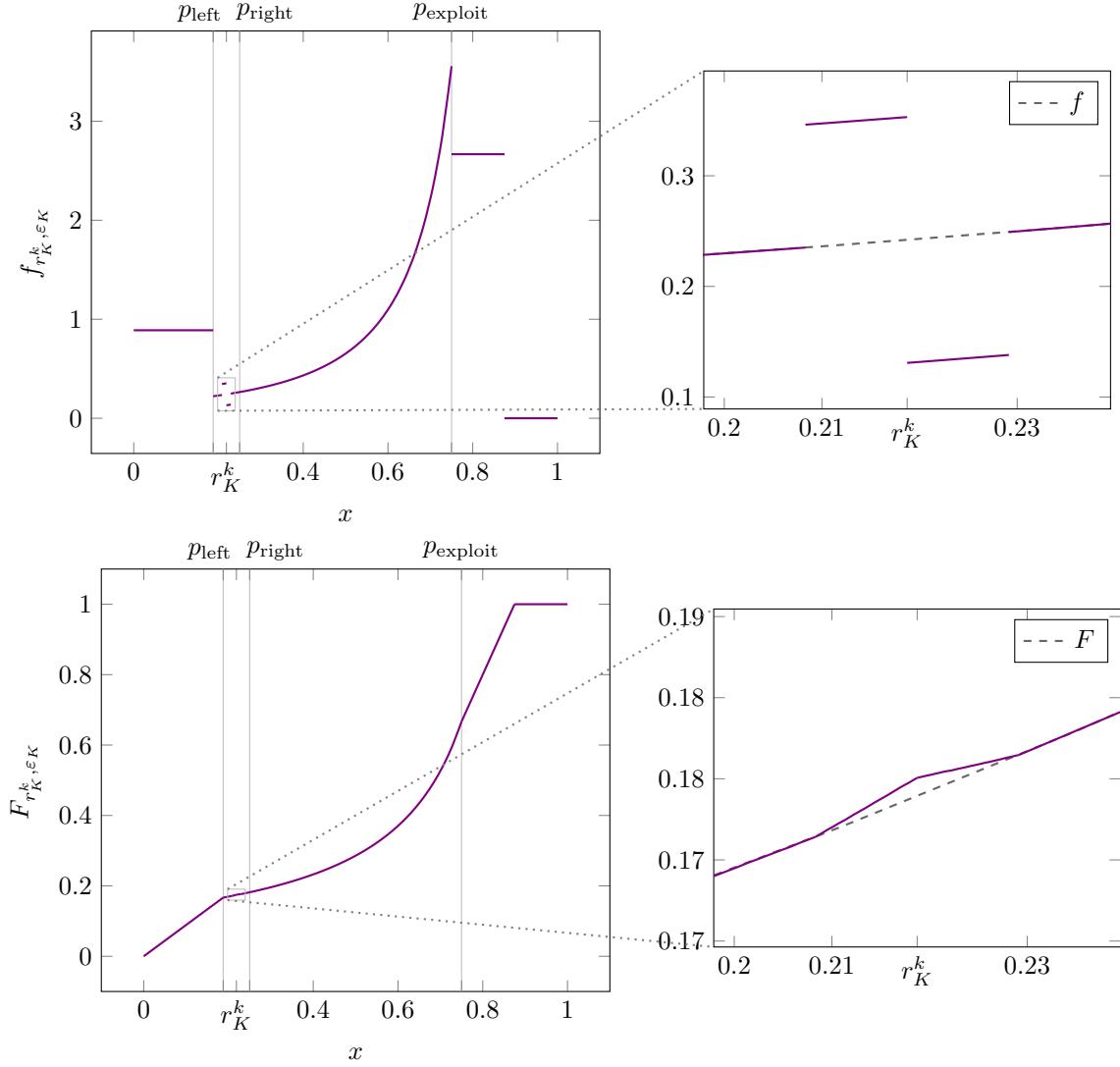
Figure 4: Above, the perturbed density $f_{r_K^k, \varepsilon_K}(x)$ and below, the perturbed *cumulative* density $F_{r_K^k, \varepsilon_K}(x)$ with $K = 3$, $k = 2$, the highlighted portions show the effects of the perturbation $\Lambda_{r_K^k, \varepsilon_K}$ on the density function $f(x)$ and on the cumulative distribution function $F(x)$ respectively. The dotted lines represent the base functions.

not-to-scale illustration).

$$R_1^{\textbf{left}} := \left\{ (b, a) \in \mathcal{U} \mid (b \le p_{\textbf{left}}) \wedge \left(p_{\textbf{right}} - \varepsilon_K \le a \le p_{\textbf{right}} \right) \right\}$$

$$R_2^{\textbf{left}} := \left\{ (b, a) \in \mathcal{U} \mid (b \le p_{\textbf{left}}) \wedge \left(p_{\textbf{right}} - 2\varepsilon_K \le a < p_{\textbf{right}} - \varepsilon_K \right) \right\}$$

$$\vdots$$

$$R_{K-1}^{\textbf{left}} := \left\{ (b, a) \in \mathcal{U} \mid (b \le p_{\textbf{left}}) \wedge \left(p_{\textbf{left}} + \varepsilon_K \le a < p_{\textbf{left}} + 2\varepsilon_K \right) \right\}$$

$$R_K^{\textbf{left}} := \left\{ (b, a) \in \mathcal{U} \mid (b \le p_{\textbf{left}}) \wedge \left(p_{\textbf{left}} \le a < p_{\textbf{left}} + \varepsilon_K \right) \right\}$$

and

$$R_1^{\textbf{top}} := \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} \le b \le p_{\textbf{left}} + \varepsilon_K) \wedge \left(p_{\textbf{right}} \le a \le p_{\textbf{exploit}} \right) \right\}$$
$$R_2^{\textbf{top}} := \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} + \varepsilon_K < b \le p_{\textbf{left}} + 2\varepsilon_K) \wedge \left(p_{\textbf{right}} \le a \le p_{\textbf{exploit}} \right) \right\}$$
$$\vdots$$
$$R_{K-1}^{\textbf{top}} := \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{right}} - 2\varepsilon_K < b \le p_{\textbf{right}} - \varepsilon_K) \wedge \left(p_{\textbf{right}} \le a \le p_{\textbf{exploit}} \right) \right\}$$
$$R_K^{\textbf{top}} := \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{right}} - \varepsilon_K < b \le p_{\textbf{right}}) \wedge \left(p_{\textbf{right}} \le a \le p_{\textbf{exploit}} \right) \right\}$$

and, for each $i, j \in [K]$ such that $i + j \le K$

$$R_{i,j}^{\textbf{square}} := \{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} + (i-1)\varepsilon_K < b \le p_{\textbf{left}} + i\varepsilon_K) \wedge$$
$$(p_{\textbf{right}} - j\varepsilon_K < a \le p_{\textbf{right}} - (j-1)\varepsilon_K) \}$$

and, for each $k \in [K]$

$$R_k^{\textbf{triangle}} := \{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} + (k-1)\varepsilon_K < b \le p_{\textbf{left}} + k\varepsilon_K) \wedge$$
$$(p_{\textbf{left}} + (k-1)\varepsilon_K < a \le p_{\textbf{left}} - k\varepsilon_K) \}$$

and

$$R_1^{\textbf{exploit}} := \left\{ (b,a) \in \mathcal{U} \mid \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} \le b \le p_{\textbf{left}} + \varepsilon_K) \wedge \left(p_{\textbf{exploit}} < a \right) \right\} \right\}$$
$$R_2^{\textbf{exploit}} := \left\{ (b,a) \in \mathcal{U} \mid \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{left}} + \varepsilon_K < b \le p_{\textbf{left}} + 2\varepsilon_K) \wedge \left(p_{\textbf{exploit}} < a \right) \right\} \right\}$$
$$\vdots$$
$$R_{K-1}^{\textbf{exploit}} := \left\{ (b,a) \in \mathcal{U} \mid \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{right}} - 2\varepsilon_K < b \le p_{\textbf{right}} - \varepsilon_K) \wedge \left(p_{\textbf{exploit}} < a \right) \right\} \right\}$$
$$R_K^{\textbf{exploit}} := \left\{ (b,a) \in \mathcal{U} \mid \left\{ (b,a) \in \mathcal{U} \mid (p_{\textbf{right}} - \varepsilon_K < b \le p_{\textbf{right}}) \wedge \left(p_{\textbf{exploit}} < a \right) \right\} \right\}$$

Let $R^{\textbf{white}}$ be the part of $\mathcal{U}$ not covered by the union of the previous regions and define

$$R^{\textbf{exploit}} := R_1^{\textbf{exploit}} \cup \cdots \cup R_K^{\textbf{exploit}}, \qquad R^{\textbf{explore}} := \mathcal{U} \backslash (R^{\textbf{white}} \cup R^{\textbf{exploit}}).$$

Notice that, for each $k \in [K]$, each $t \in \mathbb{N}$, and each $(b,a) \in \mathcal{U}$, given that $\mathbb{P}[V_{r_K^k, \varepsilon_K, t} \le M_t] = 1$, it holds that

$$\mathbb{E}\left[u(b, a, M_t, V_{r_K^k, \varepsilon_K, t}) \right] \le \mathbb{E}\left[u(b, 1, M_t, V_{r_K^k, \varepsilon_K, t}) \right]$$
$$= \mathbb{E}\left[(M_t - b)\mathbb{I}\{b > V_{r_K^k, \varepsilon_K, t}\} \right] = \left(\frac{15}{16} - b \right) F_{r_K^k, \varepsilon_K}(b)$$
$$= \frac{8}{9} b \left(\frac{15}{16} - b \right) \cdot \mathbb{I}_{[0, \frac{3}{16}]}(b) + \frac{1}{8} \cdot \mathbb{I}_{(\frac{3}{16}, \frac{3}{4}]}(b) + \frac{\varepsilon_K}{18} \left(\frac{15}{16} - b \right) \cdot \Lambda_{r_K^k, \varepsilon_K}(b)$$
$$+ \frac{8}{3} \left(b - \frac{1}{2} \right) \left(\frac{15}{16} - b \right) \cdot \mathbb{I}_{(\frac{3}{4}, \frac{7}{8}]}(b) + \left(\frac{15}{16} - b \right) \cdot \mathbb{I}_{(\frac{7}{8}, 1]}(b),$$

from which it follows that:

- The pair of prices with highest expected utility is $(r_K^k, 1)$ and

$$\max_{(b,a)\in\mathcal{U}} \mathbb{E}\big[u(b, a, M_t, V_{r_K^k, \varepsilon_K, t})\big] = \mathbb{E}\left[u\left(r_K^k, 1, M_t, V_{r_K^k, \varepsilon_K, t}\right)\right] \geq \frac{1}{8} + c_{\textbf{spike}} \cdot \varepsilon_K \ .$$

- The maximum expected utility when b is not in the perturbation is

$$\max_{(b,a)\in\mathcal{U}, b\notin[r_K^k - \frac{\varepsilon_K}{2}, r_K^k + \frac{\varepsilon_K}{2}]} \mathbb{E}\left[u(b, a, M_t, V_{r_K^k, \varepsilon_K, t})\right] = \frac{1}{8} \ .$$

- The expected utility in the exploration region is upper bounded by

$$\max_{(b,a)\in R^{\textbf{explore}}} \mathbb{E}\left[u(b, a, M_t, V_{r_K^k, \varepsilon_K, t})\right] \leq \mathbb{E}\left[u(r_K^k, p_{\textbf{exploit}}, M_t, V_{r_K^k, \varepsilon_K, t})\right] \leq \frac{1}{8} - c_{\textbf{plat}} \ .$$

Also note that:

- For each $x \in [0, 1]$, if $x \notin [r_K^k - \varepsilon_K/2, r_K^k + \varepsilon_K/2]$, then $\mathbb{P}[V_{r_K^k, \varepsilon_K, t} < x] = F(x)$.

Crucially, given that the setting is stochastic, without loss of generality, we can consider only deterministic algorithms. Fix a deterministic algorithm $(\mathscr{A}_t)_{t\in\mathbb{N}} := (\mathcal{B}_t, \mathcal{A}_t)_{t\in\mathbb{N}}$, i.e., a sequence of functions such that for each $t \in \mathbb{N}$ we have that $\mathscr{A}_t = (\mathcal{B}_t, \mathcal{A}_t) \colon ([0, 1] \times \{0, 1\} \times \{0, 1\})^{t-1} \to \mathcal{U}$, with the understanding that $\mathscr{A}_1 = (\mathcal{B}_1, \mathcal{A}_1) \in \mathcal{U}$. For each $k \in [K]$, let $(B_t^k, A_t^k)_{t\in[T]}$ be the prices posted by the algorithm when the underlying instance is $(M_t, V_{\varepsilon_K, r_K^k, t})_{t\in[T]}$, i.e., let $(B_1^k, A_1^k) := \mathscr{A}_1$ and for each $t \in [T]$ with $t \geq 2$ let $(B_t^k, A_t^k) := \mathscr{A}_t(M_1, \mathbb{I}\{V_{\varepsilon_K, r_K^k, 1} \leq B_1^k\}, \mathbb{I}\{V_{\varepsilon_K, r_K^k, 1} > A_1^k\}, \dots, M_{t-1}, \mathbb{I}\{V_{\varepsilon_K, r_K^k, t-1} \leq B_{t-1}^k\}, \mathbb{I}\{V_{\varepsilon_K, r_K^k, t-1} > A_{t-1}^k\})$. Analogously, let $(B_t, A_t)_{t\in[T]}$ be the prices posted by the algorithm when the underlying instance is $(M_t, V_t)_{t\in[T]}$. For each $k \in [K]$ and each $t \in [T]$, let also $W_t^k := (\mathbb{I}\{V_{\varepsilon_K, r_K^k, t} \leq B_t^k\}, \mathbb{I}\{V_{\varepsilon_K, r_K^k, t} > A_t^k\})$ and $W_t := (\mathbb{I}\{V_t \leq B_t\}, \mathbb{I}\{V_t > A_t\})$.

Define the following auxiliary random variables. For each $i, k \in [K]$, define

$$N_{i,k}^{\textbf{left}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{\mathscr{A}_s(M_1, W_1^k, \dots, M_{s-1}, W_{s-1}^k) \in R_i^{\textbf{left}}\right\}$$

and for each $i \in [K]$, define

$$N_i^{\textbf{left}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{\mathscr{A}_s(M_1, W_1, \dots, M_{s-1}, W_{s-1}) \in R_i^{\textbf{left}}\right\} \ .$$

Analogously, for each $i, k \in [K]$, define

$$N_{i,k}^{\textbf{top}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{\mathscr{A}_s(M_1, W_1^k, \dots, M_{s-1}, W_{s-1}^k) \in R_i^{\textbf{top}}\right\}$$

and for each $i \in [K]$, define

$$N_i^{\textbf{top}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{\mathscr{A}_s(M_1, W_1, \dots, M_{s-1}, W_{s-1}) \in R_i^{\textbf{top}}\right\} \ .$$

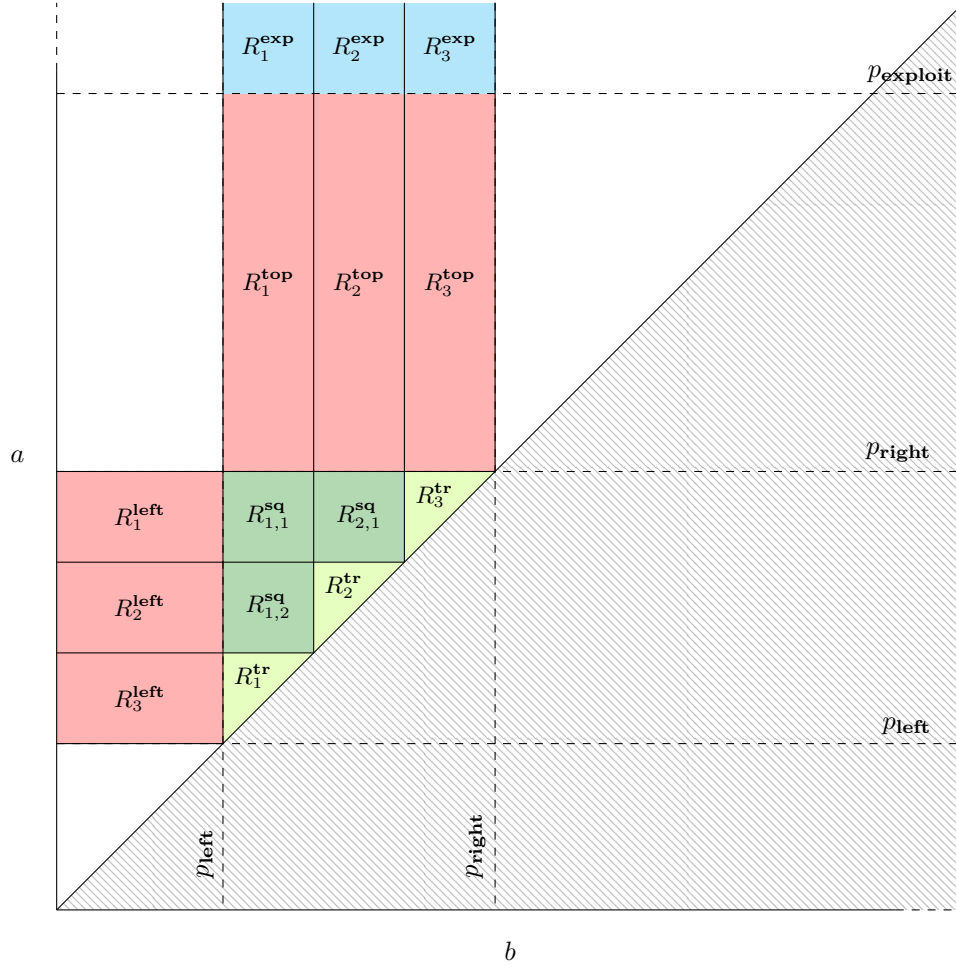Figure 5: (Not-to-scale) Division of the upper triangle $\mathcal{U}$ in regions with $K = 3$, to illustrate the relative positions of the exploitation regions (in blue) and the exploration regions (in red, green and lime).

Also, for each $i, j, k \in [K]$ with $i + j \leq K$, define

$$N_{i,j,k}^{\mathbf{square}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1^k, \ldots, M_{s-1}, W_{s-1}^k) \in R_{i,j}^{\mathbf{square}} \right\}$$

and for each $i, j \in [K]$ with $i + j \leq K$, define

$$N_{i,j}^{\mathbf{square}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1, \ldots, M_{s-1}, W_{s-1}) \in R_{i,j}^{\mathbf{square}} \right\} .$$

Then, for each $i, k \in [K]$, define

$$N_{i,k}^{\mathbf{triangle}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1^k, \ldots, M_{s-1}, W_{s-1}^k) \in R_i^{\mathbf{triangle}} \right\}$$

and for each $i \in [K]$, define

$$N_i^{\text{triangle}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1, \ldots, M_{s-1}, W_{s-1}) \in R_i^{\text{triangle}} \right\} .$$

Moreover, for each $i, k \in [K]$, define

$$N_{i,k}^{\text{exploit}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1^k, \ldots, M_{s-1}, W_{s-1}^k) \in R_i^{\text{exploit}} \right\}$$

and for each $i \in [K]$, define

$$N_i^{\text{exploit}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1, \ldots, M_{s-1}, W_{s-1}) \in R_i^{\text{exploit}} \right\} .$$

Finally, for each $k \in [K]$, define

$$N_k^{\text{white}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1^k, \ldots, M_{s-1}, W_{s-1}^k) \in R^{\text{white}} \right\}$$

and for each $i \in [K]$, define

$$N^{\text{white}}(t) := \sum_{s=1}^{t} \mathbb{I}\left\{ \mathscr{A}_s(M_1, W_1, \ldots, M_{s-1}, W_{s-1}) \in R^{\text{white}} \right\} .$$

Let R_T^k be the regret of the algorithm $\mathscr{A}$ up to the time horizon T when the underlying instance is $(V_t^k, M_t)_{t \in \mathbb{N}}$ and let R_T be the regret of the algorithm $\mathscr{A}$ up to the time horizon T when the underlying instance is $(V_t, M_t)_{t \in \mathbb{N}}$. Start by considering

$$\frac{1}{K} \sum_{k \in [K]} R_T^k \geq \frac{1}{K} \sum_{k \in [K]} \left(c_{\text{spike}} \cdot \varepsilon_K \cdot \left(T - \mathbb{E}\left[N_{k,k}^{\text{exploit}}(T) \right] \right) \right) =: (\square) .$$

We recall that if $(\mathcal{X}, \mathcal{F})$ is a measurable space and X is a $\mathcal{X}$-valued random variable, we denote by $\mathbb{P}_X$ the push-forward probability measure of $\mathbb{P}$ induced by X on $\mathcal{X}$, i.e., $\mathbb{P}_X[E] := \mathbb{P}[X \in E]$, for any $E \in \mathcal{F}$. Also, with $\|\cdot\|_{\text{TV}}$ we denote the total variation norm for measures, and for any two probability measures $\mathbb{Q}, \mathbb{Q}'$ defined on the same sample space, we denote their Kullback-Leibler divergence by $\mathcal{D}_{\text{KL}}(\mathbb{Q}, \mathbb{Q}')$. Now, for each $k \in [K]$, using Pinsker's inequality (Tsybakov, 2008, Lemma 2.5) that upper bounds the difference in the total variation $\|\cdot\|_{\text{TV}}$ of two probability measures using their Kullback-Leibler divergence $\mathcal{D}_{\text{KL}}$, we have that

$$\left| \mathbb{E}[N_{k,k}^{\text{exploit}}(T)] - \mathbb{E}[N_k^{\text{exploit}}(T)] \right| \leq \sum_{t=1}^{T} \left| \mathbb{P}[(B_t^k, A_t^k) \in R_k^{\text{exploit}}] - \mathbb{P}[(B_t, A_t) \in R_k^{\text{exploit}}] \right|$$

$$\leq \sum_{t=1}^{T-1} \left\| \mathbb{P}_{(M_1, W_1, \ldots, M_t, W_t)} - \mathbb{P}_{(M_1, W_1^k, \ldots, M_t, W_t^k)} \right\|_{\text{TV}}$$

$$\leq \sum_{t=1}^{T-1} \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}\left(\mathbb{P}_{(M_1, W_1, \ldots, M_t, W_t)}, \mathbb{P}_{(M_1, W_1^k, \ldots, M_t, W_t^k)} \right)} =: (\star_k) .$$

Now, for each $t \in [T-1]$, each $m_1, \ldots, m_t \in [0,1]$, and each $w_1, \ldots, w_t \in \{0,1\}^2$, defining $a \coloneqq \mathcal{A}_{t+1}(m_1, w_1, \ldots, m_t, w_t)$ and $b \coloneqq \mathcal{B}_{t+1}(m_1, w_1, \ldots, m_t, w_t)$, we have

$$
\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}_{(M_{t+1}, W_{t+1})|(M_1, W_1, \ldots, M_t, W_t) = (m_1, w_1, \ldots, m_t, w_t)),}
$$
$$
\mathbb{P}_{(M_{t+1}, W_{t+1}^k)|(M_1, W_1^k, \ldots, M_t, W_t^k) = (m_1, w_1, \ldots, m_t, w_t))} \big)
$$
$$
= \mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1} \le b\}, \mathbb{I}\{a < V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k \le b\}, \mathbb{I}\{a < V_{t+1}^k\})} \right)
$$
$$
\cdot \left(\mathbb{I}\{(a,b) \in R_k^{\mathbf{left}}\} + \sum_{j=1}^{k-1} \mathbb{I}\{(a,b) \in R_{k,j}^{\mathbf{square}}\} + \mathbb{I}\{(a,b) \in R_k^{\mathbf{triangle}}\} \right.
$$
$$
\left. + \sum_{i=1}^{k-1} \mathbb{I}\{(a,b) \in R_{i,k}^{\mathbf{square}}\} + \mathbb{I}\{(a,b) \in R_k^{\mathbf{top}}\} + \mathbb{I}\{(a,b) \in R_k^{\mathbf{exploit}}\} \right)
$$

A direct verification (see Lemma 9) shows that, letting $c_2 \coloneqq {}^{65}/_9$, for each $k \in [K]$, and $i, j \in [K-1]$, if $(a,b) \in R_k^{\mathbf{left}} \cup R_{k,j}^{\mathbf{square}} \cup R_k^{\mathbf{triangle}} \cup R_{i,k}^{\mathbf{square}} \cup R_k^{\mathbf{top}}$, then

$$
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1} \le b\}, \mathbb{I}\{a < V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k \le b\}, \mathbb{I}\{a < V_{t+1}^k\})} \right) \le c_2 \cdot \varepsilon_K ,
$$

likewise (see, again, Lemma 9), letting $c_1 \coloneqq {}^2/_{81}$ such that, for each $k \in [K]$, if $(a,b) \in R_k^{\mathbf{exploit}}$, then

$$
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1} \le b\}, \mathbb{I}\{a < V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k \le b\}, \mathbb{I}\{a < Z_{t+1}^k\})} \right) \le c_1 \cdot \varepsilon_K^2 .
$$

For notational convenience, set

$$
\mathbb{P}_{t+1, m_{1:t}, w_{1:t}}^k \coloneqq \mathbb{P}_{(M_{t+1}, W_{t+1}^k)|(M_1, W_1^k, \ldots, M_t, W_t^k) = (m_1, w_1, \ldots, m_t, w_t))}
$$
$$
\mathbb{P}_{t+1, m_{1:t}, w_{1:t}} \coloneqq \mathbb{P}_{(M_{t+1}, W_{t+1})|(M_1, W_1, \ldots, M_t, W_t) = (m_1, w_1, \ldots, m_t, w_t))}
$$

and notice that, for each $t \in [T-1]$ such that $t \ge 2$, using the chain rule for the Kullback-Leibler divergence (Cover and Thomas, 2006, Theorem 2.5.3), we have that

$$
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(M_1, W_1, \ldots, M_t, W_t)}, \mathbb{P}_{(M_1, W_1^k, \ldots, M_t, W_t^k)} \right)
$$
$$
= \mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(M_1, W_1, \ldots, M_{t-1}, W_{t-1})}, \mathbb{P}_{(M_1, W_1^k, \ldots, M_{t-1}, W_{t-1}^k)} \right)
$$
$$
+ \int_{([0,1] \times \{0,1\}^2)^{t-1}} \mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}_{t, m_{1:t-1}, w_{1:t-1}}, \mathbb{P}_{t, m_{1:t-1}, w_{1:t-1}}^k \big)
$$
$$
\mathrm{d}\mathbb{P}_{(M_1, W_1, \ldots, M_{t-1}, W_{t-1})}(m_1, w_1, \ldots, m_{t-1}, w_{t-1})
$$
$$
\le \mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(M_1, W_1^k, \ldots, M_{t-1}, W_{t-1}^k)}, \mathbb{P}_{(M_1, W_1, \ldots, M_{t-1}, W_{t-1})} \right)
$$
$$
+ c_1 \cdot \varepsilon_K^2 \cdot \mathbb{P}\left[(A_t, B_t) \in R_k^{\mathbf{exploit}} \right]
$$
$$
+ c_2 \cdot \varepsilon_K \cdot \left(\mathbb{P}\left[(B_t, A_t) \in R_k^{\mathbf{left}} \right] + \sum_{j=1}^{k-1} \mathbb{P}\left[(B_t, A_t) \in R_{k,j}^{\mathbf{square}} \right] \right.
$$
$$
\left. + \mathbb{P}\left[(B_t, A_t) \in R_k^{\mathbf{triangle}} \right] + \sum_{i=1}^{k-1} \mathbb{P}\left[(B_t, A_t) \in R_{i,k}^{\mathbf{square}} \right] + \mathbb{P}\left[(B_t, A_t) \in R_k^{\mathbf{top}} \right] \right)
$$

32

and iterating (and repeating essentially the same calculations in the last step where there is no conditioning), we get

$$
\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{(M_1,W_1,\ldots,M_t,W_t)},\mathbb{P}_{(M_1,W_1^k,\ldots,M_t,W_t^k)}\right) \\
&\leq c_1 \cdot \varepsilon_K^2 \cdot \mathbb{E}\left[N_k^{\textbf{exploit}}(t-1)\right] \\
&\quad + c_2 \cdot \varepsilon_K \cdot \left(\mathbb{E}\left[N_k^{\textbf{left}}(t-1)\right] + \sum_{j=1}^{k-1}\mathbb{E}\left[N_{k,j}^{\textbf{square}}(t-1)\right]\right. \\
&\qquad \left. + \mathbb{E}\left[N_k^{\textbf{triangle}}(t-1)\right] + \sum_{i=1}^{k-1}\mathbb{E}\left[N_{i,k}^{\textbf{square}}(t-1)\right] + \mathbb{E}\left[N_k^{\textbf{top}}(t-1)\right]\right) \\
&\leq c_1 \cdot \varepsilon_K^2 \cdot \mathbb{E}\left[N_k^{\textbf{exploit}}(T)\right] \\
&\quad + c_2 \cdot \varepsilon_K \cdot \left(\mathbb{E}\left[N_k^{\textbf{left}}(T)\right] + \sum_{j=1}^{k-1}\mathbb{E}\left[N_{k,j}^{\textbf{square}}(T)\right]\right. \\
&\qquad \left. + \mathbb{E}\left[N_k^{\textbf{triangle}}(T)\right] + \sum_{i=1}^{k-1}\mathbb{E}\left[N_{i,k}^{\textbf{square}}(T)\right] + \mathbb{E}\left[N_k^{\textbf{top}}(T)\right]\right)
\end{aligned}
$$

It follows that, for each $k \in [K]$,

$$
\begin{aligned}
(\star_k) &\leq T \cdot \sqrt{\frac{1}{2}} \cdot \left(c_1 \cdot \varepsilon_K^2 \cdot \mathbb{E}\left[N_k^{\textbf{exploit}}(T)\right] + c_2 \cdot \varepsilon_K \cdot \left(\mathbb{E}\left[N_k^{\textbf{left}}(T)\right] + \sum_{j=1}^{k-1}\mathbb{E}\left[N_{k,j}^{\textbf{square}}(T)\right]\right.\right. \\
&\qquad \left.\left. + \mathbb{E}\left[N_k^{\textbf{triangle}}(T)\right] + \sum_{i=1}^{k-1}\mathbb{E}\left[N_{i,k}^{\textbf{square}}(T)\right] + \mathbb{E}\left[N_k^{\textbf{top}}(T)\right]\right)\right)^{1/2} \\
&\leq T \cdot \sqrt{\frac{1}{2}} \cdot \varepsilon_K \cdot \sqrt{c_1 \mathbb{E}\left[N_k^{\textbf{exploit}}(T)\right]} \\
&\quad + T \cdot \sqrt{\frac{1}{2}} \cdot \sqrt{\varepsilon_K} \cdot \left(c_2 \cdot \left(\mathbb{E}\left[N_k^{\textbf{left}}(T)\right] + \sum_{j=1}^{k-1}\mathbb{E}\left[N_{k,j}^{\textbf{square}}(T)\right]\right.\right. \\
&\qquad \left.\left. + \mathbb{E}\left[N_k^{\textbf{triangle}}(T)\right] + \sum_{i=1}^{k-1}\mathbb{E}\left[N_{i,k}^{\textbf{square}}(T)\right] + \mathbb{E}\left[N_k^{\textbf{top}}(T)\right]\right)\right)^{1/2}
\end{aligned}
$$

For notational convenience, define

$$
N^{\textbf{explore}} := \sum_{i\in[K]} N_i^{\textbf{left}}(T) + \sum_{j\in[K]} N_j^{\textbf{top}}(T) + \sum_{k\in[K]} N_k^{\textbf{triangle}}(T) + \sum_{i,j\in[K],\,i+j\leq K} N_{i,j}^{\textbf{square}}(T)
$$

and

$$
N^{\textbf{exploit}} := \sum_{k\in[K]} N_k^{\textbf{exploit}}(T)
$$

Notice that, after bringing the summation under the square root leveraging Jensen's inequality, we sum each of the terms $N_i^{\textbf{left}}(T), N_j^{\textbf{top}}(T), N_k^{\textbf{triangle}}(T), N_{i,j}^{\textbf{square}}(T)$ at most two times, we get:

$$\frac{1}{K} \sum_{k \in [K]} (\star_k) \leq T \cdot \varepsilon_K \cdot \sqrt{\frac{c_1}{2}} \cdot \sqrt{\frac{\mathbb{E}\left[N^{\textbf{exploit}}\right]}{K}} + T \cdot \sqrt{\varepsilon_K} \cdot \sqrt{c_2} \cdot \sqrt{\frac{\mathbb{E}\left[N^{\textbf{explore}}\right]}{K}}$$

$$\leq T \cdot \varepsilon_K \cdot \sqrt{\frac{c_1}{2}} \cdot \sqrt{\frac{T}{K}} + T \cdot \sqrt{\varepsilon_K} \cdot \sqrt{c_2} \cdot \sqrt{\frac{\mathbb{E}\left[N^{\textbf{explore}}\right]}{K}} \, .$$

It follows that

$$(\square) = c_{\textbf{spike}} \cdot \varepsilon_K \cdot \left(T - \frac{1}{K} \sum_{k \in [K]} \mathbb{E}[N_{k,k}^{\textbf{exploit}}] \right)$$

$$\geq c_{\textbf{spike}} \cdot \varepsilon_K \cdot \left(T - \frac{1}{K} \sum_{k \in [K]} \mathbb{E}[N_k^{\textbf{exploit}}] - \frac{1}{K} \sum_{k \in [K]} (\star_k) \right)$$

$$= c_{\textbf{spike}} \cdot \varepsilon_K \cdot \left(T - \frac{\mathbb{E}[N^{\textbf{exploit}}]}{K} - \frac{1}{K} \sum_{k \in [K]} (\star_k) \right)$$

$$\geq c_{\textbf{spike}} \cdot \varepsilon_K \cdot T \cdot \left(1 - \frac{1}{K} - \varepsilon_K \cdot \sqrt{\frac{c_1}{2}} \cdot \sqrt{\frac{T}{K}} - \sqrt{\varepsilon_K} \cdot \sqrt{c_2} \cdot \sqrt{\frac{\mathbb{E}\left[N^{\textbf{explore}}\right]}{K}} \right) =: (\widetilde{\square}) \, .$$

Now, if $\mathbb{E}\left[N^{\textbf{explore}}\right] \geq T^{2/3}$ then

$$R_T \geq c_{\textbf{plat}} \cdot \mathbb{E}\left[N^{\textbf{explore}}\right] \geq c_{\textbf{plat}} \cdot T^{2/3} \, .$$

Instead, if $\mathbb{E}\left[N^{\textbf{explore}}\right] \leq T^{2/3}$, then

$$(\widetilde{\square}) \geq c_{\textbf{spike}} \cdot \varepsilon_K \cdot T \cdot \left(1 - \frac{1}{K} - \varepsilon_K \cdot \sqrt{\frac{c_1}{2}} \cdot \sqrt{\frac{T}{K}} - \sqrt{\varepsilon_K} \cdot \sqrt{c_2} \cdot \sqrt{\frac{T^{2/3}}{K}} \right)$$

$$= c_{\textbf{spike}} \cdot \frac{T}{16\lceil T^{1/3} \rceil} \left(1 - \frac{1}{\lceil T^{1/3} \rceil} - \frac{1}{16\lceil T^{1/3} \rceil} \cdot \sqrt{\frac{c_1}{2}} \cdot \sqrt{\frac{T}{\lceil T^{1/3} \rceil}} - \sqrt{\frac{1}{16\lceil T^{1/3} \rceil}} \cdot \sqrt{c_2} \cdot \sqrt{\frac{T^{2/3}}{\lceil T^{1/3} \rceil}} \right)$$

$$\geq 10^{-6} \cdot T^{2/3}$$

for $T \geq 42$, where in the last step we have plugged in the values of c_1 and c_2. Since we proved that

$$\frac{1}{K} \sum_{k \in [K]} R_T^k \geq 10^{-6} \cdot T^{2/3} \, ,$$

then, there exists an instance $k \in [K]$ such that the regret of the algorithm over $(M_t, V_{\varepsilon_K, r_K^k, t})_{t \in [T]}$ is at least $T^{2/3}$, concluding the result. ∎

The previous proof relies on Lemma 9, for which we need the following two lemmas.

Lemma 7 *For any pair $\varepsilon, r \in (0,1]^2$, $\Lambda_{r,\varepsilon}$ is $\frac{2}{\varepsilon}$-Lipschitz.*

Proof Fix $0 \leq b \leq a \leq 1$, we consider all possible cases with respect to the interval $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+ = [r - \varepsilon/2, r] \cup (r, r + \varepsilon/2]$ to prove that

$$|\Lambda(a) - \Lambda(b)| \leq \frac{2}{\varepsilon}|a - b|$$

- If $b, a \notin \mathcal{R}$, then $\Lambda(a) = \Lambda(b) = 0$ and the property is trivially true.

- If $b, a \in \mathcal{R}^-$ or $b, a \in \mathcal{R}^+$ then

$$|\Lambda(a) - \Lambda(b)| = \left|1 - \frac{2}{\varepsilon}(r - a) - 1 + \frac{2}{\varepsilon}(r - b)\right| = \frac{2}{\varepsilon}|a - b|$$

- If $b \in \mathcal{R}^-$ and $a \in \mathcal{R}^+$ then

$$|\Lambda(a) - \Lambda(b)| = \left|1 - \frac{2}{\varepsilon}(a - r) - 1 + \frac{2}{\varepsilon}(r - b)\right| = \frac{2}{\varepsilon}|a - b - 2r| \leq \frac{2}{\varepsilon}|a - b|$$

  because $|s - b| > 2r$.

- If $b \in \mathcal{R}^+$ and $a \notin \mathcal{R}$ then

$$|\Lambda(a) - \Lambda(b)| = \left|1 - \frac{2}{\varepsilon}(b - r)\right| = \frac{2}{\varepsilon}\left|\frac{\varepsilon}{2} + r - b\right| \leq \frac{2}{\varepsilon}|a - b|$$

  because $\varepsilon/2 + r \leq a$. The same reasoning applies to the case $b \notin \mathcal{R}$ and $a \in \mathcal{R}^+$.

- If $b \in \mathcal{R}^-$ and $a \notin \mathcal{R}$ then

$$|\Lambda(a) - \Lambda(b)| = \left|1 - \frac{2}{\varepsilon}(r - b)\right| = \frac{2}{\varepsilon}\left|\frac{\varepsilon}{2} - r + b\right| = \frac{2}{\varepsilon}\left|r - \frac{\varepsilon}{2} - b\right| \leq \frac{2}{\varepsilon}|a - b|$$

  because $r - \varepsilon/2 \leq -a$. The same reasoning applies to the case $b \notin \mathcal{R}$ and $a \in \mathcal{R}^-$.

∎

The following lemma, together with Lemma 7, will be used in the proof of Lemma 9.

Lemma 8 *For all pairs $(b, a) \in \mathcal{U} \setminus [p_{exploit}, 1]^2$, it holds that*

$$F(a) - F(b) \geq \frac{1}{6}(a - b)$$

Proof Notice that, for any pair $(b, a) \in \mathcal{U} \setminus [p_{\text{exploit}}, 1]^2$,

$$F(a) - F(b) = \int_b^a f(x)\,\mathrm{d}x \geq \min_{c \in [b,a]} f(c)(a - b)\,,$$

The minimum value of f is found when approaching $3/16$ from the right, which yields $f(c) \to 1/6$, therefore $F(a) - F(b) \geq 1/6 \cdot (a - b)$. ∎

We conclude this section by proving a key lemma used in the proof of Theorem 2.

Lemma 9 *For all $K \in \mathbb{N}$, $k \in [K]$, If $(a, b) \in R_k^{exploit}$, then we have*

$$\mathcal{D}_{KL}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1}<b\}, \mathbb{I}\{a<V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k<b\}, \mathbb{I}\{a<V_{t+1}^k\})}\right) \leq \frac{2}{81} \cdot \varepsilon_K^2 \ ,$$

while if $(a, b) \in R_k^{left} \cup R_{k,j}^{square} \cup R_k^{triangle} \cup R_{i,j}^{square} \cup R_k^{top}$, then,

$$\mathcal{D}_{KL}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1}<b\}, \mathbb{I}\{a<V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k<b\}, \mathbb{I}\{a<V_{t+1}^k\})}\right) \leq \frac{65}{9} \cdot \varepsilon_K \ .$$

Proof Fix $K \in \mathbb{N}$, $\varepsilon_K = 1/16K$ and $r_K^k = 3/16 + \varepsilon_K(k - 1/2)$, the KL divergence can be decomposed as follows

$$\mathcal{D}_{KL}\left(\mathbb{P}_{(\mathbb{I}\{V_{t+1}<b\}, \mathbb{I}\{a<V_{t+1}\})}, \mathbb{P}_{(\mathbb{I}\{V_{t+1}^k<b\}, \mathbb{I}\{a<V_{t+1}^k\})}\right)$$

$$= \log\left(\frac{F(b)}{F_{r_K^k, \varepsilon_K}(b)}\right) F(b) + \log\left(\frac{1 - F(a)}{1 - F_{r_K^k, \varepsilon_K}(a)}\right)(1 - F(a))$$

$$+ \log\left(\frac{F(a) - F(b)}{F_{r_K^k, \varepsilon_K}(a) - F_{r_K^k, \varepsilon_K}(b)}\right)(F(a) - F(b))$$

$$= \log\left(\frac{F(b)}{F(b) + \frac{\varepsilon_K}{18}\Lambda_{r_K^k, \varepsilon_K}(b)}\right) F(b) + \log\left(\frac{1 - F(a)}{1 - F(a) - \frac{\varepsilon_K}{18}\Lambda_{r_K^k, \varepsilon_K}(a)}\right)(1 - F(a))$$

$$+ \log\left(\frac{F(a) - F(b)}{F(a) - F(b) + \frac{\varepsilon_K}{18}(\Lambda_{r_K^k, \varepsilon_K}(a) - \Lambda_{r_K^k, \varepsilon_K}(b))}\right)(F(a) - F(b)) = (\star)$$

For sake of brevity, we will write Λ instead of $\Lambda_{r_K^k, \varepsilon_K}$ for the remainder of the proof. Note that the first term is non-positive for any $b \in [0, 1]$, while the second term is non-negative for any $a \in [0, 1]$. Furthermore if $\Lambda(b) = 0$, then $\log(F(b)/F(b))F(b) = 0$, likewise if $\Lambda(a) = 0$ then $\log(F(a)/F(a))F(a) = 0$, if both $\Lambda(b) = \Lambda(a) = 0$ then the whole expression is zero.

Exploitation region If $(a, b) \in R_k^{exploit}$, then $\Lambda(a) = 0$ because all perturbations happen within p_{left} and p_{right} whereas $a > p_{exploit}$, If in addition $\Lambda(b) = 0$ then the whole expression is zero, otherwise

$$(\star) \leq \log\left(\frac{F(a) - F(b)}{F(a) - F(b) - \frac{\varepsilon_K}{18}\Lambda(b)}\right)(F(a) - F(b)) + \log\left(\frac{F(b)}{F(b) + \frac{\varepsilon_K}{18}\Lambda(b)}\right) F(b)$$

$$= \log\left(\frac{1}{\left(1 - \frac{\varepsilon_K \Lambda(b)}{18(F(a)-F(b))}\right)^{(F(a)-F(b))}}\right) + \log\left(\frac{1}{\left(1 + \frac{\varepsilon_K \Lambda(b)}{18F(b)}\right)^{F(b)}}\right)$$

$$= \log\left(\frac{1}{\left(1 - \frac{1}{\frac{18(F(a)-F(b))}{\varepsilon_K \Lambda(b)}}\right)^{\frac{18(F(a)-F(b))}{\varepsilon_K \Lambda(b)} \cdot \frac{\varepsilon_K \Lambda(b)}{18}}}\right) + \log\left(\frac{1}{\left(1 + \frac{1}{\frac{18F(b)}{\varepsilon_K \Lambda(b)}}\right)^{\frac{18F(b)}{\varepsilon_K \Lambda(b)} \cdot \frac{\varepsilon_K \Lambda(b)}{18}}}\right) = (\circ)$$

We know that for $x > 1$ both $(1 - 1/x)^x$ and $(1 + 1/x)^x$ are monotonically increasing functions of x. Now, note that both $F(a) - F(b)$ and $F(b)$ are lower bounded by a constant. Indeed, $p_{left} \leq$

$b \leq p_{\text{right}}$ and $a \geq p_{\text{exploit}}$, thus $F(a) - F(b) \geq F(p_{\text{exploit}}) - F(p_{\text{right}})$ and $F(b) \geq F(p_{\text{left}})$. Let $c = \min\{F(p_{\text{exploit}}) - F(p_{\text{right}}), F(p_{\text{left}})\} = 1/6$. Therefore we have that both

$$\frac{18(F(a) - F(b))}{\varepsilon_K \Lambda(b)} \geq \frac{18c}{\varepsilon_K \Lambda(b)} > 1 \qquad \text{and} \qquad \frac{18F(b)}{\varepsilon_K \Lambda(b)} \geq \frac{18c}{\varepsilon_K \Lambda(b)} > 1$$

Now, using the monotonicity of the denominators, we can write

$$(\circ) \leq \log\left(\frac{1}{\left(1 - \frac{\varepsilon_K \Lambda(b)}{18c}\right)^c}\right) + \log\left(\frac{1}{\left(1 + \frac{\varepsilon_K \Lambda(b)}{18c}\right)^c}\right) = c \cdot \log\left(\frac{1}{1 - \left(\frac{\varepsilon_K \Lambda(b)}{18c}\right)^2}\right)$$

$$\leq c \cdot \frac{\left(\frac{\varepsilon_K \Lambda(b)}{18c}\right)^2}{1 - \left(\frac{\varepsilon_K \Lambda(b)}{18c}\right)^2} \leq \frac{4}{3} \cdot c \cdot \left(\frac{\varepsilon_K \Lambda(b)}{18c}\right)^2 = \frac{4}{3} \cdot \frac{1}{c} \cdot \frac{1}{18^2} \cdot \varepsilon_K^2$$

Here, we first use the inequality that $\log x \leq x - 1$, then we use the facts that $\Lambda(b) \leq 1$ and $\varepsilon_K \leq 9c$ (loose bound). In practice $c = 1/6$ and $\varepsilon_K \leq 9c = 3/2$ holds for any $K \in \mathbb{N}$.

Exploration regions   If $(a, b) \in R_k^{\textbf{left}} \cup R_{k,j}^{\textbf{square}} \cup R_k^{\textbf{triangle}} \cup R_{i,k}^{\textbf{square}} \cup R_k^{\textbf{top}}$, then consider each term in $(\star)$ individually.

The first term is upper bounded by zero for any $b \in [0, 1]$.

If $a > p_{\text{right}}$, then $\Lambda(a) = 0$ and the second term is zero, otherwise the second term can be bounded as follows:

$$\log\left(\frac{1 - F(a)}{1 - F(a) - \frac{\varepsilon_K}{18}\Lambda(a)}\right)(1 - F(a)) \leq \frac{\varepsilon_K}{18} \cdot \frac{\Lambda(a)}{1 - F(a) - \frac{\varepsilon_K}{18}\Lambda(a)}(1 - F(a))$$

$$= \frac{\varepsilon_K}{18} \cdot \frac{\Lambda(a)}{1 - \frac{\varepsilon_K}{18}\frac{\Lambda(a)}{1 - F(a)}} \leq \frac{\varepsilon_K}{18} \cdot \frac{1}{1 - \frac{\varepsilon_K}{18}\frac{1}{1 - F(p_{\text{right}})}} \leq \frac{\varepsilon_K}{9}$$

where we used the inequality $\log x < x - 1$ and the fact that $\frac{1}{1 - \frac{\varepsilon_K}{18}F(p_{\text{right}})} \leq 1/2$, which is always true because $\varepsilon_K \leq 9(1 - F(p_{\text{right}})) \leq 81/11$ holds for any $K \in \mathbb{N}$.

Finally, we just need to bound the third term. We consider the following cases.

- First consider the case in which $a - b \leq \varepsilon_K$, since F is Lipschitz with constant L we have $F(a) - F(b) \leq L\varepsilon_K$, thus

$$\log\left(\frac{F(a) - F(b)}{F(a) - F(b) + \frac{\varepsilon_K}{18}(\Lambda(a) - \Lambda(b))}\right)(F(a) - F(b)) = \log\left(\frac{1}{1 + \frac{\varepsilon_K}{18} \cdot \frac{\Lambda(a) - \Lambda(b)}{F(a) - F(b)}}\right)L\varepsilon_K$$

  which is maximized when $\frac{\Lambda(a) - \Lambda(b)}{F(a) - F(b)}$ is minimized, by leveraging the Lipschitzness of Λ from Lemma 7 together with the lower bound on $F(a) - F(b)$ from Lemma 8 we get

$$\frac{\Lambda(a) - \Lambda(b)}{F(a) - F(b)} \geq \frac{-2/\varepsilon_K(a - b)}{1/6(a - b)} = -\frac{12}{\varepsilon_K}$$

Substituting in the above, we get the upper bound:

$$\log\left(\frac{1}{1 - \frac{\varepsilon_K}{18}\frac{12}{\varepsilon_K}}\right)L\varepsilon_K < 2L\varepsilon_K = \frac{64}{9}\cdot\varepsilon_K$$

where $L = \frac{32}{9}$ by the definition of F in Equation (2). For the case $a = b$, note that this term is zero because $\Lambda(a) - \Lambda(b) = 0$ and can therefore be ignored.

- Next consider the opposite $a - b > \varepsilon_K$. By Lemma 8 we know that $F(a) - F(b) \geq \frac{1}{6}(a - b) \geq \frac{\varepsilon_K}{6}$. We use the inequality $\log x \leq x - 1$ to get:

$$\log\left(\frac{F(a) - F(b)}{F(a) - F(b) + \frac{\varepsilon_K}{18}(\Lambda(a) - \Lambda(b))}\right)(F(a) - F(b))$$

$$\leq -\frac{\varepsilon_K}{18}\cdot\frac{\Lambda(a) - \Lambda(b)}{F(a) - F(b) + \frac{\varepsilon_K}{18}(\Lambda(a) - \Lambda(b))}(F(a) - F(b))$$

$$\leq \frac{\varepsilon_K}{18}\cdot\frac{F(a) - F(b)}{F(a) - F(b) - \frac{\varepsilon_K}{18}} = \frac{\varepsilon_K}{18}\cdot\frac{1}{1 - \frac{\varepsilon_K}{18}\frac{1}{F(a)-F(b)}}$$

$$\leq \frac{\varepsilon_K}{18}\cdot\frac{1}{1 - \frac{\varepsilon_K}{18}\frac{6}{\varepsilon_K}} = \frac{1}{12}\cdot\varepsilon_K$$

In conclusion, the result holds with $c_1 = \frac{2}{81}$ and $c_2 = \frac{1}{9} + \frac{64}{9} = \frac{65}{9}$. ∎

Appendix E. Proof of Theorem 3

Given that the sequence is $(M_t, V_t)_{t\in[T]}$ i.i.d., without any loss of generality we can consider deterministic algorithms to post prices $(B_1, A_1), \ldots, (B_T, A_T)$. Fix a time horizon $T \in \mathbb{N}$. We build an instance where, for each $t \in [T]$, the random variable M_t takes values in $\{0, 1\}$. In this case, notice that there are at most 8^T sequences of feedback

$$(M_1, \mathbb{I}\{B_1 \geq V_1\}, \mathbb{I}\{A_1 < V_1\}), \ldots, (M_{T-1}, \mathbb{I}\{B_{T-1} \geq V_{T-1}\}, \mathbb{I}\{A_{T-1} < V_{T-1}\})$$

and, consequently, the set $\mathcal{P} \subset [0, 1]$ where the algorithm posts prices up to time contains at most $\sum_{t=1}^{T-1} 2\cdot 8^t$ different prices. Let $\varepsilon \in (0, \frac{1}{4})$. Given that $\mathcal{P}$ is finite, we can select $\frac{1}{2} - \varepsilon < c < d < \frac{1}{2} + \varepsilon$ such that the closed interval $[c, d]$ does not contain any point in $\mathcal{P}$. Now, select $(M_1, V_1), \ldots, (M_T, V_T)$ as an i.i.d. sequence drawn uniformly in the set $\{(0, d), (1, c)\}$. Notice that, for any time $t \in [T]$, a maker that posts $(b, a) \in \mathcal{U}$ such that $c < b = a < d$ gains (deterministically) at least $\frac{1}{2} - \varepsilon$ regardless of the realization of (M_t, V_t) (if $(M_t, V_t) = (0, d)$ then the maker always sells while if $(M_t, V_t) = (1, c)$ the maker always buys). On the other hand, the pair of prices (B_t, A_t) posted by the algorithm at any time $t \in [T]$ is such that one and only one of the following alternatives hold true:

- $B_t \leq A_t < c$. In this case, the maker always sells to the taker, but with probability $\frac{1}{2}$ the maker gains at most $\frac{1}{2}+\varepsilon$ (when $(M_t, V_t) = (0, d)$) and with probability $\frac{1}{2}$ the maker loses at least $\frac{1}{2} - \varepsilon$ (when $(M_t, V_t) = (1, c)$).

38

- $B_t < c$ and $d < A_t$. In this case, then the maker never sells or buys (because the proposed selling price is too high and the proposed buying price is too low) and hence the maker gains zero.

- $d < B_t \leq A_t$. In this case, then the maker always buys from the taker, but with probability $1/2$ the maker gains at most $1/2 + \varepsilon$ (when $(M_t, V_t) = (1, c)$) and with probability $1/2$ the maker loses at least $1/2 - \varepsilon$ (when $(M_t, V_t) = (0, d)$).

Overall, the algorithm gains in expectation at most 0 when the underlying instance is $(M_t, V_t)_{t \in [T]}$. Hence, for any fixed pair of prices $(b, a) \in \mathcal{U}$ such that $c < b = a < d$, we have that

$$
\begin{aligned}
R_T &\geq \mathbb{E}\left[\sum_{t=1}^{T} u(b, a, M_t, V_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} u(B_t, A_t, M_t, V_t)\right] \\
&\geq \left(\frac{1}{2} - \varepsilon\right) \cdot T - \left(\left(\frac{1}{2} + \varepsilon\right) \cdot \frac{1}{2} - \left(\frac{1}{2} - \varepsilon\right) \cdot \frac{1}{2}\right) \cdot T = \left(\frac{1}{2} - 2\varepsilon\right) \cdot T.
\end{aligned}
$$

Setting $\delta := 2\varepsilon$ and noticing that ε was chosen arbitrarily in the interval $(0, 1/4)$, the conclusion follows.