Learning Posted Prices in Bilateral Trade: Regret Guarantees under Full and Bandit Feedback

Luca Bruni

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Master's degree in Mathematics

Department of Mathematics
Faculty of Science
University of Ottawa

Prof. Maia Fraser

Table of Contents

List of Tables

List of Figures

Abstract

In this thesis we study an economically motivated sequential decision problem in which a learner repeatedly chooses an action (e.g., a posted price) and observes structured feedback. We ask how the information revealed after each decision determines whether learning is possible and what regret rates are achievable. We cast the problem in the online-learning framework and analyze two feedback models. Under full-feedback, the learner can effectively evaluate alternative actions; we give an efficient algorithm with sublinear regret and matching lower bounds, yielding sharp minimax rates. Under bandit-feedback, we show that without additional regularity, sublinear regret is impossible. We then identify natural smoothness conditions on the instance under which bandit learning becomes feasible again and derive regret guarantees. Overall, our results cleanly separate learnable from non-learnable regimes and quantify how mild structure can bridge the gap between full-feedback and bandit learning.

Dans cette thèse, nous étudions un problème de décision séquentielle motivé par l'économie, dans lequel un apprenant choisit de manière répétée une action (par exemple, un prix affiché) et observe un retour d'information structuré. Nous examinons comment l'information révélée après chaque décision détermine si l'apprentissage est possible et quels taux de regret peuvent être atteints. Nous formulons le problème dans le cadre de l'apprentissage en ligne et analysons deux modèles de retour d'information. En régime de retour complet, l'apprenant peut effectivement évaluer des actions alternatives ; nous proposons un algorithme efficace avec un regret sous-linéaire ainsi que des bornes inférieures correspondantes, ce qui fournit des taux minimax précis. En régime bandit, nous montrons qu'en l'absence de régularité supplémentaire, un regret sous-linéaire est impossible. Nous identifions ensuite des conditions naturelles de régularité (de type lissité) sous lesquelles l'apprentissage bandit redevient possible et nous en déduisons des garanties de regret. Dans l'ensemble, nos résultats distinguent nettement les régimes apprenables et non apprenables et quantifient comment une structure modeste peut réduire l'écart entre les retours complets et bandits.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Maia Fraser. Over the past two years, her guidance, patience, and steady support have been invaluable. She consistently encouraged me to aim higher, and she reassured me during the moments when I genuinely doubted I could make it to the end. This thesis would not exist without her mentorship.

I am deeply grateful to my husband, whose love and unwavering presence carried me through the most difficult parts of this journey. Thank you for keeping my morale up, for helping me overcome obstacles when I felt stuck, and for reminding me, again and again, that I was capable of finishing what I started.

I would also like to thank the members of my committee and the faculty who read this work and provided feedback that helped sharpen both the presentation and the ideas. I am grateful as well to the administrative staff of the department for their assistance throughout the program.

1

Introduction

Digital platforms routinely mediate trades between large numbers of anonymous users. Ride-sharing platforms match drivers and riders, online marketplaces match sellers and buyers of goods, and financial platforms match orders in electronic markets. In all these settings, the platform, also called *learning agent* or simply *learner*, must make *pricing decisions* under uncertainty about users' private willingness to pay or accept. A central question is:

> *Can a platform learn to post a near-optimal price*[1] *from repeated interactions with users whose valuations are initially unknown?*

This thesis studies this question in the canonical setting of *bilateral trade*. In its basic form, bilateral trade involves a single seller and a single buyer, each with a private valuation for an indivisible good. The seller's valuation S is the minimum price at which she is willing to sell; the buyer's valuation B is the maximum price that he is willing to pay. A mechanism takes as input specific information about the agents (the feedback methods will be defined later), and decides whether a trade occurs and, if so, at what price; see, e.g., the classical paper [1], the textbook [2], or the more recent monograph [3].

In the traditional mechanism-design formulation, the designer knows the distributions of S and B and is allowed to design arbitrary allocation and payment rules, subject to constraints such as incentive compatibility, individual rationality, and budget balance. The celebrated Myerson–Satterthwaite theorem [1] shows that these desiderata are fundamentally in tension: in general, no mechanism can be simultaneously efficient, truthful, individually

[1]The meaning of what is considered "near-optimal" will be explained in later sections

rational, and budget-balanced [2]. This has led to a rich literature that avoids the impossibility by relaxing at least one desideratum (e.g., approximate efficiency, weakened incentive constraints, or budget balance only in expectation), or by restricting attention to structured classes of mechanisms such as posted prices; see, for instance, [4–9].

An online learning perspective

In many modern applications, the platform repeatedly interacts with different buyer–seller pairs over time, and the most typical viewpoint is that of *online learning*: at each round $t = 1, 2, \ldots$, a new pair of agents with valuations (S_t, B_t) arrives; the platform posts a price P_t; and some feedback is observed. The platform aims to use past data to improve future prices.

This sequential interaction naturally leads to the language of online learning and regret minimization as developed in, e.g., [10], [11], and [8]. In our setting, the outcome space is the valuation space $\mathcal{X} \coloneqq [0,1]^2$, the space of possible buyer/seller pairs (S, B), the decision space is the set of posted prices $\mathcal{H} \coloneqq [0,1]$, and a reward function $v\colon \mathcal{H} \times \mathcal{X} \to \{0,1\}$ encodes whether a trade occurs at a given price[3]. At each round the platform selects a price $P_t \in \mathcal{H}$ (possibly at random, based on past observations), an outcome $X_t \coloneqq (S_t, B_t)$ is drawn from an unknown distribution on $\mathcal{X}$, and the platform receives reward $v(P_t, X_t)$. In this thesis, to lighten the notation, we will write $v(P_t, S_t, B_t)$ instead of $v(P_t, (S_t, B_t))$.

[2] These will be defined more formally in Chapter 6

[3] This binary reward structure is specific to our setting. In the general online learning framework introduced in Section 5.1, reward functions are allowed to take values in $[0, +\infty]$.

We consider two feedback models:

- In the *full-feedback* model, after posting P_t the platform observes the full valuation pair (S_t, B_t); it can therefore compute $v(p, S_t, B_t)$ for any hypothetical price p.
- In the *bandit-feedback* model, the platform only observes the reward $v(P_t, S_t, B_t)$. The underlying valuations remain unobserved, making learning substantially harder.

The agent's goal is to learn a pricing policy α with optimal performance over a time horizon T, as measured by its *regret* with respect to the best fixed price in hindsight. Let θ be the common distribution of the independent and identically distributed (i.i.d.) sequence of random pairs (S_t, B_t) and let $(S, B) \sim \theta$ be a generic valuation pair. Define

$$f_\theta(p) \coloneqq \mathbb{E}_\theta[v(p, S, B)]$$

to be the expected trading volume achieved by price p, and let

$$f_\theta^\star \coloneqq \sup_{p \in [0,1]} f_\theta(p)$$

be the benchmark value. The expected regret of α is

$$R_T(\alpha) \coloneqq T \cdot f_\theta^\star - \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(P_t, S_t, B_t) \right].$$

Intuitively, $R_T(\alpha)$ measures how much reward is lost due to not knowing the optimal price in advance. As it will be discussed in detail in Section 5, the regime of primary interest is

that of *sublinear* regret $R_T(\alpha) = o(T)$, which is equivalent (see Section 5.2) to the statement that the algorithm's average per-round reward converges to that of the best fixed benchmark value. In this thesis, we work in a stochastic setting in which the sequence $(S_t, B_t)_{t\geq 1}$ is i.i.d. drawn from an unknown distribution on $[0,1]^2$. Our central question is:

> *What rates of regret are achievable for online posted-price mechanisms in bilateral trade, under full and bandit feedback?*

In particular, we are interested in identifying structural assumptions under which bandit-feedback still permits sublinear regret, and in understanding the gap between full-information and bandit regimes.

2

Author's Contributions

	Stochastic		
	i.i.d.		+ bounded density
	Upper bound	Lower bound	Upper bound
Full-feedback	$T^{1/2}$ (Thm 1)	$T^{1/2}$ (Thm 2)	$T^{1/2}$
Bandit-feedback	–	T (Thm 5)	$T^{2/3}$ (Thm 6)

Table 2.1: Summary of regret rates in the stochastic i.i.d. settings. The rate of $T^{1/2}$ as the upper bound in full-feedback with bounded density follows directly from the upper bound in the i.i.d. setting (Thm 1).

In this thesis, we first formalize bilateral trade as an online learning problem. Our formulation cleanly separates the outcome space (valuation pairs), the decision space (prices), the reward function (trading volume), and the feedback model (full vs. bandit). Parts of this thesis are expository: we review standard background on online learning, regret minimization, and stochastic bandits, and we present classical tools used in the subsequent analyses. This perspective allows us to import standard tools from online learning and to state regret guarantees in a unified way.

Our original contributions concern the specific posted-price bilateral-trade framework with trading volume rewards and the full-feedback vs. bandit comparison developed here. In particular, while the "Follow the Best Price" strategy also appears in [12] and our full-feedback upper-bound analysis follows a similar high-level route, the lower bound in full-feedback and the bandit-feedback results (impossibility without structure and sublinear regret under bounded density) are, to the best of our knowledge, new in this setting.

- **Full-feedback setting.** In the full-feedback model, the platform observes the pair of

evaluations (S_t, B_t) after each interaction. We use a pricing algorithm "Follow the Best Price", which also appears in [12], and our full-feedback upper-bound analysis follows a similar high-level route, but adapted to our reward function (defined on page 32). This algorithm achieves sublinear regret, and we prove a matching lower bound (up to constants), using a different proof route, showing that our rate $\Omega(\sqrt{T})$ is essentially optimal. This characterizes the intrinsic difficulty of online bilateral trade when full-information is available.

- **Bandit setting: impossibility without structure.** In the bandit-feedback model without further assumptions, we prove that *no* algorithm can achieve sublinear regret: for any algorithm α, there exists a constant $c \in \mathbb{R}$ such that for all T there exists a distribution over valuations (which could potentially depend on T) such that $R_T(\alpha) \geq cT$. Thus, in the worst case, learning from pure trade/no-trade feedback is impossible, in the sense that the average performance of any algorithm remains bounded away from the optimum. This, and, in general, all the proofs in the bandit-feedback chapter, are all our original work, but done applying known techniques to our problem.

- **Bandit setting with bounded density.** We then impose a natural regularity assumption: the joint distribution of (S, B) admits a density bounded from above by a constant. Under this assumption we construct a bandit algorithm that discretizes the price space and runs a suitable bandit strategy on the discretized actions. We show that this algorithm achieves a regret bound of order $T^{2/3}$ (up to constant

factors depending on the density bound), thereby establishing that sublinear regret is attainable under mild smoothness of the valuation distribution.

Taken together, these results delineate a clear picture: in bilateral trade with a trading volume reward, full-feedback permits efficient learning, pure bandit-feedback does not, and bounded-density assumptions restore the possibility of sublinear regret at an intermediate rate.

3

Organization of the Thesis

The remainder of the thesis is organized as follows.

- In Chapter 4, we review basic machine learning paradigms, unsupervised learning, supervised learning, and then we introduce a general online learning framework
- In Chapter 5, we discuss regret and its interpretation as a notion of learning.
- In Chapter 6, we present the classical bilateral trade problem, discuss objectives such as gains from trade and trading volume, and survey relevant literature from mechanism design and online learning.
- In Chapter 7, we formalize bilateral trade as an online learning problem, define the outcome and decision spaces, the reward function, the feedback models, and the regret notion that will be used throughout the thesis.
- Chapter 8 is devoted to the full-feedback setting. We present our learning algorithm, prove an upper bound on its regret, and establish a matching lower bound.
- Chapter 9 studies the bandit-feedback setting. We first establish an impossibility result showing that, without further structure, any algorithm suffers linear regret in the worst case. We then introduce the bounded-density assumption and design an algorithm that achieves sublinear regret of order $T^{2/3}$.
- Finally, Chapter 10 summarizes our findings and outlines directions for future work, including possible extensions to other rewards and richer market models.

4

Machine Learning

In this section, we provide a brief review of the main paradigms of machine learning (ML) before formalizing the notion of regret used in this work.

4.1 Classical Learning Paradigms

Machine learning is the study of algorithms that improve their performance at a task through experience [13]. We say that a learner is able to *learn* if, given enough *experience*, its performance increases. Different learning scenarios are usually classified according to the type of feedback available to the learner and the way data are presented over time. In this section, we briefly review the classical *batch* learning paradigms (unsupervised and supervised learning). A third paradigm, *online learning*, is sufficiently central to this thesis that we devote a separate section (Section 4.2) to it.

Unsupervised learning. Let $\mathcal{X}$ be a set of *data points*, called *input space*, $\mathcal{H}$ a set of *hypotheses*, called the *hypothesis class*, and a function $\ell\colon \mathcal{H} \times \mathcal{X} \to [0, +\infty]$ called the *loss function*. In unsupervised learning, the learner, represented by an *algorithm*, observes an i.i.d. sequence of random variables $X_1, \ldots, X_n$, taking values in $\mathcal{X}$, and seeks to discover a hidden structure. The specific structure is encoded by the loss function ℓ. For example, ℓ might be the within-cluster-sum-of-squares for a cluster algorithm, i.e., an algorithm that seeks to uncover clusters. Denoting $\mathcal{X}^* \coloneqq \bigcup_{n\in\mathbb{N}} \mathcal{X}^n$ as the set of all finite sequences of elements in $\mathcal{X}$, an algorithm is defined as a function $\mathcal{A}\colon \mathcal{X}^* \to \mathcal{H}$. Performance is evaluated through the *expected loss*, called *statistical risk*, defined for any $n \in \mathbb{N}$ and any i.i.d. sequence

of random variables $X_1, X_2, \ldots, X_n$ distributed according to θ and taking values in $\mathcal{X}$, by

$$\mathbb{E}[\ell(\mathcal{A}(X_1, \ldots, X_n), X) \mid X_1, \ldots, X_n]$$

where X is a random variable also distributed according to θ. In this setting, it is customary to call X the *test point*, n the *sample size*, $X_1, \ldots, X_n$ are called *samples*, and the sequence $(X_1, \ldots, X_n)$ is called the *training set*. Canonical tasks include clustering, dimensionality reduction, and density estimation.

Supervised learning. Let $\mathcal{X}$ be an input space, $\mathcal{Y}$ a set of *labels*, called *label set*, and a loss function $\ell \colon \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to [0, +\infty]$. In supervised learning, the learner observes an i.i.d. sequence of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in $\mathcal{X} \times \mathcal{Y}$ and aims to produce a predictor, namely the label Y from a new data point X. In this case, the algorithm is a function $\mathcal{A} \colon (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$, mapping the training set to a hypothesis. Performance is, again, evaluated through the expected loss defined, for any $n \in \mathbb{N}$ and any i.i.d. sequence of random variables $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y}$, by

$$\mathbb{E}[\ell(\mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n)), (X, Y)) \mid (X_1, Y_1), \ldots, (X_n, Y_n)],$$

where the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are called *examples*, and the sequence $((X_1, Y_1), \ldots, (X_n, Y_n))$ is called the *training set*. Typical tasks include regression and classification. The training data are fully labeled: each input is accompanied by its correct output.

Both supervised and unsupervised learning are typically studied as *offline* (or *batch*)

learning paradigms: the learner is given a fixed data set in advance, and the learning procedure does not collect additional data through interaction. We now turn to the *online learning* paradigm, where data are revealed sequentially and the learner's actions may influence the information observed.

4.2 Online Learning: General Framework

Informally, the goal of online learning is to make a sequence of good decisions in an environment that is only partially known at the beginning of the process. Information about the environment is revealed gradually through interaction: the learner chooses an action, observes some feedback, and then uses this feedback to improve future decisions. Different online learning models arise depending on the nature of the data presented to the learner (stochastic or adversarial, independent or dependent over time) and on the extent of the feedback (full-information or bandit-feedback). Throughout this thesis, we focus on the i.i.d. stochastic case, and we consider both full and partial (bandit) feedback. A more detailed and problem-specific description will be given in later sections; in this subsection, we present a general abstract framework.

We define the following basic objects.

Definition 1 (Outcome space)**.** *Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, called the outcome space, where $\mathcal{X}$ is a set and $\mathcal{F}$ is a σ-algebra of measurable subsets of $\mathcal{X}$. At each round t, the environment generates an outcome $X_t \in \mathcal{X}$, modeled as an $\mathcal{F}$-measurable random variable,*

which represents the side information, signals, or state of the world revealed at that round.

Definition 2 (Decision space)**.** *Let $(\mathcal{H}, \mathcal{G})$ be a measurable space, called the decision space (also referred to as an action set or hypothesis class), where $\mathcal{H}$ is a set and $\mathcal{G}$ is a σ-algebra on $\mathcal{H}$. At each round t, the learner selects a decision $h_t \in \mathcal{H}$, modeled as a $\mathcal{G}$-measurable random element when the algorithm is randomized.*

Definition 3 (Instance / environment class)**.** *Let Θ be a set of instances (or environments). Each instance $\theta \in \Theta$ specifies a probability distribution on $(\mathcal{X}, \mathcal{F})$. Given θ, the environment generates an i.i.d. sequence $(X_t)_{t\geq 1}$ with $X_t \sim \theta$.*

Definition 4 (Reward function)**.** *Let $(\mathcal{H}, \mathcal{G})$ and $(\mathcal{X}, \mathcal{F})$ be measurable spaces. A reward function is a map*

$$v : \mathcal{H} \times \mathcal{X} \to [0, +\infty]$$

that is measurable from $(\mathcal{H} \times \mathcal{X}, \mathcal{G} \otimes \mathcal{F})$ to $([0, +\infty], \mathcal{B}([0, +\infty]))$. If the learner plays $h_t \in \mathcal{H}$ and the outcome is $X_t \in \mathcal{X}$, then the learner receives reward $v(h_t, X_t)$.

Throughout this thesis, we adopt a reward-maximization viewpoint and define regret in terms of rewards. If one prefers a loss-minimization convention, and if rewards are uniformly bounded by c, one can equivalently define $\ell(h, x) \coloneqq c - v(h, x)$; this transformation does not change the regret rates up to an additive constant.

Feedback models. After choosing h_t and the environment realizing X_t, the learner observes some feedback Z_t that depends on (h_t, X_t).

- In the *full-feedback* (or full-information) setting the learner observes the entire outcome, so $Z_t = X_t$; from this, the learner can compute $v(h, X_t)$ for any hypothetical action $h \in \mathcal{H}$.
- In the *bandit-feedback* setting the learner observes only its own reward, so $Z_t = v(h_t, X_t)$; in particular, it does not observe the full outcome X_t and cannot directly evaluate the rewards of actions it did not take.

Both settings fit into the same abstract framework; the only difference is the information contained in the feedback variable Z_t. In this thesis, we work in an online learning setting, where feedback is given at each round and the distribution of each Z_t can change with time, based on past decisions.

Online learning algorithms. A *deterministic online learning algorithm* is a sequence of measurable functions, sometimes called *decision rules*, $(\mathcal{D}_t)_{t\geq 1}$ such that

$$\mathcal{D}_1 \in \mathcal{H}, \qquad \mathcal{D}_t : (\mathcal{H} \times \mathcal{Z})^{t-1} \to \mathcal{H} \quad \text{for } t \geq 2,$$

where $\mathcal{Z}$ is the space in which the feedback Z_t takes values. At round t, the algorithm observes the past history

$$H_{t-1} \coloneqq (h_1, Z_1, \ldots, h_{t-1}, Z_{t-1}) \in (\mathcal{H} \times \mathcal{Z})^{t-1},$$

chooses $h_t = \mathcal{D}_t(H_{t-1})$, the environment then draws according to some distribution $X_t \sim \theta$, the learner receives reward $v(h_t, X_t)$, and observes feedback Z_t.

The algorithm may also be *randomized*. In this case, the algorithm has access to internal randomness, modeled by auxiliary random variables, and all expectations below are taken with respect to both the environment randomness and the algorithm's internal randomness.

More formally, a *randomized online learning algorithm* is a sequence of measurable functions $(\mathcal{A}_t)_{t\geq 1}$ such that

$$\mathcal{A}_1 : [0,1] \to \mathcal{H}, \qquad \mathcal{A}_t : (\mathcal{H} \times \mathcal{Z})^{t-1} \times [0,1] \to \mathcal{H} \quad \text{for } t \geq 2.$$

Let $(U_t)_{t\geq 1}$ be a sequence of independent random variables, independent of the environment, with $U_t \sim \text{Unif}[0,1]$. At round t, after observing the past history

$$H_{t-1} := (h_1, Z_1, \ldots, h_{t-1}, Z_{t-1}) \in (\mathcal{H} \times \mathcal{Z})^{t-1},$$

the learner plays

$$h_t = \mathcal{A}_t(H_{t-1}, U_t).$$

The environment then draws $X_t \sim \theta$, the learner receives reward $v(h_t, X_t)$, and observes feedback Z_t as above. Expectations taken under θ will always be with respect to both θ, which governs the random variables X_t, and the uniform distribution U_t, which governs the internal randomness of the algorithm (we omit making the latter explicit in our notation).

Regret. Fix an instance $\theta \in \Theta$ and a time horizon $T \in \mathbb{N}$. The performance of an online algorithm $\mathcal{A} \coloneqq (\mathcal{A}_t)_{t \geq 1}$ is measured by its *regret*, defined as the difference between the expected cumulative reward of the best fixed action $h \in \mathcal{H}$ in hindsight and the expected cumulative reward obtained by the algorithm:

$$R_T(\mathcal{A}, \theta) \coloneqq \sup_{h \in \mathcal{H}} \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h, X_t) \right] - \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h_t, X_t) \right], \tag{4.1}$$

where $h_t = \mathcal{A}_t(\mathcal{H}_{t-1})$ is the action played by the algorithm at time t, and $\mathbb{E}_\theta[\cdot]$ denotes expectation with respect to the product probability measure induced by the environment, namely $\theta^{\otimes T}$ on $(X_1, \ldots, X_T)$, and, if applicable, the independent internal randomness of the algorithm.

5

Regret and the Notion of Learning

In this section, we briefly recall the notion of regret in online learning and explain why obtaining *sublinear* regret, in this framework, is the minimum benchmark in order to consider that an algorithm is learning. Throughout, we use the same notation as in Section 4.2: $\mathcal{H}$ is the decision (action) space, $\mathcal{X}$ is the outcome space, Θ is a set of instances (distributions on $\mathcal{X}$), and $v\colon \mathcal{H} \times \mathcal{X} \to [0, +\infty]$ is the reward function.

5.1 Online Decision Problems and Static Benchmarks

Consider a generic online decision problem. Time is discrete and indexed by $t = 1, 2, \ldots, T$. Fix an instance $\theta \in \Theta$. At each round t:

1. The environment produces an outcome X_t in the outcome space $\mathcal{X}$ (in our stochastic model, X_t is drawn from θ, independently across t).
2. The learner, using an online learning algorithm $\mathcal{A} = (\mathcal{A}_t)_{t \geq 1}$ and the past history, chooses an action $h_t \in \mathcal{H}$ (possibly in a randomized way).
3. The learner receives a reward $v(h_t, X_t) \in [0, +\infty]$.

Let us write the realized cumulative reward of the learner up to time T as

$$G_T(\mathcal{A}, \theta) \coloneqq \sum_{t=1}^{T} v(h_t, X_t),$$

where h_t is the action chosen by $\mathcal{A}$ at time t (possibly depending on internal random seeds). To assess performance, we compare the learner to a simple benchmark policy. In this

exposition, we focus on the standard *static* benchmark: the best fixed action $h \in \mathcal{H}$ that is chosen once and then played at all rounds. Its cumulative reward on the same sequence $(X_t)_{t=1}^T$ is

$$G_T(h, \theta) \coloneqq \sum_{t=1}^{T} v(h, X_t).$$

Definition 5 (Regret)**.** *Let $\mathcal{A}$ denote an online learning algorithm, possibly randomized, and let $\theta \in \Theta$ be a problem instance. Let $X_t \in \mathcal{X}$ denote the outcome generated by the environment at round t, and let $h_t \in \mathcal{H}$ denote the action chosen by $\mathcal{A}$ before observing X_t. The regret of $\mathcal{A}$ on instance θ up to time T is*

$$R_T(\mathcal{A}, \theta) := \sup_{h \in \mathcal{H}} \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h, X_t) \right] - \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h_t, X_t) \right], \tag{5.1}$$

where $\mathbb{E}_\theta$ denotes expectation with respect to the randomness of the environment generated according to θ and the internal randomness of $\mathcal{A}$, if any.

By construction, the regret is always nonnegative: indeed, the benchmark term is the supremum, over all fixed actions $h \in \mathcal{H}$, of their expected cumulative reward, and is therefore at least as large as the expected cumulative reward achieved by the learner.

5.2 Average Performance and Sublinear Regret

Regret is a *cumulative* quantity. To interpret it, it is natural to divide by the time horizon and look at average rewards. Fix an instance θ and an algorithm $\mathcal{A}$, and consider the

quantities

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right], \qquad \frac{1}{T}\sup_{h\in\mathcal{H}}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h, X_t)\right].$$

This will help us define the meaning of *sublinear regret* and show how it is related to the concept of learning.

Definition 6 (Sublinear regret)**.** *An algorithm $\mathcal{A}$ has* sublinear regret *on a class of instances Θ if*

$$\sup_{\theta\in\Theta} R_T(\mathcal{A}, \theta) = o(T) \qquad \text{as } T \to \infty.$$

Equivalently, for every $\varepsilon > 0$ there exists $T_\varepsilon \in \mathbb{N}$ such that for all $T \geq T_\varepsilon$,

$$\sup_{\theta\in\Theta} \frac{R_T(\mathcal{A}, \theta)}{T} \leq \varepsilon.$$

Combining this with the definition of regret yields the basic connection between regret and learning.

Proposition 1 (Sublinear regret $\Rightarrow$ asymptotic optimality)**.** *Assume that for each $\theta \in \Theta$,*

$$\frac{1}{T}\sup_{h\in\mathcal{H}}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h, X_t)\right] \xrightarrow[T\to\infty]{} \gamma(\theta)$$

for some value $\gamma(\theta) \in \mathbb{R}$ (the optimal asymptotic average reward for that instance). If $\mathcal{A}$ enjoys sublinear regret on Θ, then for every $\theta \in \Theta$,

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right] \xrightarrow[T\to\infty]{} \gamma(\theta).$$

In words: the learner's average reward converges to the benchmark average reward.

Proof. Fix $\theta \in \Theta$. Leveraging the linearity of expectation, we can rearrange (5.1) and divide both sides by the time horizon T, obtaining

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right] = \frac{1}{T}\sup_{h\in\mathcal{H}} \mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h, X_t)\right] - \frac{R_T(\mathcal{A}, \theta)}{T}.$$

By hypothesis the first term converges to $\gamma(\theta)$ as $T \to \infty$, while sublinear regret implies $R_T(\mathcal{A}, \theta)/T \to 0$. Therefore the right-hand side converges to $\gamma(\theta)$, which proves the claim. □

Thus, *sublinear regret exactly captures the requirement that the learner asymptotically matches the performance of the benchmark*, at least in terms of expected time-averaged reward.

5.3 Why Linear Regret Means "No Learning"

Suppose instead that the regret grows *linearly* in T. That is, assume that there exist a constant $c > 0$ and an instance $\theta \in \Theta$ such that for infinitely many T,

$$R_T(\mathcal{A}, \theta) \geq cT.$$

Dividing (5.1) by T and using this lower bound gives

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right] \le \frac{1}{T}\sup_{h\in\mathcal{H}} \mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h, X_t)\right] - c.$$

Even if the benchmark average reward converges, the learner's average reward is asymptotically at least c *below* that value. In other words, the algorithm stabilizes at a uniformly suboptimal performance level: it never learns to match the benchmark.

This gap is especially clear in the i.i.d. stochastic setting. Assume $(X_t)_{t\ge1}$ is i.i.d. with law θ, and define the expected reward of a fixed action $h \in \mathcal{H}$ by

$$\mu(h) \coloneqq \mathbb{E}_\theta[v(h, X_1)], \qquad \bar{\mu} \coloneqq \sup_{h\in\mathcal{H}} \mu(h).$$

Note that, for every T,

$$\sup_{h\in\mathcal{H}} \mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h, X_t)\right] = \sup_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{E}_\theta[v(h, X_t)] = T\sup_{h\in\mathcal{H}} \mu(h) = T\bar{\mu},$$

where we used linearity of expectation and the i.i.d. property. Therefore, the regret identity becomes

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right] = \bar{\mu} - \frac{R_T(\mathcal{A}, \theta)}{T}.$$

In particular, sublinear regret $R_T(\mathcal{A}, \theta) = o(T)$ is *equivalent* to

$$\frac{1}{T}\mathbb{E}_\theta\left[\sum_{t=1}^{T} v(h_t, X_t)\right] \xrightarrow[T\to\infty]{} \bar{\mu},$$

while linear regret $R_T(\mathcal{A}, \theta) \geq cT$ for infinitely many T implies

$$\liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h_t, X_t)\right] \leq \bar{\mu} - c.$$

Hence:

- Sublinear regret means that the average per-round expected performance of the learner converges to the benchmark value. In this sense, the asymptotic price of not knowing the instance vanishes.
- Linear regret means that the average expected gap with respect to the benchmark does not vanish. Therefore, the learner fails to asymptotically match the benchmark performance.

5.4 Regret as a Learning Criterion

There are several reasons why minimizing regret is a typical goal for learning agents in sequential decision problems:

1. **Model-agnostic learner.** The learner is not assumed to know the underlying distribution θ. The regret $R_T(\mathcal{A}, \theta)$ is evaluated under the true environment θ, but the algorithm must choose its actions online using only past observations and, possibly, its own internal randomness. Thus, small regret means that the learner

performs nearly as well as the best fixed action in hindsight, despite not knowing the environment in advance.

2. **Comparison to the best fixed action.** The regret benchmark is the expected cumulative reward of the best fixed action in $\mathcal{H}$ under the true environment θ. This benchmark may be viewed as an idealized comparator that knows which fixed action is optimal for θ, whereas the learner must discover good actions online from the observed feedback. Achieving regret $o(T)$, therefore, means that the learner's average reward asymptotically matches that of the best fixed action, despite not knowing the environment in advance.

3. **Asymptotic optimality.** As shown above, sublinear regret is equivalent to asymptotically matching the average reward of the benchmark appearing in the regret definition. In the present stochastic setting, this benchmark is the best fixed action under the true instance θ, namely

$$\sup_{h \in \mathcal{H}} \mathbb{E}_\theta \left[\sum_{t=1}^{T} v(h, X_t) \right].$$

Thus, regret measures the cumulative price paid for not knowing in advance which fixed action is best for the environment. Sublinear regret means that this price, when averaged over time, vanishes.

4. **Optimal rates and lower bounds.** Regret is also the standard quantity for

impossibility results. For a class of instances Θ, one can define the minimax regret

$$\mathfrak{R}_T(\Theta) := \inf_{\mathcal{A}} \sup_{\theta \in \Theta} R_T(\mathcal{A}, \theta),$$

where the infimum is over all admissible online learning algorithms. A lower bound of the form $\mathfrak{R}_T(\Theta) \geq cf(T)$ shows that every algorithm suffers regret at least of order $f(T)$ on some instance in Θ. An upper bound of the form $\mathfrak{R}_T(\Theta) \leq Cf(T)$ shows that some algorithm attains this rate uniformly over Θ. When matching upper and lower bounds are available, they identify the optimal order of regret for the learning problem, up to constants or logarithmic factors.

For these reasons, regret is the standard performance criterion in online learning and sequential decision problems. A basic requirement for an online learning algorithm is to achieve sublinear regret, which guarantees that its expected average reward approaches that of the best fixed action under the true environment. Learning is then quantified by how fast this gap vanishes as a function of the horizon and other problem parameters. When upper and lower bounds match, they characterize the optimal achievable rate of learning for the class of instances under consideration.

6

The Bilateral Trade Problem

The classical bilateral trade problem is a model from mechanism design describing a market with a single seller and a single buyer who may trade an indivisible good. The seller has a private valuation S, representing the minimum price at which she is willing to sell, and the buyer has a private valuation B, representing the maximum price he is willing to pay. For a posted price p, a trade occurs precisely when $S \leq p \leq B$. Thus, the agents' valuations determine whether a proposed price leads to a successful trade. More generally, a mechanism specifies, as a function of the agents' valuations or reported valuations, whether a trade takes place and at what price.

In this thesis, we study an online-learning version of this problem. At each round t, a new buyer–seller pair arrives with valuations (S_t, B_t), drawn from an unknown distribution. The platform chooses a price P_t, observes feedback from the interaction, and uses past observations to improve future pricing decisions. In the basic trade-indicator version of the problem, the reward is

$$v(P_t, S_t, B_t) = \mathbb{I}\{S_t \leq P_t \leq B_t\},$$

so the platform is rewarded exactly when the posted price induces a trade.

This simple setup captures many real-world platforms that intermediate trades between two sides of a market. For example, in ride-sharing platforms such as Uber or Lyft, drivers play the role of sellers, riders play the role of buyers, and the platform chooses a price, or fare, at which a ride is offered. Similar considerations arise in online marketplaces, ad exchanges, and other two-sided platforms.

6.1 Classical Mechanism-Design Viewpoint

In the classical mechanism-design formulation, the seller's and buyer's valuations are modeled as random variables drawn from commonly known prior distributions, and agents may strategically misreport those valuations. A central goal is to design mechanisms that are *efficient* and *incentive compatible*, while also respecting participation and budget constraints. Formally, one typically asks whether there exists a mechanism that simultaneously satisfies:

- **Incentive compatibility** (IC): Truthful reporting is a dominant or Bayes–Nash equilibrium strategy for each agent;
- **Individual rationality** (IR): Each agent weakly prefers participating in the mechanism to opting out;
- **Ex post efficiency**: A trade occurs whenever it is socially beneficial, i.e., whenever $B \geq S$;
- **Budget balance** (BB): The mechanism does not require external subsidies (and often is required not to make a profit either).

The seminal Myerson–Satterthwaite theorem [1] shows that, in general, *no* mechanism can satisfy IC, IR, ex post efficiency, and BB simultaneously. This impossibility result has driven a vast literature on bilateral trade and related models, exploring relaxations and alternative objectives.

Over time, a variety of *variants* of the bilateral trade problem have been studied, including:

- **Bayesian mechanism design**, where one optimizes expected social welfare or revenue given known priors over (S, B) [2, 3];
- **Prior-free or prior-independent models**, which seek mechanisms that perform well over broad classes of distributions (see, e.g., [14]);
- **Posted-price mechanisms**, in which the mechanism offers a take-it-or-leave-it price to both agents [5];
- **Online and learning-based formulations**, where buyer–seller pairs arrive sequentially and the mechanism adapts its prices over time, often under limited feedback such as bandit observations [4, 6, 7];

6.2 Rewards: Gain From Trade, Trading Volume, and Fairness

Several performance criteria have been proposed for bilateral trade mechanisms.

Gain from trade. A natural welfare-oriented reward is the *gain from trade* (GFT), i.e., the increase in total surplus generated by the mechanism. When a trade occurs at price p,

the buyer's utility is $B - p$, the seller's utility is $p - S$, and the total gain from trade is

$$\mathrm{GFT}(p, S, B) \coloneqq (B - p)\mathbb{I}\{p \le B\} + (p - S)\mathbb{I}\{S \le p\} = (B - S)\mathbb{I}\{S \le p \le B\}.$$

Mechanisms that approximately maximize expected gains from trade have been studied under various informational assumptions and feedback models; see, for instance, [9], [12], and references therein.

Fairness-oriented rewards. More recently, fairness considerations have motivated alternative reward functions. For example, [8] introduces the fair gain from trade, defined as

$$\mathrm{FGFT}(p, S, B) = \min\{B - p, p - S\}\mathbb{I}\{S \le p \le B\},$$

which rewards the mechanism according to the smaller of the buyer's and seller's utilities. This corresponds to a Rawlsian, or max-min, view of fairness, in which the value of an allocation is determined by the utility of the worse-off participant [15]. This penalizes highly unbalanced allocations where almost all surplus accrues to one side.

Trading volume. In many applications, the platform is primarily interested in the *trading volume*, i.e., whether a trade occurs or not, rather than how the surplus is split between buyer and seller. This leads to the simpler objective

$$v(p, S, B) \coloneqq \mathbb{I}\{S \le p \le B\},$$

which counts 1 whenever a transaction is executed and 0 otherwise. Volume is a relevant metric when the platform charges a transaction fee or uses trade frequency as a proxy for liquidity. Volume-based rewards have been considered both in static mechanism design and in online learning variants of bilateral trade; see, e.g., [16].

In this thesis, we focus on the latter, and we study the extent to which a learning agent can learn to post effective prices from repeated interactions.

6.3 An Online Learning View

We now adopt an online learning perspective on the repeated interaction between a posted-price platform and a sequence of buyer–seller pairs. At each round $t = 1, 2, \ldots, T$:

1. A new pair of valuations $(S_t, B_t) \in [0, 1]^2$ is drawn from an unknown joint distribution on $[0, 1]^2$.

2. The platform, having observed the history up to round $t-1$, chooses a price $P_t \in [0, 1]$.

3. A trade occurs if and only if $S_t \leq P_t \leq B_t$. The platform then observes some feedback and receives reward $v(P_t, S_t, B_t) \in \{0, 1\}$ if the objective is trading volume (or the corresponding value of GFT or FGFT for other objectives).

This interaction can be viewed as a repeated sequence of one-period, take-it-or-leave-it mechanisms. At each round, the platform commits to a price, and the current buyer and seller decide whether to accept or reject, given their private valuations. Because the

distribution θ is fixed across rounds, observations from past interactions can be used to improve future pricing decisions. The platform's goal is to choose prices so as to maximize cumulative reward over the horizon T.

Feedback models. The information revealed after each round crucially affects the difficulty of the learning problem. We consider two canonical feedback models:

- **Full-feedback**: after posting P_t, the platform observes the full valuation pair (S_t, B_t). In this case the learner can, in principle, reconstruct the reward that would have been obtained by any hypothetical price $p \in [0, 1]$.
- **Bandit-feedback**: the platform observes only whether a trade occurred, i.e., the binary variable $v(P_t, S_t, B_t) \in \{0, 1\}$. The actual valuations (S_t, B_t) remain hidden, making the problem substantially more challenging.

Both settings fit into the general online learning framework described in Section 5; the main difference lies in how informative the feedback is for updating future prices.

6.4 Regret for Online Posted Pricing

Let (S, B) denote a generic valuation pair with the same distribution θ as each (S_t, B_t), and let

$$f_\theta(p) \coloneqq \mathbb{E}_\theta[v(p, S, B)]$$

be the expected trading volume obtained by posting a fixed price $p \in [0,1]$. The best fixed-price benchmark under θ is

$$f_\theta^\star \coloneqq \sup_{p\in[0,1]} f_\theta(p),$$

and its expected cumulative trading volume over T rounds is $Tf_\theta^\star$. This benchmark corresponds to a comparator that knows the distribution θ and chooses the best fixed price in advance; it is not a realized-sequence hindsight benchmark.

An online pricing algorithm α produces a possibly randomized sequence of prices $P_1, P_2, \ldots, P_T$ based on past feedback. Following the general regret notion in Section 5, we measure its performance on instance θ by the *expected regret*

$$R_T(\alpha,\theta) \coloneqq Tf_\theta^\star - \mathbb{E}_{\theta,\alpha}\left[\sum_{t=1}^{T} v(P_t, S_t, B_t)\right],$$

where $\mathbb{E}_{\theta,\alpha}$ denotes expectation with respect to the joint randomness induced by the valuation pairs $(S_t, B_t)_{t=1}^T$ drawn i.i.d. from θ, the feedback observed by the platform, and the internal randomness of the algorithm α, if any. Regret thus quantifies the loss in trading volume incurred by acting online and learning from partial information, as compared to an oracle that knows the valuation distribution and can commit in advance, obtaining the optimal value $f^\star$.

In summary, the bilateral trade problem provides a clean and practically motivated setting in which ideas from mechanism design, online learning, and bandit algorithms

interact. The regret framework allows us to quantify exactly when a posted-price platform can (or cannot) learn effective pricing strategies from limited feedback.

7

The Bilateral Trade Setting

In this section, we introduce the learning protocol for the sequential bilateral trade problem (see Learning Protocol 1). The reward obtained from a trade corresponds to the *trading volume* (1, if there is a trade, 0, otherwise), which is formally defined for all $p, s, b \in [0,1]$, by

$$v(p,s,b) \coloneqq \mathbb{I}\{s \le p \le b\}$$

Learning Protocol 1: Bilateral Trade

for t = 1,2,... **do**
a new seller/buyer pair arrives with (hidden) valuations $(S_t, B_t) \in [0,1]^2$;
the learner posts a price $P_t \in [0,1]$;
the learner receives a (hidden) reward $v(P_t, S_t, B_t)$;
feedback $Z_t \in \mathcal{Z}$ is revealed (where $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ is some known measurable space; see below);
end for

At each round t, a seller and a buyer arrive, each with private valuations: the seller's value is $S_t \in [0,1]$, and the buyer's is $B_t \in [0,1]$. The learner selects a price $P_t \in [0,1]$, and a trade takes place if and only if $S_t \le P_t \le B_t$. When this happens, the learner receives a reward $v(P_t, S_t, B_t)$ but, instead of observing it, the learner only observes the feedback Z_t.

The learner's objective is to design an algorithm α generating the prices $P_1, P_2, \ldots$ (as specified in Learning Protocol 1)[1] achieving sublinear *regret*

$$R_t(\alpha) \coloneqq \max_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} v(p, S_t, B_t) - \sum_{t=1}^{T} v(P_t, S_t, B_t)\right],$$

[1]If α is deterministic, then $P_1 \coloneqq \alpha_1$ and, for all $t \ge 2$, $P_t \coloneqq \alpha_t(Z_1, \ldots, Z_{t-1})$. If α is randomized, then $P_1 \coloneqq \alpha_1(U_1)$ and, for all $t \ge 2$, $P_t \coloneqq \alpha_t((Z_1, \ldots, Z_{t-1}), U_t)$.

where the expectation is taken over the sequence of valuations and, if applicable, the randomness $U_1, \ldots, U_t$ in the learner's strategy α. For simplicity, we denote by $p^\star$ any point in $[0, 1]$ that maximizes the expected cumulative trading volume in the expression above. The existence of such a $p^\star$ is guaranteed (a formal proof is provided in Appendix A). We next describe several problem variants, each characterized by different types of feedback and assumptions about the environment.

7.1 Feedback

The two specific models we're going to analyse in this thesis are the following:

Full-feedback: the feedback Z_t received at time t is the entire seller/buyer pair (S_t, B_t); in this setting, the seller and the buyer reveal their valuations at the end of a trade.

Bandit-feedback: the feedback Z_t received at time t is just the reward $v(P_t, S_t, B_t) = \mathbb{I}\{S_t \leq P_t \leq B_t\}$; in this setting, the seller and the buyer reveal nothing.

7.2 Environment

We analyze the problem under the following stochastic assumptions.

Stochastic (i.i.d.): $(S_1, B_1), (S_2, B_2), \ldots$ is an i.i.d. sequence of seller/buyer pairs, while S_t and B_t could be dependent. We will also investigate the (iid) setting under the following further assumption:

Bounded density (bd): (S_t, B_t) admits a joint density bounded by some constant M.

8

Full-Feedback Stochastic (i.i.d.)

Setting

We begin by examining the full-feedback setting. Let

$$X_t \coloneqq (S_t, B_t) \in [0,1]^2$$

denote the valuation pair observed at round t. We assume that $X_1, X_2, \ldots$ are i.i.d. with common unknown distribution θ on $[0,1]^2$. This i.i.d. assumption is across rounds only: within a given pair, the seller's valuation S_t and the buyer's valuation B_t need not be independent.

In this setup, sellers and buyers disclose their actual valuations at the end of each round, which allows the learner to access complete information about each transaction. Since the posted prices are determined solely based on prior observations and not on current declarations, agents have no incentive to misreport their values.

In Section 8.1, we show that a *Follow the Leader* approach, which we call Follow the Best Price (FBP, Algorithm 1) achieves a regret upper bound of $\mathcal{O}(\sqrt{T})$. In Section 8.2, we present a matching lower bound rate of $\Omega(\sqrt{T})$.

Algorithm 1 Follow the Best Price (FBP)

initialization: let $P_1 \leftarrow 1/2$;
for t = 1,2,... **do**
  post price P_t;
  receive feedback (S_t, B_t);
  pick $P_{t+1} \in \operatorname{argmax}_{p\in[0,1]} \frac{1}{t}\sum_{i=1}^{t} v(p, S_i, B_i)$;
end for

Algorithm 1 can be interpreted as a continuous-action analogue of the classical Follow

the Leader algorithm. In the usual full-information experts setting, the learner selects one expert at each round, observes the reward of all experts, and then, at the next round, chooses an expert with the highest empirical average reward so far. The same idea applies here. The difference is that the action set is the whole interval $[0, 1]$, so there are uncountably many possible prices rather than finitely many experts. However, in the full-feedback model, once the learner observes the valuation pair (S_t, B_t), it can compute the reward $v(p, S_t, B_t)$ for every hypothetical price $p \in [0, 1]$. Therefore, after t rounds, the learner can compare all prices through their empirical average rewards and choose any price maximizing this quantity. In this sense, Follow the Best Price is simply Follow the Leader applied to a continuous action space under full feedback.

8.1 Follow the Best Price (FBP)

We now present the Follow the Best Price (FBP) algorithm. Its core idea is to repeatedly select the price that has performed best based on observed outcomes up to the current time. Importantly, this method does not require any prior knowledge about the time horizon T. For the trading-volume reward

$$v(p, S, B) = \mathbb{I}\{S \leq p \leq B\},$$

define the empirical trading-volume function after t observations by

$$\widehat{f}_t(p) \coloneqq \frac{1}{t}\sum_{i=1}^{t} v(p, S_i, B_i).$$

Set $P_1 = \frac{1}{2}$. At each round $t \geq 1$, once the learner has observed the valuation pairs $(S_1, B_1), \ldots, (S_t, B_t)$, it chooses

$$P_{t+1} \in \operatorname*{argmax}_{p\in[0,1]} \widehat{f}_t(p).$$

To clarify the existence of this price P_{t+1}, define the finite candidate set

$$\mathcal{L}_t \coloneqq \{0, 1\} \cup \{S_i, B_i : 1 \leq i \leq t\}.$$

For fixed observations $(S_1, B_1), \ldots, (S_t, B_t)$, the function $p \mapsto \widehat{f}_t(p)$ can change only at one of the points in $\mathcal{L}_t$. Moreover, if p lies strictly between two consecutive points of $\mathcal{L}_t$, then moving p to the left endpoint of that interval cannot decrease the value of $\widehat{f}_t(p)$. Therefore,

$$\sup_{p\in[0,1]} \widehat{f}_t(p) = \max_{p\in\mathcal{L}_t} \widehat{f}_t(p).$$

Let

$$L_{t,0} < L_{t,1} < \cdots < L_{t,m_t}$$

be the elements of $\mathcal{L}_t$ sorted in increasing order. We break ties by choosing the smallest-index empirical maximizer:

$$j_t \coloneqq \min \operatorname*{argmax}_{0 \le j \le m_t} \widehat{f}_t(L_{t,j}), \qquad P_{t+1} \coloneqq L_{t,j_t}.$$

This deterministic tie-breaking rule avoids any ambiguity in the definition of P_{t+1}. In particular, P_{t+1} is a measurable function of the past observations, since it is obtained from finitely many measurable comparisons among the values $\widehat{f}_t(L_{t,j})$. More general measurable-selection approaches are discussed, for example, in [17, Section 2.4]. The main idea of the analysis of Algorithm 1 is to control the uniform deviation

$$\sup_{p \in [0,1]} \left| \widehat{f}_t(p) - f_\theta(p) \right|$$

between the empirical trading-volume function and its expectation. The bivariate DKW inequality gives a distribution-free bound on this deviation, uniformly over all prices $p \in [0,1]$ and all joint distributions θ on $[0,1]^2$. This uniform control is then used to bound the regret of the empirical maximizer.

Theorem 1. *In the full-feedback stochastic (i.i.d.) setting, the regret of Follow the Best Price satisfies, for all horizons T,*

$$R_T(\mathrm{FBP}) \le 1 + c\sqrt{T-1},$$

where $c \in (0, 572132)$ is a universal constant.

Proof. Let's first analyze the case where $T = 1$. In this case, we fix a price $p = \frac{1}{2}$ and, in the worst case (i.e. if $p \notin (S_1, B_1)$) we pay a regret of exactly 1. Assume now that $T \geq 2$. Fix any $t \in [T-1]$. For any $p \in [0,1]$, define the random variable:

$$H_t(p) \coloneqq \frac{1}{t}\sum_{i=1}^{t} v(p, S_i, B_i) - \mathbb{E}[v(p, S_1, B_1)]$$

and the quantity

$$r_t(\text{FBP}) = \mathbb{E}[v(p^*, S_{t+1}, B_{t+1})] - \mathbb{E}[v(P_{t+1}, S_{t+1}, B_{t+1})].$$

By the definition of P_{t+1} and its independence from (S_{t+1}, B_{t+1}), we obtain:

$$\begin{aligned} r_t(\text{FBP}) &\leq \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} v(P_{t+1}, S_i, B_i)\right] - \mathbb{E}[v(P_{t+1}, S_{t+1}, B_{t+1})] \\ &= \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} v(P_{t+1}, S_i, B_i) - \mathbb{E}[v(P_{t+1}, S_{t+1}, B_{t+1}) \mid P_{t+1}]\right] \qquad \text{(Tower property)} \end{aligned}$$

Now we can use the Freezing Lemma (a formal statement and proof of both the tower property and the freezing lemma are in Appendix B) to write $\mathbb{E}[v(P_{t+1}, S_{t+1}, B_{t+1}) \mid P_{t+1}] = \mathbb{E}[v(P_{t+1}, S_1, B_1) \mid P_{t+1}]$, and therefore

$$\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} v(P_{t+1}, S_i, B_i) - \mathbb{E}[v(P_{t+1}, S_{t+1}, B_{t+1}) \mid P_{t+1}]\right] = \mathbb{E}[H_t(P_{t+1})]$$

Which means we can write

$$\begin{aligned}
r_t(\text{FBP}) &\leq \mathbb{E}[H_t(P_{t+1})] \\
&\leq \mathbb{E}\left[\sup_{p\in[0,1]} H_t(p)\right] \\
&= \mathbb{E}\left[\sup_{p\in[0,1]} \left(\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq p \leq B_i\} - \mathbb{E}[\mathbb{I}\{S_1 \leq p \leq B_1\}]\right)\right] \\
&= \mathbb{E}\left[\sup_{p\in[0,1]} \left(\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq p \leq B_i\} - \mathbb{P}[S_1 \leq p \leq B_1]\right)\right] \\
&\leq \mathbb{E}\left[\sup_{x\in\mathbb{R}} \left|\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq x \leq B_i\} - \mathbb{P}[S_1 \leq x \leq B_1]\right|\right] =: (*)
\end{aligned}$$

Now, let m_0, c_1, c_2 be the constants from Theorem 7 (for the formal statement refer to Appendix C) and $\varepsilon_t = \sqrt{\frac{m_0}{t}}$. Using the Tail Integration Formula (formal statement and proof in Appendix D), we have that

$$\begin{aligned}
(*) &\leq \int_{[0,+\infty]} \mathbb{P}\left[\sup_{x\in\mathbb{R}} \left|\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq x \leq B_i\} - \mathbb{P}[S_1 \leq x \leq B_1]\right| > \varepsilon\right] d\varepsilon \\
&= \int_{[0,1]} \mathbb{P}\left[\sup_{x\in\mathbb{R}} \left|\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq x \leq B_i\} - \mathbb{P}[S_1 \leq x \leq B_1]\right| > \varepsilon\right] d\varepsilon \\
&\leq \int_{[0,\varepsilon_t]} 1 d\varepsilon + \int_{[\varepsilon_t,1]} \mathbb{P}\left[\sup_{x\in\mathbb{R}} \left|\frac{1}{t}\sum_{i=1}^{t} \mathbb{I}\{S_i \leq x \leq B_i\} - \mathbb{P}[S_1 \leq x \leq B_1]\right| > \varepsilon\right] d\varepsilon
\end{aligned}$$

Applying the bivariate DKW inequality (Theorem 7), we obtain:

$$\begin{aligned} r_t(\text{FBP}) &\leq \varepsilon_t + \int_{\varepsilon_t}^{1} c_1 e^{-c_2 t \varepsilon^2} d\varepsilon \\ &\leq \varepsilon_t + \frac{c_1}{2\sqrt{c_2 t}} \int_0^{\infty} e^{-u} u^{-1/2} du \\ &= \sqrt{\frac{m_0}{t}} + \frac{c_1}{2}\sqrt{\frac{\pi}{c_2 t}} \\ &= \frac{1}{\sqrt{t}} \left(\sqrt{m_0} + \frac{c_1}{2}\sqrt{\frac{\pi}{c_2}} \right) \end{aligned}$$

Since t was arbitrary, using Lemma 9 (statement and proof in Appendix B) and letting

$$c = 2\left(\sqrt{m_0} + \frac{c_1}{2}\sqrt{\frac{\pi}{c_2}} \right) < 572132,$$

we conclude:

$$\begin{aligned} R_T(\text{FBP}) &\leq 1 + \sum_{t=1}^{T-1} r_t(\text{FBP}) \\ &\leq 1 + \frac{c}{2} \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \\ &= 1 + c\sqrt{T-1}. \end{aligned}$$

□

Remark 1. *Theorem 1 shows that FBP attains a regret of order $O(\sqrt{T})$ in the full-feedback stochastic setting. Combined with the matching lower bound of Theorem 2 that we will prove in the next section, this characterizes the minimax regret rate in this model: $\sqrt{T}$ is not*

improvable and FBP is optimal up to constant factors. The constant c in Theorem 1 is very large because it comes from a crude multivariate Dvoretzky–Kiefer–Wolfowitz inequality; sharper empirical-process bounds could reduce this constant, but we make no attempt to optimize it, as our focus is on the dependence on T. Conceptually, the proof highlights that full-feedback allows the platform to uniformly estimate the expected trading volume function over all prices and distributions, and then track the empirically best price over time.

8.2 $\sqrt{T}$ Lower Bound

In this section, we show that the upper bound on the minimax regret we proved in Section 4.1 is tight[1]. No strategy can beat the $O(\sqrt{T})$ rate when the seller/buyer pair (S_t, B_t) is drawn i.i.d. from an unknown fixed distribution, even under the further assumptions that the valuations of the seller and buyer are independent of each other and have bounded density. Before proving the theorem, we need to prove two useful Lemmas. Lemma 1 shows the calculations needed to find the KL divergence of two Bernoulli distributions.

Lemma 1. *Fix $\varepsilon \in (0,1)$. The KL divergence of two Bernoulli distributions of parameters $\frac{1}{2}(1+\varepsilon)$ and $\frac{1}{2}(1-\varepsilon)$ is*

$$\mathcal{D}_{\mathrm{KL}}\left(\mathcal{B}\left(\frac{1}{2}(1+\varepsilon)\right)\middle\|\mathcal{B}\left(\frac{1}{2}(1-\varepsilon)\right)\right) = \varepsilon \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)$$

[1]Here, we mean to say that both upper and lower bounds have the same rate $\sqrt{T}$

Proof.

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\left(\mathcal{B}\left(\frac{1}{2}(1+\varepsilon)\right)\middle\|\mathcal{B}\left(\frac{1}{2}(1-\varepsilon)\right)\right) \\
&\quad= \frac{1}{2}(1+\varepsilon)\log\left(\frac{\frac{1}{2}(1+\varepsilon)}{\frac{1}{2}(1-\varepsilon)}\right) + \left(1-\frac{1}{2}(1+\varepsilon)\right)\log\left(\frac{1-\frac{1}{2}(1+\varepsilon)}{1-\frac{1}{2}(1-\varepsilon)}\right) \\
&\quad= \frac{1}{2}(1+\varepsilon)\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) + \frac{1}{2}(1-\varepsilon)\log\left(\frac{1-\varepsilon}{1+\varepsilon}\right) \\
&\quad= \frac{1}{2}(1+\varepsilon)\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) - \frac{1}{2}(1-\varepsilon)\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \\
&\quad= \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left[\frac{1}{2}(1+\varepsilon) - \frac{1}{2}(1-\varepsilon)\right] \\
&\quad= \varepsilon\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)
\end{aligned}$$

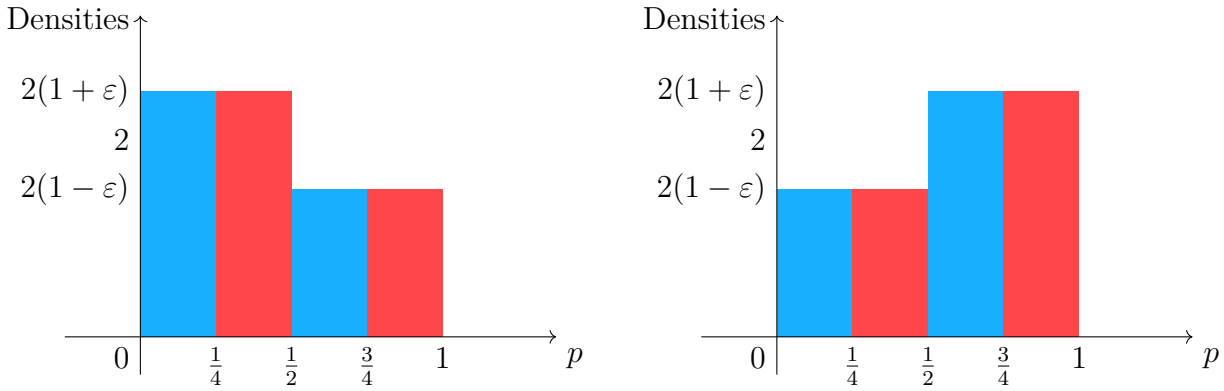
□

The lower bound will follow from choosing between two environments with slightly different optimal prices; Lemma 2 computes the expected trading volume $\mathbb{E}[v(p,S,B)]$ under each environment.

Lemma 2. *Fix $\varepsilon \in [0,1]$. Let $v(p,S,B) = \mathbb{I}\{s \le p \le b\}$ be the trading volume. Define the two functions (see Figure 8.1):*

$$f_{S,\pm\varepsilon}(s) = 2\Big[(1\pm\varepsilon)\mathbb{I}\{0 \le s \le 1/4\} + (1\mp\varepsilon)\mathbb{I}\{1/2 \le s \le 3/4\}\Big],$$

$$f_{B,\pm\varepsilon}(b) = 2\Big[(1\pm\varepsilon)\mathbb{I}\{1/4 \le b \le 1/2\} + (1\mp\varepsilon)\mathbb{I}\{3/4 \le b \le 1\}\Big].$$



(a) Distributions $f_{S,+\varepsilon}$ (blue) and $f_{B,+\varepsilon}$ (red) (b) Distributions $f_{S,-\varepsilon}$ (blue) and $f_{B,-\varepsilon}$ (red)

Figure 8.1: Distributions $f_{S,\pm\varepsilon}$ and $f_{B,\pm\varepsilon}$ that allow the model to be seen as an online learning problem with expert advice.

Then, the expected trading volume is

$$\mathbb{E}_{S,B}\left[v_{\pm\varepsilon}(S,B,p)\right] = \begin{cases} 2p(1\pm\varepsilon), & 0 \le p \le \frac{1}{4}, \\ \frac{1}{4}(1\pm\varepsilon)(3\pm\varepsilon) - p(1\pm\varepsilon)^2, & \frac{1}{4} < p \le \frac{1}{2}, \\ \frac{1}{4}(1-\varepsilon^2) + \left(p-\frac{1}{2}\right)(1\mp\varepsilon)^2, & \frac{1}{2} < p \le \frac{3}{4}, \\ 2(1-p)(1\mp\varepsilon), & \frac{3}{4} < p \le 1. \end{cases} \tag{8.1}$$

Proof. First note that $f_{S,\pm\varepsilon}, f_{B,\pm\varepsilon}$ are densities: clearly $f_{S,\pm\varepsilon}, f_{B,\pm\varepsilon} \ge 0$ if $\varepsilon \le 1$, and they are normalized

$$\int_0^1 f_{S,\pm\varepsilon}(s)ds = 2(1\pm\varepsilon)\Big(\frac{1}{4}-0\Big) + 2(1\mp\varepsilon)\Big(\frac{3}{4}-\frac{1}{2}\Big) = \frac{1\pm\varepsilon}{2} + \frac{1\mp\varepsilon}{2} = 1,$$

and similarly $\int_0^1 f_{B,\pm\varepsilon}(b)db = 1$.

From the definition of the trading volume, using independence and the decomposition $\mathbb{I}\{S \le p \le B\} = \mathbb{I}\{S \le p\} \cdot \mathbb{I}\{p \le B\}$,

$$\begin{aligned}
\mathbb{E}_{S,B}\left[v_{\pm\varepsilon}(S,B,p)\right] &= \Big(\int\int_{[0,1]^2} \mathbb{I}\{s \le p \le b\} \cdot f_{S,\pm\varepsilon}(s) \cdot f_{B,\pm\varepsilon}(b)dsdb\Big) \\
&= \Big(\int\int_{[0,1]^2} \mathbb{I}\{s \le p\} \cdot f_{S,\pm\varepsilon}(s) \cdot \mathbb{I}\{p \le b\} \cdot f_{B,\pm\varepsilon}(b)dsdb\Big) \\
&= \Big(\int_0^p f_{S,\pm\varepsilon}(s)ds\Big) \cdot \Big(\int_p^1 f_{B,\pm\varepsilon}(b)db\Big) \\
&=: F_{S,\pm\varepsilon}(p)F_{B,\pm\varepsilon}(p).
\end{aligned}$$

We now compute $F_{S,\pm\varepsilon}$ and $F_{B,\pm\varepsilon}$.

$$\begin{aligned}
F_{S,\pm\varepsilon}(p) = \mathbb{I}\{0 \le p \le \frac{1}{4}\}\int_0^p 2(1\pm\varepsilon)ds + \mathbb{I}\{\frac{1}{4} \le p \le \frac{1}{2}\}\int_0^{1/4} 2(1\pm\varepsilon)ds + \\
+\mathbb{I}\{\frac{1}{2} \le p \le \frac{3}{4}\}\left[\frac{1\pm\varepsilon}{2} + \int_{1/2}^p 2(1\mp\varepsilon)ds\right] + \mathbb{I}\{\frac{3}{4} \le p \le 1\}\left[\int_0^{1/4} 2(1\pm\varepsilon)ds + \int_{1/2}^{3/4} 2(1\mp\varepsilon)ds\right]
\end{aligned}$$

Which results in

$$F_{S,\pm\varepsilon}(p) = \begin{cases} 2(1 \pm \varepsilon)p & 0 \le p \le \frac{1}{4}, \\ \frac{1 \pm \varepsilon}{2}, & \frac{1}{4} < p \le \frac{1}{2}, \\ \frac{1 \pm \varepsilon}{2} + 2(1 \mp \varepsilon)\Big(p - \frac{1}{2}\Big), & \frac{1}{2} < p \le \frac{3}{4}, \\ 1 & \frac{3}{4} < p \le 1. \end{cases}$$

and

$$\begin{aligned} F_{B,\pm\varepsilon}(p) = \mathbb{I}\{0 \le p \le 1/4\} \left[\int_{1/4}^{1/2} 2(1 \pm \varepsilon)db + \int_{3/4}^{1} 2(1 \mp \varepsilon)db\right] + \\ + \mathbb{I}\{1/4 \le p \le 1/2\} \left[\int_{p}^{1/2} 2(1 \pm \varepsilon)db + \int_{3/4}^{1} 2(1 \mp \varepsilon)db\right] + \\ + \mathbb{I}\{1/2 \le p \le 3/4\} \int_{3/4}^{1} 2(1 \mp \varepsilon)db + \mathbb{I}\{3/4 \le p \le 1\} \int_{p}^{1} 2(1 \mp \varepsilon)db \end{aligned}$$

Which results in

$$F_{B,\pm\varepsilon}(p) = \begin{cases} 1, & 0 \le p \le \frac{1}{4}, \\ 2(1 \pm \varepsilon)\Big(\frac{1}{2} - p\Big) + \frac{1 \mp \varepsilon}{2} & \frac{1}{4} < p \le \frac{1}{2}, \\ \frac{1 \mp \varepsilon}{2} & \frac{1}{2} < p \le \frac{3}{4}, \\ 2(1 \mp \varepsilon)(1 - p) & \frac{3}{4} < p \le 1. \end{cases}$$

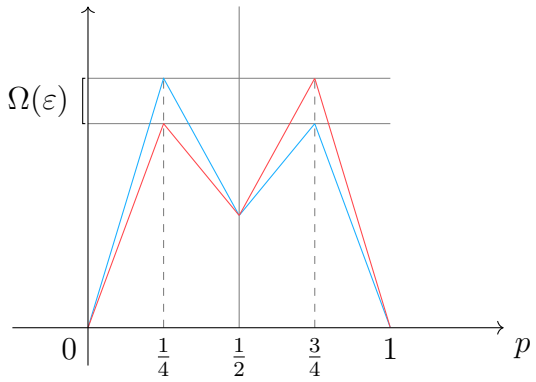


Figure 8.2: Expected trading-volume functions for the two hard instances in the lower-bound construction.

The multiplication $F_S \cdot F_B$ gives the explicit piecewise form:

$$\mathbb{E}_{S,B}\left[v_{\pm\varepsilon}(S,B,p)\right] = \begin{cases} 2p(1 \pm \varepsilon), & 0 \le p \le \frac{1}{4}, \\ \frac{1}{4}(1 \pm \varepsilon)(3 \pm \varepsilon) - p(1 \pm \varepsilon)^2, & \frac{1}{4} < p \le \frac{1}{2}, \\ \frac{1}{4}(1-\varepsilon^2) + \left(p - \frac{1}{2}\right)(1 \mp \varepsilon)^2, & \frac{1}{2} < p \le \frac{3}{4}, \\ 2(1-p)(1 \mp \varepsilon), & \frac{3}{4} < p \le 1. \end{cases}$$

□

From (8.1) we notice that in the $+\varepsilon$ case, the max over $[0, 1/2]$ is at $p = 1/4$ with value $2(1+\varepsilon) \cdot \dfrac{1}{4} = \dfrac{1}{2}(1+\varepsilon)$ and over $(1/2, 1]$ the max is at $p = 3/4$ with value $\frac{1}{4}(1-\varepsilon^2) + \frac{1}{4}(1 - 2\varepsilon + \varepsilon^2) = \frac{1}{2}(1-\varepsilon)$. Thus the gap between the best price in $[0, 1/2]$ and the best in $(1/2, 1]$ is exactly ε. By symmetry, the same reasoning can be applied to the $-\varepsilon$ case.

Theorem 2 (Lower bound for trading volume)**.** *For any horizon $T \in \mathbb{N}$ and any (possibly randomized) algorithm α, there exists an i.i.d. instance such that the expected regret in trading volume is*

$$R_T(\alpha) \geq \frac{7}{64}\sqrt{T}.$$

Proof. Let $((S_t^+, B_t^+))_{t\in\mathbb{N}}$ be an i.i.d. sequence of random variables such that, for all $t \in \mathbb{N}$, $S_t^+ \sim f_{S,+\varepsilon}$, $B_t^+ \sim f_{B,+\varepsilon}$, and S_t^+ is independent of B_t^+. Let $((S_t^-, B_t^-))_{t\in\mathbb{N}}$ be an i.i.d. sequence of random variables such that, for all $t \in \mathbb{N}$, $S_t^- \sim f_{S,-\varepsilon}$, $B_t^- \sim f_{B,-\varepsilon}$, and S_t^- is independent of B_t^-.

We consider the two instances:

- Instance "$+\varepsilon$", where $(S_t, B_t) \coloneqq (S_t^+, B_t^+)$
- Instance "$-\varepsilon$", where $(S_t, B_t) \coloneqq (S_t^-, B_t^-)$

From the piecewise characterization of the expected trading volume in Lemma 2, we see that the optimal prices are attained only at the boundary points $p = \frac{1}{4}$ and $p = \frac{3}{4}$. Any intermediate choice of p yields a strictly smaller expected gain compared to one of these boundary points. Consequently, an algorithm that assigns positive probability to prices outside $\{1/4, 3/4\}$ can only perform worse, in expectation, than an algorithm that plays exclusively from this restricted set. Therefore, without loss of generality, we can confine our analysis to algorithms α that choose prices solely in $\{1/4, 3/4\}$.

Note that under the $+\varepsilon$ instance, the per-round expected volume of price $1/4$ exceeds that of price $3/4$ by exactly ε; under the $-\varepsilon$ instance, the sign flips.

To lighten the notation, we define i.i.d. Bernoulli outcomes using the random vectors $\mathbf{Y}^+ = \left(Y_1^+, Y_2^+, \dots\right)$ and $\mathbf{Y}^- = \left(Y_1^-, Y_2^-, \dots\right)$, where:

$$Y_t^+(1) \coloneqq v(1/4, S_t^+, B_t^+) \sim \text{Bern}\left(\frac{1+\varepsilon}{2}\right), \quad Y_t^+(2) \coloneqq v(3/4, S_t^+, B_t^+) \sim \text{Bern}\left(\frac{1-\varepsilon}{2}\right),$$

$$Y_t^-(1) \coloneqq v(1/4, S_t^-, B_t^-) \sim \text{Bern}\left(\frac{1-\varepsilon}{2}\right), \quad Y_t^-(2) \coloneqq v(3/4, S_t^-, B_t^-) \sim \text{Bern}\left(\frac{1+\varepsilon}{2}\right),$$

independently across t. Let an arbitrary algorithm choose action $X_t^\pm \in \{1,2\}$ adaptively and let $N_T^+ \coloneqq \sum_{t=1}^T \mathbb{I}\{X_t^+ = 1\}$ and $N_T^- \coloneqq \sum_{t=1}^T \mathbb{I}\{X_t^- = 1\}$ (where the $\pm$ indicates whether we are in the $+\varepsilon$ or the $-\varepsilon$ instance).

The regret against the best action in each instance satisfies

$$R_T^{\mathbf{Y}^+}(\alpha) = \varepsilon\left(T - \mathbb{E}\left[N_T^+\right]\right), \qquad R_T^{\mathbf{Y}^-}(\alpha) = \varepsilon\mathbb{E}\left[N_T^-\right].$$

Therefore, using $\max\{u,v\} \geq (u+v)/2$,

$$\begin{aligned}
\max\left\{R_T^{\mathbf{Y}^+}(\alpha), R_T^{\mathbf{Y}^-}(\alpha)\right\} &\geq \frac{1}{2}\left(R_T^{\mathbf{Y}^+}(\alpha) + R_T^{\mathbf{Y}^-}(\alpha)\right) \\
&= \frac{1}{2}\left(\varepsilon\left(T - \mathbb{E}\left[N_T^+\right]\right) + \varepsilon\mathbb{E}\left[N_T^-\right]\right) \\
&= \frac{\varepsilon T}{2} - \frac{\varepsilon}{2}\left(\mathbb{E}\left[N_T^+\right] - \mathbb{E}\left[N_T^-\right]\right)
\end{aligned}$$

Taking absolute values in the last term,

$$\max\left\{R_T^{\mathbf{Y}^+}(\alpha), R_T^{\mathbf{Y}^-}(\alpha)\right\} \geq \frac{\varepsilon T}{2} - \frac{\varepsilon}{2}\left|\left(\mathbb{E}\left[N_T^+\right] - \mathbb{E}\left[N_T^-\right]\right)\right|$$

Now, by the definition of $N_T^{\pm}$ and the linearity of the expectation, we have that

$$\begin{aligned}
\left|\mathbb{E}[N_T^+] - \mathbb{E}[N_T^-]\right| &= \left|\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{X_t^+ = 1\}\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{X_t^- = 1\}\right]\right| \\
&= \left|\sum_{t=1}^{T} \Big(\mathbb{P}[X_t^+ = 1] - \mathbb{P}[X_t^- = 1]\Big)\right| \\
&\leq \sum_{t=1}^{T} \left|\mathbb{P}[X_t^+ = 1] - \mathbb{P}[X_t^- = 1]\right| \qquad \text{(Triangle inequality)}
\end{aligned}$$

Let $H_t^{\pm} = \left(Y_1^{\pm}, Y_2^{\pm}, \ldots, Y_{t-1}^{\pm}, U_1, U_2, \ldots, U_t\right)$ be the history available to the learner before picking $X_t^{\pm}$ (past observed feedback, together with the internal randomization used up to and including round t), then

$$\begin{aligned}
\sum_{t=1}^{T} \left|\mathbb{P}[X_t^+ = 1] - \mathbb{P}[X_t^- = 1]\right| &= \sum_{t=1}^{T} \left|\mathbb{P}[\alpha_t(H_t^+) = 1] - \mathbb{P}[\alpha_t(H_t^-) = 1]\right| \\
&= \sum_{t=1}^{T} \left|\mathbb{P}_{H_t^+}[\alpha_t^{-1}(1)] - \mathbb{P}_{H_t^-}[\alpha_t^{-1}(1)]\right| \\
&\leq \sum_{t=1}^{T} \left\|\mathbb{P}_{H_t^+} - \mathbb{P}_{H_t^-}\right\|_{\mathrm{TV}} \qquad \text{(Total Variation)}
\end{aligned}$$

Leveraging Pinsker's inequality (both the definition of Total Variation and the statement of

Pinsker's inequality are in Appendix E), and the chain rule of the KL divergence we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbb{P}_{H_t^+} - \mathbb{P}_{H_t^-}\|_{\mathrm{TV}} &\le \sum_{t=1}^{T} \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}_{H_t^+} \| \mathbb{P}_{H_t^-}\right)} \\
&\le \sum_{t=1}^{T} \sqrt{\frac{1}{2}\sum_{s=1}^{t} \mathcal{D}_{\mathrm{KL}}\left(\mathcal{B}\left(\frac{1}{2}(1+\varepsilon)\right) \middle\| \mathcal{B}\left(\frac{1}{2}(1-\varepsilon)\right)\right)} \\
&= \sum_{t=1}^{T} \sqrt{\frac{1}{2}\sum_{s=1}^{t} \varepsilon \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)} \qquad \text{(Lemma 1)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\ \mathbb{E}[N_T^+] - \mathbb{E}[N_T^-]\ | &\le \sum_{t=1}^{T} \sqrt{\frac{1}{2}\sum_{s=1}^{t} \varepsilon \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)} \\
&= \sum_{t=1}^{T} \sqrt{\frac{1}{2} t\varepsilon \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)} \\
&\le \sum_{t=1}^{T} \sqrt{\frac{1}{2} t\varepsilon \left(\frac{128}{49}\varepsilon\right)} \\
&= \sum_{t=1}^{T} \sqrt{\frac{64}{49}\varepsilon^2 t} \\
&\le \frac{8}{7}\varepsilon T^{\frac{3}{2}}
\end{aligned}$$

Thus,

$$\begin{aligned}\max\left(R_T^{\mathbf{Y}^+}(\alpha), R_T^{\mathbf{Y}^-}(\alpha)\right) &\geq \frac{\varepsilon T}{2} - \frac{\varepsilon}{2}\left|\mathbb{E}\left[N_T^+\right] - \mathbb{E}\left[N_T^-\right]\right| \\ &\geq \frac{\varepsilon T}{2} - \frac{\varepsilon}{2}\left(\frac{8}{7}\varepsilon T^{\frac{3}{2}}\right) \\ &= \frac{\varepsilon T}{2}\left(1 - \frac{8}{7}\varepsilon\sqrt{T}\right)\end{aligned}$$

and by choosing $\varepsilon = \frac{7}{16\sqrt{T}}$ we have obtain

$$R_T(\alpha) \geq \frac{7}{64}\sqrt{T}$$

Which proves the theorem. □

Theorem 2 shows that the $\sqrt{T}$ upper bound of Theorem 1 is essentially optimal in the full-feedback stochastic setting. Thus, in this regime, the minimax regret rate is of order $\sqrt{T}$. We now turn to the bandit-feedback setting, where the learner observes substantially less information, and where the picture changes dramatically.

9

bandit-feedback Stochastic (i.i.d.) Setting

In this section, we analyze the bandit-feedback setting, where the learner receives only partial information at the end of each round t. Specifically, the learner observes the binary outcome $\mathbb{I}\{s \leq p \leq b\}$, indicating whether a trade has occurred at the posted price p. If the indicator equals 1, the learner can infer that the posted price fell within the interval $[s, b]$, but does not observe the exact values of the seller's and buyer's evaluations. Conversely, if the indicator equals 0, the learner receives no information about which side (seller or buyer) caused the failure of the trade, and remains uncertain whether the seller's valuation exceeded p or the buyer's valuation was below p.

As in the full-feedback setting, we assume that the sequence of seller-buyer pairs $(S_1, B_1), (S_2, B_2), \ldots$ consists of independent and identically distributed (i.i.d.) random variables taking values in $[0, 1]^2$, all sampled from a common unknown distribution over $[0, 1]^2$.

In Subsection 9.1, we will show that, without additional assumptions, there is a linear lower bound and therefore no learning is achievable. In Section 9.2, we will show that by adding the bounded-density assumption, we will derive a sublinear upper bound.

9.1 Linear Lower Bound

In this section, our aim is to prove that, for any time horizon T, the minimax regret is linear in T, that is

$$\inf_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} R_T(\alpha, \theta) \geq cT$$

for some universal constant $c > 0$ (in our construction, $c = 1$). Let's first define the *Bayes risk*.

Definition 7 (Bayes risk / Bayes regret)**.** *Let $\Delta(\Theta)$ denote the set of all probability distributions on Θ. Let $y \in \Delta(\Theta)$ be a probability distribution over instances. The Bayes risk of an algorithm α up to horizon T under the prior y is defined as*

$$\mathbb{E}_{\theta \sim y}[R_T(\alpha, \theta)] = \int_{\Theta} R_T(\alpha, \theta)\, y(d\theta).$$

To obtain the linear lower bound, we will first prove that there is a linear regret for all deterministic algorithms, and then show that the regret over all algorithms is at least as large as the best Bayes regret over deterministic algorithms.

Now we define deterministic algorithms as history-dependent price-selection rules.

Definition 8 (Deterministic algorithm)**.** *A deterministic algorithm is a sequence of $[0, 1]$-valued functions β_t such that*

$$\begin{aligned} \beta_1 &\in [0, 1], \\ &\vdots \\ \beta_t &\colon [0, 1]^{t-1} \to [0, 1]. \end{aligned}$$

Next we extend the model to randomized algorithms by adding an explicit random seed input.

Definition 9 (Randomized algorithm)**.** *A randomized algorithm is a sequence of* $[0,1]$*-valued functions* α_t *such that*

$$\begin{aligned} \alpha_1 &\colon [0,1] \to [0,1], \\ &\vdots \\ \alpha_t &\colon [0,1]^{t-1} \times [0,1] \to [0,1]. \end{aligned}$$

Finally, we specify what constitutes an environment.

Definition 10 (Instance)**.** *An instance* θ *of the problem is a joint distribution on* $[0,1]^2$*.*

We denote by $\mathcal{A}$ and $\mathcal{D}$ the classes of *randomized* and *deterministic* algorithms, and by Θ the class of all *instances*, in our case represented by the pairs (S_t, B_t). For completeness, we note that deterministic algorithms are a special case of randomized ones (obtained by ignoring the random seed).

Remark 2. *The set of deterministic algorithms is included, up to a straightforward identification, in the set of randomized algorithms, i.e.* $\mathcal{D} \subset \mathcal{A}$*. Indeed, let* $\beta \in \mathcal{D}$ *be arbitrary. Recall that by Definition 8,* β *is a sequence* $(\beta_t)_{t \geq 1}$ *of* $[0,1]$*-valued functions such that*

$$\beta_1 \in [0,1] \quad \textit{and} \quad \beta_t \colon [0,1]^{t-1} \to [0,1] \quad \textit{for } t \geq 2.$$

We construct a randomized algorithm $\alpha \in \mathcal{A}$ *that behaves exactly like* β *and simply ignores*

the extra random input. Define a sequence $(\alpha_t)_{t\geq 1}$ *by*

$$\alpha_1 \colon [0,1] \to [0,1], \qquad \alpha_1(x) \coloneqq \beta_1 \quad \textit{for all } x \in [0,1],$$
$$\alpha_t \colon [0,1]^{t-1} \times [0,1] \to [0,1], \quad \alpha_t(\mathbf{x}, u) \coloneqq \beta_t(\mathbf{x}) \quad \textit{for all } (\mathbf{x}, u) \in [0,1]^{t-1} \times [0,1],\ t \geq 2.$$

In words, α_t *ignores the last coordinate (the "random seed" input) and applies the same mapping as* β_t *to the history coordinates.*

By Definition 9, $(\alpha_t)_{t\geq 1}$ *is a randomized algorithm, so* $\alpha \in \mathcal{A}$*. Since* β *was arbitrary, we have shown that every deterministic algorithm can be viewed as a (special case of a) randomized algorithm. Hence* $\mathcal{D} \subset \mathcal{A}$*.*

We now prove a lower bound on any fixed deterministic strategy.

Theorem 3. *Fix* $T \in \mathbb{N}$ *and a deterministic bandit algorithm* $\beta \in \mathcal{D}$*. Let* $\mathcal{P} = \{P_1, \ldots, P_T\}$ *be the finite set of prices that* β *plays along the trajectory where all feedbacks are* 0*. Then for all* $p \in [0,1] \setminus \mathcal{P}$*, if* θ_p *denotes the instance* $\delta_{(p,p)}$ [1]*, we have*

$$R_T(\beta, \theta_p) = T.$$

In particular, for each β*, for almost all* $x \in [0,1]$*,* $R_T(\beta, \theta_x) \geq T$*.*

Proof. Fix an arbitrary time horizon $T \in \mathbb{N}$, and fix any deterministic algorithm β. Since the algorithm is deterministic, its behavior is entirely specified by its internal rules and the

[1] By $\delta_{(p,p)}$ we mean the Dirac measure such that, for any measurable $A \subset [0,1]^2$, we have that $\delta_{(p,p)}(A) = 1$ if $(p,p) \in A$, 0 otherwise.

observed feedback. In the bandit setting, the feedback available to the algorithm at each round is only whether or not a trade occurred.

Now, suppose the learner receives no positive feedback throughout the entire time horizon. In that case, the learner's actions will be fully determined by its internal logic alone, since it cannot use any feedback to update its choices. Consequently, the algorithm will produce a fixed sequence of prices $P_1, P_2, \ldots, P_T \in [0, 1]$, which are the prices it would post if every feedback were zero.

Let us denote by $\mathcal{P} = \{P_1, P_2, \ldots, P_T\} \subset [0, 1]$ the finite set of prices played by the algorithm in this scenario.

Now, observe that there exist uncountably many prices $p \in [0, 1] \setminus \mathcal{P}$ that are not played by the learner at any round.

Let us fix any such $p \in [0, 1] \setminus \mathcal{P}$. We now define an i.i.d. distribution θ_p concentrated on the singleton pair (p, p), that is, we take $\theta_p = \delta_{(p,p)}$, where $\delta_{(p,p)}$ is the Dirac measure at $(p, p) \in [0, 1]^2$.

Under this distribution, for every round $t \in [T]$, the seller–buyer pair is exactly $(S_t, B_t) = (p, p)$ with probability 1. We now argue that, under θ_p, the realized sequence of prices posted by β is exactly the zero-feedback trajectory $(P_1, \ldots, P_T)$. This follows by induction on t. At $t = 1$, P_1 is determined without prior feedback and is therefore the same in both scenarios. Suppose that up to round $t - 1$ the realized feedback has been 0 at each round and the posted prices have been $P_1, \ldots, P_{t-1}$. Since β is deterministic and only depends on past actions and feedback, at round t it must play P_t. Because $P_t \neq p$, the feedback

at round t is 0 again, and the induction goes through. Hence, under θ_p the learner indeed plays P_t and receives feedback 0 for all t. In other words, every round has the same exact instance: a buyer and a seller who both value the good at precisely price p.

We now analyze the performance of the learner β under this instance: Since the learner plays prices $P_t \in \mathcal{P}$ and $p \notin \mathcal{P}$, we have $P_t \neq p$ for every round t. Given that $S_t = B_t = p$, the trading volume is $v(P_t, S_t, B_t) = \mathbb{I}\{S_t \leq P_t \leq B_t\} = \mathbb{I}\{p \leq P_t \leq p\}$, and therefore $v(P_t, p, p) = 0$ at every round. Thus, the learner receives zero reward on all T rounds: $\sum_{t=1}^{T} v(P_t, S_t, B_t) = 0$.

Now consider the benchmark: the optimal fixed price in hindsight. Since all $(S_t, B_t) = (p, p)$, any learner that consistently plays $p^* = p$ at each round will trigger a trade every time. So the reward of the optimal fixed price is:

$$\sum_{t=1}^{T} v(p, S_t, B_t) = \sum_{t=1}^{T} \mathbb{I}\{p \leq p \leq p\} = \sum_{t=1}^{T} 1 = T.$$

Finally, the regret of learner β under distribution θ_p is:

$$R_T(\beta, \theta_p) = \sum_{t=1}^{T} \left(v(p, S_t, B_t) - v(P_t, S_t, B_t)\right) = T - 0 = T.$$

Since β was arbitrary, this completes the proof: for any deterministic learner and any time horizon T, we can construct a distribution θ_p under which the learner incurs regret exactly equal to T. Moreover, since $\mathcal{P}$ is finite, its complement $[0, 1] \setminus \mathcal{P}$ has Lebesgue measure one, and for all p in this set we have $R_T(\beta, \theta_p) = T$. □

Claim 1 (Randomization does not improve Bayes risk)**.** *Fix a time horizon T and let $\Delta(\Theta)$ denote the set of all probability distributions on Θ. Then, for all $y \in \Delta(\Theta)$,*

$$\inf_{\alpha \in \mathcal{A}} \mathbb{E}_{\theta \sim y}[R_T(\alpha, \theta)] = \inf_{\beta \in \mathcal{D}} \mathbb{E}_{\theta \sim y}[R_T(\beta, \theta)].$$

Proof. Let $y \in \Delta(\Theta)$ be arbitrary. By the definition of regret, for any (possibly randomized) algorithm α and instance θ,

$$R_T(\alpha, \theta) = \max_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} v(p, S_t, B_t) - \sum_{t=1}^{T} v(P_t, S_t, B_t) \;\Big|\; \theta\right],$$

where the expectation is over the randomness of the environment (given θ) and of the algorithm α.

The second term does not depend on p, so for any fixed α, θ,

$$\begin{aligned} R_T(\alpha, \theta) &= \max_{p \in [0,1]} \left(\mathbb{E}\Big[\sum_{t=1}^{T} v(p, S_t, B_t) \;\Big|\; \theta\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} v(P_t, S_t, B_t) \;\Big|\; \theta\Big] \right) \\ &= \left(\max_{p \in [0,1]} \mathbb{E}\Big[\sum_{t=1}^{T} v(p, S_t, B_t) \;\Big|\; \theta\Big] \right) - \mathbb{E}\Big[\sum_{t=1}^{T} v(P_t, S_t, B_t) \;\Big|\; \theta\Big]. \end{aligned}$$

The only place α appears is through $\mathbb{E}[\sum_{t=1}^{T} v(P_t, S_t, B_t) \mid \theta]$, which is linear in the distribution of the internal randomness of α.

Now take any randomized algorithm α. Think of it as first sampling a seed ω and then running the deterministic algorithm β_ω. Equivalently, α induces a probability distribution

μ_α over deterministic algorithms:

$$\beta \sim \mu_\alpha.$$

By construction,

$$R_T(\alpha, \theta) = \mathbb{E}_{\beta\sim\mu_\alpha}[R_T(\beta, \theta)].$$

Therefore

$$\begin{aligned}\mathbb{E}_{\theta\sim y}[R_T(\alpha, \theta)] &= \mathbb{E}_{\theta\sim y}\Big[\mathbb{E}_{\beta\sim\mu_\alpha}[R_T(\beta, \theta)]\Big] \\ &= \mathbb{E}_{\beta\sim\mu_\alpha}\Big[\underbrace{\mathbb{E}_{\theta\sim y}[R_T(\beta, \theta)]}_{=:f(\beta)}\Big].\end{aligned}$$

Now, for any probability measure μ on $\mathcal{D}$,

$$\mathbb{E}_{\beta\sim\mu}[f(\beta)] \geq \inf_{\beta\in\mathcal{D}} f(\beta),$$

since an average of real numbers is always at least their infimum. Every randomized algorithm $\alpha \in \mathcal{A}$ corresponds to some such $\mu = \mu_\alpha$, so

$$\inf_{\alpha\in\mathcal{A}} \mathbb{E}_{\theta\sim y}[R_T(\alpha, \theta)] = \inf_{\alpha\in\mathcal{A}} \mathbb{E}_{\beta\sim\mu_\alpha}[f(\beta)] \geq \inf_{\beta\in\mathcal{D}} f(\beta).$$

On the other hand, $\mathcal{D} \subset \mathcal{A}$ implies

$$\inf_{\alpha\in\mathcal{A}} \mathbb{E}_{\theta\sim y}[R_T(\alpha, \theta)] \leq \inf_{\beta\in\mathcal{D}} \mathbb{E}_{\theta\sim y}[R_T(\beta, \theta)] = \inf_{\beta\in\mathcal{D}} f(\beta).$$

Combining the two inequalities yields

$$\inf_{\alpha\in\mathcal{A}} \mathbb{E}_{\theta\sim y}[R_T(\alpha,\theta)] = \inf_{\beta\in\mathcal{D}} \mathbb{E}_{\theta\sim y}[R_T(\beta,\theta)],$$

as claimed. □

Theorem 4 (Yao-type lower bound)**.** *Fix a time horizon $T \in \mathbb{N}$. Let $\mathcal{A}$ be the set of (possibly randomized) algorithms, $\mathcal{D}$ the set of deterministic algorithms, and Θ the set of instances (probability distributions on $[0,1]^2$). Let $\Delta(\Theta)$ denote the set of all probability distributions on Θ.*

Then the worst-case (over instances) regret of the best algorithm is at least as large as the best Bayes regret over deterministic algorithms, i.e.

$$\inf_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} R_T(\alpha,\theta) \geq \sup_{y\in\Delta(\Theta)} \inf_{\beta\in\mathcal{D}} \mathbb{E}_{\theta\sim y}[R_T(\beta,\theta)].$$

Proof. For all $\alpha \in \mathcal{A}$ and all $y \in \Delta(\Theta)$,

$$\begin{aligned}
\mathbb{E}_{\theta\sim y}[R_T(\alpha,\theta)] &= \int_\Theta R_T(\alpha,\theta)dy(\theta) \\
&\leq \int_\Theta \sup_{\theta'\in\Theta} R_T(\alpha,\theta')dy(\theta) \\
&= \sup_{\theta'\in\Theta} R_T(\alpha,\theta') \cdot 1,
\end{aligned}$$

therefore

$$\sup_{\theta\in\Theta} R_T(\alpha,\theta) \geq \mathbb{E}_{\theta\sim y}[R_T(\alpha,\theta)].$$

Taking $\inf_{\alpha \in \mathcal{A}}$ on both sides, we obtain

$$\begin{aligned} \inf_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} R_T(\alpha, \theta) &\geq \inf_{\alpha \in \mathcal{A}} \mathbb{E}_{\theta \sim y}[R_T(\alpha, \theta)] \\ &= \inf_{\beta \in \mathcal{D}} \mathbb{E}_{\theta \sim y}[R_T(\beta, \theta)], \end{aligned}$$

where the last equality uses Claim 1. Now take $\sup_{y \in \Delta(\Theta)}$ on both sides. The left-hand side is independent of y, hence

$$\sup_{y \in \Delta(\Theta)} \inf_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} R_T(\alpha, \theta) = \inf_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} R_T(\alpha, \theta),$$

and the right-hand side is $\sup_{y \in \Delta(\Theta)} \inf_{\beta \in \mathcal{D}} \mathbb{E}_{\theta \sim y}[R_T(\beta, \theta)]$, as required. □

Now we are going to leverage Theorem 3 and Theorem 4 to show that the regret over all algorithms is at least linear.

Theorem 5 (Linear lower bound)**.** *For all time horizons T,*

$$\inf_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} R_T(\alpha, \theta) \geq T.$$

Proof. We divide the proof into two steps.

Step 1. By Theorem 3, for all deterministic algorithms $\beta \in \mathcal{D}$, for almost all $x \in [0, 1]$ we have a linear lower bound

$$R_T(\beta, \theta_x) \geq T,$$

where θ_x denotes the instance associated with x.

Step 2. Consider the set of instances

$$\widehat{\Theta} \coloneqq \{\delta_{(x,x)} \colon x \in [0,1]\} \subset \Theta,$$

i.e., the set of all Dirac measures $\delta_{(x,x)}$ with $x \in [0,1]$.

Take a random variable $X \sim \mathrm{Unif}([0,1])$ and define the Θ-valued random variable $\delta_{X,X}$; let $\tilde{y}$ be its distribution on Θ (so $\tilde{y} \in \Delta(\Theta)$).

Using Theorem 4, we know that

$$\begin{aligned}\inf_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} R_T(\alpha,\theta) &\geq \sup_{y\in\Delta(\Theta)} \inf_{\beta\in\mathcal{D}} \mathbb{E}_{\theta\sim y}[R_T(\beta,\theta)] \\ &= \sup_{y\in\Delta(\Theta)} \inf_{\beta\in\mathcal{D}} \int_\Theta R_T(\beta,\theta) dy(\theta).\end{aligned}$$

Define

$$g(y) \coloneqq \inf_{\beta\in\mathcal{D}} \int_\Theta R_T(\beta,\theta) dy(\theta), \qquad y \in \Delta(\Theta).$$

Since $\tilde{y} \in \Delta(\Theta)$, by the definition of the supremum we have

$$\sup_{y\in\Delta(\Theta)} g(y) \geq g(\tilde{y}) = \inf_{\beta\in\mathcal{D}} \int_\Theta R_T(\beta,\theta) d\tilde{y}(\theta).$$

Combining the last two displays yields

$$\inf_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} R_T(\alpha,\theta) \geq \inf_{\beta\in\mathcal{D}}\int_{\Theta} R_T(\beta,\theta)d\tilde{y}(\theta).$$

By construction, $\tilde{y}$ is the distribution of $\delta_{X,X}$ with $X \sim \text{Unif}([0,1])$. Hence, by a standard change-of-measure argument,

$$\int_{\Theta} R_T(\beta,\theta)d\tilde{y}(\theta) = \int_{[0,1]} R_T(\beta,\delta_{x,x})dx,$$

and therefore

$$\inf_{\beta\in\mathcal{D}}\int_{\Theta} R_T(\beta,\theta)d\tilde{y}(\theta) = \inf_{\beta\in\mathcal{D}}\int_{[0,1]} R_T(\beta,\delta_{x,x})dx.$$

From Step 1, for all $\beta \in \mathcal{D}$ and for almost all $x \in [0,1]$ we have $R_T(\beta,\delta_{x,x}) \geq T$. Using the monotonicity of the integral, we obtain

$$\begin{aligned}\inf_{\beta\in\mathcal{D}}\int_{[0,1]} R_T(\beta,\delta_{x,x})dx &\geq \inf_{\beta\in\mathcal{D}}\int_{[0,1]} T dx \\ &= \inf_{\beta\in\mathcal{D}} T \\ &= T.\end{aligned}$$

Thus,

$$\inf_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} R_T(\alpha,\theta) \geq \inf_{\beta\in\mathcal{D}}\int_{[0,1]} R_T(\beta,\delta_{x,x})dx \geq T,$$

which concludes the proof. □

9.2 Upper Bound Under Bounded Density

In the previous subsection we proved a linear lower bound on the regret in the bandit-feedback setting, showing that without additional assumptions there is no hope for sublinear regret. In this subsection we show that, under a mild regularity assumption on the distribution of (S, B), one can in fact obtain a sublinear upper bound. More precisely, we assume that the joint distribution of (S, B) has bounded density (equivalently, that its cumulative distribution function is Lipschitz), and we prove that one can achieve regret of order $T^{2/3}$.

Assumption (bounded density / Lipschitz CDF). Let (S, B) be a random pair in $[0, 1]^2$ with joint distribution F. We assume that there exists a constant $\sigma > 0$ such that for all $(p, q), (p', q') \in [0, 1]^2$,

$$|F(p, q) - F(p', q')| \leq \sigma \max\{|p - p'|, |q - q'|\}.$$

Equivalently, the map

$$(p, q) \longmapsto \mathbb{P}[S \leq p, B \leq q]$$

is σ-Lipschitz on $[0, 1]^2$ in the sup–norm.

Lemma 3 (Difference of two Lipschitz functions)**.** *Let σ be a non-negative real number and $f, g\colon [0,1] \to \mathbb{R}$ be σ-Lipschitz functions, i.e.,*

$$|f(x) - f(y)| \le \sigma|x - y| \quad \textit{and} \quad |g(x) - g(y)| \le \sigma|x - y| \qquad \forall x, y \in [0,1].$$

Then the function $h \coloneqq f - g$ is 2σ-Lipschitz, that is,

$$|h(x) - h(y)| \le 2\sigma|x - y| \qquad \forall x, y \in [0,1].$$

Proof. For any $x, y \in [0,1]$ we have

$$\begin{aligned} h(x) - h(y) &= (f(x) - g(x)) - (f(y) - g(y)) \\ &= (f(x) - f(y)) - (g(x) - g(y)), \end{aligned}$$

so, by the triangle inequality,

$$|h(x) - h(y)| = |(f(x) - f(y)) - (g(x) - g(y))| \le |f(x) - f(y)| + |g(x) - g(y)|.$$

Using the Lipschitz property of f and g,

$$|h(x) - h(y)| \le \sigma|x - y| + \sigma|x - y| = 2\sigma|x - y|,$$

which proves that h is 2σ-Lipschitz. □

Recall that the reward at price p is

$$v(p, S, B) = \mathbb{I}\{S \leq p \leq B\},$$

and that the regret at time horizon T is defined as

$$R_T(\alpha) = \sup_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} v(p, S_t, B_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} v(P_t, S_t, B_t)\right],$$

where P_t is the price posted by the algorithm α at round t, and (S_t, B_t) are i.i.d. copies of (S, B).

Lemma 4 (Lipschitz continuity of the expected reward)**.** *Under the bounded-density assumption above, the function*

$$f\colon [0,1] \to [0,1], \qquad f(p) \coloneqq \mathbb{E}[v(p, S, B)]$$

is 2σ*-Lipschitz, i.e.,*

$$|f(p) - f(p')| \leq 2\sigma |p - p'| \qquad \textit{for all } p, p' \in [0,1].$$

Proof. For any $p \in [0,1]$,

$$\begin{aligned} f(p) &= \mathbb{E}[\mathbb{I}\{S \le p \le B\}] \\ &= \mathbb{E}[\mathbb{I}\{S \le p\}\mathbb{I}\{p \le B\}] \\ &= \mathbb{E}[\mathbb{I}\{S \le p\}(1 - \mathbb{I}\{p > B\})] \\ &= \mathbb{E}[\mathbb{I}\{S \le p\}] - \mathbb{E}[\mathbb{I}\{S \le p\}\mathbb{I}\{B < p\}] \\ &= \mathbb{P}[S \le p] - \mathbb{P}[S \le p, B < p]. \end{aligned}$$

Since $\mathbb{P}[B = p] = 0$ under our bounded-density assumption, we can equivalently write

$$f(p) = \mathbb{P}[S \le p] - \mathbb{P}[S \le p, B \le p].$$

The map $(p,q) \mapsto \mathbb{P}[S \le p, B \le q]$ is σ-Lipschitz on $[0,1]^2$ by assumption. Therefore, since the function $p \mapsto \mathbb{P}[S \le p, B \le p]$ is the restriction of a σ-Lipschitz function to the diagonal $\{(p,p) : p \in [0,1]\}$, it is σ-Lipschitz. Also, the function $p \mapsto \mathbb{P}[S \le p]$ can be written as $p \mapsto \mathbb{P}[S \le p, B \le 1]$, i.e., as the restriction of the same σ-Lipschitz function to the horizontal line $\{(p,1) : p \in [0,1]\}$, hence it is also σ-Lipschitz. Thus f is the difference of two σ-Lipschitz functions, so, by Lemma 3 it is 2σ-Lipschitz:

$$|f(p) - f(p')| \le 2\sigma|p - p'| \quad \forall p, p' \in [0,1].$$

□

Definition 11 (Mesh size)**.** *Let $\mathcal{X} \subset [0,1]$ be a finite discretization and define its mesh size*

$$\delta(\mathcal{X}) \coloneqq \sup_{p\in[0,1]} \inf_{b\in\mathcal{X}} |b-p|.$$

For a fixed $p \in [0,1]$, $\inf_{b\in\mathcal{X}}|b-p|$ is the distance from p to the closest grid point in $\mathcal{X}$, so the mesh size $\delta(\mathcal{X})$ is the largest possible distance between a point $p \in [0,1]$ and the nearest discretization point.

Lemma 5 (Approximation by a discretization)**.** *Let $\mathcal{X} \subset [0,1]$ be a finite discretization and let $\delta(\mathcal{X})$ be its mesh size. Let $f(p) = \mathbb{E}[v(p,S,B)]$ as above, and suppose f is L-Lipschitz. Then*

$$\sup_{p\in[0,1]} f(p) - \sup_{b\in\mathcal{X}} f(b) \le L\delta(\mathcal{X}).$$

In particular, under Lemma 4, the gap is at most $2\sigma\delta(\mathcal{X})$.

Proof. Since f is continuous (Lipschitz) on the compact interval $[0,1]$, it attains its maximum at some $p^\star \in [0,1]$:

$$f(p^\star) = \max_{p\in[0,1]} f(p).$$

By the definition of $\delta(\mathcal{X})$, there exists $\tilde{b} \in \mathcal{X}$ with $|\tilde{b} - p^\star| \le \delta(\mathcal{X})$. Then

$$\sup_{p\in[0,1]} f(p) - \sup_{b\in\mathcal{X}} f(b) = f(p^\star) - \sup_{b\in\mathcal{X}} f(b) \le f(p^\star) - f(\tilde{b}) \le L|p^\star - \tilde{b}| \le L\delta(\mathcal{X}),$$

where we used the Lipschitz property in the penultimate inequality. □

The Lipschitz property of $f(p) = \mathbb{E}[v(p, S, B)]$ implies that a sufficiently fine grid approximates the best continuous price up to a controlled additive loss. This suggests a simple strategy: choose an appropriate grid $\mathcal{X}$, treat each grid point as an arm, and apply MOSS to learn the best arm.

Algorithm 2 Discretized MOSS (D-MOSS)

Require: time horizon T, density bound σ (or any known upper bound)
  initialization:
  set $K \leftarrow \max\left\{1, \left\lceil \left(\frac{\sigma}{35}\right)^{2/3} T^{1/3} \right\rceil\right\}$;
  set grid $\mathcal{X} \leftarrow \left\{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\right\}$;
  initialize an instance of $\texttt{MOSS}(\mathcal{X}, T)$ (see Appendix F);
  for $t = 1, 2, \ldots, T$ **do**
    let $P_t \leftarrow$ the arm recommended by $\texttt{MOSS}(\mathcal{X}, T)$ at round t;
    post price P_t;
    observe bandit reward $Y_t \leftarrow v(P_t, S_t, B_t) \in \{0, 1\}$;
    feed Y_t to $\texttt{MOSS}(\mathcal{X}, T)$ (update its internal statistics);
  end for

Algorithm 2 implements this discretize-and-learn approach, with a grid size K tuned as a function of T and σ. The next theorem shows that the resulting regret is $O(\sigma^{1/3}T^{2/3})$, matching the discretization/exploration tradeoff.

Theorem 6 (Sublinear regret under bounded density)**.** *Under the bounded-density assumption above, Algorithm 2 satisfies*

$$R_T \leq C\sigma^{1/3}T^{2/3},$$

for a constant $C > 0$.

Proof. We proceed in three steps.

Step 1: discretization of the action space. Fix an integer $K \geq 1$ and consider the uniform discretization

$$\mathcal{X} \coloneqq \left\{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\right\}.$$

This is a $(K+1)$-point grid on $[0,1]$ with mesh size at most $\delta(\mathcal{X}) \leq 1/(2K)$ (the worst case is halfway between consecutive grid points). By Lemmas 4 and 5, we have, for all T,

$$\sup_{p\in[0,1]} f(p) - \sup_{b\in\mathcal{X}} f(b) \leq 2\sigma\delta(\mathcal{X}) \leq \frac{\sigma}{K}.$$

At the level of cumulative reward over T rounds, this yields a *discretization error* bounded by

$$T\left(\sup_{p\in[0,1]} f(p) - \sup_{b\in\mathcal{X}} f(b)\right) \leq \frac{\sigma T}{K}.$$

Step 2: running a stochastic bandit algorithm on the grid. We now treat $\mathcal{X}$ as the finite set of arms of a stochastic bandit problem with rewards in $[0,1]$, where arm $b \in \mathcal{X}$ has mean reward $f(b) = \mathbb{E}[v(b, S, B)]$. For example, we may run the Minimax Optimal Strategy in the Stochastic case (MOSS, Appendix F) algorithm of Bubeck and Audibert [18], which guarantees, for rewards in $[0,1]$ and $(K+1)$ arms, that

$$\sup_{b\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} v(b, S_t, B_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} v(P_t, S_t, B_t)\right] \leq 49\sqrt{(K+1)T} \leq 70\sqrt{KT},$$

where $P_t \in \mathcal{X}$ is the price chosen at round t by the bandit algorithm.

Step 3: combining discretization error and bandit regret. The *true* regret of our algorithm (with continuous benchmark over $[0,1]$) satisfies

$$\begin{aligned} R_T &= \sup_{p\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} v(p, S_t, B_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} v(P_t, S_t, B_t)\right] \\ &= \underbrace{\sup_{p\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} v(p, S_t, B_t)\right] - \sup_{b\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} v(b, S_t, B_t)\right]}_{\text{discretization error}} \\ &\quad + \underbrace{\sup_{b\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} v(b, S_t, B_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} v(P_t, S_t, B_t)\right]}_{\text{bandit regret on } \mathcal{X}}. \end{aligned}$$

By the discussion above, these two terms are bounded respectively by $\frac{\sigma T}{K}$ and $70\sqrt{KT}$. Hence, for every integer $K \geq 1$,

$$R_T \leq \frac{\sigma T}{K} + 70\sqrt{KT}.$$

We now optimize the right-hand side over $K > 0$ (treating K as a real variable for simplicity). Let

$$g(K) \coloneqq \frac{\sigma T}{K} + 70\sqrt{KT} = \frac{\sigma T}{K} + 70T^{1/2}K^{1/2}.$$

A straightforward calculus exercise shows that the minimizer is

$$K^\star = \left(\frac{2\sigma T}{70\sqrt{T}}\right)^{2/3} = \left(\frac{\sigma}{35}\right)^{2/3} T^{1/3},$$

and that, for this choice,

$$g(K^\star) \le C\sigma^{1/3}T^{2/3}$$

for some universal constant C (one can take $C \le 33$ by explicit computation). Choosing $K = \max\{1, \lfloor K^\star \rfloor\}$ yields the same bound up to a harmless multiplicative constant.

Thus there exists a bandit algorithm (MOSS run on a suitably fine discretization of $[0, 1]$) whose regret satisfies

$$R_T \le C\sigma^{1/3}T^{2/3},$$

which completes the proof. □

10

Conclusion

This thesis studied online posted-price mechanisms for bilateral trade under uncertainty about agents' private valuations. We focused on the *trading volume* reward function, where the platform receives reward $v(p, S, B) = \mathbb{I}\{S \leq p \leq B\}$, and on a stochastic model in which the valuation pairs $(S_t, B_t)_{t\geq 1}$ are i.i.d. draws from an unknown distribution on $[0, 1]^2$. Adopting an online learning perspective, we formalized the problem in terms of a decision space (prices), an outcome space (valuation pairs), feedback models, and regret with respect to the best fixed price. This framework makes it possible to directly compare what can and cannot be learned under different information structures.

Our main message is that the *extent of feedback* is the key driver of learnability. In the full-feedback model, where the platform observes (S_t, B_t) after each interaction, we designed a simple learning rule (*Follow the Best Price*) and proved that it achieves sublinear regret of order $\sqrt{T}$; moreover, we established a matching lower bound (up to constants and minor factors), showing that the $\sqrt{T}$ rate is essentially optimal in this setting. In sharp contrast, in the bandit-feedback model where the platform only observes the trade/no-trade indicator, we proved an impossibility result: without any additional structure on the valuation distribution, no algorithm can guarantee sublinear regret, and in fact the minimax regret grows linearly with T. This shows that learning from pure accept/reject feedback is information-theoretically impossible in the worst case.

We then identified a natural structural condition under which bandit learning becomes feasible. Assuming that the joint distribution of (S, B) has *bounded density* (equivalently, a Lipschitz CDF), we constructed a bandit algorithm based on discretizing the price interval and running a suitable finite-armed bandit strategy on the resulting grid. Under this

regularity assumption, we proved that the regret scales as $T^{2/3}$ (up to constants depending on the density bound). Taken together, these results delineate a clean picture for online bilateral trade with volume objective in the i.i.d. setting: full-feedback permits efficient learning at the classical $\sqrt{T}$ rate, pure bandit-feedback is impossible without assumptions, and bounded densities restore sublinear regret at an intermediate $T^{2/3}$ rate.

Relation to prior work and originality. To the best of our knowledge, the specific online posted-price bilateral-trade setting studied in this thesis, with the trading volume reward function and an explicit comparison between full-feedback and bandit-feedback models, has not been previously analyzed in the literature. While several proof techniques used throughout are standard in online learning and bandits, the results and guarantees stated here are new for this problem. In the full-feedback model, our upper bound is achieved by the *Follow-the-Best-Price* strategy, which also appears in [12]; our analysis follows a similar high-level route, adapted to our formulation and notation. We complement this with a matching $\Omega(\sqrt{T})$ lower bound proved via a different proof route. For bandit-feedback, we establish a worst-case impossibility result (linear minimax regret without additional structure), and we show that sublinear regret can be recovered under a natural regularity condition (bounded density / Lipschitz CDF) by discretization and a finite-armed bandit reduction; to our knowledge, these bandit-feasibility and bandit-impossibility statements are new in this bilateral-trade posted-price framework.

Limitations. The scope of this thesis is intentionally focused, and several modelling choices leave open directions for extensions. First, we worked in the i.i.d. stochastic setting and compared performance to the best *fixed* price in hindsight; extending the analysis to non-stationary markets, adversarial sequences, or stronger benchmarks (e.g., dynamic pricing policies) would require different tools. Second, we restricted attention to posted-price mechanisms and to the trading volume objective; other objectives such as gain from trade or fairness-aware variants may exhibit different information-theoretic behavior. Third, our positive bandit result relies on a smoothness assumption (bounded densities / Lipschitz CDF) and, as stated, uses the corresponding bound as an input to tune the discretization; weakening this requirement, or adapting to unknown smoothness, is an important practical consideration.

Future work. A first and central direction is to close the remaining gap in the bandit setting under bounded densities by proving a matching *lower bound*. Concretely, our analysis shows that $T^{2/3}$ regret is achievable under bounded densities; an important next step is to establish whether $T^{2/3}$ is also minimax-optimal in this class (up to constants), or whether faster rates are possible with a more refined algorithm or a more restrictive regularity condition. Our conjecture is that it would still be a $T^{2/3}$ rate. A quick reason could be that, if $S = 0$ was given, the problem would simplify essentially to $v(p, 0, B) = \mathbb{I}\{p \leq B\}$. This problem is very similar to a dynamic pricing problem, where the reward function is $w(p) = p\mathbb{I}\{p \leq B\}$. The intuition is that, having a multiplicative factor of p in the function, makes the problem a little easier, and for that problem there is a known worst-case lower

bound of $T^{2/3}$ ([19]).

Beyond this, several extensions appear promising: (i) designing *adaptive* procedures that do not require prior knowledge of the density/Lipschitz constant (and ideally adapt to unknown smoothness); (ii) studying alternative objectives within the same online-learning formalism and comparing how feedback constraints change the achievable regret rates; (iii) moving beyond the i.i.d. model to capture temporal dependence, regime changes, or strategic responses to posted prices; and (iv) considering richer feedback signals that may be available on platforms in practice (e.g., knowing whether the seller or the buyer rejected, or observing partial bids/asks), to quantify how incremental information improves learnability.

Overall, this thesis provides a quantitative separation between feedback models in online bilateral trade with the trading volume reward.

APPENDICES

A

Existence of the Best Price

In this appendix, we show that a price $p^\star$ maximizing the expected trading volume always exists. For any $p \in [0,1]$ and any $s, b \in [0,1]$, we define the trading volume at price p by

$$v(p,s,b) \coloneqq \mathbb{I}\{s \leq p \leq b\}.$$

so that for $[0,1]$-valued random variables (S,B) we have

$$\mathbb{E}[v(p,S,B)] = \mathbb{P}(S \leq p \leq B).$$

Lemma 6. *The function*

$$p \longmapsto \mathbb{E}[v(p,S,B)] = \mathbb{P}(S \leq p \leq B)$$

is upper semicontinuous on $[0,1]$*. In particular, there exists a maximizer* $p^\star \in [0,1]$*.*

Proof. Let $(p_n)_{n\geq 1}$ be any sequence in $\mathbb{R}$ such that $p_n \to p$. Set $X_n \coloneqq \mathbb{I}\{S \leq p_n \leq B\}$ and $X \coloneqq \mathbb{I}\{S \leq p \leq B\}$.

We claim that $\limsup_{n\to\infty} X_n \leq X$ pointwise. Indeed, fix $(s,b) \in [0,1]^2$. If $p \notin [s,b]$, then since $[s,b]$ is closed and $p_n \to p$, we have $p_n \notin [s,b]$ for all n large enough, hence $\limsup_n \mathbb{I}\{s \leq p_n \leq b\} = 0$. If $p \in [s,b]$, then trivially $\limsup_n \mathbb{I}\{s \leq p_n \leq b\} \leq 1 = \mathbb{I}\{s \leq p \leq b\}$. This proves $\limsup_n X_n \leq X$.

Since $0 \leq X_n \leq 1$, by the reverse Fatou lemma,

$$\limsup_{n\to\infty} \mathbb{E}[X_n] \leq \mathbb{E}\left[\limsup_{n\to\infty} X_n\right] \leq \mathbb{E}[X].$$

Equivalently,

$$\limsup_{n\to\infty} \mathbb{E}[v(p_n, S, B)] \leq \mathbb{E}[v(p, S, B)]$$

so $p \mapsto \mathbb{E}[v(p, S, B)]$ is upper semicontinuous.

Finally, an upper semicontinuous function on the compact set $[0, 1]$ attains its maximum, so there exists $p^\star \in [0, 1]$ maximizing $\mathbb{E}[v(p, S, B)]$. □

B

Additional lemmas

In this appendix, we state and prove some additional lemmas (known results), that we use in section 8.1.

Lemma 7 (Tower property of the conditional expectation)**.** *Let $X \in \mathcal{L}^1$ be a random variable and $\mathcal{G}_1 \subseteq \mathcal{G}_2$ be two σ-algebras, then almost surely*

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbb{E}[X \mid \mathcal{G}_1]$$

Proof. Let $Y \coloneqq \mathbb{E}[X \mid \mathcal{G}_1]$, then

- Y is $\mathcal{G}_1$-measurable by definition
- $\forall G \in \mathcal{G}_1$

$$\int_G Y d\mathbb{P} = \int_G \mathbb{E}[X \mid \mathcal{G}_1] d\mathbb{P} = \int_G X d\mathbb{P} = \int_G \mathbb{E}[X \mid \mathcal{G}_2] d\mathbb{P}$$

  where the last step is true because if $G \in \mathcal{G}_1$ then $G \in \mathcal{G}_2$.

□

Lemma 8 (Freezing Lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ and $(D, \mathcal{D})$ two measurable spaces. Let $X \colon (\Omega, \mathcal{F}) \to (D, \mathcal{D})$ and $Y \colon (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ be two random variables. Assume that $\mathcal{X}, \mathcal{Y} \subset \mathcal{F}$ are two σ-algebras such that X is $\mathcal{X}/\mathcal{D}$ measurable, Y is $\mathcal{Y}/\mathcal{E}$ measurable and $\mathcal{X}$ is independent of $\mathcal{Y}$. Then, for all bounded $(\mathcal{D} \times \mathcal{E})/\mathcal{B}(\mathbb{R})$ measurable functions $\Phi \colon D \times E \to \mathbb{R}$,*

$$\mathbb{E}[\Phi(X, Y) \mid \mathcal{X}] = [\mathbb{E}[\Phi(x, Y)]]_{x=X} = \mathbb{E}[\Phi(X, Y) \mid X] \tag{B.1}$$

Proof. Assume first that $\Phi(x,y)$ is of the form $\phi(x)\psi(y)$. Then

$$\mathbb{E}[\underbrace{\phi(X)\psi(Y)}_{=\Phi(X,Y)} \mid \mathcal{X}] = \phi(X)\mathbb{E}[\psi(Y) \mid \mathcal{X}] \overset{Y \perp\!\!\!\perp X}{=} \phi(X)\mathbb{E}[\psi(Y)] = \mathbb{E}\left[\phi(x)\psi(Y)\right]_{x=X}.$$

Now fix some $F \in \mathcal{X}$ and pick $\phi(x) = \mathbb{I}_A(x)$ and $\psi(y) = \mathbb{I}_B(y)$, where $A \in \mathcal{D}$ and $B \in \mathcal{E}$. Our argument shows that

$$\int_F \mathbb{I}_{A\times B}(X(\omega), Y(\omega))\mathbb{P}(d\omega) = \int_F \mathbb{E}[\mathbb{I}_{A\times B}(x,Y)]_{x=X(\omega)}\mathbb{P}(d\omega), \qquad F \in \mathcal{X}.$$

Both sides of this equality (can be extended from $\mathcal{D}\times\mathcal{E}$) to define measures on the product σ-algebra $\mathcal{D}\otimes\mathcal{E}$. By linearity, this becomes

$$\int_F \Phi(X(\omega), Y(\omega))\mathbb{P}(d\omega) = \int_F \mathbb{E}[\Phi(x,Y)]_{x=X(\omega)}\mathbb{P}(d\omega), \qquad F \in \mathcal{X},$$

for $\mathcal{D}\otimes\mathcal{E}$-measurable positive step functions Φ. Using standard arguments from the theory of measure and integration, we get this equality for positive measurable functions and then for all bounded measurable functions. The latter is, however, equivalent to the first equality in equation (B.1). □

Lemma 9. *For every integer $T \geq 2$,*

$$\sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \leq 2\sqrt{T-1}.$$

Proof. Consider the function $f\colon [1,+\infty) \to (0,+\infty)$ defined by

$$f(x) \coloneqq x^{-1/2}.$$

The function f is positive and strictly decreasing on $[1,+\infty)$, since

$$f'(x) = -\frac{1}{2}x^{-3/2} < 0 \qquad \text{for all } x > 0.$$

We now use a standard comparison between sums and integrals for decreasing functions. Fix an integer $T \geq 2$. For any integer t such that $2 \leq t \leq T-1$ and any $x \in [t-1,t]$, the monotonicity of f implies

$$f(t) \leq f(x).$$

Integrating this inequality over $x \in [t-1,t]$ we obtain

$$\int_{t-1}^{t} f(t)dx \leq \int_{t-1}^{t} f(x)dx.$$

Since the left-hand side is just $f(t)$ (because the interval has length 1), we get

$$f(t) \leq \int_{t-1}^{t} f(x)dx \qquad \text{for all } t = 2,\dots,T-1.$$

Summing these inequalities over $t = 2,\dots,T-1$ yields

$$\sum_{t=2}^{T-1} f(t) \leq \sum_{t=2}^{T-1} \int_{t-1}^{t} f(x)dx = \int_{1}^{T-1} f(x)dx.$$

Therefore

$$\sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} = f(1) + \sum_{t=2}^{T-1} f(t) \leq f(1) + \int_1^{T-1} f(x)dx.$$

We now compute the integral explicitly:

$$\int_1^{T-1} f(x)dx = \int_1^{T-1} x^{-1/2}dx = \left[2x^{1/2}\right]_1^{T-1} = 2\sqrt{T-1} - 2.$$

Moreover, $f(1) = 1$, so we obtain

$$\sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \leq 1 + (2\sqrt{T-1} - 2) = 2\sqrt{T-1} - 1 \leq 2\sqrt{T-1},$$

which proves the claim. □

C

DKW inequality

In this appendix, we present a bivariate DKW inequality which follows directly from the VC-type bound of [20, Theorem 4.9, see also Lemmas 4.4, 4.5, and 4.11 for the explicit constants]

Theorem 7 (Bivariate DKW inequality)**.** *There exist positive constants* $m_0 \leq 1200$*,* $c_1 \leq 13448$*,* $c_2 \geq 1/576$ *such that, if* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space and* $(X_n, Y_n)_{n\in\mathbb{N}}$ *is a* $\mathbb{P}$*-i.i.d. sequence of two-dimensional random vectors, then, for any* $\varepsilon > 0$ *and all* $m \in \mathbb{N}$ *such that* $m \geq m_0/\varepsilon^2$*, it holds*

$$\mathbb{P}\left[\sup_{x,y\in\mathbb{R}}\left|\frac{1}{m}\sum_{k=1}^{m}\mathbb{I}\{X_k \leq x, Y_k \leq y\} - \mathbb{P}[X_1 \leq x, Y_1 \leq y]\right| > \varepsilon\right] \leq c_1 \exp\left(-c_2 m\varepsilon^2\right).$$

D

Tail Integration Formula

In this appendix, we show the well-known tail integration formula. We chose to use the standard machine technique to prove it, even though it might not be the shortest method.

Theorem 8 (Tail-integration formula for nonnegative random variables)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $X\colon \Omega \to [0, \infty]$ *be a nonnegative random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x) dx,$$

Proof. **Step 1:** X **is a simple random variable.**

Let $X\colon \Omega \to [0, \infty)$ be a simple random variable taking finitely many values

$$0 \leq x_1 < x_2 < \cdots < x_n, \qquad \mathbb{P}(X = x_i) = p_i.$$

For $x \in [0, x_1)$ we have $\mathbb{P}(X \geq x) = 1$. For $x \in [x_k, x_{k+1})$ with $k = 1, \ldots, n-1$ we have

$$\mathbb{P}(X \geq x) = \mathbb{P}(X \in \{x_{k+1}, \ldots, x_n\}) = 1 - \sum_{i=1}^{k} p_i,$$

and for $x \geq x_n$ we have $\mathbb{P}(X \geq x) = 0$.

Hence the tail function is a step function and we can split the integral:

$$\begin{aligned}\int_0^\infty \mathbb{P}(X \geq x) dx &= \int_0^{x_1} 1 dx + \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} \left(1 - \sum_{i=1}^{k} p_i\right) dx + \int_{x_n}^\infty 0 dx \\ &= x_1 + \sum_{k=1}^{n-1} \left(1 - \sum_{i=1}^{k} p_i\right) (x_{k+1} - x_k).\end{aligned}$$

Expanding the sum yields a telescoping cancellation:

$$\begin{aligned} x_1 + \sum_{k=1}^{n-1} \left(1 - \sum_{i=1}^{k} p_i\right)(x_{k+1} - x_k) &= x_1 + \sum_{k=1}^{n-1}(x_{k+1} - x_k) - \sum_{k=1}^{n-1}\left(\sum_{i=1}^{k} p_i\right)(x_{k+1} - x_k) \\ &= x_n - \sum_{k=1}^{n-1}\sum_{i=1}^{k} p_i(x_{k+1} - x_k). \end{aligned}$$

Rewrite the double sum by swapping the order (finite sums, so this is purely algebraic):

$$\sum_{k=1}^{n-1}\sum_{i=1}^{k} p_i(x_{k+1} - x_k) = \sum_{i=1}^{n-1} p_i \sum_{k=i}^{n-1}(x_{k+1} - x_k) = \sum_{i=1}^{n-1} p_i(x_n - x_i).$$

Therefore

$$\begin{aligned} \int_0^\infty \mathbb{P}(X \geq x)dx &= x_n - \sum_{i=1}^{n-1} p_i(x_n - x_i) \\ &= x_n\left(1 - \sum_{i=1}^{n-1} p_i\right) + \sum_{i=1}^{n-1} p_i x_i \\ &= p_n x_n + \sum_{i=1}^{n-1} p_i x_i \\ &= \sum_{i=1}^{n} p_i x_i = \mathbb{E}[X]. \end{aligned}$$

Step 2: Approximate X from below by simple random variables. For each $m \in \mathbb{N}$ define the simple random variable

$$X_m(\omega) \coloneqq \frac{1}{2^m} \sum_{k=0}^{m2^m-1} k\mathbb{I}\left\{\frac{k}{2^m} \leq X(\omega) < \frac{k+1}{2^m}\right\} + m\mathbb{I}\left\{X(\omega) \geq m\right\}.$$

Then $0 \leq X_m(\omega) \leq X_{m+1}(\omega) \leq X(\omega)$ for all ω, and $X_m(\omega) \uparrow X(\omega)$ as $m \to \infty$. This is the standard construction of an approximating sequence of simple random variables.

Step 3: Apply the simple-case formula to X_m and use the MCT. Since each X_m is a simple nonnegative random variable, step 1 gives

$$\mathbb{E}[X_m] = \int_0^\infty \mathbb{P}(X_m \geq x)dx.$$

Because $X_m \uparrow X$ and $X_m \geq 0$, the Monotone Convergence Theorem yields

$$\lim_{m\to\infty} \mathbb{E}[X_m] = \mathbb{E}[X].$$

Now, fix $x \geq 0$. Since $X_m(\omega) \uparrow X(\omega)$ pointwise, we have

$$\mathbb{I}\{X_m(\omega) \geq x\} \uparrow \mathbb{I}\{X(\omega) \geq x\} \quad \text{as } m \to \infty.$$

Taking expectations and applying monotone convergence gives

$$\mathbb{P}(X_m \geq x) = \mathbb{E}[\mathbb{I}\{X_m \geq x\}] \uparrow \mathbb{E}[\mathbb{I}\{X \geq x\}] = \mathbb{P}(X \geq x).$$

Thus the functions $f_m(x) \coloneqq \mathbb{P}(X_m \geq x)$ are measurable, nonnegative, and satisfy $f_m(x) \uparrow f(x) \coloneqq \mathbb{P}(X \geq x)$ pointwise for all $x \geq 0$. Applying the Monotone Convergence Theorem

(in the variable x with respect to Lebesgue measure) yields

$$\lim_{m\to\infty} \int_0^\infty \mathbb{P}(X_m \geq x)dx = \int_0^\infty \mathbb{P}(X \geq x)dx.$$

Taking limits in the identity $\mathbb{E}[X_m] = \int_0^\infty \mathbb{P}(X_m \geq x)dx$ proves

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)dx,$$

as claimed. □

E

Total Variation and Pinsker's inequality

In this appendix, we state the definition of total variation between two probability measures, and we state Pinsker's inequality (a formal proof can be found in [21, Lemma 2.5]).

Definition 12 (Total Variation between two probability measures)**.** *Let P and Q be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Their total variation distance is*

$$\|P - Q\|_{\mathrm{TV}} \coloneqq \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.$$

Theorem 9 (Pinsker's inequality)**.** *Let P and Q be two probability distributions on a measurable space $(\Omega, \mathcal{F})$ with $P \ll Q$, then*

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} \mathcal{D}_{\mathrm{KL}}(P \| Q)}$$

F

MOSS (Minimax Optimal Strategy in the Stochastic case)

Let K be the number of arms, $\hat{X}_{i,s}$ be the empirical mean of arm i after s draws of this arm. Let $T_i(t)$ denote the number of times we have drawn arm i in the first t rounds.

Algorithm 3 MOSS (Minimax Optimal Strategy in the Stochastic case)

initialization: fix horizon T and number of arms K;
for each arm $i \in \{1, \ldots, K\}$ set $T_i(0) \leftarrow 0$ and $\widehat{X}_{i,0} \leftarrow 0$;
for t = 1,2,. . . ,T **do**
for each arm i, define the index

$$B_{i,T_i(t-1)} \leftarrow \begin{cases} +\infty, & \text{if } T_i(t-1) = 0, \\ \widehat{X}_{i,T_i(t-1)} + \sqrt{\dfrac{\max\left(\log\left(\dfrac{T}{KT_i(t-1)}\right), 0\right)}{T_i(t-1)}}, & \text{otherwise.} \end{cases}$$

pick an arm $I_t \in \text{argmax}_{i \in \{1,\ldots,K\}} B_{i,T_i(t-1)}$;
pull arm I_t and observe reward X_t;
update $T_{I_t}(t) \leftarrow T_{I_t}(t-1) + 1$ and $T_j(t) \leftarrow T_j(t-1)$ for all $j \neq I_t$;
update the empirical mean of the pulled arm:

$$\widehat{X}_{I_t,T_{I_t}(t)} \leftarrow \frac{(T_{I_t}(t-1))\,\widehat{X}_{I_t,T_{I_t}(t-1)} + X_{I_t,t}}{T_{I_t}(t)}.$$

end for

With this algorithm, the authors propose a policy inspired by the UCB1 policy [22], where each arm has an index measuring its performance, and at each round, they choose the arm having the highest index. The index of an arm that has been drawn more than T/K times is simply the empirical mean of the rewards obtained from the arm. For the other arms, their index is an upper confidence bound on their mean reward, which, from Hoeffding's inequality, holds with high probability.

In their paper, the authors prove the following theorem:

Theorem 10. *The MOSS algorithm satisfies*

$$\sup R_T \leq 49\sqrt{KT}$$

where the supremum is taken over all K-tuples of probability distributions on $[0, 1]$.

The analysis of the theorem can be found in [18, Theorem 5]

Bibliography

1. Myerson, R. B. & Satterthwaite, M. A. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* **29,** 265–281 (1983).
2. Nisan, N., Roughgarden, T., Tardos, E. & Vazirani, V. V. *Algorithmic Game Theory* (Cambridge University Press, 2007).
3. Hartline, J. D. *Bayesian Mechanism Design* (Cambridge University Press, 2022).
4. Blum, A., Hartline, J. D. & Kleinberg, R. D. Online learning in online auctions. *Theoretical Computer Science* **324,** 137–146 (2006).
5. Chawla, S., Hartline, J. D., Malec, D. L. & Sivan, B. *Multi-parameter mechanism design and sequential posted pricing* in *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)* (2010), 311–320.
6. Babaioff, M., Dughmi, S., Kleinberg, R. & Slivkins, A. Bandits with an edge. *SIAM Journal on Computing* **44,** 1447–1475 (2015).
7. Roth, A., Ullman, J. & Wu, Z. S. Bandit learning in repeated auctions. *Advances in Neural Information Processing Systems* **29** (2016).
8. Bachoc, F., Cesa-Bianchi, N., Cesari, T. & Colomboni, R. Fair Online Bilateral Trade. *Proceedings of the 38th Conference on Neural Information Processing Systems.* To appear; see also arXiv:2405.13919 (2024).
9. Dütting, P., Gkatzelis, V. & Roughgarden, T. The Algorithmic Foundations of Differential Pricing. *Journal of the ACM* **70,** 1–52 (2023).
10. Cesa-Bianchi, N. & Lugosi, G. *Prediction, Learning, and Games* (Cambridge University Press, 2006).
11. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, 2014).
12. Cesa-Bianchi, N., Cesari, T., Colomboni, R., Fusco, F. & Leonardi, S. Bilateral Trade: A Regret Minimization Perspective. *Mathematics of Operations Research* **49,** 171–203 (2024).
13. Mitchell, T. M. *Machine Learning* (McGraw-Hill, New York, 1997).
14. Dughmi, S. & Xu, H. Truthful learning mechanisms with experts and bandits. *Games and Economic Behavior* **104,** 624–648 (2017).
15. Rawls, J. *A Theory of Justice* (Harvard University Press, Cambridge, MA, 1971).
16. Cesari, T. & Colomboni, R. *An Online Learning Theory of Trading-Volume Maximization* in *Proceedings of the International Conference on Learning Representations (ICLR)* (2025).

17. Cesari, T. R. & Colomboni, R. A nearest neighbor characterization of Lebesgue points in metric measure spaces. *Mathematical Statistics and Learning* **3,** 71–112 (2021).

18. Audibert, J.-Y. & Bubeck, S. *Minimax Policies for Adversarial and Stochastic Bandits* in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)* (Omnipress, 2009), 217–226.

19. Kleinberg, R. D. & Leighton, F. T. *The Value of Knowing a Demand Curve: Bounds on Regret for Online Posted-Price Auctions* in *44th Symposium on Foundations of Computer Science (FOCS 2003), Cambridge, MA, USA, October 11-14, 2003, Proceedings* (IEEE Computer Society, 2003), 594–605.

20. Anthony, M. & Bartlett, P. L. *Neural Network Learning: Theoretical Foundations* ISBN: 978-0-521-11862-0 (Cambridge University Press, 2009).

21. Tsybakov, A. B. *Introduction to Nonparametric Estimation* ISBN: 978-0-387-79051-0 (Springer, New York, NY, 2009).

22. Auer, P., Cesa-Bianchi, N. & Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* **47,** 235–256 (2002).