Università degli Studi di Milano

Ph.D. Program in Computer Science

(XXXVI Cycle)

Department of Computer Science

A thesis submitted for the degree of

Doctor of Philosophy

Online Learning Methods for Digital Markets

Subject Area: INF/01

Author
Roberto Colomboni

Supervisor
Prof. Nicolò Cesa-Bianchi

Co-Supervisors
Prof. Massimiliano Pontil
Prof. Tommaso R. Cesari

PhD Coordinator
Prof. Roberto Sassi

Academic Year 2022–2023

In memoria di mio nonno Giuseppe

Abstract

We cast several problems arising from digital markets and economics into an online learning framework, where a learner sequentially interacts with an unknown environment, trying to discover its relevant features to maximize her cumulative reward.

After an introduction to online learning in Chapter 1, we start with a study of the bilateral trade problem in Chapter 2. Here, the learner plays the role of a broker whose goal is to increase the value of the market by sequentially interacting with pairs of sellers and buyers, facilitating trades between them. We show how the interplay between the feedback received by the learner and the set of available trading mechanisms affects the attainable regret regimes, devising ad hoc solutions to address the exploration/exploitation dilemma in various types of environments.

In Chapter 3, we present an analysis of transparency in repeated first-price auctions. Here, the learner participates in a sequence of first-price auctions to win objects whose exact value is revealed only when she wins the corresponding auction. We show how the level of transparency of the auctioneer (i.e., the amount of information disclosed at the end of each auction) influences the regret rates in different types of environments.

In Chapter 4, we study the problem of adaptive optimal taxation. Here, the learner plays the role of a policymaker whose goal is to increase social welfare (seen as a weighted sum of private utility and public revenue) by sequentially setting the tax rate in the labor market. Interestingly, once framed in a formal online learning setting, this problem can be seen as a non-trivial generalization of the classical dynamic pricing problem.

Finally, in Chapter 5, we propose an abstract framework generalizing bandits with delayed feedback. This framework allows us to capture scenarios arising frequently in advertising campaigns, where the feedback received comes in the form of delayed and composite income, the sources of which depend on the actions the learner took in the past, and whose exact contributions are not easily identifiable.

Acknowledgments

Non ho mai scritto ringraziamenti sulle tesi di laurea. Intendo porre rimedio a questo grave errore con la tesi di dottorato.

Ci sono molte persone che meritano di essere ringraziate per avermi supportato in questi anni.

Nicoletta, che mi è sempre stata vicina anche nelle difficoltà, incoraggiandomi e credendo in me anche quando ero io il primo a non crederci. Grazie per avermi sopportato, esserci stata, ed esserci.

I miei genitori, Sabrina ed Emanuele, che mi sono sempre stati vicino volendomi un bene immenso, e mi sono stati di sconfinato supporto nell'inseguire i miei desideri. Grazie dal cuore Cocchi.

Mio nonno Giuseppe, che purtroppo non c'è più, ma che con la sua gioia di vivere e il suo incrollabile ottimismo è stato grande fonte di forza e ispirazione. Grazie di tutto Nonno.

Mia nonna Maria, per essere sempre stata vicino ed avere sempre voluto un mondo di bene al suo "nipote preferito" (nonché unico). Grazie Nonna.

La mia famiglia, Anna, Giorgio, Steno, Marina, Elisa, Roberto, Eleonora, Giovanni, Costanza, Vittoria e Ricky. Vedervi tutti insieme mi fa sempre sentire a casa. In famiglia, appunto. Grazie.

I miei amici di sempre: Pablo, Bodo e Lalli. Sempre presenti anche quando (io) sono assente. Un porto sicuro ad ogni tempesta. Grazie per la vostra amicizia.

Il mio amico, co-supervisor e co-autore di default Tom. Grazie per le interessanti discussioni, per le opportunità e la fiducia, per essermi stato da guida, e per tutto il supporto.

Il mio supervisor Nicolò e il mio co-supervisor Massi. Grazie per la guida nel mondo accademico e della ricerca, per la disponibilità, per la libertà di ricerca concessami, e per tutto il supporto.

Co-autori e colleghi, tra cui Federico, Stefano, François, Andrea, Emmanuel, Max e Gugliemo. Grazie per le stimolanti discussioni che hanno contribuito a rendere appassionante il lavoro di ricerca.

List of Publications

Published/Accepted for Publication Papers included in this Thesis.

[55] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 289–309, New York, NY, USA, 2021. Association for Computing Machinery.

[56] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. *Journal of Machine Learning Research (JMLR)*, 23(1-24), 2022

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

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

[40] Natasa Bolić, Tommaso R. Cesari, and Roberto Colomboni. An online learning theory of brokerage. *arXiv preprint arXiv:2310.12107*, 2023. To appear at The 23rd International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2024.

[58] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. *arXiv preprint arXiv:2307.09478*, 2023. To appear at The 56th Annual ACM Symposium on Theory of Computing, STOC 2024.

Manuscripts under Review included in this Thesis.

[59] Nicolò Cesa-Bianchi, Roberto Colomboni, and Maximilian Kasy. Adaptive maximization of social welfare. *arXiv preprint arXiv:2310.09597*, 2023

Published Papers and Manuscripts not included in this Thesis.

[61] Tommaso R. Cesari and Roberto Colomboni. A nearest neighbor characterization of Lebesgue points in metric measure spaces. *Mathematical Statistics and Learning*, 3(1):71–112, 2021.

[22] François Bachoc, Tommaso R. Cesari, Roberto Colomboni, and Andrea Paudice. A near-optimal algorithm for univariate zeroth-order budget convex optimization. *arXiv preprint arXiv:2208.06720*, 2022.

[69] Roberto Colomboni, Emmanuel Esposito, and Andrea Paudice. An improved uniform convergence bound with fat-shattering dimension. *arXiv preprint arXiv:2307.06644*, 2023.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Online Learning Framework

Online learning is a subfield of machine learning that provides a framework for sequential decision-making. In contrast to batch learning, where models are trained on fixed datasets provided in advance, online learning can handle data gathered sequentially not only from stationary but also from dynamic or even adversarial environments. However, this flexibility comes with certain challenges: unlike the complete freedom in batch learning to use and transform data for optimal model performance, an online learner incurs costs for each less-than-optimal decision.

At a high level, the reader can think of online learning as a paradigm to repeatedly interact with an unknown environment in the following manner. At each time step, an agent (or, as it is commonly called, a learner) has to perform an action, selecting it from a pool of possible actions. This selection has to be performed based on the past collected information about the (otherwise unknown) environment. After having performed the action, the learner earns some reward determined by how good the action was in the current state* of the environment, observes some feedback, and then moves to the next interaction. The goal of the learner is to maximize her cumulative reward over a certain time horizon, a process formalized through the concept of *regret minimization*.

Regret measures the difference between two key quantities: the expected cumulative reward of some environment-dependent benchmark strategy, and the actual expected cumulative reward the learner earns through her actions during the learning process. Frequently, the benchmark strategy is the best possible *constant* strategy in the selected environment, which is the one selecting the best fixed action maximizing the expected cumulative reward. Notably, the best possible constant strategy is actually the best strategy in stochastic (i.i.d.) environments, i.e., when there is a fixed distribution according to which the state of the environment is drawn independently at any time.

Under full information (i.e., when the received feedback after each interaction is the reward associated with every action), batch and online learning paradigms share some methods and strategies in stochastic environments. In this setting, for example, the Empirical Risk Minimization batch learning algorithm [165] translates effortlessly into the online learning *Follow-the-Leader* strategy, together with its learning guarantees.

On the other hand, the distinction between online and batch learning paradigms becomes sharp

*In general, the evolution of the states of the environment might or might not be dependent on the actions taken by the learner. In this work, we consider only problems where this evolution is assumed to be *oblivious* to the learner's actions.

as soon as we allow the model to be more complex, as it is highlighted by the following two aspects.

Firstly, online learning measures performance using a cumulative reward function, and hence the costs for all suboptimal actions performed in the learning process sum up. This is in contrast to batch learning, where performance is evaluated based on the model accuracy *after* processing the entire dataset, without accumulating costs for individual decisions.

Secondly, in online learning, different actions may reveal different pieces of information about the unknown environment in which the learner acts. This implies that the information gathered by the learner is usually dependent on her previously performed actions. Again, this is in contrast to batch learning, where the learner has access to the full picture from the beginning.

These two characteristics necessitate that an online learner carefully devises her strategy: engaging in maybe (or even knowingly) suboptimal but informative actions could be an essential part of the learning process to better understand the environment where she acts. However, excessive indulgence in such exploration could be costly, considering that the objective to maximize is the cumulative reward.

The balancing act between gaining knowledge (exploration) and optimizing outcomes (exploitation) is a central challenge in online learning, known as the *Exploration-Exploitation Trade-Off*, and several ideas have been proposed to deal with this dilemma.

For instance, *Explore-then-Commit* strategies (see, e.g., [123]) are often a viable option in stochastic environments. Here, the learner initially allocates a period solely for exploratory purposes. Upon gathering sufficient information to confidently identify a sufficiently good action, the learner then commits to this action for subsequent decisions. While this strategy has its merits, it often falls short in optimizing performance due to the costly initial exploration phase, and the fact that the model ceases to be updated in the subsequent exploitation phase.

To address this problem, more nuanced approaches entangle exploration and exploitation.

Action elimination strategies (see, e.g., [87]) build confidence intervals around the expected rewards associated with each action and stop playing them when the learner discovers that they are suboptimal with high probability. If the set of possible actions is finite, by cycling among the still potentially optimal actions, this approach proves to be effective in the so-called *bandit* problems, i.e., problems where the information gathered during the learning process is precisely the stream of rewards associated with the actions the learner performed.

The entanglement between the exploration and exploitation phases is perhaps better embodied in the principle of *Optimism in the face of Uncertainty*. Here, the learner constructs again confidence intervals around the expected rewards associated with each action. However, at each decision point, the action with the highest upper confidence bound (representing the most optimistic potential outcome) is selected. In bandit problems where the number of actions is finite, this approach leads to the celebrated UCB algorithm and its variants [15, 17, 95, 120].

The strategies and methods previously mentioned are based on a certain level of stationarity in the environment, a premise that is not always valid in practical applications.

Drawing from the principles of convex optimization, *Follow-the-Regularized/Perturbed-Leader* strategies (see, e.g., [1] and references therein) have been developed to address the challenges of non-stationary environments. These strategies involve devising estimates of the cumulative reward function for each action and choosing at each time the one that maximizes a regularized/perturbed version of these estimates, thereby introducing some form of stability into the action selection process.

When the number of actions is $K \in \mathbb{N}$, using the Follow-the-Regularized-Leader strategy with the entropic regularizer on the probability simplex Δ^K leads to the widely used *exponential weights* algorithms. Specifically, in the full-information setting, where there is no need for an estimation procedure, we recover the classic Hedge algorithm [93]. In bandit settings, we recover the celebrated Exp3 algorithm [18] by choosing *importance weighting* as the estimation procedure, i.e., by estimating the instantaneous reward associated with a selected action using the observed reward divided by the probability of choosing that specific action.

These strategies (and many others) have been extensively studied and are now classical topics of the online learning literature, which at this point is rich in excellent books and surveys [46, 48, 97, 106, 123, 146, 164, 169].

On the applications side, the flexibility of the online learning paradigm has led to its widespread adoption across various domains.

One notable area where online learning has demonstrated its effectiveness is in the field of personalized recommendations and advertising. Online platforms, such as e-commerce websites and streaming services, leverage online learning algorithms analyzing user behaviors and preferences in real time to deliver tailored product recommendations and targeted advertisements, aiming for improved user engagement and higher conversion rates. See, e.g., [4, 62, 63, 99, 149–151, 162, 163, 176, 177, 180, 187, 188].

Additionally, online learning has found valuable applications in fraud detection and cybersecurity. Financial institutions and online payment systems can utilize online learning models to detect fraudulent transactions as they occur, enabling swift responses and enhanced security measures. See, e.g., [5, 73, 82, 112, 191].

Online learning's versatility extends also to other domains, including healthcare and clinical trials. For example, by continuously updating models with new patient data, healthcare providers can make timely decisions, predict health risks, and personalize treatment plans more effectively. On the other hand, online learning provides a principled framework for clinical trials, where the exploration (testing new drugs) versus exploitation (treating the current patient with the best drug discovered so far) trade-off is crucial, even from an ethical point of view. See, e.g., [10, 31, 32, 41, 84, 136, 137, 147, 148, 155, 157, 173, 179].

Finally, online learning has revolutionized the landscape of dynamic pricing and auctions. The fluctuation of product prices based on demand makes dynamic pricing a natural field where to apply online learning techniques, with the goal of optimizing prices in real time. E-commerce platforms, ride-hailing services, and airlines use online learning to analyze customer behavior, competitor prices, and other relevant factors to set optimal prices and maximize revenue. Notably, in the realm of auctions, online learning techniques are a fundamental tool for designing automated bidding strategies. More in general, by employing online learning strategies, online marketplaces can optimize bidding decisions, improve auction outcomes, or ensure fair and competitive pricing for buyers and sellers alike. See, e.g., [2, 3, 25, 27, 50, 52, 54, 64, 90, 98, 104, 105, 117, 126, 128, 133, 182, 189, 190].

This (non-exhaustive) list of real-world applications (for more, see, e.g., the survey [42]) highlights the adaptability and practicality of the online learning paradigm, showcasing that its potential for widespread application is arguably limited only by our imagination.

1.2 The Work in this Thesis: Online Learning for Digital Markets

In this work, we focus on problems coming from digital markets and economics, analyzing them through the lens of online learning.

Bilateral Trade. In Chapter 2, which is based on [40, 55, 57, 60], we start with bilateral trade, a classic problem in the mechanism design literature. Bilateral trade is the study of brokerage between a seller and a buyer. They want to trade a good for which they hold private valuations.

Ideally, the role of a broker is to design a trading mechanism where she does not subsidize or drain money from the trade (budget balance), while ensuring that trade happens whenever it should happen, i.e., when the seller's valuation is less than the buyer's one (efficiency). This mechanism should be designed to prevent both sellers and buyers from having strategic reasons to misreport their true valuations (incentive compatibility), while also ensuring they do not lose value in the trade (individual rationality).

Unfortunately, a classical result by Myerson and Satterthwaite [143] states that an efficient mechanism satisfying all these conditions does not exist in general, not even if we weaken most of these assumptions and we know the seller's and buyer's valuation distributions in advance. On the other hand, budget-balanced mechanisms ensuring incentive compatibility and individual rationality do exist, and they are precisely fixed-price mechanisms [67].

We investigate fixed-price mechanisms in bilateral trade, studying them from an online learning perspective where the learner plays the role of the broker.

At each time step, a new seller/buyer pair arrives. Then, the learner selects a fixed-price mechanism, i.e., proposes the same trading price to both the seller and the buyer without asking them for any information. A trade happens if and only if both the seller and the buyer accept the proposed price. The learner's goal is to bound the cumulative loss in efficiency.

We measure the efficiency quantitatively: we consider not only whether a trade happened or not, but also how much we lose by losing a trading opportunity. Specifically, we measure our reward using the *gain from trade*, which is the sum of the seller's and buyer's increase in value after the interaction. Hence, losing a trade opportunity can be cheap or costly, depending on how far the seller's and buyer's valuations are.

As we will see, the quality of the feedback received by the learner after each interaction plays a crucial role in the learning process. Specifically, we consider two different types of feedback: full feedback, where after each interaction the seller and the buyer reveal their actual valuations, and realistic/two-bit feedback, where after each interaction we only observe whether the proposed posted price was accepted or not by each party. We remark that realistic feedback is not even enough to reconstruct a bandit-type of feedback, which is the cause of many learning challenges.

We first analyze how classic online learning strategies (Follow-the-Leader/Hedge) can be used to learn under full feedback in the online bilateral trade problem. Then, in the more challenging realistic feedback setting, where full-information or bandit strategies cannot be directly implemented due to the scarcity of the feedback, we draw inspiration from online learning principles (Explore-then-Commit/exponential weights/unbiased estimation) and problem-specific tools (Decomposition/Representation/Approximation Lemmas) to devise tailored learning strategies to achieve optimal regret rates.

We conclude by providing an analysis of online learning in bilateral trade when the learner is allowed to relax the budget balance assumption: if the learner still cannot subsidize the trade, but is allowed to extract money from the trade by posting two different prices to the seller and the buyer, we demonstrate that learnability can be achieved in a larger set of environments.

First-Price Auctions. In Chapter 3, which is based on [58], we move to analyze repeated first-price auctions, which are of increasing importance due to the recent shift from second to first-auctions in the online advertising market [171, 186].

Here, we study the problem of a learner participating in a sequence of first-price auctions to increase her revenue by winning the auctioned objects paying them less than their actual value. We work under the assumption that the learner discovers the value of a certain object only when she wins the corresponding auction, a sensible assumption in a variety of scenarios arising, e.g., in the online advertising market, where click and conversion rates can be measured only after the auction is won and the ad displayed.

While this setting has been previously investigated from an online learning perspective [2, 90], we are the first to provide a systematic analysis of how the level of transparency of the auctioneer (i.e., the amount of information disclosed at the end of each auction) influences the attainable regret regimes, and we provide this analysis in combination with a variety of different assumptions about the underlying environment where the learner acts.

We stress that the level of transparency influences the feedback structure, with a natural and deep connection with feedback graphs [8]. By exploiting feedback graph ideas together with problem-specific considerations and techniques (e.g., adaptive grids), we devise algorithms providing optimal regret rates in each case.

The Optimal Taxation Problem. In Chapter 4, which is based on [59], we proceed by studying the problem of optimal taxation. Here, a policymaker aims at maximizing social welfare, defined as a weighted sum of private utility and public revenue. The social welfare weights are chosen by the policymaker, and regulate how much she values private wealth compared to public redistribution. The policymaker interacts with the environment (e.g., the labor market) by setting some policy parameters (e.g., the tax rate).

The common method in public finance to address the optimal taxation problem involves utilizing past data to determine the necessary parameters, which are subsequently inserted into equations for optimal policy selection based on theoretical models.

In contrast, we address the optimal taxation problem from an online learning perspective, where the learner plays the role of the policymaker and has to discover the relevant features along the way.

At each time step, the learner interacts with a new individual and sets a corresponding tax rate. Then, the learner observes only whether or not the individual participated in the labor market given the proposed tax rate.

Interestingly, when private welfare carries no weight in the social welfare definition, the online learning optimal taxation problem shares the same online learning structure of the classical dynamic pricing problem [117]. However, when the private welfare weight is non-zero, the problem turns out to be harder to analyze. This is due to the fact that, differently from the dynamic pricing problem, the feedback provided in the optimal taxation problem is not enough to reconstruct the reward

associated with the corresponding performed action. From this perspective, the online learning optimal taxation problem can be seen as a non-trivial generalization of the dynamic pricing problem.

Borrowing ideas from partial monitoring [49], we devise an exponential weights strategy that achieves optimal performance in adversarial and stochastic environments. Furthermore, in stochastic environments where we can assume that the expected welfare is concave, we improve on previous guarantees by devising an optimal action elimination strategy.

Nonstochastic Bandits with Composite Anonymous Feedback. In Chapter 5, which is based on [56], we conclude by proposing an abstract framework generalizing bandit problems with delayed feedback.

The *composite anonymous feedback* framework captures scenarios arising frequently, e.g., in advertising campaigns, where the campaign manager faces the challenge of distinguishing the impact of individual ads distributed across different channels from the total shift in sales.

In this framework, at each time step, each action is associated with a certain reward. However, this reward is spread across d successive rounds. Hence, at each time step, the learner observes a composite reward that is the sum of partial earnings from the last d performed actions. Furthermore, we assume that both the rewards and the way in which the earnings are spread across the successive d rounds can be chosen adversarially.

We devise a wrapper that converts algorithms for the nonstochastic bandit problem (with no delays) into algorithms operating in the composite anonymous feedback setting. Then, we show that the regret guarantees of this wrapper can be bounded in terms of the regret guarantees of the base bandit algorithm and its *stability*.

We then demonstrate that the Follow-The-Regularized-Leader algorithm, coupled with the Tsallis entropy regularizer (and importance weighting as the estimation procedure), enjoys nice stability properties. Together with its optimal regret guarantees, the use of this algorithm as the base algorithm in our wrapper ensures optimal regret guarantees for the nonstochastic composite anonymous feedback setting.

Lower Bound Techniques. We conclude this introduction with some comments about the various lower-bound constructions spread along this work. Each online learning problem we tackle has its own specific structure and several challenges arise when it comes to proving that a proposed algorithm is indeed optimal for a certain setting. That said, there are several recurring high-level ideas when devising these lower bounds.

For example, in designing linear lower bounds we repeatedly resort to the idea of finding *a needle in a haystack*, which occurs when the given feedback is not enough to detect the optimal action among other infinite possibly optimal actions in due time. Another idea for linear lower bounds is exploiting a *lack of observability* phenomenon, which occurs whenever there are two environments such that each action presents the same (action-dependent) feedback distribution in both of them, but the optimal actions in the two environments are different.

For sublinear lower bounds, we take inspiration from lower-bound constructions in other partial monitoring or feedback graph problems (e.g., expert or bandit problems, the revealing action problem, or the multi-apple tasting problem). However, we remark that recognizing these structures inside our problems, where we cannot control rewards and feedback directly (but only indirectly by carefully

devising the underlying environment), is a task rich with challenges, and requires the developing of several information theoretical methods (e.g., the Embedding and Simulations lemmas, or the one-bit/two-environments inverse-transformation representability result) to prove reductions to the aforementioned online learning problems formally.

Chapter 2

Online Learning in Bilateral Trade

2.1 Introduction

In the bilateral trade problem, two strategic agents —a seller and a buyer— wish to trade a good. They both privately hold a personal valuation for it and strive to maximize their respective quasi-linear utility. The burden of designing a mechanism to reach an agreement is usually delegated to a third party. This scenario arises naturally in brokerage in over-the-counter (OTC) markets,* where the role of the broker is to ensure that trades are executed smoothly in absence of a centralized organism, and in many internet applications, such as ridesharing systems like Uber or Lyft where trades between sellers (drivers) and buyers (riders) are managed by a mechanism designed by the platform.

In general, an ideal mechanism for the bilateral trade problem would optimize the efficiency, i.e., the social welfare resulting from trading the item, while enforcing incentive compatibility (IC) and individual rationality (IR). The assumption that makes two-sided mechanism design more complex than the one-sided counterpart is budget balance (BB): the mechanism cannot subsidize or make a profit from the market.

Unfortunately, as Vickrey observed in his seminal work [178], the optimal incentive-compatible mechanism maximizing social welfare for bilateral trade may not be budget-balanced. A more general result due to Myerson and Satterthwaite [143] shows that a fully efficient mechanism for bilateral trade that satisfies IC, IR, and BB may not exist at all. This impossibility result holds even if prior information on the buyer and seller's valuations is available, the truthful notion is relaxed to Bayesian incentive compatibility (BIC), and the exact budget balance constraint is loosened to weak budget balance (WBB).

To circumvent this obstacle, a long line of research has focused on designing approximating mechanisms that satisfy the above requirements while being nearly efficient. These approximation results build on a Bayesian assumption: seller and buyer's valuations are drawn from two distributions known to the mechanism designer. The drawback is that, while in some sense necessary —without any information on the priors there is no way to extract any meaningful approximation result [86]— this assumption is unrealistic in practice.

*OTC markets are decentralized alternatives to traditional financial exchanges that are an indispensable part of the global financial ecosystem: in the US, the value of assets traded in OTC markets surpassed a staggering 50,000 billion USD, exceeding centralized markets by over 20,000 billion USD in 2020 [183], with a steady growth trend documented since 2016 [92].

Online Pricing Protocol for Bilateral Trade

for time $t = 1, 2, \dots$ **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 $\mathrm{GFT}_t(P_t) \in [0,1]$
The learner observes some feedback Z_t

In this work, we focus on fixed-price mechanisms, a class of particular importance in bilateral trade because, on the one hand, they are the *only* direct revelation mechanisms that are IC, IR, and BB [67], and on the other hand, they enjoy the desirable features of being simple to implement and of asking the agents for very little information.

Inspired by a recent line of research [50, 72, 107, 129], we study fixed-price mechanisms in a regret minimization setting, with the aim of bounding the total loss in efficiency.

At each time step t, a seller and a buyer arrive with privately held valuations: $S_t \in [0,1]$ for the seller and $B_t \in [0,1]$ for the buyer. Then, the learner posts a price $P_t \in [0,1]$ and a trade occurs if and only if both the seller and the buyer are satisfied with the proposed price, i.e., $S_t \leqslant P_t \leqslant B_t$. The efficiency of the learner is measured by the increase in utility of the system, the so-called gain from trade. Specifically, defining the gain-from-trade function as

$$\mathrm{gft}\colon [0,1] \times [0,1]^2 \to [0,1], \qquad \big(p, (s,b)\big) \mapsto (b-s) \cdot \mathbb{I}\{s \leqslant p \leqslant b\}$$

and, for any time t, the gain from trade at time t as

$$\mathrm{GFT}_t\colon [0,1] \to [0,1], \qquad p \mapsto \mathrm{gft}\big(p, (S_t, B_t)\big)\ ,$$

the gain from trade of the market at time t if the learner posts $P_t \in [0,1]$ is defined as

$$\Big(\underbrace{B_t - P_t}_{\text{buyer's net gain}} + \underbrace{P_t - S_t}_{\text{seller's net gain}} \Big) \cdot \underbrace{\mathbb{I}\{S_t \leqslant P_t \leqslant B_t\}}_{\text{whenever a trade happens}} = (B_t - S_t) \cdot \mathbb{I}\{S_t \leqslant P_t \leqslant B_t\} = \mathrm{GFT}_t(P_t)\ .$$

After each interaction, instead of observing directly the gain from trade from having posted P_t, the learner has only access to some feedback Z_t. The nature of the sequence of valuation pairs $(S_1, B_1), (S_2, B_2), \dots$ and feedback $Z_1, Z_2, \dots$ depends on the specific instance of the problem and is described below.

Selecting the gain from trade as the target reward function,[†] the *regret* at time horizon T of a learner following a strategy α to generate the sequence of prices P_t (as in the Learning Protocol) against an environment β generating the sequence of (random) pairs (S_t, B_t) is defined by

$$R_T(\alpha, \beta) \coloneqq \max_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p) - \sum_{t=1}^{T} \mathrm{GFT}_t(P_t) \right],^{\ddagger}$$

[†]Another well-studied quantity considered in the bilateral trade literature is social welfare. It is worthwhile noticing that had we chosen social welfare $\mathrm{SW}_t(p) \coloneqq S_t + (B_t - S_t)\mathbb{I}\{S_t \leqslant p \leqslant B_t\}$ instead of the gain from trade nothing would have changed in the regret definition. In fact, since $\mathrm{SW}_t(p) = S_t + \mathrm{GFT}_t(p)$, the term S_t would have appeared twice as an additive term, with opposite signs, and hence canceled out. The reason why we chose the gain from trade is that we believe it provides a more transparent presentation in the following discussion.

[‡]a proof of the fact that this maximum is actually achieved can be found in Appendix A.1.

where the expectation is taken with respect to any randomness present in the environment and (possibly) the internal randomization used by the learner's strategy.

Notice that the regret is the difference between the expected total performance of the learner's strategy, who can only learn *sequentially* about the environment characteristics, and the expected performance of a reference benchmark $p^\star \in [0,1]$, corresponding to the best *constant* fixed-price strategy operating with *full knowledge* about the distribution governing the environment.

The goal of the learner is to determine a strategy for achieving sublinear regret in the time horizon T, *uniformly* with respect to any environment belonging to a certain class of interest. Specifically, we aim to upper bound the regret $R_T^{\mathcal{S}}(\alpha)$ of a learning strategy α in a class of environments $\mathcal{S}$, which is defined as the supremum over all environments $\beta \in \mathcal{S}$ of $R_T(\alpha, \beta)$. A lower bound on the achievable guarantees for any learner operating in a class of environment $\mathcal{S}$ is provided by the minimax regret $R_T^{\mathcal{S}}$, which is defined as the infimum over all learning strategies α of $R_T^{\mathcal{S}}(\alpha)$.

To complete the description of the problem, we need to specify the feedback obtained by the mechanism after each sequential round and the characteristics of the environment in which the learner has to operate.

Environment. We assume that the environment is *oblivious* to the learner, and we model the possible different classes of environments $\mathcal{S}$ by considering several generation models for the $[0,1]^2$-valued stochastic sequence of seller/buyer valuations $(S_t, B_t)_{t\in\mathbb{N}}$.

- **Adversarial (adv):** $(S_t, B_t)_{t\in\mathbb{N}}$ could be any deterministic sequence $(s_t, b_t)_{t\in\mathbb{N}}$.
- **Independent valuations (iv):** $(S_t, B_t)_{t\in\mathbb{N}}$ could be any stochastic sequence such that, for each $t \in \mathbb{N}$, the random variables S_t and B_t are independent of each other.
- **Bounded density (bd):** For some fixed constant $M > 0$, $(S_t, B_t)_{t\in\mathbb{N}}$ could be any stochastic sequence such that, for each $t \in \mathbb{N}$, the random pair (S_t, B_t) admits a joint density (with respect to the Lebesgue measure on $[0,1]^2$) bounded by M.
- **Independently and identically distributed (iid):** $(S_t, B_t)_{t\in\mathbb{N}}$ could be any i.i.d. sequence.

We analyze how (the various combinations of) the previous assumptions influence the regret regimes.

Feedback models. Crucial in casting the learning problem is the specification of the feedback Z_t that the platform receives after posting a price at time t. We consider the following two models.

- *Full feedback.* In the full-feedback model, the pair (S_t, B_t) is revealed as Z_t to the mechanism after the t-th trading round. The information collected by this feedback model corresponds to *direct revelation mechanisms*, where the agents publicly declare their valuations in each round, but the price proposed by the mechanism at time t only depends on past bids.
- *Realistic feedback (Two-bits feedback).* In the more challenging realistic-feedback model, only the relative order between S_t and P_t and between B_t and P_t are revealed after the t-th round: the feedback Z_t received at time t is the pair $\big(\mathbb{I}\{S_t \leqslant P_t\}, \mathbb{I}\{P_t \leqslant B_t\}\big)$. This model corresponds to *posted-price mechanisms*, where seller and buyer separately accept or refuse the posted price. The price computed at time t only depends on past bids, and the values S_t and B_t are *never* revealed to the mechanism.

	adv	iv	bd	iid	iv+bd	iv+iid	bd+iid	iv+bd+iid
Full	T	T	$T^{1/2}$	$T^{1/2}$	$T^{1/2}$	$T^{1/2}$	$T^{1/2}$	$T^{1/2}$
Realistic	T	T	T	T	T	T	T	$T^{2/3}$

Table 2.1: Summary of the regret regimes for fixed-price mechanisms. The rates are both upper and lower bounds (up to logarithmic factors).

2.1.1 Overview of our Results

In Section 2.2, we investigate the class of adversarial environments, while in Sections 2.3 and 2.4 we explore the various classes of environments arising from the combination of the (iv), (bd), and (iid) assumptions, showing how regret bounds change depending on the quality of the received feedback. In all cases, we provide matching upper and lower bounds in the time horizon (up to logarithmic factors). In particular, our positive results are constructive: explicit algorithms are given in each case. For a summary of the obtained regret regimes, see Table 2.1.

In Section 2.5, we show how to improve significantly the regret rates ($\sqrt{T}$ to $\log(T)$ in the full-feedback case, and $T^{2/3}$ to $\sqrt{T}$ in the realistic feedback case) when, on top of the (iv), (bd) and (iid) assumptions, we also assume that sellers' and buyers' valuations are identically distributed, which is a case of particular interest in certain brokerage scenarios where sellers' and buyers' roles are not strictly defined.

Finally, in Section 2.6, we depart from the budget balance setting in which the learner posts the same price to both the seller and the buyer. By considering a weak budget balance (WBB) setting in which (possibly) distinct prices $p \leqslant q$ can be posted, p to the seller, and q to the buyer, we show that we can break the linear lower bound under the (bd) assumption, achieving a $T^{3/4}$ regret rate. Surprisingly, this rate is tight in the time horizon (up to logarithmic factors), even if both the (bd) and (iid) assumptions hold.

2.1.2 Technical Challenges

In this section, we sum up the technical challenges for various instances of our problem.

Adversarial setting. When the valuations of the buyer and the seller form an arbitrary deterministic process generated by an oblivious adversary, learning is impossible. Indeed, using a construction vaguely inspired by the Cantor ternary set, we show that even when the learner receives full feedback, no strategy can lead to a sublinear worst-case regret (Theorem 1).

Full feedback. The full-feedback model fits nicely in the learning with expert advice framework [48]. Each price $p \in [0,1]$ can be viewed as an expert, and the revelation of S_t and B_t allows the mechanism to compute $\mathrm{GFT}_t(p)$ for all p, including the mechanism's own reward $\mathrm{GFT}_t(P_t)$. This unlocks several possibilities to attack the problem, e.g., exponential weights (Hedge) or Follow-the-Leader strategies.

Existing analyses for continuous versions of Hedge assume reward functions are Lipschitz [119, 130]. Unfortunately, the reward function $\mathrm{GFT}_t(\cdot)$ is not (even one-sided) Lipschitz, nor continuous (except for trivial cases). We get around this roadblock by leveraging the bounded density assumption to guarantee the Lipschitzness of the *expected* reward function $\mathbb{E}[\mathrm{GFT}_t(\cdot)]$ (Lemma 1). Then, we prove that having reward functions that are Lipschitz in expectation is enough to obtain $\widetilde{O}(\sqrt{T})$

regret guarantees for the continuous version of Hedge (Corollary 3). This seemingly small difference (Lipschitz vs Lipschitz in expectation) entails a significant technical issue in the analysis that we bypass by proving a log-exp analogous of Minkowski's integral inequality (Lemma 16), which we believe is a result of independent interest.

The Follow-the-Leader approach proves to be effective in the case where the pairs of seller and buyer's valuations form an independently and identically distributed sequence. Here, the full feedback received in each new round is used to refine the estimate of the expected gain from trade as a function of the price, while the posted prices are chosen so as to maximize this estimate. The Decomposition Lemma (Lemma 2) allows us to exploit the structure of the reward function $\mathbb{E}\big[\mathrm{GFT}_t(\cdot)\big]$ by leveraging uniform concentration inequalities to obtain a better regret bound (by a log factor, Theorem 3) with respect to the bounded-density case, even when the underlying distribution does not admit a density.

The main challenge in designing the lower bound is that the shape of the (expected) gain from trade cannot be chosen arbitrarily: we can only control it indirectly as a function of the seller/buyer pair distribution. By designing a suitable family of such distributions, we build a reduction showing that the full-feedback bilateral trade problem when the environment satisfies the (iv), (bd), and (iid) assumptions is harder than a corresponding 2-action partial monitoring game with a known $\Omega\big(\sqrt{T}\big)$ lower bound (Theorem 4).

Realistic Feedback. Here, at the end of time t, only $\mathbb{I}\{S_t \leqslant P_t\}$ and $\mathbb{I}\{P_t \leqslant B_t\}$ are revealed to the learner. In contrast to the full-feedback model, this is not enough to reconstruct the gain from trade GFT_t at time t: if the trade does not occur, it is unclear which prices would have resulted in a trade. Moreover, in contrast to bandit problems [48], this feedback is not even enough to determine $\mathrm{GFT}_t(P_t)$: if the trade occurs, there is no way to infer the difference $B_t - S_t$. Thus, we cannot directly rely on known bandits tools to tackle the two competing goals of estimating the underlying distributions (exploration) while optimizing the estimated gain from trade (exploitation). Instead, using the Decomposition Lemma (Lemma 2), we show how to decompose the expected gain from trade at any price p into a *global* part that can be uniformly estimated via a Monte Carlo method by sampling on the $[0, 1]$ interval, and a *local* part that can be learned by posting p. Theorem 5 shows that our Algorithm 3 (Scouting Bandits) can take advantage of this decomposition by relying on any bandit algorithm to learn the local part of the expected gain from trade. We derive a sublinear regret of $O(T^{2/3})$ whenever the environment satisfies the (iv), (bd), and (iid) assumptions.

The lower bounds present challenges similar to those of the full-feedback model, with additional hurdles due to the specific nature of the realistic feedback. When only realistic feedback is available and the environment satisfies the (iv), (bd), and (iid) assumptions, by designing a suitable family of distributions, we build a reduction showing that this bilateral trade problem is harder than a corresponding instance of the so-called *revealing action* partial monitoring game [48], with a known $\Omega\big(T^{2/3}\big)$ lower bound (Theorem 6). Dropping the (iv) or (iid) assumption leads to a pathological *lack of observability* phenomenon, in which it is impossible to distinguish between two scenarios with significantly different optimal prices (Theorems 7 and 9). Dropping the (bd) assumption amounts to finding a *needle in a haystack*, a different pathological phenomenon in which all prices but one suffer a high regret, and it is essentially impossible to detect this optimal price among a continuum of suboptimal prices (Theorem 8).

Faster rates when sellers and buyers share the same distribution. Central to achieving faster rates are the Approximation and Representation Lemmas (Lemmas 3 and 4). Together, they establish that if the seller and buyer's valuations are independent of each other and share the same distribution with a bounded density, the corresponding expected gain from trade is maximized at the (common) expectation of their valuations. As a consequence, when the environment satisfies the (iv), (bd), and (iid) assumptions, if sellers and buyers share the same distribution, the learner might try to follow the strategy of posting prices that are believed to be good approximations of the expected seller/buyer valuations. This can be done directly in the full-feedback case by posting the empirical mean of the observed past seller/buyer valuations. Instead, in the realistic-feedback case, due to the scarcity of available information, the learner faces an exploration/exploitation dilemma. A viable approach is to try an Exploit-then-Commit strategy, spending a certain period trying to estimate the expectation of the seller/buyer valuations, then commit to the obtained estimation when it is believed to be good enough. Despite their simplicity, both strategies are extremely effective in their respective settings, unlocking significantly better regret guarantees than the ones obtainable without assuming that sellers and buyers share the same distribution. Specifically, $O(\log(T))$ vs $O(\sqrt{T})$ in the full-feedback case (Theorem 11), and $O(\sqrt{T})$ vs $O(T^{2/3})$ in the realistic-feedback case (Theorem 14). These guarantees can be proven again by leveraging the Approximation and Representation Lemmas, which together imply that, by posting a certain price p, the learner regrets no more than a quantity proportional to the *square* of the distance of p from the (common) expected seller/buyer valuation.

We build a single family of hard distributions to show that the guarantees provided by the previous two strategies are optimal in their respective settings (up to constant terms). In the full feedback case, this family is used to show that the problem is harder than a full-feedback sequential Bayesian problem where the goal is to estimate the expectation of a certain random variable, and the loss function is the square of the distance from the expectation (Theorem 12). In the realistic feedback case, the same family is used to mimic a revealing action problem. Here, we obtain a $\Omega(\sqrt{T})$ (instead of a $\Omega(T^{2/3})$) regret lower bound due to the fact that, by posting a certain price p, the learner pays only order of the square of the distance of p from the actual optimum (Theorem 15).

Breaking Linear Lower Bounds in the Realistic Case. If the learner is allowed to post two different prices at each interaction, say p to the seller and q to the buyer, with the constraint $p \leqslant q$ to forbid subsidizing the market, we show that the (bd) assumption is enough to learn in the realistic-feedback case. Again, this result relies on the fact that the (bd) assumption implies the Lipschitzness of the expected gain from trade (Lemma 1). By discretizing the action space $[0, 1]$ and leveraging again the Decomposition Lemma (Lemma 2), we devise an exponential weight algorithm (Algorithm 6) enjoying $\widetilde{O}(T^{3/4})$ regret guarantees (Theorem 16).

We prove that this rate is optimal (up to logarithmic factors) in the time horizon, even adding the (iid) assumption (Theorem 17), by showing that the bilateral trade with partial feedback contains instances that are closely related to instances of online learning with feedback graphs [8]. The corresponding feedback graph G_K is over $2K$ actions: K of them are "exploring" and the others are "exploiting". Exploring actions are costly and reveal feedback on the corresponding exploiting actions. One of the exploiting actions is optimal, but none of them returns any feedback. We build "hard" instances so that any algorithm is forced to spend a long time playing each one of the many

exploring actions in order to learn which one of the exploiting actions is the actual optimal action. By selecting optimally the number of arms in the reduction and the difference in reward between exploiting actions, we obtain the $T^{3/4}$ rate. This proof sketch hides many technical challenges: crucially, we need to carefully design distributions with bounded density with the desired properties. This presents two problems: on the one hand, the gain from trade achievable at different prices are related (while in usual lower bound constructions for online learning with feedback graphs, the rewards can be chosen independently, [8]); on the other hand, the embedding needs to preserve the feedback structure, which is significantly different from the standard bandit or expert feedback and requires subtle arguments. To address this second challenge, we prove a general information-theoretic result (Theorem 44, in Appendix A.15) that may be of independent interest for further lower-bound constructions in related problems.

Lower Bound Techniques. Due to their technical nature, most proofs of the lower bounds are only sketched in the main text. Detailed versions are provided in the Appendix, where we also present a general partial monitoring framework for sequential games (Appendix A.5). Within this setting, we build reductions by mapping instances of our bilateral trade problem to other known partial monitoring games. These reductions rely on two key lemmas, introduced in Appendix A.6: our Embedding and Simulation Lemmas (Lemmas 19 and 20) are useful tools to manipulate rewards and feedback, allowing to build chains of progressively easier games leading to games with known minimax regrets.

Relation with Dynamic Pricing and Auctions.

Before moving on, we spend some words highlighting the differences between the one-sided dynamic pricing problem (see, e.g., [117]) and ours (in its realistic-feedback version). In the former, the learner posts a price p to a buyer with valuation b, receives the bit $\mathbb{I}\{p \leqslant b\}$, and extracts revenue $p \cdot \mathbb{I}\{p \leqslant b\}$. In our case, the learner posts price p, receives feedback $(\mathbb{I}\{s \leqslant p\}, \mathbb{I}\{p \leqslant b\})$ and obtain gain from trade $(b - s)\mathbb{I}\{s \leqslant p \leqslant b\}$. While the structures of the two feedback models share some similarities (in particular the considerations relative to the buyers in the two scenarios are exactly equivalent), the objectives are extremely different. In particular, the one-sided problem is easier than a bandit problem: if the trade happens then the learner gets the price it posts, otherwise, it gets nothing. On the other hand, our problem is harder: if the agents accept a price, let's say $1/2$, then the learner has no indication of the relative gain from trade, which could range from 1 ($s = 0, b = 1$) to 0 ($s = b = 1/2$).

2.1.3 Further Related Work

The study of the bilateral trade problem dates back to the already mentioned seminal works of Vickrey [178] and Myerson and Satterthwaite [143]. A more recent line of research focuses on Bayesian mechanisms that achieve the IC, BB, and IR requirements while approximating the optimal social welfare or the gain form trade. Blumrosen and Dobzinski [38] proposed the *median mechanism* that sets a posted price equal to the median of the seller distribution and shows that this mechanism obtains an approximation factor of 2 to the optimal social welfare. Subsequent work by the same authors [39] improved the approximation guarantee to $e/(e-1)$ through a randomized mechanism

whose prices depend on the seller distribution in a more intricate way. Kang et al. [111] recently showed that it is possible in general to strictly improve on the $e/(e-1)$ guarantee, but that such bound is tight when the mechanism only knows the seller's distribution. In Colini-Baldeschi et al. [67] it is demonstrated that all IC mechanisms that are BB and IR must post a fixed price to the buyer and to the seller. The same result has been previously proven under stronger assumptions in [100]. Recently, Braun and Kesselheim [43] have used tools from the prophet inequality literature to tackle welfare maximization in two-sided markets, i.e., the natural generalization of bilateral trade where multiple sellers intend to trade with multiple buyers.

In a different research direction aimed to characterize the information theoretical requirements of two-sided market mechanisms, Dütting et al. [86] prove that setting the price equal to a single sample from the seller distribution gives a 2-approximation to the optimal social welfare; the same mechanism is shown to yield a 4/3-approximation when seller and buyer share the same distribution [111]. In a parallel line of work, the harder objective of approximating the *gain from trade* has been considered. An asymptotically tight fixed-price $O\big(\log \frac{1}{r}\big)$ approximation bound is also achieved in [68], with r being the probability that a trade happens (i.e., the value of the buyer is higher than the value of the seller). A BIC 2-approximation of the second best with a simple mechanism is obtained in [45]. Very recently, Deng et al. [79] have presented the first (BIC) constant factor approximation to the first best. Their analysis has then been tightened by Fei [88].

In the following, we discuss the relationship between the approximation results mentioned above and the regret analysis we develop in this work that compares online learning mechanisms against the best ex-ante fixed-price mechanism. First of all, in the realistic feedback setting, the approximation mechanisms for bilateral trade cannot be easily implemented. For example, the single sample 2-approximation to the optimal social welfare [86] requires multiple rounds of interaction in order to obtain, approximately, a random sample from the distribution. The median mechanism of Blumrosen and Dobzinski [38] requires an even larger number of rounds in order to estimate the median of the seller distribution. Furthermore, here we note that these two more demanding approaches may yield worse performances than the best ex-ante fixed price[§]. This implies that there are instances where our online learning approach converges to a mechanism that is strictly better than the median or sample mechanisms, even assuming they have full knowledge of the underlying distributions.

There is a vast body of literature on regret analysis in (one-sided) dynamic pricing and online posted price auctions — see, e.g., the excellent survey published by den Boer [76] and the tutorial slides by Slivkins and Zeevi [170]. In their seminal paper, Kleinberg and Leighton prove a $O(T^{2/3})$ upper bound (ignoring logarithmic factors) on the regret in the adversarial setting [117]. Later works show simultaneous multiplicative and additive bounds on the regret when prices have range $[1, h]$ [36, 37]. These bounds have the form $\varepsilon\, G_T^\star + O\big((h \ln h)/\varepsilon^2\big)$ ignoring $\ln \ln h$ factors, where $G_T^\star$ is the total revenue of the optimal price $p^\star$. Recent improvements on these results prove that the additive term can be made $O(p^\star \big(\ln h)/\varepsilon^2\big)$, where the linear scaling is now with respect to the optimal price rather than the maximum price h [47]. Other variants consider settings in which the number of copies of the item to sell is limited [7, 21, 24], buyers act strategically in order to maximize their utility in future rounds [11, 80, 83, 142], or there are features associated with the goods on sale [66]. In the stochastic setting, previous works typically assume parametric [44], locally smooth [117],

[§]Consider a seller with value $\varepsilon > 0$ or 0 with equal probability and a buyer with value 1. The best fixed price has welfare of 1. For small ε, the median and the sample mechanism, respectively, obtains a welfare close to 1/2 and 3/4.

or piecewise constant demand curves [54, 77]. For further related work on this literature, see also Section 3.1.3.

Finally, in recent follow-up work, Azar et al. [20] study the sequential bilateral trade problem in the adversarial setting. In particular, they generalize Theorem 1 to hold for α-regret for any $\alpha \in [1, 2)$ and show that sublinear 2-regret is achievable even with realistic feedback.

2.2 The Adversarial Setting

In this section, we prove that, even in the full-feedback case, no strategy can achieve worst-case sublinear regret in an adversarial environment. The idea of the proof is to build, for any strategy, a *hard* sequence of sellers and buyers' valuations $(s_1, b_1), (s_2, b_2), \dots$ which causes the learner to suffer linear regret for any horizon T. This sequence is built in a way that, at each time step, the achievable gain from trade is approximately $\frac{1}{2}$, the probability that the learner's strategy misses the corresponding trading opportunity is at least $\frac{1}{2}$ and, finally, that there exists a fixed price $p^\star \in [0, 1]$ that allows the trades between sellers and buyers at every time step.

Theorem 1. *In the full-feedback adversarial (adv) setting where the class of environments $\mathcal{S}$ is represented by all deterministic sequences $(s_1, b_1), (s_2, b_2), \dots \in [0, 1]^2$ of sellers' and buyers' valuations, the minimax regret $R_T^{\mathcal{S}}$ satisfies*

$$R_T^{\mathcal{S}} \geqslant cT \ ,$$

where $c \geqslant 1/4$.

Proof. We begin by fixing any strategy α of the learner. This is a sequence of functions $(\alpha_t)_{t\in\mathbb{N}}$, such that, for each t, α_t maps the past feedback $(s_1, b_1), \dots, (s_{t-1}, b_{t-1})$, together with some internal randomization, to the price P_t to be posted by the learner at time t. In other words, the strategy maintains a distribution ν_t over the prices that is updated after observing each new pair (s_t, b_t) and used to draw each new price P_t. We will show how to constructively determine a sequence of seller/buyer valuations that is hard for α to learn. This sequence is oblivious to the prices $P_1, P_2, \dots$ posted by α, in the sense it does not have access to the realizations of its internal randomization. The idea is, at any time t, to determine a seller/buyer pair (s_t, b_t) either of the form $(c_t, 1)$ or $(0, d_t)$, with $c_t \approx \frac{1}{2} \approx d_t$, such that the probability ν_t that the strategy picks a price $P_t \in [s_t, b_t]$ (i.e., that there is a trade) is at most $1/2$ and, at the same time, there is common price $p^\star$ which belongs to $[s_t, b_t]$ for all times t.[¶] This way, since $b_t - s_t \approx \frac{1}{2}$ for all t, the regret of α with respect to $(s_1, b_1), (s_2, b_2), \dots$ is at least (approximately) greater than or equal to $T/4$.

The construction proceeds inductively as follows. Let $\varepsilon \in \left(0, \frac{1}{18}\right)$. Let

$$\begin{cases} c_1 \coloneqq \frac{1}{2} - \frac{3}{2}\varepsilon,\ d_1 \coloneqq \frac{1}{2} - \frac{1}{2}\varepsilon,\ s_1 \coloneqq 0,\ b_1 \coloneqq d_1, & \text{if } \nu_1\left[\left[0, \frac{1}{2} - \frac{1}{2}\varepsilon\right]\right] \leqslant \frac{1}{2} \ , \\ c_1 \coloneqq \frac{1}{2} + \frac{1}{2}\varepsilon,\ d_1 \coloneqq \frac{1}{2} + \frac{3}{2}\varepsilon,\ s_1 \coloneqq c_1,\ b_1 \coloneqq 1, & \text{otherwise.} \end{cases}$$

Then, for any time t, given that c_i, d_i, s_i, b_i are defined for all $i \leqslant t$ and recalling that ν_{t+1} is the distribution over the prices at time $t + 1$ (of the strategy α after observing the feedback

[¶] If the reader has doubts about the obliviousness of the $(s_t, b_t)_{t\in\mathbb{N}}$ sequence, see Appendix A.11 where the (oblivious) formal definition of the distributions $\nu_1, \nu_2, \dots$ is provided.

$(s_1, b_1), \dots, (s_t, b_t))$, let

$$\begin{cases} c_{t+1} \coloneqq c_t,\ d_{t+1} \coloneqq d_t - \frac{2\varepsilon}{3^t},\ s_{t+1} \coloneqq 0,\ b_{t+1} \coloneqq d_{t+1}, & \text{if } \nu_{t+1}\big[\big[0, c_t + \frac{\varepsilon}{3^t}\big]\big] \leqslant \frac{1}{2}\ , \\ c_{t+1} \coloneqq c_t + \frac{2\varepsilon}{3^t},\ d_{t+1} \coloneqq d_t,\ s_{t+1} \coloneqq c_{t+1},\ b_{t+1} \coloneqq 1, & \text{otherwise.} \end{cases}$$

Then the sequence of seller/buyer valuations $(s_1, b_1), (s_2, b_2), \dots$ defined above by induction satisfies:

- $\nu_t\big[[s_t, b_t]\big] \leqslant \frac{1}{2}$, for each time t.
- There exists $p^\star \in [0, 1]$ such that $p^\star \in [s_t, b_t]$, for each time t (e.g., $p^\star \coloneqq \lim_{t\to\infty} c_t$).
- $b_t - s_t \geqslant \frac{1-3\varepsilon}{2}$, for each time t.

This implies, for any horizon T,

$$\begin{aligned} R_T\big(\alpha, (s_t, b_t)_{t\in\mathbb{N}}\big) &= \sum_{t=1}^{T} \mathrm{gft}\big(p^\star, (s_t, b_t)\big) - \sum_{t=1}^{T} \mathbb{E}\big[\mathrm{gft}\big(P_t, (s_t, b_t)\big)\big] \\ &\geqslant \sum_{t=1}^{T} (b_t - s_t)\big(1 - \nu_t\big[[s_t, b_t]\big]\big) \geqslant \frac{1-3\varepsilon}{4} T. \end{aligned}$$

The fact that ε and α were chosen arbitrarily yields immediately $R_T^{\mathcal{S}} \geqslant T/4$. □

A more detailed analysis can be found in Appendix A.11.

Before concluding this section, we notice that Theorem 1 immediately implies two things. First, if $\mathcal{S}$ is the set of all (iv) environments, then also $R_T^{\mathcal{S}} \geqslant T/4$. In fact, if $(S_t, B_t)_{t\in\mathbb{N}}$ is a deterministic sequence $(s_t, b_t)_{t\in\mathbb{N}}$, then S_t is clearly independent of B_t, which means that any (adv) environment is also an (iv) environment. Second, given that full feedback is enough to reconstruct realistic feedback, the same impossibility results hold under realistic feedback. Hence, we have already filled the columns under (adv) and (iv) in Table 2.1.

2.3 The Full Feedback Case

In this section, we explore various learning strategies in the full feedback case. We propose two different strategies to learn depending on the properties of the environment where the learner has to act. When the environment satisfies the (bd) assumption we show that the Hedge algorithm in the continuum achieves $\widetilde{O}(\sqrt{T})$ regret guarantees, while if the environment satisfies the (iid) assumption we show that a Follow-the-Leader strategy achieves $O(\sqrt{T})$ regret guarantees. We complement these results by showing that any strategy that strives to compete against all environments satisfying simultaneously the (iv), (bd) and (iid) assumptions has to suffer at least $\Omega(\sqrt{T})$ regret in some instances.

2.3.1 Bounded Density yields Lipschitzness in Expectation

We first prove that, although the gain from trade is discontinuous in general, its expectation is M-Lipschitz, whenever the underlying pair of seller/buyer valuations admits a density bounded by some constant $M > 0$.

Lemma 1 (Lipschitzness)**.** *Let (S, B) be a pair of random variables on $[0,1]^2$ admitting a density (with respect to the Lebesgue measure on $[0,1]^2$) bounded by $M > 0$. Then, the induced gain from trade $\mathrm{GFT}(\cdot) \coloneqq \mathrm{gft}\big(\cdot, (S,B)\big)$ is such that its expectation is M-Lipschitz:*

$$|\mathbb{E}\left[\mathrm{GFT}(y)\right] - \mathbb{E}\left[\mathrm{GFT}(x)\right]| \leqslant M|y-x|, \quad \forall x, y \in [0,1] \tag{2.1}$$

Proof. Without loss of generality we may (and do!) assume that $x > y$. Let U and V two independent uniform random variables in $[0,1]$. We have the following chain of inequalities:

$$\begin{aligned}
|\mathbb{E}\left[\mathrm{GFT}(y)\right] - \mathbb{E}\left[\mathrm{GFT}(x)\right]| &= |\mathbb{E}\left[(B-S)(\mathbb{I}\{S \leqslant y \leqslant B\} - \mathbb{I}\{S \leqslant x \leqslant B\})\right]| \\
&= |\mathbb{E}\left[(B-S)(\mathbb{I}\{S \leqslant y \leqslant B \leqslant x\} - \mathbb{I}\{y \leqslant S \leqslant x \leqslant B\})\right]| \\
&\leqslant \mathbb{P}\left[S \leqslant y \leqslant B \leqslant x\right] + \mathbb{P}\left[y \leqslant S \leqslant x \leqslant B\right] \\
&= \mathbb{P}\big[(S,B) \in [0,y] \times [y,x]\big] + \mathbb{P}\big[(S,B) \in [y,x] \times [x,1]\big] \\
&\leqslant M \cdot \mathbb{P}\big[(U,V) \in [0,y] \times [y,x]\big] + M \cdot \mathbb{P}\big[(U,V) \in [y,x] \times [x,1]\big] \\
&= M \cdot \big(y \cdot (x-y) + (1-x)(x-y)\big) \leqslant M(x-y) = M|x-y|
\end{aligned}$$

Note that in the second to last inequality we used the fact that (S, B) admits a bounded density bounded by M with respect to the Lebesgue measure on $[0,1]^2$. □

2.3.2 Hedge in the Continuum

We now propose an algorithm to deal with the full-feedback case when the environment is only known to satisfy the (bd) assumption, i.e., when there exists a certain constant $M > 0$ such that the seller/buyer pairs $(S_1, B_1), (S_2, B_2), \dots$ form a sequence of $[0,1]^2$-valued random variables, and each of these pairs admits a (possibly different) density bounded by M, without any further assumptions on their distribution (in particular, the sequence $(S_1, B_1), (S_2, B_2), \dots$ is not necessarily i.i.d. and, for any time t, S_t and B_t could be arbitrarily correlated).

We show that running Hedge [93] on the continuum of arms/prices in $[0,1]$ gives a regret rate of order $\widetilde{O}(\sqrt{T})$, featuring also a mild dependence in the upper bound M on the densities of the buyer/seller pairs. The algorithm Continuous-Price Hedge (CPH) is a version of the classic Hedge algorithm played on a continuum of prices where, at time t, a price P_t is drawn according to the *continuous* distribution μ_t with density f_t defined on $[0,1]$ as follows:

$$f_t(p) = \frac{\exp\big(\eta \cdot \sum_{s=1}^{t-1} \mathrm{GFT}_s(p)\big)}{\int_{[0,1]} \exp\big(\eta \cdot \sum_{s=1}^{t-1} \mathrm{GFT}_s(x)\big)\,\mathrm{d}x} = \frac{\exp\big(\eta \cdot \sum_{s=1}^{t-1} \mathrm{GFT}_s(p)\big)}{\Big\|\exp\big(\eta \cdot \sum_{s=1}^{t-1} \mathrm{GFT}_s(\cdot)\big)\Big\|_1}$$

We refer to the pseudocode for further details. Crucially, it is possible to efficiently sample prices from the distributions f_t because the function $\sum_{s=1}^{t-1} \mathrm{GFT}_s$ (and consequently, the density f_t) is piecewise constant with $\Theta(t)$ discontinuities.

While continuous versions of Hedge have already been studied, to the best of our knowledge, we are the first to provide positive results under the assumption that *expected* rewards are Lipschitz. Previous work [119, 130] assumes Lipschitzness of the rewards for *any realization*. The latter assumption is, however, not applicable to the gain from trade, which is discontinuous and not even one-sided Lipschitz in general. This seemingly small difference —from a rewards family that is

realization-wise Lipschitz to one that is regular only in expectation— entails a technical issue in the analysis that we bypass by proving a log-exp analogous of Minkowski's integral inequality (Lemma 16 in Appendix A.3) that we believe is of independent interest. Using this tool, we prove Theorem 35 in Appendix A.2, whose corollary (Corollary 3) provides the general guarantees of Hedge when rewards are Lipschitz in expectation. The proof of the following result is thus an immediate corollary of Lemma 1 and Corollary 3.

Algorithm 1 Continuous-Price Hedge (CPH) - Full Feedback

Input: Learning rate $\eta \in (0,1)$
Initialization: Initialize $W_1(x) \coloneqq 1$, for all $x \in [0,1]$
for time $t = 1, 2, \dots$ **do**
Let μ_t be a distribution with pdf defined by $f_t(x) \coloneqq \frac{W_t(x)}{\|W_t\|_1}$, for all $x \in [0,1]$
Post price P_t drawn according to distribution μ_t
Update $W_{t+1}(x) \coloneqq W_t(x) \cdot \exp\big(\eta\, \mathrm{GFT}_t(x)\big)$, for each $x \in [0,1]$

Theorem 2. *Consider the problem of repeated bilateral trade in the full-feedback model. Let $M > 0$. Suppose that $\mathcal{S}$ is the set of environments such that, for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a density bounded by M (with respect to the Lebesgue measure on $[0,1]^2$). If we run* Continuous-Price Hedge *with learning rate $\eta \in (0,1)$, then, for each time horizon $T \in \mathbb{N}$, we have that*

$$R_T^{\mathcal{S}}(\mathrm{CPH}) \leqslant \frac{1}{\eta} \ln\left(\frac{\eta T \max(M,2)}{1 - e^{-\eta T}}\right) + (e-2)\eta T \,.$$

In particular, if $\eta = \sqrt{\frac{\ln(2T)}{(e-2)T}}$ we have

$$R_T^{\mathcal{S}}(\mathrm{CPH}) \leqslant \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln\big(\max(M,2)\big)}{\ln(2T)}\right) .$$

The bound in Theorem 2 is optimal in the time horizon (see Theorem 4) up to logarithmic terms. Furthermore, we note that the bound exhibits an extremely mild dependence on M *without* requiring any knowledge of M to tune the parameter learning rate η. On the other hand, the learning parameter η does depend on the time horizon T. If the time horizon is unknown, we can obtain the same order of regret with a standard doubling trick [48].

We conclude this section with a brief discussion of an alternative approach. The reader might question why we do not adopt a simpler method to address this problem, which involves initially constructing a uniform grid on the interval $[0,1]$ with K elements, and then applying the standard Hedge algorithm to these points, treating them as arms. In fact, in the light of Lemma 1, this more elementary approach actually works (see also Claim 1), and the corresponding guarantees on the regret are of the form $O\left(\frac{MT}{K} + \sqrt{\log(K)T}\right)$. Even though it is true that this would lead to optimal guarantees in the time horizon T, it is equally clear that, when it also comes to the dependence on the density parameter M, comparable guarantees to those provided by Theorem 2 require the learner to tune K knowing M *in advance*, which is a clear drawback of this alternative approach.

2.3.3 The Decomposition Lemma

In this section, we present a key lemma whose purpose is to decompose the gain from trade into terms that depend only on the outcome of yes/no questions. This result allows leveraging DKW inequalities (see Appendix A.12) in the proofs of our upper bounds. Moreover, it shows how to use the limited feedback available to reconstruct the expected gain from trade in the realistic feedback settings. Furthermore, it leads to an easy proof of the existence of the maximum of the expected gain from trade, under no assumptions on the seller and buyer distributions (see Appendix A.1).

Lemma 2 (Decomposition lemma)**.** *Fix any price $p \in [0,1]$. Then, for any $s, b \in [0,1]$,*

$$\operatorname{gft}\big(p,(s,b)\big) = \int_{[p,1]} \mathbb{I}\{s \leqslant p \leqslant \lambda \leqslant b\}\,\mathrm{d}\lambda + \int_{[0,p]} \mathbb{I}\{s \leqslant \lambda \leqslant p \leqslant b\}\,\mathrm{d}\lambda. \tag{2.2}$$

Furthermore, let S and B be two $[0,1]$-valued random variables.

- *Then*

$$\mathbb{E}\big[\operatorname{gft}\big(p,(S,B)\big)\big] = \int_{[p,1]} \mathbb{P}[S \leqslant p \leqslant \lambda \leqslant B]\,\mathrm{d}\lambda + \int_{[0,p]} \mathbb{P}[S \leqslant \lambda \leqslant p \leqslant B]\,\mathrm{d}\lambda. \tag{2.3}$$

- *If U is uniform on $[0,1]$ and independent of (S,B), then*

$$\mathbb{E}\big[\operatorname{gft}\big(p,(S,B)\big)\big] = \mathbb{P}[S \leqslant p \leqslant U \leqslant B] + \mathbb{P}[S \leqslant U \leqslant p \leqslant B]. \tag{2.4}$$

- *If U is uniform on $[0,1]$ and S, B, U are independent, then*

$$\mathbb{E}\big[\operatorname{gft}\big(p,(S,B)\big)\big] = \mathbb{P}[S \leqslant p]\mathbb{P}[p \leqslant U \leqslant B] + \mathbb{P}[p \leqslant B]\mathbb{P}[S \leqslant U \leqslant p]. \tag{2.5}$$

- *If U is uniform on $[p,1]$, V is uniform on $[0,p]$ and (U,V) is independent of (S,B), then*

$$\mathbb{E}\big[\operatorname{gft}\big(p,(S,B)\big)\big] = \mathbb{E}\big[(1-p)\mathbb{I}\{S \leqslant p \leqslant U \leqslant B\}\big] + \mathbb{E}\big[p\mathbb{I}\{S \leqslant V \leqslant p \leqslant B\}\big]. \tag{2.6}$$

Proof. We begin by proving Equation (2.2). For any $s, b \in [0,1]$, we have

$$\begin{aligned}\operatorname{gft}\big(p,(s,b)\big) &= (b-s)\mathbb{I}\{s \leqslant p \leqslant b\} = \int_s^b \mathrm{d}\lambda \cdot \mathbb{I}\{s \leqslant p \leqslant b\} = \int_{[0,1]} \mathbb{I}\{s \leqslant p \leqslant b\}\mathbb{I}\{s \leqslant \lambda \leqslant b\}\,\mathrm{d}\lambda \\ &= \int_{[p,1]} \mathbb{I}\{s \leqslant p \leqslant \lambda \leqslant b\}\,\mathrm{d}\lambda + \int_{[0,p]} \mathbb{I}\{s \leqslant \lambda \leqslant p \leqslant b\}\,\mathrm{d}\lambda.\end{aligned}$$

Equation (2.3) is an immediate consequence of Equation (2.2) and Fubini's theorem.

We now prove Equation (2.4). Under the assumptions, Equation (2.3) implies

$$\begin{aligned}\mathbb{P}[S \leqslant p \leqslant U \leqslant B] &= \mathbb{P}\Big[\{S \leqslant p\} \cap \{U \leqslant B\} \cap \big\{U \in [p,1]\big\}\Big] \\ &= \int_{[p,1]} \mathbb{P}\big[\{S \leqslant p\} \cap \{U \leqslant B\} \mid U = \lambda\big]\,\mathrm{d}\mathbb{P}_U(\lambda) = \int_{[p,1]} \mathbb{P}[S \leqslant p \leqslant \lambda \leqslant B]\,\mathrm{d}\lambda.\end{aligned}$$

The equality $\mathbb{P}[S \leqslant U \leqslant p \leqslant B] = \int_{[0,p]} \mathbb{P}[S \leqslant \lambda \leqslant p \leqslant B]\,\mathrm{d}\lambda$ can be shown analogously, proving Equation (2.4).

Equation (2.5) is an immediate consequence of Equation (2.4), leveraging independence.

We now prove Equation (2.6). If $p \in \{0,1\}$, the result follows from Equation (2.5). Thus, assume $p \in (0,1)$. Then

$$\mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big] = \int_{[p,1]} \mathbb{P}\big[S \leqslant p \leqslant \lambda \leqslant B\big]\,\mathrm{d}\lambda + \int_{[0,p]} \mathbb{P}\big[S \leqslant \lambda \leqslant p \leqslant B\big]\,\mathrm{d}\lambda.$$

For the first addend, we have,

$$\begin{aligned}
\int_{[p,1]} \mathbb{P}\big[S \leqslant p \leqslant \lambda \leqslant B\big]\,\mathrm{d}\lambda &= (1-p)\int_{[p,1]} \mathbb{P}\big[S \leqslant p \leqslant \lambda \leqslant B\big]\,\mathrm{d}\mathbb{P}_U(\lambda) \\
&= (1-p)\int_{[p,1]} \mathbb{P}[S \leqslant p \leqslant U \leqslant B \mid U = \lambda]\,\mathrm{d}\mathbb{P}_U(\lambda) \\
&= (1-p)\mathbb{P}[S \leqslant p \leqslant U \leqslant B] \\
&= \mathbb{E}\big[(1-p)\mathbb{I}\{S \leqslant p \leqslant U \leqslant B\}\big].
\end{aligned}$$

Analogously, one shows $\int_{[0,p]} \mathbb{P}\big[S \leqslant \lambda \leqslant p \leqslant B\big]\,\mathrm{d}\lambda = \mathbb{E}\big[p \cdot \mathbb{I}\{S \leqslant V \leqslant p \leqslant B\}\big]$, which gives Equation (2.6). □

2.3.4 Follow the Best Price (FPB)

We now consider the full-feedback model in an (iid) environment, i.e., when the seller/buyer pairs $(S_1,B_1),(S_2,B_2),\dots$ form an i.i.d. sequence of $[0,1]^2$-valued random variables, all with the same law as some (S,B), without any further assumptions on their common distribution (in particular, S and B could be arbitrarily correlated and the pair (S,B) does not necessarily admit a bounded density).

We show that a *Follow-the-Leader* approach, which we call Follow the Best Price (FBP, Algorithm 2), achieves a $O\big(\sqrt{T}\big)$ regret upper bound. The Follow the Best Price (FBP) algorithm consists in posting the best price with respect to the samples that have been observed so far. Notably, it does not need preliminary knowledge of the time horizon T.

Algorithm 2 Follow the Best Price (FBP) - Full Feedback

Let $P_1 \coloneqq 1/2$
for $t = 1,2,\dots$ **do**
  Post price P_t
  Pick $P_{t+1} \in \arg\max_{p\in[0,1]} \frac{1}{t}\sum_{i=1}^{t} \mathrm{gft}\big(p,(S_i,B_i)\big)$

For each time t, given $(S_1,B_1),\dots,(S_t,B_t)$, one can reconstruct the gain from trade function $\mathrm{gft}\big(\cdot,(S_i,B_i)\big)$ at each time step $i \leqslant t$ and compute (one of) the best price(s) $P_{t+1} \in \arg\max_{p\in[0,1]} \frac{1}{t}\sum_{i=1}^{t} \mathrm{gft}\big(p,(S_i,B_i)\big)$. Note that $\frac{1}{t}\sum_{i=1}^{t} \mathrm{gft}\big(\cdot,(S_i,B_i)\big)$ is a step-wise constant function that attains its maximum at one of the observed sellers' valuations $S_1\dots,S_t$.[‖] Hence, even a naive enumeration approach is computationally efficient.

On a technical note, prices P_{t+1} should be defined in a measurable way in order for the regret to be well defined. For example, this can be done by picking P_{t+1} as S_i, where i is the smallest index among all the indices j such that $S_j \in \arg\max_{p\in[0,1]} \frac{1}{t}\sum_{t=1}^{t} \mathrm{gft}\big(p,(S_i,B_i)\big)$.

[‖] By the symmetry of the problem, the maximum is also attained at one of the buyers' valuations.

The main idea of the analysis of Algorithm 2 is to show that the approximation of the expected gain from trade with its empirical means is uniform over all possible seller/buyer distributions and prices. An alternative way to achieve this result is through a pseudo-dimension argument (e.g., see Li et al. [125, Introduction and Theorem 5]). However, this approach requires subtler measurability considerations. We will show that one could get around these measurability issues altogether by leveraging the Decomposition lemma (Lemma 2) and a bivariate DKW inequality (Theorem 42).

Theorem 3. *Consider the problem of repeated bilateral trade in the full-feedback model. Suppose that $\mathcal{S}$ is the set of environments such that the sequence $(S_1, B_1), (S_2, B_2), \dots$ is independent and identically distributed (iid), all with the same law as some (S, B). If we run Follow the Best Price, then, for each time horizon $T \in \mathbb{N}$, we have that*

$$R_T^{\mathcal{S}}(\mathrm{FBP}) \leqslant \frac{1}{2} + c\sqrt{T-1}.$$

where $c \in (0, 1144240)$ is a universal constant.

Proof. Without loss of generality, assume that $T \geqslant 2$. Fix any $t \in [T-1]$. For any $p \in [0,1]$ define the random variable

$$H_t(p) \coloneqq \frac{1}{t}\sum_{i=1}^{t} \mathrm{GFT}_i(p) - \mathbb{E}\big[\mathrm{GFT}(p)\big]\,,$$

where we recall that $\mathrm{GFT}_i(p) = \mathrm{gft}\big(p, (S_i, B_i)\big)$, while we defined $\mathrm{GFT}(p) \coloneqq \mathrm{gft}\big(p, (S, B)\big)$. Leveraging the definition of P_{t+1} and the independence of P_{t+1} and (S_{t+1}, B_{t+1}), the Freezing Lemma (Lemma 17) yields

$$\begin{aligned}\mathbb{E}\big[\mathrm{GFT}_{t+1}(p^\star)\big] - \mathbb{E}\big[\mathrm{GFT}_{t+1}(P_{t+1})\big] &\leqslant \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t}\mathrm{GFT}_i(P_{t+1})\right] - \mathbb{E}\big[\mathrm{GFT}_{t+1}(P_{t+1})\big] \\ &= \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t}\mathrm{GFT}_i(P_{t+1}) - \mathbb{E}\big[\mathrm{GFT}_{t+1}(P_{t+1}) \mid P_{t+1}\big]\right] = \mathbb{E}\big[H_t(P_{t+1})\big] =: (*).\end{aligned}$$

Then, by the Decomposition lemma (2.2)-(2.3), we get

$$\begin{aligned}H_t(P_{t+1}) &\leqslant \sup_{p\in[0,1]}\left(\frac{1}{t}\sum_{i=1}^{t}\mathrm{GFT}_i(p) - \mathbb{E}\big[\mathrm{GFT}(p)\big]\right) \qquad (2.7)\\
&= \sup_{p\in[0,1]}\left(\frac{1}{t}\sum_{i=1}^{t}\left(\int_{[p,1]}\mathbb{I}\left\{S_i \leqslant p \leqslant \lambda \leqslant B_i\right\}\mathrm{d}\lambda + \int_{[0,p]}\mathbb{I}\left\{S_i \leqslant \lambda \leqslant p \leqslant B_i\right\}\mathrm{d}\lambda\right)\right.\\
&\qquad\qquad \left. - \left(\int_{[p,1]}\mathbb{P}[S \leqslant p \leqslant \lambda \leqslant B]\,\mathrm{d}\lambda + \int_{[0,p]}\mathbb{P}[S \leqslant \lambda \leqslant p \leqslant B]\,\mathrm{d}\lambda\right)\right)\\
&= \sup_{p\in[0,1]}\left(\int_{[0,p]}\left(\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i \leqslant \lambda, -B_i \leqslant -p\} - \mathbb{P}[S \leqslant \lambda, -B \leqslant -p]\right)\mathrm{d}\lambda\right.\\
&\qquad\qquad \left. + \int_{[p,1]}\left(\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i \leqslant p, -B_i \leqslant -\lambda\} - \mathbb{P}[S \leqslant p, -B \leqslant -\lambda]\right)\mathrm{d}\lambda\right)\\
&\leqslant 2\sup_{x,y\in\mathbb{R}}\left|\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i \leqslant x, -B_i \leqslant y\} - \mathbb{P}[S \leqslant x, -B \leqslant y]\right|. \qquad (2.8)\end{aligned}$$

Letting m_0, c_1, c_2 as in Theorem 42, $\varepsilon_t \coloneqq \sqrt{m_0/t}$, taking expectations to the left and right hand side of Equation (2.8), and applying the bivariate DKW inequality (Theorem 42), we get

$$\begin{aligned}(*) &\leqslant \mathbb{E}\left[2\sup_{x,y\in\mathbb{R}}\left|\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i\leqslant x,-B_i\leqslant y\}-\mathbb{P}[S\leqslant x,-B\leqslant y]\right|\right]\\ &\leqslant 2\varepsilon_t+2\int_{[\varepsilon_t,1]}\mathbb{P}\left[\sup_{x,y\in\mathbb{R}}\left|\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i\leqslant x,-B_i\leqslant y\}-\mathbb{P}[S\leqslant x,-B\leqslant y]\right|>\varepsilon\right]\mathrm{d}\varepsilon \qquad (2.9)\\ &\leqslant 2\varepsilon_t+2\int_{\varepsilon_t}^{1}c_1\exp\bigl(-c_2t\varepsilon^2\bigr)\,\mathrm{d}\varepsilon\leqslant 2\varepsilon_t+\frac{c_1}{\sqrt{c_2t}}\int_0^\infty e^{-u}u^{-1/2}\,\mathrm{d}u=\left(2\sqrt{m_0}+c_1\sqrt{\frac{\pi}{c_2}}\right)\frac{1}{\sqrt{t}}.\end{aligned}$$

Being t arbitrary, using the fact that $\sum_{t=1}^{T-1}t^{-1/2}\leqslant 2\sqrt{T-1}$, and letting $c\coloneqq 2\left(2\sqrt{m_0}+c_1\sqrt{\frac{\pi}{c_2}}\right)<1144265$, we have that

$$R_T\bigl(\mathrm{FBP},(S_t,B_t)_{t\in\mathbb{N}}\bigr)\leqslant\frac{1}{2}+\sum_{t=1}^{T-1}\Bigl(\mathbb{E}\bigl[\mathrm{GFT}_{t+1}(p^\star)\bigr]-\mathbb{E}\bigl[\mathrm{GFT}_{t+1}(P_{t+1})\bigr]\Bigr)\leqslant\frac{1}{2}+\frac{c}{2}\sum_{t=1}^{T-1}\frac{1}{\sqrt{t}}=\frac{1}{2}+c\sqrt{T-1}\,,$$

which concludes the proof. □

The loose bound on the constant c appearing in the statement is due to the (likely suboptimal) large constants appearing in Theorem 42: any improvement on the bivariate DKW inequality would result in an improvement of this constant. For example, it is conjectured (see, e.g., Naaman [144, Section 5]) that the tightest bound for the bivariate DKW inequality is (with the same notation as Theorem 42), for all $m\in\mathbb{N}$ and $\varepsilon>0$, $\mathbb{P}\left[\sup_{x,y\in\mathbb{R}}\left|\frac{1}{m}\sum_{k=1}^{m}\mathbb{I}\{X_k\leqslant x,Y_k\leqslant y\}-\mathbb{P}[X\leqslant x,Y\leqslant y]\right|>\varepsilon\right]\leqslant 4\exp\bigl(-2m\varepsilon^2\bigr)$. If this was the case, we could replace Equation (2.9) with

$$(*)\leqslant 2\int_{[0,1]}\mathbb{P}\left[\sup_{x,y\in\mathbb{R}}\left|\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}\{S_i\leqslant x,-B_i\leqslant y\}-\mathbb{P}[S\leqslant x,-B\leqslant y]\right|>\varepsilon\right]\mathrm{d}\varepsilon\leqslant 2\sqrt{2\pi}\frac{1}{\sqrt{t}}.$$

leading to a significantly smaller constant $c\coloneqq 2\cdot 2\sqrt{2\pi}<11$.

2.3.5 $\sqrt{T}$ Lower Bound (iv) + (bd) + (iid) in Full Feedback

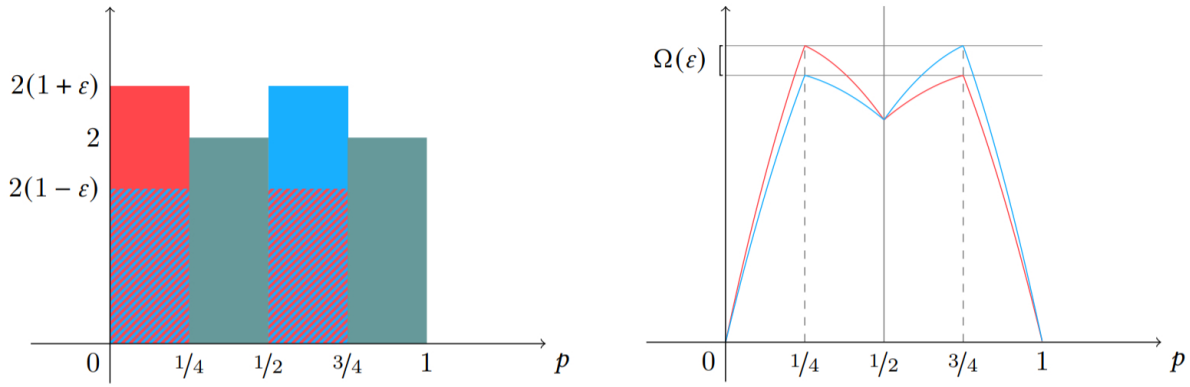
In this section, we show that the upper bounds on the minimax regret we proved in Section 2.3.2 and Section 2.3.4 are essentially tight. No strategy can beat the $O\bigl(\sqrt{T}\bigr)$ rate when the seller/buyer pair (S_t,B_t) is drawn i.i.d. from an unknown fixed distribution, even under the further assumptions that the valuations of the seller and buyer are independent of each other and have bounded densities. In particular, this implies that we have completed the first row in Table 2.1. For a full proof of the following theorem, see Appendix A.7.

Theorem 4. *In the full-feedback model, for all horizons T, the minimax regret $R_T^{\mathcal{S}}$ satisfies*

$$R_T^{\mathcal{S}}\geqslant c\sqrt{T}\,,$$

where $c\geqslant 1/\bigl(8\sqrt{2\pi}\bigr)$, and $\mathcal{S}$ is the set of all environments such that

(iv) *for each $t\in\mathbb{N}$, S_t and B_t are independent of each other.*



(a) Distributions $f_{S_{\pm\varepsilon}}$ (red/blue) and f_B (green) (b) Expected GFT relative to $+\varepsilon$ (red) and $-\varepsilon$ (blue)

Figure 2.1: The best posted price is $1/4$ (resp., $3/4$) in the $+\varepsilon$ (resp., $-\varepsilon$) case. By posting $1/4$, the player suffers a $\Omega(\varepsilon)$ regret in the $-\varepsilon$ case, and the same is true posting $3/4$ if in $+\varepsilon$ case.

(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a joint density bounded by $M \geqslant 4$.*

(iid) *$(S_1, B_1), (S_2, B_2), \ldots$ is an i.i.d. sequence.*

Proof sketch. We build a family of distributions $\mu_{\pm\varepsilon}$ for the seller/buyer pair parameterized by $\varepsilon \in [0, 1]$. For the seller, for any $\varepsilon \in [0, 1]$, we define the density

$$f_{S,\pm\varepsilon} \coloneqq 2(1 \pm \varepsilon)\mathbb{I}_{\left[0,\frac{1}{4}\right]} + 2(1 \mp \varepsilon)\mathbb{I}_{\left[\frac{1}{2},\frac{3}{4}\right]}. \qquad \text{(Figure 2.1a, in red/blue)}$$

For the buyer, we define a single density (independently of ε)

$$f_B \coloneqq 2\mathbb{I}_{\left[\frac{1}{4},\frac{1}{2}\right]\cup\left[\frac{3}{4},1\right]}. \qquad \text{(Figure 2.1a, in green)}$$

In the $+\varepsilon$ (resp., $-\varepsilon$) case, the optimal price belongs to the region $[0, 1/2]$ (resp., $(1/2, 1]$, see Figure 2.1b). By posting prices in the wrong region $(1/2, 1]$ (resp., $[0, 1/2]$) in the $+\varepsilon$ (resp., $-\varepsilon$) case, the learner incurs a $\Omega(\varepsilon)$ regret. Thus, the only way to avoid suffering $\Omega(\varepsilon T)$ regret is to identify the sign of $\pm\varepsilon$ and play accordingly. However, by information-theoretic arguments, this task requires $\approx \frac{1}{\varepsilon^2}$ rounds, during which we pay at least $\approx \varepsilon$ in each of them. Tuning $\varepsilon \approx \frac{1}{\sqrt{T}}$, leads to the conclusion.

The reader might have noticed that this construction closely resembles the lower bound of online learning with expert advice. Actually, a technical proof (see Appendix A.7), shows that our setting is harder (i.e., it has a higher minimax regret) than an instance of an expert problem (with two experts), which has a known lower bound on its minimax regret of $\frac{1}{8\sqrt{2\pi}}\sqrt{T}$ [70]. □

2.4 The Realistic Feedback Case

In this section, we tackle the problem in the more challenging realistic-feedback model. We recall that in the realistic-feedback model, the only information collected by the learner at the end of each round t consists of $\mathbb{I}\{S_t \leqslant P_t\}$ and $\mathbb{I}\{P_t \leqslant B_t\}$.

2.4.1 Scouting Bandits: from Realistic Feedback to Multi-Armed Bandits

We start by studying the problem under the assumption that the seller/buyer pairs $(S_1, B_1), (S_2, B_2), \dots$ are $[0,1]^2$-valued i.i.d. random variables (iid), all with the same law as some (S, B), where S and B are independent (iv) and have bounded densities (bd).

The main challenge in designing low-regret algorithms with realistic feedback lies in the fact that posting a price *does not* reveal the corresponding gain from trade. We can observe this phenomenon by looking at the Decomposition lemma (2.5). While the local terms $\mathbb{P}[S \leqslant p]$ and $\mathbb{P}[p \leqslant B]$ can be reconstructed by simply posting the same price p multiple times, the integral terms are inherently global: they depend on all values in $(p, 1]$ and $[0, p)$, and thus estimating them requires posting prices that are *far* from p. This prevents direct application of well-established algorithms, such as action elimination or UCB [169], and suggests that this problem is harder than multiarmed bandits (as in fact it is: see Section 2.4.2).

A naive approach to tackle this issue could be estimating the CDFs of S and B on a suitable grid of prices and using this information to reconstruct both the global and the local terms of $\mathbb{E}\big[\mathrm{gft}\big(\cdot, (S,B)\big)\big]$. This would lead to an $\widetilde{O}\big(T^{3/4}\big)$ regret. Instead, our Algorithm 3 (Scouting Bandits) exploits better the decomposition in Equation (2.5) by learning *separately* the global and local parts of the gain from trade. First, a global exploration phase is run (scouting phase), in which prices uniformly sampled in $[0,1]$ are posted and used to simultaneously estimate the integral terms on a suitable grid of K points. Once this is done, by replacing the integrals in Equation (2.5) with their approximations $\widehat{F}_k$ and $\widehat{G}_k$ for each price q_k in the grid, we obtain the estimate

$$\begin{aligned}\mathbb{E}\big[\mathrm{gft}\big(q_k, (S,B)\big)\big] &\approx \mathbb{P}[S \leqslant q_k]\,\widehat{F}_k + \mathbb{P}[q_k \leqslant B]\,\widehat{G}_k \\ &= \mathbb{E}\big[\mathbb{I}\{S \leqslant q_k\}\,\widehat{F}_k + \mathbb{I}\{q_k \leqslant B\}\,\widehat{G}_k \mid H\big] =: \mathbb{E}\big[Z(k) \mid H\big]\,,\end{aligned}$$

where H consists of the estimates $\widehat{F}_j, \widehat{G}_j$ (for all j) at the the end of the global exploration phase. We are now only left to solve a bandit problem on K arms with reward function Z: the only quantities to learn are the two local terms $\mathbb{P}[S \leqslant q_k]$ and $\mathbb{P}[q_k \leqslant B]$, which can be estimated with the available feedback by posting the price q_k.

Algorithm 3 Scouting Bandits - Realistic Feedback

input: exploration time T_0, grid size K, and K-armed bandit algorithm α
initialization: $q_k \coloneqq k/(K+1)$, $\widehat{F}_k \coloneqq 0$, $\widehat{G}_k \coloneqq 0$, for all $k \in [K]$
for $t = 1, 2, \dots, T_0$ **do** ▷ (scouting phase)
  post U_t drawn uniformly at random in $[0,1]$
  let $\widehat{F}_k \coloneqq \widehat{F}_k + \frac{1}{T_0}\mathbb{I}\{q_k \leqslant U_t \leqslant B_t\}$, and $\widehat{G}_k \coloneqq \widehat{G}_k + \frac{1}{T_0}\mathbb{I}\{S_t \leqslant U_t \leqslant q_k\}$, for all $k \in [K]$
for $t = T_0 + 1, T_0 + 2, \dots$ **do** ▷ (bandit phase)
  generate the next arm I_t with α
  post price q_{I_t}
  feed α the reward $Z_t(I_t) \coloneqq \mathbb{I}\{S_t \leqslant q_{I_t}\}\widehat{F}_{I_t} + \mathbb{I}\{q_{I_t} \leqslant B_t\}\widehat{G}_{I_t}$

The independence of S and B (iv) is required for applying Equation (2.5), while the bounded density assumption (bd) implies the Lipschitzness of the expected gain from trade (Lemma 1), which in turns allows to discretize the problem.

We are now ready to state and prove the main result of this section.

Theorem 5. *Consider the problem of repeated bilateral trade in the realistic-feedback model. Suppose that $\mathcal{S}$ is the set of environments such that the sequence of evaluations $(S_1, B_1), (S_2, B_2), \dots$ is independent and identically distributed (iid), all with the same law as some (S, B) admitting a density (with respect to the Lebesgue measure on $[0,1]^2$) bounded by some constant M (bd), and such that S and B are independent of each other (iv). If we run Scouting Bandits (SB) with parameters T_0, K, and α, then, for any time horizon $T \geqslant T_0$, we have*

$$R_T^{\mathcal{S}}(\mathrm{SB}) \leqslant T_0 + \left(\frac{M}{K+1} + \sqrt{\frac{2\pi}{T_0}}\right)(T - T_0) + \mathcal{R}_{T-T_0}(\alpha)\ ,$$

where $\mathcal{R}_\tau(\alpha)$ is a distribution-free upper bound on the regret after τ rounds of α in the stochastic i.i.d. setting with $[0,1]$-valued rewards.

In particular, if for each K we have a bandit algorithm α^K over K arms such that $\mathcal{R}_\tau(\alpha^K) = O\big(\sqrt{K\tau}\big)$ (e.g., if α^K is the MOSS algorithm over K arms [15]), then tuning the parameters $T_0 \coloneqq \lceil T^{2/3} \rceil$ and $K \coloneqq \lceil T^{1/3} \rceil$ gives the regret bound $R_T(\mathrm{SB}) = O\big(MT^{2/3}\big)$.

Proof. Let $H \coloneqq (\widehat{F}_k, \widehat{G}_k)_{k\in[K]}$ and denote its range space $[0,1]^{2K}$ by $\mathcal{H}$. For each $h = (f_k, g_k)_{k\in[K]} \in \mathcal{H}$, let $(I_{h,t})_{t\geqslant T_0+1}$ be the sequence of arms pulled by α (possibly using some internal randomization) on the sequence of rewards $(Z_{h,t})_{t\geqslant T_0+1}$ defined for any time $t \geqslant T_0+1$ and all arms $k \in [K]$ by

$$Z_{h,t}(k) \coloneqq \mathbb{I}\{S_t \leqslant q_k\} f_k + \mathbb{I}\{q_k \leqslant B_t\} g_k.$$

Let $p^\star \in \arg\max_{p\in[0,1]} \mathbb{E}\big[\mathrm{gft}\big(p, (S,B)\big)\big]$ and $k^\star$ be the index of a point in the grid $\{q_1, \dots, q_K\}$ closest to $p^\star$. Let P_t be the price posted by SB at each time t. Then

$$\begin{aligned}
R_T\big(\mathrm{SB}, (S_t, B_t)_{t\in\mathbb{N}}\big) &\leqslant T_0 + \sum_{t=T_0+1}^{T} \mathbb{E}\big[\mathrm{GFT}_t(p^\star) - \mathrm{GFT}_t(P_t)\big] \\
&= T_0 + \sum_{t=T_0+1}^{T} \Big(\mathbb{E}\big[\mathrm{GFT}_t(p^\star)\big] - \mathbb{E}\big[\mathrm{GFT}_t(q_{k^\star})\big]\Big) + \sum_{t=T_0+1}^{T} \Big(\mathbb{E}\big[\mathrm{GFT}_t(q_{k^\star})\big] - \mathbb{E}\big[Z_{H,t}(k^\star)\big]\Big) \\
&\quad + \mathbb{E}\left[\sum_{t=T_0+1}^{T} Z_{H,t}(k^\star) - \sum_{t=T_0+1}^{T} Z_{H,t}(I_{H,t})\right] + \sum_{t=T_0+1}^{T} \Big(\mathbb{E}\big[Z_{H,t}(I_{H,t})\big] - \mathbb{E}\big[\mathrm{GFT}_t(P_t)\big]\Big) \\
&=: T_0 + (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}) + (\mathrm{IV}).
\end{aligned} \tag{2.10}$$

We bound the four terms separately.

For the term (I), by the M-Lipschitzness of the gain from trade (Lemma 1) and the fact that the step size of the grid is $1/(K+1)$, we get

$$(\mathrm{I}) = \sum_{t=T_0+1}^{T} \Big(\mathbb{E}\big[\mathrm{GFT}_t(p^\star)\big] - \mathbb{E}\big[\mathrm{GFT}_t(q_{k^\star})\big]\Big) \leqslant M|p^\star - q_{k^\star}|(T - T_0) \leqslant \frac{M}{K+1}(T - T_0).$$

For the term (II), for any $t \geqslant T_0 + 1$, by the independence of H and (S_t, B_t), we have

$$\begin{aligned}
\mathbb{E}\big[Z_{H,t}(k^\star)\big] &= \mathbb{E}\big[\mathbb{I}\{S_t \leqslant q_{k^\star}\}\widehat{F}_{k^\star} + \mathbb{I}\{q_{k^\star} \leqslant B_t\}\widehat{G}_{k^\star}\big] \\
&= \mathbb{P}[S_t \leqslant q_{k^\star}]\mathbb{P}[q_{k^\star} \leqslant U_t \leqslant B_t] + \mathbb{P}[q_{k^\star} \leqslant B_t]\mathbb{P}[S_t \leqslant U_t \leqslant q_{k^\star}] = \mathbb{E}\big[\mathrm{GFT}_t(q_{k^\star})\big]\ ,
\end{aligned}$$

where the last identity follows from Equation (2.5), and in turn implies that (II) = 0.

For the term (III), using the fact that for $\mathbb{P}_H$-almost every $h \in \mathcal{H}$, the sequence $(Z_{h,t})_{t \geqslant T_0+1}$ is included in $[0,1]$, we obtain

$$\begin{aligned}(\mathrm{III}) &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=T_0+1}^{T} Z_{H,t}(k^\star) - \sum_{t=T_0+1}^{T} Z_{H,t}(I_{H,t})\right] \mid H\right] \\ &\overset{(*)}{\leqslant} \int_{\mathcal{H}} \mathbb{E}\left[\sum_{t=T_0+1}^{T} Z_{h,t}(k^\star) - \sum_{t=T_0+1}^{T} Z_{h,t}(I_{h,t})\right] \mathrm{d}\mathbb{P}_H(h) \leqslant \mathcal{R}_{T-T_0}(\alpha)\end{aligned}$$

where $(*)$ follows from the independence of $(I_{h,t}, S_t, B_t)$ and H (for any $h \in \mathcal{H}$ and all $t \geqslant T_0 + 1$) and in the last inequality we upper bounded (for $\mathbb{P}_H$-almost every $h \in \mathcal{H}$) the regret of α when run on the sequence of rewards $(Z_{h,t})_{t \geqslant T_0+1}$ with $\mathcal{R}_{T-T_0}(\alpha)$.

Finally, we upper bound the last term (IV). If the K-armed bandit algorithm α is randomized, let V_t be its internal randomization of at each time step $t \geqslant T_0 + 1$; otherwise, omit all references to $(V_t)_{t \geqslant T_0+1}$. Define, for each time step $t \geqslant T_0 + 1$,

$$L_t \coloneqq (H, V_{T_0+1}, S_{T_0+1}, B_{T_0+1}, \ldots, V_{t-1}, S_{t-1}, B_{t-1}, V_t)$$

and $\mathbb{P}_t \coloneqq \mathbb{P}[\cdot \mid L_t]$. Then take a uniform random variable U_t on $[0,1]$ independent of (L_t, B_t, S_t). Now, for all $t \geqslant T_0 + 1$, leveraging the measurability of $q_{I_{H,t}}, \widehat{F}_{I_{H,t}}, \widehat{G}_{I_{H,t}}$ with respect to $\sigma(L_t)$, the independence of L_t and (S_t, B_t), and the Decomposition lemma (2.5), we get

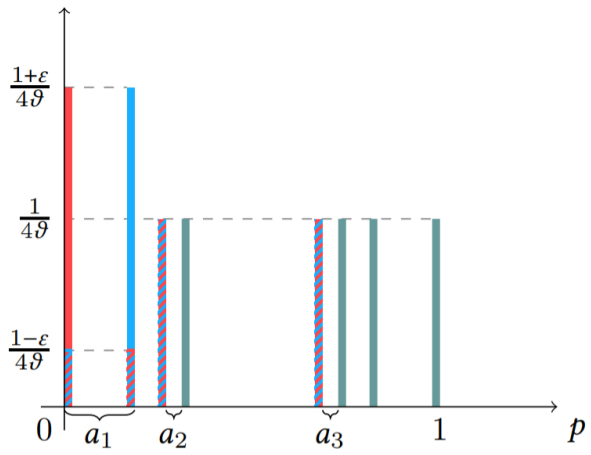
$$\begin{aligned}&\mathbb{E}\big[Z_{H,t}(I_{H,t})\big] - \mathbb{E}\big[\mathrm{GFT}_t(P_t)\big] \\ &\quad= \mathbb{E}\Big[\mathbb{E}\big[\big(\mathbb{I}\{S_t \leqslant q_{I_{H,t}}\}\widehat{F}_{I_{H,t}} + \mathbb{I}\{q_{I_{H,t}} \leqslant B_t\}\widehat{G}_{I_{H,t}}\big) - \mathrm{GFT}(q_{I_{H,t}}, S_t, B_t) \mid L_t\big]\Big] \\ &\quad= \mathbb{E}\big[\mathbb{P}_t[S_t \leqslant q_{I_{H,t}}]\big(\widehat{F}_{I_{H,t}} - \mathbb{P}_t[q_{I_{H,t}} \leqslant U_t \leqslant B_t]\big)\big] \\ &\qquad+ \mathbb{E}\big[\mathbb{P}_t[q_{I_{H,t}} \leqslant B_t]\big(\widehat{G}_{I_{H,t}} - \mathbb{P}_t[S_t \leqslant U_t \leqslant q_{I_{H,t}}]\big)\big] \\ &\quad\leqslant \mathbb{E}\left[\max_{k\in[K]}\big|\widehat{F}_k - \mathbb{P}[q_k \leqslant U_1 \leqslant B_1]\big|\right] + \mathbb{E}\left[\max_{k\in[K]}\big|\widehat{G}_k - \mathbb{P}[S_1 \leqslant U_1 \leqslant q_k]\big|\right] =: (\mathrm{V}) + (\mathrm{VI})\end{aligned}$$

For the first addend, applying the univariate DKW inequality (Theorem 41), we have

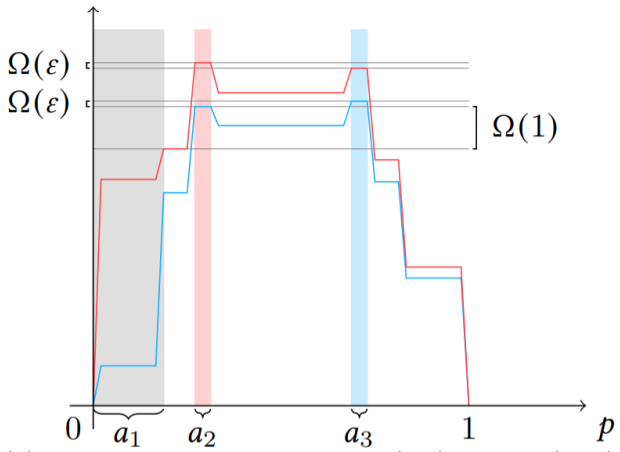
$$\begin{aligned}(\mathrm{V}) &= \int_{[0,1]} \mathbb{P}\left[\max_{k\in[K]}\big|\widehat{F}_k - \mathbb{P}[q_k \leqslant U_1 \leqslant B_1]\big| > \varepsilon\right] \mathrm{d}\varepsilon \\ &= \int_{[0,1]} \mathbb{P}\left[\max_{k\in[K]}\left|\frac{1}{T_0}\sum_{i=1}^{T_0}\mathbb{I}\big\{-U_i\mathbb{I}\{U_i \leqslant B_i\} \leqslant -q_k\big\} - \mathbb{P}\big[-U_1\mathbb{I}\{U_1 \leqslant B_1\} \leqslant -q_k\big]\right| > \varepsilon\right] \mathrm{d}\varepsilon \\ &\leqslant \int_{[0,1]} \mathbb{P}\left[\sup_{x\in\mathbb{R}}\left|\frac{1}{T_0}\sum_{i=1}^{T_0}\mathbb{I}\big\{-U_i\mathbb{I}\{U_i \leqslant B_i\} \leqslant x\big\} - \mathbb{P}\big[-U_1\mathbb{I}\{U_1 \leqslant B_1\} \leqslant x\big]\right| > \varepsilon\right] \mathrm{d}\varepsilon \\ &\leqslant \int_0^1 2\exp\big(-2T_0\varepsilon^2\big)\,\mathrm{d}\varepsilon \leqslant \frac{1}{\sqrt{2T_0}}\int_0^\infty e^{-u}u^{-1/2}\,\mathrm{d}u = \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{T_0}}.\end{aligned}$$

Similarly, one can show that $(\mathrm{VI}) \leqslant \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{T_0}}$ which in turn yields $(\mathrm{IV}) \leqslant \sqrt{\frac{2\pi}{T_0}}(T - T_0)$.

Putting the bounds on (I)-(IV) together in (2.10) gives the first part of the result. Substituting the stated choice of the parameters yields the second. □



(a) Distributions $f_{S,\pm\varepsilon}$ (red/blue) and f_B (green)



(b) Expected GFT relative to $+\varepsilon$ (red) and $-\varepsilon$ (blue)

Figure 2.2: The only three regions where it makes sense for the learner to post prices are a_1, a_2, a_3. Prices in a_1 reveal information about the sign of $\pm\varepsilon$ suffering a $\Omega(1)$ regret; prices in a_2 are optimal if the distribution of the seller is the red one $(+\varepsilon)$ but incur $\Omega(\varepsilon)$ regret if it is the blue one $(-\varepsilon)$; the converse happens in a_3.

Note that to achieve a regret of order $O(MT^{2/3})$ we tuned the parameters T_0 and K of Scouting Bandits as a function of T. If the time horizon is unknown, we can obtain the same order of regret with a standard doubling trick [48]. In addition, note that if we allow tuning the parameters as a function of the Lipschitz constant M (which is however unknown in general), the regret rate would improve to order $O(M^{1/3}T^{2/3})$. This can be achieved by taking $T_0 \coloneqq \lceil T^{2/3} \rceil$ and $K \coloneqq \lceil M^{2/3}T^{1/3} \rceil$.

2.4.2 $T^{2/3}$ Lower Bound under Realistic Feedback (iv) + (bd) + (iid)

In this section, we show that the upper bound on the minimax regret we proved in Section 2.4.1 is tight in the time horizon. No strategy can beat the $O(T^{2/3})$ 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 densities. For a full proof of the following theorem, see Appendix A.8.

Theorem 6. *In the realistic-feedback model, for all horizons T, the minimax regret satisfies*

$$R_T^{\mathcal{S}} \geqslant cT^{2/3} ,$$

where $c \geqslant 11/672$, and $\mathcal{S}$ is the set of all environments such that

(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a joint density bounded by $M \geqslant 24$.*

(iid) *$(S_1, B_1), (S_2, B_2), \ldots$ is an i.i.d. sequence.*

Proof sketch. We build a family of distributions $\mu_{\pm\varepsilon}$ of the seller/buyer pair parameterized by $\varepsilon \in [0, 1]$. For the seller, for any $\varepsilon \in [0, 1]$, we define the density

$$f_{S,\pm\varepsilon} \coloneqq \frac{1}{4\vartheta}\left((1 \pm \varepsilon)\mathbb{I}_{[0,\vartheta]} + (1 \mp \varepsilon)\mathbb{I}_{\left[\frac{1}{6},\frac{1}{6}+\vartheta\right]} + \mathbb{I}_{\left[\frac{1}{4},\frac{1}{4}+\vartheta\right]} + \mathbb{I}_{\left[\frac{2}{3},\frac{2}{3}+\vartheta\right]}\right) , \quad \text{(Figure 2.2a, in red/blue)}$$

where $\vartheta \coloneqq 1/48$ is a normalization constant. For the buyer, we define a single density (independently of ε)

$$f_B \coloneqq \frac{1}{4\vartheta}\left(\mathbb{I}_{\left[\frac{1}{3}-\vartheta,\,\frac{1}{3}\right]} + \mathbb{I}_{\left[\frac{3}{4}-\vartheta,\,\frac{3}{4}\right]} + \mathbb{I}_{\left[\frac{5}{6}-\vartheta,\,\frac{5}{6}\right]} + \mathbb{I}_{[1-\vartheta,\,1]}\right). \qquad \text{(Figure 2.2a, in green)}$$

In the $+\varepsilon$ (resp., $-\varepsilon$) case, the optimal price belongs to a region a_2 (resp., a_3, see Figure 2.2b). By posting prices in the wrong region a_3 (resp., a_2) in the $+\varepsilon$ (resp., $-\varepsilon$) case, the learner incurs $\Omega(\varepsilon)$ regret. Thus, the only way to avoid suffering $\Omega(\varepsilon T)$ regret is to identify the sign of $\pm\varepsilon$ and play accordingly. Clearly, the feedback received from the buyer gives no information on $\pm\varepsilon$. Since the feedback received from the seller at time t by posting a price p is $\mathbb{I}\{S_t \leqslant p\}$, one can obtain information about (the sign of) $\pm\varepsilon$ only by posting prices in the costly ($\Omega(1)$-regret each time) sub-optimal region a_1, where, by information-theoretic arguments, we need to post $\approx \frac{1}{\varepsilon^2}$ times to collect reliable information about (the sign of) $\pm\varepsilon$. Tuning $\varepsilon \approx T^{-1/3}$ leads to the desired lower bound.

The reader might have noticed that this construction closely resembles the learning dilemma present in the so-called *revealing action* partial monitoring game [48]. Actually, a technical proof (see Appendix A.8), shows that our setting is harder (i.e., it has a higher minimax regret) than an instance of a revealing action problem, which has a known lower bound on its minimax regret of $\frac{11}{96}\left(\frac{1}{7}T^{2/3}\right)$ [49]. □

2.4.3 Linear Lower Bound under Realistic Feedback (bd) + (iid)

In this section, we show that no strategy that can achieve worst-case sublinear regret when the seller/buyer pair (S_t, B_t) is drawn i.i.d. from an unknown fixed distribution, even under the further assumption that the valuations of the seller and buyer have bounded densities. This is due to a lack of observability. For a full proof of the following theorem, see Appendix A.9.

Theorem 7. *In the realistic-feedback model, for all horizons T, the minimax regret $R_T^{\mathcal{S}}$ satisfies*

$$R_T^{\mathcal{S}} \geqslant cT\,,$$

where $c \geqslant 1/24$, and $\mathcal{S}$ is the set of all environments such that

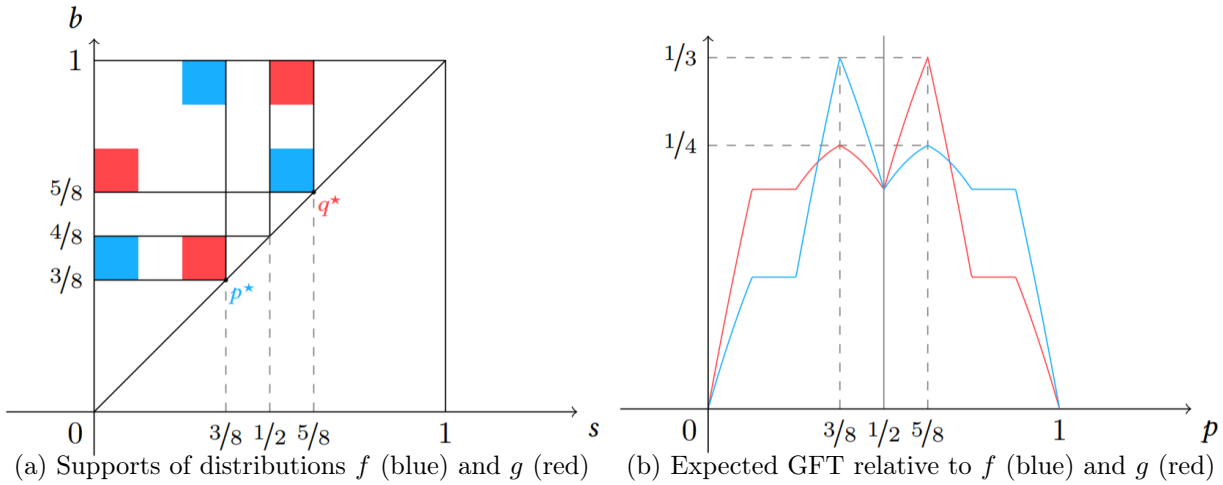
(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a joint density bounded by $M \geqslant 24$.*

(iid) *$(S_1, B_1), (S_2, B_2), \ldots$ is an i.i.d. sequence.*

Proof sketch. Consider the two joint densities f and g of the seller/buyer pair as the normalized indicator functions of the red and blue squares in Figure 2.3a. Formally

$$f \coloneqq \frac{64}{3}\left(\mathbb{I}_{[0/8,\,1/8]\times[3/8,\,4/8]} + \mathbb{I}_{[2/8,\,3/8]\times[7/8,\,8/8]} + \mathbb{I}_{[4/8,\,5/8]\times[5/8,\,6/8]}\right)$$

and $g(s,b) \coloneqq f(1-b, 1-s)$. In the f (resp., g) case, the optimal price belongs to the region $[0, 1/2]$ (resp., $(1/2, 1]$, see Figure 2.3b). By posting prices in the wrong region $(1/2, 1]$ (resp., $[0, 1/2]$) in the f (resp., g) case, the learner incurs at least a $1/3 - 1/4 = 1/12$ regret. Thus, the only way to avoid suffering linear regret is to determine if the valuations of the seller and buyer are generated by f or g. For each price $p \in [0,1]$, consider the four rectangles with opposite vertices (p,p) and (u_i, v_i), where $\{(u_i, v_i)\}_{i=1,\ldots,4}$ are the four vertices of the unit square. Note that the only information on the



(a) Supports of distributions f (blue) and g (red) (b) Expected GFT relative to f (blue) and g (red)

Figure 2.3: Under realistic feedback, the two densities f and g are indistinguishable. The optimal price $p^\star$ for f gives constant regret under g and $q^\star$ does the converse.

distribution of (S_t, B_t) that the learner can gather from the realistic feedback $\big(\mathbb{I}\{S_t \leqslant p\}, \mathbb{I}\{p \leqslant B_t\}\big)$ received after posting a price p is (an estimate of) the area of the portion of the support of the distribution included in each of these four rectangles. However, these areas coincide in the cases f and g. Hence, under realistic feedback, f and g are completely indistinguishable. Therefore, given that the optimal price in the f (resp., g) case is $3/8$ (resp., $5/8$), the best that the learner can do is to sample prices uniformly at random in the set $\{3/8, 5/8\}$, incurring a regret of $T/24$. For a formalization of this argument leveraging the techniques we described in the introduction, see Appendix A.9. □

2.4.4 Linear Lower Bound under Realistic Feedback (iv) + (iid)

In this section, we prove that in the realistic-feedback case, no strategy can achieve sublinear regret without any limitations on how concentrated the distributions of the valuations of the seller and buyer are, not even if they from an independent and identically distributed sequence (iid) and are independent of each other (iv).

At a high level, if the two distributions of the seller and the buyer are very concentrated in a small region, finding an optimal price is like finding a needle in a haystack. For a full proof of the following theorem, see Appendix A.10.

Theorem 8. *In the realistic-feedback model, for all horizons T, the minimax regret satisfies*

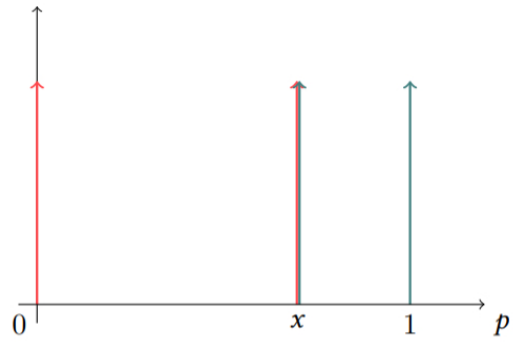
$$R_T^{\mathcal{S}} \geqslant cT\ ,$$

where $c \geqslant 1/8$, and $\mathcal{S}$ is the set of all environments such that

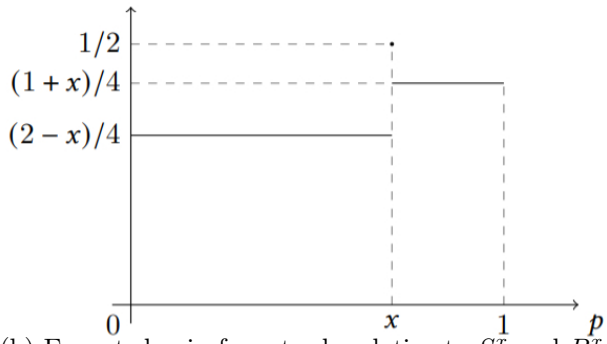
(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(iid) *$(S_1, B_1), (S_2, B_2), \ldots$ is an i.i.d. sequence.*

Proof sketch. Consider the family of seller/buyer i.i.d. sequences $(S^x, B^x), (S_1^x, B_1^x), (S_2^x, B_2^x), \ldots$, parameterized by $x \in I$, where I is a small interval centered in $1/2$, S^x and B^x are independent of



(a) Distributions of S^x (red) and B^x (green)



(b) Expected gain from trade relative to S^x and B^x

Figure 2.4: All prices but x have high regret. However, under realistic feedback, finding x in finite time is impossible.

each other, and they satisfy

$$S^x = \begin{cases} x & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{1}{2} \end{cases}, \qquad B^x = \begin{cases} x & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}.$$

The distributions and the corresponding gain from trade are represented in Figure 2.4a and Figure 2.4b, respectively. A direct verification shows that the function $p \mapsto \mathbb{E}\big[\mathrm{gft}\big(p, (S^x, B^x)\big)\big]$ is maximized at $p = x$. Furthermore, by posting any other prices, the learner incurs a regret of approximately $1/2$ with probability $1/4$. Now, under realistic feedback, no strategy can locate (exactly!) each possible $x \in I$ in a finite number of steps. This results, for any strategy, in regret of at least (approximately) $T/8$. See Appendix A.10 for a more detailed analysis. □

2.4.5 Linear Lower Bound under Realistic Feedback (iv) + (bd)

In this section, we prove that in the realistic-feedback case, assuming only that at each time t the evaluation of the seller S_t is independent of the valuation of the buyer B_t (iv) and that their distribution admits densities that are uniformly bounded (bd) is not enough to allow sublinear regret. The construction is based on ideas analogous to the one already proposed in section Section 2.4.3.

Theorem 9. *In the realistic-feedback model, for all horizons T, the minimax regret satisfies*

$$R_T^{\mathcal{S}} \geqslant cT \,,$$

where $c \geqslant 1/24$, and $\mathcal{S}$ is the set of all environments such that

(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a joint density bounded by $M \geqslant 64$.*

Proof. Consider the six squares depicted in Figure 2.3:

$$Q_1 \coloneqq \left[0, \tfrac{1}{8}\right] \times \left[\tfrac{3}{8}, \tfrac{1}{2}\right], \qquad Q_2 \coloneqq \left[\tfrac{1}{4}, \tfrac{3}{8}\right] \times \left[\tfrac{7}{8}, 1\right], \qquad Q_3 \coloneqq \left[\tfrac{1}{2}, \tfrac{5}{8}\right] \times \left[\tfrac{5}{8}, \tfrac{3}{4}\right],$$
$$Q_4 \coloneqq \left[\tfrac{1}{2}, \tfrac{5}{8}\right] \times \left[\tfrac{7}{8}, 1\right], \qquad Q_5 \coloneqq \left[0, \tfrac{1}{8}\right] \times \left[\tfrac{5}{8}, \tfrac{3}{4}\right], \qquad Q_6 \coloneqq \left[\tfrac{1}{4}, \tfrac{3}{8}\right] \times \left[\tfrac{3}{8}, \tfrac{1}{2}\right].$$

To each square Q_i, we associate a uniform probability distribution over it: we say that the random valuations (S, B) are distributed uniformly over Q_i under $\mathbb{P}^i$ and $\mathbb{E}^i$, for each $i = 1, \dots, 6$. Starting from these distributions, we construct two other distributions: the "red" one and the "blue" one. When (S, B) is sampled from the blue one, it is sampled uniformly at random from the union of the blue squares Q_1, Q_2 and Q_3. In formula, the probability measure $\mathbb{P}^{\text{blue}}$ is just a uniform mixture of $\mathbb{P}^1$, $\mathbb{P}^2$ and $\mathbb{P}^3$. The same can be done for the red distribution over the red squares Q_4, Q_5 and Q_6. Note that both the red and the blue distributions admit a density bounded by 64.

From Theorem 7, we know that any learning strategy α to post prices P_t suffers linear regret against at least one of the following i.i.d. instance: the environment is the one corresponding to the red or the blue distribution, and extracts valuations from it i.i.d. over the rounds. In formula:

$$\max_{\text{color}\in\{\text{red,blue}\}} \left(\max_{p\in[0,1]} \sum_{t=1}^{T} \mathbb{E}^{\text{color}} \big[\text{GFT}_t(p) - \text{GFT}_t(p_t) \big] \right) \geqslant \frac{1}{24} T. \tag{2.11}$$

We cannot use this construction directly for our result, as seller and buyer valuations are not independent in the blue and red distributions. However, we can generate an equivalent random sequence of distributions admitting a bounded density such that each one of them has *independent* seller and buyer valuations.

Consider the following family $\mathcal{S}_0 \subset \mathcal{S}$ of environments: each $\beta \in \mathcal{S}_0$ is characterized by a color red or blue, and a sequence $(i_t)_{t=1}^T$ of T indices, where red environments have $i_t \in \{4, 5, 6\}$ and blue environments have $i_t \in \{1, 2, 3\}$. We denote with $\mathcal{S}_0^{\text{red}}$ the set of all red environments and with $\mathcal{S}_0^{\text{blue}}$ the set of all blue environments. Any β in the sequence generates the valuations as follows: (S_t, B_t) is drawn independently and uniformly at random from Q_{i_t}. Note that any $\beta \in \mathcal{S}_0$ enjoys the property that the distribution chosen at each time step has independent seller and buyer. We argue that any learning strategy α suffers linear regret against at least one of these adversaries. In fact:

$$\begin{aligned} R_T^{\mathcal{S}}(\alpha) &\geqslant \max_{\beta\in\mathcal{S}_0} \left[\max_{p\in[0,1]} \left(\sum_{t=1}^{T} \mathbb{E}^{i_t} \left[\text{GFT}_t(p) - \text{GFT}_t(p_t) \right] \right) \right] \\ &= \max_{\text{color}\in\{\text{red,blue}\}} \max_{\beta\in\mathcal{S}_0^{\text{color}}} \left[\max_{p\in[0,1]} \left(\sum_{t=1}^{T} \mathbb{E}^{i_t} \left[\text{GFT}_t(p) - \text{GFT}_t(p_t) \right] \right) \right] \\ &\geqslant \max_{\text{color}\in\{\text{red,blue}\}} \left[\max_{p\in[0,1]} \left(\sum_{t=1}^{T} \mathbb{E}^{\text{color}} \left[\text{GFT}_t(p) - \text{GFT}_t(p_t) \right] \right) \right] \end{aligned} \tag{2.12}$$

Note that the i_t are the indices induced by β. The previous inequality, combined with Equation (2.11) concludes the proof. The only delicate step we need to clarify is the last inequality in Equation (2.12). To this end, fix any color, say red (the same argument holds for blue). The regret of α against the worst sequence in $\mathcal{S}_0^{\text{red}}$ is at least the expected regret of α against a randomized environment that is obtained by drawing uniformly at random β from $\mathcal{S}_0^{\text{red}}$. Now, the crucial argument is that the sequence of valuations (S_t, B_t) obtained by choosing uniformly at random an environment β from $\mathcal{S}_0^{\text{red}}$ follows the exact same distribution as drawing (S_t, B_t) i.i.d. from the red distribution. In fact, the valuations at different steps are independent and every square has the same probability of being chosen at each time step. □

2.4.6 Beyond Realistic Feedback - Learning with One Bit

In this brief section, we discuss the (im)possibility of learning with less than realistic feedback. Specifically, we investigate whether the single bit $\mathbb{I}\{S_t \leqslant P_t \leqslant B_t\}$ is sufficient for obtaining sublinear regret bounds. *This is not the case*: even under the (iv), (bd), and (iid) assumptions simultaneously, a single bit is not enough to provide sufficient observability. Indeed, consider a first instance in which the seller S and buyer B have uniform distributions on $[0, 1]$, independent of each other. In this case, the only maximizer of the expected gain from trade is $p^\star = 1/2$. As a second instance, consider two independent distributions of the seller S' and buyer B' with densities (bounded by 2 and even infinitely differentiable) $f_{S'}(s) \coloneqq 4(4 - 2s^3 + s^2)/(s^3 - s^2 + 4)^2$ and $f_{B'}(b) \coloneqq b(b - 1/2)(b - 1) + 1$ respectively. Then, for all $p \in [0, 1]$, we have $\mathbb{P}[S \leqslant p \leqslant B] = \mathbb{P}[S' \leqslant p \leqslant B']$. Therefore, the two instances are indistinguishable under the single-bit feedback, but a direct verification shows that in the second instance, $p^\star = 1/2$ is *not* a maximizer of the expected gain from trade. Leveraging these facts and the continuity of the gain from trade in the two instances leads to a linear minimax regret, using, for example, the same ideas as in Theorem 7.

2.5 Unlocking Faster Rates in Bilateral Trade: the $S \sim B$ Case.

In this section, we explore the case when sellers and buyers share the same distribution. In this case, we show that under the (iv), (bd), and (iid) assumptions faster rates are achievable both under full and realistic feedback. Interestingly, the upper and the lower bounds proposed in this section are matching (up to constants) not only in the time horizon but even in the density parameter M.

2.5.1 Brokerage with no Designated Seller's and Buyer's Roles

Sellers and buyers play starkly different roles in bilateral trade. It is understandable, therefore, that a reader might question the significance of the arguably artificial scenario where sellers and buyers possess identical distributions for their valuations. To address these concerns, we start by proposing real-world motivations for analyzing this case.

In several over-the-counter (OTC) markets (see Footnote $*$), traders do not have definite roles of sellers and buyers, but they may decide to sell or buy depending on the prevailing market conditions [168]. These markets encompass a wide array of asset trades, including stocks, derivatives, art, collectibles, precious metals and minerals, energy commodities like gas and oil, as well as digital currencies (cryptocurrencies), among others.

The interaction between brokers and traders in this scenario can be modeled in the following way. At each time $t \in \mathbb{N}$,

1. Two traders arrive with private valuations V_{2t-1} and V_{2t}.
2. The broker proposes a trading price P_t.
3. If the price P_t falls between the lowest** $V_{2t-1} \wedge V_{2t}$ and highest $V_{2t-1} \vee V_{2t}$ valuations (i.e., if the trader with the smallest valuation is eager to sell at price P_t and the other is willing to buy at P_t), the trader with the highest valuation buys the item from the trader with the lowest valuation paying the brokerage price P_t.

**We denote the minimum (resp., maximum) of any two real numbers $x, y \in \mathbb{R}$ by $x \wedge y$ (resp., $x \vee y$).

4. Some feedback is revealed (full or realistic).

Accordingly, the gain from trade function for this problem becomes

$$\widetilde{\mathrm{gft}}\colon [0,1] \times [0,1]^2 \to [0,1], \qquad \big(p, (v_1, v_2)\big) \mapsto |v_2 - v_1| \cdot \mathbb{I}\{v_1 \wedge v_2 \leqslant p \leqslant v_1 \vee v_2\}$$

while the gain from trade at time t becomes

$$\widetilde{\mathrm{GFT}}_t\colon [0,1] \to [0,1], \qquad p \mapsto \widetilde{\mathrm{gft}}\big(p, (V_{2t-1}, V_{2t})\big) \ .$$

Notice that this problem can be cast into the previously proposed bilateral trade framework in Section 2.1 by defining, for each time t, $S_t \coloneqq V_{2t-1} \wedge V_{2t}$ and $B_t \coloneqq V_{2t-1} \vee V_{2t}$. This way, for each $t \in \mathbb{N}$, we have the identity

$$\mathrm{GFT}_t = \widetilde{\mathrm{GFT}}_t \ .$$

Furthermore, the corresponding notions of full (resp., realistic) feedback for this new setting are enough to reconstruct the corresponding full (resp., realistic) feedback in the old setting. This is surely a feasible way to solve the problem with the tools we already provided in previous sections, but we now provide a different and fruitful perspective.

Suppose the sequence $(V_t)_{t\in\mathbb{N}}$ is independent and identically distributed (a sensible way to model large and stable markets). Then, for any $p \in [0,1]$

$$\begin{aligned}\mathbb{E}[\widetilde{\mathrm{GFT}}_t(p)] &= \mathbb{E}\big[(V_{2t-1} - V_{2t})\mathbb{I}\{V_{2t} \leqslant p \leqslant V_{2t-1}\} + (V_{2t} - V_{2t-1})\mathbb{I}\{V_{2t-1} \leqslant p \leqslant V_{2t}\}\big] \\ &= 2\mathbb{E}\big[(V_{2t} - V_{2t-1})\mathbb{I}\{V_{2t-1} \leqslant p \leqslant V_{2t}\}\big] = 2\mathbb{E}\Big[\mathrm{gft}\big(p, (V_{2t-1}, V_{2t})\big)\Big] \ .\end{aligned}$$

Hence, by defining instead $S_t \coloneqq V_{2t-1}$ and $B_t \coloneqq V_{2t}$, we obtain, for each $p \in [0,1]$,

$$\mathbb{E}[\widetilde{\mathrm{GFT}}_t(p)] = 2\mathbb{E}[\mathrm{GFT}_t(p)] \ .$$

With this different definition, the two sequences $(S_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$ are i.i.d. and, for each $t \in \mathbb{N}$, S_t and B_t share the same distribution and they are independent of each other. Moreover, notice that if the elements in the sequence $(V_t)_{t\in\mathbb{N}}$ admit a bounded density, the same holds also for those in the sequences $(S_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$.

We believe that this last reduction provides sufficient motivation to study the standard bilateral trade problem of Section 2.1 under the (iv) and (iid) assumptions when sellers' and buyers' evaluations share the same distribution, and to further investigate the role played by the (bd) assumption in this setting.

2.5.2 The Approximation and Representation Lemmas

In this section, we present two key results for the case when sellers and buyers are independent of each other, and they share the same distribution that admits a bounded density.

We first fix some notation. The Dirac measure based at $x \in \mathbb{R}$ is denoted by δ_x, i.e., δ_x is the measure defined via the equation $\delta_x[A] = \mathbb{I}\{x \in A\}$ for any set A. For any (signed) measure μ and any measurable set E, we will write μE rather than $\mu[E]$ whenever this does not cause confusion. For any measure μ over $[0,1]$, we let $\bar{\mu} \coloneqq \int_{[0,1]} x \,\mathrm{d}\mu(x)$, and we define the functions $\widetilde{\rho}(\mu)$ and $\rho(\mu)$,

for all $p \in [0,1]$, by

$$\widetilde{\rho}(\mu)(p) \coloneqq \int_0^p \big(\mu[0,\lambda] + \mu[0,\lambda)\big)\,\mathrm{d}\lambda + \big(\mu[0,p] + \mu[0,p)\big)(\bar{\mu} - p)\ ,$$
$$\rho(\mu)(p) \coloneqq \widetilde{\rho}(\mu)(p) + \mu\{p\}\left(\int_0^p \mu[0,\lambda]\,\mathrm{d}\lambda + \int_p^1 \mu[\lambda,1]\,\mathrm{d}\lambda\right)\ .$$

The following lemma shows that $\bar{\mu}$ maximizes $\widetilde{\rho}(\mu)$ (in general) and $\rho(\mu)$ (if μ has a bounded density), and the cost of approximating $\bar{\mu}$.

Lemma 3 (Approximation)**.** *If μ is a probability measure on $[0,1]$, then $\widetilde{\rho}(\mu)(\bar{\mu}) = \max_{p\in[0,1]} \widetilde{\rho}(\mu)(p)$ and, for any $p \in [0,1]$, $\widetilde{\rho}(\mu)(\bar{\mu}) - \widetilde{\rho}(\mu)(p) \leqslant 2|\bar{\mu} - p|$. If μ has a density bounded by $M > 0$, then $\rho(\mu) = \widetilde{\rho}(\mu)$ and*

$$0 \leqslant \rho(\mu)(\bar{\mu}) - \rho(\mu)(p) \leqslant M\,|\bar{\mu} - p|^2\ , \qquad \forall p \in [0,1]\ .$$

Proof. For $\lambda \in [0,1]$, let $m(\lambda) \coloneqq \mu[0,\lambda] + \mu[0,\lambda)$ and note that m is a $[0,1]$-valued non-decreasing function of λ. For any $p \in [0,1]$,

$$0 \leqslant \int_p^{\bar{\mu}} \big(m(\lambda) - m(p)\big)\,\mathrm{d}\lambda = \widetilde{\rho}(\mu)(\bar{\mu}) - \widetilde{\rho}(\mu)(p) = \int_p^{\bar{\mu}} \big(m(\lambda) - m(p)\big)\,\mathrm{d}\lambda \leqslant 2|\bar{\mu} - p|\ , \tag{2.13}$$

which implies that $\widetilde{\rho}(\mu)(\bar{\mu}) = \max_{p\in[0,1]} \widetilde{\rho}(\mu)(p)$. Next, note that for all $p \in [0,1]$, $\big|\widetilde{\rho}(\mu)(p) - \rho(\mu)(p)\big| \leqslant \mu\{p\}$, which, if μ has a density f bounded by a constant M, implies $\widetilde{\rho}(\mu)(p) = \rho(\mu)(p)$ and

$$\begin{aligned}\rho(\mu)(\bar{\mu}) - \rho(\mu)(p) = \widetilde{\rho}(\mu)(\bar{\mu}) - \widetilde{\rho}(\mu)(p) &\overset{(2.13)}{=} \int_p^{\bar{\mu}} \big(m(\lambda) - m(p)\big)\,\mathrm{d}\lambda \\ &= 2\int_p^{\bar{\mu}} \int_p^{\lambda} f(x)\,\mathrm{d}x\,\mathrm{d}\lambda \leqslant 2M\left|\int_p^{\bar{\mu}} |\lambda - p|\ \mathrm{d}\lambda\right| = M\,|\bar{\mu} - p|^2\ .\end{aligned}$$

□

The next lemma provides a crucial representation of the objective $p \mapsto \mathbb{E}\big[\mathrm{GFT}_t(p)\big]$. Its long and (somewhat) tedious proof is deferred to Appendix A.13.

Lemma 4 (Representation)**.** *Let S and B be two $[0,1]$-valued random variables independent of each other with common distribution ν. Then, the induced gain from trade $\mathrm{GFT}(\cdot) \coloneqq \mathrm{gft}\big(\cdot,(S,B)\big)$ is such that, for any $p \in [0,1]$,*

$$\mathbb{E}\big[\mathrm{GFT}(p)\big] = \frac{1}{2}\rho(\nu)(p)\ .$$

The following is an immediate corollary of Lemmas 3 and 4.

Theorem 10. *Suppose the sequence $(S_t, B_t)_{t\in N}$ of sellers and buyers' evaluations is i.i.d., and that, for each $t \in \mathbb{N}$, the evaluation S_t and B_t are independent of each other and they share the same distribution ν having a density bounded by some constant $M > 0$. Then, for any $t \in \mathbb{N}$ and any $p \in [0,1]$, it holds that*

$$0 \leqslant \mathbb{E}\big[\mathrm{GFT}_t\big(\bar{\nu}\big)\big] - \mathbb{E}\big[\mathrm{GFT}_t(p)\big] \leqslant \frac{M}{2} \cdot \big|\bar{\nu} - p\big|^2\ ,$$

and, in particular, $\max_{p\in[0,1]} \mathbb{E}\big[\mathrm{GFT}_t(p)\big] = \mathbb{E}\big[\mathrm{GFT}_t\big(\bar{\nu}\big)\big]$.

The previous theorem gives much intuition on the problem under the (iv), (bd) and (iid) assumptions, when sellers' and buyers' evaluations share the same distribution ν. It proves that the optimal action is to post the (unknown) expected value $\bar{\nu}$ of the valuations. Moreover, it suggests the strategy of approximating this value (on the basis of the available feedback), since posting a price close to the expectation has only a quadratic cost in the approximation.

2.5.3 Full Feedback - Upper Bound

In this section, we provide a logarithmic upper bound in the full feedback case under the (iv), (bd) and (iid) assumptions when sellers and buyers share the same distribution.

Following the intuition provided by Theorem 10, we introduce the Follow the Mean algorithm (FTM), which simply posts the empirical average of the past valuations.

Algorithm 4 Follow the Mean (FTM) - Full Feedback

Let $P_1 \coloneqq 1/2$
for $t = 1, 2, \dots$ **do**
  Post price P_t
  Let $P_{t+1} \coloneqq \frac{1}{2t} \sum_{i=1}^{t} (S_i + B_i)$

The next theorem shows that FTM enjoys an $M \log T$ regret, where M is an upper bound on the density of the distribution shared by sellers and buyers.

Theorem 11. *Consider the problem of repeated bilateral trade in the full-feedback model. Suppose that $\mathcal{S}$ is the set of environments such that the sequence of evaluations $(S_1, B_1), (S_2, B_2), \dots$ is independent and identically distributed (iid), all with the same law as some (S, B), where S and B are independent (iv) of each other and share the same distribution μ having a density (with respect to the Lebesgue measure on $[0,1]$), bounded by some constant M (bd). If we run Follow the Mean (FTM), then, for any time horizon $T \geqslant 2$, we have*

$$R_T^{\mathcal{S}}(\mathrm{FTM}) \leqslant \frac{1}{2} + \frac{M}{8}\big(1 + \ln(T-1)\big) \, .$$

Proof. For notational convenience, let V a random variable with the same distribution of S (and hence, B). For all time horizons $T \geqslant 2$, we have

$$\begin{aligned}
R_T^{\mathcal{S}}\big(\mathrm{FTM}, (S_t, B_t)_{t\in\mathbb{N}}\big) - \frac{1}{2} &\leqslant \sum_{t=2}^{T} \Big(\mathbb{E}\big[\mathrm{GFT}_t\big(\mathbb{E}[V]\big)\big] - \mathbb{E}\big[\mathrm{GFT}_t(P_t)\big] \Big) \\
&\overset{(l)}{=} \sum_{t=2}^{T} \mathbb{E}\left[\Big[\mathbb{E}\big[\mathrm{GFT}_t\big(\mathbb{E}[V]\big) - \mathrm{GFT}_t(p)\big] \Big]_{p=P_t} \right] \\
&\overset{(t)}{\leqslant} \sum_{t=2}^{T} \mathbb{E}\left[\Big[\frac{M}{2} \big| p - \mathbb{E}[V] \big|^2 \Big]_{p=P_t} \right] = \frac{M}{2} \sum_{t=2}^{T} \mathbb{E}\Big[\big| P_t - \mathbb{E}[V] \big|^2 \Big] \\
&\overset{(f)}{=} \frac{M}{2} \sum_{t=2}^{T} \int_0^\infty \mathbb{P}\Big[\big| P_t - \mathbb{E}[V] \big|^2 \geqslant \varepsilon \Big] \, \mathrm{d}\varepsilon \overset{(h)}{\leqslant} \frac{M}{2} \sum_{t=1}^{T-1} \int_0^\infty 2e^{-8t\varepsilon} \, \mathrm{d}\varepsilon \\
&= \frac{M}{8} \sum_{t=1}^{T-1} \frac{1}{t} \leqslant \frac{M}{8} \left(1 + \int_1^{T-1} \frac{1}{s} \, \mathrm{d}s \right) \leqslant \frac{M}{8} \big(1 + \ln(T-1) \big) \, ,
\end{aligned}$$

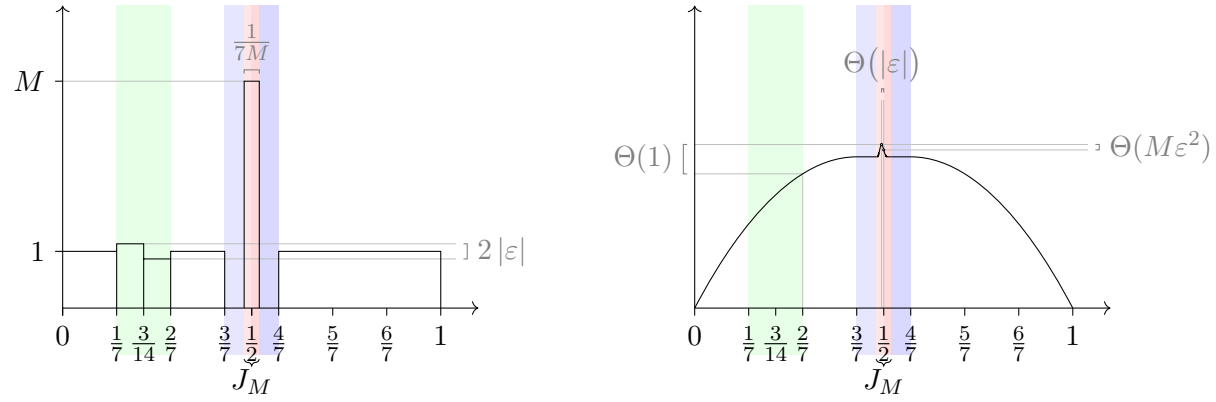


Figure 2.5: On the left, the density f_ε of a "hard" instance used to prove the lower bounds in Theorems 12 and 15. A base uniform distribution is warped in the intervals $[1/7, 2/7]$ (green) and $[3/7, 4/7]$ (blue+red). The density on $[1/7, 2/7]$ is split into two uneven parts, differing by ε from the original. The mass on $[3/7, 4/7]$ is concentrated in a small set J_M of size $\Theta(1/M)$ around $1/2$. The corresponding gain from trade, on the right, has a smooth spike of height $\Theta(M|\varepsilon|^2)$ situated in J_M, at a distance $\Theta(|\varepsilon|)$ from $1/2$. When $\varepsilon < 0$ (resp., $\varepsilon > 0$), the spike is left (resp., right) of $1/2$, and posting $1/2$ is better than posting any price after (resp., before) $1/2$. In the two-bit feedback lower bound, the only way to gather usable feedback is to post prices in $[1/7, 2/7]$, which give rewards $\Theta(1)$-away from the optimal one.

where (l) follows from the Freezing Lemma (Lemma 17) after observing that P_t is independent of (S_t, B_t) and $\mathrm{GFT}_t(P_t) = \mathrm{gft}\big(P_t, (S_t, B_t)\big)$; (t) from Theorem 10; (f) follows from Fubini's Theorem; and (h) from Hoeffding's inequality. □

2.5.4 Full Feedback - Lower Bound

In this section, we prove the optimality of the FTM algorithm by showing a matching $M \log T$ lower bound. To this end, we build a family of distributions to show that the problem is harder than a full-feedback sequential Bayesian problem where the goal is to estimate the expectation of a certain random variable, and the loss function is the square of the distance from the expectation.

Theorem 12. *Consider the problem of repeated bilateral trade in the realistic-feedback model. There exists two numerical constants $c_1, c_2 > 0$ such that, for any $M \geqslant 2$ and any time horizon $T \geqslant c_2 M^4$, the minimax regret satisfies*

$$R_T^{\mathcal{S}} \geqslant c_1 M \log T \ ,$$

where $\mathcal{S}$ is the set of all environments such that, for each $t \in \mathbb{N}$, S_t and B_t share the same distribution ν and

(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(bd) *for each $t \in \mathbb{N}$, ν admits a density bounded by M.*

(iid) *$(S_1, B_1), (S_2, B_2), \dots$ is an i.i.d. sequence.*

Proof. Given that we are in a stochastic i.i.d. setting, we can restrict this proof to deterministic strategies without loss of generality. Let $M \geqslant 2$, $J_M \coloneqq \left[\frac{1}{2} - \frac{1}{14M}, \frac{1}{2} + \frac{1}{14M}\right]$, $f \coloneqq \mathbb{I}_{\left[0, \frac{3}{7}\right]} + M\mathbb{I}_{J_M} + \mathbb{I}_{\left[\frac{4}{7}, 1\right]}$, and, for any $\varepsilon \in [-1, 1]$, $g_\varepsilon \coloneqq -\varepsilon\mathbb{I}_{\left[\frac{1}{7}, \frac{3}{14}\right]} + \varepsilon\mathbb{I}_{\left(\frac{3}{14}, \frac{2}{7}\right]}$ and $f_\varepsilon \coloneqq f + g_\varepsilon$ (see Figure 2.5, left). For any $\varepsilon \in [-1, 1]$, note that $0 \leqslant f_\varepsilon \leqslant M$ and $\int_0^1 f_\varepsilon(x)\,\mathrm{d}x = 1$, hence f_ε is a valid density on $[0, 1]$

bounded by M, and we will denote the corresponding probability measure by ν_ε. Consider for each $q \in [0,1]$, an i.i.d. sequence $(D_{q,t})_{t\in\mathbb{N}}$ of Bernoulli random variables of parameter q, an i.i.d. sequence $(\widetilde{D}_t)_{t\in\mathbb{N}}$ of Bernoulli random variables of parameter $1/7$, an i.i.d. sequence $(U_t)_{t\in\mathbb{N}}$ of uniform random variables on $[0,1]$, a uniform random variable E on $[-\varepsilon_M, \varepsilon_M]$ where $\varepsilon_M := \frac{7}{M}$, such that $\Big((D_{q,t})_{t\in\mathbb{N}, q\in[0,1]}, (\widetilde{D}_t)_{t\in\mathbb{N}}, (U_t)_{t\in\mathbb{N}}, E\Big)$ is an independent family. Let $\varphi\colon [0,1] \to [0,1]$ be such that, if U is a uniform random variable on $[0,1]$, then the distribution of $\varphi(U)$ has density $\frac{7}{6} \cdot f \cdot \mathbb{I}_{[0,1]\setminus[1/7,2/7]}$ (which exists by the Skorokhod representation theorem [185, Section 17.3]). For each $\varepsilon \in [-1,1]$ and $t \in \mathbb{N}$, define

$$\begin{aligned} S_{\varepsilon,t} &\coloneqq \left(\frac{2+U_{2t-1}}{14}(1 - D_{\frac{1+\varepsilon}{2},2t-1}) + \frac{3+U_{2t-1}}{14} D_{\frac{1+\varepsilon}{2},2t-1}\right)\widetilde{D}_{2t-1} + \varphi(U_{2t-1})(1-\widetilde{D}_{2t-1})\,, \\ B_{\varepsilon,t} &\coloneqq \left(\frac{2+U_{2t}}{14}(1 - D_{\frac{1+\varepsilon}{2},2t}) + \frac{3+U_{2t}}{14} D_{\frac{1+\varepsilon}{2},2t}\right)\widetilde{D}_{2t} + \varphi(U_{2t})(1-\widetilde{D}_{2t})\,. \end{aligned} \tag{2.14}$$

Straightforward computations show that, for each $\varepsilon \in [-1,1]$ the sequence $(S_{\varepsilon,t}, B_{\varepsilon,t})_{t\in\mathbb{N}}$ is i.i.d. and for each $t \in \mathbb{N}$ the random variables $S_{\varepsilon,t}$ and $B_{\varepsilon,t}$ are independent of each other with commonly shared distribution given by ν_ε. Furthermore, for each $\varepsilon \in [-1,1]$ the sequence $(S_{\varepsilon,t}, B_{\varepsilon,t})_{t\in\mathbb{N}}$ is independent of E. For any $\varepsilon \in [-1,1]$, $p \in [0,1]$, and $t \in \mathbb{N}$, let $\mathrm{GFT}_{\varepsilon,t}(p) \coloneqq \mathrm{gft}\big(p, (S_{\varepsilon,t}, B_{\varepsilon,t})\big)$ (for a qualitative representation of its expectation, see Figure 2.5, right). For any $\varepsilon \in [-1,1]$ and $t \in \mathbb{N}$, a direct computation shows that $\bar{\nu}_\varepsilon = \mathbb{E}[V_{\varepsilon,t}] = \frac{1}{2} + \frac{\varepsilon}{196}$. By Lemmas 3 and 4, we have, for all $\varepsilon \in [-1,1], t \in \mathbb{N}$, and $p \in [0,1]$,

$$\mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] = \int_0^p \int_0^\lambda f_\varepsilon(s)\,\mathrm{d}s\,\mathrm{d}\lambda + (\bar{\nu}_\varepsilon - p)\int_0^p f_\varepsilon(s)\,\mathrm{d}s\,,$$

which, together with the fundamental theorem of calculus —[29, Theorem 14.16], noting that $p \mapsto \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big]$ is absolutely continuous with derivative defined a.e. by $p \mapsto (\bar{\nu}_\varepsilon - p) f_\varepsilon(p)$— yields, for any $p \in J_M$,

$$\mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(\bar{\nu}_\varepsilon)\big] - \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] = \frac{M}{2}|\bar{\nu}_\varepsilon - p|^2\,. \tag{2.15}$$

Note also that for all $\varepsilon \in [-\varepsilon_M, \varepsilon_M]$, $t \in \mathbb{N}$, and $p \in [0,1]\setminus J_M$,

$$\mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] \leqslant \mathbb{E}\left[\mathrm{GFT}_{\varepsilon,t}\left(1/2\right)\right]\,. \tag{2.16}$$

Fix any arbitrary deterministic strategy for the full feedback setting $(\widetilde{\alpha}_t)_{t\in\mathbb{N}}$, i.e., a sequence of functions $\widetilde{\alpha}_t\colon \big([0,1]\times[0,1]\big)^{t-1} \to [0,1]$ mapping past feedback into prices (with the convention that $\widetilde{\alpha}_1$ is just a number in $[0,1]$). For each $t \in \mathbb{N}$, define $\alpha_t\colon \big([0,1]\times[0,1]\big)^{t-1} \to J_M$ equal to $\widetilde{\alpha}_t$ whenever $\widetilde{\alpha}_t$ takes values in J_M, and equal to $1/2$ otherwise. Defining $Z \coloneqq \frac{1+E}{2}$, and R_T^ν as the regret of the strategy $(\widetilde{\alpha}_t)_{t\in\mathbb{N}}$ at time T when the underlying sequence of sellers' and buyers' evaluations follows the distribution ν, we have that the regret $R_T^{\mathcal{S}}\left((\widetilde{\alpha}_t)_{t\in\mathbb{N}}\right)$ is lower bounded by

$$\begin{aligned} &\sup_{\varepsilon\in[-\varepsilon_M,\varepsilon_M]} \sum_{t=1}^T \mathbb{E}\Big[\mathrm{GFT}_{\varepsilon,t}(\bar{\nu}_\varepsilon) - \mathrm{GFT}_{\varepsilon,t}\left(\widetilde{\alpha}_t\big((S_{\varepsilon,1},B_{\varepsilon,1}),\dots,(S_{\varepsilon,t-1},B_{\varepsilon,t-1})\big)\right)\Big] \\ &\overset{(2.16)}{\geqslant} \sup_{\varepsilon\in[-\varepsilon_M,\varepsilon_M]} \sum_{t=1}^T \mathbb{E}\Big[\mathrm{GFT}_{\varepsilon,t}(\bar{\nu}_\varepsilon) - \mathrm{GFT}_{\varepsilon,t}\left(\alpha_t\big((S_{\varepsilon,1},B_{\varepsilon,1}),\dots,(S_{\varepsilon,t-1},B_{\varepsilon,t-1})\big)\right)\Big] \end{aligned}$$

$$
\begin{aligned}
&\overset{\spadesuit}{=} \frac{M}{2} \sup_{\varepsilon \in [-\varepsilon_M, \varepsilon_M]} \sum_{t=1}^{T} \mathbb{E}\Big[\big|\bar{\nu}_\varepsilon - \alpha_t\big((S_{\varepsilon,1}, B_{\varepsilon,1}), \dots, (S_{\varepsilon,t-1}, B_{\varepsilon,t-1})\big)\big|^2\Big] \\
&\geqslant \frac{M}{2} \sum_{t=1}^{T} \mathbb{E}\Big[\big|\bar{\nu}_E - \alpha_t\big((S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})\big)\big|^2\Big] \\
&\overset{\heartsuit}{\geqslant} \frac{M}{2} \sum_{t=1}^{T} \mathbb{E}\Big[\big|\bar{\nu}_E - \mathbb{E}[\bar{\nu}_E \mid (S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})]\big|^2\Big] \\
&= \frac{M}{384} \sum_{t=1}^{T} \mathbb{E}\Big[\big|E - \mathbb{E}[E \mid (S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})]\big|^2\Big] \\
&\overset{\blacklozenge}{\geqslant} \frac{M}{384} \sum_{t=1}^{T} \mathbb{E}\Big[\big|E - \mathbb{E}[E \mid D_{\frac{1+E}{2},1}, \dots, D_{\frac{1+E}{2},2(t-1)}]\big|^2\Big] = \\
&= \frac{M}{196} \sum_{t=1}^{T} \mathbb{E}\Big[\big|Z - \mathbb{E}[Z \mid D_{Z,1}, \dots, D_{Z,2(t-1)}]\big|^2\Big]
\end{aligned}
$$

where ♠ follows from (2.15) and the fact that α_t takes values in J_M; ♥ from the fact that the minimizer of the $L^2(\mathbb{P})$-distance from $\bar{\nu}_E$ in $\sigma\left((S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})\right)$ is the random variable $\mathbb{E}[\bar{\nu}_E \mid (S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})]$ (see, e.g., [185, Section 9.4]); ♦ follows from the fact that, by Equation (2.14) and the independence of E from $\Big((D_{q,t})_{t\in\mathbb{N}, q\in[0,1]}, (\widetilde{D}_t)_{t\in\mathbb{N}}, (U_t)_{t\in\mathbb{N}}\Big)$, the conditional expectation $\mathbb{E}[E \mid (S_{E,1}, B_{E,1}), \dots, (S_{E,t-1}, B_{E,t-1})]$ is a measurable function of $D_{\frac{1+E}{2},1}, \dots, D_{\frac{1+E}{2},2(t-1)}$, together with the same observation made in ♥ about the minimization of $L^2(\mathbb{P})$ distance.

Finally, the general term of this last sum is the expected squared distance between the random parameter (drawn uniformly over $[(1-\varepsilon_M)/2, (1+\varepsilon_M)/2]$) of an i.i.d. sequence of Bernoulli random variables and the conditional expectation of this random parameter given $2(t-1)$ independent realizations of these Bernoullis. A probabilistic argument shows that there exist two universal constants $\widetilde{c}, c_2 > 0$ such that, for all $t \geqslant c_2 M^4$,

$$
\mathbb{E}\Big[\big|Z - \mathbb{E}[Z \mid D_{Z,1}, \dots, D_{Z,2(t-1)}]\big|^2\Big] \geqslant \widetilde{c}\frac{1}{t-1} \,. \tag{2.17}
$$

At a high level, this is because, in an event of probability $\Omega(1)$, if t is large enough, the conditional expectation $\mathbb{E}[Z \mid D_{Z,1}, \dots, D_{Z,2(t-1)}]$ is very close to the empirical average $\frac{1}{2(t-1)}\sum_{s=1}^{2(t-1)} D_{Z,s}$, whose expected squared distance from Z is $\Omega\big(1/(t-1)\big)$. For a formal proof (2.17) with explicit constants, see the appendix Appendix A.14. Summing over t and putting everything together gives the result. □

What if we remove the (bd) assumption?

We conclude this section by showing that assuming that sellers' and buyers' evaluations are independent of each other, together with the (iv) and (iid) assumptions, is not enough to achieve faster rates than those already obtained in Section 2.3.

Theorem 13. *Consider the problem of repeated bilateral trade in the full-feedback model. There exists a numerical constant $c > 0$ such that, for any time horizon T, the minimax regret satisfies*

$$
R_T^{\mathcal{S}} \geqslant c\sqrt{T} \,,
$$

where $\mathcal{S}$ is the set of all environments such that, for each $t \in \mathbb{N}$, S_t and B_t share the same distribution ν and

(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(iid) *$(S_1, B_1), (S_2, B_2), \dots$ is an i.i.d. sequence.*

Proof sketch. For each $\varepsilon \in \left[-\frac{1}{4}, \frac{1}{4}\right]$, consider the distribution

$$\nu_\varepsilon \coloneqq \frac{1}{4}\delta_0 + \left(\frac{1}{4} + \varepsilon\right)\delta_{1/3} + \left(\frac{1}{4} - \varepsilon\right)\delta_{2/3} + \frac{1}{4}\delta_1 \ .$$

Consider for each $\varepsilon \in \left[-\frac{1}{4}, \frac{1}{4}\right]$ an i.i.d. sequence $(D_{\varepsilon,t})_{t\in\mathbb{N}}$ of Bernoulli random variables of parameter $\frac{1}{2} + 2\varepsilon$, and consider two i.i.d. sequences $(D_t)_{t\in\mathbb{N}}, (\widetilde{D}_t)_{t\in\mathbb{N}}$ of parameter $1/2$, such that the family of random variables $\Big((D_{\varepsilon,t})_{\varepsilon\in[-\frac{1}{4},\frac{1}{4}],t\in\mathbb{N}}, (D_t)_{t\in\mathbb{N}}, (\widetilde{D}_t)_{t\in\mathbb{N}}\Big)$ is an independent family. For each $t \in \mathbb{N}$ and each $\varepsilon \in [-\frac{1}{4}, \frac{1}{4}]$, define

$$\begin{aligned} S_{\varepsilon,t} &\coloneqq \frac{1}{3}(1 - D_{2t-1})D_{\varepsilon,2t-1} + \frac{2}{3}(1 - D_{2t-1})(1 - D_{\varepsilon,2t-1}) + D_{2t-1}\widetilde{D}_{2t-1} \ , \\ B_{\varepsilon,t} &\coloneqq \frac{1}{3}(1 - D_{2t})D_{\varepsilon,2t} + \frac{2}{3}(1 - D_{2t})(1 - D_{\varepsilon,2t}) + D_{2t}\widetilde{D}_{2t} \ . \end{aligned}$$

and, for each $\varepsilon \in \left[-\frac{1}{4}, \frac{1}{4}\right]$, notice that $(S_{\varepsilon,t})_{t\in\mathbb{N}}$ and $(B_{\varepsilon,t})_{t\in\mathbb{N}}$ are two i.i.d sequences, such that for each $t \in \mathbb{N}$ the two random variables $S_{\varepsilon,t}$ and $B_{\varepsilon,t}$ are independent of each other with common distribution ν_ε. For any $\varepsilon \in \left[-\frac{1}{4}, \frac{1}{4}\right]$, $p \in [0,1]$, and $t \in \mathbb{N}$, let $\mathrm{GFT}_{\varepsilon,t}(p) \coloneqq \mathrm{gft}\left(p, (S_{\varepsilon,t}, B_{\varepsilon,t})\right)$. For each $\varepsilon \in \left[-\frac{1}{8}, \frac{1}{8}\right]$ and each $t \in \mathbb{N}$, note that:

$$\max_{p\in\left\{\frac{1}{3},\frac{2}{3}\right\}} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] = \max_{p\in[0,1]} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] \tag{2.18}$$

$$\min_{p\in\left\{\frac{1}{3},\frac{2}{3}\right\}} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] - \max_{p\in[0,1]\setminus\left\{\frac{1}{3},\frac{2}{3}\right\}} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] = \Omega(1) \tag{2.19}$$

$$\mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(1/3)\big] - \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(2/3)\big] = \mathrm{sgn}(\varepsilon) \cdot \Omega\big(|\varepsilon|\big) \tag{2.20}$$

Fix a time horizon $T \in \mathbb{N}$ and select $\varepsilon \coloneqq T^{-1/2}$. We will show that for each strategy for the full-feedback setting and each time horizon T, if R_T^ν is the regret of the strategy at time horizon T when the underlying distribution of the traders' valuations is ν, then $\max\big(R_T^{\nu-\varepsilon}, R_T^{\nu+\varepsilon}\big) = \Omega\big(\sqrt{T}\big)$. Notice that, by posting prices in the wrong region $[0,1]\setminus\{1/3\}$ (resp., $[0,1]\setminus\{2/3\}$) in the $+\varepsilon$ (resp., $-\varepsilon$) case, the learner incurs a $\Omega(\varepsilon) = \Omega\big(1/\sqrt{T}\big)$ instantaneous regret by (2.18), (2.19), and (2.20). Then, in order to attempt suffering less than $\Omega\big(1/\sqrt{T} \cdot T\big) = \Omega\big(\sqrt{T}\big)$ regret, the algorithm would have to detect the sign of $\pm\varepsilon$ and play accordingly. However, the algorithm has no means to gather enough information to accomplish this task in due time. In fact, notice that the feedback received from the two traders at time t after having posted a price p is $S_{\pm\varepsilon,t}$ and $B_{\pm\varepsilon,t}$, which can't give more information about $\pm\varepsilon$ than the information carried by the two Bernoullis $D_{\pm\varepsilon,2t-1}$ and $D_{\pm\varepsilon,2t}$. Since (via an information-theoretic argument) in order to distinguish the sign of $\pm\varepsilon$ having access to i.i.d. Bernoulli random variables of parameter $\frac{1}{2} \pm 2\varepsilon$ requires $\Omega(1/\varepsilon^2) = \Omega(T)$ samples, the algorithm will have already suffered a regret $\Omega\big(T\big) \cdot \Omega(1/\sqrt{T}) = \Omega\big(\sqrt{T}\big)$ before having the chance to distinguish the sign of $\pm\varepsilon$. □

2.5.5 Realistic Feedback - Upper Bound

In this section, we provide a $\sqrt{MT}$ upper bound in the realistic feedback case under the (iv), (bd) and (iid) assumptions when sellers and buyers share the same distribution.

Motivated once more by the intuition provided by Theorem 10, we begin this section by giving a way to approximate the expected value of traders' valuations on the basis of the two-bit feedback and quantify the approximation power of this strategy.

Lemma 5. *For any random variable X on $[0,1]$ and any $T_0 \in \mathbb{N}$,*

$$0 \leqslant \mathbb{E}[X] - \frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{P}\left[\frac{t}{T_0} \leqslant X\right] \leqslant \frac{1}{T_0}$$

Proof. Notice that

$$\begin{aligned}
\frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{P}\left[\frac{t}{T_0} \leqslant X\right] &= \sum_{t=1}^{T_0} \int_{\frac{t-1}{T_0}}^{\frac{t}{T_0}} \mathbb{P}\left[\frac{t}{T_0} \leqslant X\right] \mathrm{d}\lambda \\
&\leqslant \sum_{t=1}^{T_0} \int_{\frac{t-1}{T_0}}^{\frac{t}{T_0}} \mathbb{P}\left[\lambda \leqslant X\right] \mathrm{d}\lambda \\
&\leqslant \sum_{t=1}^{T_0} \int_{\frac{t-1}{T_0}}^{\frac{t}{T_0}} \mathbb{P}\left[\frac{t-1}{T_0} \leqslant X\right] \mathrm{d}\lambda = \frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{P}\left[\frac{t-1}{T_0} \leqslant X\right] \mathrm{d}\lambda \,.
\end{aligned}$$

Since by Fubini's Theorem,

$$\mathbb{E}[X] = \int_0^1 \mathbb{P}[\lambda \leqslant X]\,\mathrm{d}\lambda = \sum_{t=1}^{T_0} \int_{\frac{t-1}{T_0}}^{\frac{t}{T_0}} \mathbb{P}[\lambda \leqslant X]\,\mathrm{d}\lambda \,,$$

we obtain

$$\begin{aligned}
0 \leqslant T_0\mathbb{E}[X] - \sum_{t=1}^{T_0} \mathbb{P}\left[\frac{t}{T_0} \leqslant X\right] &\leqslant \sum_{t=1}^{T_0} \left(\mathbb{P}\left[\frac{t-1}{T_0} \leqslant X\right] - \mathbb{P}\left[\frac{t}{T_0} \leqslant X\right]\right) \\
= \sum_{t=1}^{T_0} \mathbb{P}\left[\frac{t-1}{T_0} \leqslant X < \frac{t}{T_0}\right] &= \mathbb{P}[0 \leqslant X < 1] \leqslant 1 \,.
\end{aligned}$$

□

The previous lemma suggests the design of a simple Explore-then-Commit (ETC) strategy (Algorithm 5), where the learner spends an initial phase of length T_0 trying the common expectation of the sellers' and buyers' evaluations and then posts this estimate every round up to the time horizon T.

ETC algorithms are of great practical importance due to their easy implementability and interpretability. This usually comes at a cost of performance. As the following result (together with Theorem 15, in the next section) will show, our ECT algorithm is free of this flaw.

Theorem 14. *Consider the problem of repeated bilateral trade in the realistic-feedback model. Suppose that $\mathcal{S}$ is the set of environments such that the sequence of evaluations $(S_1, B_1), (S_2, B_2), \ldots$ is independent and identically distributed (iid), all with the same law as some (S, B), where S and B*

Algorithm 5 Explore-then-Commit (ETC) - Realistic Feedback

input: Exploration time $T_0 \in \mathbb{N}$
for $t = 1, 2, \dots, T_0$ **do**
  Post price $P_t \coloneqq t/T_0$
for $t = T_0 + 1, T_0 + 2, \dots$ **do**
  Post $P_t \coloneqq \frac{1}{2T_0} \sum_{i=1}^{T_0} \big(\mathbb{I}\{P_s \leqslant S_i\} + \mathbb{I}\{P_s \leqslant B_i\}\big)$

are independent (iv) and share the same distribution μ having a density (with respect to the Lebesgue measure on $[0,1]$), bounded by some constant M (bd). If we run Explore-then-Commit (ETC) with parameter $T_0 \in \mathbb{N}$, then, for any time horizon T, we have

$$R_T^{\mathcal{S}}(\mathrm{ETC}) \leqslant T_0 - \frac{1}{2} + \frac{M}{2}(T - T_0)\left(\frac{2}{T_0^2} + \frac{1}{T_0}\right)$$

Tuning the parameter $T_0 \coloneqq \left\lceil \sqrt{\frac{MT}{2}} \right\rceil$ yields

$$R_T^{\mathcal{S}}(\mathrm{ETC}) \leqslant 2.5 + \sqrt{2MT}\,.$$

Proof. For notational convenience, let V be another random variable with the same distribution as S (and hence, also as B). Fix any $T_0 \in \mathbb{N}$ and let $p_0 \coloneqq \frac{1}{T_0} \sum_{s=1}^{T_0} \mathbb{P}\left[\frac{s}{T_0} \leqslant V\right]$. By Hoeffding's inequality and Fubini's theorem, we get

$$\mathbb{E}\left[\left|p_0 - P_{T_0+1}\right|^2\right] = \int_0^{+\infty} \mathbb{P}\left[\left|p_0 - P_{T_0+1}\right|^2 \geqslant \varepsilon\right] \mathrm{d}\varepsilon \leqslant \int_0^{+\infty} 2\exp(-4\varepsilon T_0)\,\mathrm{d}\varepsilon = \frac{1}{2T_0}\,,$$

from which, leveraging also Lemma 5, it follows that

$$\mathbb{E}\Big[\left|\mathbb{E}[V] - P_{T_0+1}\right|^2\Big] \leqslant 2\left|\mathbb{E}[V] - p_0\right|^2 + 2\mathbb{E}\Big[\left|p_0 - P_{T_0+1}\right|^2\Big] \leqslant \frac{2}{T_0^2} + \frac{1}{T_0}\,.$$

Proceeding as in the proof of Theorem 11, we obtain, for all $t \in \mathbb{N}$,

$$\mathbb{E}\Big[\mathrm{GFT}_t\big(\mathbb{E}[V]\big) - \mathrm{GFT}_t(P_t)\Big] \leqslant \frac{M}{2}\mathbb{E}\Big[\left|\mathbb{E}[V] - P_t\right|^2\Big]\,.$$

Putting everything together, we get, for all $T \geqslant T_0 + 1$

$$\begin{aligned} R_T\big(\mathrm{ETC}, (S_t, B_t)_{t\in\mathbb{N}}\big) - T_0 + \frac{1}{2} &\leqslant \sum_{t=T_0+1}^{T} \mathbb{E}\Big[\mathrm{GFT}_t\big(\mathbb{E}[V]\big) - \mathrm{GFT}_t(P_t)\Big] \\ &\leqslant \frac{M}{2}\sum_{t=T_0+1}^{T} \mathbb{E}\Big[\left|\mathbb{E}[V] - P_t\right|^2\Big] = \frac{M}{2}\sum_{t=T_0+1}^{T} \mathbb{E}\Big[\left|\mathbb{E}[V] - P_{T_0+1}\right|^2\Big] \\ &\leqslant \frac{M}{2}(T - T_0)\left(\frac{2}{T_0^2} + \frac{1}{T_0}\right)\,. \end{aligned}$$

Substituting the selected parameters in the final expression yields the second part of the result. □

2.5.6 Realistic Feedback - Lower Bound

In this section, we prove the optimality of the ETC algorithm by showing a matching $\sqrt{MT}$ lower bound. The same family of distribution used in Theorem 12 is here used to mimic a revealing action problem. The reason why we obtain a $\Omega(\sqrt{T})$ regret regime (instead of a $\Omega(T^{2/3})$) is due to the fact that, as shown in Theorem 10, by posting a certain price p, the learner pays only order of the square of the distance of p from the actual optimum.

Theorem 15. *Consider the problem of repeated bilateral trade in the realistic-feedback model. There exists two numerical constants $c_1, c_2 > 0$ such that, for any $M \geqslant 2$ and any time horizon $T \geqslant c_2 M^3$, the minimax regret satisfies*

$$R_T^{\mathcal{S}} \geqslant c_1 \sqrt{MT}\ ,$$

where $\mathcal{S}$ is the set of all environments such that, for each $t \in \mathbb{N}$, S_t and B_t share the same distribution ν and

(iv) *for each $t \in \mathbb{N}$, S_t and B_t are independent of each other.*

(bd) *for each $t \in \mathbb{N}$, ν admits a density bounded by M.*

(iid) *$(S_1, B_1), (S_2, B_2), \dots$ is an i.i.d. sequence.*

Proof sketch. Fix $M \geqslant 2$ and $T \in \mathbb{N}$. We will use the same random variables, distributions, densities, and notation as in the proof of Theorem 12. We will show that for each strategy for the realistic feedback setting and each time horizon T, if R_T^ν is the regret of that strategy at time horizon T when the underlying common distribution of the sellers' and buyers' evaluations is ν, then $\max\big(R_T^{\nu-\varepsilon}, R_T^{\nu+\varepsilon}\big) = \Omega\big(\sqrt{MT}\big)$ if $T = \Omega(M^3)$.

Note that for all $\varepsilon > 0$, $t \in \mathbb{N}$, and $p < \frac{1}{2}$

$$\mathbb{E}\left[\mathrm{GFT}_{\varepsilon,t}\left(1/2\right)\right] \geqslant \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big]\ . \tag{2.21}$$

Similarly, for all $\varepsilon < 0$, $t \in \mathbb{N}$, and $p > \frac{1}{2}$,

$$\mathbb{E}\left[\mathrm{GFT}_{\varepsilon,t}\left(1/2\right)\right] \geqslant \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big]\ . \tag{2.22}$$

Furthermore, a direct verification shows that, for each $\varepsilon \in [-1, 1]$ and $t \in \mathbb{N}$,

$$\max_{p\in[0,1]} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] - \max_{p\in[\frac{1}{7},\frac{2}{7}]} \mathbb{E}\big[\mathrm{GFT}_{\varepsilon,t}(p)\big] \geqslant \frac{1}{100} = \Omega(1)\ . \tag{2.23}$$

Now, assume that $T \geqslant M^3/14^4$ so that, defining $\varepsilon \coloneqq (MT)^{-1/4}$, we have that the maximizer of the expected gain from trade $\frac{1}{2} + \frac{\varepsilon}{196}$ belongs to the spike region J_M. In the $+\varepsilon$ (resp., $-\varepsilon$) case, the optimal price belongs to the region $\big(\frac{1}{2}, \frac{1}{2} + \frac{1}{14M}\big]$ (resp., $\big[\frac{1}{2} - \frac{1}{14M}, \frac{1}{2}\big)$). By posting prices in the wrong region $\big[0, \frac{1}{2}\big]$ (resp., $\big[\frac{1}{2}, 1\big]$) in the $+\varepsilon$ (resp., $-\varepsilon$) case, the learner incurs a $\Omega(M\varepsilon^2) = \Omega\big(\sqrt{M/T}\big)$ instantaneous regret by (2.15) and (2.21) (resp., (2.15) and (2.22)). Then, in order to attempt suffering less than $\Omega\big(\sqrt{M/T} \cdot T\big) = \Omega\big(\sqrt{MT}\big)$ regret, the algorithm would have to detect the sign of $\pm\varepsilon$ and play accordingly. We show now that even this strategy will not improve the regret of the algorithm (by more than a constant) because of the cost of determining the sign of $\pm\varepsilon$ with the

available feedback. Since the feedback received from the two traders at time t by posting a price p is $\mathbb{I}\{p \leqslant S_{\pm\varepsilon,t}\}$ and $\mathbb{I}\{p \leqslant B_{\pm\varepsilon,t}\}$, the only way to obtain information about (the sign of) $\pm\varepsilon$ is to post in the costly ($\Omega(1)$-instantaneous regret by Equation (2.23)) sub-optimal region $\left[\frac{1}{7}, \frac{2}{7}\right]$ (see Figure 2.5). However, posting prices in the region $\left[\frac{1}{7}, \frac{2}{7}\right]$ at time t can't give more information about $\pm\varepsilon$ than the information carried by $S_{\pm\varepsilon,t}$ and $B_{\pm\varepsilon,t}$, which, in turn, can't give more information about $\pm\varepsilon$ than the information carried by the two Bernoullis $D_{\frac{1\pm\varepsilon}{2},2t-1}$ and $D_{\frac{1\pm\varepsilon}{2},2t}$. Since information-theoretic arguments imply that in order to distinguish the sign of $\pm\varepsilon$ having access to i.i.d. Bernoulli random variables of parameter $\frac{1\pm\varepsilon}{2}$ requires $\Omega(1/\varepsilon^2) = \Omega\big(\sqrt{MT}\big)$ samples, we are forced to post at least $\Omega\big(\sqrt{MT}\big)$ prices in the costly region $\left[\frac{1}{7}, \frac{2}{7}\right]$ suffering a regret of $\Omega\big(\sqrt{MT}\big)\cdot\Omega(1) = \Omega\big(\sqrt{MT}\big)$. Putting everything together, each strategy pays at least $\Omega\big(\sqrt{MT}\big)$ regret. □

What if we remove the (bd) assumption?

We conclude this section by showing that assuming that sellers' and buyers' evaluations are independent of each other, together with the (iv) and (iid) assumptions, is not enough to achieve sublinear regret rates.

The lower-bound construction is another needle in a haystack pathology and closely resembles the one in Theorem 8. In fact, with the same notation as in the proof sketch of Theorem 8, it is enough to modify the distribution of the random variables S^x and B^x such that they have the following common distribution:

$$\nu_x \coloneqq \frac{1}{3}\delta_0 + \frac{1}{3}\delta_x + \frac{1}{3}\delta_1 \,,$$

where, for any $a \in \mathbb{R}$, we recall that δ_a is the Dirac measure centered in a. This construction leads to a minimax regret $R_T^{\mathcal{S}} \geqslant \frac{T}{9}$. A formalization of these ideas can be carried out following the same proof scheme of Appendix A.9.

2.6 Weakly Budget Balanced Results

In this section, we propose a way to break linear lower bounds in the realistic feedback case.

We start by recalling that in Section 2.4 we showed that Algorithm 3 achieves learnability assuming (iv), (bd), and (iid). There, we saw also that there were two major obstructions preventing us from achieving learnability if we wanted to remove even just one of these assumptions. One obstruction was the lack observability (Theorems 7 and 9), while the other one was the needle in a haystack phenomenon (Theorem 8).

On the other hand, since the Lipschitzness of the expected gain from trade opens us the door to the use discretization methods, Lemma 1 guarantees that we can get rid of the needle in a haystack pathology just assuming (bd). Moreover, if we get a closer look at Lemma 2, we see that the feedback available in the realistic feedback case is *almost* enough to achieve the observability of the expected gain from trade. Specifically, if U is a random variable uniform on $[p, 1]$ and V is a random variable uniform on $[0, p]$, both independent on the $[0, 1]^2$-valued random pair (S, B), we have that Equation (2.6) tells us that $\mathbb{E}\big[\mathrm{gft}\,(p, (S, B))\big] = \mathbb{E}[(1-p)\mathbb{I}\{S \leqslant U \leqslant p \leqslant B\}] + \mathbb{E}[p\mathbb{I}\{S \leqslant p \leqslant V \leqslant B\}]$. This suggests that the missing piece to achieve observability under realistic feedback is allowing the learner to ask two different prices, one to the seller and the other to the buyer, *in the same interaction*. While this would violate the budget-balanced condition, we see that we need just to

propose two different prices $p_1 \leqslant p_2$, where p_1 is proposed to the seller while p_2 is proposed to the buyer. In other words, to achieve observability, we do not need to subsidize but we might need to extract money from the trade. Mechanisms enjoying this weaker form of the budget balance condition are called *weakly* budget-balanced mechanisms. It is then natural to ask whether allowing the learner to use weakly budget-balanced posted price mechanisms yields learnability in the realistic feedback case just under the (bd) assumption.

We now proceed to make this problem formal. In this new setting, rather than posting a single price, the learner can post two (possibly distinct) prices $0 \leqslant p \leqslant q \leqslant 1$, p to the seller, and q to the buyer. Naturally, this changes the benchmark: if the learner posts a pair $(p, q) \in [0,1]^2$ and the valuations of the seller and the buyer are $(s, b) \in [0,1]^2$, the net gain of the seller is $p - s$ while that of the buyer is $b - q$. Thus, if we define the upper triangle $\mathcal{U} \coloneqq \{(p, q) \in [0,1]^2 \mid p \leqslant q\}$, we can *overload* the gain from trade function by defining

$$\begin{aligned} \mathrm{gft} \colon \mathcal{U} \times [0,1]^2 &\to [0,1] \,, \\ (p, q, s, b) &\mapsto (b - q + p - s)\mathbb{I}\{s \leqslant p \leqslant q \leqslant b\}, \end{aligned}$$

we can *overload* the gain from trade at any time t by defining

$$\mathrm{GFT}_t \colon \mathcal{U} \to [0,1], \qquad (p, q) \mapsto \mathrm{gft}\big((p, q), (S_t, B_t)\big) \,,$$

and finally, with these definitions, the gain from trade of the market at time t if the learner posts $(P_t, Q_t) \in \mathcal{U}$ becomes

$$\Big(\underbrace{B_t - Q_t}_{\text{buyer's net gain}} + \underbrace{P_t - S_t}_{\text{seller's net gain}} \Big) \cdot \underbrace{\mathbb{I}\{S_t \leqslant P_t \leqslant Q_t \leqslant B_t\}}_{\text{whenever a trade happens}} = \mathrm{GFT}_t(P_t, Q_t) \,.^{\dagger\dagger}$$

Now, the following observation is crucial.

Remark 1. *The only reason for a budget-balanced strategy to post two different prices is to obtain more information. A direct verification shows that the expected gain from trade can always be maximized by posting the same price to both the seller and the buyer.*

This last remark has two crucial consequences.

First, there is no point in posting two different prices when full feedback is available. Hence, all the full feedback results we have achieved in Section 2.3 apply verbatim to weakly budget-balanced mechanisms, and the only interesting case to study is the realistic feedback one.

Second, the notion of regret stays nearly unchanged. Precisely, noting that for $(s, b) \in [0,1]^2$ and any $p \in [0,1]$ it holds that $\mathrm{gft}\big((p, p), (s, b)\big) = \mathrm{gft}\big(p, (s, b)\big)$, the *regret* at time horizon T of a learner following a strategy α to generate the sequence of prices $(P_t, Q_t) \in \mathcal{U}$ against an environment β generating the sequence of (random) pairs (S_t, B_t) becomes

$$R_T(\alpha, \beta) \coloneqq \max_{(p,q) \in \mathcal{U}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p, q) - \sum_{t=1}^{T} \mathrm{GFT}_t(P_t, Q_t) \right]$$

[††] Other works consider the following alternative definition for the gain from trade: $\Big(\underbrace{B_t - Q_t}_{\text{buyer's net gain}} + \underbrace{Q_t - P_t}_{\text{broker's net gain}} + \underbrace{P_t - S_t}_{\text{seller's net gain}} \Big) \cdot \underbrace{\mathbb{I}\{S_t \leqslant P_t \leqslant Q_t \leqslant B_t\}}_{\text{whenever a trade happens}} = (B_t - S_t)\mathbb{I}\{S_t \leqslant P_t \leqslant Q_t \leqslant B_t\}$. Our results translate with minimal effort to this definition.

Estimation procedure of GFT using two prices and one-bit feedback

Input: price p
Environment: fixed pair of seller and buyer valuations (s, b)
Toss a biased coin with probability p of Heads
if Heads **then** draw V uniformly at random in $[0, p]$ and set $\widehat{p} \coloneqq V$, $\widehat{q} \coloneqq p$
else draw U uniformly at random in $[p, 1]$ and set $\widehat{p} \coloneqq p$, $\widehat{q} \coloneqq U$
Post price $\widehat{p}$ to the seller and $\widehat{q}$ to the buyer and observe the one-bit feedback $\mathbb{I}\{s \leqslant \widehat{p} \leqslant \widehat{q} \leqslant b\}$
Return $\widehat{\mathrm{GFT}}(p) \coloneqq \mathbb{I}\{s \leqslant \widehat{p} \leqslant \widehat{q} \leqslant b\}$ ▷ Unbiased estimator of $\mathrm{GFT}(p)$

$$= \max_{p\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p) - \sum_{t=1}^{T} \mathrm{GFT}_t(P_t, Q_t)\right] .$$

where, again, the expectation is taken with respect to any randomness present in the environment and (possibly) the internal randomization used by the learner's strategy.

2.6.1 Realistic Feedback - $T^{3/4}$ Upper Bound via Blind-Exp3

In this section, we introduce the algorithm Blind-Exp3, which achieves a $\widetilde{O}(T^{3/4})$ regret rate whenever it works in a (bd) environment.

We first formalize, as an easy corollary of Lemma 1, that the (bd) assumption allows us to discretization methods. In fact, for any fixed grid of prices G in $[0, 1]$, it is possible to relate the expected gain from trade of the best price in G with that of the best fixed price in $[0, 1]$, paying a discretization error that depends on the upper bound M on the densities of the elements in the sequence $(S_t, B_t)_{t\in\mathbb{N}}$. For notational convenience, for any finite grid G, we define the parameter $\delta(G)$ as follows:

$$\delta(G) = \max_{p\in[0,1]} \min_{g\in G} |p - g|.$$

Claim 1 (Discretization error)**.** *Let G be any finite grid of prices in $[0, 1]$. Then, for any sequence of $[0, 1]^2$-valued random variables $(S_1, B_1), \ldots, (S_T, B_T)$, each of them admitting a density (with respect to the Lebesgue measure on $[0, 1]^2$) bounded by some $M > 0$, we have the following:*

$$\max_{p\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p)\right] - \max_{g\in G} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(g)\right] \leqslant M\delta(G)T .$$

Proof. Let $p^\star$ be the best fixed price in hindsight in $[0, 1]$ with respect to the sequence $(S_1, B_1), \ldots, (S_T, B_T)$. We have two cases. If $p^\star \in Q$, then there is nothing to prove. If this is not the case, then there exists $p_G \in G$, such that $|p^\star - p_G| \leqslant \delta(G)$. We have the following:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p^\star)\right] - \max_{p\in G} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p)\right] &\leqslant \sum_{t=1}^{T} \big(\mathbb{E}\left[\mathrm{GFT}_t(p^\star)\right] - \mathbb{E}\left[\mathrm{GFT}_t(p_G)\right]\big) \\ &\leqslant M|p^\star - p_Q|T \leqslant M\delta(G)T,\end{aligned}$$

where, in the second to last inequality, we used the Lipschitz property of the expected gain from trade as in Lemma 1. □

A key technique that we use is a Monte Carlo estimation procedure $\widehat{\mathrm{GFT}}$ (see pseudocode for

details) that allows us to estimate the expected gain from trade $\mathbb{E}\big[\mathrm{GFT}\big(p,(S_t,B_t)\big)\big]$ of a price p, by posting *two* different prices $(\widehat{p},\widehat{q})$ and receiving *one* bit of feedback.

Lemma 6 (Lemma 1 of Azar et al. [20])**.** *Fix any agents' valuations $(s,b)\in[0,1]^2$. For any price $p\in[0,1]$, it holds that $\widehat{\mathrm{GFT}}(p)$ is an unbiased estimator of $\mathrm{gft}\big(p,(s,b)\big)$, i.e., $\mathbb{E}\left[\widehat{\mathrm{GFT}}(p)\right]=\mathrm{gft}\big(p,(s,b)\big)$, where the expectation is with respect to the randomness of the estimation procedure.*

The proof of Lemma 6 follows immediately from Equation (2.6).

Once we have this procedure, we can present our algorithm. At a high level, the algorithm mimics the behavior of Exp3 on a fixed discretization of K prices, but the estimation procedure is used to perform the uniform exploration step. Our algorithm is "blind" because—unlike what happens in the bandit case—posting a price does not reveal the corresponding gain from trade. With a careful analysis, we show that the uniform exploration term is indeed enough to achieve the tight regret bound of order $\widetilde{O}(T^{3/4})$ whenever the (bd) assumption holds.

Algorithm 6 Blind-Exp3 - Realistic Feedback

input: Learning rate $\eta>0$, exploration rate $\gamma\in(0,1)$, grid of prices G, with $|G|=K$
initialization: Set $w_1(i)$ to 1 for all $i\in[K]$ and $W_1\coloneqq K$
for time $t=1,2,\dots$ **do**
  Let $\pi_t(i)\coloneqq\frac{w_t(i)}{W_t}$ for all $i\in[K]$
  Toss a biased coin with probability γ of Heads
  if Tails **then** ▷ Exploitation step
    Post price P_t drawn according to distribution π_t and set $\widehat{r}_t(i)\coloneqq 0$ for all $i\in[K]$
  else ▷ Exploration step
    Draw a price g_{I_t} uniformly at random in G
    Use the estimation procedure on price g_{I_t} and receive $\widehat{\mathrm{GFT}}_t(g_{I_t})$
    Set $\widehat{r}_t(I_t)\coloneqq\frac{K}{\gamma}\cdot\widehat{\mathrm{GFT}}_t(g_{I_t})$ and $\widehat{r}_t(j)\coloneqq 0$ for all $j\neq I_t$.
  Let $w_{t+1}(i)\coloneqq w_t(i)\cdot\exp\big(\eta\widehat{r}_t(i)\big)$ for all $i\in[K]$ ▷ Exponential weight update
  Let $W_{t+1}\coloneqq\sum_{p_i\in G}w_{t+1}(i)$

Theorem 16. *Consider the problem of repeated bilateral trade in the weakly budget-balanced realistic-feedback model.[‡‡] Let $M>0$. Suppose that $\mathcal{S}$ is the set of environments such that, for each $t\in\mathbb{N}$, the pair (S_t,B_t) admits a density bounded by M (with respect to the Lebesgue measure on $[0,1]^2$). If we run* Blind-Exp3 *with exploration rate $\gamma\in(0,1)$, learning rate $\eta>0$, and the uniform K-grid G such that $\frac{2\eta K}{\gamma}\leqslant 1$ then, for each time horizon $T\in\mathbb{N}$, we have that*

$$R_T^{\mathcal{S}}(\text{Blind-Exp3})\leqslant\frac{\ln K}{\eta}+\left(\gamma+\eta\frac{K}{\gamma}+\frac{M}{K}\right)T.$$

In particular, if $T\geqslant 16$, tuning the number of grid points $K=\lfloor T^{1/4}\rfloor$, the exploration rate $\gamma=\frac{(\ln T)^{1/3}}{T^{1/4}}$, and the learning rate $\eta=\frac{1}{2}\frac{(\ln T)^{2/3}}{T^{3/4}}$, then $R_T^{\mathcal{S}}(\text{Blind-Exp3})\leqslant 2\left(M+(\ln T)^{1/3}\right)\cdot T^{3/4}$.

Proof. The analysis of Blind-Exp3 needs to carefully take into account many sources of randomness: the internal randomness of the algorithm, of the estimation procedures, and the randomness governing

[‡‡]Interestingly —and in contrast to what happens in the budget balanced case (Section 2.4.6)— Blind-Exp3 would work even having access to just one-bit of feedback, i.e., observing just $\mathbb{I}\{S_t\leqslant P_t\leqslant Q_t\leqslant B_t\}$ after each interaction. The same guarantees stated in this theorem holds also for the one-bit feedback case.

the sequence of sellers' and buyers' evaluations. Fix any exploration rate $\gamma \in (0,1)$, learning rate $\eta > 0$ and number of grid points $K \in \mathbb{N}$ such that ${}^{2\eta K}/_{\gamma} \leqslant 1$. Fix also any time horizon $T \in \mathbb{N}$. In the following, we use the random variables (P_t, Q_t) to denote the randomized prices posted by the algorithm at time t.

Fix any history of the algorithm (i.e. realization of all the randomness involved). We have the following:

$$\begin{aligned}
\ln\left(\frac{W_{T+1}}{W_1}\right) = \ln\left(\prod_{t=1}^{T}\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T}\ln\left(\frac{W_{t+1}}{W_t}\right) &= \sum_{t=1}^{T}\ln\left(\sum_{i\in[K]}\pi_t(i)\exp\left(\eta\widehat{r}_t(i)\right)\right) \\
&\leqslant \sum_{t=1}^{T}\ln\left(1+\eta\sum_{i\in[K]}\pi_t(i)\widehat{r}_t(i)+\eta^2\sum_{i\in[K]}\pi_t(i)\big(\widehat{r}_t(i)\big)^2\right) \\
&\leqslant \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\widehat{r}_t(i)+\eta^2\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\big(\widehat{r}_t(i)\big)^2 \qquad \text{(using } \widehat{r}_t(i)\leqslant\tfrac{K}{\gamma}\text{)} \\
&\leqslant \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\widehat{r}_t(i)\left(1+\eta\frac{K}{\gamma}\right). \qquad (2.24)
\end{aligned}$$

Crucially, we can use the standard exponential and logarithmic inequalities $\exp(x) \leqslant 1 + x + x^2$ (valid whenever $x \leqslant 1$), and $\ln(1+x) \leqslant x$ (valid whenever $x > -1$) only because the particular choice of the parameters (${}^{2\eta K}/_{\gamma} \leqslant 1$) implies that $\eta\widehat{r}_t(i) \leqslant 1$ and

$$\eta\sum_{i\in[K]}\pi_t(i)\widehat{r}_t(i)+\eta^2\sum_{i\in[K]}\pi_t(i)\big(\widehat{r}_t(i)\big)^2 \leqslant 2\eta\sum_{i\in[K]}\pi_t(i)\widehat{r}_t(i) \leqslant \frac{K}{\gamma}.$$

Inequality 2.24 is the pivot of our analysis, as we construct upper and lower bounds to its two extremes. We start from its first term, take the expectation with respect to the whole randomness of the process and consider any price g_i in the grid G:

$$\begin{aligned}
\mathbb{E}\left[\ln\left(\frac{W_{T+1}}{W_1}\right)\right] = \mathbb{E}\left[\ln\left(W_{T+1}\right)\right] - \ln K &\geqslant \mathbb{E}\left[\ln\left(w_{T+1}(i)\right)\right] - \ln K \\
&= \eta\sum_{t=1}^{T}\mathbb{E}\left[\widehat{r}_t(i)\right] - \ln K = \eta\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(g_i)\right] - \ln K. \qquad (2.25)
\end{aligned}$$

The only delicate passage of the previous formula is the last equality, where we used that $\mathbb{E}\left[\widehat{r}_t(i)\right] = \mathbb{E}\left[\mathrm{GFT}_t(g_i)\right]$. To see why the latter holds, consider the filtration $\{\mathcal{F}_t\}_t$ relative to the story of the process: $\mathcal{F}_t$ is the σ-algebra generated by all the random variables involved in the process up to time t (excluded). Moreover, let $\mathcal{E}_t^i$ be the event that at round t the coin toss results in Heads and the price selected u.a.r. for exploration is g_i. We have the following:

$$\begin{aligned}
\mathbb{E}\left[\widehat{r}_t(i) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_t^i}\widehat{r}_t(i) \mid \mathcal{F}_t\right] && \widehat{r}_t(i) = \mathbb{I}_{\mathcal{E}_t^i}\widehat{r}_t(i) \\
&= \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_t^i}\mathbb{E}\left[\widehat{r}_t(i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Law of total exp.} \\
&= \frac{K}{\gamma}\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_t^i}\mathbb{E}\left[\widehat{\mathrm{GFT}}_t(g_i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Def. of } \widehat{r}_t(i)
\end{aligned}$$

$$= \frac{K}{\gamma}\mathbb{P}[\mathcal{E}_t^i \mid \mathcal{F}_t]\mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right] \qquad \text{Lemma 6 and } (S_t, B_t) \text{ indep. of } \mathcal{E}_t^i$$
$$= \mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right]$$

For the final step, note that, conditioned on $\mathcal{F}_t$, the event $\mathcal{E}_t^i$ has probability γ/K: the random coin gives Tails with probability γ and price g_i is chosen (independently) u.a.r. as the one to be actually explored with probability $1/K$. Taking the expectation with respect to $\mathcal{F}_t$ gives that $\mathbb{E}\left[\widehat{r}_t(i)\right] = \mathbb{E}\left[\mathrm{GFT}_t(g_i)\right]$.

Let's go back to Equation (2.24) and focus on the last term. Conditioning with respect to $\mathcal{F}_t$:

$$\mathbb{E}\left[\pi_t(i)\widehat{r}_t(i) \mid \mathcal{F}_t\right] = \pi_t(i)\mathbb{E}\left[\widehat{r}_t(i) \mid \mathcal{F}_t\right] = \pi_t(i)\mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right].$$

Taking the expectation with respect to $\mathcal{F}_t$ and summing over all the $g_i \in G$, we have the following:

$$\mathbb{E}\left[\mathrm{GFT}_t(P_t, Q_t)\right] \geqslant (1-\gamma)\sum_{i\in[K]}\mathbb{E}\left[\pi_t(i)\mathrm{GFT}_t(g_i)\right] = (1-\gamma)\sum_{i\in[K]}\mathbb{E}\left[\pi_t(i)\widehat{r}_t(i)\right], \tag{2.26}$$

where the first inequality follows from the fact that with probability $1-\gamma$ the learner at time t chooses exploitation and thus posts a price in the grid G according to distribution π_t. We can plug Equation (2.25) and Equation (2.26) into Equation (2.24) to obtain the following:

$$\eta\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(g_i)\right] - \ln K \leqslant \frac{\eta}{1-\gamma}\left(1+\eta\frac{K}{\gamma}\right)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(P_t, Q_t)\right]$$

Multiplying everything by $(1-\gamma)/\eta$, rearranging, and using that the gain from trade is always upper bounded by 1, we get:

$$\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(g_i)\right] - \sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(P_t, Q_t)\right] \leqslant \frac{\ln K}{\eta} + \left(\gamma + \eta\frac{K}{\gamma}\right)T$$

The argument so far holds for any environment β and any choice of price on the grid g_i. This, together with the discretization result Claim 1 gives the desired bound:

$$R_T(\text{Blind-Exp3}, \beta) \leqslant \frac{\ln K}{\eta} + \left(\gamma + \eta\frac{K}{\gamma} + \frac{M}{K}\right)T$$

when the environment β is such that, for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a density bounded by M. □

2.6.2 Realistic Feedback - $T^{3/4}$ Lower Bound via Multi-Apple Tasting

In this section, we prove that Blind-Exp3 is an almost optimal algorithm for the realistic feedback case under the (bd) assumption. This result is proven by the means of an exotic lower bound of order $T^{3/4}$, which has two notable implications. First, it provides a formalization to the intuition that the *realistic* feedback is strictly less informative than the *bandit* feedback (see the discussion in Section 2.1.2), being the regret of the latter of order at most $T^{2/3}$.[§§] Second, noting that the hard

[§§]Although our decision space is two-dimensional, one can see that, with bandit feedback in a (bd) environment, a regret of order $T^{2/3}$ can be obtained by running an optimal bandit algorithm (e.g., MOSS Audibert and Bubeck

instances constructed in the proof of Theorem 17 are i.i.d., we see that adding the (iid) assumption to the (bd) assumption does not help to improve regret rates.¶¶

Theorem 17. *Consider the problem of repeated bilateral trade in the weakly budget-balanced realistic-feedback model. There exists a numerical constants $c > 50^{-3}$ such that, for any time horizon $T \geqslant 8008$, the minimax regret satisfies*

$$R_T^{\mathcal{S}} \geqslant cT^{3/4} ,$$

where $\mathcal{S}$ is the set of all environments such that

(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a density bounded above by $M \geqslant 9$.*

(iid) *$(S_1, B_1), (S_2, B_2), \dots$ is an i.i.d. sequence.*

The rest of the section is devoted to sketching the proof of the theorem (for a full proof, see Appendix A.16). The sketch is divided into three steps: first, we construct a hard instance of the repeated bilateral trade problem; then, we present a related problem on a discrete set of actions that preserves the relevant features of the original problem while allowing for an easier analysis of the regret; finally, we show how the minimax regret of the second problem leads to a $T^{3/4}$ regret for bilateral trade.

The construction of a hard family of adversaries

Here, we construct the family of distributions with bounded densities for the seller/buyer evaluation pair that we use to prove the lower bound. We consider an i.i.d. environment: i.e., the valuations (S_t, B_t) are drawn i.i.d. according to a fixed distribution. We build this family of distributions by suitable perturbations over a base distribution, whose support is given by the union of the six squares $Q_1, \dots, Q_6$ (see Figure 2.6, left). The squares are obtained by translating $[0, 1/6]^2$, respectively, by $(0, \frac{1}{3}), (0, \frac{1}{2}), (0, \frac{5}{6}), (\frac{5}{6}, \frac{5}{6}), (\frac{5}{6}, 0), (\frac{1}{2}, \frac{2}{3})$. Letting $a \coloneqq 2\ln(27/16)$, the probability density function f of the base distribution is

$$f(x,y) \coloneqq \frac{36}{1+8a} \cdot \left(\frac{5-6(y+x)}{6(y-x)} \mathbb{I}_{Q_1}(x,y) + a\mathbb{I}_{Q_2}(x,y) + 2a\mathbb{I}_{Q_3 \cup Q_4 \cup Q_5}(x,y) + \mathbb{I}_{Q_6}(x,y) \right) .$$

The perturbations to this base distribution are parametrized by two terms: a translation $v \in \left(\frac{1}{3}, \frac{1}{2}\right)$ and a scale $\varepsilon \in \left(0, \frac{1}{12}\right)$ such that $\frac{1}{3} + \varepsilon \leqslant v \leqslant \frac{1}{2} - \varepsilon$. We denote the set of these parameters by Ξ. Each perturbed distribution has density $f_{v,\varepsilon} := f + g_{v,\varepsilon}$, where $g_{v,\varepsilon}$ is defined as follows:

$$g_{v,\varepsilon}(x,y) \coloneqq \frac{36}{1+8a} \cdot \left(\mathbb{I}_{R^1_{v,\varepsilon} \cup R^4_{v,\varepsilon}}(x,y) - \mathbb{I}_{R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon}}(x,y) \right) ,$$

and the rectangles $R^i_{v,\varepsilon}$ (see Figure 2.6, left/center) have the following analytic expression: $R^1_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\frac{3}{4}, \frac{5}{6}\right]$, $R^2_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\frac{2}{3}, \frac{3}{4}\right)$, $R^3_{v,\varepsilon} = [v, v+\varepsilon] \times \left[\frac{3}{4}, \frac{5}{6}\right]$, $R^4_{v,\varepsilon} = [v, v+\varepsilon] \times \left[\frac{2}{3}, \frac{3}{4}\right)$. Note that the rectangles $R^i_{v,\varepsilon}$ are included in Q_6 for all $i \in [4]$ and $(v, \varepsilon) \in \Xi$.

15, whose upper bound on the regret is of order $\sqrt{KT}$) on a discretization of $K = \Theta(T^{1/3})$ equispaced prices on the diagonal $\{(p,q) \in \mathcal{U} \mid p = q\}$, thanks to Claim 1. Similar results appeared, e.g., in Auer et al. [19], Kleinberg [116].

¶¶Actually, the (iid) assumption does not seem to play any role for weakly budget balanced posted prices when it comes to time horizon regret guarantees: the same needle in a haystack pathology in Theorem 7 implies that the (iid) assumption alone is not enough to achieve sublinear regret even if weakly budget balanced posted prices are allowed.

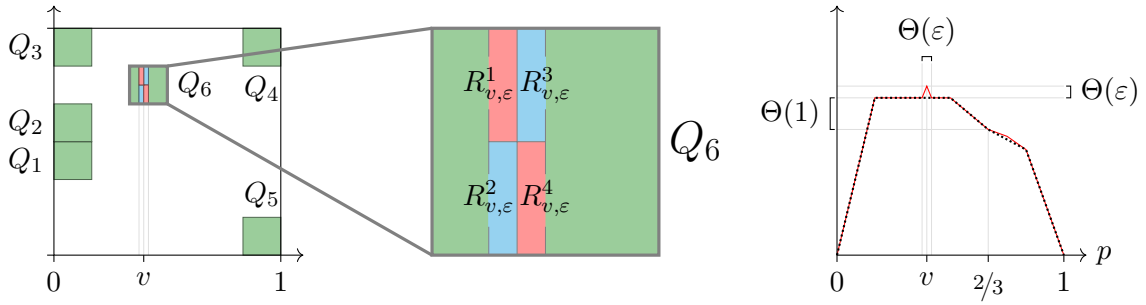


Figure 2.6: Left/center: The six squares $Q_1, \dots, Q_6$ (in green) are the support of the base density f, and the four rectangles $R^1_{v,\varepsilon}, \dots, R^4_{v,\varepsilon}$ (in red and blue) inside Q_6 are the regions where the density is perturbed with $g_{v,\varepsilon}$. Right: The corresponding qualitative plots of $p \mapsto \mathbb{E}[\mathrm{GFT}(p, S, B)]$ (black, dotted) and $p \mapsto \mathbb{E}^{v,\varepsilon}[\mathrm{GFT}(p, S, B)]$ (red, solid).

Let $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$, for all $(v, \varepsilon) \in \Xi$) be a probability measure such that the sequence of seller/buyer evaluations (S, B), (S_1, B_1), (S_2, B_2), $\dots$ is i.i.d. and the distribution of (S, B) has probability density function f (resp., $f_{v,\varepsilon}$). We denote the expectation with respect to $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$) by $\mathbb{E}$ (resp., $\mathbb{E}^{v,\varepsilon}$). Note that, for each $(v, \varepsilon) \in \Xi$, the density $f_{v,\varepsilon}$ is upper bounded by $M = 9$. Given the explicit form for the base distribution, we can compute the corresponding expected value of the gain from trade $\mathbb{E}\big[\mathrm{gft}\big(p, (S, B)\big)\big]$ obtained by posting price $p \in [0, 1]$ to both agents, when (S, B) is drawn from the base distribution. The analytic expression of $\mathbb{E}\big[\mathrm{gft}\big(\cdot, (S, B)\big)\big]$ can be found in Appendix A.16 (Equation (A.8)), and a plot is reported in Figure 2.6 (right, dotted black). What is relevant to our argument is that the function $p \mapsto \mathbb{E}\big[\mathrm{gft}\big(p, (S, B)\big)\big]$ is continuous, maximized at every point of the plateau region $\left[\frac{1}{6}, \frac{1}{2}\right]$, and its value at $\frac{2}{3}$ is bounded away from the maximum. We can explicitly compute the expected gain from trade $\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p, (S, B)\big)\big]$ obtainable by posting any price $p \in [0, 1]$ to both agents, when (S, B) is drawn from the distribution with perturbation parameters v and ε. We have the following:

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p, (S, B)\big)\big] = \mathbb{E}\big[\mathrm{gft}\big(p, (S, B)\big)\big] + \frac{1}{864(1 + 8a)}\big(\varepsilon \cdot \Lambda_{v,\varepsilon}(p) + 12\varepsilon^2 \cdot \Lambda_{\frac{3}{4},\frac{1}{12}}(p)\big)$$

where $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as $\Lambda_{u,r}(x) \coloneqq (1 - |x - u|/r)^+$. Thus, for each $(v, \varepsilon) \in \Xi$, the plot of $\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(\cdot, (S, B)\big)\big]$ coincides with that of $\mathbb{E}\big[\mathrm{gft}\big(\cdot, (S, B)\big)\big]$ up to two small deviations (around v and $3/4$), and it is maximized at v (see Figure 2.6, right).

We now focus our attention on the feedback received by a learner that posts prices (p, q), when the underlying distribution corresponds to perturbations parameters $(v, \varepsilon) \in \Xi$.

Claim 2. *Fix any $(v, \varepsilon) \in \Xi$, $(p, q) \in \mathcal{U} \setminus \bigcup_{i \in [4]} R^i_{v,\varepsilon}$, and let $Z \coloneqq \big(\mathbb{I}\{S \leqslant p\}, \mathbb{I}\{q \leqslant B\}\big)$. Then Z follows the same distribution under both $\mathbb{P}$ and $\mathbb{P}^{v,\varepsilon}$.*

Proof. Here we consider only the event $\big\{Z = (0, 0)\big\}$; for a full proof, see Claim 6 in Appendix A.16.

$$\mathbb{P}^{v,\varepsilon}\big[Z = (0, 0)\big] = \mathbb{P}^0\big[Z = (0, 0)\big] + \int_{(p,1] \times [0,q)} g_{v,\varepsilon}(x, y)\, \mathrm{d}x\, \mathrm{d}y\ .$$

If (p, q) is not in $R^{v,\varepsilon}$, by symmetry, the integral term is 0. □

Claim 2 implies that if the learner wants to locate $v \in \left[\frac{1}{3} + \varepsilon, \frac{1}{2} - \varepsilon\right]$ observing samples of the two-bit feedback Z drawn according to the distribution $\mathbb{P}^{v,\varepsilon}$, they have to post prices in the region

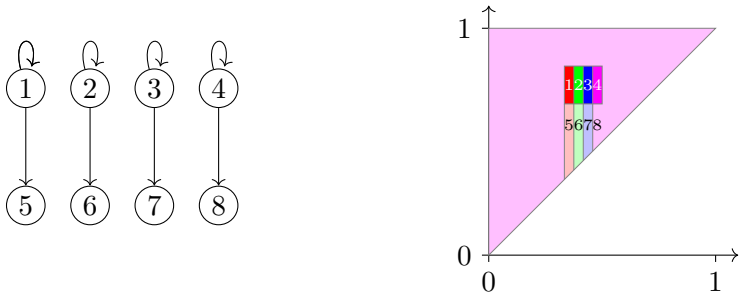


Figure 2.7: Left: The feedback graph of multi-apple tasting for $K = 4$. Right: The map ι.

Q_6. However, in doing so, they suffer constant instantaneous regret. Indeed, a direct verification shows that for any $(v, \varepsilon) \in \Xi$ and all $(p, q) \in Q_6$,

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(v, (S, B)\big)\big] - \mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big((p, q), (S, B)\big)\big] \geqslant \mathbb{E}\left[\mathrm{gft}\big(\tfrac{1}{2}, (S, B)\big) - \mathrm{gft}\big(\tfrac{2}{3}, (S, B)\big)\right] = \Theta(1) \,.$$

So far, we built a family of i.i.d. adversaries for our bilateral trade problem such that the optimal pair of prices belongs to $\mathscr{D}_{\mathrm{opt}} \coloneqq \big\{(p, q) \in \mathcal{U} \mid p = q \in [1/3, 1/2]\big\}$, but, when the underlying distribution is determined by one of the probability measures $\mathbb{P}^{v,\varepsilon}$, in order not to suffer regret $\Omega(\varepsilon T)$, the learner has to detect an ε-spike inside $\mathscr{D}_{\mathrm{opt}}$. As observed in Claim 2, this can only be accomplished by posting prices in Q_6, which, as shown above, has an instantaneous regret of order $\Omega(1)$. The missing piece is now to quantify how long the learner can be forced to spend time posting prices in Q_6. To this end, we build a reduction from a simplified *online learning with feedback graph* problem on $2K$ arms that highlights the underlying structure of our problem. Our goal is to show that for any algorithm α for the repeated bilateral trade problem there exists an algorithm $\widetilde{\alpha}$ for the new problem such that the regret suffered by the latter is a lower bound on the regret suffered by the former.

The multi-apple tasting problem

In this section, we introduce an auxiliary online learning problem on a discrete set of actions that we call *multi-apple tasting*: it will be easier to analyze than our original bilateral trade problem while still capturing its difficulties. The multi-apple tasting problem has the following form: there are $2K$ actions, the first K are called the *exploration* arms, while the others are the *exploitation* arms. Playing an exploitation arm yields no feedback, while an exploration arm i gives information about the performance of the corresponding exploitation arm $i + K$. The reader familiar with the notion of online learning with directed feedback graphs [8] will recognize that the feedback model described here corresponds to the simple (weakly observable) feedback graph in Figure 2.7 (left).

The rewards. We now describe the random rewards of $K + 1$ instances of the multi-apple tasting problem associated to $K + 1$ probability measures $\mathbb{P}^0, \dots, \mathbb{P}^K$. Set c_{prob} to be $7/(2a)$ and consider the i.i.d. sequence of random vectors $Y, Y_1, Y_2, \dots, Y_T$ such that $Y \in \{0, 1\}^{2K}$ and, for each $k \in \{0, \dots, K\}$ and $i \in [K]$, it holds that $Y(i + K) \coloneqq Y_1(i + K) \coloneqq \cdots \coloneqq Y_T(i + K) = 0$ and

$$\mathbb{P}^k\big[Y(i) = 1\big] \coloneqq \begin{cases} \frac{1}{2} & \text{if } i \in [K] \backslash \{k\} \\ \frac{1}{2} + c_{\mathrm{prob}} \cdot \varepsilon & \text{if } i = k \end{cases}$$

The random vectors $Y_1, Y_2, \dots, Y_T$ control the rewards the learner gets in this new problem. Formally, a learner playing action $i \in [2K]$ at time t gets reward $\rho_t(i) \coloneqq \rho(i, Y_t)$ where

$$\rho(i,y) \coloneqq \begin{cases} 0 & \text{if } j \in [K] \\ c_{\text{plat}} + \frac{c_{\text{spike}}}{c_{\text{prob}}} \cdot \big(y(j-K) - \frac{1}{2}\big) & \text{otherwise} \end{cases}$$

$c_{\text{plat}} \coloneqq \frac{a}{2(1+8a)}$, $c_{\text{spike}} \coloneqq \frac{1}{6(1+8a)} \cdot \frac{1}{144}$, and, for any $i \in [K]$, we denoted i-th component of y by $y(i)$. Observe that for all $k \in \{0, \dots, K\}$ and $i \in \{K+1, \dots, 2K\}$, we have

$$\mathbb{E}^k\big[\rho(i,Y)\big] = \begin{cases} c_{\text{plat}} & \text{if } k \neq i-K \\ c_{\text{plat}} + c_{\text{spike}} \cdot \varepsilon & \text{otherwise} \end{cases}$$

The feedback. The learner in multi-apple tasting receives two types of feedback. If they play action $i \geqslant K+1$ (an exploitation arm) at time t, then they receive no feedback (modeled by $Y_t(i) = 0$). If instead, they play action $i \leqslant K$ (an exploration arm), they receive feedback $Y_t(i)$. This feedback structure describes an instance of online learning with feedback graphs, where the underlying graph is the one in Figure 2.7 (left). The rewards incurred by the exploring arms are fixed and known irregardless of the action played, while the only way to learn the expected value of $\rho_t(i)$ for $i > K$ is to play the corresponding exploring action $i - K$.

The minimax regret. Leveraging a standard information-theoretic argument, it can be proved that any algorithm for the multi-apple tasting problem has to suffer a regret of order at least $\Omega\big(\min(\frac{K}{\varepsilon^2}, \varepsilon T)\big)$ on at least one of the instances induced by $\mathbb{P}^0, \dots, \mathbb{P}^K$. Intuitively, in order to prevent losing εT, the learner has to play each one of the K exploring arms at least $\Omega(1/\varepsilon^2)$ times.

Relating the two problems

We have described multi-apple tasting, and $K+1$ distributions to generate the sequence of rewards for it. We now show how to simulate any distribution of the feedback in instances $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$ of the bilateral trade problem using the random variables Y (and some extra random seeds). Let $K = \lceil T^{1/4} \rceil$ and $\varepsilon = \frac{1}{2K}$, and consider the baseline instance and the K perturbed instances of the repeated bilateral trade problem above, each corresponding to $(v_k, \frac{\varepsilon}{6})$ for $v_k = \frac{1}{3} + (2k-1)\frac{\varepsilon}{6}$ and $k \in [K]$. For each one of these instances, we construct an instance of multi-apple tasting that can be used to *simulate* it.

As a first step, we explain how to associate each pair of prices in the upper triangle (i.e., the set of actions in the bilateral trade problem) to one of the $2K$ actions in the feedback graph problem. We partition the upper triangle $\mathcal{U}$ of the unit square $[0,1]^2$ into $2K$ subsets, each corresponding to areas of "similar" behavior:

- $J_k \coloneqq [v_k - \frac{\varepsilon}{6}, v_k + \frac{\varepsilon}{6}) \times [\frac{2}{3}, \frac{5}{6}]$, $\forall k \in [K-1]$, and $J_K \coloneqq [v_K - \frac{\varepsilon}{6}, v_K + \frac{\varepsilon}{6}] \times [\frac{2}{3}, \frac{5}{6}]$.
- $J_{k+K} \coloneqq \{(p,q) \in \mathcal{U} \mid v_k - \frac{\varepsilon}{6} \leqslant p < v_k + \frac{\varepsilon}{6} \text{ and } q < \frac{2}{3}\}$, $\forall k \in [K-1]$, and $J_{2K} \coloneqq \mathcal{U} \setminus \bigcup_{k=1}^{2K-1} J_k$.

Given the partition, we can introduce the map ι which associates each $(p,q) \in \mathcal{U}$ with the unique $i \in [2K]$ such that $(p,q) \in J_i$ (see Figure 2.7, right, for a pictorial representation of ι). Then, we introduce an i.i.d. sequence $V, V_1, V_2, \dots, V_T$ of uniform random variables in $[0,1]$, independent of the sequence of Ys. Both the Y and the V sequences are independent of the sequence of valuations $(S_1, B_1), (S_2, B_2), \dots, (S_T, B_T)$.

The next claim is the core of our reduction: it can be proved by applying our novel information-theoretic result (Theorem 44, Appendix A.15). To do it, one can verify that, for all $k \in [K]$, the Radon-Nikodym derivative of the distribution of the feedback $\big(\mathbb{I}(S \leqslant p), \mathbb{I}\{q \leqslant B\}\big)$ under $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$ with respect to its distribution under $\mathbb{P}$ is bounded from above (resp., below) by the maximum (resp., minimum) of the Radon-Nikodym derivative of the distribution of $Y(\iota(p,q))$ under $\mathbb{P}^k$ with respect to its distribution under $\mathbb{P}^0$. For a proof, see Claim 7 in Appendix A.16.

Claim 3. *For any $(p,q) \in \mathcal{U}$ there exists a function $\varphi_{p,q} : \{0,1\} \times [0,1] \to \{0,1\}^2$ such that, for all $k \in [K]$, the distribution of $\varphi_{p,q}(Y(\iota(p,q)), V)$ under $\mathbb{P}^0$ (resp., $\mathbb{P}^k$, for all $k \in [K]$) is the same as that of $\big(\mathbb{I}(S \leqslant p), \mathbb{I}\{q \leqslant B\}\big)$ under $\mathbb{P}$ (resp., $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$).*

We now proceed as follows. Let α be any strategy for the original bilateral trade problem. We show how to simulate its behavior over the instances $\mathbb{P}$ and $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$, for $k \in [K]$, using a strategy $\widetilde{\alpha}$ for multi-apple tasting (together with the sequence of random seeds $V_1, V_2, \dots, V_T$) over the distributions $\mathbb{P}^0$ and $\mathbb{P}^k$, for $k \in [K]$. When the strategy α chooses prices $(p_t, q_t) \in \mathcal{U}$ at time t, then $\widetilde{\alpha}$ plays the action $\iota(p_t, q_t) \in [2K]$, receives reward $\rho_t(\iota(p_t, q_t))$ and observes the feedback $Y_t(\iota(p_t, q_t))$. The strategy α is then fed the feedback $\varphi_{p_t,q_t}\big(Y_t(\iota(p_t, q_t)), V_t\big) \in \{0,1\}^2$ which it uses to select its new action (p_{t+1}, q_{t+1}). Crucially, leveraging Claim 3 and the structure of the rewards in two problems, one can prove that the regret $R_T^0(\alpha)$ (resp., $R_T^k(\alpha)$, for any $k \in [K]$) that algorithm α suffers under probability $\mathbb{P}$ (resp., $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$) in the repeated bilateral trade problem is at least the regret $\widetilde{R}_T^0(\widetilde{\alpha})$ (resp., $\widetilde{R}_T^k(\widetilde{\alpha})$) that the strategy $\widetilde{\alpha}$ suffers under probability $\mathbb{P}^0$ (resp., $\mathbb{P}^k$) in the multi-apple tasting problem. Finally, the proof can be concluded by putting together the lower bound $\Omega\big(\min(\frac{K}{\varepsilon^2}, \varepsilon T)\big)$ for the multi-apple tasting problem with our choices of K and ε to obtain that the minimax regret for the bilateral trade problem is at least of order $\Omega(T^{3/4})$.

2.7 Conclusions

In this chapter, we provided a thorough study of the bilateral trade problem in a regret minimization framework. We proved tight bounds on the regret rates that can be achieved under various feedback, private valuation models, and various budget-balanced conditions. Our work opens several possibilities for future investigation.

First, with the exceptions of the (iv) + (bd) + (iid) case when sellers and buyers share the same distribution, and the full-feedback (bd) case, we have obtained tight (up to constant or logarithmic factors) regret rates in the time horizon only. In the other cases where the (bd) assumption plays a crucial role in learning, the problem of obtaining tight regret dependencies in the time horizon and the (bd) density parameter M simultaneously is still open.

Second, it would be interesting to study the contextual version of the bilateral trade problem, where a context related to the traders' valuations is available to the broker/learner before making a decision, and, possibly, multiple traders arrive at each time step.

Finally, an interesting research direction is the study of different reward functions. For example, in the case of weakly budget-balanced mechanisms, a sensible problem is to consider the amount of money extracted in the trade by the broker as the reward function, or perhaps weighted versions of the gain from trade.

Chapter 3

The Role of Transparency in Repeated First-Price Auctions with Unknown Valuations

3.1 Introduction

The online advertising market has recently transitioned from second to first-price auctions. A recent remarkable example is Google AdSense's move at the end of 2021 [186], following the switch made by Google AdManager and AdMob. Earlier examples also include OpenX, AppNexus, Index Exchange, and Rubicon [171]. With the purpose of increasing transparency, some platforms (like AdManager) have a single bidding session for each available impression (unified bidding) and require all partners to share and receive bid data; in particular, bidders receive the minimum bid price which would have won them the impression after the first-price auction closes [33].

In practice, advertisers face multiple sources of uncertainty at the moment of bidding. Besides ignoring the value of the competing bids, they also ignore the actual value of the impression they are bidding on. Indeed, clicks and conversion rates can only be measured *after* the auction is won and the ad is displayed, can vary wildly over time, or be highly correlated with competing bids. We remark that ignoring the value of the impression has a strong effect on the utility of the bidder: it may lead to overbidding for an impression of low value or, conversely, underbidding and losing a valuable one. To cope with this uncertainty, advertisers rely on auto-bidders that use the feedback provided in the auctions to learn good bidding strategies. We study the learning problem faced by a single bidder within the framework of regret minimization according to the following protocol:

Online Bidding Protocol for Repeated First-Price Auctions with Unknown Valuations

for time $t = 1, 2, \dots$ **do**
The valuation $V_t \in [0, 1]$ and the highest competing bid $M_t \in [0, 1]$ are privately generated
The learner posts a bid $B_t \in [0, 1]$ and gets utility $\text{Util}_t(B_t) = (V_t - B_t)\mathbb{I}\{B_t \geqslant M_t\}$
The learner observes some feedback Z_t

In this work, we are specifically interested in understanding how the "transparency" of the auctions—i.e., the amount of information on competing bids disclosed by the auctioneer *after* the auction takes place—affects the learning process. There is a clear tension regarding transparency:

on the one hand, bidders want to receive as much information as possible about the environment to learn the competitor's bidding strategies, while revealing as little as possible about their (private) bids. On the other hand, the publisher may not want to publicly reveal her revenue (i.e., the winning bid). It is the auctioneer's choice to decide the level of transparency to motivate bidders and publishers to participate in the auctions. The role of transparency in repeated first-price auctions has been investigated by Bergemann and Hörner [30], but mostly from a game-theoretic viewpoint. In particular, they study the impact of the feedback policy on the bidders' strategy and show how disclosing the bids at the end of each round affects the equilibria of a bidding game with infinite horizon. In contrast, we want to characterize the impact of different amounts of feedback (or degrees of transparency) on the learner's regret, which is measured against the optimal fixed bid in hindsight. To model the level of transparency, we distinguish four natural types of feedback Z_t (see the table below here), specifying the conditions under which the highest competing bid M_t and the bidder's valuation V_t are revealed to the bidder after each round t.

<table>
<tr><th></th><th>M_t</th><th>V_t</th></tr>
<tr><td>Full</td><td colspan="2">Always observed</td></tr>
<tr><td>Transparent</td><td>Always observed</td><td rowspan="3">Observed if the auction is won</td></tr>
<tr><td>Semi-Transparent</td><td>Observed if the auction is lost</td></tr>
<tr><td>Bandit</td><td>Never observed</td></tr>
</table>

In the transparent feedback setting, M_t is always observed after the auction is concluded, while V_t is only known if the auction is won, that is when $B_t \geqslant M_t$. In the semi-transparent setting, instead, M_t is only observed when the auction is lost. In other words, in the semi-transparent setting, each bidder only observes the highest bid, whereas, in the transparent setting, the winning bidder also observes the second highest bid. We also consider two extreme settings: full feedback (M_t and V_t are always observed irrespective of the auction's outcome) and bandit feedback (M_t is never observed while V_t is only observed by the winning bidder). Note that the learner can compute the value of the utility $\mathrm{Util}_t(B_t)$ at time t with any type of feedback, including bandit feedback. In this chapter, we characterize the learner's minimax regret not only with respect to the degree of transparency of the auction, but also with respect to the nature of the process generating the sequence of pairs (V_t, M_t). In particular, we consider four types of environments: stochastic i.i.d., adversarial, and their smooth versions (see the end of Section 3.1.3 for a discussion about smoothness, and Section 3.2 for the formal definition). We refer to Table 3.1 for a summary of our results.

3.1.1 Overview of our Results

In the following discussion we ignore logarithmic factors.

Stochastic i.i.d. settings

- In both the full and transparent feedback models, the minimax regret is of order $\sqrt{T}$ (Theorems 21 and 22), and adding the smoothness requirement leaves this rate unchanged.
- In the semi-transparent feedback model, the minimax regret is of order $T^{2/3}$ (Theorems 19 and 20). Also in this case, adding the smoothness requirement leaves this rate unchanged.
- In the bandit feedback model, smoothness is crucial to achieve a sublinear regret (Theorem 18). In particular, smoothness implies a minimax regret of $T^{2/3}$ (this is obtained by combining the

	Stochastic i.i.d.		Adversarial	
	Smooth	General	Smooth	General
Full	Thm 22: $\Omega(\sqrt{T})$			Thm 25: $\Omega(T)$
Transparent		Thm 21: $O(\sqrt{T})$	Thm 24: $\widetilde{O}(\sqrt{T})$	
Semi-Transparent	Thm 20: $\Omega\big(T^{2/3}\big)$	Thm 19: $\widetilde{O}\big(T^{2/3}\big)$		
Bandit		Thm 18: $\Omega(T)$	Thm 23: $O\big(T^{2/3}\big)$	

Table 3.1: Summary of our results. The rows correspond to feedback models and the columns to environments. The minimax regret of every problem has been characterized, resulting in one of the following three regimes: $\widetilde{\Theta}(\sqrt{T})$ (green squares), $\widetilde{\Theta}(T^{2/3})$ (yellow squares) and $\widetilde{\Theta}(T)$ (red squares).

upper bound in Theorem 23 and the lower bound in Theorem 20).

Adversarial settings

- Without smoothness, sublinear regret cannot be achieved, even with full feedback (Theorem 25).
- In both the full and transparent feedback model, the minimax regret in a smooth environment is of order $\sqrt{T}$ (combining the lower bound in Theorem 22 and the upper bound in Theorem 24).
- Both with semi-transparent and bandit feedback, the minimax regret in a smooth environment is of order $T^{2/3}$ (combining the lower bound in Theorem 20 and the upper bound in Theorem 23).

The minimax regret rates for first-price auctions mirror the allowed regret regimes in finite partial monitoring games [28] and in online learning with feedback graphs [9]. However, as shown by Lattimore [122] —and as we already seen in Section 2.6— games with continuous outcome/action spaces allow for a much larger set of regret rates.

Table 3.1 reveals some interesting properties of the minimax regret for this problem: full feedback and transparent feedback are essentially equivalent while semi-transparent feedback and bandit feedback differ only in the stochastic i.i.d. setting. Moreover, while smoothness is key for learning in the adversarial setting, in the stochastic case smoothness is only relevant for bandit feedback.

3.1.2 Technical Challenges

The utility function. The utility functions $b \mapsto \mathrm{Util}_t(b) = (V_t - b)\mathbb{I}\{M_t \leqslant b\}$ are defined over a continuous decision space $[0,1]$ and are not Lipschitz (even the weaker property that the expected cumulative reward $b \mapsto \sum_{t\in[T]} \mathbb{E}\big[\mathrm{Util}_t(b)\big]$ is one-sided Lipschitz does not hold in general). We address this problem by developing techniques designed to control the approximation error incurred when discretizing the bidding space. In the stochastic i.i.d. setting, the approximation error is controlled by adaptively building a *non-uniform* grid. This allows us to estimate the distribution of these competing bids, uniformly over the subintervals of $[0,1]$. In the adversarial setting, instead, we use the smoothness assumption to guarantee that the expected utility is Lipschitz. In this case, the approximation error is controlled using a uniform grid with an appropriate grid-size (Lemma 10).

The feedback models. Our feedback models interpolate between bandit (only the bidder's utility is observed) and full feedback (V_t and M_t are always observed). In the stochastic i.i.d. case, the different levels of transparency are crucial to the process of building the non-uniform grids used to control the discretization error. In the adversarial case, when there are only K allowed bids, the optimal rates are $\sqrt{T \ln K}$ and $\sqrt{KT}$ under full and bandit feedback, respectively. While the

semi-transparent feedback is not enough to improve the bandit rate, the transparent one can be exploited via a more sophisticated approach. To this end, we design an algorithm, Exp3.FPA, enjoying the full feedback regret rate $\sqrt{T \ln K}$ while only relying on the weaker transparent feedback.

Lower bounds. The linear lower bounds (Theorems 18 and 25) exploit a "needle in a haystack" phenomenon, where there is a hidden optimal bid $b^\star$ (the needle) in the $[0, 1]$ interval (the haystack) and the learner has no way of finding $b^\star$ using the feedback she has access to. This is indeed the case in the non-smooth adversarial full-feedback setting and in the non-smooth i.i.d. bandit setting. To prove the remaining lower bounds, we design careful embeddings of known hard instances into our framework. In particular, in Theorem 22 we embed the hard instance for prediction with two experts and in Theorem 20 the hard instance for K (with $K = \Theta(T^{1/3})$) bandits.

3.1.3 Related Work

The role of transparency in first-price auctions, where the winning bid is disclosed at the end of each auction, has been studied in Bergemann and Hörner [30] with a focus on how transparency affects the equilibria of the repeated bidding game.

Although the problem of regret minimization in first-price auctions has been studied before, only few papers consider the setting of unknown valuations. Feng et al. [90] introduce a general framework for the study of regret in auctions where a bidder's valuation is only observed when the auction is won. In the special case of first-price auctions, their setting is equivalent to our transparent feedback when the sequence of pairs (V_t, M_t) is adversarially generated. Following a parameterization introduced by Weed et al. [182], Feng et al. [90] provide a $O\big(\sqrt{T \ln \max\{\Delta_0^{-1}, T\}}\big)$ regret bound, where $\Delta_0 = \min_{t<t'} |M_t - M_{t'}|$ is controlled by the environment. In the stochastic i.i.d. case, their results translate into *distribution-dependent* guarantees not providing any worst-case sublinear bound (we obtain a $\sqrt{T}$ rate). In the adversarial case, their guarantees are still worst-case linear (we obtain $\sqrt{T}$ bounds leveraging the smoothness assumption). Achddou et al. [2] consider a stochastic i.i.d. setting with the additional assumption that V_t and M_t are independent. Their main result is a bidding algorithm with *distribution-dependent* regret rates (of order $T^{1/3+\varepsilon}$ or $\sqrt{T}$, depending on the assumptions on the underlying distribution) in the transparent setting. Again, this result is not comparable to ours because of the independence assumption and the distribution-dependent rates (which do not allow to recover our minimax rates).

Other works consider regret minimization in repeated second-price auctions with unknown valuations. Dikkala and Tardos [81] investigate a repeated bidding setting, but do not consider regret minimization. Weed et al. [182] derive regret bounds for the case when M_t are adversarially generated, while V_t are stochastically or adversarially generated and the feedback is transparent.

Considerably more work study first-price auctions when the valuation V_t is known to the bidder at the beginning of each round t. Note that these results are not directly comparable to ours. Balseiro et al. [27] look at the case when the V_t are adversarial and the M_t are either stochastic i.i.d. or adversarial. In the bandit feedback case (when M_t is never observed), they show that the minimax regret is $\widetilde{O}\big(T^{2/3}\big)$ in the stochastic case and $\widetilde{O}\big(T^{3/4}\big)$ in the adversarial case. Han et al. [105] prove a $\widetilde{O}\big(\sqrt{T}\big)$ regret bound in the semi-transparent setting (M_t observed only when the auction is lost) with adversarial valuations and stochastic bids. Han et al. [104] focus on the adversarial case, when V_t and M_t are both generated adversarially. They prove a $\widetilde{O}\big(\sqrt{T}\big)$ regret bound in the full feedback

setting (M_t always observed) when the regret is defined with respect to all Lipschitz shading policies. This setup is extended in Zhang et al. [190] where the authors consider the case in which the bidder is provided access to hints before each auction. Zhang et al. [189] also study the full information feedback setting and design a space-efficient variant of the algorithm proposed by Han et al. [104]. Badanidiyuru et al. [25] introduce a contextual model in which V_t is adversarial and $M_t = \langle \theta, x_t \rangle + \varepsilon_t$ where $x_t \in \mathbb{R}^d$ is contextual information available at the beginning of each round t, $\theta \in \mathbb{R}^d$ is an unknown parameter, and ε_t is drawn from an unknown log-concave distribution. They study regret in bandit and full feedback settings.

A different thread of research is concerned with the convergence property of the regret minimization dynamics in first-price auctions (or, more specifically, with the learning dynamics of mean-based regret minimization algorithms). Feldman et al. [89] show that with continuous bid levels, coarse-correlated equilibria exist whose revenue is below the second price. Feng et al. [91] prove that regret minimizing bidders converge to a Bayesian Nash equilibrium in first-price auctions when bidder values are drawn i.i.d. from a uniform distribution on $[0, 1]$. Kolumbus and Nisan [118] show that if two bidders with finitely many bid values converge, then the equilibrium revenue of the bidder with the highest valuation is the second price. Deng et al. [78] provide a characterization of the equilibria of the learning dynamics depending on the number of bidders with the highest valuation. Their characterization is for both time-average and last-iterate convergence.

Finally, smoothed analysis of algorithms, originally introduced by Spielman and Teng [172] and later formalized for online learning by Rakhlin et al. [153] and Haghtalab et al. [101], is a known approach to the analysis of algorithms in which the instances at every round are generated from a distribution that is not too concentrated. Recent works on the smoothed analysis of online learning algorithms include Haghtalab et al. [101], Haghtalab et al. [103], Block et al. [35], Durvasula et al. [85], and the papers on bilateral trade [40, 55, 57, 60] upon which Chapter 2 is based (where we used a smoothness parameterization via the bounded density parameter M).

3.2 The Learning Model

We introduce formally the repeated bidding problem in first-price auctions. At each time step t, a new item arrives for sale, for which the learner holds some unknown valuation $V_t \in [0, 1]$. The learner bids some $B_t \in [0, 1]$ and, at the same time, a set of competitors bid for the same object. We denote their highest competing bid by $M_t \in [0, 1]$. The learner gets the item at cost B_t if she wins the auction (i.e., if $B_t \geqslant M_t$), and does not get it otherwise. Then, the learner observes some feedback Z_t and gains utility $\mathrm{Util}_t(B_t)$, where, for all $b \in [0, 1]$, $\mathrm{Util}_t(b) = (V_t - b)\mathbb{I}\{b \geqslant M_t\}$ (see the online bidding protocol in Section 3.1). Crucially, at time t the learner does not know her valuation V_t for the item before bidding, implying that her bid B_t only depends on her past observations $Z_1, \ldots, Z_{t-1}$ (and, possibly, some internal randomization). The goal of the learner is to design a learning algorithm α that maximizes her utility. More precisely, we measure the performance of an algorithm α by its *regret* $R_T^{\mathcal{S}}(\alpha)$ against the worst environment β in a certain class $\mathcal{S}$, where

$$R_T^{\mathcal{S}}(\alpha) \coloneqq \sup_{\beta \in \mathcal{S}} R_T(\alpha, \beta) \,, \qquad R_T(\alpha, \beta) \coloneqq \sup_{b \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(b) - \sum_{t=1}^{T} \mathrm{Util}_t(B_t) \right] \,,$$

and the expectation is taken with respect to the randomness of the algorithm α which selects B_t, and (possibly) the randomness of the environment β generating the (V_t, M_t) pairs.

The environments. In this chapter we consider both stochastic i.i.d. and adversarial environments.

- Stochastic i.i.d.: The pairs $(V_1, M_1), (V_2, M_2), \dots$ are a stochastic i.i.d. process.
- Adversarial: The sequence $(V_1, M_1), (V_2, M_2), \dots$ is generated by an oblivious adversary.

Following previous works in online learning (see Section 3.1.3), we also study versions of the above environments that are constrained to generate the sequence of (V_t, M_t) values using distributions that are "not too concentrated". To this end, we introduce the notion of smooth distributions.

Definition 1 ([102]). *Let $\mathcal{X}$ be a domain that supports a uniform distribution ν. A measure μ on $\mathcal{X}$ is said to be σ-smooth if for all measurable subsets $A \subseteq \mathcal{X}$, we have $\mu(A) \leqslant \frac{\nu(A)}{\sigma}$.*[*]

We thus also consider the following two types of environments.

- The σ-smooth stochastic i.i.d. environment, which is a stochastic i.i.d. environment where the distribution of each pair $(V_1, M_1), (V_2, M_2), \dots$ is σ-smooth.
- The σ-smooth adversarial setting, where the pairs $(V_1, M_1), (V_2, M_2), \dots$ form a stochastic process such that, for each t, the distribution of the pair (V_t, M_t) is σ-smooth.

The feedback. Once we have described the types of environments we study, we specify the types of feedback the learner receives at the end of each round, from the richest to the less informative.

- Full Feedback. The learner observes her valuation and the highest competing bid: $Z_t = (V_t, M_t)$.
- Transparent Feedback. The learner always observes M_t, but V_t is only revealed if she gets the item: Z_t is equal to $(\star, M_t)$ if $B_t < M_t$ and (V_t, M_t) otherwise.
- Semi-Transparent Feedback[†]. The learner observes V_t if she gets the item and M_t otherwise: Z_t is equal to $(\star, M_t)$ if $B_t < M_t$ and $(V_t, \star)$ otherwise.
- The bandit feedback[‡]. The learner observes V_t if she gets the item and the symbol $\star$ otherwise: Z_t is $\star$ if $B_t < M_t$ and V_t otherwise.

3.3 The Stochastic i.i.d. Setting

In this section, we investigate the problem of repeated bidding in first-price auctions with unknown valuations, when the pairs of valuations and highest competing bids are drawn i.i.d. from a fixed but unknown distribution. We study the different feedback models separately. We start by proving in Section 3.3.1 that it is not possible to achieve sublinear regret under the bandit feedback model without any assumption on the distribution of the environment. Then, in Section 3.3.2 we give matching upper and lower bounds of order $T^{2/3}$ in the semi-transparent feedback model. Notably, the latter lower bound holds for smooth distributions, while the upper bound works for any (possibly

[*]It is worthwhile noticing that μ being σ-smooth is equivalent to μ being absolutely continuous with respect to ν with the further requirement that the Radon-Nikodym derivative satisfies $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \leqslant \frac{1}{\sigma}$. In Chapter 2, to control the smoothness of a distribution we used the (bd) assumption, with the different but equivalent parameterization $M = \frac{1}{\sigma}$.

[†]This feedback is similar to the winner-only feedback in Han et al. [105].

[‡]We call this the bandit feedback because it is equivalent to receiving $\mathrm{Util}_t(B_t)$ (with the extra information $\star$ to distinguish between losing the item and winning it with $V_t = B_t$).

non-smooth) distributions. Finally, in Section 3.3.3 we prove that both the full and transparent feedback yield the same minimax regret regime of order $\sqrt{T}$, regardless of the regularity of the distribution.

3.3.1 Stochastic i.i.d. Environment with Bandit Feedback

In the bandit feedback model, at each time step, the learner observes the valuation V_t (and nothing else) when she wins, she observes nothing at all when she loses the auction. The crucial difference with the other (richer) types of feedback is the amount of information received about M_t, which, in the bandit case, is just the relative position with respect to B_t (i.e., whether $M_t \leqslant B_t$ or $B_t < M_t$). This allows to hide in the interval $[0, 1]$ an optimal bid $b^\star$ which cannot be uncovered by the learner over a finite time horizon. Following this idea, a difficult environment should be one which randomizes between two scenarios: a good scenario with large value $V_t = 1$ and M_t slightly smaller than $b^\star$ and a bad one with poor value $V_t = 0$ and M_t slightly larger than $b^\star$. This way, not to suffer linear regret, the learner has to find this tiny interval around $b^\star$ (the "needle in a haystack").

Theorem 18. *Consider the problem of repeated bidding in first-price auctions with bandit feedback. Suppose that $\mathcal{S}$ is the set of stochastic i.i.d. environments. Then, for any learning algorithm α and any time horizon T, it holds*

$$R_T^{\mathcal{S}}(\alpha) \geqslant \frac{1}{20} T \, .$$

Proof. We construct a randomized i.i.d. environment β, such that any deterministic algorithm α suffers linear regret against it. By Yao's Minimax principle, this concludes the proof.

The randomized environment is simple: before starting the sequence, a uniform seed $b^\star$ is drawn uniformly at random in $(\frac{1}{3}, \frac{1}{2} - \varepsilon)$, where ε is a small parameter we set later. Then the i.i.d. sequence $(V_1, M_1), (V_2, M_2), \dots$ is drawn as follows: at each time step t with probability $\frac{1}{2}$ we have $(V_t, M_t) = (1, b^\star)$, otherwise $(0, b^\star + \varepsilon)$. The bid $b^\star$ is the best bid in hindsight, yielding an overall expected utility of $\frac{T}{2}(1 - b^\star)$, which is at least $\frac{T}{4}$ because $b^\star$ belongs to the interval $(\frac{1}{3}, \frac{1}{2})$.

We now upper bound the utility achievable by any deterministic algorithm α against β. Fix any such algorithm, and consider its bids against any environment that selects the valuations V_t to be either 0 or 1 (as the one constructed). At each time step, the feedback that α receives is either 0, 1 or $\star$ (when the item is allocated to one of the competitors), so that the history of the bids posted by α is naturally described by a ternary decision tree of height T, where each level corresponds to a time step and any node to a bid. Crucially, the leaves of this tree are finite (at most 3^T), which means that the algorithm α only posts bids in a finite subset N of $[0, 1]$. Now, let $\varepsilon = 3^{-2T}/12$; we have that, with probability at least $1 - \frac{6N\varepsilon}{1-6\varepsilon} \geqslant 1 - e^{-T}$, the set $[b^\star, b^\star + \varepsilon]$ does not intersect N. Note: the randomness is with respect to the uniform seed $b^\star$ drawn by β, while the bound on the probability holds independently to the choice of the deterministic algorithm α.

The total utility of α when $[b^\star, b^\star + \varepsilon]$ does not intersect N is easy to analyze: every time that α posts bids smaller than $b^\star$, then it never wins the item (zero utility). Instead, if it posts bids larger than $b^\star + \varepsilon$, then it always gets the item (whose average value is $\frac{1}{2}$), paying at least $b^\star + \varepsilon \geqslant \frac{1}{3}$. Putting these two cases together, we have proved that at each time step the expected utility earned by the learner is at most $\frac{1}{6} = \frac{1}{2} - \frac{1}{3}$, when $[b^\star, b^\star + \varepsilon] \cap N = \varnothing$ does not intersect N (which happens with probability at least $1 - e^{-T}$). Finally, by combining the lower bound

Collect Bids (CB) - Semi-Transparent Feedback

input: Time horizon T_0
Let $X_0 \coloneqq 0$ and $M^{(0)} \coloneqq 0$
for time $t = 1, 2, \ldots, T_0$ **do**
    Post bid $B_t \coloneqq 0$ and observe the highest competing bid M_t
Sort the observed highest competing bids in increasing order: $M^{(1)} \leqslant M^{(2)} \leqslant \cdots \leqslant M^{(T_0)}$
if $M^{(T_0)} = 0$ **then return** candidate bid X_0
for $i = 1, 2, \ldots$ **do**
    Let $j^\star_{i-1} \coloneqq \max\{j \in \{0, \ldots, T_0\} \mid X_{i-1} = M^{(j)}\}$, $j_i \coloneqq \min\{j^\star_{i-1} + \lceil\sqrt{T_0}\rceil, T_0\}$, $X_i \coloneqq M^{(j_i)}$
    if $j_i = T_0$ **then** let $K \coloneqq i$ and **break**;
return Candidate bids $X_0, X_1, X_2, \ldots, X_K$

on the performance of $b^\star$ with the upper bound on the expected utility of the learner, we get $R_T(\alpha, \beta) \geqslant (1 - e^{-T})(T/4 - T/6) \geqslant T/20$. □

3.3.2 Stochastic i.i.d. Environment with Semi-Transparent Feedback

In this section, we prove two results settling the minimax regret for the semi-transparent feedback where the environment is i.i.d. (and, possibly, smooth). First, we construct a learning algorithm, Collecting Bandit, achieving $T^{2/3}$ regret against any i.i.d. environment. Then, we complement it with a lower bound of the same order (up to log terms) obtained even in a smooth i.i.d. environment.

A $T^{2/3}$ upper bound for the i.i.d. environment

Our learning algorithm Collecting Bandit is composed of two phases. First, for $T_0 = \Theta(T^{2/3})$ rounds, it collects samples from the highest competing bid random variables $M_1, M_2, \ldots, M_{T_0}$ by posting dummy bids $B_1 = B_2 = \cdots = B_{T_0} = 0$. Among these values (plus the value $X_0 = 0$), the algorithm selects $\Theta(\sqrt{T_0})$ bids according to their ordering, in a manner that the empirical frequencies of bids $M_1, M_2, \ldots, M_{T_0}$ landing strictly in between two consecutive selected values are at most $\Theta(1/\sqrt{T_0})$ (see the pseudo-code of Collect Bids for details). Second, for the remaining time steps, it runs any bandit algorithm, using as candidate bids the ones collected in the first phase (see Collecting Bandit for details). Note that, in this second phase, the (less informative) bandit feedback would be enough to run the algorithm: we only used the additional information provided by the semi-transparent feedback in the initial "collecting bids" phase.

We first state a simple concentration result pertaining the i.i.d. process $M, M_1, M_2, \ldots, M_{T_0}$, for $T_0 \in \mathbb{N}$. If $\mathcal{I}$ is the family of all the subintervals of $[0, 1]$ and $\delta \in (0, 1)$, we define

$$\mathcal{E}_\delta^{T_0} \coloneqq \bigcap_{I \in \mathcal{I}} \left\{ \left| \tfrac{1}{T_0} \textstyle\sum_{t=1}^{T_0} \mathbb{I}\{M_t \in I\} - \mathbb{P}[M \in I] \right| < 8\sqrt{\tfrac{\ln(1/\delta)}{T_0}} \right\}.$$

Lemma 7. *For every $T_0 \in \mathbb{N}$ and $\delta \in (0, 1)$, we have $\mathbb{P}[\mathcal{E}_\delta^{T_0}] \geqslant 1 - \delta$.*

Proof. The family $\mathcal{I}$ of all the subintervals of $[0, 1]$ has VC dimension 2 (see, e.g., Chapter 14.2. of Mitzenmacher and Upfal [141]). Therefore we get the desired result by directly applying the standard sample complexity bound for ε-samples (see, e.g., Theorem 14.15 of Mitzenmacher and Upfal [141]) for T_0 samples and $\varepsilon = 8\sqrt{\frac{\ln(1/\delta)}{T_0}}$. □

To lighten future notation, we introduce the following

Notation 1. *If $K \in \mathbb{N}$, $0 = x_0 < x_1 < \cdots < x_K \leqslant 1 < x_{K+1} = 2$, and $\mathcal{X} = \{x_0, \dots, x_K\}$, we denote by $k_{\mathcal{X}} \colon [0,1] \to \{0, 1, \dots, K\}$ the function that maps each $b \in [0,1]$ to the unique $k \in \{0, 1, \dots, K\}$ such that $x \in [x_k, x_{k+1})$.*

We now prove another lemma that allows us to control the expected cumulative utility of any bid in $[0,1]$ with that of the best bid in a discretization (without relying on any smoothness assumption).

Lemma 8. *Assume that the process $M, M_1, M_2, \dots$ of the highest competing bids form an i.i.d. sequence. Let also $0 = x_0 < x_1 < \cdots < x_K \leqslant 1 < x_{K+1} = 2$ and $\mathcal{X} = \{x_0, \dots, x_K\}$. For all $b \in [0,1]$ and $T_0, T_1 \in \mathbb{N}$ with $T_0 < T_1$, we have:*

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(b)\right] \leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t\big(x_{k_{\mathcal{X}}(b)}\big)\right] + (T_1 - T_0)\mathbb{P}\big[x_{k_{\mathcal{X}}(b)} < M < x_{k_{\mathcal{X}}(b)+1}\big] .$$

Proof. Fix any $b \in [0,1]$, $T_0, T_1 \in \mathbb{N}$ with $T_0 < T_1$, and a time step $t \in \{T_0 + 1, \dots, T_1\}$. Then

$$\begin{aligned}\mathbb{E}\big[\mathrm{Util}_t(b)\big] &= \mathbb{E}\big[(V_t - b)\mathbb{I}\{b \geqslant M_t\}\big] \leqslant \mathbb{E}\big[(V_t - x_{k_{\mathcal{X}}(b)})\big(\mathbb{I}\{x_{k_{\mathcal{X}}(b)} \geqslant M_t\} + \mathbb{I}\{b \geqslant M_t > x_{k_{\mathcal{X}}(b)}\}\big)\big] \\ &\leqslant \mathbb{E}\big[\mathrm{Util}_t(x_{k_{\mathcal{X}}(b)})\big] + \mathbb{P}[x_{k_{\mathcal{X}}(b)} < M_t \leqslant b] \leqslant \mathbb{E}\big[\mathrm{Util}_t(x_{k_{\mathcal{X}}(b)})\big] + \mathbb{P}[x_{k_{\mathcal{X}}(b)} < M_t < x_{k_{\mathcal{X}}(b)+1}] .\end{aligned}$$

Summing over t and recalling that M_t and M shares the same distribution, yields the conclusion. □

As a corollary of Lemmas 7 and 8 we obtain similar discretization error guarantees when the grid of points $\mathcal{X}$ is random.

Lemma 9. *Fix any $T_0 \in \mathbb{N}$ and $\delta \in (0,1)$. Let $\mathcal{X} = \{X_0, \dots, X_K\}$ be a random set containing a random number K of points satisfying $0 = X_0 < X_1 < \cdots < X_K \leqslant 1 < X_{K+1} = 2$. Assume that the random variables $K, X_0, X_1, \dots, X_{K+1}$ are $\mathcal{H}_{T_0}$-measurable, where $\mathcal{H}_{T_0}$ is the history up to and including time T_0. Assume that the process $(V_1, M_1), (V_2, M_2), \dots$ of the valuations/highest competing bids form an i.i.d. sequence. Then, for all $b \in [0,1]$ and $T_1 \in \mathbb{N}$ with $T_1 > T_0$, we have:*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(b)\right] &\leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t\big(X_{k_{\mathcal{X}}(b)}\big)\right] \\ &\quad + (T_1 - T_0)\mathbb{E}\left[\frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{I}\big\{X_{k_{\mathcal{X}}(b)} < M_t < X_{k_{\mathcal{X}}(b)+1}\big\}\right] + (T_1 - T_0)\left(8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) .\end{aligned}$$

We are now ready to present the main theorem of this section.

Theorem 19. *Consider the problem of repeated bidding in first-price auctions with semi-transparent feedback. Suppose that $\mathcal{S}$ is the set of stochastic i.i.d. environments. Then there exists a learning algorithm α such that*

$$R_T^{\mathcal{S}}(\alpha) \leqslant 16\big(13 + \sqrt{\ln T}\big)T^{2/3} .$$

Proof. We prove that Collecting Bandit yields the desired bound when its learning routine $\widetilde{\alpha}$ is (a rescaled version of) MOSS [15]: since MOSS is designed to run with gains in $[0,1]$ while the utilities we observe are in $[-1,1]$, we first apply the reward transformation $x \mapsto \frac{x+1}{2}$ to the observed utilities.

Collecting Bandit (CoBa) - Semi-Transparent Feedback

input: Time horizon T and a bandit algorithm $\widetilde{\alpha}$ for gains in $[-1,1]$
$T_0 \coloneqq \lceil T^{2/3} \rceil$
Run Collect Bids with time horizon T_0 and obtain $X_0, X_1, \dots, X_K$
Initialize $\widetilde{\alpha}$ on $K+1$ actions (one for each candidate bid X_i) and $T - T_0$ as time horizon
for time $t = T_0 + 1, T_0 + 2, \dots, T$ **do**
    Receive from $\widetilde{\alpha}$ the arm $I_t \in \{0, 1, \dots, K\}$
    Post bid $B_t \coloneqq X_{I_t}$ and observe semi-transparent feedback Z_t
    Reconstruct $\mathrm{Util}_t(B_t)$ from Z_t and feed it to $\widetilde{\alpha}$ as the reward associated to I_t

This will cost a multiplicative factor of 2 on the regret guarantees of MOSS. Leveraging the fact that the empirical frequency between two consecutive X_k and X_{k+1} generated by Collect Bids is at most $2/\sqrt{T_0}$ by design and applying Lemma 9 with $T_1 = T$ to the random variables $X_0, X_1, \dots, X_K$, we obtain, for all $b \in [0,1]$

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t(b)\right] \leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t\big(X_{k_{\mathcal{X}}(b)}\big)\right] + (T - T_0)\left(\frac{2}{\sqrt{T_0}} + 8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) = (\star) \,.$$

Now, applying the tower rule to the expectation on the right-hand side conditioning to the history $\mathcal{H}_{T_0}$ up to time T_0, we can use the fact that the regret of the rescaled version of MOSS is upper bounded by $98\sqrt{(K+1)(T-T_0)}$ and the number of points $K+1$ collected by Collect Bids is at most $\sqrt{T_0} + 1$ to obtain

$$(\star) \leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t(B_t)\right] + 98\sqrt{(\sqrt{T_0}+1)(T-T_0)} + (T - T_0)\left(\frac{2}{\sqrt{T_0}} + 8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) \,.$$

Finally, tuning $\delta = 1/T_0$, upper bounding the cumulative regret over the first T_0 rounds with T_0, and recalling that $T_0 = \lceil T^{2/3} \rceil$, yields the conclusion. □

A $T^{2/3}$ lower bound for the smooth i.i.d. environment

We prove here that the $\widetilde{O}(T^{2/3})$ bound achieved by Collecting Bandit is indeed optimal in the i.i.d. setting (up to logarithmic terms), even if we further impose that the environment is smooth. Our lower bound consists in carefully embedding into our model a hard multiarmed bandit instance with $K = \Theta(T^{1/3})$ arms, which entails a lower bound of order $\Omega(\sqrt{KT}) = \Omega(T^{2/3})$. Note that the proof agenda we have presented is rich of challenges: we want to embed a discrete construction on K independent actions into our continuous framework, where the utility of different bids are correlated, while enforcing smoothness. Furthermore, the feedback models are different. We report here a proof sketch and refer the interested reader to Appendix B.1 for the missing details.

Theorem 20. *Consider the problem of repeated bidding in first-price auctions with semi-transparent feedback. Suppose that $\mathcal{S}$ is the set of σ-smooth i.i.d. environments, with $\sigma \in (0, \frac{1}{66}]$. Then, for any learning algorithm α and any time horizon $T \geqslant 8$, it holds*

$$R_T^{\mathcal{S}}(\alpha) \geqslant \frac{3}{10^4} T^{2/3} \,.$$

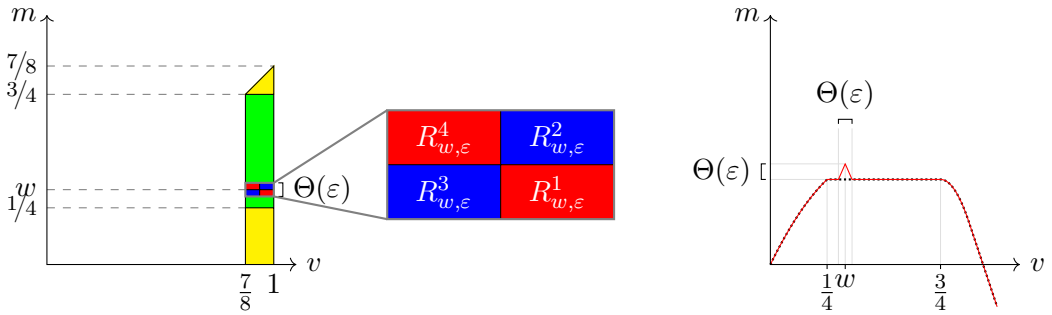


Figure 3.1: Left/center: The support of the base density f lies inside the yellow and green regions. The perturbation $g_{w,\varepsilon}$ of f occurs inside the green region, where the four rectangles $R^1_{w,\varepsilon}, \dots, R^4_{w,\varepsilon}$ (in red and blue) lie. Right: The corresponding qualitative plots of $b \mapsto \mathbb{E}[\mathrm{Util}_t(b)]$ (black, dotted) and $p \mapsto \mathbb{E}^{w,\varepsilon}[\mathrm{Util}_t(b)]$ (red, solid).

Proof sketch. Define, for all $v, m \in [0,1]$, the density

$$f(v,m) \coloneqq \mathbb{I}_{[\frac{7}{8},1]}(v)\left(\frac{1}{(v-m)^2}\mathbb{I}_{[\frac{1}{4},v-\frac{1}{8}]}(m) + \frac{4}{v-1/4}\mathbb{I}_{[0,\frac{1}{4})}(m)\right) .$$

Let $\mathbb{P}^0$ be a probability measure such that $(V,M),(V_1,M_1),(V_2,M_2),\dots$ is a $\mathbb{P}$-i.i.d. sequence where each pair (V,M) has common probability density function f. Denoting by $\mathbb{E}^0$ the expectation with respect to $\mathbb{P}^0$, we have, for any bid $b \in [0,1]$ and any t

$$\mathbb{E}^0\big[\mathrm{Util}_t(b)\big] = b\left(\frac{1}{2} + (1-4b)\ln\frac{6}{5}\right)\mathbb{I}_{[0,\frac{1}{4})}(b) + \frac{1}{8}\mathbb{I}_{[\frac{1}{4},\frac{3}{4})}(b) - \left(4b^2 - 6b + \frac{17}{8}\right)\mathbb{I}_{[\frac{3}{4},\frac{7}{8})}(b) + \left(\frac{15}{16} - b\right)\mathbb{I}_{[\frac{7}{8},1]}(b) .$$

This function grows with b on $[0,1/4)$, it has a plateau of maximizers $[1/4,3/4]$, then decreases on $(3/4,1]$ (see Figure 3.1, right). Now, let $\Xi \coloneqq \big\{(w,\varepsilon) \in [0,1]^2 : w-\varepsilon \geqslant 1/4 \text{ and } w+\varepsilon \leqslant 3/4\big\}$ and define, for all $(w,\varepsilon) \in \Xi$, the four rectangles $R^1_{w,\varepsilon} \coloneqq [15/16,\,1] \times [w-\varepsilon,\,w)$, $R^2_{w,\varepsilon} \coloneqq [15/16,\,1] \times [w,\,w+\varepsilon)$, $R^3_{w,\varepsilon} \coloneqq [7/8,\,15/16) \times [w-\varepsilon,\,w)$, $R^4_{w,\varepsilon} \coloneqq [7/8,\,15/16) \times [w,\,w+\varepsilon)$, and, for all $v,m \in [0,1]$, the perturbation

$$g_{w,\varepsilon}(v,m) \coloneqq \frac{16}{9}\big(\mathbb{I}_{R^1_{w,\varepsilon}\cup R^4_{w,\varepsilon}}(v,m) - \mathbb{I}_{R^2_{w,\varepsilon}\cup R^3_{w,\varepsilon}}(v,m)\big) .$$

For all $(w,\varepsilon) \in \Xi$, define $f_{w,\varepsilon} \coloneqq f + g_{w,\varepsilon}$ (see Figure 3.1, left/center) and note that it is a valid probability density function, i.e., $f_{w,\varepsilon} \geqslant 0$ and $\int_{[0,1]^2} f_{w,\varepsilon}(v,m)\,\mathrm{d}v\mathrm{d}m = 1$. For all $(w,\varepsilon) \in \Xi$, let $\mathbb{P}^{w,\varepsilon}$ be a probability measure such that $(V,M),(V_1,M_1),(V_2,M_2),\dots$ is a $\mathbb{P}^{w,\varepsilon}$-i.i.d. sequence where each pair (V,M) has common probability density function $f_{w,\varepsilon}$. Denoting by $\mathbb{E}^{w,\varepsilon}$ the expectation with respect to $\mathbb{P}^{w,\varepsilon}$, we have, for any bid $b \in [0,1]$ and any t

$$\mathbb{E}^{w,\varepsilon}\big[\mathrm{Util}_t(b)\big] = \mathbb{E}^0\big[\mathrm{Util}_t(b)\big] + \frac{\varepsilon}{144}\Lambda_{w,\varepsilon}(b)$$

where $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as $\Lambda_{u,r}(x) \coloneqq \max\{1 - |x-u|/r, 0\}$. In words, in a perturbed scenario $\mathbb{P}^{w,\varepsilon}$ the expected utility is maximized at the peak of a spike centered at w with length and height $\Theta(\varepsilon)$ perturbing the plateau area $[1/4,\,3/4]$ of maximum height (see Figure 3.1, right). Define, for all times $t \in \mathbb{N}$, the feedback function

$$\psi_t \colon [0,1] \to \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big) , \qquad b \mapsto \begin{cases} (V_t, \star) & \text{if } b \geqslant M_t \\ (\star, M_t) & \text{if } b < M_t \end{cases}$$

and note that, in our semi-transparent feedback model, the feedback Z_t received after bidding B_t at time t is $\psi_t(B_t)$. Then, for each $(w, \varepsilon) \in \Xi$ and each $b \in [0,1] \backslash [w-\varepsilon, w+\varepsilon]$, note that the distribution of $\psi_t(b)$ under $\mathbb{P}^{w,\varepsilon}$ coincides with the distribution of $\psi_t(b)$ under $\mathbb{P}^0$, i.e., in push-forward notation (for a refresher on push-forward measures, see Appendix A.15),

$$\mathbb{P}^{w,\varepsilon}_{\psi_t(b)} = \mathbb{P}^0_{\psi_t(b)} \,. \tag{3.1}$$

Now, let $K \in \mathbb{N}$, $\varepsilon = 1/(4K)$, $w_k = 1/4 + (2k-1)\varepsilon$ and $\mathbb{P}^k = \mathbb{P}^{w_k,\varepsilon}$ (for each $k \in [K]$). At a high level, we built a problem in which we know in advance the region where the optimal bid belongs to (i.e., the interval $[1/4, 3/4]$), but, when the underlying scenario is determined by the probability measure $\mathbb{P}^k$ for some $k \in [K]$, in order not to suffer regret $\Omega(\varepsilon T)$, the learner has to detect inside this potentially optimal region where a spike of height (and length) $\Theta(\varepsilon)$ in the reward occurs. This last task can be accomplished only by locating where the perturbation in the base probability measure occurs, which, given the feedback structure, can only be done by playing in the interval $[w_k - \varepsilon, w_k + \varepsilon)$ if the underlying probability is $\mathbb{P}^k$, suffering instantaneous regret of order ε whenever the underlying probability is $\mathbb{P}^j$, with $j \neq k$. Given that we partitioned the potentially optimal region $[1/4, 3/4]$ into $\Theta(\frac{1}{\varepsilon})$ disjoint intervals where these perturbations can occur, the feedback structure implies that each of these intervals deserves its own dedicated exploration.

To better highlight this underlying structure, we will show (see Appendix B.1) that our problem is no easier than a simplified K-armed stochastic bandit problem, where the instances we consider are determined by the probability measures $\mathbb{P}^1, \dots, \mathbb{P}^K$. In this bandit problem, when the underlying probability measure is induced by some $\mathbb{P}^k$, the corresponding arm k has an expected reward $\Theta(\varepsilon)$ larger than the others. Then, via an information-theoretic argument, we can show that any learner would need to spend at least order of $1/\varepsilon^2$ rounds to explore each of the K arms (paying $\Omega(\varepsilon)$ each time) or else, she would pay a regret $\Omega(\varepsilon T)$. Hence, the regret of any learner, in the worst case, is lower bounded by $\Omega\big(\frac{K}{\varepsilon^2}\varepsilon + \varepsilon T\big) = \Omega\big(K^2 + T/K\big)$ (recalling our choice of $\varepsilon = 1/(4K)$). Picking $K = \Theta(T^{1/3})$ yields a lower bound of order $T^{2/3}$. For all missing technical details, see Appendix B.1. □

3.3.3 Stochastic i.i.d. Environment with Transparent and Full Feedback

This section completes the study of the stochastic i.i.d. environment by determining the minimax regret when the learner has access to full or transparent feedback.

A $\sqrt{T}$ upper bound for the i.i.d. environment

While with semi-transparent feedback, we had to rely on dummy bids $B_1 = \cdots = B_{T_0} = 0$ to gather information about the distribution of the highest competing bids, with the transparent one, this information is collected for free at each bidding round. To use this extra information, we present a wrapper W.T.FPA (for a sequence of base learning algorithms for the transparent feedback model) whose purpose is restarting the learning process with a geometric cadence to update the set of candidate bids. We assume that each of the wrapped base algorithms $\widetilde{\alpha}_\tau$ can take as input any finite subset $\mathcal{X} \subset [0,1]$ and returns bids in $\mathcal{X}$. Furthermore, for all T', we let $\mathcal{R}_{T'}(\widetilde{\alpha}_\tau, \mathcal{X})$ be an upper bound on the regret over T' rounds of $\widetilde{\alpha}_\tau$ with input $\mathcal{X}$ against the best fixed $x \in \mathcal{X}$. Formally, we require that for any two times $T_0 < T_1$ such that $T' = T_1 - T_0$, the quantity $\mathcal{R}_{T'}(\widetilde{\alpha}_\tau, \mathcal{X})$ is an upper

upper bound on $\max_{x\in\mathcal{X}} \mathbb{E}\big[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(x) - \sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(B_t)\big]$, where $B_t \in \mathcal{X}$ is the sequence of prices played by $\widetilde{\alpha}_\tau$ (with input $\mathcal{X}$) when started at round $t = T_0 + 1$ and ran up to time T_1. Without loss of generality, we assume that $T' \mapsto \mathcal{R}_{T'}(\widetilde{\alpha}_\tau, \mathcal{X})$ is non-decreasing.

W.T.FPA (Wrapper for Transparent First-Price Auctions) - Transparent Feedback

input: Base algorithms $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \dots$
initialization: Let $s \coloneqq 0$
for each epoch $\tau = 1, 2, \dots$ **do**
    Let $\mathcal{X}_\tau \coloneqq \{0\} \cup \{M_1, \dots, M_s\}$ (with the understanding that $\mathcal{X}_1 \coloneqq \{0\}$)
    Start $\widetilde{\alpha}_\tau$ with input $\mathcal{X}_\tau$ and run it for rounds $t = s+1, \dots, s+2^{\tau-1}$
    Update $s \coloneqq s + 2^{\tau-1}$

Proposition 1. *Consider the problem of repeated bidding in first-price auctions with transparent feedback. Suppose that $\mathcal{S}$ is the set of stochastic i.i.d. environments. Then, for any time horizon T, the regret of W.T.FPA run with base algorithms $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \dots$ satisfies*

$$R_T^{\mathcal{S}}(\text{W.T.FPA}) \leqslant \sum_{\tau=2}^{\lceil \log_2(T+1)\rceil} \mathcal{R}_{2^{\tau-1}}\big(\widetilde{\alpha}_\tau, \mathcal{X}_\tau\big) + 3 + 16\big(\sqrt{2}+2\big)\sqrt{T\ln T}\,.$$

Proof. Fix an arbitrary epoch $\tau \in \big\{2, \dots, \lceil \log_2(T+1)\rceil\big\}$ (the first epoch will be upper bounded separately). With respect to the notation in Lemma 9, let $\mathcal{X} = \mathcal{X}_\tau$, $K+1 = |\mathcal{X}|$, $T_0 = \sum_{\tau'=1}^{\tau-1} 2^{\tau'-1} = 2^{\tau-1} - 1$ (the time passed from the beginning of epoch 1 up to and including the end of epoch $\tau - 1$), $T_1 = \min\{T_0 + 2^{\tau-1}, T\}$ (the end of epoch τ), and let $X_0 < X_1 < \cdots < X_K$ be the distinct elements of $\mathcal{X}$ in increasing order, where we note that $X_0 = 0$, $X_K \leqslant 1$, and we set $X_{K+1} = 2$. Let also $\mathcal{H}_{T_0}$ be the history up to and including time T_0 and recall Notation 1. Applying first Lemma 9 (together with the fact that the empirical frequency between any two consecutive values X_k and X_{k+1} is 0 by design), then exploiting the monotonicity of $T' \mapsto \mathcal{R}_{T'}(\widetilde{\alpha}_\tau, \mathcal{X}_\tau)$ for the last epoch (if $T_0 + 2^{\tau-1} > T$), we obtain, for all $b \in [0,1]$ and $\delta \in (0,1)$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_0+1}^{\min\{T_0+2^{\tau-1},T\}} \mathrm{Util}_t(b)\right] &\leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{\min\{T_0+2^{\tau-1},T\}} \mathrm{Util}_t\Big(X_{k_{\mathcal{X}}(b)}\Big)\right] + 2^{\tau-1}\left(8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) \\
&\leqslant \mathbb{E}\left[\sum_{t=T_0+1}^{\min\{T_0+2^{\tau-1},T\}} \mathrm{Util}_t(B_t)\right] + \mathcal{R}_{2^{\tau-1}}\big(\widetilde{\alpha}_\tau, \mathcal{X}_\tau\big) + 2^{\tau-1}\left(8\sqrt{\frac{\ln(1/\delta)}{2^{\tau-1}-1}} + \delta\right)\,.
\end{aligned}$$

Summing over epochs $\tau \in \big\{2, \dots, \lceil \log_2(T+1)\rceil\big\}$, upper bounding by 1 the regret incurred in the first epoch, and tuning $\delta = 1/T$, yields the conclusion. □

Now we are only left to design appropriate base algorithms $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \dots$ for the transparent feedback to wrap W.T.FPA around.

The Exp3.FPA algorithm. To this end, we introduce the Exp3.FPA algorithm (designed to run with transparent feedback), which borrows ideas from online learning with feedback graphs [9]. Similar algorithms for related settings have been previously proposed by Weed et al. [182] and Feng et al. [90]. For the familiar reader, note that our setting can be seen as an instance of online learning

with strongly observable feedback graphs. In contrast to a black-box application of feedback-graph results, we shave off a logarithmic term (in the time horizon) by using a dedicated analysis. For any $x \in [0,1]$, we denote by δ_x the Dirac distribution centered at x.

Exp3.FPA - Transparent Feedback

1: **input:** Finite set $\mathcal{X} \subset [0,1]$ with maximum $\bar{x}$ and exploration rate $\gamma \in (0,1)$
2: For all $x \in \mathcal{X}$, let $w_1(x) \coloneqq 1$
3: **for** time $t = 1, 2, \dots$ **do**
4: Post bid $B_t \sim p_t \coloneqq (1-\gamma)\frac{w_t}{\|w_t\|_1} + \gamma\delta_{\bar{x}}$
5: For all $x \in \mathcal{X}$, define the reward estimate $\widehat{g}_t(x) \coloneqq (V_t - x)\mathbb{I}\{x \geqslant M_t\}\frac{\mathbb{I}\{M_t \leqslant B_t\}}{\sum_{y \geqslant M_t} p_t(y)}$
6: For all $x \in \mathcal{X}$, update the weight $w_{t+1}(x) \coloneqq w_t(x)\exp\big(\gamma\widehat{g}_t(x)\big)$

Note that the transparent feedback is sufficient to compute the reward estimates in Line 5. We defer the proof of the following proposition to Appendix B.2.

Proposition 2. *Let $\mathcal{X} \subset [0,1]$ be a finite set, $T \in \mathbb{N}$ a time horizon, and tune the exploration rate as $\gamma = \sqrt{\ln(|\mathcal{X}|)/(e-1)T}$. Then, the regret of Exp3.FPA against the best fixed bid in $\mathcal{X}$ is*

$$\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x) - \sum_{t=1}^{T} \mathrm{Util}_t(B_t)\right] \leqslant 2\sqrt{(e-1)\ln\big(|\mathcal{X}|\big)T}$$

Putting together Propositions 1 and 2 yields the desired optimal rate.

Theorem 21. *Consider the problem of repeated bidding in first-price auctions with transparent feedback. Suppose that $\mathcal{S}$ is the set of stochastic i.i.d. environments. Then the regret of W.T.FPA run with the base algorithm of each epoch τ being Exp3.FPA tuned with $\gamma = \gamma(\tau) = \sqrt{\ln(|\mathcal{X}_\tau|)/\left((e-1)2^{\tau-1}\right)}$, satisfies*

$$R_T^{\mathcal{S}}(\text{W.T.FPA}) \leqslant 3 + 2\big(\sqrt{2}+2\big)\big(\sqrt{2(e-1)}+8\big)\sqrt{T\ln T}\,.$$

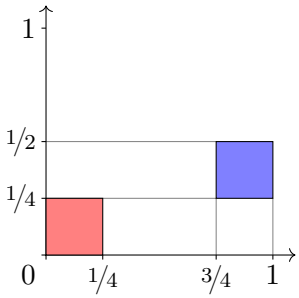
Proof. Plugging the guarantees of Proposition 2 into those of Proposition 1 and recalling that $|\mathcal{X}_\tau| \leqslant 2^{\tau-1}$ for each epoch $\tau = 2, 3, \dots$, gives the result (after straightforward computations). □

A $\sqrt{T}$ lower bound for the i.i.d. environment

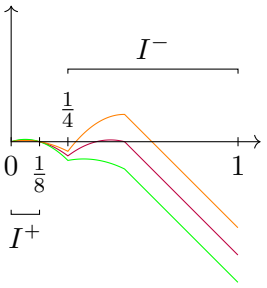
We complement the positive result of Theorem 21 with a matching lower bound of order $\sqrt{T}$, that holds even if we further assume that the underlying environment is smooth. The idea underlying our hard instance is to embed the well-known lower bound for prediction with (two) experts into our framework: we construct two smooth distributions that are "similar" but have two different optimal bids whose performance is separated. We then formally prove that no learner can identify the correct distribution without suffering less than $\sqrt{T}$ regret.

Theorem 22. *Consider the problem of repeated bidding in first-price auctions with full feedback. Suppose that $\mathcal{S}$ is the set of σ-smooth i.i.d. environments, with $\sigma \in (0, \frac{1}{19}]$. Then, for any learning algorithm α and for time horizon T, it holds*

$$R_T^{\mathcal{S}}(\alpha) \geqslant \frac{1}{2048}\sqrt{T}\,.$$



(a) Support of $f^{\pm\varepsilon}$.



(b) Utility under $\mathbb{P}^0$, $\mathbb{P}^{+\varepsilon}$, and $\mathbb{P}^{-\varepsilon}$.

Figure 3.2: Supporting figures of the Proof of Theorem 22. Figure 3.2a represents the support of the density functions $f^{\pm\varepsilon}$: the two squares Q_+ and Q_-. Figure 3.2b represents the expected utility function for three different distributions: $\mathbb{P}^0$ in violet, $\mathbb{P}^+$ in orange and $\mathbb{P}^+$ in green.

Proof. We prove the Theorem by Yao's principle: we show that there exists a distribution over stochastic σ-smooth environments such that any deterministic learning algorithm α suffers $\Omega(\sqrt{T})$ regret against it, in expectation. We do that in two steps. First, for every $\varepsilon \in (0, \frac{1}{2})$ we construct a pair of $\frac{1}{9}$-smooth distributions that are hard to discriminate for the learner. Then, we prove that, for the right choice of ε, any learner suffers the desired regret against at least one of them. For visualization, we refer to Figure 3.2.

As a tool for our construction, we introduce a baseline probability measure $\mathbb{P}^0$, such that the sequence $(V, M), (V_1, M_1), (V_2, M_2), \dots$ is $\mathbb{P}^0$-i.i.d., and (V, M) has a distribution $\mathbb{P}^0_{(V,M)}$ (for a refresher on push-forward measures, see Appendix A.15) whose density function is as follows:

$$f^0(v, m) \coloneqq 8\mathbb{I}\{(v, m) \in Q_+\} + 8\mathbb{I}\{(v, m) \in Q_-\},$$

where $Q_+ \coloneqq (0, \frac{1}{4}) \times (0, \frac{1}{4})$ and $Q_- \coloneqq (\frac{3}{4}, 1) \times (\frac{1}{4}, \frac{1}{2})$ (see Figure 3.2a). A convenient way to visualize this distribution is to draw a uniform random variable U_t in the square Q_+ and then toss an unbiased coin. If the coin yields heads, then (V_t, M_t) is equal to U_t, otherwise (V_t, M_t) coincides with U_t translated by $(\frac{3}{4}, \frac{1}{4})$. With some simple (but tedious) computation, it is possible to explicitly compute the expected utility of posting any bid $b \in [0, 1]$, when (V_t, M_t) is drawn following the distribution $\mathbb{P}^0$ (with expectation $\mathbb{E}^0$):

$$\mathbb{E}^0[\mathrm{Util}_t(b)] = \begin{cases} \frac{b}{4}(1 - 8b) & \text{if } b \in [0, \frac{1}{4}) \\ -\frac{1}{8}(16b^2 - 14b + 3) & \text{if } b \in [\frac{1}{4}, \frac{1}{2}) \\ \frac{1}{2}(1 - 2b) & \text{if } b \in [\frac{1}{2}, 1] \end{cases}$$

The function $\mathbb{E}^0[\mathrm{Util}_t(b)]$ has two global maxima in $[0, 1]$, of value $\frac{1}{128}$, attained in $\frac{1}{16}$ and $\frac{7}{16}$ (see the red line in Figure 3.2b).

For any $\varepsilon \in (0, \frac{1}{2})$, we also define two additional (perturbed) probability measures $\mathbb{P}^{\pm\varepsilon}$, such that the sequence $(V, M), (V_1, M_1), (V_2, M_2), \dots$ is $\mathbb{P}^{\pm\varepsilon}$-i.i.d. and the distribution $\mathbb{P}^{\pm\varepsilon}_{(V,M)}$ of (V, M) has density:

$$f^{\pm\varepsilon}(v, m) \coloneqq 8(1 \pm \varepsilon)\mathbb{I}\{(v, m) \in Q_+\} + 8(1 \mp \varepsilon)\mathbb{I}\{(v, m) \in Q_-\}.$$

Note, $||f^{\pm\varepsilon}||_\infty < 9$, while $||f^0||_\infty = 8$, therefore all the distributions considered in this proof are

$\frac{1}{9}$-smooth. To visualize this new perturbed distributions, recall the construction of $\mathbb{P}^0_{(V,M)}$ using the coin toss and the uniform random variable U: in this case the coin is biased and the probability of getting tail is $(1 \pm \varepsilon)/2$. It is still possible to compute explicitly the expected utility under these perturbed distributions for any bid $b \in [0, 1]$:

$$\mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(b)] = \begin{cases} \frac{b}{4}(1-8b) \pm \varepsilon\frac{b}{4}(1-8b) & \text{if } b \in [0, \frac{1}{4}) \\ -\frac{1}{8}(16b^2 - 14b + 3) \pm \frac{\varepsilon}{4}(8b^2 - 11b + 2) & \text{if } b \in [\frac{1}{4}, \frac{1}{2}) \\ \frac{1}{2}(1 - 2b \mp \frac{3}{4}\varepsilon) & \text{if } b \in [\frac{1}{2}, 1] \end{cases} \tag{3.2}$$

We refer to Figure 3.2b for visualization. The crucial property of the distributions we constructed is that the instantaneous regret of not playing in the "correct" region is $\Omega(\varepsilon)$; formally we have the following result. For the sake of readability, we postpone the proof of this Claim to Appendix B.3.

Claim 4. *There exists two disjoint intervals I_+ and I_- in $[0, 1]$ such that, for any $\varepsilon \in (0, \frac{1}{2})$ and any time t, the following inequalities hold:*

$$\max_{x\in[0,1]} \mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(x)] \geqslant \mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(b)] + \frac{1}{128}\varepsilon, \text{ for all } b \notin I_\pm$$

Since the two distributions are "ε-close"[§], any learner needs at least $\frac{1}{\varepsilon^2}$ rounds to discriminate which ones of the two distributions she is actually facing, paying each error with an instantaneous regret of $\Omega(\varepsilon)$ (Claim 4). All in all, any learner suffers a regret that is $\Omega(\varepsilon \cdot \frac{1}{\varepsilon^2} + \varepsilon T)$, which is of the desired $\Omega(\sqrt{T})$ order for the right choice of $\varepsilon \approx T^{-1/2}$.

As the last step of the proof, we formalize the above argument. Fix $\varepsilon = 1/(4\sqrt{T})$ and rename $\mathbb{P}^{+\varepsilon} = \mathbb{P}^1$ and $\mathbb{P}^{-\varepsilon} = \mathbb{P}^2$, given our choice of ε; similarly, denote with I_1 and I_2 the two intervals I_+ and I_- as in the statement of Claim 4. For each $j \in \{0, 1, 2\}$, consider the run of α against the stochastic environment which draws $(V_1, M_1), (V_2, M_2), \dots$ i.i.d. from $\mathbb{P}^j$. Let N_1 be the random variable that counts the number of times that algorithm α posts a bid in I_1. Similarly, N_2 counts the number of times that it posts a bid in I_2. For $i = 1, 2$, we have the following crucial relation between the expected value of N_i under $\mathbb{P}^i$. Note, the results hold because the two distributions are so similar that the deterministic algorithm α bids in the wrong region a costant fraction of the time steps. For the formal proof of we refer the reader to Appendix B.3.

Claim 5. *The following inequality hold:*

$$\frac{1}{2}\sum_{i=1,2} \mathbb{E}^i[N_i] \leqslant \frac{3}{4}T.$$

We finally have all the ingredients to conclude the proof. Consider an environment that selects uniformly at random either $\mathbb{P}^1$ or $\mathbb{P}^2$ and then draws the (V_t, M_t) i.i.d. following it. We prove that the algorithm α suffers linear regret against this randomized environment and, by a simple averaging argument, against at least one of them. Specifically, if $b_i^\star$ is the optimal bid in the scenario

[§]In Appendix B.3 we formally prove that their total variation is at most $\Theta(\varepsilon)$.

determined by $\mathbb{P}^i$, for $i \in \{1,2\}$, we have

$$\begin{aligned} R_T(\alpha) &\geqslant \frac{1}{2}\sum_{i=1,2}\mathbb{E}^i\left[\sum_{t=1}^{T}\mathrm{Util}_t(b_i^\star) - \sum_{t=1}^{T}\mathrm{Util}_t(B_t)\right] \\ &\geqslant \frac{1}{1024\sqrt{T}}\sum_{i=1,2}\mathbb{E}^i\left[T - N_i\right] && \text{(By Claim 4 and choice of } \varepsilon) \\ &\geqslant \frac{1}{512\sqrt{T}}\left(T - \frac{3}{4}T\right) && \text{(By Claim 5)} \\ &= \frac{\sqrt{T}}{2048}. \end{aligned}$$

□

3.4 The Adversarial Setting

In this section we complete the perspective on repeated bidding in first-price auction by investigating the adversarial model. In particular, we consider two models: the standard one, where the sequence $(V_1, M_1), (V_2, M_2), \dots$ is chosen up front in a deterministic oblivious way, and the smooth environment, where the sequence $(V_1, M_1), (V_2, M_2), \dots$ is any σ-smooth stochastic process. In Section 3.4.1 we construct an algorithm achieving $T^{2/3}$ regret in the bandit feedback model under the smoothness assumption; this result, together with the lower bound of the same order for the semi-transparent feedback (Theorem 20) settles the problem for these two feedback regimes. Then, in Section 3.4.2 we provide another upper bound, namely an algorithm achieving $\sqrt{T}$ regret in the transparent feedback model under the smoothness assumption; this result, together with the lower bound of the same order for the semi-transparent feedback (Theorem 22) settles the problem for these two feedback regimes. Finally, in Section 3.4.3 we provide a lower bound proving that the non-smooth adversarial environment is too hard to learn, even when the learner has access to full feedback.

3.4.1 Bandit Feedback against the Smooth Environment

The smoothness assumption regularizes the objective function. In particular, if (V_t, M_t) is smooth, then the corresponding expected utility is Lipschitz.

Lemma 10 (Lipschitzness)**.** *Let (V_t, M_t) be a σ-smooth random variable in $[0,1]$. Then the induced expected utility function $\mathbb{E}\left[\mathrm{Util}_t(\cdot)\right]$ is $2/\sigma$-Lipschitz in $[0,1]$:*

$$\left|\mathbb{E}\left[\mathrm{Util}_t(y) - \mathrm{Util}_t(x)\right]\right| \leqslant \frac{2}{\sigma}|y-x|, \quad \forall x, y \in [0,1]. \tag{3.3}$$

Proof. Let $x > y$ be any two bids in $[0,1]$, we have the following:

$$\begin{aligned} \left|\mathbb{E}\left[\mathrm{Util}_t(x) - \mathrm{Util}_t(y)\right]\right| &= \left|\mathbb{E}\left[(V_t - x)\mathbb{I}\{M_t \leqslant x\} - (V_t - y)\mathbb{I}\{M_t \leqslant y\}\right]\right| \\ &= \left|\mathbb{E}\left[(V_t - x)\mathbb{I}\{y < M_t \leqslant x\} + (y - x)\mathbb{I}\{M_t \leqslant y\}\right]\right| \\ &\leqslant \mathbb{P}\big[M_t \in [x,y]\big] + (x-y) \leqslant \frac{2}{\sigma}(x-y). \end{aligned}$$

□

As an interesting fact, note that we only need the marginal distribution of M_t to be σ-smooth for the previous lemma to hold.

This Lipshitzness property has the immediate corollary that any fine enough discretization of $[0,1]$ contains a bid whose utility is close the the optimal one.

Lemma 11 (Discretization Lemma)**.** *Let $\mathcal{X}$ be any finite grid of bids in $[0,1]$, and let $\delta(\mathcal{X})$ be the largest distance of a point in $[0,1]$ to $\mathcal{X}$ (i.e., $\delta(\mathcal{X}) = \max_{p\in[0,1]} \min_{x\in\mathcal{X}} |p-x|$), then if each pair of random variables $(V_1, M_1), \dots, (V_T, M_T)$ is σ-smooth, we have the following:*

$$\max_{b\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(b)\right] - \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x)\right] \leqslant 2\frac{\delta(\mathcal{X})}{\sigma} T \,.$$

Proof. Fix any such sequence and let $b^\star$ be the corresponding best fixed bid in hindsight. If $b^\star$ is in $\mathcal{X}$ there is nothing to prove, otherwise these exists $x^\star \in \mathcal{X}$ such that $|b^\star - x^\star| \leqslant \delta(\mathcal{X})$ (by definition of $\delta(\mathcal{X})$). We have the following:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(b^\star)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x^\star)\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{Util}_t(b^\star) - \mathrm{Util}_t(x^\star)\right] \\
&\leqslant \sum_{t=1}^{T} \frac{2}{\sigma} |b^\star - x^\star| \qquad \text{(By Lipschitzness, Lemma 10)} \\
&\leqslant 2\frac{\delta(\mathcal{X})}{\sigma} T
\end{aligned}$$

□

We can combine in a natural way the above discretization Lemma with any (optimal) bandit algorithm to obtain the desired bound on the regret. For details we refer to the pseudocode of Discretized Bandit.

Discretized Bandit - Bandit Feedback

1: **input:** Time horizon T, bandit algorithm $\widetilde{\alpha}$ for gains in $[-1,1]$, and grid of K bids $\mathcal{X}$
2: Initialize $\widetilde{\alpha}$ on K actions, one for each bid $x \in \mathcal{X}$, and time horizon T
3: **for** time $t = 1, 2, \dots, T$ **do**
4: Receive from $\widetilde{\alpha}$ the bid $B_t \in \mathcal{X}$
5: Post bid B_t and observe bandit feedback Z_t
6: Reconstruct $\mathrm{Util}_t(B_t)$ from Z_t and feed it to $\widetilde{\alpha}$ as the reward associated to the arm B_t

Theorem 23. *Consider the problem of repeated bidding in first-price auctions with bandit feedback. Suppose that $\mathcal{S}$ is the set of adversarial σ-smooth environments. Then there exists a learning algorithm α such that*

$$R_T^{\mathcal{S}}(\alpha) \leqslant \frac{27}{\sigma} T^{2/3} \,.$$

Proof. We prove that algorithm Discretized Bandit with the right choice of learning algorithm $\widetilde{\alpha}$ and grid of bids $\mathcal{X}$ achieves the desired bound on the regret. As learning algorithm $\widetilde{\alpha}$ we use (a rescaled version of) the Poly INF algorithm [16]: since Poly INF is designed to run with gains in $[0,1]$ while the utilities we observe are in $[-1,1]$, we first apply the reward transformation $x \mapsto \frac{x+1}{2}$ to the

observed utilities. This transformation will cost a multiplicative factor of 2 in the regret guarantees of Poly INF.

The analysis builds on the discretization result in Lemma 11, by choosing as $\mathcal{X}$ the uniform grid of $\lceil T^{2/3} \rceil + 1$ equally spaced bids on $[0,1]$ (note, $\delta(\mathcal{X})$ becomes $T^{-1/3}$). Fix any σ-smooth environment β, we have the following:

$$\begin{aligned}\max_{b\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(b)\right] &\leqslant \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x)\right] + \frac{4}{\sigma}T^{2/3} && \text{(Lemma 11)}\\ &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(B_t)\right] + \frac{4}{\sigma}T^{2/3} + 23T^{2/3} \leqslant \frac{27}{\sigma}T^{2/3},\end{aligned}$$

where the second inequality follows from the guarantees of (the rescaled version of) Poly INF (Theorem 11 of Audibert and Bubeck [16]). □

3.4.2 Transparent Feedback against the Smooth Environment

For transparent feedback we combine two tools we have already used: the discretization Lemma (11) and the algorithm Exp3.FPA for learning with transparent feedback on a finite grid. Note: using any other $\sqrt{KT}$ black box learning algorithm (like in the previous section for bandits) would yield a suboptimal regret bound of $T^{2/3}$.

Theorem 24. *Consider the problem of repeated bidding in first-price auctions with transparent feedback. Suppose that $\mathcal{S}$ is the set of adversarial σ-smooth environments. Then there exists a learning algorithm α such that*

$$R_T^{\mathcal{S}}(\alpha) \leqslant 4\left(\frac{1}{\sigma} + \sqrt{\ln T}\right)\sqrt{T}\,.$$

Proof. Consider algorithm Exp3.FPA on the uniform grid $\mathcal{X}$ of $\lceil \sqrt{T} \rceil + 1$ bids, with $\delta(\mathcal{X}) \leqslant \sqrt{T}$. For any fixed σ-smooth environment β, we have the following:

$$\begin{aligned}\max_{b\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(b)\right] &\leqslant \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x)\right] + \frac{4}{\sigma}\sqrt{T} && \text{(Lemma 11)}\\ &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(B_t)\right] + 2\sqrt{(e-1)T\ln T} + \frac{4}{\sigma}\sqrt{T}\\ &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(B_t)\right] + 4\left(\frac{1}{\sigma} + \sqrt{\ln T}\right)\sqrt{T},\end{aligned}$$

where the second inequality follows from the guarantees of Proposition 2. □

3.4.3 The (Non-Smooth) Adversarial Model is Hopeless

In the previous sections, we have been able to provide positive results under one of two conditions: either the environment is stochastic and the learner has at least the semi-transparent feedback (Theorem 18 says that bandit feedback is not enough) or the environment uses smooth distributions. Both these settings allow the learner to compute efficiently a discrete class of representative bids

where the learning may happen. In this final section, we formally complete the proof of the fact that learning is impossible if any of these assumptions are dropped. Specifically, the standard adversarial environment that generates arbitrarily the sequence without any smoothness constraint is too strong. In particular, we construct a randomized sequence $(V_1, M_1), (V_2, M_2), \dots$ that induces any learner to suffer at least linear regret. This construction shares some similarities with the lower bound construction in Theorem 18, the main difference being that the best bid $b^\star$ is randomized and hidden in such a way that even a learner having access to full feedback cannot pin-point it.

Theorem 25. *Consider the problem of repeated bidding in first-price auctions with full feedback. Suppose that $\mathcal{S}$ is the set of adversarial environments. Then, for any learning algorithm α and any time horizon T, it holds*

$$R_T^{\mathcal{S}}(\alpha) \geqslant \frac{T}{24} .$$

Proof. We prove the result via Yao's principle, showing that there exists a randomized environment β such that any deterministic learning algorithm suffers $\frac{1}{24} \cdot T$ regret against it.

The random sequence posted by β is based on two randomized auxiliary sequences $L_1, L_2, \dots$ and $U_1, U_2, \dots$ defined as follows. They are initiated to $L_0 = \frac{1}{2}$, $U_0 = \frac{2}{3}$. Then, they evolve recursively following the rule

$$\begin{cases} L_t = L_{t-1} + \frac{2}{3}\Delta_{t-1} \text{ and } U_t = U_{t-1}, \text{ with probability } \frac{1}{2}, \\ U_t = U_{t-1} - \frac{2}{3}\Delta_{t-1} \text{ and } L_t = L_{t-1}, \text{ with probability } \frac{1}{2}, \end{cases}$$

where $\Delta_{t-1} = U_{t-1} - L_{t-1}$.

For each realized sequence of the (L_t, U_t) pairs, the actual sequence of the (M_t, V_t) selected by β is constructed as follows. At each time step t, the environment selects $(M_t, V_t) = (L_t, 1)$ or $(U_t, 0)$, uniformly at random. Note, the distribution is characterized by two levels of independent randomness: the auxiliary sequence of shrinking intervals and the choice between $(L_t, 1)$ and $(U_t, 0)$.

We move our attention to the expected performance of the best fixed price in hindsight. For each realization of the random auxiliary sequence, there exists a bid B^* such that (i) it wins all the auctions (V_t, M_t) of the form $(L_t, 1)$ (which we may call "good auctions" because they bring positive utility when won) and (ii) it loses all the auctions (V_t, M_t) of the form $(U_t, 0)$ (which we may call "bad auctions" because they bring negative utility). Thus its expected utility at each time step is at least $\frac{1}{6}$: with probability $\frac{1}{2}$ the environment selects a good auction, which induces an utility of $(1 - L_t) \geqslant \frac{1}{3}$. All in all, the optimal bid achieves an expected utility of at least $\frac{T}{6}$.

Consider now the performance of any deterministic algorithm α: for any fixed time $t > 1$ and possible realization of the past observations, the learner posts some deterministic bid B_t. If $B_t < L_{t-1}$, then it gets 0 utility, so we only consider the following cases:

- If $B_t \in [L_{t-1}, L_{t-1} + \frac{1}{3}\Delta_{t-1})$, then the bidder gets the item with probability $\frac{1}{4}$ ($L_t = L_{t-1}$, V_t is set to 1 and $M_t = L_t$) with an expected utility of $\frac{1}{4}(1 - L_t) \leqslant \frac{1}{8}$.
- If $B_t \in [L_{t-1} + \frac{1}{3}\Delta_{t-1}, L_{t-1} + \frac{2}{3}\Delta_{t-1})$, the bidder gets the item with probability $\frac{1}{2}$ (when $L_t = L_{t-1}$ and $U_t = U_{t-1} - \frac{2}{3}\Delta_{t-1}$) for an expected utility of $\frac{1}{4}(1 - L_{t-1}) - \frac{1}{4}(L_{t-1} + \frac{1}{3}\Delta_{t-1}) \leqslant 0 \leqslant \frac{1}{8}$
- If $B_t \in [L_{t-1} + \frac{2}{3}\Delta_{t-1}, U_{t-1})$ the bidder gets the item with probability $\frac{3}{4}$ (when $L_t = L_{t-1}$ and when $U_t = U_{t-1}$, $V_t = 1$ and $M_t = L_t$) for an expected utility of $\frac{1}{4}(1 - L_{t-1}) - \frac{1}{4}(L_{t-1} + \frac{1}{3}\Delta_{t-1}) + \frac{1}{4}(1 - L_{t-1} - \frac{2}{3}\Delta_{t-1}) \leqslant 0 + \frac{1}{8} = \frac{1}{8}$

- If $B_t \geqslant U_{t-1}$ then the bidder always gets the item, with an expected utility smaller than 0 (which is in turn smaller than $\frac{1}{8}$).

All in all, we have that the expected utility of any deterministic algorithm is at most $\frac{1}{8}T$. If we compare this quantity with the lower bound on the expected utility of the best bid in hindsight we get the desired result:

$$\mathbb{E}\left[R_T(\alpha, \beta)\right] \geqslant \frac{T}{6} - \frac{T}{8} = \frac{T}{24} \, .$$

□

A final observation: the crucial ingredient in the proof is the possibility of constructing this elaborate auxiliary sequence. To this end, we only needed the non-smoothness of M_t, while we may have chosen the valuations V_t to be smooth (and even i.i.d.), say uniformly in $[0, \frac{1}{4}]$ for the bad auctions and in $[\frac{3}{4}, 1]$ for the good ones.

3.5 Conclusions

Motivated by the recent shift from second to first-price auctions in online advertising market, in this chapter we offered a comprehensive analysis of the online learning problem of repeated bidding in first-price auction under the realistic assumption that the bidder does not know her valuation before bidding. We have characterized the minimax regret achievable for different levels of transparency in the auction format and for different data generation models, considering both the stochastic i.i.d. and the standard adversarial model, with a focus also on smoothness. Although all our regret rates are tight in their dependence on the time horizon T, a natural open problem consists in studying their minimax dependence in the smoothness parameter σ.

This work belongs to the long line of research that studies economic problems from the online learning perspective; an intriguing open problem there is to offer a unified framework to characterize in a satisfying way all these games with partial feedback, similar to what has been done for partial monitoring and feedback graph.

Chapter 4

Adaptive Maximization of Social Welfare

4.1 Introduction

Consider a policymaker who aims to maximize social welfare, defined as a weighted sum of utility across individuals. The policymaker can choose a policy parameter such as a sales tax rate, an unemployment benefit level, a health-insurance copay rate, etc. The policymaker does *not* directly observe the welfare resulting from her policy choices. She does, however, observe behavioral outcomes such as the consumption of taxed goods, labor market participation, or health care expenditures. She can revise her policy choices over time in light of observed outcomes.

How should such a policymaker act? To address this question, we bring together insights from welfare economics (in particular optimal taxation [26, 65, 140, 154, 159]) with insights from machine learning (in particular online learning and multi-armed bandits [46, 123, 169]).

In our baseline model, individuals arrive sequentially and make a single binary decision. In each period, the policymaker chooses a tax rate that applies to this binary decision. Then, she observes the individual's response. We remark that the policymaker never observes the individual's private utility. Social welfare is given by a weighted sum of private utility and public revenue. Later, we extend our model to nonlinear income taxation, where welfare weights vary as a function of individual earnings capacity, and sketch an extension to commodity taxation, where individual decisions involve a continuous consumption vector.

Our goal is to give guidance to the policymaker. We propose algorithms to maximize cumulative social welfare, and we provide guarantees for the performance of these algorithms. In doing so, we also show that welfare maximization is a harder learning problem than reward maximization in the multi-armed bandit setting. Private utility in our baseline model is equal to consumer surplus, which is given by the integral of the demand function. To learn this integral, we need to learn the demand for counterfactual, suboptimal tax rates. This drives the difficulty of the learning problem.

A lower bound on regret Our main theorems provide lower and upper bounds on the *regret*. The regret is defined as the difference in cumulative welfare between the *chosen* sequence of policies and the *best* possible constant policy. We consider both stochastic and adversarial regret. The former assumes that preference parameters are drawn i.i.d. from some distribution, whereas the latter allows for arbitrary sequences of preference parameters.

We first prove a stochastic (and thus also adversarial) lower bound on the regret, for any

possible algorithm. Our proof of this lower bound constructs a family of possible distributions for preferences. This family is such that two candidate policies are potentially optimal. The difference in welfare between these two policies depends on the integral of the demand function over intermediate policy values. To learn which of these two candidate policies is optimal, we need to learn behavioral responses for intermediate strictly suboptimal policies. Because of the need to probe these suboptimal policies sufficiently often, we obtain a lower bound on regret which grows at a rate of $T^{2/3}$, even if we restrict our attention to settings with finite, known support for preference parameters and policies. We remark that this rate is worse than the worst-case rate for bandits of $T^{1/2}$.

A matching upper bound on adversarial regret for modified Exp3 We next propose an algorithm for the adaptive maximization of social welfare. Our algorithm is a modification of the well-known Exp3 algorithm [18]. Exp3 is based on an unbiased estimate of cumulative welfare for each policy. The probability of choosing a given policy is proportional to the exponential of this estimate of cumulative welfare, times some rate parameter. Relative to Exp3, we require two modifications for our setting. First, we need to discretize the continuous policy space. Second, and more interestingly, we need additional exploration of counterfactual policies, including some policies that are clearly sub-optimal, in order to learn welfare for the policies that are contenders for the optimum. This need for additional exploration again arises because of the dependence of welfare on the integral of the demand over counterfactual policy choices. For our modified Exp3 algorithm, we prove an adversarial (and thus also stochastic) upper bound on the regret. We show that, for an appropriate choice of tuning parameters, the worst-case cumulative regret over all possible sequences of preference parameters grows at a rate of $T^{2/3}$, up to a logarithmic term. The algorithm thus achieves the best possible rate.*

Improved stochastic bounds for concave social welfare The proof of our lower bound on regret is based on the construction of a distribution of preferences that delivers a non-concave social welfare function. If we restrict attention to the stochastic setting, where preferences are i.i.d. over time, and if we assume that social welfare is concave, then we can improve upon this bound on regret. In the stochastic case, assuming the concavity of the expected utility function, we prove a $\Omega(\sqrt{T})$ lower bound on the regret. We then propose a dyadic search algorithm achieving this rate, up to logarithmic terms. This dyadic search algorithm maintains an "active interval", containing the optimal policy with high probability, which is narrowed down over time. Only policies within the active interval are sampled.

Extensions to non-linear income taxation and to commodity taxation Our discussion up to this point focuses on the somewhat stylized case of an optimal tax problem, where individual actions are binary, and the policy imposes a tax on this binary action. Our arguments generalize, however, to more complicated and practically relevant settings. This includes optimal nonlinear income taxation (see Section 4.5), as in Mirrlees [140] and Saez [159]. For nonlinear income taxation,

*Since stochastic regret (averaged over sequences of willingness to pay) is always less or equal to adversarial regret (for the worst-case sequence), the stochastic lower bound immediately implies a corresponding adversarial lower bound, and the adversarial upper bound implies a corresponding stochastic upper bound. Since the rates for our stochastic lower and adversarial upper bound coincide, up to a logarithmic term, we have a complete characterization of learning rates for the welfare maximization problem.

different tax rates apply at different income levels, and welfare weights depend on individual earnings capacity. In Section 4.5, we discuss an extension of our tempered Exp3 algorithm to nonlinear income taxation, and characterize its regret.

We propose a different generalization in Appendix C.1, where commodity taxation for a bundle of goods is discussed, as in Ramsey [154]. For commodity taxation, different tax rates apply to different goods, and consumption decisions are continuous vectors. However, we still do not know whether our arguments generalize beyond the one-dimensional case for this problem (see the discussion in Appendix C.1), and we leave its investigation open for future research.

Roadmap The rest of this chapter proceeds as follows. We conclude this introduction with a discussion of some related work and relevant references. Section 4.2 introduces our setup, formally defines the adversarial and stochastic settings, and compares our setup to related learning problems. Section 4.3 provides lower and upper bounds on regret in the adversarial and stochastic settings. Section 4.4 restricts attention to the stochastic setting with concave social welfare, and provides improved regret bounds for this setting. Section 4.5 discusses an extension of our baseline model to non-linear income taxation. Appendix C.1 sketches another extension of our baseline model to commodity taxation. All proofs can be found in Appendix C.2.

4.1.1 Background and Literature

To put our work in context, it is useful to contrast our framework with the standard approach in public finance and optimal tax theory, and with the frameworks considered in machine learning and the multi-armed bandit literature.

Optimal taxation Optimal tax theory, and optimal policy theory more generally, is concerned with the maximization of social welfare, where social welfare is understood as a (weighted) sum of subjective utility across individuals [26, 65, 113, 140, 154, 159]. A key tradeoff in such models is between, first, *redistribution* to those with higher welfare weights, and second, the efficiency cost of behavioral responses to tax increases. Such *behavioral responses* might reduce the tax base.

Optimal tax problems are defined by normative parameters (such as welfare weights for different individuals), as well as empirical parameters (such as the elasticity of the tax base with respect to tax rates). The typical approach in public finance uses historical or experimental variation to estimate the relevant empirical parameters (causal effects, elasticities). These estimated parameters are then plugged into formulas for optimal policy choice, which are derived from theoretical models. The implied optimal policies are finally implemented, without further experimental variation.

Multi-armed bandits The standard approach of public finance, which separates elasticity estimation from policy choice, contrasts with the adaptive approach that characterizes decision-making in many branches of AI, including online learning, multi-armed bandits, and reinforcement learning. In particular, multi-armed bandit algorithms trade off *exploration* and *exploitation* over time to maximize a stream of rewards [46, 123, 169]. Exploration here refers to the acquisition of information for better future policy decisions, while exploitation refers to the use of currently available information for optimal policy decisions at the present moment.

Online Taxation Protocol

for time $i = 1, 2, \ldots$ **do**
A new individual arrives with (hidden) valuation $v_i \in [0, 1]$
The learner posts a tax rate $x_i \in [0, 1]$
The learner receives a (hidden) reward $U_i(x_i)$, where $U_i(x) \coloneqq x \cdot \mathbb{I}\{x \leqslant v_i\} + \lambda \cdot \max(v_i - x, 0)$
The learner observes feedback $y_i \coloneqq \mathbb{I}\{x_i \leqslant v_i\} \in \{0, 1\}$

We remark that bandit algorithms (and similarly, adaptive experimental designs for informing policy choice, as in [114, 158]) are not directly applicable to social welfare maximization problems, such as those of optimal tax theory. The reason is that bandit algorithms maximize a stream of *observed* rewards. By contrast, social welfare as conceived in welfare economics is based on *unobserved* subjective utility.

Bandit approaches for economic problems Even though we already discussed this topic in previous chapters, we briefly recall the relevant information for the sake of the reader.

Bandit-type approaches have been applied to a number of other economic and financial scenarios in the literature where rewards *are* observable. These include dynamic pricing [117] (see also the survey [76] and the related work Section 2.1.3), second-price auctions [50, 52, 182], first-price auctions [2, 90, 91, 104, 105, 118]—see also Chapter 3 and the related work Section 3.1.3 therein—and combinatorial auctions [72]. Bandit-type approaches have also been applied to settings where rewards are *not* directly observable, including bilateral trade (that we already discussed in Chapter 2) and, e.g., the newsvendor problem [127].

4.2 Setup

At each time $i = 1, 2, \ldots, T$, one individual arrives who is characterized by an unknown willingness to pay $v_i \in [0, 1]$. This individual is exposed to a tax rate x_i, and makes a binary decision $y_i = \mathbb{I}\{x_i \leqslant v_i\}$. The implied public revenue is $x_i \cdot y_i$. The implied private welfare is $\max(v_i - x_i, 0)$. We define social welfare as a weighted sum of public revenue and private welfare, with a weight $\lambda \in (0, 1)$ for the latter, fixed by the policymaker depending on her preferences for redistribution. Social welfare for time period i is therefore given by $U_i(x_i)$, where x_i is the policy chosen by the learner at time i, and for any $x \in [0, 1]$, we have defined

$$U_i(x) \coloneqq \underbrace{x \cdot \mathbb{I}\{x \leqslant v_i\}}_{\text{Public revenue}} \quad + \quad \lambda \cdot \underbrace{\max(v_i - x, 0)}_{\text{Private welfare}}. \tag{4.1}$$

After period i, the learner observes y_i and nothing else. In particular, the learner does *not* observe welfare $U_i(x_i)$. See also the Learning Protocol.

We can rewrite social welfare $U_i(x)$ as follows. Denote the individual demand function by $G_i(x) = \mathbb{I}\{v_i \geqslant x\}$, so that $y_i = G_i(x_i)$. Then, private welfare can be written as $\max(v_i - x, 0) = \int_x^1 G_i(x')\mathrm{d}x'$. That is, private welfare can be obtained by integrating the demand function.[†] This

[†]This reflects the absence of income effects in our model, which implies that private utility, consumer surplus, compensating variation, and equivalent variation all coincide.

representation of private welfare implies

$$U_i(x) = \underbrace{x \cdot G_i(x)}_{\text{Public revenue}} + \lambda \cdot \underbrace{\int_x^1 G_i(x')\mathrm{d}x'}_{\text{Private welfare}} . \tag{4.2}$$

We consider algorithms for the choice of x_i which might depend on the observable history $(x_j, y_j)_{j=1}^{i-1}$, as well as possibly a randomization device.

Notation For the *adversarial* setting, we will consider cumulative demand and welfare, denoted by boldface letters, summing across $j = 1, \ldots, i$. Specifically,

$$\boldsymbol{G}_i(x) \coloneqq \sum_{j \leqslant i} G_i(x), \qquad \boldsymbol{U}_i(x) \coloneqq \sum_{j \leqslant i} U_i(x), \qquad \boldsymbol{U}_i \coloneqq \sum_{j \leqslant i} U_j(x_j).$$

$\boldsymbol{G}_i(x)$ and $\boldsymbol{U}_i(x)$ are cumulative demand and welfare for a counterfactual, fixed policy x. $\boldsymbol{U}_i$, without an argument, is the cumulative welfare for the policies x_j actually chosen.

For the *stochastic* setting (where we recall that the sequence $(v_i)_{i\in\mathbb{N}}$ is independent and identically distributed), we will analogously consider expected demand and expected welfare, denoted by blackboard bold letters. The expectation is taken across some stationary distribution μ of v_i, where v_i is statistically independent of x_i, and of v_j for $j \neq i$. Specifically, for any x, we define the expected demand and the expected utility as

$$\mathbb{G}(x) \coloneqq \mathbb{E}[G_i(x)], \qquad \mathbb{U}(x) \coloneqq \mathbb{E}[U_i(x)] ,$$

and we note explicitly that this definition is independent of the choice of the time i.

4.2.1 Regret

The adversarial case Following the literature, we consider regret for both the adversarial and the stochastic setting. In the adversarial setting, we allow for arbitrary sequences of willingness to pay, $\{v_i\}_{i=1}^T$. We compare the expected performance of any given algorithm α for choosing $\{x_i\}_{i=1}^T$ to the performance of the best possible constant policy x. This comparison yields cumulative expected regret, which is given by

$$R_T\big(\alpha, \{v_i\}_{i=1}^T\big) \coloneqq \sup_{x\in[0,1]} \mathbb{E}\left[\boldsymbol{U}_T(x) - \boldsymbol{U}_T\right]. \tag{4.3}$$

The expectation in this expression is taken over any possible randomness in the tax rates x_i chosen by the algorithm; there is no other source of randomness.

The stochastic case We also consider the stochastic setting. In this setting, we add structure by assuming that the v_i are i.i.d. draws from some distribution μ on $[0, 1]$, with implied demand function $\mathbb{G}(x) = \mathbb{P}[v_i \geqslant x]$. This demand function is identified by the regression

$$\mathbb{G}(x) = \mathbb{E}[y_i | x_i = x].$$

The expectation in this expression is taken over the distribution of v_i, which is presumed to be statistically independent of the tax rate x_i. Expected welfare for this distribution of v_i is given by

$$\mathbb{U}(x) = x \cdot \mathbb{G}(x) + \lambda \cdot \int_x^1 \mathbb{G}(x')\mathrm{d}x'.$$

For an algorithm α, cumulative expected regret in the stochastic case equals

$$R_T(\alpha, \mathbb{G}) \coloneqq \sup_{x\in[0,1]} \mathbb{E}\left[\boldsymbol{U}_T(x) - \boldsymbol{U}_T\right] = T \cdot \sup_{x\in[0,1]} \mathbb{U}(x) - \mathbb{E}\left[\sum_{i\leqslant T} \mathbb{U}(x_i)\right]. \tag{4.4}$$

The expectation in this expression is taken over any possible randomness in the tax rates x_i, and over the i.i.d. draws of v_i.

Lower and upper bounds Below, we will derive lower and upper bounds for adversarial and stochastic regret. A lower bound on adversarial (resp., stochastic) regret requires that, for any algorithm, there exists some sequence $\{v_i\}_{i=1}^T$ (resp., some stationary distribution μ) over which the algorithm has to suffer at least a certain amount of regret. A lower bound on stochastic regret immediately implies a lower bound on adversarial regret, since the supremum over sequences $\{v_i\}_{i=1}^T$ exceeds the expectation over such sequences, generated from any distribution μ.

An adversarial upper bound on regret has to hold for a given algorithm and any sequence $\{v_i\}_{i=1}^T$. Such an adversarial upper bound again immediately implies a stochastic upper bound on regret, by the same argument as above. When an adversarial upper bound coincides with a stochastic lower bound, in terms of rates of regret, it follows that the proposed algorithm is rate efficient, for both stochastic and adversarial regret.

4.2.2 Comparison to Related Learning Problems

Before proceeding with our analysis of regret, we take a step back, and compare our learning problem to two related problems that have received some attention in the literature. The first of these is the *dynamic pricing* problem; see for instance [117]. This problem is equivalent to our setting when we set $\lambda = 0$, interpret x as a price, and U_i^{DP} as monopolist profits (neglecting production costs):

$$U_i^{\mathrm{DP}}(x) \coloneqq x_i \cdot \mathbb{I}\{x_i \leqslant v_i\} = \underbrace{x \cdot G_i(x)}_{\text{Monopolist revenue}}. \tag{4.5}$$

As in our adaptive taxation setting, the feedback received at the end of period i is

$$y_i = G_i(x_i) = \mathbb{I}\{x_i \leqslant v_i\}.$$

The other related problem is price setting for *bilateral trade*, which was the topic of Chapter 2. We recall that, in this problem, welfare $U_i^{\mathrm{BT}}(x)$ is given by the sum of seller and buyer welfare. Trade happens if and only if both sides agree to transact at the proposed price. Buyer willingness to

pay is given by v_i^b, while the seller is willing to trade at prices above v_i^s.

$$\begin{aligned} U_i^{\mathrm{BT}}(x) &\coloneqq \mathbb{I}\left\{v_i^b \geqslant x\right\} \cdot \max(x - v_i^s, 0) &+ \mathbb{I}\left\{v_i^s \leqslant x\right\} \cdot \max(v_i^b - x, 0) \\ &= G_i^b(x) \cdot \underbrace{\int_0^x G_i^s(x')\mathrm{d}x'}_{\text{Seller welfare}} &+ G_i^s(x) \cdot \underbrace{\int_x^1 G_i^b(x')\mathrm{d}x'}_{\text{Buyer welfare}}. \end{aligned} \tag{4.6}$$

Realistic feedback for bilateral trade is a little richer with respect to the one available in the taxation problem. We observe whether the buyer and the seller accepted the posted price,

$$y_i^b \coloneqq G_i^b(x_i) \coloneqq \mathbb{I}\{x_i \leqslant v_i^b\} \qquad \text{and} \qquad y_i^s \coloneqq G_i^s(x_i) \coloneqq \mathbb{I}\{x_i \geqslant v_i^s\}.$$

Lipschitzness and information requirements The difficulty of the learning problem in each of these models critically depends on (i) the Lipschitz properties of the welfare function, and (ii) the information required to evaluate welfare at a point. We say that a generic welfare function $W : [0,1] \to \mathbb{R}$ is one-sided Lipschitz if $W(x+\varepsilon) \leqslant W(x) + \varepsilon$ for all $0 \leqslant x \leqslant 1$ and all $0 \leqslant \varepsilon \leqslant 1 - x$. We say that learning $W(\cdot)$ requires only pointwise information if $W(x)$ is a function of $G(x)$, and does not depend on $G(\cdot)$ otherwise. One-sided Lipschitzness allows us to bound the approximation error of a learning algorithm operating on a finite subset of the set of policies. Pointwise information allows us to avoid exploring policies that are clearly suboptimal, when we aim to learn the optimal policy.

Now, it can be easily seen that the following holds:

1. For *dynamic pricing*, welfare $U_i^{\mathrm{DP}}(x)$ is one-sided Lipschitz and only depends on $G_i(x)$ pointwise.

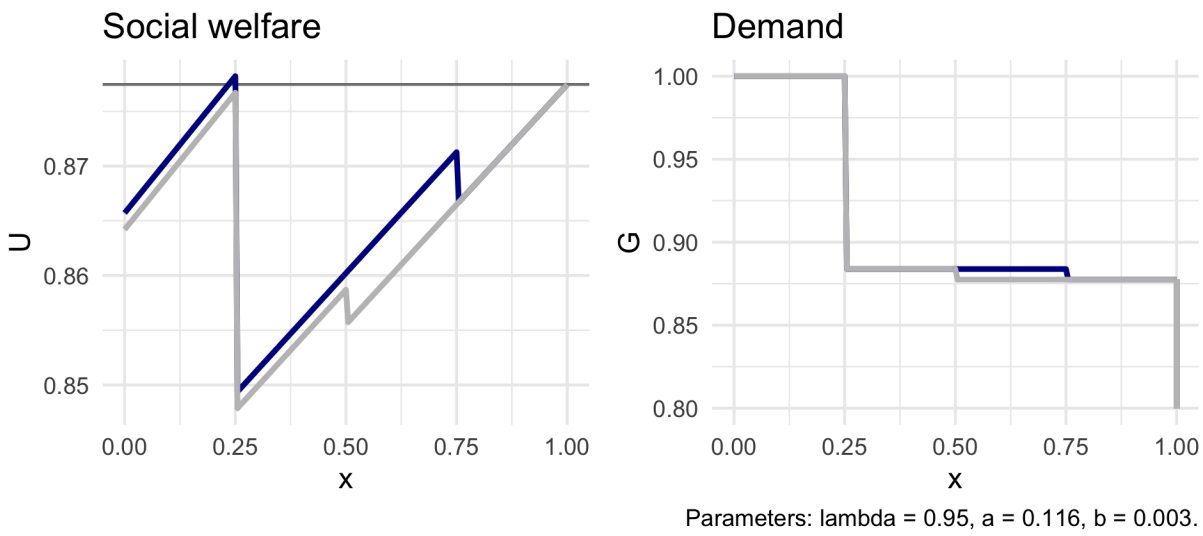
2. For *optimal taxation*, welfare $U_i(x)$ is one-sided Lipschitz and depends on both $G_i(x)$ at the given x (pointwise), and on an integral of $G_i(x')$ for a range of values of x' (non-pointwise).

3. For *bilateral trade*, welfare $U_i^{\mathrm{BT}}(x)$ is not one-sided Lipschitz and depends on both $G_i^b(x)$ and $G_i^s(x)$ (pointwise), as well as the integrals of $G_i^b(x')$ and $G_i^s(x')$ (non-pointwise).

These properties suggest a ranking in terms of the difficulty of the corresponding learning problems, and in particular in terms of the rates of divergence of cumulative regret: the information requirements of optimal taxation are stronger than those of dynamic pricing, but its continuity properties are more favorable than those of bilateral trade.

Comparison with multi-armed bandits We may also compare these problems to conventional multi-armed bandits. It is worth emphasizing that there are two distinct reasons for the slower regret rates with respect to multi-armed bandits. First, the continuous support of x, as opposed to a finite number of arms, which alone is already enough to slow down convergence with respect to the $\sqrt{T}$ bandit rate. Second, the requirement of additional exploration of clearly sub-optimal policies to estimate the reward of other more promising policies. This happens both in bilateral trade and in optimal taxation but is completely missing in the dynamic price problem. Even more, this phenomenon shows up even if we restrict our attention to a discrete set of feasible policies x. In fact, while then dynamic pricing reduces to a multi-armed bandit problem (with a minimax regret

Figure 4.1: Construction for proving the lower bound on regret



Notes: This figure illustrates our construction for proving the lower bound on regret. The relative social welfare of policies 1 and .25 depends on the sign of ε. The dark line corresponds to $\varepsilon = -1$, the bright line to $\varepsilon = 1$. In order to distinguish between these two, we must learn demand in the intermediate interval $[.5, .75]$.

rate of $\sqrt{T}$), the optimal tax problem still has a regret rate of $T^{2/3}$ even if we restrict our attention to the case of finite known support for v and x, as shown by the proof of Theorem 26 below.

4.3 Stochastic and Adversarial Regret Bounds

We now turn to our main theoretical results, lower and upper bounds on stochastic and adversarial regret for the problem of social welfare maximization. We first prove a lower bound on stochastic regret, which applies to any algorithm, and which immediately implies a lower bound on adversarial regret. We then introduce the algorithm Tempered Exp3 for Social Welfare. We show that, for an appropriate choice of tuning parameters, this algorithm achieves the rates of the lower bound on regret, up to a logarithmic term. Formal proofs of these bounds can be found in section C.2.

4.3.1 Lower Bound

Theorem 26 (Lower bound on regret)**.** *Consider the setup of Section 4.2. There exists a constant $C > 0$ such that, for any randomized algorithm α for the choice of $x_1, x_2, \ldots$ and any time horizon $T \in \mathbb{N}$, the following holds.*

1. *There exists a distribution μ on $[0,1]$ with associated expected demand function $\mathbb{G}$ for which the stochastic cumulative expected regret $R_T(\alpha, \mathbb{G})$ is at least $C \cdot T^{2/3}$.*

2. *There exists a sequence $(v_1, \ldots, v_T) \in [0,1]^T$ for which the adversarial cumulative expected regret $R_T\big(\alpha, \{v_i\}_{i=1}^T\big)$ is at least $C \cdot T^{2/3}$.*

The proof of Theorem 26 can be found in section C.2. The adversarial lower bound follows immediately from the stochastic lower bound, since worst case regret (over possible sequences of v_i)

is bounded below by average regret (over i.i.d. draws of v_i), for any distribution of v_i.

Sketch of proof To prove the stochastic lower bound we construct a family of distributions $(\mu^\varepsilon)_{\varepsilon\in[-1,1]}$ for v_i, indexed by a parameter $\varepsilon \in [-1, 1]$. The distributions in this family have four points of support, $(1/4, 1/2, 3/4, 1)$. The probability of these points is given by

$$\big(a, (1+\varepsilon)b, (1-\varepsilon)b, 1-a-2b\big).$$

The values of a and b are chosen such that (i) the two middle points $1/2, 3/4$ are far from optimal, for any value of ε, and (ii) learning which of the two end points $(1/4, 1)$ is optimal requires sampling from the middle.[‡] For each $\varepsilon \in [-1, 1]$, denote the demand function associated to μ^ε by $\mathbb{G}^\varepsilon$, and the expected social welfare associated to $\mathbb{G}^\varepsilon$ by $\mathbb{U}^\varepsilon$. Property (ii) holds because of the integral term $\int_{\frac{1}{4}}^{1} \mathbb{G}^\varepsilon(x')\mathrm{d}x'$, which shows up in $\mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(1/4)$. This construction is illustrated in Figure 4.1. This figure shows plots of $\mathbb{G}^\varepsilon$ and of $\mathbb{U}^\varepsilon$ for $\lambda = .95$ and $\varepsilon \in \{\pm 1\}$.

The difference in welfare $\mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(1/4)$ of the two candidates optimal policies $1/4$ and 1 depends on the sign of ε. In order not to suffer regret of order $|\varepsilon| \cdot T$, any learning algorithm needs to sample policies from points that are informative about the sign of ε. The only points that are informative are those in the region $(1/2, 3/4]$, where welfare is bounded away from optimal welfare.

More specifically, due to information-theoretic arguments, a learning algorithm has to sample on the order of $|\varepsilon|^{-2}$ times from the region $(1/2, 3/4]$ to be able to detect the sign of ε, incurring regret on the order of $|\varepsilon|^{-2}$ in the process. Any learning algorithm therefore incurs regret on the order of $\min\big(|\varepsilon|^{-2}, |\varepsilon| \cdot T\big)$, which, for $\varepsilon = \Theta(T^{-1/3})$, leads to the conclusion.

4.3.2 An Algorithm that Achieves the Lower Bound

We next introduce an algorithm that allows us to essentially achieve the lower bound on regret, in terms of rates. Algorithm 7 is a modification of the well-known Exp3 algorithm. Conventional Exp3, for the multi-armed bandit setting, uses inverse probability weighting to construct an unbiased estimator $\widehat{\boldsymbol{U}}_k$ of the cumulative payoff of each arm k. Apart from some constant fixed probability of exploration, any given arm is then chosen with probability proportional to $\exp(\eta \cdot \widehat{\boldsymbol{U}}_{ik})$, where η is a tuning parameter.

Modifications relative to standard Exp3 Relative to this standard algorithm, we require three modifications. First, we discretize the continuous support $[0, 1]$ of x, restricting attention to the grid of policy values $\widetilde{x}_k = (k-1)/K$. Second, since welfare $U_i(x)$ is not directly observed for the chosen policy x, we need to estimate it indirectly. In particular, we first form an estimate $\widehat{\boldsymbol{G}}_{ik}$ of cumulative demand for each of the policy values $\widetilde{x}_k$, using inverse probability weighting. We then use this estimated demand, interpolated using a step-function, to form estimates of cumulative social welfare, $\widehat{\boldsymbol{U}}_{ik} = \widetilde{x}_k \cdot \widehat{\boldsymbol{G}}_{ik} + \frac{\lambda}{K} \cdot \sum_{k'>k} \widehat{\boldsymbol{G}}_{ik'}$. Third, we require additional exploration, relative to Exp3. Since social welfare depends on demand for counterfactual policy choices, we need to explore policies that are away from the optimum, in order to learn the relative welfare of approximately optimal policy choices. The mixing weight γ, which determines the share of policies sampled from the uniform

[‡]Specifically, $a \coloneqq \frac{(1-\lambda)\cdot(136-99\cdot\lambda)}{2\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)}$, and $b \coloneqq \frac{1-\lambda}{2\cdot(24-17\cdot\lambda)}$. These two constants are strictly greater than zero, and satisfy $1 - a - 2 \cdot b > 0$.

Algorithm 7 Tempered Exp3 for Social Welfare

input: Tuning parameters K, γ and η
initialization: Calculate evenly spaced grid-points $\widetilde{x}_k \coloneqq (k-1)/K$ and initialize $\widehat{\boldsymbol{G}}_{1k} \coloneqq 0$ and $\widehat{\boldsymbol{U}}_{1k} = 0$ for $k = 1, \dots, K+1$
for individual $i = 1, 2, \dots, T$ **do**
For all $k = 1, 2, \dots, K+1$, set ▷ Assignment probabilities

$$p_{ik} \coloneqq (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\boldsymbol{U}}_{ik})}{\sum_{k'} \exp(\eta \cdot \widehat{\boldsymbol{U}}_{ik'})} + \frac{\gamma}{K+1} \tag{4.7}$$

Choose k_i at random according to the probability distribution $(p_{i,1}, \dots, p_{i,K+1})$
Set $x_i \coloneqq \widetilde{x}_{k_i}$, and query y_i accordingly
For all $k = 1, 2, \dots, K+1$, set ▷ Estimated demand

$$\widehat{\boldsymbol{G}}_{i+1,k} \coloneqq \widehat{\boldsymbol{G}}_{i,k} + y_i \cdot \frac{\mathbb{I}\{k_i = k\}}{p_{ik}} \tag{4.8}$$

For all $k = 1, 2, \dots, K+1$, set ▷ Estimated welfare

$$\widehat{\boldsymbol{U}}_{i+1,k} \coloneqq \widetilde{x}_k \cdot \widehat{\boldsymbol{G}}_{i+1,k} + \tfrac{\lambda}{K} \cdot \sum_{k'>k} \widehat{\boldsymbol{G}}_{i+1,k'} \tag{4.9}$$

distribution, needs to be larger relative to conventional Exp3, to ensure sufficient exploration away from the optimum.

Theorem 27 (Adversarial upper bound on regret of Tempered Exp3 for Social Welfare)**.** *Consider the setup of Section 4.2, and let α be Algorithm 7. Assume that $(K+1)\eta < \gamma$.*
Then for any sequence $(v_1, \dots, v_T) \in [0,1]^T$ the regret $R_T\big(\alpha, \{v_i\}_{i=1}^T\big)$ is bounded above by

$$\left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \tfrac{\lambda}{K}\right) \cdot T + \tfrac{\log(K+1)}{\eta}. \tag{4.10}$$

Suppose additionally that $c_1, c_2, c_3 > 0$ are constants. Then, there exists a constant c_4 such that, if we set $\gamma = c_1 \cdot \left(\frac{\log(T)}{T}\right)^{1/3}$, $\eta = c_2 \cdot \gamma^2$, and $K = \lfloor c_3/\gamma \rfloor$, the regret $R_T\big(\alpha, \{v_i\}_{i=1}^T\big)$ is bounded above by

$$c_4 \cdot \log(T)^{1/3} T^{2/3}. \tag{4.11}$$

As an immediate corollary of the previous theorem we get the following.

Corollary 1 (Stochastic upper bound on regret of Tempered Exp3 for Social Welfare)**.** *Under the assumptions of Theorem 27, suppose additionally that v_i is drawn i.i.d. from some distribution with associated expected demand function $\mathbb{G}$. Then the regret $R_T(\alpha, \mathbb{G})$ is bounded above by the same expressions as in Theorem 27.*

The proof of Theorem 27 can again be found in section C.2.

Tuning The statement of the theorem leaves the constants c_1, c_2, c_3 in the definition of the tuning parameters unspecified. Suppose we wish to choose the tuning parameters so as to optimize the upper bound obtained in Theorem 27. Ignoring the rounding of K, an approximate solution to this

problem is given by

$$\eta \coloneqq 1/a \cdot (\log(T)/T)^{2/3}$$
$$\gamma \coloneqq \lambda\sqrt{(e-2)/a} \cdot (\log(T)/T)^{1/3}$$
$$K \coloneqq \sqrt{3\lambda a/(e-2)} \cdot (T/\log(T))^{1/3}$$

where

$$a \coloneqq (9(e-2))^{1/3} \, (\sqrt{\lambda/3} + \lambda)^{2/3}.$$

This solution is obtained by taking the upper bound in Equation (4.10), approximating $(K+1)/K \approx 1$ and $(2K+1)/6 \approx K/3$, and solving the first order conditions with respect to the three tuning parameters. This approximation, and the tuning parameters specified above, yield an approximate upper bound on regret of $6 \cdot \log(T)^{1/3} T^{2/3}$.

Unknown time horizon Note that the proposed tuning depends crucially on knowledge of the time horizon T at which regret is evaluated. In order to extend our rate results to the case of unknown time horizons, we can use the so-called doubling trick; cf. Section 2.3 of [48]. Consider a sequence of epochs (intervals of time-periods) of exponentially increasing length, and re-run Algorithm 7 for each time-period separately, tuning the parameters over the current epoch length. This construction converts Algorithm 7 into an "anytime algorithm" which enjoys the same regret guarantees of Theorem 27, up to a multiplicative constant factor. Another more efficient strategy to achieve the same goal is to modify Algorithm 7, allowing the parameters η and γ to change at each iteration, and splitting each bin associated with the discretization parameter K whenever more precision is required.

4.4 Stochastic Regret Bounds for Concave Social Welfare

Theorem 26 in Section 4.3 provides a lower bound proportional to $T^{2/3}$ for adversarial and stochastic regret in social welfare maximization. The proof of this lower bound constructs a distribution for the v_i. This distribution is such that expected social welfare $\mathbb{U}(x)$ is non-concave, as a function of x; two global optima are separated by a region of lower welfare. In order to learn which of two candidates for the globally optimal policy is actually optimal, it is necessary to sample policies in between. These intermediate policies yield lower welfare, and sampling them contributes to cumulative regret. This construction is illustrated in Figure 4.1.

Given that the construction relies on non-concavity of expected social welfare, could we achieve lower regret if we knew that social welfare is actually concave? The answer turns out to be yes, for the stochastic setting (in the adversarial setting, cumulative welfare is necessarily non-concave). One reason is that concavity ensures that the function is unimodal. To estimate the difference in social welfare between two policies it therefore suffices to sample policies that lie in the interval between them. These in-between policies yield social welfare exceeding the minimum of the two boundary policies. A second reason is that concavity prevents unexpected spikes in social welfare. This property allows us to test carefully chosen triples of points for extended periods, to ensure that one of them is suboptimal, without incurring significant regret.

For the stochastic setting with concave social welfare, we present an algorithm that achieves a

bound on regret of order $T^{1/2}$, up to logarithmic terms. Before describing our proposed algorithm, Dyadic Search for Social Welfare, let us formally state the improved regret bounds. The proofs of these lower and upper bounds can again be found in Appendix C.2.

Theorem 28 (Lower bound on regret for the concave case)**.** *Consider the setup of Section 4.2. There exists a constant $C > 0$ such that, for any randomized algorithm α for the choice of $x_1, x_2, \dots$ and any time horizon $T \in \mathbb{N}$, there exists a distribution μ on $[0, 1]$ with associated expected demand function $\mathbb{G}$ and* concave *expected social welfare function $\mathbb{U}$, for which the regret $R_T(\alpha, \mathbb{G})$ is at least $C \cdot T^{1/2}$.*

Theorem 29 (Stochastic upper bound on regret of Dyadic Search for Social Welfare)**.** *Consider the stochastic setup of Section 4.2 and time horizon $T \in \mathbb{N}$. If α is Algorithm 8 run with confidence parameter $\delta = \frac{1}{T^{5/2}}$, and if the expected social welfare function $\mathbb{U}$ is concave, then, the regret $R_T(\alpha, \mathbb{G})$ is of order at most $T^{1/2}$, up to logarithmic terms.*

Dyadic search Our algorithm is based on a modification of dyadic search, as discussed in [22, 23]. At any point in time, this algorithm maintains an active interval I_τ, which contains the optimal policy with high probability. Only policies within this interval are sampled going forward. As evidence accumulates, this interval is trimmed down, by excluding policies that are sub-optimal with high probability.

The algorithm proceeds in epochs τ. At the start of each epoch, a sub-interval $[l, r] \subset I_\tau$ is formed, with mid-point $c = (l + r)/2$. The points l, c, r are in a dyadic grid, that is, they are of the form $k/2^m$. After sampling from $[l, r]$, we calculate confidence intervals $J_t(l, c)$, $J_t(c, r)$, and $J_t(l, r)$ for the welfare differences $\Delta(l, c)$, $\Delta(c, r)$, and $\Delta(l, r)$, where $\Delta(x, x') \coloneqq \mathbb{U}(x') - \mathbb{U}(x)$.

If the confidence interval $J_t(l, c)$ or $J_t(l, r)$ lies above 0, concavity implies that the optimal policy cannot lie to the left of l; we can thus trim the active interval I_τ by dropping all points to the left of l. Symmetrically, if the confidence interval $J_t(c, r)$ or $J_t(l, r)$ lies below 0, we can trim I_τ by dropping all points to the right of r.

Confidence intervals for welfare differences This procedure requires the construction of confidence intervals for welfare differences of the form

$$\Delta(x, x') \coloneqq \mathbb{U}(x') - \mathbb{U}(x) = x' \cdot \mathbb{G}(x') - x \cdot \mathbb{G}(x) - \lambda \int_x^{x'} \mathbb{G}(x'')\mathrm{d}x''. \tag{4.12}$$

At time t, we estimate demand $\mathbb{G}(x)$, for policies x chosen in previous periods, as[§]

$$\widehat{\mathbb{G}}_t(x) \coloneqq \frac{1}{n_t(x)} \sum_{i \leqslant t} y_i \cdot \mathbb{I}\{x_i = x\}, \qquad n_t(x) \coloneqq \sum_{i \leqslant t} \mathbb{I}\{x_i = x\}.$$

We similarly estimate integrated demand $\int_x^{x'} \mathbb{G}(x'')\mathrm{d}x''$ by $(x' - x)$ times the average of realized demand y_i for observations x_i in the open interval (x, x'). We have to be careful, however, to use a sample of x_i that is (approximately) uniformly distributed over this interval. This can be achieved

[§]We use the convention $0/0 = 0$ and $a/0 = +\infty$ whenever $a > 0$. Furthermore, every summation over an empty set of indices is understood to have value 0.

for our dyadic search procedure, as specified in Algorithm 8, by truncating the time index used to estimate this average.[¶] Let

$$s(x,x',t) \coloneqq \max\left\{ s \leqslant t :\ \log_2\left(1+\sum_{i\leqslant s}\mathbb{I}\left\{x_i\in(x,x')\right\}\right)\in\mathbb{N}\right\}.$$

We define

$$\widehat{\mathbb{G}}_t(x,x') \coloneqq \frac{1}{n_t(x,x')+1}\sum_{i\leqslant s(x,x',t)} y_i\cdot\mathbb{I}\left\{x_i\in(x,x')\right\},\qquad n_t(x,x') \coloneqq \sum_{i\leqslant s(x,x',t)}\mathbb{I}\left\{x_i\in(x,x')\right\}.$$

At each round, Algorithm 8 maintains estimates for welfare differences among three points l,c,r (for left, center and right, respectively). The estimate of the welfare difference between $x'=c$ and $x=l$ (or between $x'=r$ and $x=c$) is given by

$$\widehat{\Delta}_t(x,x') \coloneqq x'\cdot\widehat{\mathbb{G}}_t(x') - x\cdot\widehat{\mathbb{G}}_t(x) - \lambda\cdot(x'-x)\cdot\widehat{\mathbb{G}}_t(x,x'). \tag{4.13}$$

while the estimate of the welfare difference between r and l is given by

$$\widehat{\Delta}_t(l,r) \coloneqq \widehat{\Delta}_t(l,c) + \widehat{\Delta}_t(c,r). \tag{4.14}$$

To construct confidence intervals for $\Delta(x,x')$, we also need to quantify the uncertainty of our demand estimates. We use the following interval half-lengths for confidence intervals for tax revenue at x, and for the private welfare difference between x' and x:

$$\Gamma_t(x) \coloneqq x\cdot\sqrt{\tfrac{1}{2n_t(x)}\log\left(\tfrac{2}{\delta}\right)},\quad \Gamma_t(x,x') \coloneqq \lambda\cdot(x'-x)\cdot\left(\sqrt{\frac{1}{2\left(n_t(x,x')+1\right)}\log\left(\tfrac{2}{\delta}\right)}+\tfrac{2}{n_t(x,x')+1}\right).$$

Using the shorthand $a\pm b=[a-b,a+b]$, our confidence interval for $\Delta(x,x')$, where $x'=c$ and $x=l$ (or $x'=r$ and $x=c$) is given by

$$J_t(x,x') \coloneqq \widehat{\Delta}_t(x,x') \pm \left(\Gamma_t(x')+\Gamma_t(x)+\Gamma_t(x,x')\right), \tag{4.15}$$

while our confidence interval for $\Delta(l,r)$ is given by

$$J_t(l,r) \coloneqq \widehat{\Delta}_t(l,r) \pm \left(\Gamma_t(r)+\Gamma_t(l)+\Gamma_t(l,c)+\Gamma_t(c,r)\right). \tag{4.16}$$

With these preliminaries, we are now ready to state our algorithm, Dyadic Search for Social Welfare.

Before concluding this section, we highlight two features of Algorithm 8. First, two of the three points l,c,r, and the corresponding estimates of demand, are kept from each epoch to the next. Second, estimation of the integral term is performed by querying points following a fixed and balanced design on the dyadic grid – instead of, for example, using a randomized Monte Carlo procedure which may lead to unbalanced exploration. This implies that the points queried to estimate the integral terms can be easily reused to obtain other integral estimates from each epoch to the next.

[¶]The sampling procedure in Algorithm 8 samples sequentially from the dyadic grid in the active interval, refining the grid in subsequent iterations. $s(x,x',t)$ provides a truncation of the time index such that one round of such dyadic sampling has been completed.

Algorithm 8 Dyadic Search for Social Welfare

input: Confidence parameter $\delta \in (0,1)$
initialization: Set $I_1 \coloneqq [0,1]$, $t_0 \coloneqq 0$, $k \coloneqq 0$
for epochs $\tau = 1, 2, \ldots$ **do**
  Let $c \coloneqq (\sup I_\tau + \inf I_\tau)/2$, and $d \coloneqq \sup I_\tau - \inf I_\tau$ ▷ Subinterval for sampling
  if τ is odd **then**
    Let $l \coloneqq c - \frac{1}{4}d$, $r \coloneqq c + \frac{1}{4}d$
  else
    Let $l \coloneqq c - \frac{1}{6}d$, $r \coloneqq c + \frac{1}{6}d$
  for $t = t_{\tau-1} + 1, t_{\tau-1} + 2, \ldots$ **do**
    Select $w \in \arg\max_{w' \in \{l,c,r,(l,c),(c,r)\}} \Gamma_{t-1}(w')$, ▷ Sampling
      (breaking ties following the order $l, c, r, (l,c), (c,r)$)
    if $w \in \{l, c, r\}$ **then**
      Set $x_t \coloneqq w$.
    else
      Set $x_t \coloneqq w_1 + (w_2 - w_1) \cdot \frac{k+1/2}{n_{t-1}(w_1,w_2)+1}$, and $k \coloneqq (k+1) \mod n_{t-1}(w_1, w_2) + 1$.
    Calculate $J_t(l,c)$, $J_t(c,r)$, and $J_t(l,r)$, as in Equations (4.15) and (4.16) ▷ Inference
    if $\inf\big(J_t(l,c)\big) \geqslant 0$ or $\inf\big(J_t(l,r)\big) \geqslant 0$ **then**
      let $I_{\tau+1} \coloneqq I_\tau \cap [l,1]$ and $t_\tau \coloneqq t$ and **break** ▷ Shrinking the active interval
    else if $\sup\big(J_t(c,r)\big) \leqslant 0$ or $\sup\big(J_t(l,r)\big) \leqslant 0$ **then**
      let $I_{\tau+1} \coloneqq I_\tau \cap [0,r]$ and $t_\tau \coloneqq t$ and **break** ▷ Shrinking the active interval

These two features combined ensure that Algorithm 8 recycles information very efficiently to prune the active interval quickly.

4.5 Income Taxation

We discuss two extensions of the baseline model of optimal taxation that we introduced in Section 4.2. These extensions incorporate features that are important in more realistic models of optimal taxation. The first extension, discussed in this section, is a variant of the Mirrlees model of optimal income taxation [140, 159, 160]. The second extension, discussed in Appendix C.1 is a variant of the Ramsey model of commodity taxation [154].

Our model of income taxation generalizes our baseline model by allowing for heterogeneous wages w_i, welfare weights $\omega(w_i)$, extensive-margin labor supply responses determined by the cost of participation v_i, and non-linear income taxes $x_i = \mathbf{x}(w_i)$. Two simplifications are maintained in this model, relative to a more general model of income taxation. First, only extensive margin responses (participation decisions) by individuals are allowed; there are no intensive margin responses (hours adjustments). Second, as in the baseline model of Section 4.2, there are no income effects. In imposing these assumptions, our model mirrors the model of optimal income taxation discussed in Section II.2 of [160].

Setup At each time $i = 1, 2, \ldots, T$, one individual arrives who is characterized by (i) a potential wage $w_i \in [0,1]$, and (ii) an unknown cost of participation $v_i \in [0,1]$. This individual makes a binary labor supply decision y_i. If she participates in the labor market ($y_i = 1$), she earns w_i, but pay a tax according to the tax rate $x_i = \mathbf{x}(w_i)$ on her earnings w_i. She furthermore incurs a non-monetary

cost of participation v_i.

Her optimal labor supply decision is therefore given by $y_i = \mathbb{I}\{v_i \leqslant w_i \cdot (1 - x_i)\}$, and private welfare equals $\max(w_i \cdot (1 - x_i) - v_i, 0)$. The implied public revenue is equal to the tax on earnings $x_i \cdot w_i$ if $y_i = 1$, and 0 otherwise.

We define social welfare as a weighted sum of public revenue and private welfare, with a weight $\omega(w_i)$ for the latter. Typically, ω is a decreasing function of w chosen by the policymaker (and hence, we assume it is known to the learner), reflecting a preference for redistribution towards those with lower earnings potential, cf. [161]. Social welfare for time period i, as a function of the tax schedule $\mathbf{x}(\cdot)$ chosen by the learner, is therefore given by

$$U_i\big(\mathbf{x}(\cdot)\big) \coloneqq \underbrace{\mathbf{x}(w_i) \cdot w_i \cdot \mathbb{I}\{v_i \leqslant w_i \cdot (1 - \mathbf{x}(w_i))\}}_{\text{Public revenue}} \quad + \quad \omega(w_i) \cdot \underbrace{\max(w_i \cdot (1 - \mathbf{x}(w_i)) - v_i, 0)}_{\text{Private welfare}}. \tag{4.17}$$

After period i, we observe y_i. If $y_i = 1$, we also observe w_i. Nothing else is observed.[‖]

Piecewise constant tax schedules We next construct a generalization of Algorithm 7 based on piecewise constant tax schedules, with tax rates changing at the grid-points $\mathcal{W}$, where $0 \in \mathcal{W} \subset [0, 1]$. Formally, define $\lfloor w \rfloor \coloneqq \max\{w' \in \mathcal{W} : \ w' \leqslant w\}$, rounding the wage w down to the nearest grid-point in $\mathcal{W}$,[**] Denote $H \coloneqq |\mathcal{W}|$, and let

$$\mathcal{X}_{\mathcal{W}} \coloneqq \{\mathbf{x}(\cdot) : \ \forall w \in [0, 1], \ \mathbf{x}(w) = \mathbf{x}(\lfloor w \rfloor)\}.$$

For $w \in \mathcal{W}$ and any $x \in [0, 1]$, denote

$$G_i(w, x) \coloneqq w_i \cdot \mathbb{I}\{v_i \leqslant w_i \cdot (1 - x)\} \cdot \mathbb{I}\{\lfloor w_i \rfloor = w\},$$

so that $y_i \cdot w_i = G_i\big(w_i, \mathbf{x}_i(w_i)\big)$. $G_i(w, x)$ is the individual labor supply function, in monetary units, interacted with an indicator for whether the wage w_i falls into the tax bracket starting at w. With this notation, we can rewrite

$$\max\big(w_i \cdot (1 - x) - v_i, 0\big) = \int_x^1 G_i(\lfloor w_i \rfloor, x')\mathrm{d}x'.$$

For piecewise constant tax rates $\mathbf{x}(\cdot)$ we get

$$U_i(\mathbf{x}(\cdot)) = \sum_{w \in \mathcal{W}} \left[\mathbf{x}(w) \cdot G_i(w, \mathbf{x}(w)) + \omega(w_i) \cdot \int_{\mathbf{x}(w)}^1 G_i(w, x')\mathrm{d}x' \right]. \tag{4.18}$$

[‖]It should be noted that in this model we take the transfer x_0 for individuals without other income as given. The effective tax owed by an employed individual equals $\mathbf{x}(w_i) \cdot w_i - x_0$. The "unconditional basic income" x_0 does not affect labor supply, given our assumption that there are no income effects, and it enters social welfare additively. It is therefore without loss of generality to omit x_0 from our model.

[**]Here we use slightly non-standard notation, where $\lfloor \cdot \rfloor$ denotes rounding down to the nearest grid-point, rather than the nearest integer.

Cumulative social welfare is given by $\boldsymbol{U}_i \coloneqq \sum_{j \leqslant i} U_i(\mathbf{x}_i(\cdot))$, and we correspondingly define cumulative expected regret of an algorithm α, in the adversarial setting, as

$$R_T\big(\alpha, (v_i, w_i)_{i=1}^T\big) \coloneqq \sup_{\mathbf{x}(\cdot) \in \mathcal{X}_\mathcal{W}} \mathbb{E}\left[\boldsymbol{U}_T(\mathbf{x}(\cdot)) - \boldsymbol{U}_T\right].$$

The supremum here is taken over all tax schedules $\mathbf{x}(\cdot)$ that are piecewise constant between the gridpoints $w \in \mathcal{W}$.

Algorithm Algorithm 9 generalizes Algorithm 7 to this setting. As before, we form an unbiased estimate $\widehat{G}_i$ of G_i using inverse probability weighting, map this estimate into a corresponding estimate $\widehat{U}_i$ of U_i, based on Equation (4.18), and cumulate across time periods to obtain $\widehat{\boldsymbol{U}}_i$. Note that w_i is observed whenever $y_i = 1$. This implies that the estimate $\widehat{G}_i$ is in fact a function of observables, and the same holds for $\widehat{U}_i$.

Algorithm 9 keeps track of estimated demand and social welfare for each bin ("tax bracket"), as defined by the gridpoints $w \in \mathcal{W}$. The algorithm then constructs a distribution $p_i(x|w)$ over tax rates $x \in \mathcal{X}$ given w, using the tempered Exp3 distribution. The tax schedule $\mathbf{x}(\cdot)$ is sampled according to these (marginal) distributions of tax rates for each bracket. Though immaterial for the following theorem, we choose the perfectly correlated coupling, across brackets, of these marginal distributions, which is implemented using the random variable A_i in Algorithm 9.

Algorithm 9 Tempered Exp3 for Optimal Income Taxation

input: Tuning parameters K, γ and η, and set of gridpoints $\mathcal{W} \subset [0,1]$
initialization: Calculate evenly spaced grid-points $\mathcal{X} \coloneqq \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\}$, initialize $\widehat{\boldsymbol{U}}_1(w,x) \coloneqq 0$ for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$
for individual $i = 1, 2, \dots, T$ **do**
For all $x, w \in \mathcal{X}$, set $\lfloor w \rfloor = \max\{w' \in \mathcal{W} : \ w' \leqslant w\}$, and ▷ Assignment probabilities

$$p_i(x|w) \coloneqq (1-\gamma) \cdot \frac{\exp(\eta \cdot \widehat{\boldsymbol{U}}_i(x, \lfloor w \rfloor))}{\sum_{x' \in \mathcal{X}} \exp(\eta \cdot \widehat{\boldsymbol{U}}_i(x', \lfloor w \rfloor))} + \frac{\gamma}{K+1}. \tag{4.19}$$

Draw $A_i \sim U[0,1]$, for all $w \in [0,1]$, set

$$\mathbf{x}_i(w) \coloneqq \max\left\{ x \in \mathcal{X} : \sum_{x' \in \mathcal{X}, x' < x} p_i(x'|w) \leqslant A_i \right\}, \tag{4.20}$$

and query y_i accordingly.
For all $w \in \mathcal{W}$ and $x \in \mathcal{X}$, set ▷ Estimated labor supply

$$\widehat{G}_i(x,w) \coloneqq y_i \cdot w_i \cdot \frac{\mathbb{I}\{\lfloor w_i \rfloor = w, \mathbf{x}_i(w_i) = x\}}{p_i(x|w)}. \tag{4.21}$$

For all $w \in \mathcal{W}$ and $x \in \mathcal{X}$, set ▷ Estimated welfare

$$\widehat{\boldsymbol{U}}_{i+1}(x,w) \coloneqq \widehat{\boldsymbol{U}}_i(x,w) + x \cdot \widehat{G}_i(x,w) + \frac{\omega(w_i)}{K} \cdot \sum_{x' \in \mathcal{X}, x' > x} \widehat{G}_i(x', w). \tag{4.22}$$

Theorem 30 (Adversarial upper bound on regret of Tempered Exp3 for Optimal Income Taxation)**.** *Consider the setup of Section 4.5, and let α be Algorithm 9. Assume that $(K+1)\eta < \gamma$, and that $\omega(w) \leqslant 1$ for all w.*
Then for any sequence $\big((v_1, w_1), \dots, (v_T, w_T)\big) \in [0,1]^{2T}$, the regret $R_T\big(\alpha, (v_i, w_i)_{i=1}^T\big)$ is bounded above by

$$\left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{1}{\gamma}\right) + \tfrac{1}{K}\right) \cdot T + \tfrac{H \log(K+1)}{\eta}. \tag{4.23}$$

Suppose additionally[††] *that $K = c_1 \cdot (T/H)^{1/3}$, $\gamma = c_2/(K+1)$, and $\eta = c_3/(K+1)^2$, for some constants c_1, c_2, c_3. Then the regret $R_T\big(\alpha, (v_i, w_i)_{i=1}^T\big)$ is bounded above by*

$$c_4 \cdot H^{1/3} \cdot \log(T)^{1/3} T^{2/3}, \tag{4.24}$$

for some constant c_4.

4.6 Conclusions

In this chapter, we investigated the problem of adaptive optimal taxation in a regret minimization framework from the perspective of a policymaker whose goal is to maximize social welfare. We compared optimal taxation to dynamic pricing, bilateral trade, and (finite) multi-armed bandits, discussing the similarities and differences of these problems in the process. We provided tight upper and lower bounds for the optimal taxation problem under various stochastic and adversarial assumptions. Finally, we proposed two interesting generalizations of the optimal taxation problem —the income taxation problem and the commodity taxation problem— whose regret regimes characterization we left for future research.

[††]for simplicity, we assume that in the following tuning K is an integer. If not, round K to the closest integer.

Chapter 5

Nonstochastic Bandits with Composite Anonymous Feedback

5.1 Introduction

Multiarmed bandits, originally proposed for managing clinical trials, are now routinely applied to a variety of other tasks, including computational advertising, e-commerce, and beyond. Typical examples of e-commerce applications include content recommendation systems, like the recommendation of products to visitors of merchant websites and social media platforms. A common pattern in these applications is that the response elicited in a user by the recommendation system is typically not instantaneous, and might occur some time in the future, well after the recommendation was issued. This delay, which might depend on several unknown factors, implies that the reward obtained by the recommender at time t can actually be seen as the combined effect of many previous recommendations to that user.

The more specific scenario of bandits with delayed rewards has been investigated in the literature under the assumption that the contributions of past recommendations to the combined reward is individually discernible —see, e.g., [51, 109, 145, 176]. Pike-Burke et al. [149] revisited the problem of bandits with delayed feedback under the more realistic assumption that only the combined reward is available to the system, while the individual reward components remain unknown. This model captures a much broader range of practical settings where bandits are successfully deployed. Consider for example an advertising campaign which is spread across several channels simultaneously (e.g., radio, tv, web, social media). A well-known problem faced by the campaign manager is to disentangle the contribution of individual ads deployed in each channel from the overall change in sales. Pike-Burke et al. [149] formalized this harder delayed setting in a bandit framework with stochastic rewards, where they introduced the notion of *delayed anonymous feedback* to emphasize the fact that the reward received at any point in time is the sum of rewards of an unknown subset of past selected actions. More specifically, choosing action $I_t \in [K]$ at time t generates a stochastic reward $Y_t(I_t) \in [0,1]$ and a stochastic delay $\tau_t \in \{0,1,\dots\}$, where $\big\{Y_t(i), \tau_t\big\}_{i\in[K], t\in\mathbb{N}}$ is a family of independent random variables such that $Y_1(i), Y_2(i), \dots$ have a common distribution $\nu_Y(i)$ (for all arms $i \in [K]$) and $\tau_1, \tau_2, \dots$ have a common distribution ν_τ with expectation μ_τ. The delayed anonymous feedback assumption entails that the reward observed at time t by the algorithm is the sum of t components of the form $Y_s(I_s)\mathbb{I}\{\tau_s = t - s\}$ for $s \in \{1, \dots, t\}$. The main result in [149] is

NO DELAY	DELAYED FEEDBACK	ANONYMOUS COMPOSITE FEEDBACK
$\sqrt{KT}$ [18]	$\sqrt{(d+K)T}$ [51]	$\sqrt{(d+1)KT}$ (this chapter)

Table 5.1: Summary of the regret regimes in delayed multi-armed bandits

that, when the expected delay μ_τ is known, the regret is at most of order of $K\big((\ln T)/\Delta + \mu_\tau\big)$, where Δ is the suboptimality gap. This bound is of the same order as the corresponding bound for the setting where the feedback is stochastically delayed, but not anonymous [109], and cannot be improved in general.

In this chapter, we study a bandit setting similar to delayed anonymous feedback, but with two important differences. First, we work in a nonstochastic bandit setting, where rewards (or losses, in our case) are generated by some unspecified deterministic mechanism. Second, we relax the assumption that the loss of an action is charged to the player at a single instant in the future. More precisely, we assume that the loss for choosing an action at time t is adversarially spread over at most $d+1$ consecutive time steps $t, t+1, \dots, t+d$. Hence, the loss observed by the player at time t is a *composite loss*, that is, the sum of $(d+1)$-many loss components $\ell_t^{(0)}(I_t), \ell_{t-1}^{(1)}(I_{t-1}), \dots, \ell_{t-d}^{(d)}(I_{t-d})$, where $\ell_{t-s}^{(s)}(I_{t-s})$ defines the s-th loss component from the selection of action I_{t-s} at time $t-s$. Note that in the special case when $\ell_t^{(s)}(i) = 0$ for all $s \neq d_t$, and $\ell_t^{(d_t)}(i) = \ell_t(i)$, we recover the model of nonstochastic bandits with delays $d_1, d_2, \dots \leqslant d$ (which, in particular, reduces to the standard nonstochastic bandits when $d = 0$). Our setting, which we call *composite anonymous feedback*, can accomodate scenarios where actions have a lasting effect which combines additively over time. Online businesses provide several use cases for this setting. For instance, an impression that results in an immediate clickthrough, later followed by a conversion, or a user that interacts with a recommended item —such as media content— multiple times over several days, or the free credit assigned to a user of a gambling platform which might not be used all at once.

Our main contribution is a general reduction technique (Composite Loss Wrapper, or CoLoWr, Algorithm 10) turning a base nonstochastic bandit algorithm into one operating within the composite anonymous feedback setting. We then show that the regret of CoLoWr can be upper bounded in terms of the stability and the regret of the base algorithm (Theorem 31). Choosing as a base algorithm Follow the Regularized Leader (FTRL) with Tsallis entropy, Theorem 31 gives an upper bound of order $\sqrt{(d+1)KT}$ on the regret of nonstochastic bandits with composite anonymous feedback (Corollary 2), where $d \geqslant 0$ is a known upper bound on the delay, K is the number of actions, and T is the time horizon. This result relies on a nontrivial stability analysis of FTRL with Tsallis entropy that could be of independent interest (Theorem 32). Finally, we show the optimality of the $\sqrt{(d+1)KT}$ rate by proving a matching lower bound (up to a logarithmic factor, Theorem 34). In particular, this shows that, in the nonstochastic case with delay d, anonymous feedback is strictly harder than nonanonymous feedback, whose minimax regret was characterized by Cesa-Bianchi et al. [51] as $\sqrt{(d+K)T}$. See Table 5.1 for a summary of results for nonstochastic K-armed bandits (all rates are optimal ignoring logarithmic factors). We now give an idea of the proof techniques. Similar to [149], we play the same action for a block of at least $2d+1$ time steps, hence the feedback we get in the last $d+1$ steps contains only loss components pertaining to the same action, so that we can estimate in those steps the "true loss" of that action. Unfortunately, although the original

losses are in $[0,1]$, the composite losses can be as large as $d+1$ (a composite loss sums $d+1$ loss components, and each component can be as large as 1). This causes a corresponding scaling in the regret, compromising optimality. However, we observe that the total composite loss relative to the same action over any $d+1$ consecutive steps can be at most $2d+1$ (Lemma 12). Hence, we can normalize the total composite loss relative to the same action over $d+1$ consecutive steps, simply dividing by $2d+1$, obtaining an average loss in the range $[0,1]$. This idea leads to the right dependence on d in the regret. The last problem is how to avoid suffering a big regret in the first d steps of each block, where the composite losses mix loss components belonging to more than one action. We solve this issue by borrowing an idea of Dekel et al. [75]. We build blocks with random endpoints so that their length is (always) at least $2d+1$ and (on average) not much bigger. This random positioning and length of the blocks is the key to prevent the oblivious adversary from causing a large regret in the first half of each block. Moreover, as we prove in Theorem 31, if the distribution over actions maintained by the base algorithm is *stable* (Definition 3), then the algorithm is not significantly affected by the uncertainty in the positioning of the blocks. Extending our results to the case where d is unknown, [181] show a regret bound of order $T^{2/3}$. When d is known, however, their analysis does not guarantee our faster $\sqrt{T}$ rate.

Further related work Online learning with delayed feedback was studied in the full information (non-bandit) setting by Garrabrant et al. [96], Joulani et al. [109, 110], Khashabi et al. [115], Langford et al. [121], Mann et al. [132], Mesterharm [138], Quanrud and Khashabi [152], Weinberger and Ordentlich [184], see also [167] for an interesting variant. The bandit setting with delay was investigated in [6, 51, 94, 108, 109, 124, 131, 145, 149, 174, 176, 177, 192]. Our delayed composite loss function generalizes the composite loss function setting of Dekel et al. [74]—see the discussion at the end of Section 5.2 for details—and is also related to the notion of loss functions with memory. This latter setting has been investigated, e.g., by [14], who showed how to turn an online algorithm with regret guarantee of $O(T^q)$ into one attaining $O(T^{1/(2-q)})$-policy regret, also adopting a blocking scheme. A more recent paper in this direction is [12], where the authors considered a more general loss framework than ours, though with the benefit of counterfactual feedback, in that the algorithm is aware of the loss it would incur had it played any sequence of d decisions in the previous d rounds, thereby making their results incomparable to ours.

5.2 Preliminaries

We denote the set of positive integers by $\mathbb{N}$ and the set of integers by $\mathbb{Z}$. For all $n \in \mathbb{N}$ we denote the set $\{1,\dots,n\}$ of the first n integers by $[n]$. We will use the handy convention that, if $(c_t)_{t\in\mathbb{Z}} \subset \mathbb{R}$ and $m,n \in \mathbb{Z}$ are such that $m > n$, then $\sum_{t=m}^{n} c_t = 0$ and $\prod_{t=m}^{n} c_t = 1$. For any $x \in \mathbb{R}$, we denote its positive part $\max\{x,0\}$ by x^+.

We start by considering a nonstochastic multiarmed bandit problem on K actions with oblivious losses in which the loss $\ell_t(i) \in [0,1]$ at time t of an action $i \in [K]$ is defined by the sum

$$\ell_t(i) \coloneqq \sum_{s=0}^{d} \ell_t^{(s)}(i)$$

of $(d+1)$-many components $\ell_t^{(s)}(i) \geqslant 0$ for $s \in \{0,\dots,d\}$. Let I_t denote the action chosen by the

player at the beginning of round t. If $I_t = i$, then the player incurs loss $\ell_t^{(0)}(i)$ at time t, loss $\ell_t^{(1)}(i)$ at time $t+1$, and so on until time $t+d$. Yet, what the player observes at time t is only the combined loss incurred at time t, which is the sum

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \cdots + \ell_{t-d}^{(d)}(I_{t-d})$$

of the past $d+1$ loss contributions, where $\ell_t^{(s)}(i) = 0$ for all i and s when $t \leqslant 0$. Then, we define the d-delayed *composite* loss at time t of a sequence of $d+1$ actions $i_{t-d}, \ldots, i_t \in [K]$ as

$$\ell_t^{\circ}(i_{t-d}, \ldots, i_t) \coloneqq \sum_{s=0}^{d} \ell_{t-s}^{(s)}(i_{t-s}) \,. \tag{5.1}$$

With this notation, the d-delayed composite anonymous feedback assumption states that what the player observes at the end of each round t is only the composite loss $\ell_t^{\circ}(I_{t-d}, \ldots, I_t)$. The goal of the algorithm is to bound its regret R_T against the best fixed action in hindsight,

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{\circ}(I_{t-d}, \ldots, I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t^{\circ}(i, \ldots, i) \,.$$

We define the regret in terms of the composite losses ℓ_t° rather than the true losses ℓ_t because in our model ℓ_t° is what the algorithm pays overall on round t. It is easy to see that a bound on R_T implies a bound on the more standard notion of regret $\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_k \sum_{t=1}^{T} \ell_t(k)$ up to an additive term of at most $O(d)$.

Our setting generalizes the composite loss function setting of Dekel et al. [74]. Specifically, the linear composite loss function therein can be seen as a special case of the composite loss (5.1) once we remove the superscripts s from the loss function components. In fact, in the linear case, the feedback in [74] allows one to easily reconstruct each individual loss component in a recursive manner. This is clearly impossible in our more involved scenario, where the new loss components that are observed in round t need not have occurred in past rounds.

5.3 The CoLoWr Algorithm

Our Composite Loss Wrapper algorithm (Algorithm 10) takes as input a standard K-armed bandit algorithm α and a Boolean sequence $\boldsymbol{B}$. The base algorithm α operates on standard (noncomposite) losses with values in $[0,1]$, producing probability distributions $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ over the action set $[K]$. The wrapper calls the base algorithm α only in a subset of rounds determined by the Boolean sequence $\boldsymbol{B}$, which we call update rounds.

Definition 2 (Update round)**.** *We say that $t \in \mathbb{N}$ is an update round with respect to a Boolean sequence $\boldsymbol{B} = (b_t)_{t\in\mathbb{N}} \subset \{0,1\}^{\mathbb{N}}$ if $t \geqslant 2d+1$ and $b_t \prod_{s=1}^{2d}(1 - b_{t-s}) = 1$.*

Note that if $d > 0$, the condition is equivalent to $b_t = 1$, and $b_{t-1} = \ldots = b_{t-2d} = 0$. If $d = 0$, by our convention, the condition is equivalent to $b_t = 1$.

To help understand our algorithm, we will also define two other types of rounds. We say that t is a *draw round* if $t = 1$ or the previous round $t-1$ was an update round. If t is not a draw round,

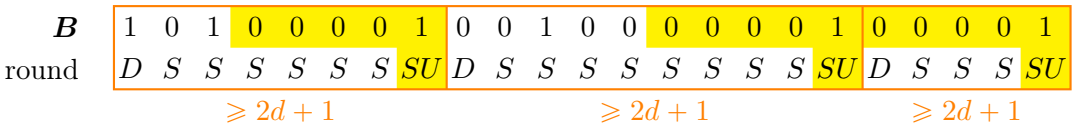


Figure 5.1: Sequence of rounds the algorithm is undergoing when $d = 2$. The top line contains the values of the Boolean sequence $\boldsymbol{B} = (B_t)_{t\in\mathbb{N}}$. The bottom line shows the corresponding types of rounds: each block begins with a (D)raw round, followed by a variable number of (S)tay rounds, the last of which is also an (U)pdate round. Since a round t is an update round only if $B_t = 1$ and $B_s = 0$ for the $2d$ previous rounds s, the length of each block is at least $2d + 1$.

we say that it is a *stay round*. Note that, if $d = 0$, both draw and stay rounds can be update rounds, while if $d \geqslant 1$, only stay rounds can be update rounds.

The CoLoWr algorithm proceeds in blocks of (random) length of at least $2d + 1$ rounds in which it constantly plays the same action (Figure 5.1). Blocks in Algorithm 10 are counted by variable n_t. Each block n_t consists of a draw round followed by ($2d$ or more) stay rounds, with the last round of the block being also an update round. During a draw round t, CoLoWr uses its current distribution $\boldsymbol{p}_t$ to draw and play an action I_t. During stay rounds, it keeps playing the action that was drawn during the latest draw round. After playing the action I_t for the current round t, if t is an update round, CoLoWr asks the base algorithm α to make an update of its base distribution $\boldsymbol{q}_{n_t} \to \boldsymbol{q}_{n_t+1}$ as if α played action I_t and observed as the loss of I_t the quantity $\frac{1}{2d+1}\sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d},\dots,I_s)$. Then, the block ends and the distribution of CoLoWr at the beginning of the next block $n_{t+1} = n_t + 1$ is $\boldsymbol{p}_{t+1} = \boldsymbol{q}_{n_t+1}$.

Note that if t is an update round, the quantity $\frac{1}{2d+1}\sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d},\dots,I_s)$ that is fed back to α relates only to the current action I_t, because blocks contain at least $2d + 1$ rounds and the same action is played in all of them.

Algorithm 10 CoLoWr (Composite Loss Wrapper)

input: Base K-armed bandit algorithm A and Boolean sequence $\boldsymbol{B}$
initialization: let $n_0 \coloneqq 0$ and $\boldsymbol{q}_1$ be the initial distribution over $[K]$ of α
for round $t = 1, 2, \dots$ **do**
  if either $t = 1$ or $t - 1$ was an update round (w.r.t. $\boldsymbol{B}$) **then**
    let $n_t \coloneqq n_{t-1} + 1$, $\boldsymbol{p}_t \coloneqq \boldsymbol{q}_{n_t}$, and draw $I_t \sim \boldsymbol{p}_t$ ▷ draw
  else
    let $n_t \coloneqq n_{t-1}$, $\boldsymbol{p}_t \coloneqq \boldsymbol{p}_{t-1}$, and $I_t \coloneqq I_{t-1}$ ▷ stay
  play I_t and observe loss $\ell_t^\circ(I_{t-d},\dots,I_t)$
  if t is an update round (w.r.t. $\boldsymbol{B}$) **then** ▷ update
    feed α with arm I_t and loss $\frac{1}{2d+1}\sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d},\dots,I_s)$
    use the update rule $\boldsymbol{q}_{n_t} \to \boldsymbol{q}_{n_t+1}$ of α to obtain a new base distribution $\boldsymbol{q}_{n_t+1}$

The following lemma shows that this quantity is indeed in $[0, 1]$, so that it is a legitimate feedback to pass to the base algorithm α.

Lemma 12. *For all $t \geqslant 2d + 1$ and $i \in [K]$,*

$$\sum_{\tau=t-d}^{t} \ell_\tau^\circ(i,\dots,i) \leqslant 2d + 1\,.$$

Proof. For all $t \geqslant 2d+1$ and $i \in [K]$,

$$\sum_{\tau=t-d}^{t} \ell_\tau^\circ(i,\dots,i) = \sum_{\tau=t-d}^{t}\sum_{s=0}^{d} \ell_{\tau-s}^{(s)}(i) = \sum_{s=0}^{d}\sum_{\tau=t-d}^{t} \ell_{\tau-s}^{(s)}(i) = \sum_{s=0}^{d}\sum_{\rho=t-d-s}^{t-s} \ell_\rho^{(s)}(i)$$
$$\leqslant \sum_{s=0}^{d}\sum_{\rho=t-2d}^{t} \ell_\rho^{(s)}(i) = \sum_{\rho=t-2d}^{t}\sum_{s=0}^{d} \ell_\rho^{(s)}(i) = \sum_{\rho=t-2d}^{t} \ell_\rho(i) \leqslant t-(t-2d)+1 = 2d+1\,.$$

□

As a final remark, we point out that, albeit the algorithm is parameterized with an entire sequence $\boldsymbol{B}$, at each time t, it does not require the knowledge of the sequence at future times $t+1, t+2, \dots$. This implies in particular that these Boolean values could be produced and fed to CoLoWr in an on-line fashion.

5.4 Upper Bound

We begin by formalizing the notion of stability (of the base algorithm), in terms of which we express the performance of the CoLoWr algorithm.

Definition 3 (ξ-stability)**.** *Let $\xi > 0$, α be a K-armed bandit algorithm, and $(\boldsymbol{q}_n)_{n\in\mathbb{N}}$ be the (random) sequence of probability distributions over actions $[K]$ produced by α during a run over rounds $\{1,2,\dots\}$. We say that α is ξ-stable if for any round n, we have*

$$\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+\right] \leqslant \xi\,.$$

In the previous definition, note that since $\sum_{i\in[K]} \boldsymbol{q}_{n+1}(i) = 1 = \sum_{i\in[K]} \boldsymbol{q}_n(i)$, then

$$\big\|\boldsymbol{q}_{n+1} - \boldsymbol{q}_n\big\|_1 = \big\|\boldsymbol{q}_{n+1} - \boldsymbol{q}_n\big\|_1 + \sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big) = 2\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+\,.$$

Therefore, the ξ-stability of an algorithm is equivalent to controlling the expected $\|\cdot\|_1$-distance between any two consecutive probability distributions produced by the algorithm. We stick to the positive part definition as this is the quantity that naturally appears in the analysis.

We can now state our main result of this section.

Theorem 31. *If we run CoLoWr with a ξ-stable base K-armed bandit algorithm α and an i.i.d. sequence $\boldsymbol{B} = (B_t)_{t\in\mathbb{N}}$ of Bernoulli random variables with bias $\beta \in (0,1)$ (independent of the randomization of α), then, for any time horizon $T \geqslant 2d+1$, the regret R_T satisfies*

$$R_T \leqslant 2d + \frac{2d+1}{d+1}\left(3d + 2d\beta(1-\beta)^{2d}\xi T + \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}\right)$$

where $\mathcal{R}_{\lfloor T/(2d+1)\rfloor}$ is the worst-case regret after $\big\lfloor T/(2d+1)\big\rfloor$ rounds of α (for an adversarial setting with $[0,1]$-valued losses).

Proof. Fix an arbitrary horizon $T \geqslant 2d+1$ and an arm $i^\star \in [K]$. Let, for all* $t \geqslant 2d+1$,

$$c_t \coloneqq \mathbb{E}\big[\ell_t^\circ(I_{t-d},\dots,I_t) - \ell_t^\circ(i^\star,\dots,i^\star)\big]\,, \qquad a \coloneqq 2d+1\,, \qquad b \coloneqq T\,.$$

Applying the elementary identity in Lemma 28 (Appendix D.2) in step $(*)$ below, we obtain

$$\begin{aligned}
R_T &= \sum_{t=1}^{2d} c_t + \sum_{t=2d+1}^{T} c_t \leqslant \sum_{t=1}^{2d} \mathbb{E}\big[\ell_t(I_t)\big] + \sum_{t=2d+1}^{T} c_t \leqslant 2d + \sum_{t=2d+1}^{T} c_t \\
&= 2d + \frac{1}{d+1}\left(\sum_{t=a-d}^{a-1}(t-a+d+1)\,c_t + (d+1)\sum_{t=a}^{b} c_t + \sum_{t=b+1}^{b+d}(b+d+1-t)\,c_t\right) \\
&\quad - \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)\,c_t - \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)\,c_t \\
&\overset{(*)}{=} 2d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t - \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)c_t - \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)c_t \\
&\leqslant 2d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t + \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)\ell_t^\circ(i^\star,\dots,i^\star) \\
&\quad + \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)\ell_t^\circ(i^\star,\dots,i^\star) \eqqcolon (\blacktriangledown)
\end{aligned}$$

Now, applying Lemma 12 in steps $(\circ)$ below, we get

$$\begin{aligned}
(\blacktriangledown) &\overset{(\circ)}{\leqslant} 2d + 2\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t \\
&= 2d + 2\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{t=\tau-d}^{\tau}\mathbb{E}\big[\ell_t^\circ(I_{t-d},\dots,I_t) - \ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d})\big] \\
&\quad + \frac{2d+1}{d+1}\mathbb{E}\left[\sum_{\tau=2d+1}^{T+d}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\big(\ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d}) - \ell_t^\circ(i^\star,\dots,i^\star)\big)\right] \\
&\overset{(\circ)}{\leqslant} 2d + 3\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{t=\tau-d}^{\tau}\mathbb{E}\big[\ell_t^\circ(I_{t-d},\dots,I_t) - \ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d})\big] \\
&\quad + \frac{2d+1}{d+1}\mathbb{E}\left[\sum_{\tau=2d+1}^{T}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\big(\ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d}) - \ell_t^\circ(i^\star,\dots,i^\star)\big)\right] \\
&\eqqcolon 2d + 3\frac{2d+1}{d+1}d + (\mathrm{I}) + \frac{2d+1}{d+1}\times(\mathrm{II})\,.
\end{aligned}$$

We upper bound the two terms (I) and (II) separately. First, let $\mathcal{U}$ be the (random) set of update rounds

$$\mathcal{U} \coloneqq \big\{\tau \in [T] : \tau \text{ is an update round (w.r.t. } \boldsymbol{B})\big\}\,.$$

*Here, we refer to the infinite sequence of losses $(\ell_t^{(s)})_{s\in\{0,\dots,d\},t\in\mathbb{N}}$ (and the respective composite losses $(\ell_t^\circ)_{t\in\mathbb{N}}$). If the problem is formalized only with a finite sequence $(\ell_t^{(s)})_{s\in\{0,\dots,d\},t\in[T]}$, it is sufficient to define an arbitrary sequence $(\ell_t^{(s)})_{s\in\{0,\dots,1\},t>T}$ with $\ell_t^{(s)} \geqslant 0$ and $\sum_{s=0}^{d}\ell_t^{(s)} \leqslant 1$ (and the corresponding composite losses $(\ell_t^\circ)_{t>T}$) and proceed as we do. This trick is needed to invoke Lemma 12, for which it is handy to sum d rounds into the future.

For the first term (I), we have

$$\begin{aligned}
\text{(I)} &= \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\sum_{t=\tau-d}^{\tau}\sum_{s=0}^{d}\ell_{t-s}^{(s)}(i)\big(\boldsymbol{p}_{t-s}(i)-\boldsymbol{p}_{\tau-2d}(i)\big)\right] \\
&\leqslant \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\max_{t\in\{\tau-d,\dots,\tau\},s\in\{0,\dots,d\}}\big(\boldsymbol{p}_{t-s}(i)-\boldsymbol{p}_{\tau-2d}(i)\big)\right]\sum_{t=\tau-d}^{\tau}\ell_t^{\circ}(i,\dots,i) \\
&\overset{(\circ)}{\leqslant} \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\max_{t\in\{\tau-d,\dots,\tau\},s\in\{0,\dots,d\}}\big(\boldsymbol{p}_{t-s}(i)-\boldsymbol{p}_{\tau-2d}(i)\big)\right] \\
&\overset{(\heartsuit)}{=} \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\mathbb{I}\left\{\bigcup_{\sigma=\tau-2d}^{\tau-1}\{\sigma\in\mathcal{U}\}\right\}\big(\boldsymbol{p}_{\tau}(i)-\boldsymbol{p}_{\tau-2d}(i)\big)^+\right] \\
&= \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\sum_{\sigma=\tau-2d}^{\tau-1}\mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\}\big(\boldsymbol{p}_{\tau}(i)-\boldsymbol{p}_{\tau-2d}(i)\big)^+\right] =: (\blacklozenge)
\end{aligned}$$

where $(\circ)$ follows by Lemma 12 together with the fact that the max in the previous line is always nonnegative (to see this, simply observe that picking $t=\tau-d$ and $s=d$ within the max makes $\boldsymbol{p}_{t-s}(i)=\boldsymbol{p}_{\tau-2d}(i)$), and $(\heartsuit)$ follows by the facts that the max in the previous line is always greater than or equal to zero, it can be strictly positive only if there is an update in a round σ with $\tau-2d\leqslant\sigma\leqslant\tau-1$, and there can be at most a single update in $2d+1$ consecutive time steps. Now, using the facts that $\boldsymbol{p}_\tau=\boldsymbol{q}_{n_\tau}$ and that $n_\tau=n_{\tau-2d}+1$ on the event $\{\sigma\in\mathcal{U}\}$, we get

$$\begin{aligned}
(\blacklozenge) &= \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\sum_{\sigma=\tau-2d}^{\tau-1}\mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\}\big(\boldsymbol{q}_{n_{\tau-2d}+1}(i)-\boldsymbol{q}_{n_{\tau-2d}}(i)\big)^+\right] \\
&= \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\sum_{\sigma=\tau-2d}^{\tau-1}\sum_{n\in\mathbb{N}}\mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\}\big(\boldsymbol{q}_{n+1}(i)-\boldsymbol{q}_n(i)\big)^+\right] \\
&\overset{(\clubsuit)}{=} \frac{2d+1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{\sigma=\tau-2d}^{\tau-1}\sum_{n\in\mathbb{N}}\mathbb{E}\big[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\}\big]\,\mathbb{E}\left[\sum_{i\in[K]}\big(\boldsymbol{q}_{n+1}(i)-\boldsymbol{q}_n(i)\big)^+\right] \\
&\overset{(\spadesuit)}{\leqslant} \frac{2d+1}{d+1}\xi\sum_{\tau=2d+1}^{T+d}\sum_{\sigma=\tau-2d}^{\tau-1}\sum_{n\in\mathbb{N}}\mathbb{E}\big[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\}\big] \\
&= \frac{2d+1}{d+1}\,\xi\sum_{\tau=2d+1}^{T+d}\sum_{\sigma=\tau-2d}^{\tau-1}\mathbb{P}[\sigma\in\mathcal{U}] \leqslant \frac{2d+1}{d+1}2d\beta(1-\beta)^{2d}\xi T
\end{aligned}$$

where $(\clubsuit)$ follows by the independence of the Bernoulli sequence $\boldsymbol{B}$ and the randomization of base algorithm α; $(\spadesuit)$ by the ξ-stability of α; and the last inequality by the fact that the probability that a time step σ is an update round is 0 if $\sigma\leqslant 2d$ and $\beta(1-\beta)^{2d}$ otherwise.

For the second term (II), set for brevity

$$r_\tau\colon[K]\to\mathbb{R}\,,\qquad i\mapsto\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\big(\ell_t^{\circ}(i,\dots,i)-\ell_t^{\circ}(i^\star,\dots,i^\star)\big)$$

for all $\tau \geqslant 2d+1$. We first note that

$$\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_\tau)\right] = \mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_{\tau-2d})\right] = \mathbb{E}\left[\sum_{\tau=2d+1}^{T} r_\tau(I_{\tau-2d})B_\tau\prod_{s=1}^{2d}(1-B_{\tau-s})\right]$$
$$= \beta(1-\beta)^{2d}\sum_{\tau=2d+1}^{T}\mathbb{E}\big[r_\tau(I_{\tau-2d})\big]$$

where the first identity is a consequence of the fact that if $\tau \in \mathcal{U}$, then the action I_τ played by Algorithm 10 at round τ coincides with the actions played in the previous $2d$ rounds, i.e., $I_\tau = \cdots = I_{\tau-2d}$, while the last equality follows by the independence of $I_{\tau-2d}$ and the vector of Bernoulli random variables $(B_\tau, \ldots, B_{\tau-2d})$. Thus,

$$\text{(II)} = \mathbb{E}\left[\sum_{\tau=2d+1}^{T} r_\tau(I_{\tau-2d})\right] = \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_\tau)\right]$$
$$= \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\ell_t^\circ(I_\tau,\ldots,I_\tau) - \sum_{\tau\in\mathcal{U}}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\ell_t^\circ(i^\star,\ldots,i^\star)\right] \tag{5.2}$$

Now define for any $\tau \geqslant 2d+1$,

$$\widetilde{\ell}_\tau \colon [K] \to [0,1]\,, \quad i \mapsto \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\ell_t^\circ(i,\ldots,i)\,.$$

(Note that $\widetilde{\ell}_\tau(i) \in [0,1]$ for all $i \in [K]$ by Lemma 12.) Leveraging again the fact that $\tau \in \mathcal{U}$ implies $I_\tau = \cdots = I_{\tau-2d}$, yields, for each $\tau \in \mathcal{U}$,

$$\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\ell_t^\circ(I_{t-d},\ldots,I_t) = \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\ell_t^\circ(I_\tau,\ldots,I_\tau) = \widetilde{\ell}_\tau(I_\tau)\,.$$

This shows that the loss Algorithm 10 feeds the base algorithm α (in Line 10) at each update round τ is a bandit feedback (for α) for the arm I_τ with respect to the $[0,1]$-valued loss $\widetilde{\ell}_\tau$. Therefore, by Equation (5.2) and the regret guarantees of the base algorithm, we obtain

$$\text{(II)} = \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}}\widetilde{\ell}_\tau(I_\tau) - \sum_{\tau\in\mathcal{U}}\widetilde{\ell}_\tau(i^\star)\right] \leqslant \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}$$

where in the last inequality we also used the monotonicity of the worst-case regret $\tau \mapsto \mathcal{R}_\tau$ and the fact that $|\mathcal{U}| \leqslant \left\lceil \frac{T-(2d+1)+1}{2d+1} \right\rceil \leqslant \left\lfloor \frac{T}{2d+1} \right\rfloor$.

In conclusion, we have

$$R_T \leqslant 2d + 3\frac{2d+1}{d+1}d + \text{(I)} + \frac{2d+1}{d+1}\cdot\text{(II)}$$
$$\leqslant 2d + \frac{2d+1}{d+1}\left(3d + 2d\beta(1-\beta)^{2d}\xi T + \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}\right)$$

hence concluding the proof. □

We can derive corollaries for various algorithms using Theorem 31. Consider for instance as

a base algorithm α the well-known Exp3 algorithm of [18]. It can be easily shown that Exp3 is η-stable (where η is its learning rate, see Lemma 31 in Appendix D.4). Combining this fact with its worst-case regret bound $\mathcal{R}_N \leqslant \frac{\ln K}{\eta} + \frac{\eta}{2}KN$, we obtain that the horizon-T regret of the CoLoWr algorithm with $\alpha = \mathrm{Exp3}(\eta)$, $\eta = \sqrt{\frac{d \ln K}{KT}}$, and an i.i.d. sequence $\boldsymbol{B}$ of Bernoulli random variables with bias $\beta = \frac{1}{2d+1}$ (independent of the randomization of α), satisfies $R_T = O\big(\sqrt{(d+1)KT \log K}\big)$.

Using Follow the Regularized Leader (FTRL) with $\frac{1}{2}$-Tsallis Entropy (Algorithm 11), we can remove the $\log K$ term in the above bound.[†]

Algorithm 11 Follow The Regularized Leader (FTRL) with (1/2)-Tsallis entropy

input: learning rate $\eta \geqslant 0$
initialization: $\widehat{L}_0 = 0$
for round $n = 1, 2, \ldots$ **do**
Play action J_n drawn according to

$$\boldsymbol{q}_n \in \operatorname*{arg\,min}_{\boldsymbol{q} \in \Delta_K} \left(\sum_{i \in [K]} \widehat{L}_{n-1}(i) \boldsymbol{q}(i) - 2\eta \sum_{i \in [K]} \sqrt{\boldsymbol{q}(i)} \right)$$

where Δ_K is the probability simplex in $\mathbb{R}^K$
Observe loss $\ell_n(J_n)$ and update $\widehat{L}_n = \widehat{L}_{n-1} + \widehat{\ell}_n$, where

$$\widehat{\ell}_n(i) = \frac{\ell_n(i)}{\boldsymbol{q}_n(i)} \mathbb{I}\{J_n = i\} \qquad \forall i \in [K]$$

However, we still need to prove the stability of Algorithm 11. This is established by the following result, whose (non-trivial) proof is given in Appendix D.3.

Theorem 32. *Algorithm 11 run with any learning rate $\eta > 0$ is ξ-stable, with*

$$\xi \leqslant 2 \frac{1 + \ln K}{\eta}$$

The worst-case regret guarantees of FTRL with $\frac{1}{2}$-Tsallis entropy are as follows.[‡]

Theorem 33 (Abernethy et al. 1)**.** *For each $N \in \mathbb{N}$, the worst-case regret $\mathcal{R}_N$ after N rounds of Algorithm 11 run with learning rate $\eta = \sqrt{\frac{N}{2}}$ (in an adversarial setting with $[0,1]$-valued losses) satisfies*

$$\mathcal{R}_N \leqslant 2\sqrt{2KN}$$

Combining Theorems 31 to 33, we obtain the following regret bound for composite losses.

Corollary 2. *For any time horizon $T \in \mathbb{N}$, if we run CoLoWr using:*

- *As α, Algorithm 11 with learning rate $\eta = \sqrt{\frac{1}{2}\lfloor T/(2d+1) \rfloor}$*
- *As $\boldsymbol{B}$, an i.i.d. sequence of Bernoulli random variables with bias $\beta = \frac{1}{2d+1}$ (independent of the randomization of α)*

[†]The arg min where Algorithm 11 picks each distribution $\boldsymbol{q}_n$ is always a singleton if $\eta > 0$. This is a consequence of Lemma 29, in Appendix D.3. For the sake of convenience, we also allow the learning rate $\eta = 0$, corresponding to a (non-regularized) Follow The Leader algorithm. In this case, multiple minimizers could exist but, for the sake of our results, ties could be broken in any (measurable) way.

[‡]The analysis of [1] is presented for a more general class of algorithms. A straightforward application of their Corollary 3.2 shows the validity of Theorem 33 for Algorithm 11.

then, its regret satisfies

$$R_T \leqslant c\sqrt{(d+1)KT}$$

where $c = 2\sqrt{2}$ *if* $d = 0$*, and* $c = 28$ *otherwise.*

Proof. If $d = 0$, then $\beta = 1$ and Algorithm 10 reduces to the base algorithm. The result in this case is therefore implied immediately by Theorem 33.

For the second part, we can assume without loss of generality that $T \geqslant 2d + 1$, so that $\eta > 0$. Then, plugging $\xi \leqslant 2\frac{1+\ln K}{\eta}$ (Theorem 32) and $\mathcal{I}_{\lfloor T/(2d+1)\rfloor} \leqslant 2\sqrt{2K\lfloor T/(2d+1)\rfloor}$ (Theorem 33) into Theorem 31 gives the result. □

5.5 Lower Bound

In this section we derive a lower bound for bandits with composite anonymous feedback. We do that through a reduction from the setting of linear bandits (in the probability simplex) to our setting. This reduction allows us to upper bound the regret of a linear bandit algorithm in terms of (a suitably scaled version of) the regret of an algorithm in our setting. Since the reduction applies to any instance of a linear bandit problem, we can use a known lower bound for the linear bandit setting to derive a corresponding lower bound for our composite setting.

Let Δ_K be the probability simplex in $\mathbb{R}^K$. At each round t, an algorithm A for linear bandit optimization chooses an action $\boldsymbol{q}_t \in \Delta_K$ and suffers loss $\boldsymbol{\ell}_t^\top \boldsymbol{q}_t$, where $\boldsymbol{\ell}_t \in [0,1]^K$ is some unknown loss vector. The feedback observed by the algorithm at the end of round t is the scalar $\boldsymbol{\ell}_t^\top \boldsymbol{q}_t$. The regret suffered by algorithm A playing actions $\boldsymbol{q}_1, \dots, \boldsymbol{q}_T$ is

$$R_T^{\text{lin}} = \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q}_t - \min_{\boldsymbol{q}\in\Delta_K} \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q} = \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q}_t - \min_{i=1,\dots,K} \sum_{t=1}^T \boldsymbol{\ell}^\top \boldsymbol{e}_i \tag{5.3}$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_K$ are the elements of the canonical basis of $\mathbb{R}^K$ and we used the fact that a linear function on the simplex is minimized at one of the corners. Let $\mathcal{R}_T^{\text{lin}}(A, \Delta_K)$ denote the worst case regret (over the oblivious choice of $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$) of algorithm A. Similarly, let $\mathcal{R}_T(A_d, K, d)$ be the worst case regret (over the oblivious choice of loss components $\ell_t^{(s)}(i)$ for all t, s, and i) of algorithm A_d for nonstochastic K-armed bandits with d-delayed composite anonymous feedback. For the sake of clarity, we assume below that the time horizon T is a multiple of $d + 1$. If this is not the case, we can straightforwardly stop at the highest multiple of $d + 1$ (smaller than T) up to paying an additive $O(d + 1)$ regret. Our reduction shows the following.

Lemma 13. *For any algorithm* A_d *for* K*-armed bandits with* d*-delayed composite anonymous feedback, there exists an algorithm* A *for linear bandits in* Δ_K *such that* $\mathcal{R}_T(A_d, K, d) \geqslant (d + 1)\,\mathcal{R}_{T/(d+1)}^{\text{lin}}(A, \Delta_K)$.

Our reduction, described in detail in the proof of the above lemma (see Appendix D.5), essentially builds the probability vectors $\boldsymbol{q}_t$ played by A based on the empirical distribution of actions played by A_d during blocks of size $d + 1$. Now, an additional lemma is needed (whose proof is also given in the Appendix D.5).

Lemma 14. *The regret of any algorithm* A *for linear bandits on* Δ_K *satisfies* $\mathcal{R}_T^{\text{lin}}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{KT}\big)$.

In the previous lemma, as well as the following result, the $\widetilde{\Omega}$ notation is only hiding a $\sqrt{\log T}$ denominator. Using the two lemmas above we can finally prove the lower bound.

Theorem 34. *For any algorithm A_d for K-armed bandits with d-delayed composite anonymous feedback, $\mathcal{R}_T(A_d, K, d) = \widetilde{\Omega}\big(\sqrt{(d+1)KT}\big)$.*

Proof. Fix an algorithm A_d. Using the reduction of Lemma 13 gives an algorithm A such that $\mathcal{R}_T(A_d, K, d) \geqslant (d+1)\,\mathcal{R}^{\mathrm{lin}}_{T/(d+1)}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{(d+1)KT}\big)$, where we used Lemma 14 with horizon $T/(d+1)$ to prove the $\widetilde{\Omega}$-equality. □

Although the loss sequence used to prove the lower bound for linear bandits in the simplex is stochastic i.i.d., the loss sequence achieving the lower bound in our delayed setting is not independent due to the deterministic loss transformation in the proof of Lemma 13 (which is defined independently of the algorithm, thus preserving the oblivious nature of the adversary).

5.6 Conclusions

In this final chapter, we investigated the setting of d-delayed composite anonymous feedback as applied to nonstochastic bandits. Composite anonymous feedback lends itself to formalize scenarios where the actions performed by the online decision-maker produce long-lasting effects that combine additively over time. A general reduction technique was introduced that enables the conversion of a stable algorithm working in a standard bandit framework into one working in the composite feedback framework. Applying our reduction to the FTRL algorithm with Tsallis entropy, we obtain an upper bounded on the regret of order $\sqrt{(d+1)KT}$, which we showed to be optimal.

Bibliography

[1] Jacob D. Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[2] Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Fast rate learning in stochastic first price bidding. In *ACML*, volume 157 of *Proceedings of Machine Learning Research*, pages 1754–1769. PMLR, 2021.

[3] Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Efficient algorithms for stochastic repeated second-price auctions. In *Algorithmic Learning Theory*, pages 99–150. PMLR, 2021.

[4] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Explore/exploit schemes for web content optimization. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1–10. IEEE, 2009.

[5] Palvi Aggarwal, Marcus Gutierrez, Christopher D. Kiekintveld, Branislav Bosanskỳ, and Cleotilde Gonzalez. Evaluating adaptive deception strategies for cyber defense with human adversaries. *Game Theory and Machine Learning for Cyber Security*, pages 77–96, 2021.

[6] Priyank Agrawal and Theja Tulabandula. Learning by repetition: Stochastic multi-armed bandits under priming effect. In *Conference on Uncertainty in Artificial Intelligence*, pages 470–479. PMLR, 2020.

[7] Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics and Computation, EC'14*, pages 989–1006, New York, NY, USA, 2014. ACM, Association for Computing Machinery.

[8] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 23–35. JMLR.org, 2015.

[9] Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM J. Comput.*, 46(6):1785–1826, 2017.

[10] Mawulolo K. Ameko, Miranda L. Beltzer, Lihua Cai, Mehdi Boukhechba, Bethany A. Teachman, and Laura E. Barnes. Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 249–258, 2020.

[11] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *NIPS*, pages 1169–1177, 2013.

[12] Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

[13] Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[14] Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proc. 29th ICML*, 2012.

[15] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[16] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

[17] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[18] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[19] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 454–468. Springer, 2007.

[20] Yossi Azar, Amos Fiat, and Federico Fusco. An α-regret analysis of adversarial bilateral trade. In *NeurIPS*, 2022.

[21] Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):4, 2015.

[22] François Bachoc, Tommaso R. Cesari, Roberto Colomboni, and Andrea Paudice. A near-optimal algorithm for univariate zeroth-order budget convex optimization. *arXiv preprint arXiv:2208.06720*, 2022.

[23] François Bachoc, Tommaso R. Cesari, Roberto Colomboni, and Andrea Paudice. Regret analysis of dyadic search. *arXiv preprint arXiv:2209.00885*, 2022.

[24] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3):13:1–13:55, 2018.

[25] Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. Learning to bid in contextual first price auctions. In *Proceedings of the ACM Web Conference 2023*, pages 3489–3497, 2023.

[26] Martin N. Baily. Some aspects of optimal unemployment insurance. *Journal of Public Economics*, 10(3):379–402, 1978.

[27] Santiago R. Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab S. Mirrokni, and Jon Schneider. Contextual bandits with cross-learning. *NeurIPS*, pages 9676–9685, 2019.

[28] Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

[29] Richard F. Bass. *Real analysis for graduate students*. Createspace Ind Pub, USA, 2013.

[30] Dirk Bergemann and Johannes Hörner. Should first-price auctions be transparent? *American Economic Journal: Microeconomics*, 10(3):177–218, 2018.

[31] Donald A. Berry. The application of two-armed bandit strategies to clinical trials. Technical report, University of Minnesota, 1976.

[32] Donald A. Berry. Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association*, 73(362):339–345, 1978.

[33] Jason Bigler. Rolling out first price auctions to Google Ad Manager partners. `https://www.blog.google/products/admanager/rolling-out-first-price-auctions-google-ad-manager-partners/`, 2019. Accessed April 7, 2023.

[34] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons: New York, 1995.

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

[36] Avrim Blum and Jason D. Hartline. Near-optimal online auctions. In *ACM-SIAM Symposium on Discrete Algorithms, SODA'05*, pages 1156–1163, USA, 2005. Society for Industrial and Applied Mathematics, Society for Industrial and Applied Mathematics.

[37] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2-3):137–146, 2004.

[38] Liad Blumrosen and Shahar Dobzinski. Reallocation mechanisms. In *EC*, page 617. ACM, 2014.

[39] Liad Blumrosen and Shahar Dobzinski. (almost) efficient mechanisms for bilateral trading. *Games Econ. Behav.*, 130:369–383, 2021.

[40] Natasa Bolić, Tommaso R. Cesari, and Roberto Colomboni. An online learning theory of brokerage. *arXiv preprint arXiv:2310.12107*, 2023.

[41] Djallel Bouneffouf, Irina Rish, and Guillermo A. Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *Artificial General Intelligence: 10th International Conference, 2017, Proceedings 10*, pages 237–248. Springer, 2017.

[42] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

[43] Alexander Braun and Thomas Kesselheim. Truthful mechanisms for two-sided markets via prophet inequalities. In *ACM Conference on Economics and Computation (EC)*, pages 202–203, 2021.

[44] Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

[45] Johannes Brustle, Yang Cai, Fa Wu, and Mingfei Zhao. Approximating gains from trade in two-sided markets via simple mechanisms. In *EC*, pages 589–590. ACM, 2017.

[46] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.*, 5(1):1–122, 2012.

[47] Sébastien Bubeck, Nikhil R. Devanur, Zhiyi Huang, and Rad Niazadeh. Online auctions and multi-scale online learning. In *EC*, pages 497–514. ACM, 2017.

[48] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

[50] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Trans. Inf. Theory*, 61(1):549–564, 2015.

[51] Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622, 2016.

[52] Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, pages 465–481. PMLR, 2017.

[53] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773. PMLR, 2018.

[54] Nicolò Cesa-Bianchi, Tommaso R. Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In *ALT*, volume 98 of *Proceedings of Machine Learning Research*, pages 247–273. PMLR, 2019.

[55] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 289–309, New York, NY, USA, 2021. Association for Computing Machinery.

[56] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. *Journal of Machine Learning Research (JMLR)*, 23(1-24), 2022.

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

[58] Nicolò Cesa-Bianchi, Tommaso R. Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. *arXiv preprint arXiv:2307.09478*, 2023.

[59] Nicolò Cesa-Bianchi, Roberto Colomboni, and Maximilian Kasy. Adaptive maximization of social welfare. *arXiv preprint arXiv:2310.09597*, 2023.

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

[61] Tommaso R. Cesari and Roberto Colomboni. A nearest neighbor characterization of Lebesgue points in metric measure spaces. *Mathematical Statistics and Learning*, 3(1):71–112, 2021.

[62] Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105, 2014.

[63] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[64] Ningyuan Chen and Guillermo Gallego. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69(3):974–984, 2021.

[65] Raj Chetty. Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics*, 1(1):451–488, 2009.

[66] Maxime C. Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Manag. Sci.*, 66(11):4921–4943, 2020.

[67] Riccardo Colini-Baldeschi, Bart de Keijzer, Stefano Leonardi, and Stefano Turchetta. Approximately efficient double auctions with strong budget balance. In *SODA*, pages 1424–1443. SIAM, 2016.

[68] Riccardo Colini-Baldeschi, Paul W. Goldberg, Bart de Keijzer, Stefano Leonardi, and Stefano Turchetta. Fixed price approximability of the optimal gain from trade. In *WINE*, volume 10660 of *Lecture Notes in Computer Science*, pages 146–160. Springer, 2017.

[69] Roberto Colomboni, Emmanuel Esposito, and Andrea Paudice. An improved uniform convergence bound with fat-shattering dimension. *arXiv preprint arXiv:2307.06644*, 2023.

[70] Thomas Cover. Behavior of sequential predictors of binary sequences. In *Proc. of the 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, 1965.

[71] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2006.

[72] Constantinos Daskalakis and Vasilis Syrgkanis. Learning in auctions: Regret is hard, envy is easy. In *FOCS*, pages 219–228. IEEE Computer Society, 2016.

[73] Liad Dekel, Ilia Leybovich, Polina Zilberman, and Rami Puzis. Mabat: A multi-armed bandit approach for threat-hunting. *IEEE Transactions on Information Forensics and Security*, 18: 477–490, 2022.

[74] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Online learning with composite loss functions. In *Conference on Learning Theory*, pages 1214–1231, 2014.

[75] Ofer Dekel, Elad Hazan, and Tomer Koren. The blinded bandit: Learning with adaptive feedback. *Advances in Neural Information Processing Systems*, 27:1610–1618, 2014.

[76] Arnoud V. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

[77] Arnoud V. den Boer and N. Bora Keskin. Discontinuous demand functions: estimation and pricing. *Management Science*, 66(10):4516–4534, 2020.

[78] Xiaotie Deng, Xinyan Hu, Tao Lin, and Weiqiang Zheng. Nash convergence of mean-based learning algorithms in first price auctions. In *Proceedings of the ACM Web Conference 2022*, pages 141–150, 2022.

[79] Yuan Deng, Jieming Mao, Balasubramanian Sivan, and Kangning Wang. Approximately efficient bilateral trade. *CoRR*, abs/2111.03611, 2021.

[80] Nikhil R. Devanur, Yuval Peres, and Balasubramanian Sivan. Perfect bayesian equilibria in repeated sales. *Games Econ. Behav.*, 118:570–588, 2019.

[81] Nishanth Dikkala and Éva Tardos. Can credit increase revenue? In *WINE*, volume 8289 of *Lecture Notes in Computer Science*, pages 121–133. Springer, 2013.

[82] Kaize Ding, Jundong Li, and Huan Liu. Interactive anomaly detection on attributed networks. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 357–365, 2019.

[83] Alexey Drutsa. Weakly consistent optimal pricing algorithms in repeated posted-price auctions with strategic buyer. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1318–1327. PMLR, 2018.

[84] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

[85] Naveen Durvasula, Nika Haghtalab, and Manolis Zampetakis. Smoothed analysis of online non-parametric auctions. *EC*, 2023.

[86] Paul Dütting, Federico Fusco, Philip Lazos, Stefano Leonardi, and Rebecca Reiffenhäuser. Efficient two-sided markets with limited information. In *STOC*, pages 1452–1465. ACM, 2021.

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

[88] Yumou Fei. Improved approximation to first-best gains-from-trade. In *Web and Internet Economics*, pages 204–218, Cham, 2022. Springer International Publishing. ISBN 978-3-031-22832-2.

[89] Michal Feldman, Brendan Lucier, and Noam Nisan. Correlated and coarse equilibria of single-item auctions. In *WINE*, volume 10123 of *Lecture Notes in Computer Science*, pages 131–144. Springer, 2016.

[90] Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to bid without knowing your value. In *EC*, pages 505–522. ACM, 2018.

[91] Zhe Feng, Guru Guruganesh, Christopher Liaw, Aranyak Mehta, and Abhishek Sethi. Convergence analysis of no-regret bidding algorithms in repeated auctions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5399–5406, 2021.

[92] Bank for International Settlements. Otc derivatives statistics at end-june 2022, 2022. URL `https://www.bis.org/publ/otc_hy2211.pdf`.

[93] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[94] Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

[95] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[96] Scott Garrabrant, Nate Soares, and Jessica Taylor. Asymptotic convergence in online learning with unbounded delays. *arXiv preprint arXiv:1604.05280*, 2016.

[97] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[98] Ramki Gummadi, Peter B. Key, and Alexandre Proutiere. Optimal bidding strategies in dynamic auctions with budget constraints. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 588–588. IEEE, 2011.

[99] Hongbo Guo, Ruben Naeff, Alex Nikulkov, and Zheqing Zhu. Optimism based exploration in large-scale recommender systems. *arXiv preprint arXiv:2304.02572*, 2023.

[100] Kathleen M. Hagerty and William P. Rogerson. Robust trading mechanisms. *Journal of Economic Theory*, 42(1):94–107, 1987.

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

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

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

[104] Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*, 2020.

[105] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*, 2020.

[106] Elad Hazan. *Introduction to online convex optimization*. MIT Press, 2022.

[107] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *J. Artif. Intell. Res.*, 55:317–359, 2016.

[108] Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

[109] Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

[110] Pooria Joulani, András György, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *AAAI*, volume 16, pages 1744–1750, 2016.

[111] Zi Yang Kang, Francisco Pernice, and Jan Vondrák. Fixed-price approximations in bilateral trade. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2964–2985, 2022.

[112] Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

[113] Maximilian Kasy. Optimal taxation and insurance using machine learning – sufficient statistics and beyond. *Journal of Public Economics*, 167, 2018.

[114] Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

[115] Daniel Khashabi, Kent Quanrud, and Amirhossein Taghvaei. Adversarial delays in online strongly-convex optimization. *arXiv preprint arXiv:1605.06201*, 2016.

[116] Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS*, pages 697–704, 2004.

[117] Robert D. Kleinberg and Frank Thomson Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *FOCS*, pages 594–605. IEEE Computer Society, 2003.

[118] Yoav Kolumbus and Noam Nisan. Auctions between regret-minimizing agents. In *WWW*, pages 100–111. ACM, 2022.

[119] Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The Hedge algorithm on a continuum. In *International Conference on Machine Learning*, pages 824–832. PMLR, 2015.

[120] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, pages 1091–1114, 1987.

[121] John Langford, Alexander J. Smola, and Martin Zinkevich. Slow learners are fast. *Advances in Neural Information Processing Systems*, 22:2331–2339, 2009.

[122] Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 1547–1575. PMLR, 2022.

[123] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, UK, 2020.

[124] Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 993–1002. PMLR, 2019.

[125] Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

[126] Allen Liu, Renato Paes Leme, and Jon Schneider. Optimal contextual pricing and extensions. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1059–1078. SIAM, 2021.

[127] Gábor Lugosi, Mihalis G. Markakis, and Gergely Neu. On the hardness of learning from censored and nonstationary demand. *INFORMS Journal on Optimization*, 2023.

[128] Yiyun Luo, Will Wei Sun, and Yufeng Liu. Distribution-free contextual dynamic pricing. *Mathematics of Operations Research*, 2023.

[129] Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. Learning and efficiency in games with dynamic population. In *SODA*, pages 120–129. SIAM, 2016.

[130] Odalric-Ambrym Maillard and Rémi Munos. Online learning in adversarial lipschitz environments. In *ECML/PKDD (2)*, volume 6322 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2010.

[131] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *AAAI*, pages 2849–2856, 2015.

[132] Timothy Arthur Mann, Sven Gowal, Andras Gyorgy, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *International Conference on Machine Learning*, pages 4324–4332. PMLR, 2019.

[133] Jieming Mao, Renato Leme, and Jon Schneider. Contextual pricing for lipschitz buyers. *Advances in Neural Information Processing Systems*, 31, 2018.

[134] Takesaki Masamichi. *Theory of Operator Algebras I*. Springer New York, NY, 1979.

[135] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[136] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

[137] Aditya Mate, Andrew Perrault, and Milind Tambe. Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *AAMAS*, pages 880–888, 2021.

[138] Chris Mesterharm. On-line learning with delayed label feedback. In *Algorithmic Learning Theory*, pages 399–413. Springer, 2005.

[139] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70 (2):583–601, 2002.

[140] James A. Mirrlees. An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, pages 175–208, 1971.

[141] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Second Edition*. Cambridge University Press, 2017.

[142] Mehryar Mohri and Andres Munoz Medina. Optimal regret minimization in posted-price auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1871–1879, USA, 2014. NeurIPS.

[143] Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. *J. Econ. Theory*, 29(2):265–281, 1983.

[144] Michael Naaman. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters*, 173:109088, 2021.

[145] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23*, pages 1804–1812. Curran Associates, Inc., 2010.

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

[147] Jean Czerlinski Whitmore Ortega. Learning with insufficient data: a multi-armed bandit perspective on covid-19 interventions. *Mind & Society*, 21(2):183–193, 2022.

[148] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

[149] Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4105–4113. PMLR, 10–15 Jul 2018.

[150] Akshay Pilani, Kritagya Mathur, Himanshu Agrawald, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. Contextual bandit approach-based recommendation system for personalized web-based services. *Applied Artificial Intelligence*, 35(7):489–504, 2021.

[151] Xin Qiu and Risto Miikkulainen. Enhancing evolutionary conversion rate optimization via multi-armed bandit algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9581–9588, 2019.

[152] Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, pages 1270–1278, 2015.

[153] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, pages 1764–1772, 2011.

[154] Frank P. Ramsey. A contribution to the theory of taxation. *The economic journal*, 37(145): 47–61, 1927.

[155] Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18:241–252, 2020.

[156] Daniel Reem, Simeon Reich, and Alvaro De Pierro. Re-examination of bregman functions and new properties of their divergences. *Optimization*, 68(1):279–348, 2019.

[157] Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(6):527–535, 1952.

[158] Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68 (6):1625–1647, 2020.

[159] Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229, 2001.

[160] Emmanuel Saez. Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, 117(3):1039–1073, 2002.

[161] Emmanuel Saez and Stefanie Stantcheva. Generalized social welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45, 2016. URL `https://dl.dropboxusercontent.com/u/12222201/Saez_Stantcheva_GSWW.pdf`.

[162] Eric M. Schwartz, Eric T. Bradlow, and Peter S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

[163] Steven L. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

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

[165] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[166] Ohad Shamir. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pages 1523–1551, 2015.

[167] Ohad Shamir and Liran Szlak. Online learning with local permutations and delayed feedback. In *Proc. 34th ICML*, 2017.

[168] Katerina Sherstyuk, Krit Phankitnirundorn, and Michael J. Roberts. Randomized double auctions: gains from trade, trader roles, and price discovery. *Experimental Economics*, 24(4): 1–40, 2020.

[169] Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12 (1-2):1–286, 2019.

[170] Aleksandrs Slivkins and Assaf Zeevi. Dynamic Pricing Under Model Uncertainty. Tutorial given at the 16th ACM Conference on Economics and Computation, 2015.

[171] Sarah Sluis. Big changes coming to auctions, as exchanges roll the dice on first-price. `https://adexchanger.com/platforms/big-changes-coming-auctions-exchanges-roll-dice-first-price/`, 2017. Accessed July 3, 2023.

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

[173] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[174] Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6541–6550, 2019.

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

[176] Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI*, 2017.

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

[178] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[179] Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[180] Yongquan Wan, Junli Xian, and Cairong Yan. A contextual multi-armed bandit approach based on implicit feedback for online recommendation. In *Knowledge Management in Organizations: 15th International Conference, KMO 2021, Kaohsiung, Taiwan, July 20-22, 2021, Proceedings 15*, pages 380–392. Springer, 2021.

[181] Siwei Wang, Haoyun Wang, and Longbo Huang. Adaptive algorithms for multi-armed bandit with composite and anonymous feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10210–10217, 2021.

[182] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory, COLT'16*, pages 1562–1583. PMLR, 2016.

[183] Pierre-Olivier Weill. The search theory of over-the-counter markets. *Annual Review of Economics*, 12:747–773, 2020.

[184] Marcelo J. Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

[185] David Williams. *Probability with martingales*. Cambridge University Press, 1991.

[186] Matt Wong. Moving AdSense to a first-price auction. `https://blog.google/products/ads-commerce/our-move-to-a-first-price-auction/`, 2021. Accessed July 6, 2023.

[187] Ding Xiang, Becky West, Jiaqi Wang, Xiquan Cui, and Jinzhou Huang. Adaptively optimize content recommendation using multi armed bandit algorithms in e-commerce. *arXiv preprint arXiv:2108.01440*, 2021.

[188] Ding Xiang, Rebecca West, Jiaqi Wang, Xiquan Cui, and Jinzhou Huang. Multi armed bandit vs. a/b tests in e-commerce-confidence interval and hypothesis test power perspectives. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4204–4214, 2022.

[189] Wei Zhang, Brendan Kitts, Yanjun Han, Zhengyuan Zhou, Tingyu Mao, Hao He, Shengjun Pan, Aaron Flores, San Gultekin, and Tsachy Weissman. MEOW: A space-efficient nonparametric bid shading algorithm. In *KDD*, pages 3928–3936. ACM, 2021.

[190] Wei Zhang, Yanjun Han, Zhengyuan Zhou, Aaron Flores, and Tsachy Weissman. Leveraging the hints: Adaptive bidding in repeated first-price auctions. *NeurIPS*, 2022.

[191] Yinglun Zhu, Sumeet Katariya, and Robert Nowak. Robust outlier arm identification. In *International Conference on Machine Learning*, pages 11566–11575. PMLR, 2020.

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

[193] Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28–1, 2021.

Appendix A

Online Learning in Bilateral Trade

A.1 Existence of the Best Price

In this section, we show that a price $p^\star$ maximizing the expected regret always exists.

Lemma 15. *If $T \in \mathbb{N}$ and $(S_1, B_1), \dots, (S_T, B_T)$ is a sequence of $[0,1]^2$-valued random variables, the function $p \mapsto \mathbb{E}\big[\sum_{t=1}^T \operatorname{gft}\big(p, (S_t, B_t)\big)\big]$ is upper semicontinuous. In particular, there exists a maximizer $p^\star \in [0,1]$.*

Proof. Let $t \in [T]$ and let U_t be a $[0,1]$-valued uniform random variable independent of (S_t, B_t). By the Decomposition lemma (2.4) and the fact that the sum of two upper semicontinuous functions is upper semicontinuous, it is sufficient to show that

$$f_t \colon \mathbb{R} \to [0,1],\ p \mapsto \mathbb{P}[S_t \leqslant p \leqslant U_t \leqslant B_t] \qquad \text{and} \qquad g_t \colon \mathbb{R} \to [0,1],\ p \mapsto \mathbb{P}[S_t \leqslant U_t \leqslant p \leqslant B_t]$$

are both upper semicontinuous in order to prove that $p \mapsto \mathbb{E}\big[\operatorname{gft}\big(p, (S_t, B_t)\big)\big]$ is upper semicontinuous. We now prove that f_t is upper semicontinuous, i.e., that for any $p \in \mathbb{R}$, we have

$$\limsup_{q \to p} f_t(q) \leqslant f_t(p)\ .$$

To do so, we show that for any $p \in \mathbb{R}$ and any two sequences $q_n \uparrow p$, $r_n \downarrow p$, we have that

$$\limsup_{q_n \uparrow p} f_t(q_n) \leqslant f_t(p) \qquad \text{and} \qquad \limsup_{r_n \downarrow p} f_t(r_n) \leqslant f_t(p)\ .$$

If $p \in \mathbb{R}\backslash[0,1]$, the result is trivially true. Thus, let $p \in [0,1]$, $q_n \uparrow p$ and $r_n \downarrow p$. Then,

$$\begin{aligned}
\mathbb{I}\{S_t \leqslant q_n \leqslant U_t \leqslant B_t\} &\to \mathbb{I}\{S_t < p \leqslant U_t \leqslant B_t\}\ , \qquad n \to \infty\ , \\
\mathbb{I}\{S_t \leqslant r_n \leqslant U_t \leqslant B_t\} &\to \mathbb{I}\{S_t \leqslant p < U_t \leqslant B_t\}\ , \qquad n \to \infty\ ,
\end{aligned}$$

pointwise everywhere. By Lebesgue's dominated convergence theorem, it follows that, if $n \to \infty$,

$$\begin{aligned}
f_t(q_n) &\to \mathbb{P}[S_t < p \leqslant U_t \leqslant B_t] \leqslant \mathbb{P}[S_t \leqslant p \leqslant U_t \leqslant B_t] = f_t(p)\ , \\
f_t(r_n) &\to \mathbb{P}[S_t \leqslant p < U_t \leqslant B_t] = \mathbb{P}[S_t \leqslant p \leqslant U_t \leqslant B_t] = f_t(p)\ .
\end{aligned}$$

Being p, $(q_n)_{n\in\mathbb{N}}$ and $(r_n)_{n\in\mathbb{N}}$ arbitrarily chosen, it follows that f_t is upper semicontinuous. Analo-

gously, one can prove that g_t is upper semicontinuous. Hence, $p \mapsto \mathbb{E}\big[\mathrm{gft}\big(p, (S_t, B_t)\big)\big] = f_t(p) + g_t(p)$ is an upper semicontinuous function. Being t arbitrarily chosen, the same conclusion holds for any $t \in [T]$. Hence, our target function is upper semicontinuous as well, since, by the linearity of the expectation, it is a sum of T upper semicontinuous functions. Finally, being our target function defined on the compact set $[0,1]$, it attains its maximum at some $p^\star \in [0,1]$ by the Weierstrass theorem. □

A.2 An Improved Analysis of Continuous Hedge

In what follows, we denote with $\mathcal{B}_{[0,1]}$, respectively $\mathcal{B}_{[0,+\infty]}$, the Borel σ-algebra of $[0,1]$, respectively $[0,+\infty]$, while $\mathcal{B}$ stands for the Borel σ-algebra of $\mathbb{R}$. For any any measurable function $g\colon [0,1] \to \mathbb{R}$, we denote with $\|g\|_1$ the integral with respect the Lebesgue measure of $|g|$ on $[0,1]$.

The following result implies directly theoretical guarantees for Hedge. We state the theorem in an abstract way to highlight that its claims are really about the properties of some stochastic processes rather than specific online learning protocols.

Theorem 35. *Let $(\mathcal{Y}, \mathcal{E}_\mathcal{Y})$ be a measurable space. Let $\rho\colon [0,1] \times \mathcal{Y} \to [0,1]$ be a $(\mathcal{E}_\mathcal{Y} \otimes \mathcal{B}_{[0,1]})/\mathcal{B}_{[0,1]}$-measurable function. Let $(X_t, Y_t)_{t\in\mathbb{N}}$ be a $[0,1] \times \mathcal{Y}$-valued stochastic process. For any $t \in \mathbb{N}$, let $\mathcal{H}_t = \sigma(X_1, Y_1, \dots, X_{t-1}, Y_{t-1})$ be the σ-algebra generated by the history up to the end of time $t-1$ (with the understanding that $\mathcal{H}_1 = \sigma\big(\{\varnothing\}\big)$). Let $M \geqslant 2$ and $\eta \in (0,1)$. Assume that:*

- *For any $t \in \mathbb{N}$, the conditional law $\mathbb{P}_{X_t|\mathcal{H}_t}$ of X_t given $\mathcal{H}_t$ admits as a density (w.r.t. the Lebesgue measure on $[0,1]$) the (random) function $f_t(\cdot) = \frac{\sum_{s=1}^{t-1} \exp\big(\eta\rho(\cdot, Y_s)\big)}{\int_{[0,1]} \sum_{s=1}^{t-1} \exp\big(\eta\rho(x, Y_s)\big)\,\mathrm{d}x}$ (for $t = 1$, $f_1 = \mathbb{I}_{[0,1]}$) .*
- *For any $t \in \mathbb{N}$, the two random variables X_t and Y_t are conditionally independent given $\mathcal{H}_t$.*
- *For any $t \in \mathbb{N}$, the function $[0,1] \to [0,1]$, $x \mapsto \mathbb{E}\big[\rho(x, Y_t)\big]$ is M-Lipschitz.*

Then, for any $T \in \mathbb{N}$,

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \leqslant \frac{1}{\eta} \ln\left(\frac{\eta T M}{1 - e^{-\eta T}}\right) + (e-2)\eta T \;.$$

In particular, if $\eta = \sqrt{\frac{\ln(2T)}{(e-2)T}}$ we have

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \leqslant \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln(M)}{\ln(2T)}\right) \;.$$

Proof. Define $W_1(x) = 1$ for all $x \in [0,1]$ and, for each $t \in \mathbb{N}$, define by induction $W_{t+1}(\cdot) = W_t(\cdot)\exp(\eta\rho(\cdot, Y_t))$. Then, denoting for any measurable function $g\colon [0,1] \to \mathbb{R}$, the integral with respect the Lebesgue measure of $|g|$ on $[0,1]$ by $\|g\|_1$, we have

$$\ln\big(\|W_{T+1}\|_1\big) = \ln\left(\prod_{t=1}^{T} \frac{\|W_{t+1}\|_1}{\|W_t\|_1}\right) = \sum_{t=1}^{T} \ln\left(\int_{[0,1]} \exp\big(\eta\rho(x, Y_t)\big) f_t(x)\,\mathrm{d}x\right)$$

$$
\begin{aligned}
&\leqslant \sum_{t=1}^{T} \ln\left(\int_{[0,1]} \Big(1 + \eta\rho(x,Y_t) + (e-2)\eta^2\big(\rho(x,Y_t)\big)^2\Big) f_t(x)\,\mathrm{d}x\right)\\
&= \sum_{t=1}^{T} \ln\left(1 + \int_{[0,1]} \Big(\eta\rho(x,Y_t) + (e-2)\eta^2\big(\rho(x,Y_t)\big)^2\Big) f_t(x)\,\mathrm{d}x\right)\\
&\leqslant \eta\sum_{t=1}^{T}\int_{[0,1]} \rho(x,Y_t) f_t(x)\,\mathrm{d}x + (e-2)\eta^2\sum_{t=1}^{T}\int_{[0,1]} \big(\rho(x,Y_t)\big)^2 f_t(x)\,\mathrm{d}x\\
&\leqslant \eta\sum_{t=1}^{T}\int_{[0,1]} \rho(x,Y_t) f_t(x)\,\mathrm{d}x + (e-2)\eta^2 T\\
&= \eta\sum_{t=1}^{T} \mathbb{E}\big[\rho(X_t,Y_t) \mid \sigma(Y_t,\mathcal{H}_t)\big] + (e-2)\eta^2 T\;,
\end{aligned}
$$

where the last equality follows from the Generalized Freezing Lemma (see Lemma 18 in Appendix A.4) noticing that, for each $t \in [T]$, Φ_t defined for each Borel subset $A \subset [0,1]$ via $\Phi_t[A] = \int_A f_t(x)\,\mathrm{d}x$ is a regular conditional probability for $\mathbb{P}_{X_t|\mathcal{H}_t}$ and $\int_{[0,1]} \rho(x,Y_t) f_t(x)\,\mathrm{d}x = \int_{[0,1]} \rho(x,Y_t)\,\mathrm{d}\Phi_t(x)$. Hence, using the tower rule,

$$
\mathbb{E}\left[\ln\big(\|W_{T+1}\|_1\big)\right] \leqslant \eta\mathbb{E}\left[\sum_{t=1}^{T}\rho(X_t,Y_t)\right] + (e-2)\eta^2 T\;.
$$

On the other hand, let $x^\star \in [0,1]$ be a point belonging to $\arg\max_{x\in[0,1]} \sum_{t=1}^{T} \mathbb{E}\big[\rho(x,Y_t)\big]$, which does exist due to the fact that this last sum, as a function of x, is MT-Lipschitz (hence continuous on the compact set $[0,1]$). Then, for any $x \in [0,1]$,

$$
\sum_{t=1}^{T}\mathbb{E}\big[\rho(x^\star,Y_t)\big] - \sum_{t=1}^{T}\mathbb{E}\big[\rho(x,Y_t)\big] \leqslant T\min\big(1, M|x-x^\star|\big)\;. \tag{A.1}
$$

Let X be a uniform random variable on $[0,1]$ independent of $Y_1,\ldots,Y_T$. It follows that

$$
\begin{aligned}
\mathbb{E}\Big[\ln\big(\|W_{T+1}\|_1\big)\Big] &= \mathbb{E}\left[\ln\left(\int_{[0,1]} \exp\left(\eta\sum_{t=1}^{T}\rho(x,Y_t)\right)\,\mathrm{d}x\right)\right]\\
&= \mathbb{E}\left[\ln\mathbb{E}\left[\exp\left(\eta\sum_{t=1}^{T}\rho(X,Y_t)\right) \mid X\right]\right]\\
&\geqslant \ln\mathbb{E}\left[\exp\left(\mathbb{E}\left[\eta\sum_{t=1}^{T}\rho(X,Y_t) \mid (Y_1,\ldots,Y_T)\right]\right)\right]\\
&= \ln\left(\int_{[0,1]} \exp\left(\mathbb{E}\left[\eta\sum_{t=1}^{T}\rho(x,Y_t)\right]\right)\,\mathrm{d}x\right)\\
&= \eta\sum_{t=1}^{T}\mathbb{E}\big[\rho(x^\star,Y_t)\big] + \ln\left(\int_{[0,1]} \exp\left(\eta\left(\sum_{t=1}^{T}\mathbb{E}\big[\rho(x,Y_t)\big] - \sum_{t=1}^{T}\mathbb{E}\big[\rho(x^\star,Y_t)\big]\right)\right)\,\mathrm{d}x\right)\\
&\geqslant \eta\sum_{t=1}^{T}\mathbb{E}\big[\rho(x^\star,Y_t)\big] + \ln\left(\int_{[0,1]} \exp\left(-\eta T\min(1, M|x-x^\star|)\right)\,\mathrm{d}x\right) = (\star)\;,
\end{aligned}
$$

where

- the second and the third equalities follow from the Freezing Lemma (see Lemma 17 in Appendix A.4).
- the first inequality follows from the log-exp analogous of Minkowski's integral inequality, in the form of Corollary 4, with $(\mathcal{V}, \mathcal{E}_{\mathcal{V}}) = \big([0,1], \mathcal{B}_{[0,1]}\big)$, $(\mathcal{W}, \mathcal{E}_{\mathcal{W}}) = (\mathcal{Y}^T, \otimes^T \mathcal{E}_{\mathcal{Y}})$, $V = X$, $W = (Y_1, \dots, Y_T)$, and $g\colon [0,1] \times \mathcal{Y}^T \to [0, +\infty]$, $\big(x, (y_1, \dots, y_T)\big) \mapsto \eta \sum_{t=1}^T \rho(x, y_t)$.
- the last inequality follows from Equation (A.1).

Now, if $x^\star \leqslant \frac{1}{2}$, then, for any $x \in \left[x^\star, x^\star + \frac{1}{M}\right]$ we have that

$$\min(1, M|x - x^\star|) = M|x - x^\star|$$

and then, recalling that $M \geqslant 2$,

$$\begin{aligned}(\star) &\geqslant \eta \sum_{t=1}^T \mathbb{E}\big[\rho(x^\star, Y_t)\big] + \ln\left(\int_{\left[x^\star, x^\star + \frac{1}{M}\right]} \exp\left(-\eta T \min(1, M|x - x^\star|)\right) \mathrm{d}x\right) \\ &= \eta \sum_{t=1}^T \mathbb{E}\big[\rho(x^\star, Y_t)\big] + \ln\left(\frac{1 - \exp\left(-\eta T\right)}{\eta T M}\right)\end{aligned}$$

The case $x^\star > \frac{1}{2}$ can be worked out analogously obtaining the same result. In any case, putting everything together, we get

$$\eta \mathbb{E}\left[\sum_{t=1}^T \rho(X_t, Y_t)\right] + (e-2)\eta^2 T \geqslant \eta \mathbb{E}\left[\sum_{t=1}^T \rho(x^\star, Y_t)\right] + \ln\left(\frac{1 - \exp\left(-\eta T\right)}{\eta T M}\right)$$

which, dividing by η and rearranging, becomes

$$\mathbb{E}\left[\sum_{t=1}^T \rho(x^\star, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^T \rho(X_t, Y_t)\right] \leqslant \frac{1}{\eta} \ln\left(\frac{\eta T M}{1 - e^{-\eta T}}\right) + \eta(e-2)T$$

So, if $\eta = \sqrt{\frac{\ln(2T)}{(e-2)T}}$, we have

$$\mathbb{E}\left[\sum_{t=1}^T \rho(x^\star, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^T \rho(X_t, Y_t)\right] \leqslant \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln(M)}{\ln(2T)}\right) .$$

□

In the same spirit of the previous theorem, we now obtain an immediate corollary that provides theoretical guarantees for Hedge run for $[0,1]$-armed experts (see the general online protocol of $\mathcal{X}$-armed experts and the corresponding definition of Hedge when $\mathcal{X} = [0,1]$) with Lipschitz *expected* rewards.

Corollary 3. *If there exists $M \geqslant 2$ such that, for all $t \in \mathbb{N}$, $x \mapsto \mathbb{E}[G_t(x)]$ is an M-Lipschitz function, then, for any time horizon $T \in \mathbb{N}$, the regret of Hedge for $[0,1]$-Armed Experts run with*

Online Protocol: $\mathcal{X}$-Armed Experts

Instance parameters: Known action space $\mathcal{X}$, unknown environment's action space $\mathcal{Y}$, unknown reward function $\rho\colon \mathcal{X} \times \mathcal{Y} \to [0,1]$

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

The environment secretly selects an action $Y_t \in \mathcal{Y}$ (possibly at random)

The learner secretly selects an action $X_t \in \mathcal{X}$ (possibly at random)

The learner gains reward $\rho(X_t, Y_t)$

X_t is revealed to the environment and $G_t(\cdot) = \rho(\cdot, Y_t)$ is revealed to the learner

Learning algorithm with full feedback: Hedge for $[0,1]$-Armed Experts

Input: $\eta \in (0,1)$

Initialization: Initialize $W_1(x) = 1$, for all $x \in [0,1]$

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

Play $X_t \sim \mu_t$, where μ_t is a distribution with density defined, for all $x \in [0,1]$, by $f_t(x) = \frac{W_t(x)}{\|W_t\|_1}$

Update $W_{t+1}(x) = W_t(x) \cdot \exp\big(\eta G_t(x)\big)$, for each $x \in [0,1]$

parameter $\eta \in (0,1)$ *is*[*]

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \leqslant \frac{1}{\eta} \ln\left(\frac{\eta T M}{1 - e^{-\eta T}}\right) + (e-2)\eta T \ .$$

In particular, if $\eta = \sqrt{\frac{\ln(2T)}{(e-2)T}}$ *we have*

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \leqslant \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln(M)}{\ln(2T)}\right) \ .$$

We remark that Hedge achieves an extremely mild dependence on M —disappearing completely if T is larger than M— without requiring the knowledge of M to tune the parameter η.

Finally, we highlight a key feature of our Theorem 35 and Corollary 3: they only assume that *expected* rewards are Lipschitz. This is in contrast with the classic assumption that the rewards themselves are Lipschitz. This seemingly small difference entails a technical issue in the analysis that we bypassed by proving a log-exp analogous of Minkowski's integral inequality, Lemma 16.

A.3 A Log-Exp Minkowski's Integral Inequality

In this section, we prove a log-exp analogous to Minkowski's integral inequality. In its original form, Minkowski's inequality states that

$$\int_{\mathcal{V}} \left(\int_{\mathcal{W}} \big(g(v,w)\big)^p \, \mathrm{d}\mu_{\mathcal{W}}(w)\right)^{1/p} \mathrm{d}\mu_{\mathcal{V}}(v) \geqslant \left(\int_{\mathcal{W}} \left(\int_{\mathcal{V}} g(v,w) \, \mathrm{d}\mu_{\mathcal{V}}(v)\right)^p \mathrm{d}\mu_{\mathcal{W}}(w)\right)^{1/p} ,$$

[*]Formally, we are assuming that $(\mathcal{Y}, \mathcal{E}_{\mathcal{Y}})$ is a measurable space; for all $t \in \mathbb{N}$, Y_t is chosen in a measurable way as a function of the information available to the environment at the beginning of time t, including its possible randomization; and ρ is a $(\mathcal{B}_{[0,1]} \otimes \mathcal{E}_{\mathcal{Y}})/\mathcal{B}_{[0,1]}$-measurable function.

where $p \geqslant 1$, $(\mathcal{V}, \mathcal{E}_{\mathcal{V}}, \mu_{\mathcal{V}})$ and $(\mathcal{W}, \mathcal{E}_{\mathcal{W}}, \mu_{\mathcal{W}})$ are two σ-finite measure spaces[†] and $g\colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$ is a measurable function.

We now prove a log-exp analogous of Minkowski's Integral Inequality. To the best of our knowledge, the following result has not been previously presented in the literature, and we believe it may be of independent interest.

We recall that $\mathcal{B}_{[0,+\infty]}$ denotes the Borel σ-algebra of $[0, +\infty]$.

Lemma 16 (Log-Exp Minkowski's Integral Inequality)**.** *Let $(\mathcal{V}, \mathcal{E}_{\mathcal{V}}, \mu_{\mathcal{V}})$ and $(\mathcal{W}, \mathcal{E}_{\mathcal{W}}, \mu_{\mathcal{W}})$ be two σ-finite measure spaces such that $\mu_{\mathcal{V}}[\mathcal{V}] \neq 0 \neq \mu_{\mathcal{W}}[\mathcal{W}]$. Let $g\colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$ be a $(\mathcal{E}_{\mathcal{V}} \otimes \mathcal{E}_{\mathcal{Y}})/\mathcal{B}_{[0,+\infty]}$ measurable function. Then (with the understanding that $0 \cdot \infty = 0$):*

$$\int_{\mathcal{V}} \ln\left(\int_{\mathcal{W}} \exp\big(g(v,w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w)\right) \mathrm{d}\mu_{\mathcal{V}}(v) \geqslant \mu_{\mathcal{V}}[\mathcal{V}] \ln\left(\int_{\mathcal{W}} \exp\left(\int_{\mathcal{V}} g(v,w)\, \mathrm{d}\mu_{\mathcal{V}}(v)\right) \mathrm{d}\mu_{\mathcal{W}}(w)\right)$$

Proof. Assume first that both $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{W}}$ are finite measures. Let $L^\infty(\mathcal{W})$ be the set of bounded $\mathcal{E}_{\mathcal{W}}/\mathcal{B}$-measurable functions. Define

$$\Phi\colon L^\infty(\mathcal{W}) \to \mathbb{R} \qquad f \mapsto \ln \int_{\mathcal{W}} \exp\big(f(w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w)$$

Notice that Φ is convex. In fact, for any $f_1, f_2 \in L^\infty(\mathcal{W})$ and any $\lambda \in (0,1)$, we have

$$\begin{aligned}
\Phi\big((1-\lambda)f_1 + \lambda f_2\big) &= \ln \int_{\mathcal{W}} \exp\big((1-\lambda)f_1(w) + \lambda f_2(w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w) \\
&= \ln \int_{\mathcal{W}} \Big(\exp\big(f_1(w)\big)\Big)^{1-\lambda} \Big(\exp\big(f_2(w)\big)\Big)^{\lambda} \mathrm{d}\mu_{\mathcal{W}}(w) \\
&\leqslant \ln \left(\left(\int_{\mathcal{W}} \exp\big(f_1(w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w)\right)^{1-\lambda} \left(\int_{\mathcal{W}} \exp\left(f_2(w)\right)\, \mathrm{d}\mu_{\mathcal{W}}(w)\right)^{\lambda}\right) \\
&= (1-\lambda) \ln \left(\int_{\mathcal{W}} \exp\big(f_1(w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w)\right) + \lambda \ln \left(\int_{\mathcal{W}} \exp\left(f_2(w)\right)\, \mathrm{d}\mu_{\mathcal{W}}(w)\right) \\
&= (1-\lambda)\Phi(f_1) + \lambda\Phi(f_2)\ ,
\end{aligned}$$

where the inequality follows from Hölder inequality with $p = \frac{1}{1-\lambda}$ and $q = \frac{1}{\lambda}$, the monotonicity of the integral, and the fact that ln is monotonically increasing. Now, notice that Φ is differentiable from the Banach space $(L^\infty(\mathcal{W}), \|\cdot\|_\infty)$ to $\mathbb{R}$ (where $\|f\|_\infty = \sup_{w \in \mathcal{W}} |f(w)|$), and for each $f \in L^\infty(\mathcal{W})$ the differential of Φ at any $f \in L^\infty(\mathcal{W})$ satisfies

$$\mathrm{d}\Phi(f)(h) = \frac{\int_{\mathcal{W}} \exp\big(f(w)\big) h(w)\, \mathrm{d}\mu_{\mathcal{W}}(w)}{\int_{\mathcal{W}} \exp\big(f(w)\big)\, \mathrm{d}\mu_{\mathcal{W}}(w)}\ , \qquad \text{for each } h \in L^\infty(\mathcal{W})\ .$$

The convexity and the differentiability of Φ together implies that for any $f_1, f_2 \in L^\infty(\mathcal{W})$ it holds that

$$\Phi(f_1) \geqslant \Phi(f_2) + \mathrm{d}\Phi(f_2)(f_1 - f_2)\ .$$

[†]We recall that a measure space $(\mathcal{A}, \mathcal{E}_{\mathcal{A}}, \mu_{\mathcal{A}})$ is σ-finite if there exist a countable family $A_1, A_2, \cdots \in \mathcal{E}_{\mathcal{A}}$ such that $\mu_{\mathcal{A}}(A_k) < +\infty$ for all $k \in \mathbb{N}$ and $\bigcup_{k \in \mathbb{N}} A_k = \mathcal{A}$.

Now, if $g \in L^\infty(\mathcal{V} \times \mathcal{W})$ (i.e., if g is bounded and $(\mathcal{E}_\mathcal{V} \otimes \mathcal{E}_\mathcal{W})/\mathcal{B}_{[0,+\infty]}$ measurable), define

$$G\colon \mathcal{V} \to L^\infty(\mathcal{W}) \,, \qquad v \mapsto g(v,\cdot) \,,$$

and define also

$$f_2(\cdot) = \int_\mathcal{V} g(v',\cdot)\,\mathrm{d}\mu_\mathcal{V}(v') \in L^\infty(\mathcal{W}) \,.$$

It follows that, for any $v \in \mathcal{V}$,

$$\begin{aligned}
\ln \int_\mathcal{W} \exp\big(g(v,w)\big)\,\mathrm{d}\mu_\mathcal{W}(w) &= \ln \int_\mathcal{W} \exp\big(G(v)(w)\big)\,\mathrm{d}\mu_\mathcal{W}(w) = \Phi\big(G(v)\big) \\
&\geqslant \Phi(f_2) + \mathrm{d}\Phi(f_2)\big(G(v) - f_2\big) \\
&= \ln\left(\int_\mathcal{W} \exp\left(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\right)\,\mathrm{d}\mu_\mathcal{W}(w)\right) \\
&\quad + \frac{\int_\mathcal{W}\Big(\exp\big(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\,\big(g(v,w) - \int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\Big)\,\mathrm{d}\mu_\mathcal{W}(w)}{\int_\mathcal{W} \exp\big(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\,\mathrm{d}\mu_\mathcal{W}(w)} \,.
\end{aligned}$$

Given that this last inequality holds for any $v \in \mathcal{V}$, we can integrate both sides with respect to $\mathrm{d}\mu_\mathcal{V}(v)$ and get

$$\begin{aligned}
&\int_\mathcal{V} \ln\left(\int_\mathcal{W} \exp\big(g(v,w)\big)\,\mathrm{d}\mu_\mathcal{W}(w)\right)\,\mathrm{d}\mu_\mathcal{V}(v) \\
&\quad\geqslant \mu_\mathcal{V}[\mathcal{V}] \ln\left(\int_\mathcal{W} \exp\left(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\right)\,\mathrm{d}\mu_\mathcal{W}(w)\right) \\
&\qquad + \int_\mathcal{V} \frac{\int_\mathcal{W}\Big(\exp\big(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\,\big(g(v,w) - \int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\Big)\,\mathrm{d}\mu_\mathcal{W}(w)}{\int_\mathcal{W} \exp\big(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\big)\,\mathrm{d}\mu_\mathcal{W}(w)}\,\mathrm{d}\mu_\mathcal{V}(v) \\
&\quad= \mu_\mathcal{V}[\mathcal{V}] \ln\left(\int_\mathcal{W} \exp\left(\int_\mathcal{V} g(v',w)\,\mathrm{d}\mu_\mathcal{V}(v')\right)\,\mathrm{d}\mu_\mathcal{W}(w)\right)
\end{aligned}$$

where the last equality follows from Fubini's theorem. Notice that we have proved the theorem under the assumption that $g \in L^\infty(\mathcal{V} \times \mathcal{W})$ and that $\mu_\mathcal{V}$ and $\mu_\mathcal{W}$ are finite measures.

Now, if $g \notin L^\infty(\mathcal{V} \times \mathcal{W})$ but $\mu_\mathcal{V}$ and $\mu_\mathcal{W}$ are finite, given that $g \geqslant 0$, we can find a sequence $(g_n)_{n\in\mathbb{N}} \subset L^\infty(\mathcal{V} \times \mathcal{W})$ such that $g_n \uparrow g$ pointwise, and obtain the conclusion from the monotone convergence theorem. If $\mu_\mathcal{V}[\mathcal{V}] = +\infty$ but $\mu_\mathcal{W}$ is finite, given that $\mu_\mathcal{V}$ is σ-finite, we can find a sequence $A_1 \subset A_2 \subset \ldots$ such that $\bigcup_{n\in\mathbb{N}} A_n = \mathcal{V}$ and, for each $n \in \mathbb{N}$ it holds that $A_n \in \mathcal{E}_\mathcal{V}$ and $\mu_\mathcal{V}[A_n] < +\infty$ and apply the theorem to the restriction of $\mu_\mathcal{V}$ to A_n and let $n \to \infty$ to obtain the conclusion via the monotone convergence theorem. Finally, if $\mu_\mathcal{W}[\mathcal{W}] = +\infty$, given that $\mu_\mathcal{W}$ is σ-finite, we can find a sequence $B_1 \subset B_2 \subset \ldots$ such that $\bigcup_{n\in\mathbb{N}} B_n = \mathcal{W}$ and, for each $n \in \mathbb{N}$ it holds that $B_n \in \mathcal{E}_\mathcal{W}$ and $\mu_\mathcal{W}[B_n] < +\infty$ and apply the theorem to the restriction of $\mu_\mathcal{W}$ to B_n and let $n \to \infty$ to obtain the conclusion via the monotone convergence theorem again. □

As an immediate corollary of the previous lemma, we get the following.

Corollary 4 (Log-Exp Minkowski's Integral Inequality for probability measures)**.** *Let $(\mathcal{V}, \mathcal{E}_\mathcal{V})$ and $(\mathcal{W}, \mathcal{E}_\mathcal{W})$ be two measurable spaces and let $g\colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$ be a $\mathcal{E}_\mathcal{V} \otimes \mathcal{E}_\mathcal{W}/\mathcal{B}_{[0,+\infty]}$-measurable function. Assume that V and W are an $\mathcal{V}$-valued and a $\mathcal{W}$-valued random variables, respectively,*

independent of each other. Then

$$\mathbb{E}\Big[\ln \mathbb{E}\Big[\exp\big(g(V,W)\big) \mid V\Big]\Big] \geqslant \ln \mathbb{E}\Big[\exp\Big(\mathbb{E}\big[g(V,W) \mid W\big]\Big)\Big]$$

A.4 A Generalized Freezing Lemma

The classic "freezing lemma" (see, e.g., Cesari and Colomboni 61, Lemma 8) states that the conditional expectation of a measurable function of two independent random variables given one of them can be computed as an expectation only with respect to the other random variable followed by a composition with the random variable in the conditioning.

Lemma 17 (The *freezing lemma*)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(\mathcal{V}, \mathcal{F}_{\mathcal{V}})$ and $(\mathcal{W}, \mathcal{F}_{\mathcal{W}})$ be two measurable spaces. Let $f\colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$, $V\colon \Omega \to \mathcal{V}$, $W\colon \Omega \to \mathcal{W}$ be three measurable functions. If V and W are $\mathbb{P}$-independent, then*

$$\mathbb{E}\big[f(V,W) \mid V\big] = \Big[\mathbb{E}\big[f(v,W)\big]\Big]_{v=V} \tag{A.2}$$

$\mathbb{P}$-almost surely, where the right hand side is the composition

$$\Big[\mathbb{E}\big[f(v,W)\big]\Big]_{v=V} = \Big(v \mapsto \mathbb{E}\big[f(v,W)\big]\Big) \circ V\ .$$

The freezing lemma is extremely useful in derivations as it allows one to isolate the random parts that are being averaged while keeping the others fixed. Unfortunately, the freezing lemma does not cover the case where the expectations are replaced with conditional expectation on some σ-algebra, which is often the case in online learning, where expectations and probabilities are typically intended as conditional on the history up to the present time. This problem cannot be solved by simply replacing expectations with conditional expectations everywhere because of the fact that versions of conditional expectations remain as such if changed on a probability-zero event, making the naive extension to the right-hand side of Equation (A.2) not even well-defined. To aid us in giving a sound statement of such a generalization of the freezing lemma, we begin by recalling the definition of regular conditional probability.

Definition 4 (Regular conditional probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(\mathcal{X}, \mathcal{E}_{\mathcal{X}})$ be a measurable space. Let $X\colon \Omega \to \mathcal{X}$ be a $\mathcal{F}/\mathcal{E}_{\mathcal{X}}$-measurable. Let $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. We say that $\Phi\colon \mathcal{E}_{\mathcal{X}} \to [0,1]^{\Omega}$ is a regular conditional probability for $\mathbb{P}_{X|\mathcal{H}}$ if:*

- *For each $A \in \mathcal{E}_{\mathcal{X}}$, the function $\omega \mapsto \Phi[A](\omega)$ is $\mathcal{H}/\mathcal{B}_{[0,1]}$-measurable.*
- *For each $\omega \in \Omega$, the function $A \mapsto \Phi[A](\omega)$ is a probability measure.*
- *For each $A \in \mathcal{E}_{\mathcal{X}}$ and each $H \in \mathcal{H}$, it holds that $\mathbb{P}\big[H \cap \{X \in A\}\big] = \mathbb{E}\big[\mathbb{I}_H \Phi[A]\big]$.*

Notice that the first and the third bullet imply that $\Phi[A] = \mathbb{E}[\mathbb{I}_{X \in A} \mid \mathcal{H}]$ for each $A \in \mathcal{E}_{\mathcal{X}}$.

We can now state and prove a generalized version of the freezing lemma, which we believe may be of independent interest.

We recall that $\mathcal{B}_{[0,+\infty]}$ denotes the Borel σ-algebra of $[0, +\infty]$.

Lemma 18 (Generalized Freezing Lemma)**.** *Let $(\mathcal{X}, \mathcal{E}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{E}_\mathcal{Y})$ be two measurable spaces. Let $g\colon \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ be a $(\mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y})/\mathcal{B}_{[0,+\infty]}$- measurable function. Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space and $\mathcal{F}, \mathcal{G}, \mathcal{H}$ be three sub-σ-algebras of $\mathcal{E}$. Let $X\colon \Omega \to \mathcal{X}$ be a $\mathcal{F}/\mathcal{E}_\mathcal{X}$-measurable random variable. Let $Y\colon \Omega \to \mathcal{Y}$ be a $\mathcal{G}/\mathcal{E}_\mathcal{Y}$-measurable random variable. Assume that $\mathcal{F}$ and $\mathcal{G}$ are $\mathbb{P}$-conditionally independent given $\mathcal{H}$. Assume that Φ is a regular conditional probability for $\mathbb{P}_{X|\mathcal{H}}$. Then*

$$\int_\mathcal{X} g(x, Y)\, \mathrm{d}\Phi(x) = \mathbb{E}\big[g(X, Y) \mid \sigma(\mathcal{G}, \mathcal{H})\big]\,.$$

Proof. First, notice that the random variable $\int_\mathcal{X} g\big(x, Y\big)\, \mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable. In fact, if $A \in \mathcal{E}_\mathcal{X}$ and $B \in \mathcal{E}_\mathcal{Y}$ we have

$$\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y)\, \mathrm{d}\Phi(x) = \Phi[A]\mathbb{I}_B(Y),$$

which implies that $\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y)\, \mathrm{d}\Phi(x)$, as a product of a $\mathcal{H}$-measurable function and a $\mathcal{G}$-measurable function is $\sigma(\mathcal{G}, \mathcal{H})$-measurable. Now, consider the family

$$\mathcal{C} = \left\{ C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y} \mid \int_\mathcal{X} \mathbb{I}_C(x, Y)\, \mathrm{d}\Phi(x) \text{ is } \sigma(\mathcal{G}, \mathcal{H})\text{-measurable} \right\}.$$

Notice that $\mathcal{X} \times \mathcal{Y} \in \mathcal{C}$, that $\mathcal{C}$ is closed under complementation and that if $(C_n)_{n\in\mathbb{N}} \subset \mathcal{C}$ is such that $C_1 \subset C_2 \subset \dots$ then $\bigcup_{n\in\mathbb{N}} C_n \in \mathcal{C}$. Hence, $\mathcal{C}$ is a λ-system which contains the π-system $\mathcal{D} = \{C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y} \mid \exists A \in \mathcal{E}_\mathcal{X}, \exists B \in \mathcal{E}_\mathcal{Y}, C = A \times B\}$. Hence, by the π-λ theorem [34, Theorem 3.2] it holds that $\sigma(\mathcal{D}) \subset \mathcal{C}$, and since $\sigma(\mathcal{D}) = \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ it holds that $\mathcal{C} = \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$. It follows that for each $C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ the random variable $\int_\mathcal{X} \mathbb{I}_C(x, Y)\, \mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable. By pointwise monotone increasing approximation via $\mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$-measurable simple functions[‡], we get that the random variable $\int_\mathcal{X} g\big(x, Y\big)\, \mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable.

Now, pick $A \in \mathcal{E}_\mathcal{X}, B \in \mathcal{E}_\mathcal{Y}, G \in \mathcal{G}$ and $H \in \mathcal{H}$. Notice that

$$\begin{aligned}
\mathbb{E}\left[\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y)\, \mathrm{d}\Phi(x)\mathbb{I}_{G\cap H}\right] &= \mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\Phi[A]\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)} \mid \mathcal{H}\right]\Phi[A]\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)} \mid \mathcal{H}\right]\mathbb{E}[\mathbb{I}_{X\in A} \mid \mathcal{H}]\mathbb{I}_H\right] \\
(\mathcal{F} \text{ and } \mathcal{G} \text{ are conditionally independent given } \mathcal{H}) \qquad &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\mathbb{I}_{X\in A} \mid \mathcal{H}\right]\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\mathbb{I}_{X\in A}\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{I}_A(X)\mathbb{I}_B(Y)\mathbb{I}_{G\cap H}\right]\,.
\end{aligned}$$

Applying twice a π-λ argument as done above, we can prove that for each $C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ and each $K \in \sigma(\mathcal{G}, \mathcal{H})$, it holds that

$$\mathbb{E}\left[\int_\mathcal{X} \mathbb{I}_C(x, Y)\, \mathrm{d}\Phi(x)\mathbb{I}_K\right] = \mathbb{E}\left[\mathbb{I}_C(X, Y)\mathbb{I}_K\right]\,.$$

Applying again a pointwise monotone approximation argument using $\mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$-measurable simple

[‡]We recall that simple functions are linear combinations of indicator functions.

functions, we can prove that for each $K \in \sigma(\mathcal{G}, \mathcal{H})$ it holds that

$$\mathbb{E}\left[\int_{\mathcal{X}} g(x, Y)\, \mathrm{d}\Phi(x) \mathbb{I}_K\right] = \mathbb{E}\left[g(X, Y)\mathbb{I}_K\right] .$$

Given that we have already proved that the random variable $\int_{\mathcal{X}} g(x, Y)\, \mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable, the conclusion follows. □

A.5 Model and Notation

For all $T \in \mathbb{N}$, we denote the set of the first T integers $\{1, \dots, T\}$ by $[T]$. If $\mathbb{P}$ is a probability measure and X is a random variable, we denote by $\mathbb{P}_X$ the probability measure defined for any (measurable) set E, by $\mathbb{P}_X[E] := \mathbb{P}[X \in E]$. We denote the expectation of a random variable X with respect to the probability measure $\mathbb{P}$ by $\mathbb{E}_{\mathbb{P}}[X]$. If a measure ν is absolutely continuous with respect to another measure μ with density f, we denote ν by $f\mu$, so that for any (measurable) set E, $(f\mu)[E] := \nu[E] = \int_E f(x)\, \mathrm{d}\mu(x)$. We denote the Lebesgue measure on the interval $[0, 1]$ by μ_L and the product Lebesgue measure on $[0, 1]^{\mathbb{N}}$ by $\boldsymbol{\mu}_L$. For any set E and $x \in E$, we denote the Dirac measure on x by δ_x (the dependence on E will always be clear from context).

A.5.1 The Learning Model

In this section, we introduce an abstract notion of sequential games which encompasses all the settings we discussed in the main part of the bilateral trade section, providing a unified perspective. This will be especially useful when proving lower bounds.

Definition 5 (Sequential game)**.** *A* (sequential) game *is a tuple* $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$*, where:*

- $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ *are sets called the* player's action space*,* adversary's action space*, and* feedback space*.*
- $\rho\colon \mathcal{X} \times \mathcal{Y} \to [0, 1]$ *and* $\varphi\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ *are called the* reward *and* feedback *functions*[§]*.*
- $\mathcal{P}$ *is a set of probabilities on the set* $\mathcal{Y}^{\mathbb{N}}$ *of sequences in* $\mathcal{Y}$*, called the* adversary's behavior*.*

This definition generalizes the partial monitoring games of Bartók et al. [28], Lattimore and Szepesvári [123] to settings with infinitely many arms and is able to model adversarial, i.i.d., and more general stochastic settings all at once. Before proceeding, we introduce another few extra handy definitions.

Definition 6. *If* $\mathscr{G} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$ *is a game, then we say the following.* The sample space *is the set* $\Omega := \mathcal{Y}^{\mathbb{N}} \times [0, 1]^{\mathbb{N}}$*.* The adversary's actions $\left(Y_t\right)_{t\in\mathbb{N}}$ *and* the player's randomization $\left(U_t\right)_{t\in\mathbb{N}}$ *are sequences of random variables defined, for all* $t \in \mathbb{N}$ *and* $\omega = \left((y_n)_{n\in\mathbb{N}}, (u_n)_{n\in\mathbb{N}}\right) \in \Omega$*, by* $Y_t(\omega) := y_t$ *and* $U_t(\omega) := u_t$*.* The set of environments $\mathcal{S}$ *is the set of probability measures* $\mathbb{P}$ *on* Ω *of the form* $\mathbb{P} = \boldsymbol{\mu} \otimes \boldsymbol{\mu}_L$*, where* $\boldsymbol{\mu} \in \mathbb{P}$*.*

[§]More precisely, we need $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to be non-empty measurable spaces and ρ, φ to be measurable functions. To avoid clutter, in the following we will never mention explicitly these types of standard measurability assumptions unless strictly needed.

For the sake of conciseness, whenever we fix a game $\mathscr{G}$, we will assume that all the objects (sets, functions, random variables) presented in Definitions 5–6 are fixed and denoted by the same letters without declaring them explicitly each time, unless strictly needed.

Note that this setting models an *oblivious* environment since the adversary's actions are independent of the player's past randomization, i.e., for all $t \in \mathbb{N}$, $\mathbb{P}_{Y_{t+1}|Y_1,\dots,Y_t,U_1,\dots,U_t} = \mathbb{P}_{Y_{t+1}|Y_1,\dots,Y_t}$. Note also that we are assuming that the randomization of the player's strategy is carried out by drawing numbers in the interval $[0,1]$ independently and uniformly at random. We can restrict ourselves to this case in light of the Skorokhod Representation Theorem [185, Section 17.3] without losing (much) generality. We now introduce formally the strategies of the player, the resulting played actions, and the corresponding feedback.

Definition 7 (Player's strategies, actions, and feedback)**.** *Given a game $\mathscr{G}$, we define a* player's strategy *as a sequence of functions $\alpha \coloneqq (\alpha_t)_{t\in\mathbb{N}}$ such that, for each $t \in \mathbb{N}$, $\alpha_t \colon [0,1]^t \times \mathcal{Z}^{t-1} \to \mathcal{X}$.*¶ *Given a player's strategy α, we define inductively (on t) the corresponding sequences of* player's actions $(X_t)_{t\in\mathbb{N}}$ *and* player's feedback $(Z_t)_{t\in\mathbb{N}}$ *by $X_t := \alpha_t(U_1,\dots,U_t,Z_1,\dots,Z_{t-1})$, $Z_t := \varphi(X_t,Y_t)$. In the sequel, we will denote the set of all strategies for a game $\mathscr{G}$ by $\mathscr{A}(\mathscr{G})$.*

To lighten the notation, we will write $\mathscr{A}$ instead of $\mathscr{A}(\mathscr{G})$ if it is clear from context. We can now extend the standard notions of regret, worst-case regret, and minimax regret to our general setting.

Definition 8 (Regret)**.** *Given a game $\mathscr{G}$ and a horizon $T \in \mathbb{N}$, we define the* regret *(of $\alpha \in \mathscr{A}$ in an environment $\mathbb{P} \in \mathcal{S}$), the* worst-case regret *(of $\alpha \in \mathscr{A}$), and the* minimax regret *(of $\mathscr{G}$), respectively, by*

$$R_T(\alpha,\mathbb{P}) := \sup_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T}\rho(x,Y_t) - \sum_{t=1}^{T}\rho(X_t,Y_t)\right],\ R_T^{\mathcal{S}}(\alpha) := \sup_{\mathbb{P}\in\mathcal{S}} R_T(\alpha,\mathbb{P}),\ R_T^{\star}(\mathscr{G}) := \inf_{\alpha\in\mathscr{A}(\mathscr{G})} R_T^{\mathcal{S}}(\alpha)\,.$$

If $\mathscr{G}$ and $\widetilde{\mathscr{G}}$ are two games and $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\widetilde{\mathscr{G}})$, we say that $\widetilde{\mathscr{G}}$ is *easier* than $\mathscr{G}$ (or equivalently, that $\mathscr{G}$ is *harder* than $\widetilde{\mathscr{G}}$). When it is clear from the context, we will omit the dependence on $\mathscr{G}$ in $R_T^\star(\mathscr{G})$.

A.5.2 Bilateral Trade as a Game

We now formally cast the various instances of bilateral trade we introduced in the main body into our sequential game setting.‖ In this context, we think of the learner as the *player* and the environments as the corresponding *adversaries*.

Player's Actions, Adversary's Actions, and Reward

The player's action space $\mathcal{X}$ is the unit interval $[0,1]$. This corresponds to the player posting the same price to both the seller and the buyer (budget balance). The adversary's action space $\mathcal{Y}$ is $[0,1]^2$. They are the pairs of valuations of the seller and buyer. The reward function ρ is the gain from trade $\mathrm{gft}\colon [0,1]\times[0,1]^2 \to [0,1]$, $\big(p,(s,b)\big) \mapsto (b-s)\mathbb{I}\{s \leqslant p \leqslant b\}$.

¶When $t=1$, $[0,1]^t \times \mathcal{Z}^{t-1} := [0,1]$. In the following, we will always adopt this type of convention without mention it.

‖Straightforwardly, the same can be done for the weak budget balance setting we studied in Section 2.6.

Available Feedback

Full: the feedback space $\mathcal{Z}$ is the unit square $[0,1]^2$ and the feedback function is $\varphi\colon [0,1]\times[0,1]^2 \to [0,1]^2,\ \big(p,(s,b)\big) \mapsto (s,b)$. This corresponds to the seller and the buyer revealing their valuations at the end of a trade.

Realistic: the feedback space $\mathcal{Z}$ is the boolean square $\{0,1\}^2$ and the feedback function is $\varphi\colon [0,1]\times[0,1]^2 \to \{0,1\}^2,\ \big(p,(s,b)\big) \mapsto \big(\mathbb{I}\{s \leqslant p\}, \mathbb{I}\{p \leqslant b\}\big)$. This corresponds to the seller and the buyer accepting or rejecting a trade at a price p.

Adversary's Behavior

Independent and identically distributed (iid): the adversary's behavior $\mathcal{P} = \mathcal{P}_{\text{iid}}$ consists of products of a single probability on $\mathcal{Y} = [0,1]^2$, i.e., $\boldsymbol{\mu} \in \mathcal{P}_{\text{iid}}$ if and only if there exists a probability measure μ on $[0,1]^2$ such that $\boldsymbol{\mu} = \otimes_{t\in\mathbb{N}}\,\mu$. This corresponds to a stochastic i.i.d. environment, where however the valuations of the seller and the buyer could be correlated.

In this appendix, we will also investigate the following stronger assumptions.

(iid) + independent valuations (iv): the adversary's behavior $\mathcal{P} = \mathcal{P}_{\text{iid+iv}}$ is the subset of $\mathcal{P}_{\text{iid}}$ in which the valuations of the seller and the buyer are independent, i.e., $\boldsymbol{\mu} \in \mathcal{P}_{\text{iid+iv}}$ if and only if there exist two probability measures μ_S, μ_B on $[0,1]$ such that $\boldsymbol{\mu} = \otimes_{t\in\mathbb{N}}\,(\mu_S \otimes \mu_B)$.

(iid) + bounded density (bd): for a fixed $M \geqslant 1$, the adversary's behavior $\mathcal{P} = \mathcal{P}^M_{\text{iid+bd}}$ is the subset of $\mathcal{P}_{\text{iid}}$ in which the joint distribution of the valuations of buyer and seller has a density bounded by M, i.e., $\boldsymbol{\mu} \in \mathcal{P}^M_{\text{iid+bd}}$ if and only if there exists a density $f\colon [0,1]^2 \to [0,M]$ such that $\boldsymbol{\mu} = \otimes_{t\in\mathbb{N}}\,(f\mu)$, where $\mu = \mu_L \otimes \mu_L$.

(iid) + independent valuations with bounded density (iv) + (bd): for a fixed $M \geqslant 1$, the adversary's behavior $\mathcal{P} = \mathcal{P}^M_{\text{iid+iv+bd}}$ is the subset $\mathcal{P}_{\text{iid+iv}} \cap \mathcal{P}^M_{\text{iid+bd}}$ of $\mathcal{P}_{\text{iid}}$.

Adversarial (adv): the adversary's behavior $\mathcal{P} = \mathcal{P}_{\text{adv}}$ consists of products of Dirac measures on $\mathcal{Y} = [0,1]^2$, i.e., $\boldsymbol{\mu} \in \mathcal{P}_{\text{adv}}$ if and only if there exists a sequence $(s_t, b_t)_{t\in\mathbb{N}} \subset [0,1]^2$ such that $\boldsymbol{\mu} = \otimes_{t\in\mathbb{N}}\,\delta_{(s_t,b_t)}$. This corresponds to a deterministic, oblivious, and adversarial environment.

A.6 Two Key Lemmas on Simplifying Sequential Games

In this section we introduce some useful techniques that could be of independent interest for proving lower bounds in sequential games. The idea is to give sufficient conditions for a given game to be harder than another, where the second one has a known lower bound on its minimax regret.

At a high level, the first lemma shows that if the adversary's actions are independent of each other, a game $\widetilde{\mathscr{G}}$ is easier than game $\mathscr{G}$ if $\widetilde{\mathscr{G}}$ can be embedded in $\mathscr{G}$ in such a way that:

1. The optimal player's actions of $\widetilde{\mathscr{G}}$ are no better than the ones in $\mathscr{G}$.
2. The suboptimal player's actions of $\widetilde{\mathscr{G}}$ no worse than the ones in $\mathscr{G}$.
3. At distributional level, the quality of the feedback in $\widetilde{\mathscr{G}}$ is no worse than that in $\mathscr{G}$.

The proof is deferred to Appendix A.6.1.

Lemma 19 (Embedding)**.** *Let $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$ and $\widetilde{\mathscr{G}} := (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}, \widetilde{\mathcal{Z}}, \widetilde{\rho}, \widetilde{\varphi}, \widetilde{\mathcal{P}})$ be two games, $\mathcal{S}, \widetilde{\mathcal{S}}$ their respective sets of environments, $(Y_t)_{t\in\mathbb{N}}, (\widetilde{Y}_t)_{t\in\mathbb{N}}$ their adversaries' actions, and $T \in \mathbb{N}$ a horizon. Assume that $Y_1, \dots, Y_T$ are $\mathbb{P}$-independent for any environment $\mathbb{P} \in \mathcal{S}$, $\widetilde{Y}_1, \dots, \widetilde{Y}_T$ are $\widetilde{\mathbb{P}}$-independent for any environment $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$, and that there exist $\widetilde{f}\colon \mathcal{X} \to \widetilde{\mathcal{X}}$, $g\colon \widetilde{\mathcal{Z}} \to \mathcal{Z}$, and $h\colon \widetilde{\mathcal{S}} \to \mathcal{S}$ satisfying:*

1. $\sup_{\widetilde{x}\in\widetilde{\mathcal{X}}} \sum_{t=1}^{T} \mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{x}, \widetilde{Y}_t)\big] \leqslant \sup_{x\in\mathcal{X}} \sum_{t=1}^{T} \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big]$ *for any environment $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$.*
2. $\mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}\big(\widetilde{f}(x), \widetilde{Y}_t\big)\big] \geqslant \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big]$ *for any time $t \in [T]$, environment $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$, and action $x \in \mathcal{X}$.*
3. $\widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(x), \widetilde{Y}_t\right)\right)} = \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(x, Y_t)}$ *for any time $t \in [T]$, environment $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$, and action $x \in \mathcal{X}$.*

Then $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\widetilde{\mathscr{G}})$.

The second lemma addresses feedback with uninformative (i.e., environment-independent) components. At a high level, if the feedback of some of the player's actions has one or more uninformative components, the game can be simplified by getting rid of them. The player can achieve this by simulating the uninformative parts of the feedback using her randomization. The proof is deferred to Appendix A.6.1.

Lemma 20 (Simulation)**.** *Let $\mathcal{V}, \mathcal{W}$ be two sets, $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$ a game with $\mathcal{Z} = \mathcal{V} \times \mathcal{W}$, $\mathcal{S}$ its set of environments, $(Y_t)_{t\in\mathbb{N}}$ its adversary's actions, $\pi\colon \mathcal{Z} \to \mathcal{V}$ the projection on $\mathcal{V}$, and $T \in \mathbb{N}$ a horizon. Assume that $Y_1, \dots, Y_T$ are $\mathbb{P}$-independent for any environment $\mathbb{P} \in \mathcal{S}$ and that there exist disjoint sets $\mathcal{I}$,$\mathcal{U} \subset \mathcal{X}$ such that $\mathcal{I} \cup \mathcal{U} = \mathcal{X}$ and*

1. *For any time $t \in [T]$ and action $x \in \mathcal{I}$ there exists $\psi_{t,x}\colon [0,1] \to \mathcal{W}$ such that, for all $\mathbb{P} \in \mathcal{S}$,*

$$\mathbb{P}_{\varphi(x,Y_t)} = \mathbb{P}_{\pi\left(\varphi(x,Y_t)\right)} \otimes (\mu_L)_{\psi_{t,x}} \,.$$

2. *For any time $t \in [T]$ and action $x \in \mathcal{U}$, there exists $\gamma_{t,x}\colon [0,1] \to \mathcal{Z}$ such that, for all $\mathbb{P} \in \mathcal{S}$,*

$$\mathbb{P}_{\varphi(x,Y_t)} = (\mu_L)_{\gamma_{t,x}} \,.$$

Let $* \in \mathcal{V}$ and define

$$\widetilde{\varphi}\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{V},\ (x,y) \mapsto \begin{cases} \pi\big(\varphi(x,y)\big)\,, & \text{if } x \in \mathcal{I}, \\ *\,, & \text{if } x \in \mathcal{U}. \end{cases}$$

Define the game $\widetilde{\mathscr{G}} := (\mathcal{X}, \mathcal{Y}, \mathcal{V}, \rho, \widetilde{\varphi}, \mathcal{P})$. Then $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\widetilde{\mathscr{G}})$.

A.6.1 Proofs of the Lemmas

In this section, we will give a full proof of the two useful Embedding and Simulation lemmas introduces in Appendix A.6. To lighten the notation, for any $m, n \in \mathbb{N}$, with $m \leqslant n$ and a family $(\lambda_k)_{k\in\mathbb{N}}$ we let $\lambda_{m:n} := (\lambda_m, \lambda_{m+1}, \dots, \lambda_n)$ and similarly $\lambda_{n:m} := (\lambda_n, \lambda_{n-1} \dots, \lambda_m)$.

We begin by proving the Embedding lemma, that we restate for ease of reading.

Lemma 19 (Embedding)**.** *Let $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$ and $\widetilde{\mathscr{G}} := (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}, \widetilde{\mathcal{Z}}, \widetilde{\rho}, \widetilde{\varphi}, \widetilde{\mathcal{P}})$ be two games, $\mathcal{S}, \widetilde{\mathcal{S}}$ their respective sets of environments, $(Y_t)_{t\in\mathbb{N}}, (\widetilde{Y}_t)_{t\in\mathbb{N}}$ their adversaries' actions, and $T \in \mathbb{N}$ a horizon. Assume that $Y_1, \dots, Y_T$ are $\mathbb{P}$-independent for any environment $\mathbb{P} \in \mathcal{S}$, $\widetilde{Y}_1, \dots, \widetilde{Y}_T$ are $\widetilde{\mathbb{P}}$-independent for any environment $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$, and that there exist $\widetilde{f}\colon \mathcal{X} \to \widetilde{\mathcal{X}}$, $g\colon \widetilde{\mathcal{Z}} \to \mathcal{Z}$, and $h\colon \widetilde{\mathcal{S}} \to \mathcal{S}$ satisfying:*

1. $\sup_{\widetilde{x}\in\widetilde{\mathcal{X}}} \sum_{t=1}^{T} \mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{x}, \widetilde{Y}_t)\big] \leqslant \sup_{x\in\mathcal{X}} \sum_{t=1}^{T} \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big]$ *for any environment* $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$.
2. $\mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}\big(\widetilde{f}(x), \widetilde{Y}_t\big)\big] \geqslant \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big]$ *for any time* $t \in [T]$*, environment* $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$*, and action* $x \in \mathcal{X}$.
3. $\widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(x), \widetilde{Y}_t\right)\right)} = \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(x, Y_t)}$ *for any time* $t \in [T]$*, environment* $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$*, and action* $x \in \mathcal{X}$.

Then $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\widetilde{\mathscr{G}})$.

Proof. Fix any strategy $\alpha \in \mathscr{A}(\mathscr{G})$. For each time $t \in \mathbb{N}$, define

$$\widetilde{\alpha}_t\colon [0,1]^t \times \widetilde{\mathcal{Z}}^{t-1} \to \widetilde{\mathcal{X}}, (u_1, \dots, u_t, \widetilde{z}_1, \dots, \widetilde{z}_{t-1}) \mapsto \widetilde{f}\Big(\alpha_t\big(u_1, \dots, u_t, g(\widetilde{z}_1), \dots, g(\widetilde{z}_{t-1})\big)\Big).$$

Then $\widetilde{\alpha} := (\widetilde{\alpha}_t)_{t\in\mathbb{N}} \in \mathscr{A}(\widetilde{\mathscr{G}})$. As usual, let $(Y_t)_{t\in\mathbb{N}}$ and $(U_t)_{t\in\mathbb{N}}$ be the adversary's actions and the player's randomization in game $\mathscr{G}$ and $(X_t)_{t\in\mathbb{N}}$ and $(Z_t)_{t\in\mathbb{N}}$ the player's actions and the feedback according to the strategy α. Let $(\widetilde{Y}_t)_{t\in\mathbb{N}}, (\widetilde{U}_t)_{t\in\mathbb{N}}, (\widetilde{X}_t)_{t\in\mathbb{N}}, (\widetilde{Z}_t)_{t\in\mathbb{N}}$ be the corresponding objects for the game $\widetilde{\mathscr{G}}$ and the strategy $\widetilde{\alpha}$. Furthermore, define

$$\widehat{X}_1 = \alpha_1(\widetilde{U}_1), \quad \widehat{Z}_1 = g\big(\widetilde{\varphi}(\widetilde{X}_1, \widetilde{Y}_1)\big), \quad \widehat{X}_2 = \alpha_2(\widetilde{U}_1, \widetilde{U}_2, \widehat{Z}_1), \quad \widehat{Z}_2 = g\big(\widetilde{\varphi}(\widetilde{X}_2, \widetilde{Y}_2)\big), \dots .$$

Fix $\widetilde{\mathbb{P}} \in \widetilde{\mathcal{S}}$, where $\widetilde{\mathcal{S}}$ is the set of environments of the game $\widetilde{\mathscr{G}}$. Then $\widetilde{\mathbb{P}}_{\widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{U_1}$. Now, since $X_1 = \alpha_1(U_1)$ and $\widehat{X}_1 = \alpha_1(\widetilde{U}_1)$, we also have that $\widetilde{\mathbb{P}}_{\widehat{X}_1, \widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{X_1, U_1} =: \mathbb{Q}_1$. Now, up to a set with $\mathbb{Q}_1$-probability zero, if $x_1 \in \mathcal{X}$ and $u_1 \in [0,1]$, we get, using Item 3:

$$\begin{aligned}
\widetilde{\mathbb{P}}_{\widehat{Z}_1 | \widehat{X}_1 = x_1, \widetilde{U}_1 = u_1} &= \widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(\widehat{X}_1), \widetilde{Y}_1\right)\right) | \widehat{X}_1 = x_1, \widetilde{U}_1 = u_1} = \widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(x_1), \widetilde{Y}_1\right)\right)} \\
&= \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(x_1, Y_1)} = \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(X_1, Y_1) | X_1 = x_1, U_1 = u_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_1 | X_1 = x_1, U_1 = u_1} .
\end{aligned}$$

So, if $A_1 \subset \mathcal{Z}$ and $D \subset \mathcal{X} \times [0,1]$, then

$$\begin{aligned}
\widetilde{\mathbb{P}}_{\widehat{Z}_1, \left(\widehat{X}_1, \widetilde{U}_1\right)}(A_1 \times D) &= \int_D \mathbb{P}_{\widehat{Z}_1 | \widehat{X}_1 = x_1, \widetilde{U}_1 = u_1}(A_1) \, \mathrm{d}\mathbb{P}_{\widehat{X}_1, \widetilde{U}_1}(x_1, u_1) \\
&= \int_D \big(h(\widetilde{\mathbb{P}})\big)_{Z_1 | X_1 = x_1, U_1 = u_1}(A_1) \, \mathrm{d}\big(h(\widetilde{\mathbb{P}})\big)_{X_1, U_1}(x_1, u_1) = \big(h(\widetilde{\mathbb{P}})\big)_{Z_1, (X_1, U_1)}(A_1 \times D) ,
\end{aligned}$$

from which it follows that $\widetilde{\mathbb{P}}_{\widehat{Z}_1, \widehat{X}_1, \widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_1, X_1, U_1}$. By induction, suppose that for time $t \in [T-1]$ we have that

$$\widetilde{\mathbb{P}}_{\widehat{Z}_t, \dots, \widehat{Z}_1, \widehat{X}_t, \dots, \widehat{X}_1, \widetilde{U}_t, \dots, \widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_t, \dots, Z_1, X_t, \dots, X_1, U_t, \dots, U_1} .$$

Then, using independence we have that

$$\widetilde{\mathbb{P}}_{\widehat{Z}_t, \dots, \widehat{Z}_1, \widehat{X}_t, \dots, \widehat{X}_1, \widetilde{U}_{t+1}, \widetilde{U}_t, \dots, \widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_t, \dots, Z_1, X_t, \dots, X_1, U_{t+1}, U_t, \dots, U_1} .$$

Furthermore, since $X_{t+1} = \alpha_{t+1}(U_1, \dots, U_{t+1}, Z_1, \dots, Z_t)$ and $\widehat{X}_{t+1} = \alpha_{t+1}(\widetilde{U}_1, \dots, \widetilde{U}_{t+1}, \widehat{Z}_1, \dots, \widehat{Z}_t)$, we have that

$$\widetilde{\mathbb{P}}_{\widehat{Z}_t,\dots,\widehat{Z}_1,\widehat{X}_{t+1},\widehat{X}_t,\dots,\widehat{X}_1,\widetilde{U}_{t+1},\widetilde{U}_t,\dots,\widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_t,\dots,Z_1,X_{t+1},X_t,\dots,X_1,U_{t+1},U_t,\dots,U_1} =: \mathbb{Q}_{t+1} \,.$$

Now, up to a set with $\mathbb{Q}_{t+1}$-probability zero, if $x_1, \dots, x_{t+1} \in \mathcal{X}$, $u_1, \dots, u_{t+1} \in [0,1]$, and $z_1, \dots, z_t \in \mathcal{Z}$, by the $\widetilde{\mathbb{P}}$-independence of $\widetilde{Y}_1, \dots, \widetilde{Y}_{t+1}$, Item 3, and the $h(\widetilde{\mathbb{P}})$-independence of $Y_1, \dots, Y_{t+1}$, we have

$$\begin{aligned}
&\widetilde{\mathbb{P}}_{\widehat{Z}_{t+1}|\widehat{Z}_t=z_t,\dots,\widehat{Z}_1=z_1,\widehat{X}_{t+1}=x_{t+1},\dots,\widehat{X}_1=x_1,\widetilde{U}_{t+1}=u_{t+1},\dots,\widetilde{U}_1=u_1} \\
&\quad= \widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(\widehat{X}_{t+1}),\widetilde{Y}_{t+1}\right)\right)|\widehat{Z}_t=z_t,\dots,\widehat{Z}_1=z_1,\widehat{X}_{t+1}=x_{t+1},\dots,\widehat{X}_1=x_1,\widetilde{U}_{t+1}=u_{t+1},\dots,\widetilde{U}_1=u_1} = \widetilde{\mathbb{P}}_{g\left(\widetilde{\varphi}\left(\widetilde{f}(x_{t+1}),\widetilde{Y}_{t+1}\right)\right)} \\
&\quad= \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(x_{t+1},Y_{t+1})} = \big(h(\widetilde{\mathbb{P}})\big)_{\varphi(X_{t+1},Y_{t+1})|Z_t=z_t,\dots,Z_1=z_1,X_{t+1}=x_{t+1},\dots,X_1=x_1,U_{t+1}=u_{t+1},\dots,U_1=u_1} \\
&\quad= \big(h(\widetilde{\mathbb{P}})\big)_{Z_{t+1}|Z_t=z_t,\dots,Z_1=z_1,X_{t+1}=x_{t+1},\dots,X_1=x_1,U_{t+1}=u_{t+1},\dots,U_1=u_1} \,.
\end{aligned}$$

So, if $A_{t+1} \subset \mathcal{Z}, D \subset \mathcal{Z}^t \times \mathcal{X}^{t+1} \times [0,1]^{t+1}$, we have that

$$\begin{aligned}
&\widetilde{\mathbb{P}}_{\widehat{Z}_{t+1},\left(\widehat{Z}_{t:1},\widehat{X}_{t+1:1},\widetilde{U}_{t+1:1}\right)}(A_{t+1} \times D) \\
&= \int_D \widetilde{\mathbb{P}}_{\widehat{Z}_{t+1}|\widehat{Z}_{t:1}=z_{t:1},\widehat{X}_{t+1:1}=x_{t+1:1},\widetilde{U}_{t+1:1}=u_{t+1:1}}(A_{t+1})\,\mathrm{d}\widetilde{\mathbb{P}}_{\widehat{Z}_{t:1},\widehat{X}_{t+1:1},\widetilde{U}_{t+1:1}}(z_{t:1}, x_{t+1:1}, u_{t+1:1}) \\
&= \int_D \big(h(\widetilde{\mathbb{P}})\big)_{Z_{t+1}|Z_{t:1}=z_{t:1},C_{t+1:1}=x_{t+1:1},U_{t+1:1}=u_{t+1:1}}(A_{t+1})\,\mathrm{d}\big(h(\widetilde{\mathbb{P}})\big)_{Z_{t:1},X_{t+1:1},U_{t+1:1}}(z_{t:1}, x_{t+1:1}, u_{t+1:1}) \\
&= \big(h(\widetilde{\mathbb{P}})\big)_{Z_{t+1},\left(Z_{t:1},X_{t+1:1},U_{t+1:1}\right)}(A_{t+1} \times D) \,,
\end{aligned}$$

from which follows that $\widetilde{\mathbb{P}}_{\widehat{Z}_{t+1},\dots,\widehat{Z}_1,\widehat{X}_{t+1},\dots,\widehat{X}_1,\widetilde{U}_{t+1},\dots,\widetilde{U}_1} = \big(h(\widetilde{\mathbb{P}})\big)_{Z_{t+1},\dots,Z_1,X_{t+1},\dots,X_1,U_{t+1},\dots,U_1}$. In particular, for each $t \in [T]$ we have that $\widetilde{\mathbb{P}}_{\widehat{X}_t} = \big(h(\widetilde{\mathbb{P}})\big)_{X_t}$. Hence, using the $h(\widetilde{\mathbb{P}})$-independence of $Y_1, \dots, Y_T$, Item (2), and the $\widetilde{\mathbb{P}}$-independence of $\widetilde{Y}_1, \dots, \widetilde{Y}_T$, we get

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(X_t, Y_t)\big] &= \sum_{t=1}^T \int_{\mathcal{X}} \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big]\,\mathrm{d}\big(h(\widetilde{\mathbb{P}})\big)_{X_t}(x) \\
&\leqslant \sum_{t=1}^T \int_{\mathcal{X}} \mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{\rho}(\widetilde{f}(x), \widetilde{Y}_t)]\,\mathrm{d}\big(h(\widetilde{\mathbb{P}})\big)_{X_t}(x) \\
&= \sum_{t=1}^T \int_{\mathcal{X}} \mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{\rho}(\widetilde{f}(x), \widetilde{Y}_t)]\,\mathrm{d}\widetilde{\mathbb{P}}_{\widehat{X}_t}(x) \\
&= \sum_{t=1}^T \mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{f}(\widehat{X}_t), \widetilde{Y}_t)\big] = \sum_{t=1}^T \mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{X}_t, \widetilde{Y}_t)\big] \,.
\end{aligned}$$

Then, using Item (1), we have

$$R_T\big(\alpha, h(\widetilde{\mathbb{P}})\big) = \sup_{x \in \mathcal{X}} \left(\sum_{t=1}^T \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(x, Y_t)\big] - \sum_{t=1}^T \mathbb{E}_{h(\widetilde{\mathbb{P}})}\big[\rho(X_t, Y_t)\big] \right)$$

$$\geqslant \sup_{\widetilde{x}\in\widetilde{\mathcal{X}}}\Bigg(\sum_{t=1}^{T}\mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{x},\widetilde{Y}_t)\big]-\sum_{t=1}^{T}\mathbb{E}_{\widetilde{\mathbb{P}}}\big[\widetilde{\rho}(\widetilde{X}_t,\widetilde{Y}_t)\big]\Bigg)=R_T(\widetilde{\alpha},\widetilde{\mathbb{P}})\ .$$

Since $\widetilde{\mathbb{P}}$ was arbitrary, we get

$$R_T^\star(\widetilde{\mathscr{G}})=\inf_{\beta\in\mathscr{A}(\widetilde{\mathscr{G}})}R_T^{\widetilde{\mathcal{S}}}(\beta)\leqslant R_T^{\widetilde{\mathcal{S}}}(\widetilde{\alpha})=\sup_{\widetilde{\mathbb{P}}\in\widetilde{\mathcal{S}}}R_T(\widetilde{\alpha},\widetilde{\mathbb{P}})\leqslant\sup_{\widetilde{\mathbb{P}}\in\widetilde{\mathcal{S}}}R_T\big(\alpha,h(\widetilde{\mathbb{P}})\big)\leqslant\sup_{\mathbb{P}\in\mathcal{S}}R_T\big(\alpha,\mathbb{P}\big)=R_T^{\mathcal{S}}(\alpha)\ ,$$

and since α was arbitrary, we get

$$R_T^\star(\widetilde{\mathscr{G}})\leqslant\inf_{\alpha\in\mathscr{A}(\mathscr{G})}R_T^{\mathcal{S}}(\alpha)=R_T^\star(\mathscr{G})\ .$$

□

We now prove the Simulation lemma we introduced in Appendix A.6 showing how to get rid of uninformative feedback.

Lemma 20 (Simulation)**.** *Let $\mathcal{V},\mathcal{W}$ be two sets, $\mathscr{G}:=(\mathcal{X},\mathcal{Y},\mathcal{Z},\rho,\varphi,\mathcal{P})$ a game with $\mathcal{Z}=\mathcal{V}\times\mathcal{W}$, $\mathcal{S}$ its set of environments, $(Y_t)_{t\in\mathbb{N}}$ its adversary's actions, $\pi\colon\mathcal{Z}\to\mathcal{V}$ the projection on $\mathcal{V}$, and $T\in\mathbb{N}$ a horizon. Assume that $Y_1,\dots,Y_T$ are $\mathbb{P}$-independent for any environment $\mathbb{P}\in\mathcal{S}$ and that there exist disjoint sets $\mathcal{I}$,$\mathcal{U}\subset\mathcal{X}$ such that $\mathcal{I}\cup\mathcal{U}=\mathcal{X}$ and*

1. *For any time $t\in[T]$ and action $x\in\mathcal{I}$ there exists $\psi_{t,x}\colon[0,1]\to\mathcal{W}$ such that, for all $\mathbb{P}\in\mathcal{S}$,*

$$\mathbb{P}_{\varphi(x,Y_t)}=\mathbb{P}_{\pi\left(\varphi(x,Y_t)\right)}\otimes(\mu_L)_{\psi_{t,x}}\ .$$

2. *For any time $t\in[T]$ and action $x\in\mathcal{U}$, there exists $\gamma_{t,x}\colon[0,1]\to\mathcal{Z}$ such that, for all $\mathbb{P}\in\mathcal{S}$,*

$$\mathbb{P}_{\varphi(x,Y_t)}=(\mu_L)_{\gamma_{t,x}}\ .$$

Let $*\in\mathcal{V}$ and define

$$\widetilde{\varphi}\colon\mathcal{X}\times\mathcal{Y}\to\mathcal{V},\ (x,y)\mapsto\begin{cases}\pi\big(\varphi(x,y)\big)\ , & \text{if } x\in\mathcal{I},\\ *\ , & \text{if } x\in\mathcal{U}.\end{cases}$$

Define the game $\widetilde{\mathscr{G}}:=(\mathcal{X},\mathcal{Y},\mathcal{V},\rho,\widetilde{\varphi},\mathcal{P})$. Then $R_T^\star(\mathscr{G})\geqslant R_T^\star(\widetilde{\mathscr{G}})$.

Proof. For each number $a\in[0,1]$, fix a binary representation $0.a_1a_2a_3\dots$ of a and define $\xi(a):=0.a_1a_3a_5\dots$, $\zeta(a):=0.a_2a_4a_6\dots$. Note that the two resulting functions $\xi,\zeta\colon[0,1]\to[0,1]$ are μ_L-independent with common (uniform) push-forward distribution $(\mu_L)_\xi=\mu_L=(\mu_L)_\zeta$.

Let $(Y_t)_{t\in\mathbb{N}},(U_t)_{t\in\mathbb{N}}$ be the sequences of adversary's actions and player's randomization for the sequential game $\mathscr{G}$ and note that they are also the same for the sequential game $\widetilde{\mathscr{G}}$. For each $t\in\mathbb{N}$ define $\beta_t\colon\mathcal{X}\times\mathcal{V}\times[0,1]\to\mathcal{Z}$ via

$$(x,v,u)\mapsto\begin{cases}\big(v,\psi_{t,x}(u)\big)\ , & \text{if } x\in\mathcal{I}\ ,\\ \gamma_{t,x}(u)\ , & \text{if } x\in\mathcal{U}\ ,\end{cases}$$

if $t \leqslant T$, and in an arbitrary manner if $t \geqslant T+1$. Fix $\alpha = (\alpha_t)_{t\in\mathbb{N}} \in \mathscr{A}(\mathscr{G})$. Let $(X_t)_{t\in\mathbb{N}}, (Z_t)_{t\in\mathbb{N}}$ be the sequences of player's actions and feedback associated to the strategy α.

Fix $(u_t)_{t\in\mathbb{N}} \subset [0,1]$ and $(v_t)_{t\in\mathbb{N}} \subset \mathcal{V}$. Define by induction (on t) the sequences $(x_t)_{t\in\mathbb{N}}$ and $(z_t)_{t\in\mathbb{N}}$ via the relationships

$$x_t = \alpha_t\big(\xi(u_1), \dots, \xi(u_t), z_1, \dots, z_{t-1}\big), \qquad z_t = \beta_t\big(x_t, v_t, \zeta(u_t)\big).$$

Note that for each $t \in \mathbb{N}$, we have that x_t depends only on $u_1, \dots, u_t, v_1, \dots, v_{t-1}$, so we can define

$$\widetilde{\alpha}_t(u_1, \dots, u_t, v_1, \dots, v_{t-1}) := x_t.$$

Being $(u_t)_{t\in\mathbb{N}}$ and $(v_t)_{t\in\mathbb{N}}$ arbitrary, this defines a sequence of functions $(\widetilde{\alpha}_t)_{t\in\mathbb{N}}$ such that, for all $t \in \mathbb{N}$,

$$\widetilde{\alpha}_t \colon [0,1]^t \times \mathcal{V}^{t-1} \to \mathcal{X}$$

i.e., $\widetilde{\alpha} := (\widetilde{\alpha}_t)_{t\in\mathbb{N}} \in \mathscr{A}(\widetilde{\mathscr{G}})$. Let $(\widetilde{X}_t)_{t\in\mathbb{N}}$ and $(\widetilde{V}_t)_{t\in\mathbb{N}}$ be respectively the sequence of player's actions and the feedback sequence associated with the strategy $\widetilde{\alpha}$. For each $t \in \mathbb{N}$, define also $\widetilde{Z}_t := \beta_t\big(\widetilde{X}_t, \widetilde{V}_t, \zeta(U_t)\big)$. Note that for each $t \in \mathbb{N}$ it holds that $\widetilde{X}_t = \alpha_t\big(\xi(U_1), \dots, \xi(U_t), \widetilde{Z}_1, \dots, \widetilde{Z}_{t-1}\big)$.

Fix a environment $\mathbb{P} \in \mathcal{S}$. Note first that $\mathbb{P}_{\xi(U_1)} = \mathbb{P}_{U_1}$, and since $X_1 = \alpha_1(U_1)$ and $\widetilde{X}_1 = \widetilde{\alpha}_1(U_1) = \alpha_1\big(\xi(U_1)\big)$, we also have that $\mathbb{P}_{\widetilde{X}_1,\xi(U_1)} = \mathbb{P}_{X_1,U_1} =: \mathbb{Q}_1$. Now, up to a set with $\mathbb{Q}_1$-probability zero, if $x_1 \in \mathcal{X}$ and $u_1 \in [0,1]$, using Items (1) and (2), we have that

$$\begin{aligned}
&\mathbb{P}_{\widetilde{Z}_1|\widetilde{X}_1=x_1,\xi(U_1)=u_1} = \mathbb{P}_{\beta_1\left(\widetilde{X}_1,\widetilde{\varphi}(\widetilde{X}_1,Y_1),\zeta(U_1)\right)|\widetilde{X}_1=x_1,\xi(U_1)=u_1} = \mathbb{P}_{\beta_1\left(x_1,\widetilde{\varphi}(x_1,Y_1),\zeta(U_1)\right)} \\
&= \begin{cases} \mathbb{P}_{\beta_1\left(x_1,\pi\left(\varphi(x_1,Y_1)\right),\zeta(U_1)\right)} & \text{if } x_1 \in \mathcal{I} \\ \mathbb{P}_{\beta_1\left(x_1,*,\zeta(U_1)\right)} & \text{if } x_1 \in \mathcal{U} \end{cases} = \begin{cases} \mathbb{P}_{\left(\pi\left(\varphi(x_1,Y_1)\right),\psi_{1,x_1}\left(\zeta(U_1)\right)\right)} & \text{if } x_1 \in \mathcal{I} \\ \mathbb{P}_{\gamma_{1,x_1}\left(\zeta(U_1)\right)} & \text{if } x_1 \in \mathcal{U} \end{cases} \\
&= \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_1,Y_1)\right)} \otimes \mathbb{P}_{\psi_{1,x_1}\left(\zeta(U_1)\right)} & \text{if } x_1 \in \mathcal{I} \\ \mathbb{P}_{\gamma_{1,x_1}\left(\zeta(U_1)\right)} & \text{if } x_1 \in \mathcal{U} \end{cases} = \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_1,Y_1)\right)} \otimes \left(\mathbb{P}_{\zeta(U_1)}\right)_{\psi_{1,x_1}} & \text{if } x_1 \in \mathcal{I} \\ \left(\mathbb{P}_{\zeta(U_1)}\right)_{\gamma_{1,x_1}} & \text{if } x_1 \in \mathcal{U} \end{cases} \\
&= \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_1,Y_1)\right)} \otimes \left(\mu_L\right)_{\psi_{1,x_1}} & \text{if } x_1 \in \mathcal{I} \\ \left(\mu_L\right)_{\gamma_{1,x_1}} & \text{if } x_1 \in \mathcal{U} \end{cases} = \mathbb{P}_{\varphi(x_1,Y_1)} = \mathbb{P}_{\varphi(X_1,Y_1)|X_1=x_1,U_1=u_1} = \mathbb{P}_{Z_1|X_1=x_1,U_1=u_1}\,.
\end{aligned}$$

So, if $A_1 \subset \mathcal{Z}$ and $D \subset \mathcal{X} \times [0,1]$, then

$$\begin{aligned}
\mathbb{P}_{\widetilde{Z}_1,\left(\widetilde{X}_1,\xi(U_1)\right)}(A_1 \times D) &= \int_D \mathbb{P}_{\widetilde{Z}_1|\widetilde{X}_1=x_1,\xi(U_1)=u_1}(A_1)\,\mathrm{d}\mathbb{P}_{\widetilde{X}_1,\xi(U_1)}(x_1,u_1) \\
&= \int_D \mathbb{P}_{Z_1|X_1=x_1,U_1=u_1}(A_1)\,\mathrm{d}\mathbb{P}_{X_1,U_1}(x_1,u_1) = \mathbb{P}_{Z_1,(X_1,U_1)}(A_1 \times D)\,,
\end{aligned}$$

from which it follows that $\mathbb{P}_{\widetilde{Z}_1,\widetilde{X}_1,\xi(U_1)} = \mathbb{P}_{Z_1,X_1,U_1}$. By induction, suppose that for $t \in [T-1]$ we have that

$$\mathbb{P}_{\widetilde{Z}_t,\dots,\widetilde{Z}_1,\widetilde{X}_t,\dots,\widetilde{X}_1,\xi(U_t),\dots,\xi(U_1)} = \mathbb{P}_{Z_t,\dots,Z_1,X_t,\dots,X_1,U_t,\dots,U_1}\,.$$

Then, using independence we have that

$$\mathbb{P}_{\widetilde{Z}_t,\ldots,\widetilde{Z}_1,\widetilde{X}_t,\ldots,\widetilde{X}_1,\xi(U_{t+1}),\xi(U_t),\ldots,\xi(U_1)} = \mathbb{P}_{Z_t,\ldots,Z_1,X_t,\ldots,X_1,U_{t+1},U_t,\ldots,U_1} \,.$$

Furthermore, since $X_{t+1} = \alpha_{t+1}(U_1,\ldots,U_{t+1},Z_1,\ldots,Z_t)$ and $\widetilde{X}_{t+1} = \widetilde{\alpha}_{t+1}(U_1,\ldots,U_{t+1},\widetilde{V}_1,\ldots,\widetilde{V}_t) = \alpha_{t+1}(\xi(U_1),\ldots,\xi(U_{t+1}),\widetilde{Z}_1,\ldots,\widetilde{Z}_t)$, we have that

$$\mathbb{P}_{\widetilde{Z}_t,\ldots,\widetilde{Z}_1,\widetilde{X}_{t+1},\widetilde{X}_t,\ldots,\widetilde{X}_1,\xi(U_{t+1}),\xi(U_t),\ldots,\xi(U_1)} = \mathbb{P}_{Z_t,\ldots,Z_1,X_{t+1},X_t,\ldots,X_1,U_{t+1},U_t,\ldots,U_1} =: \mathbb{Q}_{t+1} \,.$$

Now, up to a set with $\mathbb{Q}_{t+1}$-probability zero, if $x_1,\ldots,x_{t+1} \in \mathcal{X}$, $u_1,\ldots,u_{t+1} \in [0,1]$ and $z_1,\ldots,z_t \in \mathcal{Z}$, using the $\mathbb{P}$-independence of $Y_1,\ldots,Y_{t+1}$ and Items (1)–(2), we have that

$$\begin{aligned}
&\mathbb{P}_{\widetilde{Z}_{t+1}|\widetilde{Z}_t=z_t,\ldots,\widetilde{Z}_1=z_1,\widetilde{X}_{t+1}=x_{t+1},\ldots,\widetilde{X}_1=x_1,\xi(U_{t+1})=u_{t+1},\ldots,\xi(U_1)=u_1} \\
&\quad= \mathbb{P}_{\beta_{t+1}\left(\widetilde{X}_{t+1},\widetilde{\varphi}(\widetilde{X}_{t+1},Y_{t+1}),\zeta(U_{t+1})\right)|\widetilde{Z}_t=z_t,\ldots,\widetilde{Z}_1=z_1,\widetilde{X}_{t+1}=x_{t+1},\ldots,\widetilde{X}_1=x_1,\xi(U_{t+1})=u_{t+1},\ldots,\xi(U_1)=u_1} \\
&\quad= \mathbb{P}_{\beta_{t+1}\left(x_{t+1},\widetilde{\varphi}(x_{t+1},Y_{t+1}),\zeta(U_{t+1})\right)} = \begin{cases} \mathbb{P}_{\beta_{t+1}\left(x_{t+1},\pi\left(\varphi(x_{t+1},Y_{t+1})\right),\zeta(U_{t+1})\right)} & \text{if } x_{t+1} \in \mathcal{I} \\ \mathbb{P}_{\beta_{t+1}\left(x_{t+1},*,\zeta(U_{t+1})\right)} & \text{if } x_{t+1} \in \mathcal{U} \end{cases} \\
&\quad= \begin{cases} \mathbb{P}_{\left(\pi\left(\varphi(x_{t+1},Y_{t+1})\right),\psi_{t+1,x_{t+1}}\left(\zeta(U_{t+1})\right)\right)} & \text{if } x_{t+1} \in \mathcal{I} \\ \mathbb{P}_{\gamma_{t+1,x_{t+1}}\left(\zeta(U_{t+1})\right)} & \text{if } x_{t+1} \in \mathcal{U} \end{cases} \\
&\quad= \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_{t+1},Y_{t+1})\right)} \otimes \mathbb{P}_{\psi_{t+1,x_{t+1}}\left(\zeta(U_{t+1})\right)} & \text{if } x_{t+1} \in \mathcal{I} \\ \mathbb{P}_{\gamma_{t+1,x_{t+1}}\left(\zeta(U_{t+1})\right)} & \text{if } x_{t+1} \in \mathcal{U} \end{cases} \\
&\quad= \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_{t+1},Y_{t+1})\right)} \otimes \left(\mathbb{P}_{\zeta(U_{t+1})}\right)_{\psi_{t+1,x_{t+1}}} & \text{if } x_{t+1} \in \mathcal{I} \\ \left(\mathbb{P}_{\zeta(U_{t+1})}\right)_{\gamma_{t+1,x_{t+1}}} & \text{if } x_{t+1} \in \mathcal{U} \end{cases} \\
&\quad= \begin{cases} \mathbb{P}_{\pi\left(\varphi(x_{t+1},Y_{t+1})\right)} \otimes \left(\mu_L\right)_{\psi_{t+1,x_{t+1}}} & \text{if } x_{t+1} \in \mathcal{I} \\ \left(\mu_L\right)_{\gamma_{t+1,x_{t+1}}} & \text{if } x_{t+1} \in \mathcal{U} \end{cases} \\
&\quad= \mathbb{P}_{\varphi(x_{t+1},Y_{t+1})} = \mathbb{P}_{\varphi(X_{t+1},Y_{t+1})|Z_t=z_t,\ldots,Z_1=z_1,X_{t+1}=x_{t+1},\ldots,X_1=x_1,U_{t+1}=u_{t+1},\ldots,U_1=u_1} \\
&\quad= \mathbb{P}_{Z_{t+1}|Z_t=z_t,\ldots,Z_1=z_1,X_{t+1}=x_{t+1},\ldots,X_1=x_1,U_{t+1}=u_{t+1},\ldots,U_1=u_1} \,.
\end{aligned}$$

So, if $A_{t+1} \subset \mathcal{Z}$, $D \subset \mathcal{Z}^t \times \mathcal{X}^{t+1} \times [0,1]^{t+1}$, we have that

$$\begin{aligned}
&\mathbb{P}_{\widetilde{Z}_{t+1},\left(\widetilde{Z}_t\ldots,\widetilde{Z}_1,\widetilde{X}_{t+1},\ldots,\widetilde{X}_1,\xi(U_{t+1}),\ldots,\xi(U_1)\right)}(A_{t+1} \times D) \\
&\quad= \int_D \mathbb{P}_{\widetilde{Z}_{t+1}|\widetilde{Z}_{t:1}=z_{t:1},\widetilde{X}_{t+1:1}=x_{t+1:1},\left(\xi(U_{t+1}),\ldots,\xi(U_1)\right)=u_{t+1:1}}(A_{t+1}) \,\mathrm{d}\mathbb{Q}_{t+1}(z_{t:1},x_{t+1:1},u_{t+1:1}) \\
&\quad= \int_D \mathbb{P}_{Z_{t+1}|Z_{t:1}=z_{t:1},X_{t+1:1}=x_{t+1:1},U_{t+1:1}=u_{t+1:1}}(A_{t+1}) \,\mathrm{d}\mathbb{Q}_{t+1}(z_{t:1},x_{t+1:1},u_{t+1:1}) \\
&\quad= \mathbb{P}_{Z_{t+1},\left(Z_t,\ldots,Z_1,X_{t+1},\ldots,X_1,U_{t+1},\ldots,U_1\right)}(A_{t+1} \times D)
\end{aligned}$$

from which it follows that $\mathbb{P}_{\widetilde{Z}_{t+1:1},\widetilde{X}_{t+1:1},\left(\xi(U_{t+1}),\ldots,\xi(U_1)\right)} = \mathbb{P}_{Z_{t+1:1},X_{t+1:1},U_{t+1:1}}$. In particular, for each

$t \in [T]$ we have that $\mathbb{P}_{X_t} = \mathbb{P}_{\widetilde{X}_t}$. So, for each $t \in [T]$, using the $\mathbb{P}$-independence of $Y_1, \dots, Y_t$, we have that

$$\mathbb{P}_{X_t,Y_t} = \mathbb{P}_{X_t} \otimes \mathbb{P}_{Y_t} = \mathbb{P}_{\widetilde{X}_t} \otimes \mathbb{P}_{Y_t} = \mathbb{P}_{\widetilde{X}_t,Y_t} ,$$

and then

$$\mathbb{E}_{\mathbb{P}}\big[\rho(X_t, Y_t)\big] = \mathbb{E}_{\mathbb{P}_{X_t,Y_t}}\big[\rho\big] = \mathbb{E}_{\mathbb{P}_{\widetilde{X}_t,Y_t}}\big[\rho\big] = \mathbb{E}_{\mathbb{P}}\big[\rho(\widetilde{X}_t, Y_t)\big] .$$

In conclusion

$$R_T(\alpha, \mathbb{P}) = \sup_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \rho(x, Y_t) - \sum_{t=1}^{T} \rho(X_t, Y_t)\right] = \sup_{x\in\mathcal{X}}\left(\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}}\left[\rho(x, Y_t)\right] - \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}}\left[\rho(X_t, Y_t)\right]\right)$$
$$= \sup_{x\in\mathcal{X}}\left(\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}}\left[\rho(x, Y_t)\right] - \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}}\left[\rho(\widetilde{X}_t, Y_t)\right]\right) = \sup_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \rho(x, Y_t) - \sum_{t=1}^{T} \rho(\widetilde{X}_t, Y_t)\right] = R_T(\widetilde{\alpha}, \mathbb{P}) .$$

Since $\mathbb{P}$ was arbitrary, it follows that $R_T^{\mathcal{S}}(\alpha) = R_T^{\mathcal{S}}(\widetilde{\alpha})$. Since α was arbitrary, it follows that

$$R_T^\star(\mathscr{G}) = \inf_{\alpha\in\mathscr{A}(\mathscr{G})} R_T^{\mathcal{S}}(\alpha) = \inf_{\alpha\in\mathscr{A}(\mathscr{G})} R_T^{\mathcal{S}}(\widetilde{\alpha}) \geqslant \inf_{\alpha'\in\mathscr{A}(\widetilde{\mathscr{G}})} R_T^{\mathcal{S}}(\alpha') = R_T^\star(\widetilde{\mathscr{G}}) .$$

□

A.7 $\sqrt{T}$ Lower Bound under Full-Feedback (iid+iv+bd)

In this section, we prove that in the full-feedback case, no strategy can beat the $\sqrt{T}$ rate that we proved in Theorem 3 when the seller/buyer pair (S_t, B_t) is drawn i.i.d. from an unknown fixed distribution, not even under the further assumptions that the valuations of the seller and buyer are independent of each other and have bounded densities.

The idea of the proof is to build a family of environments $\mathbb{P}^{\pm\varepsilon}$ parameterized by $\varepsilon \in [0, 1]$, like in Figure 2.1. The only way to avoid suffering $\Omega(\varepsilon T)$ regret in an environment $\mathbb{P}^{\pm\varepsilon}$ is to identify the sign of $\pm\varepsilon$. Leveraging the Embedding and Simulation lemmas (Lemmas 19 and 20), this construction leads to a reduction to a two-action expert problem, which has a know lower bound on the regret of order $\sqrt{T}$.

Theorem 36 (Theorem 4, restated)**.** *With the same notation as in Appendix A.5.2, in the full-feedback stochastic (iid) setting with independent valuations (iv) and densities bounded by a constant $M \geqslant 4$ (bd), for all horizons $T \in \mathbb{N}$, the minimax regret satisfies*

$$R_T^\star = \Omega\big(\sqrt{T}\big) .$$

Proof. Fix an arbitrary horizon $T \in \mathbb{N}$ and any $M \geqslant 4$. Recalling Appendix A.5.2, the full-feedback stochastic (iid) setting with independent valuations (iv) and densities bounded (bd) by M is a game $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$, where $\mathcal{X} = [0, 1]$, $\mathcal{Y} = [0, 1]^2$, $\mathcal{Z} = [0, 1]^2$, $\rho = \mathrm{gft}$, $\varphi\colon \big(p, (s, b)\big) \mapsto (s, b)$, and $\mathcal{P} = \mathcal{P}^M_{\text{iid+iv+bd}}$. Define, for each $\varepsilon \in [-1, 1]$, the densities $f_{S,\varepsilon} = 2(1+\varepsilon)\mathbb{I}_{[0,\frac{1}{4}]} + 2(1-\varepsilon)\mathbb{I}_{[\frac{1}{2},\frac{3}{4}]}$ and $f_B = 2\mathbb{I}_{[\frac{1}{4},\frac{1}{2}]\cup[\frac{3}{4},1]}$. Fix the adversary's behavior $\mathcal{P}_1$ as the subset of $\mathcal{P}$ whose elements have the form $\boldsymbol{\mu}_\varepsilon := \otimes_{t\in\mathbb{N}}(f_{S,\varepsilon}\mu_L \otimes f_B\mu_L)$, for some $\varepsilon \in [-1, 1]$. Since $\mathcal{P}_1 \subset \mathcal{P}$, the game $\mathscr{G}_1 := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P}_1)$ is easier than $\mathscr{G}$ (i.e., $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\mathscr{G}_1)$) by the Embedding lemma

(Lemma 19) with $\widetilde{f}$ and g as the identities, and h as the inclusion. Now, define $\rho_1 \colon \mathcal{X} \times \mathcal{Y} \to [0,1]$, $(p,(s,b)) \mapsto (b-s)\mathbb{I}\{s \leqslant \frac{1}{4} \leqslant b\}\,\mathbb{I}\{p \leqslant \frac{1}{2}\} + (b-s)\mathbb{I}\{s \leqslant \frac{3}{4} \leqslant b\}\,\mathbb{I}\{p > \frac{1}{2}\}$ and note that, defining $\mathscr{G}_2 := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho_1, \varphi, \mathcal{P}_1)$, by the Embedding lemma with $\widetilde{f}, g, h$ as the identities, we have that the game $\mathscr{G}_2$ is easier than the game $\mathscr{G}_1$ (i.e., $R_T^\star(\mathscr{G}_1) \geqslant R_T^\star(\mathscr{G}_2)$). Then, let $\mathcal{Z}_3 := \{0,1\} \times \big[0, \frac{1}{4}\big] \times [0,1]$ and $\varphi_3 \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_3$, $\big(p,(s,b)\big) \mapsto \big(\mathbb{I}\{s \leqslant 1/4\},\, s\mathbb{I}\{s \leqslant 1/4\} + (s-1/2)\mathbb{I}\{1/2 \leqslant s \leqslant 3/4\},\, b\big)$. Define the game $\mathscr{G}_3 := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}_3, \rho_1, \varphi_3, \mathcal{P}_1)$. By the Embedding lemma with $\widetilde{f}, h$ as the identities and $g \colon \mathcal{Z}_3 \to \mathcal{Z}$, $(i, \widetilde{s}, b) \mapsto \big(\widetilde{s}i + (1/2 + \widetilde{s})(1-i),\, b\big)$, we have that the game $\mathscr{G}_3$ is easier than the game $\mathscr{G}_2$ (i.e., $R_T^\star(\mathscr{G}_2) \geqslant R_T^\star(\mathscr{G}_3)$). Next, let $\varphi_4 \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_3$, $\big(p,(s,b)\big) \mapsto \mathbb{I}\{s \leqslant \frac{1}{4}\}$, and define the game $\mathscr{G}_4 := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}_3, \rho_1, \varphi_4, \mathcal{P}_1)$. Let $(Y_t)_{t\in\mathbb{N}}$ be the adversary's actions in $\mathscr{G}_4$. A tedious computation verifies that for all $t \in \mathbb{N}$, $p \in \mathcal{X}$, and environments $\mathbb{P}$ of game $\mathscr{G}_3$, $\mathbb{P}_{\varphi_3(p,Y_t)} = \mathbb{P}_{\pi(\varphi_3(p,Y_t))} \otimes (\nu \otimes f_B \mu_L)$, where $\pi \colon \mathcal{Z}_3 \to \{0,1\}$ is the projection on the first component $\{0,1\}$ of $\mathcal{Z}_3$ and ν is the uniform distribution on $[0, 1/4]$. By the well-known Skorokhod representation [185, Section 17.3], there exists $\psi \colon [0,1] \to [0,1/4] \times [0,1]$ such that $\nu \otimes f_B \mu_L = (\mu_L)_\psi$. Thus, by the Simulation lemma (Lemma 20) with $\mathcal{I} = \mathcal{X}$ and $\mathcal{U} = \varnothing$, the game $\mathscr{G}_4$ is easier than $\mathscr{G}_3$ (i.e., $R_T^\star(\mathscr{G}_3) \geqslant R_T^\star(\mathscr{G}_4)$). Finally, consider the game $\mathscr{G}_5 := \big(\{1,2\}, \{1,2\}, \{0,1\}, \rho_5, \varphi_5, \mathcal{P}_5\big)$, where in matrix notation, $\rho_5 = \big[\rho_5(i,j)\big]_{i,j\in\{1,2\}}$ and $\varphi_5 = \big[\varphi_5(i,j)\big]_{i,j\in\{1,2\}}$ are given by

$$\rho_5 := \begin{bmatrix} 1/2 & 3/8 \\ 3/8 & 1/2 \end{bmatrix}, \qquad \varphi_5 := \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

and $\mathbb{P}_5$ is the set of all measures $\widetilde{\boldsymbol{\mu}}_\varepsilon$ of the form $\widetilde{\boldsymbol{\mu}}_\varepsilon = \otimes_{t=1}^\infty \big(\frac{1+\varepsilon}{2}\delta_1 + \frac{1-\varepsilon}{2}\delta_2\big)$ for some $\varepsilon \in [-1,1]$, where δ_i is the Dirac measure at $i \in \{1,2\}$. Thus, letting $\mathcal{S}_4$ and $\mathcal{S}_5$ be the two sets of environments in games $\mathscr{G}_4$ and $\mathscr{G}_5$ respectively (note that $\mathcal{S}_4$ coincides with the set of environments of $\mathscr{G}_1$) and using again the Embedding lemma, this time with $\widetilde{f} \colon [0,1] \to \{1,2\}$, $p \mapsto \mathbb{I}\{p \leqslant 1/2\} + 2\mathbb{I}\{p > 1/2\}$, $g \colon \{0,1\} \to \{0,1\}$, $i \mapsto i$, and $h \colon \mathcal{S}_5 \to \mathcal{S}_4$, $\widetilde{\boldsymbol{\mu}}_\varepsilon \otimes \boldsymbol{\mu}_L \mapsto \boldsymbol{\mu}_\varepsilon \otimes \boldsymbol{\mu}_L$, we obtain that $\mathscr{G}_5$ is easier than $\mathscr{G}_4$ (i.e., $R_T^\star(\mathscr{G}_4) \geqslant R_T^\star(\mathscr{G}_5)$). This last game $\mathscr{G}_5$ is an online learning problem with full information (also known as learning with expert advice), whose minimax regret is known to be lower bounded by $\frac{1}{8\sqrt{2\pi}}\sqrt{T}$ [70]. In conclusion, we proved that $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\mathscr{G}_5) \geqslant \frac{1}{8\sqrt{2\pi}}\sqrt{T}$. □

A.8 Proof of $T^{2/3}$ Lower Bound under Realistic Feedback (iid+iv+bd)

In this section we give a detailed proof of our $T^{2/3}$ lower bound of Section 2.4.2 which hinges in a non-trivial way on our Embedding and Simulation lemmas (Lemmas 19 and 20). We denote Bernoulli distributions with parameter λ by Ber_λ.

Theorem 37 (Theorem 6, restated)**.** *With the same notation as in Appendix A.5.2, in the realistic-feedback stochastic (iid) setting with independent valuations (iv) and densities bounded by a constant $M \geqslant 24$ (bd), for all horizons $T \in \mathbb{N}$, the minimax regret satisfies*

$$R_T^\star \geqslant \frac{11}{672} T^{2/3}.$$

Proof. Fix an arbitrary horizon $T \in \mathbb{N}$ and any $M \geqslant 24$. Recalling Appendix A.5.2, the realistic-feedback stochastic (iid) setting with independent valuations (iv) and densities bounded (bd) by M is a game $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$, where $\mathcal{X} = [0,1]$, $\mathcal{Y} = [0,1]^2$, $\mathcal{Z} = \{0,1\}^2$, $\rho = \mathrm{gft}$,

$\varphi\colon \big(p,(s,b)\big) \mapsto \big(\mathbb{I}\{s \leqslant p\}, \mathbb{I}\{p \leqslant b\}\big)$, and $\mathcal{P} = \mathcal{P}^M_{\text{iid+iv+bd}}$. The idea of the proof is to build a sequence of games, each one easier than the former, the last of which has a known lower bound on its minimax regret. In the first step we limit the adversary's behavior to a parametric family which is easily manageable and well-represents the difficulty of the problem (see Figure 2.2). In the second step, we increase the reward of suboptimal actions in order to have only three possible expected-reward values in each environment. In the third and fifth steps we increase the feedback, presenting it in a way that highlights that only its first component is informative. In step four and six, we simulate-away the uninformative parts of the feedback. Finally, in step 7 we show that the resulting game is harder than a known partial monitoring game with minimax regret of order at least $T^{2/3}$.

Step 1. Let $\vartheta := 1/48$. Define the following densities of the seller and buyer, respectively, by

$$f_{S,\varepsilon} := \frac{1}{4\vartheta}\left((1+\varepsilon)\mathbb{I}_{[0,\vartheta]} + (1-\varepsilon)\mathbb{I}_{\left[\frac{1}{6},\frac{1}{6}+\vartheta\right]} + \mathbb{I}_{\left[\frac{1}{4},\frac{1}{4}+\vartheta\right]} + \mathbb{I}_{\left[\frac{2}{3},\frac{2}{3}+\vartheta\right]}\right) , \forall \varepsilon \in [-1,1] ,$$
$$f_B := \frac{1}{4\vartheta}\left(\mathbb{I}_{\left[\frac{1}{3}-\vartheta,\frac{1}{3}\right]} + \mathbb{I}_{\left[\frac{3}{4}-\vartheta,\frac{3}{4}\right]} + \mathbb{I}_{\left[\frac{5}{6}-\vartheta,\frac{5}{6}\right]} + \mathbb{I}_{[1-\vartheta,1]}\right) .$$

Note that $f_{S,\varepsilon}$ corresponds to red/blue in Figure 2.2, while f_B to the green part. Define $\mathcal{P}_1$ as the subset of $\mathcal{P}$ whose elements have the form $\boldsymbol{\mu}_\varepsilon := \otimes_{t\in\mathbb{N}}(f_{S,\varepsilon}\mu_L \otimes f_B\mu_L)$ for $\varepsilon \in [-1,1]$. Since $\mathcal{P}_1 \subset \mathcal{P}$, the game $\mathscr{G}_1 := (\mathcal{X},\mathcal{Y},\mathcal{Z},\rho,\varphi,\mathcal{P}_1)$ is easier than $\mathscr{G}$ (i.e., $R^\star_T(\mathscr{G}) \geqslant R^\star_T(\mathscr{G}_1)$) by the Embedding lemma (Lemma 19) with $\widetilde{f}$ and g as the identities, and h as the inclusion.

Step 2. Define $\rho_2\colon \mathcal{X}\times\mathcal{Y} \to [0,1]$, $\big(p,(s,b)\big) \mapsto \mathrm{gft}\big(\frac{1}{6}+\vartheta,(s,b)\big)\mathbb{I}\big\{p<\frac{1}{4}\big\} + \mathrm{gft}\big(\frac{1}{4}+\vartheta,(s,b)\big)\mathbb{I}\big\{\frac{1}{4}\leqslant p<\frac{1}{3}\big\} + \mathrm{gft}\big(\frac{2}{3}+\vartheta,(s,b)\big)\mathbb{I}\big\{\frac{1}{3}<p\big\}$. By the Embedding lemma with $\widetilde{f}$, g, and h as the identities, we have that the game $\mathscr{G}_2 := (\mathcal{X},\mathcal{Y},\mathcal{Z},\rho_2,\varphi,\mathcal{P}_1)$ is easier than $\mathscr{G}_1$ (i.e., $R^\star_T(\mathscr{G}_1) \geqslant R^\star_T(\mathscr{G}_2)$).

Step 3. Define $\mathcal{Z}_3 := \big\{0,\frac{1}{6},\frac{1}{4},\frac{2}{3}\big\}\times[0,\vartheta]\times\{0,1\}\times\{0,1\}\times\mathcal{X}$ and $\varphi_3\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}_3$,

$$\big(p,(s,b)\big) \mapsto \begin{cases} \big(\eta(s), s-\eta(s), 0, \mathbb{I}\{p\leqslant b\}, p\big) , & \text{if } p<\frac{1}{4} , \\ \big(0,0,\mathbb{I}\{s\leqslant p\},\mathbb{I}\{p\leqslant b\},p\big) , & \text{if } p\geqslant\frac{1}{4} , \end{cases}$$

where $\eta\colon[0,1]\to\big\{0,\frac{1}{6},\frac{1}{4},\frac{2}{3}\big\}$, $s\mapsto \frac{1}{6}\mathbb{I}\big\{\frac{1}{6}\leqslant s\leqslant\frac{1}{6}+\vartheta\big\} + \frac{1}{4}\mathbb{I}\big\{\frac{1}{4}\leqslant s\leqslant\frac{1}{4}+\vartheta\big\} + \frac{2}{3}\mathbb{I}\big\{\frac{2}{3}\leqslant s\leqslant\frac{2}{3}+\vartheta\big\}$. Define the game $\mathscr{G}_3 := (\mathcal{X},\mathcal{Y},\mathcal{Z}_3,\rho_2,\varphi_3,\mathcal{P}_1)$. By the Embedding lemma with $\widetilde{f}, h$ as the identities and

$$g\colon \mathcal{Z}_3\to\mathcal{Z} , \quad (v,u,i,j,p)\mapsto \begin{cases} \big(\mathbb{I}\{v+u\leqslant p\},j\big) & \text{if } p<\frac{1}{4} , \\ (i,j) , & \text{if } p\geqslant\frac{1}{4} , \end{cases}$$

we have that the game $\mathscr{G}_3$ is easier than $\mathscr{G}_2$ (i.e., $R^\star_T(\mathscr{G}_2)\geqslant R^\star_T(\mathscr{G}_3)$).

Step 4. Let $\mathcal{Z}_4 := \big\{0,\frac{1}{6},\frac{1}{4},\frac{2}{3}\big\}$ and $\varphi_4\colon\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}_4$, $\big(p,(s,b)\big)\mapsto \eta(s)\mathbb{I}\big\{p<\frac{1}{4}\big\}$. Define the game $\mathscr{G}_4 := (\mathcal{X},\mathcal{Y},\mathcal{Z}_4,\rho_2,\varphi_4,\mathcal{P}_1)$. Let $(Y_t)_{t\in\mathbb{N}} = (S_t,B_t)_{t\in\mathbb{N}}$ be the adversary's actions in $\mathscr{G}_4$, $E := \big[0,\vartheta\big]\cup\big[\frac{1}{6},\frac{1}{6}+\vartheta\big]\cup\big[\frac{1}{4},\frac{1}{4}+\vartheta\big]\cup\big[\frac{2}{3},\frac{2}{3}+\vartheta\big]$ and $F := \big[\frac{1}{3}-\vartheta,\frac{1}{3}\big]\cup\big[\frac{3}{4}-\vartheta,\frac{3}{4}\big]\cup\big[\frac{5}{6}-\vartheta,\frac{5}{6}\big]\cup\big[1-\vartheta,1\big]$. A long and tedious computation verifies that for all $t\in\mathbb{N}$,

- For each $p\in[0,1/4)$ and any environment $\mathbb{P}$ of game $\mathscr{G}_3$, $\mathbb{P}_{\varphi_3(p,Y_t)} = \mathbb{P}_{\eta(S_t)}\otimes(\nu\otimes\delta_0\otimes\mathrm{Ber}_{\lambda_{F,p}}\otimes\delta_p)$,

where ν is the uniform distribution on $[0,\vartheta]$ and $\lambda_{F,p} := \frac{1}{4\vartheta}\mu_L\big([p,1]\cap F\big)$. By the well-known Skorokhod representation [185, Section 17.3], there exists $\psi_p\colon [0,1] \to [0,\vartheta]\times\{0,1\}\times\{0,1\}\times\mathcal{X}$ such that $\nu\otimes\delta_0\otimes\mathrm{Ber}_{\lambda_{F,p}}\otimes\delta_p = (\mu_L)_{\psi_p}$.

- For each $p\in[1/4,1]$ and any environment $\mathbb{P}$ of game $\mathscr{G}_3$, $\mathbb{P}_{\varphi_3(p,Y_t)} = \delta_0\otimes\delta_0\otimes\mathrm{Ber}_{\lambda_{E,p}}\otimes\mathrm{Ber}_{\lambda_{F,p}}\otimes\delta_p$, where $\lambda_{E,p} := \frac{1}{4\vartheta}\mu_L\big([0,p]\cap E\big]$ and $\lambda_{F,p} := \frac{1}{4\vartheta}\mu_L\big([p,1]\cap F\big]$. By the Skorokhod representation, there exists $\gamma_p\colon[0,1]\to\mathcal{Z}_3$ such that $\delta_0\otimes\delta_0\otimes\mathrm{Ber}_{\lambda_{E,p}}\otimes\mathrm{Ber}_{\lambda_{F,p}}\otimes\delta_p = (\mu_L)_{\gamma_p}$.

Thus, by the Simulation lemma (Lemma 20) with $\mathcal{I}=[0,1/4)$ and $\mathcal{U}=[1/4,1]$, the game $\mathscr{G}_4$ is easier than $\mathscr{G}_3$ (i.e., $R_T^\star(\mathscr{G}_3)\geqslant R_T^\star(\mathscr{G}_4)$).

Step 5. Let $\mathcal{Y}_5 := \mathcal{Y}^{\mathbb{N}}$, $\mathcal{Z}_5 := \{0,1\}\times\big(\mathbb{N}\cup\{\infty\}\big)\times\{0,1\}\times\mathcal{X}$, $\rho_5\colon\mathcal{X}\times\mathcal{Y}_5\to[0,1]$, $\big(p,(s_k,b_k)_{k\in\mathbb{N}}\big)\mapsto\rho_2(p,s_1,b_1)$,

$$\varphi_5\colon\mathcal{X}\times\mathcal{Y}_5\to\mathcal{Z}_5\ ,\quad \big(p,(s_k,b_k)_{k\in\mathbb{N}}\big)\mapsto\begin{cases}\Big(\mathbb{I}\big\{\eta(s_\tau)=0\big\},\tau,\mathbb{I}\big\{\eta(s_1)=\frac14\big\},p\Big)\ , & \text{if } p\in\big[0,\frac14\big)\ ,\\ (0,1,0,p)\ , & \text{if } p\in\big[\frac14,1\big]\ ,\end{cases}$$

where η is defined in game $\mathscr{G}_3$, $\tau := \inf\big\{k\in\mathbb{N}\mid\eta(s_k)\in\{0,1/6\}\big\}\in\mathbb{N}\cup\{\infty\}$, and $s_\infty := 0$. Let $\mathcal{P}_5$ be the set of measures on $\mathcal{Y}_5^{\mathbb{N}}$ of the form $\widetilde{\boldsymbol{\mu}}_\varepsilon := \otimes_{t\in\mathbb{N}}\big(\otimes_{k\in\mathbb{N}}(f_{S,\varepsilon}\mu_L\otimes f_B\mu_L)\big)$ for $\varepsilon\in[-1,1]$, and define the game $\mathscr{G}_5 := (\mathcal{X},\mathcal{Y}_5,\mathcal{Z}_5,\rho_5,\varphi_5,\mathcal{P}_5)$. By the Embedding lemma with $\widetilde{f}$ as the identity,

$$g\colon\mathcal{Z}_5\to\mathcal{Z}_4\ ,\quad (z,k,j,p)\mapsto\frac16(1-z)\mathbb{I}\left\{p<\frac14,k=1\right\}+\left(\frac14 j+\frac23(1-j)\right)\mathbb{I}\left\{p<\frac14,k>1\right\}\ ,$$

and $h\colon\widetilde{\boldsymbol{\mu}}_\varepsilon\otimes\boldsymbol{\mu}_L\mapsto\boldsymbol{\mu}_\varepsilon\otimes\boldsymbol{\mu}_L$, we have that the game $\mathscr{G}_5$ is easier than $\mathscr{G}_4$ (i.e., $R_T^\star(\mathscr{G}_4)\geqslant R_T^\star(\mathscr{G}_5)$).

Step 6. Now, define $\pi\colon\mathcal{Z}_5\to\{0,1\}$ as the projection on the first component $\{0,1\}$ of $\mathcal{Z}_5$, $\mathcal{Z}_6:=\{0,1\}$, $\varphi_6:=\pi\circ\varphi_5$, and the game $\mathscr{G}_6 := (\mathcal{X},\mathcal{Y}_5,\mathcal{Z}_6,\rho_5,\varphi_6,\mathcal{P}_5)$. Let $(\widetilde{Y}_t)_{t\in\mathbb{N}}$ be the adversary's actions in $\mathscr{G}_5$. A straightforward verification shows that for all $t\in\mathbb{N}$,

- For each $p\in[0,1/4)$ and any environment $\mathbb{P}$ of game $\mathscr{G}_5$, $\mathbb{P}_{\varphi_5(p,\widetilde{Y}_t)} = \mathbb{P}_{\pi\left(\varphi_5(p,\widetilde{Y}_t)\right)}\otimes(\nu\otimes\delta_p)$, where ν is the unique distribution on $\big(\mathbb{N}\cup\{\infty\}\big)\times\{0,1\}$ such that, for all $k\in\mathbb{N}\cup\{\infty\}$, $j\in\{0,1\}$, $\nu\big[\{(k,j)\}\big] = \frac12\mathbb{I}\{k=1,j=0\}+\frac{1}{2^{k+1}}\mathbb{I}\{1<k<\infty\}$. Using again the Skorokhod representation, there exists $\psi_p\colon[0,1]\to\big(\mathbb{N}\cup\{\infty\}\big)\times\{0,1\}\times[0,1]$ such that $\nu\otimes\delta_p=(\mu_L)_{\psi_p}$.
- For each $p\in[1/4,1]$ and any environment $\mathbb{P}$ of game $\mathscr{G}_5$, $\mathbb{P}_{\varphi_5(p,\widetilde{Y}_t)} = \delta_{(0,1,0,p)} = (\mu_L)_{\gamma_p}$, where $\gamma_p\colon[0,1]\to\mathcal{Z}_5$, $\lambda\mapsto(0,1,0,p)$.

Thus, by the Simulation lemma with $\mathcal{I}=[0,1/4)$ and $\mathcal{U}=[1/4,1]$, the game $\mathscr{G}_6$ is easier than $\mathscr{G}_5$ (i.e., $R_T^\star(\mathscr{G}_5)\geqslant R_T^\star(\mathscr{G}_6)$).

Step 7. Finally, consider the game $\mathscr{G}_7 := \big(\{1,2,3\},\{1,2\},\{0,1\},\rho_7,\varphi_7,\mathcal{P}_7\big)$, where in matrix notation, $\rho_7 = \big[\rho(i,j)\big]_{i\in\{1,2,3\},j\in\{1,2\}}$ and $\varphi_7 = \big[\varphi(i,j)\big]_{i\in\{1,2,3\},j\in\{1,2\}}$ are given by

$$\rho_7 := \frac{1}{96}\begin{bmatrix}34 & 34\\ 45 & 37\\ 38 & 44\end{bmatrix}\ ,\qquad \varphi_7 := \begin{bmatrix}1 & 0\\ 0 & 0\\ 0 & 0\end{bmatrix}\ ,$$

and $\mathcal{P}_7$ is the set of all measures of the form $\otimes_{t\in\mathbb{N}}\big(\frac{1+\varepsilon}{2}\delta_1+\frac{1-\varepsilon}{2}\delta_2\big)$, for $\varepsilon\in[-1,1]$. Thus, using again the Embedding lemma, this time with $\widetilde{f}\colon[0,1]\to\{1,2,3\}$, $p\mapsto\mathbb{I}\{p<1/4\}+2\mathbb{I}\{1/4\leqslant p\leqslant 1/3\}+3\mathbb{I}\{1/3<p\}$, $g\colon\{0,1\}\to\{0,1\}$, $i\mapsto i$, and $h\colon\otimes_{t\in\mathbb{N}}\big(\frac{1+\varepsilon}{2}\delta_1+\frac{1-\varepsilon}{2}\delta_2\big)\otimes\boldsymbol{\mu}_L\mapsto\tilde{\boldsymbol{\mu}}_\varepsilon\otimes\boldsymbol{\mu}_L$, we obtain that $\mathscr{G}_7$ is easier than $\mathscr{G}_6$ (i.e., $R_T^\star(\mathscr{G}_6)\geqslant R_T^\star(\mathscr{G}_7)$). This last game is an instance of the so-called revealing action partial monitoring game, whose minimax regret is known to be lower bounded by $\frac{11}{96}\big(\frac{1}{7}T^{2/3}\big)$ [49]. In conclusion, we proved that $R_T^\star(\mathscr{G})\geqslant R_T^\star(\mathscr{G}_7)\geqslant\frac{11}{672}T^{2/3}$. □

A.9 Linear Lower Bound under Realistic Feedback (iid+bd)

In this section, we prove that in the realistic-feedback case, no strategy can achieve sublinear worst-case regret in the independent and identically distributed case when the valuations of the buyer and the seller may be dependent, not even if they have a bounded density.

The idea of the proof is to exploit the lack of observability in this setting, building a family of environments $\mathbb{P}^\lambda$ (parameterized by $\lambda\in[0,1]$) as convex combinations of the two measures in Figure 2.3. If $\lambda<1/2$, the optimal action is $3/8$, while if $\lambda>1/2$, the optimal action becomes $5/8$. This family is built is such a way that the feedback gives no information on λ, making it impossible to distinguish between the two cases. Leveraging the Embedding and Simulation lemmas (Lemmas 19 and 20), this construction leads to a reduction to an instance of a non-observable partial monitoring game, whose regret is trivially lower bounded by $T/24$.

Theorem 38 (Theorem 7, restated)**.** *With the same notation as in Appendix A.5.2, in the realistic-feedback stochastic (iid) setting with joint density bounded by a constant $M\geqslant 64/3$ (bd), for all horizons $T\in\mathbb{N}$, the minimax regret satisfies*

$$R_T^\star\geqslant\frac{1}{24}T\,.$$

Proof. Fix any horizon $T\in\mathbb{N}$ and $M\geqslant 64/3$. Recalling Appendix A.5.2, the realistic-feedback stochastic (iid) setting with joint density bounded by M (bd) is a game $\mathscr{G}:=(\mathcal{X},\mathcal{Y},\mathcal{Z},\rho,\varphi,\mathcal{P})$, where $\mathcal{X}=[0,1]$, $\mathcal{Y}=[0,1]^2$, $\mathcal{Z}=\{0,1\}^2$, $\rho=\mathrm{gft}$, $\varphi\colon\big(p,(s,b)\big)\mapsto\big(\mathbb{I}\{s\leqslant p\},\mathbb{I}\{p\leqslant b\}\big)$, and $\mathcal{P}=\mathcal{P}^M_{\mathrm{iid+bd}}$. Define the two joint densities $f=\frac{64}{3}\big(\mathbb{I}_{[0/8,1/8]\times[3/8,4/8]}+\mathbb{I}_{[2/8,3/8]\times[7/8,8/8]}+\mathbb{I}_{[4/8,5/8]\times[5/8,6/8]}\big)$ and $g\colon[0,1]^2\to[0,M]$, $(s,b)\mapsto f(1-b,1-s)$ (see Figure 2.3, left). Let $\mathcal{P}_1$ be the subset of $\mathcal{P}^M_{\mathrm{iid+bd}}$ whose elements have the form $\boldsymbol{\mu}_\lambda:=\otimes_{t\in\mathbb{N}}\big(\big((1-\lambda)f+\lambda g\big)(\mu_L\otimes\mu_L)\big)$ for $\lambda\in[0,1]$. Since $\mathcal{P}_1\subset\mathcal{P}$ the game $\mathscr{G}_1:=(\mathcal{X},\mathcal{Y},\mathcal{Z},\rho,\varphi,\mathcal{P}_1)$ is easier than $\mathscr{G}$ (i.e., $R_T^\star(\mathscr{G})\geqslant R_T^\star(\mathscr{G}_1)$) by the Embedding lemma (Lemma 19) with $\widetilde{f}$ and g as the identities, and h as the inclusion. Define $\mathcal{Z}_1:=\{0\}$ and $\varphi_1\colon\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}_1$, $\big(p,(s,b)\big)\mapsto 0$. Let $(Y_t)_{t\in\mathbb{N}}$ be the adversary's actions in $\mathcal{G}_1$. Now, since for all $t\in\mathbb{N}$, any two environments $\mathbb{P}$ and $\mathbb{Q}$ of game $\mathscr{G}_1$, and each $p\in[0,1]$, $\mathbb{P}_{\varphi(p,Y_t)}=\mathbb{Q}_{\varphi(p,Y_t)}$, then by the well-known Skorokhod representation [185, Section 17.3], for each $t\in\mathbb{N}$ and each $p\in[0,1]$ there exists $\gamma_{t,p}\colon[0,1]\to\{0,1\}^2$ such that for any environment $\mathbb{P}$ of game $\mathscr{G}_1$, $\mathbb{P}_{\varphi(x,Y_t)}=(\mu_L)_{\gamma_{t,x}}$. Thus, the Simulation lemma (Lemma 20) with $\mathcal{I}=\varnothing$ and $\mathcal{U}=\mathcal{X}$ implies that the game $\mathscr{G}_2:=(\mathcal{X},\mathcal{Y},\mathcal{Z}_2,\rho,\varphi_2,\mathcal{P}_1)$ is easier than $\mathscr{G}_1$ (i.e., $R_T^\star(\mathscr{G}_1)\geqslant R_T^\star(\mathscr{G}_2)$). Define $\rho_3\colon\mathcal{X}\times\mathcal{Y}\to[0,1]$, $\big(p,(s,b)\big)\mapsto(b-s)\mathbb{I}\big\{s\leqslant\frac{3}{8}\leqslant b\big\}\mathbb{I}\big\{p\leqslant\frac{1}{2}\big\}+(b-s)\mathbb{I}\big\{s\leqslant\frac{5}{8}\leqslant b\big\}\,\mathbb{I}\big\{p>\frac{1}{2}\big\}$ and $\mathscr{G}_3:=(\mathcal{X},\mathcal{Y},\mathcal{Z}_2,\rho_3,\varphi_2,\mathcal{P}_1)$. By the Embedding lemma with $\widetilde{f},g,h$ as the identities, we have that the game $\mathscr{G}_3$ is easier than the game $\mathscr{G}_2$ (i.e., $R_T^\star(\mathscr{G}_2)\geqslant R_T^\star(\mathscr{G}_3)$). Finally, consider the game $\mathscr{G}_4:=\big(\{1,2\},\{1,2\},\{0\},\rho_4,\varphi_4,\mathcal{P}_4\big)$, where in matrix notation, $\rho_4=\big[\rho(i,j)\big]_{i,j\in\{1,2\}}$ and

$\varphi_4 = \big[\varphi(i,j)\big]_{i,j\in\{1,2\}}$ are given by

$$\rho_4 := \begin{bmatrix} 1/3 & 1/4 \\ 1/4 & 1/3 \end{bmatrix}, \qquad \varphi_4 := \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and $\mathcal{P}_4$ is the set of all measures of the form $(1-\lambda)\delta_1 + \lambda\delta_2$, for $\lambda \in [0,1]$. Using again the Embedding lemma, this time with $\widetilde{f}\colon [0,1] \to \{1,2\}$, $p \mapsto \mathbb{I}\{p \leqslant 1/2\} + 2\mathbb{I}\{1/2 < p\}$, $g\colon \{0\} \to \{0\}$, $i \mapsto i$, and $h\colon \otimes_{t\in\mathbb{N}} \big((1-\lambda)\delta_1 + \lambda\delta_2\big) \otimes \boldsymbol{\mu}_L \mapsto \boldsymbol{\mu}_\lambda \otimes \boldsymbol{\mu}_L$, we obtain that $\mathscr{G}_4$ is easier than $\mathscr{G}_3$ (i.e., $R_T^\star(\mathscr{G}_3) \geqslant R_T^\star(\mathscr{G}_4)$). This last game has (trivially) minimax regret at most $\big(\frac{1}{3} - \frac{1}{4}\big)\frac{T}{2}$. In conclusion, we proved that $R_T^\star(\mathscr{G}) \geqslant R_T^\star(\mathscr{G}_4) \geqslant \frac{1}{24}T$. □

A.10 Linear Lower Bound under Realistic Feedback (iid+iv)

In this section, we prove that in the realistic-feedback case, no strategy can achieve sublinear regret without any limitations on how concentrated the distributions of the valuations of the seller and buyer are, not even if they are independent of each other (iv) and the process of valuations is independent and identically distributed (iid).

The idea of the proof is that if the two distributions are very concentrated in a small region, finding an optimal price is like finding a needle in a haystack. Each strategy that (at each time step) receives as feedback only a finite number of bits, as in our realistic setting, can assign positive probability to at most a countable set of points. Thus one could find concentrated distributions of the buyer and seller that have a unique optimal point in which the strategy has zero probability of posting prices at all time steps, and such that *all* other prices suffer large regret.

Theorem 39 (Theorem 8, restated)**.** *With the same notation as in Appendix A.5.2, in the realistic-feedback stochastic (iid) setting with independent valuations (iv), for all horizons $T \in \mathbb{N}$, the minimax regret satisfies*

$$R_T^\star \geqslant \frac{1}{8}T\,.$$

Proof. To lighten the notation, for any $n \in \mathbb{N}$ and a family $(\lambda_k)_{k\in\mathbb{N}}$, we let $\lambda_{1:n} := (\lambda_1, \dots, \lambda_n)$. Fix an arbitrary horizon $T \in \mathbb{N}$. Recalling Appendix A.5.2, the realistic-feedback stochastic (iid) setting with independent valuations (iv) is a game $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$, where $\mathcal{X} = [0,1]$, $\mathcal{Y} = [0,1]^2$, $\mathcal{Z} = \{0,1\}^2$, $\rho = \mathrm{gft}$, $\varphi\colon \big(p, (s,b)\big) \mapsto \big(\mathbb{I}\{s \leqslant p\}, \mathbb{I}\{p \leqslant b\}\big)$, and $\mathcal{P} = \mathcal{P}_{\text{iid+iv}}$. Let $\mathcal{S}$ be the set of environments of $\mathscr{G}$. Fix a strategy α for game $\mathscr{G}$ and let $\varepsilon \in (0,1)$. Define $\bar{\alpha}_1 := \alpha_1$, $\nu_1 := (\mu_L)_{\bar{\alpha}_1}$, and for each $t \in \mathbb{N}$ and $z_1, \dots, z_t \in \{0,1\}^2$,

$$\bar{\alpha}_{t+1,z_{1:t}}\colon [0,1]^{t+1} \to [0,1], \quad u_{1:t+1} \mapsto \alpha_{t+1}(u_{1:t+1}, z_{1:t}) \qquad \text{and} \qquad \nu_{t+1,z_{1:t}} := (\otimes_{s=1}^{t+1}\mu_L)_{\bar{\alpha}_{t+1,z_{1:t}}}\,.$$

Define also the set $A_1 := \big\{x \in [0,1] \mid \nu_1[\{x\}] > 0\big\}$ and, for each $t \in \mathbb{N}$, the union $A_{t+1} := \bigcup_{z_{1:t}\in\{0,1\}^2}\big\{x \in [0,1] \mid \nu_{t,z_{1:t}}[\{x\}] > 0\big\}$. Note that, for each $t \in \mathbb{N}$, A_t is countable, being the union of 4^{t-1} countable sets. Then $A := \bigcup_{t\in\mathbb{N}} A_t$ is countable. Since $B := [\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}]$ has the power of continuum, we have that the same holds for $B\backslash A$. In particular, $B\backslash A$ is non-empty. Pick $x^\star \in B\backslash A$ and define $\mu_S := \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{x^\star}$, $\mu_B := \frac{1}{2}\delta_{x^\star} + \frac{1}{2}\delta_1$, and $\mathbb{P} := (\otimes_{t\in\mathbb{N}}(\mu_S \otimes \mu_B)) \otimes \boldsymbol{\mu}_L \in \mathcal{S}$. Then for each $t \in \mathbb{N}$, we have that

$$\mathbb{E}_{\mathbb{P}}\big[\rho(x^\star, Y_t)\big] = \frac{x^\star + (1-x^\star) + 1}{4}\,.$$

On the other hand, $\mathbb{P}[X_1 = x^\star] = \nu_1[\{x^\star\}] = 0$ and for each $t \in \mathbb{N}$, we have that

$$\begin{aligned}\mathbb{P}[X_{t+1} = x^\star] &= \mathbb{P}\left[\alpha_{t+1}(U_1, \dots, U_{t+1}, Z_1, \dots, Z_t) = x^\star\right] \\ &= \sum_{z_1,\dots,z_t \in \{0,1\}^2} \mathbb{P}\left[\alpha_{t+1}(U_1, \dots, U_{t+1}, z_1, \dots, z_t) = x^\star \cap Z_1 = z_1 \cap \dots \cap Z_t = z_t\right] \\ &\leqslant \sum_{z_1,\dots,z_t \in \{0,1\}^2} \mathbb{P}\left[\alpha_{t+1}(U_1, \dots, U_{t+1}, z_1, \dots, z_t) = x^\star\right] = \sum_{z_1,\dots,z_t \in \{0,1\}^2} \nu_{t+1,z_1,\dots,z_t}\left[\{x^\star\}\right] = 0\,,\end{aligned}$$

which in turn gives

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\big[\rho(X_t, Y_t)\big] &= \frac{\mathbb{E}_{\mathbb{P}}\left[\rho\big(X_t, (0, x^\star)\big)\right] + \mathbb{E}_{\mathbb{P}}\left[\rho\big(X_t, (x^\star, 1)\big)\right] + \mathbb{E}_{\mathbb{P}}\left[\rho\big(X_t, (0, 1)\big)\right] + \mathbb{E}_{\mathbb{P}}\left[\rho\big(X_t, (x^\star, x^\star)\big)\right]}{4} \\ &= \frac{x^\star \mathbb{P}_{X_t}\big[[0, x^\star]\big] + (1 - x^\star)\mathbb{P}_{X_t}\big[[x^\star, 1]\big] + 1}{4} = \frac{x^\star \mathbb{P}_{X_t}\big[[0, x^\star)\big] + (1 - x^\star)\mathbb{P}_{X_t}\big[(x^\star, 1]\big] + 1}{4} \\ &\leqslant \frac{\max(x^\star, 1 - x^\star) + 1}{4} = \frac{x^\star + (1 - x^\star) + 1 - \min(x^\star, 1 - x^\star)}{4}\,.\end{aligned}$$

So, if $T \in \mathbb{N}$ we get

$$R_T(\alpha, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \rho(x^\star, Y_t) - \sum_{t=1}^{T} \rho(X_t, Y_t)\right] \geqslant \frac{\min(x^\star, 1 - x^\star)}{4} T \geqslant \frac{1 - \varepsilon}{8} T.$$

Since ε was arbitrary, we get, for all $T \in \mathbb{N}$, $R_T^{\mathcal{S}}(\alpha) = \sup_{\mathbb{P} \in \mathcal{S}} R_T(\alpha, \mathbb{P}) \geqslant \sup_{\varepsilon \in (0,1)} \frac{1-\varepsilon}{8} T = T/8$. Since α was arbitrary we get, for each $T \in \mathbb{N}$, $R_T^\star = \inf_{\alpha \in \mathscr{A}} R_T^{\mathcal{S}}(\alpha) \geqslant T/8$. □

A.11 Adversarial Setting: Linear Lower Bound under Full Feedback

In this section, we give a more detailed proof of Theorem 1 with a notation consistent with our abstract setting of sequential games.

Theorem 40 (Theorem 1, restated)**.** *With the same notation as in Appendix A.5.2, in the full-feedback adversarial (adv) setting, for all horizons $T \in \mathbb{N}$, we have*

$$R_T^\star \geqslant \frac{1}{4} T\,.$$

Proof. Recalling Appendix A.5.2, the full-feedback adversarial (adv) bilateral trade setting is a game $\mathscr{G} := (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \rho, \varphi, \mathcal{P})$, where $\mathcal{X} = [0, 1]$, $\mathcal{Y} = [0, 1]^2$, $\mathcal{Z} = [0, 1]^2$, $\rho = \mathrm{gft}$, $\varphi\colon \big(p, (s, b)\big) \mapsto (s, b)$, and $\mathcal{P} = \mathcal{P}_{\mathrm{adv}}$. Let $\mathcal{S}$ be the set of environments of $\mathscr{G}$. Fix a strategy $\alpha \in \mathscr{A}$ and an $\varepsilon \in (0, 1/18)$. Define $\bar{\alpha}_1 := \alpha_1$, $\nu_1 := (\mu_L)_{\bar{\alpha}_1}$, and

$$\begin{cases} c_1 := \frac{1}{2} - \frac{3}{2}\varepsilon,\ d_1 := \frac{1}{2} - \frac{1}{2}\varepsilon,\ s_1 := 0,\ b_1 := d_1, & \text{if } \nu_1\big[\big[0, \frac{1}{2} - \frac{1}{2}\varepsilon\big]\big] \leqslant \frac{1}{2}\,, \\ c_1 := \frac{1}{2} + \frac{1}{2}\varepsilon,\ d_1 := \frac{1}{2} + \frac{3}{2}\varepsilon,\ s_1 := c_1,\ b_1 := 1, & \text{otherwise.} \end{cases}$$

If $t \in \mathbb{N}$, suppose we defined $\bar{\alpha}_t, \nu_t, c_t, d_t, s_t, b_t$ and let

$$\bar{\alpha}_{t+1} : [0, 1]^{t+1} \to [0, 1], (u_1, \dots, u_{t+1}) \mapsto \alpha_{t+1}\left(u_1, \dots, u_{t+1}, (s_1, b_1), \dots, (s_t, b_t)\right),$$

$\nu_{t+1} := \big(\otimes_{s=1}^{t+1}\mu_L\big)_{\bar{\alpha}_{t+1}}$, and

$$\begin{cases} c_{t+1} := c_t,\ d_{t+1} := d_t - \frac{2\varepsilon}{3^t},\ s_{t+1} := 0,\ b_{t+1} := d_{t+1}, & \text{if } \nu_{t+1}\big[\big[0, c_t + \frac{\varepsilon}{3^t}\big]\big] \leqslant \frac{1}{2}\ , \\ c_{t+1} := c_t + \frac{2\varepsilon}{3^t},\ d_{t+1} := d_t,\ s_{t+1} := c_{t+1},\ b_{t+1} := 1, & \text{otherwise.} \end{cases}$$

Then $(\bar{\alpha}_t)_{t\in\mathbb{N}}, (\nu_t)_{t\in\mathbb{N}}, (c_t)_{t\in\mathbb{N}}, (d_t)_{t\in\mathbb{N}}, (s_t)_{t\in\mathbb{N}}, (b_t)_{t\in\mathbb{N}}$ are well-defined by induction and satisfy:

- For each $t \in \mathbb{N}$, $d_t - c_t = \frac{\varepsilon}{3^{t-1}}$.
- For each $t \in \mathbb{N}$, $c_1 \leqslant c_2 \leqslant c_3 \leqslant \cdots \leqslant c_t \leqslant d_t \leqslant \cdots \leqslant d_3 \leqslant d_2 \leqslant d_1$.
- $\exists! x^\star \in \bigcap_{t=1}^{\infty}[c_t, d_t]$.
- For each $t \in \mathbb{N}$, $\rho\left(x^\star, (s_t, b_t)\right) = b_t - s_t \geqslant \frac{1-3\varepsilon}{2}$.
- For each $t \in \mathbb{N}$, $\mathbb{P}\big[\alpha_t\big(U_1, \ldots, U_t, (s_1, b_1), \ldots, (s_{t-1}, b_{t-1})\big) \in [s_t, b_t]\big] \leqslant \frac{1}{2}$.

Now, define $\mathbb{P} := \big(\otimes_{t\in\mathbb{N}}\delta_{(s_t,b_t)}\big) \otimes \boldsymbol{\mu}_L \in \mathcal{S}$. Then, for each $t \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[\rho\left(X_t, Y_t\right)] &= \mathbb{E}_{\mathbb{P}}\Big[\rho\Big(\alpha_t\big(U_1, \ldots, U_t, (s_1, b_1), \ldots, (s_{t-1}, b_{t-1})\big), (s_t, b_t)\Big)\Big] \\ &\leqslant \left(\frac{1}{2} + \frac{3\varepsilon}{2}\right)\mathbb{P}\big[\alpha_t\big(U_1, \ldots, U_t, (s_1, b_1), \ldots, (s_{t-1}, b_{t-1})\big) \in [s_t, b_t]\big] \leqslant \frac{1}{4} + \frac{3\varepsilon}{4}\ , \end{aligned}$$

and so, for each $T \in \mathbb{N}$

$$\begin{aligned} R_T(\alpha, \mathbb{P}) &= \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T}\rho(x^\star, Y_t) - \sum_{t=1}^{T}\rho(X_t, Y_t)\right] = \sum_{t=1}^{T}\rho(x^\star, (s_t, b_t)) - \sum_{t=1}^{T}\mathbb{E}_{\mathbb{P}}\left[\rho\left(X_t, Y_t\right)\right] \\ &\geqslant \sum_{t=1}^{T}(b_t - s_t)\big(1 - \mathbb{P}\big[\alpha_t\big(U_1, \ldots, U_t, (s_1, b_1), \ldots, (s_{t-1}, b_{t-1})\big) \in [s_t, b_t]\big]\big) \geqslant \frac{1-3\varepsilon}{4}T\ . \end{aligned}$$

Since ε was arbitrary, we get, for all $T \in \mathbb{N}$, $R_T^{\mathcal{S}}(\alpha) = \sup_{\mathbb{P}\in\mathcal{S}} R_T(\alpha, \mathbb{P}) \geqslant \sup_{\varepsilon\in(0,1/18)} \frac{1-3\varepsilon}{4}T = \frac{T}{4}$. Since α arbitrarity, we get, for each $T \in \mathbb{N}$, $R_T^\star = \inf_{\alpha\in\mathscr{A}} R_T^{\mathcal{S}}(\alpha) \geqslant \frac{T}{4}$. □

A.12 DKW Inequalities

We begin this section by presenting the univariate DKW inequality as proved by Massart [135].

Theorem 41. *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X, X_1, X_2, \ldots$ is a $\mathbb{P}$-i.i.d. sequence of random variables, then, for any $\varepsilon > 0$ and all $m \in \mathbb{N}$, it holds*

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}}\left|\frac{1}{m}\sum_{k=1}^{m}\mathbb{I}\{X_k \leqslant x\} - \mathbb{P}[X \leqslant x]\right| > \varepsilon\right] \leqslant 2\exp\big(-2m\varepsilon^2\big)\ .$$

We now present a bivariate DKW inequality which can be proved by applying the VC-type bound of [13, Theorem 4.9; see also Lemmas 4.4, 4.5, and 4.11 for the explicit constants].

Theorem 42. *There exist positive constants $m_0 \leqslant 1200$, $c_1 \leqslant 13448$, $c_2 \geqslant 1/576$ such that, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ 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* $m \geqslant m_0/\varepsilon^2$*, it holds*

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

A.13 Proof of the Representation Lemma

In this section, we fill in the missing details we left unproven in Section 2.5.2. Specifically, we prove now that, if S, B are two independent random variables supported in $[0,1]$ that share the same distribution μ and (hence have) common expectation $\bar{\mu}$, then, for each $t \in \mathbb{N}$ and each $p \in [0,1]$ it holds that:

$$2\mathbb{E}\Big[\mathrm{gft}\big(p,(S,B)\big)\Big] = \widetilde{\rho}\,(\mu)\,(p) + \mu\,\{p\}\left(\int_0^p \mu\,[0,\lambda]\;\mathrm{d}\lambda + \int_p^1 \mu\,[\lambda,1]\;\mathrm{d}\lambda\right)$$

where

$$\widetilde{\rho}\,(\mu)\,(p) = \int_0^p \left(\mu\,[0,\lambda] + \mu\,[0,\lambda)\right)\,\mathrm{d}\lambda + \left(\mu\,[0,p] + \mu\,[0,p)\right)(\bar{\mu} - p)\,.$$

For notational convenience, let V be another random variable with distribution μ.

In what follows, we will use the following observation. For any $0 \leqslant a < b \leqslant 1$ we have

- $\int_a^b \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda = \int_a^b \mathbb{P}\,[V > \lambda]\;\mathrm{d}\lambda$,
- $\int_a^b \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda = \int_a^b \mathbb{P}\,[V < \lambda]\;\mathrm{d}\lambda$.

This is due to the fact that the two functions $\lambda \mapsto \mathbb{P}\,[V \geqslant \lambda]$ and $\lambda \mapsto \mathbb{P}\,[V > \lambda]$ are different only in a set that is at most countable. Hence, the set where they differ have measure zero and the first two integral coincides. The same reasoning applies to the second two integrals.

Now, notice that, by the decomposition lemma Lemma 2, for all $p \in [0,1]$,

$$\begin{aligned}
2\mathbb{E}\Big[\mathrm{gft}\big(p,(S,B)\big)\Big] &= 2\mathbb{P}\,[V \leqslant p]\int_p^1 \mathbb{P}\,[\lambda \leqslant V]\;\mathrm{d}\lambda + 2\mathbb{P}\,[V \geqslant p]\int_0^p \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda \\
&= \left(\mathbb{P}\,[V \leqslant p] + \mathbb{P}\,[V < p]\right)\int_p^1 \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda + \left(\mathbb{P}\,[V \geqslant p] + \mathbb{P}\,[V > p]\right)\int_0^p \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\,[V \leqslant p] - \mathbb{P}\,[V < p]\right)\int_p^1 \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda + \left(\mathbb{P}\,[V \geqslant p] - \mathbb{P}\,[V > p]\right)\int_0^p \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda \\
&= \left(\mathbb{P}\,[V \leqslant p] + \mathbb{P}\,[V < p]\right)\mathbb{E}\,[V] - \left(\mathbb{P}\,[V \leqslant p] + \mathbb{P}\,[V < p]\right)\int_0^p \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda \\
&\quad + \left(1 - \mathbb{P}\,[V < p] + 1 - \mathbb{P}\,[V \leqslant p]\right)\int_0^p \left(1 - \mathbb{P}\,[V > \lambda]\right)\mathrm{d}\lambda + \left(\mathbb{P}\,[V \leqslant p] - \mathbb{P}\,[V < p]\right)\int_p^1 \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\,[V \geqslant p] - \mathbb{P}\,[V > p]\right)\int_0^p \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda \\
&= \left(\mathbb{P}\,[V \leqslant p] + \mathbb{P}\,[V < p]\right)\left(\mathbb{E}\,[V] - p\right) - \left(\mathbb{P}\,[V \leqslant p] + \mathbb{P}\,[V < p]\right)\int_0^p \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda \\
&\quad + 2p - \left(1 - \mathbb{P}\,[V < p] + 1 - \mathbb{P}\,[V \leqslant p]\right)\int_0^p \mathbb{P}\,[V > \lambda]\;\mathrm{d}\lambda + \left(\mathbb{P}\,[V \leqslant p] - \mathbb{P}\,[V < p]\right)\int_p^1 \mathbb{P}\,[V \geqslant \lambda]\;\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\,[V \geqslant p] - \mathbb{P}\,[V > p]\right)\int_0^p \mathbb{P}\,[V \leqslant \lambda]\;\mathrm{d}\lambda
\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\left(\mathbb{E}\left[V\right] - p\right) - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\int_0^p \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad + 2p - \left(1 - \mathbb{P}\left[V < p\right] + 1 - \mathbb{P}\left[V \leqslant p\right]\right)\int_0^p \left(1 - \mathbb{P}\left[V \leqslant \lambda\right]\right)\,\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= 2\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\left(\mathbb{E}\left[V\right] - p\right) - \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\int_0^p \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)p \\
&\quad + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= \int_0^p \left(\mu\left[0,\lambda\right] + \mu\left[0,\lambda\right)\right)\,\mathrm{d}\lambda + \left(\mu\left[0,p\right] + \mu\left[0,p\right)\right)\left(\bar{\mu} - p\right) - \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\int_0^p \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)p \\
&\quad + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&=: \widetilde{\rho}\left(\mu\right)\left(p\right) + \left(\mathrm{I}\right).
\end{aligned}
$$

It is left to prove that $\left(\mathrm{I}\right) = \mu\left\{p\right\}\left(\int_0^p \mu\left[0,\lambda\right]\,\mathrm{d}\lambda + \int_p^1 \mu\left[\lambda,1\right]\,\mathrm{d}\lambda\right)$. In fact

$$
\begin{aligned}
\left(\mathrm{I}\right) &= -\left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\int_0^p \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)p + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= -\left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)\int_0^p \left(1 - \mathbb{P}\left[V \leqslant \lambda\right]\right)\,\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)p + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= \left(\left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right) - \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&\quad - \left(\mathbb{P}\left[V \leqslant p\right] + \mathbb{P}\left[V < p\right]\right)p + \left(\mathbb{P}\left[V < p\right] + \mathbb{P}\left[V \leqslant p\right]\right)p \\
&\quad + \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= \left(\mathbb{P}\left[V \leqslant p\right] - \mathbb{P}\left[V < p\right]\right)\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \left(\mathbb{P}\left[V \geqslant p\right] - \mathbb{P}\left[V > p\right]\right)\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= \mathbb{P}\left[V = p\right]\int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda + \mathbb{P}\left[V = p\right]\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda \\
&= \mathbb{P}\left[V = p\right]\left(\int_0^p \mathbb{P}\left[V \leqslant \lambda\right]\,\mathrm{d}\lambda + \int_p^1 \mathbb{P}\left[V \geqslant \lambda\right]\,\mathrm{d}\lambda\right)
\end{aligned}
$$

$$= \mu\{p\}\left(\int_0^p \mu[0,\lambda]\,\mathrm{d}\lambda + \int_p^1 \mu[\lambda,1]\,\mathrm{d}\lambda\right)$$

which concludes the proof of the claim.

A.14 Missing Details in the Proof of Theorem 12

We will show now that, with the notation of the proof of the Theorem 12, for any $M \geqslant 2$, if $t \geqslant 580M^4$, it holds that

$$\mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid D_{Z,1},\ldots,D_{Z,2(t-1)}\right]\right)^2\right] \geqslant \frac{1}{147}\cdot\frac{1}{t-1}.$$

For any $t \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{E}\left[(Z - \mathbb{E}[Z \mid D_{Z,1},\ldots,D_{Z,t}])^2\right] &\geqslant \mathbb{E}\left[(Z - \mathbb{E}[Z \mid D_{Z,1},\ldots,D_{Z,t}])^2\,\mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\}\right] \\
&= \mathbb{E}\left[\left(\underbrace{\left(Z - \sum_{k=1}^t D_{Z,k}\right)}_{a} + \underbrace{\left(\sum_{k=1}^t D_{Z,k} - \mathbb{E}[Z \mid D_{Z,1},\ldots,D_{Z,t}]\right)}_{b}\right)^2 \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\}\right] \\
&\geqslant \mathbb{E}\left[\left(Z - \frac{1}{t}\sum_{k=1}^t D_{Z,k}\right)^2 \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\}\right] \\
&\quad - 2\mathbb{E}\left[\left|Z - \frac{1}{t}\sum_{k=1}^t D_{Z,k}\right|\left|\frac{1}{t}\sum_{k=1}^t D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\ldots,D_{Z,t}\right]\right| \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\}\right] \\
&=: (\mathrm{I}) - 2\cdot(\mathrm{II}),
\end{aligned}$$

where the last inequality follows from $(a+b)^2 \geqslant a^2 - 2|ab|$. Now, if W is a uniform random variable on $\left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]$ independent of $(D_{q,t})_{q\in[0,1],t\in\mathbb{N}}$, we have that

$$\begin{aligned}
(\mathrm{I}) &= \mathbb{E}\left[\left(Z - \frac{1}{t}\sum_{k=1}^t D_{Z,k}\right)^2 \mid Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right]\mathbb{P}\left[Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right] \\
&= \frac{1}{9}\mathbb{E}\left[\left(Z - \frac{1}{t}\sum_{k=1}^t D_{Z,k}\right)^2 \mid Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9},\frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right] = \frac{1}{9}\mathbb{E}\left[\left(W - \frac{1}{t}\sum_{k=1}^t D_{W,k}\right)^2\right] =: (\star).
\end{aligned}$$

It follows that

$$\begin{aligned}
(\star) &= \frac{1}{9}\int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \mathbb{E}\left[\left(w - \frac{1}{t}\sum_{k=1}^t D_{w,k}\right)^2\right]\mathrm{d}\mathbb{P}_W(w) = \frac{1}{9}\int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \mathrm{Var}\left[\frac{1}{t}\sum_{k=1}^t D_{w,k}\right]\mathrm{d}\mathbb{P}_W(w) \\
&= \frac{1}{9}\int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \frac{w(1-w)}{t}\,\mathrm{d}\mathbb{P}_W(w) \leqslant \frac{1}{9}\frac{3}{7}\frac{4}{7}\frac{1}{t} = \frac{4}{147}\cdot\frac{1}{t}.
\end{aligned}$$

About the term (II), we have

$$\begin{aligned}(\mathrm{II}) &\leqslant \mathbb{P}\left[Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9}, \frac{1}{2}+\frac{\varepsilon_M}{9}\right] \cap \frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \notin \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right] \\ &\quad + \mathbb{E}\left[\left|Z-\frac{1}{t}\sum_{k=1}^{t} D_{Z,k}\right| \left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right| \cdot \right. \\ &\qquad \left. \cdot\, \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9}, \frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\} \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right] =: (\mathrm{III}) + (\mathrm{IV})\end{aligned}$$

About the term (III), we have

$$\begin{aligned}(\mathrm{III}) &= \int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \mathbb{P}\left[\frac{1}{t}\sum_{k=1}^{t} D_{z,k} \notin \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right] \mathrm{d}\mathbb{P}_Z(z) \\ &= \int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \mathbb{P}\left[\left\{\frac{1}{t}\sum_{k=1}^{t} D_{z,k} - z < \frac{1}{2}-\frac{\varepsilon_M}{6}-z\right\} \cup \left\{\frac{1}{t}\sum_{k=1}^{t} D_{z,k} - z > \frac{1}{2}+\frac{\varepsilon_M}{6}-z\right\}\right] \mathrm{d}\mathbb{P}_Z(z) \\ &\leqslant \int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \left(\exp\left(-2\left(\frac{1}{2}-\frac{\varepsilon_M}{6}-z\right)^2 t\right) + \exp\left(-2\left(\frac{1}{2}+\frac{\varepsilon_M}{6}-z\right)^2 t\right)\right) \mathrm{d}\mathbb{P}_Z(z) \\ &\leqslant \int_{\frac{1}{2}-\frac{\varepsilon_M}{9}}^{\frac{1}{2}+\frac{\varepsilon_M}{9}} \left(\exp\left(-2\left(\frac{1}{2}-\frac{\varepsilon_M}{6}-\frac{1}{2}+\frac{\varepsilon_M}{9}\right)^2 t\right) + \exp\left(-2\left(\frac{1}{2}+\frac{\varepsilon_M}{6}-\frac{1}{2}-\frac{\varepsilon_M}{9}\right)^2 t\right)\right) \mathrm{d}\mathbb{P}_Z(z) \\ &= \frac{2}{9}\exp\left(-2\left(\frac{\varepsilon_M}{18}\right)^2 t\right) = \frac{2}{9}\exp\left(-\frac{\varepsilon_M^2}{162}t\right) = \frac{2}{9}\exp\left(-\frac{\left(\frac{7}{M}\right)^2}{162}t\right) = \frac{2}{9}\exp\left(-\frac{49}{162}\cdot\frac{t}{M^2}\right),\end{aligned}$$

where the first inequality follows from Hoeffding's inequality. About the term (IV), we have

$$\begin{aligned}(\mathrm{IV}) &\leqslant \sqrt{\mathbb{E}\left[\left|Z-\frac{1}{t}\sum_{k=1}^{t} D_{Z,k}\right|^2 \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9}, \frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\} \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]} \cdot \\ &\quad \cdot \sqrt{\mathbb{E}\left[\left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right|^2 \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9}, \frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\} \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]} \\ &\leqslant \sqrt{\mathbb{E}\left[\left|Z-\frac{1}{t}\sum_{k=1}^{t} D_{Z,k}\right|^2 \mathbb{I}\left\{Z \in \left[\frac{1}{2}-\frac{\varepsilon_M}{9}, \frac{1}{2}+\frac{\varepsilon_M}{9}\right]\right\}\right]} \cdot \\ &\quad \cdot \sqrt{\mathbb{E}\left[\left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right|^2 \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]} \\ &= \sqrt{\frac{4}{147}\cdot\frac{1}{t}} \cdot \sqrt{\mathbb{E}\left[\left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right|^2 \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]} =: (\circ),\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz and the last inequality follows from $(\star)$. Now, using that $(a-b)^2 \leqslant 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get:

$$\begin{aligned}&\left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right|^2 \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} \in \left[\frac{1}{2}-\frac{\varepsilon_M}{6}, \frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\} \\ &\qquad \leqslant 2\left|\frac{1}{t}\sum_{k=1}^{t} D_{Z,k} - \frac{1+\sum_{k=1}^{t} D_{Z,k}}{t+2}\right|^2 +\end{aligned}$$

$$2\left|\frac{1+\sum_{k=1}^{t} D_{Z,k}}{t+2}-\mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]\right|^2 \mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t} D_{Z,k}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}$$
$$=: (\mathrm{V})+(\mathrm{VI})\,.$$

Simple calculations show that

$$(\mathrm{V})\leqslant\frac{18}{t^2}.$$

About (VI), we first compute $\mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]$ using Bayes' formula and get

$$\mathbb{E}\left[Z \mid D_{Z,1},\dots,D_{Z,t}\right]=\frac{\int_{\left[\frac{1}{2}-\varepsilon_M,\frac{1}{2}+\varepsilon_M\right]} p^{1+\sum_{k=1}^{t} D_{Z,k}}(1-p)^{t-\sum_{k=1}^{t} D_{Z,k}}\,\mathrm{d}p}{\int_{\left[\frac{1}{2}-\varepsilon_M,\frac{1}{2}+\varepsilon_M\right]} p^{\sum_{k=1}^{t} D_{Z,k}}(1-p)^{t-\sum_{k=1}^{t} D_{Z,k}}\,\mathrm{d}p},$$

then, we select, for any $n\in\mathbb{N}$ and $x\in(0,1)$, a binomial random variable $\mathrm{Bin}\,(n,x)$ of parameters n and x, to get

$$(\mathrm{VI})=2\left|\frac{1+\sum_{k=1}^{t} D_{Z,k}}{t+2}-\frac{\int_{\left[\frac{1}{2}-\varepsilon_M,\frac{1}{2}+\varepsilon_M\right]} p^{1+\sum_{k=1}^{t} D_{Z,k}}(1-p)^{t-\sum_{k=1}^{t} D_{Z,k}}\,\mathrm{d}p}{\int_{\left[\frac{1}{2}-\varepsilon_M,\frac{1}{2}+\varepsilon_M\right]} p^{\sum_{k=1}^{t} D_{Z,k}}(1-p)^{t-\sum_{k=1}^{t} D_{Z,k}}\,\mathrm{d}p}\right|^2 \mathbb{I}\left\{\frac{\sum_{k=1}^{t} D_{Z,k}}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}$$
$$=\left[2\left|\frac{s+1}{t+2}\frac{t-s+1}{t+2}\frac{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)=s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)=s+1\right]\right|}{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)\geqslant s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\right|}\right|^2 \mathbb{I}\left\{\frac{s}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]_{|s=\sum_{k=1}^{t} D_{Z,k}}$$
$$\leqslant\left[2\left|\frac{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)=s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)=s+1\right]\right|}{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)\geqslant s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\right|}\right|^2 \mathbb{I}\left\{\frac{s}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]_{|s=\sum_{k=1}^{t} D_{Z,k}}$$
$$\leqslant\left[2\left|\frac{\max\left(\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)=s+1\right],\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)=s+1\right]\right)}{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)\geqslant s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\right|}\right|^2 \mathbb{I}\left\{\frac{s}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]_{|s=\sum_{k=1}^{t} D_{Z,k}}$$
$$\leqslant\left[2\left|\frac{\max\left(\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)\leqslant s+1\right],\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\right)}{\left|\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)\geqslant s+1\right]-\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\right|}\right|^2 \mathbb{I}\left\{\frac{s}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\right]_{|s=\sum_{k=1}^{t} D_{Z,k}}$$
$$=:(\heartsuit)\,.$$

Now, since, for any $s,t\in\mathbb{N}$, if $\frac{s}{t}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]$ and $t\geqslant\frac{6}{7}M$ we have that

$$\frac{s+1}{t+1}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{3},\frac{1}{2}+\frac{\varepsilon_M}{3}\right],\quad\frac{s+1}{t+2}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{3},\frac{1}{2}+\frac{\varepsilon_M}{3}\right]$$

we get, using Hoeffding inequality in each of the following inequalities, that

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)\geqslant s+1\right]\\
&\quad=\mathbb{P}\left[\frac{1}{t+1}\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)-\left(\frac{1}{2}+\varepsilon_M\right)\geqslant\frac{s+1}{t+1}-\left(\frac{1}{2}+\varepsilon_M\right)\right]\\
&\quad=1-\mathbb{P}\left[\frac{1}{t+1}\mathrm{Bin}\left(t+1,\frac{1}{2}+\varepsilon_M\right)-\left(\frac{1}{2}+\varepsilon_M\right)<-\left(\left(\frac{1}{2}+\varepsilon_M\right)-\frac{s+1}{t+1}\right)\right]\\
&\quad\geqslant 1-\exp\left(-2\left(\left(\frac{1}{2}+\varepsilon_M\right)-\frac{s+1}{t+1}\right)^2(t+1)\right)\geqslant 1-\exp\left(-\frac{8}{9}\varepsilon_M^2(t+1)\right)
\end{aligned}\tag{A.3}$$

while

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\\
&\quad=\ \mathbb{P}\left[\frac{1}{t+1}\,\mathrm{Bin}\left(t+1,\frac{1}{2}-\varepsilon_M\right)-\left(\frac{1}{2}-\varepsilon_M\right)\geqslant\frac{s+1}{t+1}-\left(\frac{1}{2}-\varepsilon_M\right)\right]\\
&\quad\leqslant\exp\left(-2\left(\frac{s+1}{t+1}-\left(\frac{1}{2}-\varepsilon_M\right)\right)^2(t+1)\right)\leqslant\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+1)\right)
\end{aligned}\tag{A.4}$$

and

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)\leqslant s+1\right]\\
&\quad=\mathbb{P}\left[\frac{1}{t+2}\,\mathrm{Bin}\left(t+2,\frac{1}{2}+\varepsilon_M\right)-\left(\frac{1}{2}+\varepsilon_M\right)\leqslant\frac{s+1}{t+2}-\left(\frac{1}{2}+\varepsilon_M\right)\right]\\
&\quad\leqslant\exp\left(-2\left(\frac{s+1}{t+2}-\left(\frac{1}{2}+\varepsilon_M\right)\right)^2(t+2)\right)\leqslant\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+2)\right)
\end{aligned}\tag{A.5}$$

and, finally

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)\geqslant s+1\right]\\
&\quad=\mathbb{P}\left[\frac{1}{t+2}\,\mathrm{Bin}\left(t+2,\frac{1}{2}-\varepsilon_M\right)-\left(\frac{1}{2}-\varepsilon_M\right)\geqslant\frac{s+1}{t+2}-\left(\frac{1}{2}-\varepsilon_M\right)\right]\\
&\quad\leqslant\exp\left(-2\left(\frac{s+1}{t+2}-\left(\frac{1}{2}-\varepsilon_M\right)\right)^2(t+2)\right)\leqslant\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+2)\right).
\end{aligned}\tag{A.6}$$

Plugging the inequalities (A.3), (A.4), (A.5), (A.6) into (♥), we get

$$\begin{aligned}
(\heartsuit)&\leqslant 2\left(\frac{\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+2)\right)}{1-2\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+1)\right)}\right)^2\mathbb{I}\left\{\frac{1}{t}\sum_{k=1}^{t}D_{Z,k}\in\left[\frac{1}{2}-\frac{\varepsilon_M}{6},\frac{1}{2}+\frac{\varepsilon_M}{6}\right]\right\}\\
&\leqslant 2\left(\frac{\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+2)\right)}{1-2\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+1)\right)}\right)^2
\end{aligned}$$

and hence

$$\begin{aligned}
(\mathrm{IV})\leqslant(\circ)&\leqslant\sqrt{\frac{4}{147}\cdot\frac{1}{t}}\cdot\sqrt{\mathbb{E}\left[(\mathrm{V})+(\mathrm{VI})\right]}\leqslant\sqrt{\frac{4}{147}\cdot\frac{1}{t}}\cdot\sqrt{\frac{18}{t^2}+2\left(\frac{\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+2)\right)}{1-2\exp\left(-\frac{8}{9}\varepsilon_M^2\,(t+1)\right)}\right)^2}\\
&=\sqrt{\frac{4}{147}}\sqrt{18+2\left(t\frac{\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}{1-2\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}\right)^2\cdot\frac{1}{t^{3/2}}}
\end{aligned}$$

Putting everything together, we have:

$$\mathbb{E}\left[\left(Z-\mathbb{E}\left[Z\mid D_{Z,1},\ldots,D_{Z,t}\right]\right)^2\right]\geqslant(\mathrm{I})-2\cdot(\mathrm{II})\geqslant\frac{4}{147}\cdot\frac{1}{t}-2\cdot\big((\mathrm{III})+(\mathrm{IV})\big)$$

$$\geqslant \frac{4}{147}\cdot\frac{1}{t} - 2\cdot\left(\frac{2}{9}\exp\left(-\frac{49}{162}\cdot\frac{t}{M^2}\right) + \sqrt{\frac{4}{147}}\sqrt{18 + 2\left(t\frac{\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}{1-2\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}\right)^2}\cdot\frac{1}{t^{3/2}}\right) =: (\spadesuit)$$

Elementary computations show that:

- if $t \geqslant 274M^4$ then $\frac{4}{9}\exp\left(-\frac{49}{162}\cdot\frac{t}{M^2}\right) \leqslant \frac{1}{147}\cdot\frac{1}{t}$
- if $t \geqslant 2M^4$ then $\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right) \leqslant \frac{1}{t}$
- if $t \geqslant \frac{4}{100}M^2$ then $1 - 2\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right) \geqslant \frac{1}{2}$
- therefore, if $t \geqslant \max\left(2M^4, \frac{4}{100}M^2, 61152\right)$, we have

$$2\sqrt{\frac{4}{147}}\sqrt{18 + 2\left(t\frac{\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}{1-2\exp\left(-\frac{392}{9}\frac{t+2}{M^2}\right)}\right)^2}\cdot\frac{1}{t^{3/2}} \leqslant \frac{1}{147}\cdot\frac{1}{t}.$$

These together with $(\spadesuit)$ implies that, if $t \geqslant \max\left(274M^4, 61152\right)$ (which is in particular implied by $t \geqslant 580M^4$), we have that

$$\mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid D_{Z,1}, \ldots, D_{Z,t}\right]\right)^2\right] \geqslant \frac{2}{147}\cdot\frac{1}{t}.$$

In conclusion, if $2\left(t-1\right) \geqslant 580M^4$ (which is again implied by $t \geqslant 580M^4$), we have that

$$\mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid D_{Z,1}, \ldots, D_{Z,2(t-1)}\right]\right)^2\right] \geqslant \frac{1}{147}\cdot\frac{1}{t-1}.$$

A.15 Inverse-Transformation Representability with One Bit and Two Environments

In this section, we denote the Lebesgue measure on $[0,1]$ by $\mathbb{L}$.

We recall that given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(\Omega, \mathcal{F})$, we say that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and we write $\mathbb{Q} \ll \mathbb{P}$ if for all $E \in \mathcal{F}$ such that $\mathbb{P}[E] = 0$, it holds that $\mathbb{Q}[E] = 0$. Moreover, if $\mathbb{Q} \ll \mathbb{P}$, the Radon-Nikodym theorem states that there exists a density (called Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ and denoted by) $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \colon \Omega \to [0,\infty)$ such that, for all $E \in \mathcal{F}$, it holds that

$$\mathbb{Q}[E] = \int_E \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega)\,\mathrm{d}\mathbb{P}(\omega)\,.$$

For a reference of the previous result, see [29, Theorem 13.4].

Moreover, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ is a measurable space, and X is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, we denote by $\mathbb{P}_X$ the push-forward measure of $\mathbb{P}$ by X, i.e., the probability measure defined on $\mathcal{F}_{\mathcal{X}}$ by $\mathbb{P}_X[F] \coloneqq \mathbb{P}[X \in F]$, for all $F \in \mathcal{F}_{\mathcal{X}}$.

If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are two measurable spaces, we denote by $\mathcal{F} \otimes \mathcal{F}'$ the σ-algebra of subsets of $\Omega \times \Omega'$ generated by the collection of subsets of the form $F \times F'$, where $F \in \mathcal{F}$ and $F' \in \mathcal{F}'$. If $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega', \mathcal{F}', \mathbb{P}')$ are two probability spaces, we denote the product measure of $\mathbb{P}$ and

$\mathbb{P}'$ by $\mathbb{P} \otimes \mathbb{P}'$, i.e., $\mathbb{P} \otimes \mathbb{P}'$ is the unique probability measure defined on $\mathcal{F} \otimes \mathcal{F}'$ which satisfies $(\mathbb{P} \otimes \mathbb{P}')[F \times F'] = \mathbb{P}[F]\mathbb{P}'[F']$, for all $E \in \mathcal{F}$ and $E' \in \mathcal{F}'$.

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ are measurable spaces, X is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, and Y is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, we denote the conditional probability of X given Y by $\mathbb{P}_{X|Y}$, i.e., $\mathbb{P}_{X|Y}[E] = \mathbb{P}[X \in E \mid Y]$, for each $E \in \mathcal{F}_{\mathcal{X}}$. In this case, for each $E \in \mathcal{F}_{\mathcal{X}}$, we recall that $\mathbb{P}_{X|Y}[E]$ is a $\sigma(Y)$-measurable random variable. Furthermore, if X' is another random variable from $(\Omega, \mathcal{F})$ to some measurable space $(\mathcal{X}', \mathcal{F}_{\mathcal{X}'})$, f and g are two real-valued bounded measurable functions (respectively from $(\mathcal{X} \otimes \mathcal{Y}, \mathcal{F}_{\mathcal{X}} \otimes \mathcal{F}_{\mathcal{Y}})$ to the reals and from $(\mathcal{X}' \otimes \mathcal{Y}, \mathcal{F}_{\mathcal{X}'} \otimes \mathcal{F}_{\mathcal{Y}})$ to the reals), and both $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{X}, \mathcal{F}_{\mathcal{X}'})$ are measurable spaces that arise from considering the Borel subsets of separable and complete metric space $(\mathcal{X}, d)$ and $(\mathcal{X}', d')$ respectively, it holds that

$$\mathbb{E}\big[f(X,Y)g(X',Y) \mid Y\big] = \mathbb{E}\big[f(X,Y) \mid Y\big] \cdot \mathbb{E}\big[g(X',Y) \mid Y\big]$$

whenever

$$\mathbb{P}_{(X,X')|Y} = \mathbb{P}_{X|Y} \otimes \mathbb{P}_{X'|Y} \ .$$

A.15.1 Our Inverse-Transformation Result

In this section, we present a theorem that extends, in spirit, the classic inverse transformation method. This result that can be of independent interest for replacing a type of feedback with another of better quality in lower-bound constructions based on reductions to simpler games.

Definition 9 (Inverse-transformation representability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{B}$ be the Borel σ-algebra of $[0,1]$. We say that $\mathbb{P}$ is* inverse-transformation-representable *if there exists a measurable function ψ from $\big([0,1], \mathcal{B}\big)$ to $(\Omega, \mathcal{F})$ such that $\mathbb{P} = \mathbb{L}_{\psi}$.*[**]

The following theorem is a simple consequence of [134, Corollary A.11], and shows "inverse-transformation representability in separable and complete metric spaces".

Theorem 43. *Suppose that $(\mathcal{Y}, d)$ is a separable and complete metric space, with $\mathcal{F}_{\mathcal{Y}}$ as the Borel σ-algebra of $(\mathcal{Y}, d)$. Then any probability measure defined on $\mathcal{F}_{\mathcal{Y}}$ is inverse-transformation-representable.*

We are now ready to state the main theorem of this section. When we are uncertain about the underlying probability according to which some samples are drawn, and the uncertainty is between two probability measure $\mathbb{P}$ and $\mathbb{Q}$, the theorem provides a characterization under which we can simulate a random variable Y using some independent random seed U and having access to a 1-bit random variable X. This theorem can be of independent interest as a tool for lower bound reductions in online learning problems, as we used for example in Theorem 17. It establishes "One-bit/two-environments inverse-transformation representability in separable and complete metric spaces".

Theorem 44. *Suppose that $(\mathcal{Y}, d)$ is a separable and complete metric space with $\mathcal{F}_{\mathcal{Y}}$ as the Borel σ-algebra of $(\mathcal{Y}, d)$. Let $(\Omega, \mathcal{F})$ be a measurable space, X a random variable from $(\Omega, \mathcal{F})$ to $\big(\{0,1\}, 2^{\{0,1\}}\big)$, Y a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, and U random variable from $(\Omega, \mathcal{F})$ to $\big([0,1], \mathcal{B}\big)$,*

[**]We recall that $\mathbb{L}$ is the Lebesgue measure on $\mathcal{B}$.

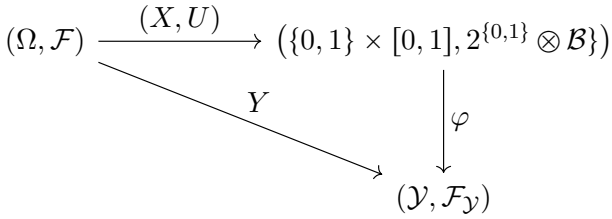


Figure A.1: Pictorial representation of Theorem 44. The way to interpret it is not event by event but in probability: the probability of a measurable set in $\mathcal{F}_Y$ can be computed in Ω equivalently via the pullback of Y, or of $\varphi \circ (X, U)$.

where $\mathcal{B}$ is the Borel σ-algebra of $[0,1]$. Suppose that $\mathbb{P}, \mathbb{Q}$ are probability measures defined on $\mathcal{F}$, and $p \in (0,1)$, $q \in [0,1]$ are such that:

- $\mathbb{P}[X = 1] = p$ *and* $\mathbb{Q}[X = 1] = q$.
- U *is a uniform random variable on* $[0,1]$ *both under* $\mathbb{P}$ *and* $\mathbb{Q}$, *i.e., we have that* $\mathbb{P}_U = \mathbb{L} = \mathbb{Q}_U$.
- U *is independent of* X *both under* $\mathbb{P}$ *and* $\mathbb{Q}$, *i.e.,* $\mathbb{P}_{(X,U)} = \mathbb{P}_X \otimes \mathbb{P}_U$ *and* $\mathbb{Q}_{(X,U)} = \mathbb{Q}_X \otimes \mathbb{Q}_U$.

Then, the following are equivalent:

1. *There exists a measurable function φ from $\big(\{0,1\} \times [0,1], 2^{\{0,1\}} \otimes \mathcal{B}\big)$ to $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$ such that*

$$\mathbb{P}_Y = \mathbb{P}_{\varphi(X,U)} \qquad \text{and} \qquad \mathbb{Q}_Y = \mathbb{Q}_{\varphi(X,U)} \, .$$

2. *$\mathbb{Q}_Y \ll \mathbb{P}_Y$, and $\mathbb{P}_Y$-almost-surely it holds that*

$$\min \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \leqslant \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} \leqslant \max \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \, .$$

Proof. We divide the proof in two parts, depending on whether or not $p = q$.

Assume first that $p \neq q$. In this case, we will prove the chain of equivalencies

$$\text{Item 1} \quad \Leftrightarrow \quad \text{Item a} \quad \Leftrightarrow \quad \text{Item b} \quad \Leftrightarrow \quad \text{Item c} \quad \Leftrightarrow \quad \text{Item 2} \ ,$$

where Item a, Item b, and Item c are the following propositions:

a) There exists two probability measures μ_0 and μ_1 over $\mathcal{F}_\mathcal{Y}$ such that

$$\mathbb{P}_Y = (1-p)\mu_0 + p\mu_1 \qquad \text{and} \qquad \mathbb{Q}_Y = (1-q)\mu_0 + q\mu_1 \, .$$

b) $\frac{q}{q-p}\mathbb{P}_Y - \frac{p}{q-p}\mathbb{Q}_Y \geqslant 0$ and $\frac{1-p}{q-p}\mathbb{P}_Y - \frac{1-q}{q-p}\mathbb{Q}_Y \geqslant 0$.

c) $\mathbb{Q}_Y \ll \mathbb{P}_Y$ and $\min\big(\frac{q}{p}, \frac{1-q}{1-p}\big) \leqslant \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} \leqslant \max\big(\frac{q}{p}, \frac{1-q}{1-p}\big)$ for all $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] > 0$.

We begin by proving that Item 1 is equivalent to Item a. Assume Item 1. Define $\mu_0 \coloneqq \mathbb{P}_{\varphi(0,U)}$ and $\mu_1 \coloneqq \mathbb{P}_{\varphi(1,U)}$. Since U is uniform under both under $\mathbb{P}$ and $\mathbb{Q}$, it also holds that $\mu_0 = \mathbb{Q}_{\varphi(0,U)}$ and $\mu_1 = \mathbb{Q}_{\varphi(1,U)}$. Thus

$$\mathbb{P}_Y = \mathbb{P}_{\varphi(X,U)} = (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-p)\mu_0 + p\mu_1$$

$$\mathbb{Q}_Y = \mathbb{Q}_{\varphi(X,U)} = (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = (1-q)\mu_0 + q\mu_1 \ ,$$

where we used that fact that X and U are independent both under $\mathbb{P}$ and $\mathbb{Q}$ and that $\mathbb{P}[X=1]=p$, $\mathbb{Q}[X=1]=q$. This proves Item a.

Vice versa, assume Item a. By Theorem 43, we can find two measurable functions ψ_0, ψ_1 from $\big([0,1],\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$ such that $\mu_0 = \mathbb{L}_{\psi_0}$ and $\mu_1 = \mathbb{L}_{\psi_1}$ and define

$$\varphi(x,u) \coloneqq \begin{cases} \psi_0(u) & \text{if } x=0 \\ \psi_1(u) & \text{if } x=1 \end{cases}$$

for all $x \in \{0,1\}$ and $u \in [0,1]$. Then φ is a measurable function from $\big(\{0,1\}\times[0,1], 2^{\{0,1\}}\otimes\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$, and since X is independent of U and U is uniform on $[0,1]$ both under $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}
\mathbb{P}_{\varphi(X,U)} &= (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-p)\mathbb{P}_{\psi_0(U)} + p\mathbb{P}_{\psi_1(U)} \\
&= (1-p)\mathbb{L}_{\psi_0} + p\mathbb{L}_{\psi_1} = (1-p)\mu_0 + p\mu_1 = \mathbb{P}_Y \\
\mathbb{Q}_{\varphi(X,U)} &= (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = (1-q)\mathbb{Q}_{\psi_0(U)} + q\mathbb{Q}_{\psi_1(U)} \\
&= (1-q)\mathbb{L}_{\psi_0} + q\mathbb{L}_{\psi_1} = (1-q)\mu_0 + q\mu_1 = \mathbb{Q}_Y
\end{aligned}$$

This proves Item 1 and in turn yields that Item 1 is equivalent to Item a.

We now prove that Item a is equivalent to Item b. Assume Item a. Then, for each $A \in \mathcal{F}_\mathcal{Y}$ we have that the pair $\big(\mu_0[A], \mu_1[A]\big)$ is the (only) solution of the linear system

$$\begin{cases} (1-p)x_0 + px_1 &= \mathbb{P}_Y[A] \\ (1-q)x_0 + qx_1 &= \mathbb{Q}_Y[A] \end{cases}$$

in the two variables (x_0, x_1), which implies

$$\mu_0[A] = \frac{q}{q-p}\mathbb{P}_Y[A] - \frac{p}{q-p}\mathbb{Q}_Y[A] \qquad \text{and} \qquad \mu_1[A] = \frac{1-p}{q-p}\mathbb{Q}_Y[A] - \frac{1-q}{q-p}\mathbb{P}_Y[A] \ .$$

Since μ_0 and μ_1 are (non-negative) measures, this implies Item b.

Vice versa, assume Item b. Define

$$\mu_0 \coloneqq \frac{q}{q-p}\mathbb{P}_Y - \frac{p}{q-p}\mathbb{Q}_Y \qquad \text{and} \qquad \mu_1 \coloneqq \frac{1-p}{q-p}\mathbb{Q}_Y - \frac{1-q}{q-p}\mathbb{P}_Y \ .$$

Since μ_0 and μ_1 are a linear combination of measures, they are signed measures and, by Item b, actually, they are (non-negative) measures. The fact that they are also probability measures follows trivially from $\mathbb{P}_Y[\mathcal{Y}] = 1 = \mathbb{Q}_Y[\mathcal{Y}]$. Now, a direct verification shows that $\mathbb{P}_Y = (1-p)\mu_0 + p\mu_1$ and $\mathbb{Q}_Y = (1-q)\mu_0 + q\mu_1$, i.e., that Item a holds. We have then proved that Item a is equivalent to Item b.

We now prove that Item b is equivalent to Item c. Firstly, note that by elementary linear-algebra (dividing by $\widetilde{p}$ and solving by $\widetilde{q}/\widetilde{p}$ the linear system of inequalities), for each $\widetilde{q} \in [0,1]$ and $\widetilde{p} \in (0,1]$,

the following equivalence holds

$$\begin{cases} \frac{q}{q-p}\widetilde{p} - \frac{p}{q-p}\widetilde{q} \geqslant 0 \\ \frac{1-p}{q-p}\widetilde{q} - \frac{1-q}{q-p}\widetilde{p} \geqslant 0 \end{cases} \iff \min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \leqslant \frac{\widetilde{q}}{\widetilde{p}} \leqslant \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \tag{A.7}$$

Assume Item b. Note that if $p < q$ (resp., $q < p$), then if $A \in \mathcal{F}_{\mathcal{Y}}$ is such that $\mathbb{P}_Y[A] = 0$, the first (resp., second) inequality in Item b implies that also $\mathbb{Q}_Y[A] = 0$, which in turn yields $\mathbb{Q}_Y \ll \mathbb{P}_Y$. Furthermore, for each $A \in \mathcal{F}_{\mathcal{Y}}$ such that $\mathbb{P}_Y[A] \neq 0$, the equivalence in (A.7) with $\widetilde{p} \coloneqq \mathbb{P}_Y[A]$ and $\widetilde{q} \coloneqq \mathbb{Q}_Y[A]$ implies that

$$\min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \leqslant \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} \leqslant \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right)$$

which yields Item c.

Vice versa, assume Item c. Note that Item b holds

- For all $A \in \mathcal{F}_{\mathcal{Y}}$ such that $\mathbb{P}_Y[A] = 0$, because in this case also $\mathbb{Q}_Y[A] = 0$
- For all $A \in \mathcal{F}_{\mathcal{Y}}$ such that $\mathbb{P}_Y[A] \neq 0$, by the equivalence in (A.7) with $\widetilde{p} \coloneqq \mathbb{P}_Y[A]$ and $\widetilde{q} \coloneqq \mathbb{Q}_Y[A]$

This proves that Item b and Item c are equivalent.

We now prove that Item c is equivalent to Item 2. Assume Item c. Assume by contradiction that Item 2 does not hold. Then, there exists $A \in \mathcal{F}_{\mathcal{Y}}$ such that $\mathbb{P}_Y[A] > 0$ such that either for all $y \in A$ it holds that $\max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) < \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}(y)$ or it holds that $\min\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) > \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}(y)$. In the first case

$$\max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) = \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \geqslant \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} = \frac{1}{\mathbb{P}_Y[A]}\int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}\,\mathrm{d}\mathbb{P}_Y > \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right)\,,$$

yielding the contradiction we were seeking. The second case yields a contradiction in an analogous manner.

Vice versa, assume Item 2. Then, if $A \in \mathcal{F}_{\mathcal{Y}}$ is such that $\mathbb{P}_Y[A] > 0$, notice that

$$\min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) = \min\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \leqslant \frac{1}{\mathbb{P}_Y[A]}\int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}\,\mathrm{d}\mathbb{P}_Y \leqslant \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) = \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right)$$

which together with

$$\frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} = \frac{1}{\mathbb{P}_Y[A]}\int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}\,\mathrm{d}\mathbb{P}_Y$$

(since $\mathbb{Q}_Y \ll \mathbb{P}_Y$), implies Item c. This proves that Item c and Item 2 are equivalent and shows in turn that Item 1 is equivalent to Item 2 whenever $p \neq q$.

Assume now that $p = q$. Assume Item 1. Since X is independent of U and U is uniform on $[0, 1]$ both under $\mathbb{P}$ and $\mathbb{Q}$, we get

$$\mathbb{P}_Y = \mathbb{P}_{\varphi(X,U)} = (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = \mathbb{Q}_{\varphi(X,U)} = \mathbb{Q}_Y\,.$$

Hence, in particular $\mathbb{Q}_Y \ll \mathbb{P}_Y$ and $\frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} = 1$ $\mathbb{P}_Y$-almost-surely, which, together with the fact

$$\min\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \leqslant 1 \leqslant \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right)$$

implies Item 2.

Vice versa, assume Item 2. Fix a measurable function ψ from $\big([0,1],\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$ such that $\mathbb{P}_Y = \mathbb{L}_\psi$ (whose existence is guaranteed by Theorem 43). Let $\varphi(x,u) \coloneqq \psi(u)$ for all $x \in \{0,1\}$ and $u \in [0,1]$. Being U uniform both under $\mathbb{P}$ and $\mathbb{Q}$, we get that $\mathbb{P}_{\varphi(X,U)} = \mathbb{P}_{\psi(U)} = \mathbb{L}_\psi = \mathbb{Q}_{\psi(U)} = \mathbb{Q}_{\varphi(X,U)}$. Moreover, since $p = q$, we have that $\min \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} = 1 = \max \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}$, which, together with Item 2, yields that, for any $A \in \mathcal{F}_\mathcal{Y}$,

$$\mathbb{Q}_Y[A] = \int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}\,\mathrm{d}\mathbb{P}_Y = \int_A 1\,\mathrm{d}\mathbb{P}_Y = \mathbb{P}_Y[A]\,,$$

thus $\mathbb{P}_Y = \mathbb{Q}_Y$. Putting everything together, since we proved that all distributions $\mathbb{P}_{\varphi(X,U)}$, $\mathbb{P}_Y$, $\mathbb{Q}_{\varphi(X,U)}$, $\mathbb{Q}_Y$ are equal to each other, we obtain Item 1, concluding the proof. □

A.16 Missing Proofs from Section 2.6.2

This section is devoted to proving the main result of the weakly budget-balanced Section 2.6: under the realistic feedback model, every learner suffers at least $\Omega(T^{3/4})$ regret, even if it is allowed to post two different prices, one to the seller and one (larger) to the buyer, and the sequence of valuations is independent and identically distributed (iid) with a shared bounded density (bd).

Theorem 17. *Consider the problem of repeated bilateral trade in the weakly budget-balanced realistic-feedback model. There exists a numerical constants $c > 50^{-3}$ such that, for any time horizon $T \geqslant 8008$, the minimax regret satisfies*

$$R_T^\mathcal{S} \geqslant cT^{3/4}\,,$$

where $\mathcal{S}$ is the set of all environments such that

(bd) *for each $t \in \mathbb{N}$, the pair (S_t, B_t) admits a density bounded above by $M \geqslant 9$.*

(iid) *$(S_1,B_1),(S_2,B_2),\dots$ is an i.i.d. sequence.*

Proof. We prove this result in several steps: we begin by constructing a hard instance of the learning problem, then we present a related (easier) learning problem and, finally, we show that the minimax regret of the latter (and therefore, the former) is at least $\Omega(T^{3/4})$.

The construction of a hard family of adversaries

Fix any $M \in [9,\infty)$ and $T \geqslant 8008$. Since the regret against an i.i.d. environment is entirely characterized by the distribution that drives the drawing of seller/buyer valuations, we model the environment with probability measures. More precisely, we model the environment with a single sequence of seller/buyer valuations $(S,B),(S_1,B_1),(S_2,B_2),\dots$ whose distribution changes when we change the underlying probability measure. For any strategy α of the learner, we will find an underlying probability measure such that the elements in the process $(S,B),(S_1,B_1),(S_2,B_2)\dots$ are such that their distribution with respect to this probability measure admits a density bounded above by M, the whole process is i.i.d., independent of the player's randomization, and it satisfies

$$\max_{p\in[0,1]} T\mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big] - \mathbb{E}\left[\sum_{t=1}^T \mathrm{gft}\big((P_t,Q_t),(S_t,B_t)\big)\right] \geqslant \frac{1}{50^3}T^{3/4}\,.$$

Let $a \coloneqq 2 \cdot \ln(27/16)$. Define the six disjoint squares (Figure 2.6, left)

$$Q_1 \coloneqq \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{1}{3}, \tfrac{1}{2}\right) , \qquad Q_2 \coloneqq \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{1}{2}, \tfrac{2}{3}\right] , \qquad Q_3 \coloneqq \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{5}{6}, 1\right] ,$$
$$Q_4 \coloneqq \left[\tfrac{5}{6}, 1\right] \times \left[\tfrac{5}{6}, 1\right] , \qquad Q_5 \coloneqq \left[\tfrac{5}{6}, 1\right] \times \left[0, \tfrac{1}{6}\right] , \qquad Q_6 \coloneqq \left[\tfrac{1}{3}, \tfrac{1}{2}\right] \times \left[\tfrac{2}{3}, \tfrac{5}{6}\right] .$$

Fix the base probability density function $f \colon [0,1]^2 \to [0,\infty)$ defined for all $(x,y) \in [0,1]^2$ by

$$f(x,y) \coloneqq \frac{36}{1+8a} \cdot \left(\frac{5-6(y+x)}{6(y-x)} \mathbb{I}_{Q_1}(x,y) + a\mathbb{I}_{Q_2}(x,y) + 2a\mathbb{I}_{Q_3 \cup Q_4 \cup Q_5}(x,y) + \mathbb{I}_{Q_6}(x,y) \right) .$$

We define a set of perturbations of f parameterized by the elements of

$$\Xi \coloneqq \left\{ (v,\varepsilon) \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right) \times \left(0, \tfrac{1}{12}\right) \mid \tfrac{1}{3} + \varepsilon \leqslant v \leqslant \tfrac{1}{2} - \varepsilon \right\} .$$

For all $(v,\varepsilon) \in \Xi$, define the four disjoint rectangles (Figure 2.6, left)

$$R^1_{v,\varepsilon} \coloneqq [v-\varepsilon, v) \times \left[\tfrac{3}{4}, \tfrac{5}{6}\right] , \qquad R^2_{v,\varepsilon} \coloneqq [v-\varepsilon, v) \times \left[\tfrac{2}{3}, \tfrac{3}{4}\right) ,$$
$$R^3_{v,\varepsilon} \coloneqq [v, v+\varepsilon] \times \left[\tfrac{3}{4}, \tfrac{5}{6}\right] , \qquad R^4_{v,\varepsilon} \coloneqq [v, v+\varepsilon] \times \left[\tfrac{2}{3}, \tfrac{3}{4}\right) .$$

and the corresponding perturbation $g_{v,\varepsilon} \colon [0,1]^2 \to \mathbb{R}$ defined for all $(x,y) \in [0,1]^2$ by

$$g_{v,\varepsilon}(x,y) \coloneqq \frac{36}{1+8a} \cdot \left(\mathbb{I}_{R^1_{v,\varepsilon} \cup R^4_{v,\varepsilon}}(x,y) - \mathbb{I}_{R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon}}(x,y) \right) .$$

Note that the rectangles $R^i_{v,\varepsilon}$ are included in Q_6 for all $i \in [4]$ and $(v,\varepsilon) \in \Xi$. We define perturbed density functions by summing together the base probability density function f and one of the perturbations above. Formally, for all $(v,\varepsilon) \in \Xi$, we let

$$f_{v,\varepsilon} \coloneqq f + g_{v,\varepsilon} .$$

Let $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$, for all $(v,\varepsilon) \in \Xi$) be a probability measure such that the sequence of seller/buyer evaluations (S,B), (S_1,B_1), (S_2,B_2), $\dots$ is i.i.d. and the distribution of (S,B) has density f (resp., $f_{v,\varepsilon}$) with respect to the Lebesgue measure. We denote the expectation with respect to $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$, for all $(v,\varepsilon) \in \Xi$) by $\mathbb{E}$ (resp., $\mathbb{E}^{v,\varepsilon}$). Note that f (resp., $f_{v,\varepsilon}$, for all $(v,\varepsilon) \in \Xi$) is bounded above by 9, and hence, by M. Note also that, for each $(v,\varepsilon) \in \Xi$, and $p \in [0,1]$,

$$\begin{aligned}
\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p,(S,B)\big)\big] &= \mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big] + \int_{[0,p]\times[p,1]} (y-x) g_{v,\varepsilon}(x,y) \,\mathrm{d}x\mathrm{d}y \\
&= \mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big] + \frac{1}{6(1+8a)} \cdot \frac{\varepsilon}{144} \cdot \Lambda_{v,\varepsilon}(p) + \frac{1}{6(1+8a)} \cdot \frac{\varepsilon^2}{12} \cdot \Lambda_{\frac{3}{4},\frac{1}{12}}(p) ,
\end{aligned}$$

where, for each $u \in \mathbb{R}$ and each $r > 0$, $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as

$$\Lambda_{u,r} \colon \mathbb{R} \to \mathbb{R} , \qquad x \mapsto \left(1 - \frac{|x-u|}{r}\right)^+ .$$

A direct computation shows that, for each $p \in [0,1]$

$$\mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big] = \frac{1}{6(1+8a)} \cdot \begin{cases} 3p\big(5+29a-6(1+3a)p\big) & \text{if } p \in \big[0,\frac{1}{6}\big] \\ 2+13a & \text{if } p \in \big(\frac{1}{6},\frac{1}{2}\big] \\ -18ap^2+3ap+2(1+8a) & \text{if } p \in \big(\frac{1}{2},\frac{2}{3}\big] \\ -18p^2+15p+10a & \text{if } p \in \big(\frac{2}{3},\frac{5}{6}\big] \\ 72ap(1-p) & \text{if } p \in \big(\frac{5}{6},1\big] \end{cases} \tag{A.8}$$

from which it can be seen that the function $p \mapsto \mathbb{E}\big[\mathrm{gft}\big(p,(S,B)\big)\big]$ is continuous and maximized at every point of the plateau region $\big[\frac{1}{6},\frac{1}{2}\big]$ (Figure 2.6, right). Putting everything together, we see that, for each $(v,\varepsilon) \in \Xi$, the point v is the unique maximizer of the perturbed function $p \mapsto \mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p,(S,B)\big)\big]$, which is increasing on $\big[0,\frac{1}{6}\big]$, constant on $\big[\frac{1}{6},v-\varepsilon\big]$, has a symmetric spike on $[v-\varepsilon,v+\varepsilon]$, becomes constant again on $\big[v+\varepsilon,\frac{1}{2}\big]$, and decreases on $\big[\frac{1}{2},1\big]$. Given that, regardless which is the underlying distribution, the expected gain from trade is maximized on the diagonal $\{(p,q) \in [0,1]^2 \mid p=q\}$, it follows that for each $(v,\varepsilon) \in \Xi$,

$$\max_{(p,q)\in\mathcal{U}} \mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big((p,q),(S,B)\big)\big] = \mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(v,(S,B)\big)\big]\,,$$

where we recall that $\mathcal{U}$ is the upper triangle.

Now, we show that the distribution of the realistic feedback $\big(\mathbb{I}\{S \leqslant p\},\mathbb{I}\{q \leqslant B\}\big)$ is the same regardless of the underlying perturbed probability measure unless the learner selects a pair of prices (p,q) in one of the four rectangles where the perturbations occur.

Claim 6. *For all $(v,\varepsilon) \in \Xi$, $(p,q) \in \mathcal{U} \setminus \bigcup_{k\in[4]} R^k_{v,\varepsilon}$, and $(i,j) \in \{0,1\}^2$, it holds*

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \leqslant p\},\,\mathbb{I}\{q \leqslant B\}\big) = (i,j)\Big] = \mathbb{P}\Big[\big(\mathbb{I}\{S \leqslant p\},\,\mathbb{I}\{q \leqslant B\}\big) = (i,j)\Big]\,.$$

Proof. For each $(v,\varepsilon) \in \Xi$, and each $(p,q) \in \mathcal{U}$, the distribution under $\mathbb{P}^{v,\varepsilon}$ of the 2-bit feedback $\big(\mathbb{I}\{S \leqslant p\},\mathbb{I}\{q \leqslant B\}\big)$ is given, for all $(i,j) \in \{0,1\}^2$, by

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \leqslant p\},\,\mathbb{I}\{q \leqslant B\}\big) = (i,j)\Big] = \begin{cases} \mathbb{P}^{v,\varepsilon}\big[S > p \cap B < q\big] & \text{if } (i,j) = (0,0) \\ \mathbb{P}^{v,\varepsilon}\big[S > p \cap B \geqslant q\big] & \text{if } (i,j) = (0,1) \\ \mathbb{P}^{v,\varepsilon}\big[S \leqslant p \cap B < q\big] & \text{if } (i,j) = (1,0) \\ \mathbb{P}^{v,\varepsilon}\big[S \leqslant p \cap B \geqslant q\big] & \text{if } (i,j) = (1,1) \end{cases}$$

$$= \begin{cases} \int_{(p,1]\times[0,q)} f(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{(p,1]\times[0,q)} g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (0,0) \\ \int_{(p,1]\times[q,1]} f(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{(p,1]\times[q,1]} g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (0,1) \\ \int_{[0,p]\times[0,q)} f(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{[0,p]\times[0,q)} g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (1,0) \\ \int_{[0,p]\times[q,1]} f(x,y)\,\mathrm{d}x\mathrm{d}y + \int_{[0,p]\times[q,1]} g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (1,1) \end{cases}$$

and noting that, by symmetry, all integrals of $g_{v,\varepsilon}$ in the previous formula vanish if (p,q) does not belong to one of the four rectangles $R^1_{v,\varepsilon}, R^2_{v,\varepsilon}, R^3_{v,\varepsilon}, R^4_{v,\varepsilon}$, we get that $(p,q) \notin R^1_{v,\varepsilon} \cup R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon} \cup R^4_{v,\varepsilon}$

implies

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \leqslant p\},\, \mathbb{I}\{q \leqslant B\}\big) = (i,j)\Big] = \mathbb{P}\Big[\big(\mathbb{I}\{S \leqslant p\},\, \mathbb{I}\{q \leqslant B\}\big) = (i,j)\Big]\,.$$

□

It follows that, for any fixed $\varepsilon \in \big(0, \frac{1}{12}\big)$, if the learner wants to locate $v \in \big[\frac{1}{3}+\varepsilon, \frac{1}{2}-\varepsilon\big]$ observing samples of the realistic feedback drawn according to the distribution $\mathbb{P}^{v,\varepsilon}$, since $R^1_{v,\varepsilon} \cup R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon} \cup R^4_{v,\varepsilon} \subset Q_6$, she has to post prices in the region Q_6. However, note that for each $(v,\varepsilon) \in \Xi$ and $(p,q) \in Q_6$

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big((p,q),(S,B)\big)\big] \leqslant \mathbb{E}^{v,\varepsilon}\left[\mathrm{gft}\left(\left(\frac{1}{2},\frac{2}{3}\right),(S,B)\right)\right] \leqslant \mathbb{E}^{v,\varepsilon}\left[\mathrm{gft}\left(\frac{2}{3},(S,B)\right)\right]$$

while posting prices (p',p') for p' belonging to the potentially optimal region $\big[\frac{1}{3},\frac{1}{2}\big]$ would return

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p',(S,B)\big)\big] \geqslant \mathbb{E}^{v,\varepsilon}\left[\mathrm{gft}\left(\frac{1}{2},(S,B)\right)\right]\,.$$

Hence, for each $(v,\varepsilon) \in \Xi$, each $p' \in \big[\frac{1}{3},\frac{1}{2}\big]$ and each $(p,q) \in Q_6$, we have

$$\begin{aligned}\mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big(p',(S,B)\big)\big] &- \mathbb{E}^{v,\varepsilon}\big[\mathrm{gft}\big((p,q),(S,B)\big)\big] \\ &\geqslant \mathbb{E}^{v,\varepsilon}\left[\mathrm{gft}\left(\frac{1}{2},(S,B)\right)\right] - \mathbb{E}^{v,\varepsilon}\left[\mathrm{gft}\left(\frac{2}{3},(S,B)\right)\right] = \frac{a}{2(1+8a)} \in [0.05,0.06] = \Theta(1)\end{aligned}$$

which means that the learner suffers an instantaneous regret of order $\Theta(1)$ when trying to locate where the perturbation occurs.

Define $K \coloneqq \big\lceil T^{1/4}\big\rceil$ and $\varepsilon \coloneqq \frac{1}{2K}$. For each $k \in \{0,\dots,K\}$, define $v_k \coloneqq \frac{1}{3} + (2k-1)\frac{\varepsilon}{6}$. For the sake of convenience, for each $k \in [K]$ denote $\mathbb{P}^{v_k,\frac{\varepsilon}{6}}$ by $\mathbb{P}^k$ and the corresponding expectation by $\mathbb{E}^k$, and similarly, denote $\mathbb{P}$ by $\mathbb{P}^0$ and the corresponding expectation by $\mathbb{E}^0$.

Interlude

Before proceeding further, let's recap what we have obtained so far and where we plan to go. At a high level, we built a problem in which we know in advance the region where the optimal pair of prices belongs (i.e., the diagonal $\{(p,q) \in [0,1]^2 \mid p = q \in [\frac{1}{3},\frac{1}{2}]\}$), but, when the underlying environment is determined by the probability measure $\mathbb{P}^k$ for some $k \in [K]$, in order not to suffer regret $\Omega(\varepsilon T)$, the learner has to detect inside this potentially optimal region where a spike of height (and base) $\Theta(\varepsilon)$ in the reward occurs. This last task can be accomplished only by locating where the perturbation in the base probability measure occurs, which, given the feedback structure, can only be done by playing in the costly region Q_6, suffering instantaneous regret of order $\Omega(1)$ whenever doing so. However, the region Q_6 can be further partitioned into $\Theta(\frac{1}{\varepsilon})$ disjoint rectangles where these perturbations can occur, and again, given the feedback structure, this implies that each of these rectangles deserves its own dedicated exploration. To better highlight this underlying structure, we will show that the bilateral trade problem is no easier than a simplified problem (that we call multi-apple tasting) where the learner can play $2K$ actions, which we may identify with the set $[2K]$, and where the instances we consider are determined by the probability measures $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$. Each (exploring) action $i \in [K]$ gives zero reward (and corresponds to one of the $\Theta(\frac{1}{\varepsilon})$ rectangles

inside the region Q_6), but, if played at time $t \in \mathbb{N}$, it reveals the realization of a Bernoulli random variable $Y_t(i)$ which is, up to a rescaling and a shifting, the reward of the corresponding (exploiting) action $i+K$ at time t. (The reader familiar with the notion of online learning with directed feedback graphs [8] can see that the feedback model described here corresponds to the weakly observable feedback graph in Figure 2.7, left). The biases of these Bernoullis depend on which is the underlying probability measure among $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$. Specifically, for each $i \in [K]$, each $k \in \{0, \dots, K\}$, and each $t \in \mathbb{N}$, the bias of $Y_t(i)$ under $\mathbb{P}^k$ is $\frac{1}{2}$ if $i \neq k$, while it is $\frac{1}{2} + \Theta(\varepsilon)$ if $i = k$. This way, the exploiting actions $K+1, \dots, 2K$ (which correspond to the regions where the spike in the expected gain from trade can occur) have an expected reward of order $\Omega(1)$ regardless of the underlying probability measure, so that the potentially optimal arm is among them. The catch is that no informative feedback is revealed by these K exploiting actions, and only one of them is optimal when the underlying probability measure is one among $\mathbb{P}^1, \dots, \mathbb{P}^K$. Specifically, the arm $i+K$ is the only optimal action when the underlying probability measure is $\mathbb{P}^i$, having an expected reward that is $\Theta(\varepsilon)$ higher that the other potentially optimal actions. Therefore, since spotting the Bernoulli random variable with bias $\frac{1}{2} + \Theta(\varepsilon)$ among the other $K-1$ unbiased Bernoullis requires playing the K exploring actions $\Theta\big(\frac{1}{\varepsilon^2}\big)$ times each, any algorithm for this new problem (and hence, for the bilateral trade problem) should suffer a regret of order $\Omega\big(\min\big(\frac{K}{\varepsilon^2}, \varepsilon T\big)\big) = \Omega(T^{3/4})$ in at least one environment among $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$, given our choices of K and ε. We will now formalize this idea.

The multi-apple tasting problem

We now described the multi-apple tasting problem on $2K$ arms.

Pick a sequence of $\{0,1\}^{2K}$-valued random variables $Y, Y_1, \dots, Y_T$ and a sequence of $[0,1]$-valued random variables $U, U_1, \dots, U_T, V, V_1, \dots, V_T$ such that:

- For each $k \in \{0, \dots, K\}$ the sequence $Y, Y_1, \dots, Y_T$ is $\mathbb{P}^k$-i.i.d.
- Letting $c_{\text{prob}} \coloneqq \frac{7}{2a}$, for each $k \in \{0, \dots, K\}$ and each $i \in [K]$ we have that $Y(i+K) = Y_1(i+K) = \dots = Y_T(i+K) = 0$ and

$$\mathbb{P}^k[Y(i) = 1] = \begin{cases} \frac{1}{2} & \text{if } i \in [K] \backslash \{k\} \\ \frac{1}{2} + c_{\text{prob}} \cdot \varepsilon & \text{if } i = k \end{cases}$$

- For each $k \in \{0, \dots, K\}$ the sequence $V, V_1, \dots, V_T$ is $\mathbb{P}^k$-i.i.d. and $\mathbb{P}^k_V = \mathbb{L}$.
- For each $k \in \{0, \dots, K\}$, we have

$$\mathbb{P}^k_{((S,B),(S_1,B_1),\dots,(S_T,B_T)),(U,U_1,\dots U_T),(Y,Y_1,\dots Y_T),(V,V_1,\dots V_T)\big)} = \mathbb{P}^k_{((S,B),(S_1,B_1),\dots,(S_T,B_T))} \otimes \mathbb{P}^k_{(U,U_1,\dots U_T)} \otimes \mathbb{P}^k_{(Y,Y_1,\dots Y_T)} \otimes \mathbb{P}^k_{(V,V_1,\dots V_T)}$$

The multi-apple tasting problem proceeds as follows. At each time $t \in [T]$, the player can play any action i in the set $[2K]$, receiving no feedback if $i \geqslant K+1$ (modeled by $Y(i) = Y_1(i) = \dots = Y_T(i) = 0$) and feedback $Y_t(i)$ if $i \in [K]$, obtaining in any case (but not observing) a reward $\rho(i, Y_t)$,

where letting $c_{\text{plat}} \coloneqq \frac{a}{2(1+8a)}$ and $c_{\text{spike}} \coloneqq \frac{1}{6(1+8a)} \cdot \frac{1}{144}$,

$$\rho\colon [2K] \times \{0,1\}^{2K} \to \mathbb{R}\,, \qquad (j,y) \mapsto \begin{cases} 0 & \text{if } j \in [K] \\ c_{\text{plat}} + \frac{c_{\text{spike}}}{c_{\text{prob}}} \cdot \big(y(j-K) - \frac{1}{2}\big) & \text{otherwise} \end{cases}$$

Observe that for all $k \in \{0,\dots,K\}$ and $i \in \{K+1,\dots,2K\}$, we have

$$\mathbb{E}^k\big[\rho(i,Y)\big] = \begin{cases} c_{\text{plat}} & \text{if } k \neq i-K \\ c_{\text{plat}} + c_{\text{spike}} \cdot \varepsilon & \text{otherwise} \end{cases}$$

Relating the two problems

To map the bilateral trade problem into the multi-apple tasting problem, we first partition the upper triangle $\mathcal{U}$ in the following $2K$ disjoint regions:

- $\forall k \in [K-1], J_k \coloneqq [v_k - \frac{\varepsilon}{6}, v_k + \frac{\varepsilon}{6}) \times [\frac{2}{3}, \frac{5}{6}]$
- $J_K \coloneqq [v_K - \frac{\varepsilon}{6}, v_K + \frac{\varepsilon}{6}] \times [\frac{2}{3}, \frac{5}{6}]$
- $\forall k \in [K-1], J_{k+K} \coloneqq \{(p,q) \in \mathcal{U} \mid v_k - \frac{\varepsilon}{6} \leqslant p < v_k + \frac{\varepsilon}{6} \text{ and } q < \frac{2}{3}\}$
- $J_{2K} \coloneqq \mathcal{U} \setminus \bigcup_{k=1}^{2K-1} J_k$

Define $\iota\colon \mathcal{U} \to [2K]$ as the map that associates to each $(p,q) \in \mathcal{U}$ the unique $i \in [2K]$ such that $(p,q) \in J_i$ (Figure 2.7, right).

Claim 7. *For any $(p,q) \in \mathcal{U}$ there exists a function $\varphi_{p,q} : \{0,1\} \times [0,1] \to \{0,1\}^2$ such that, for all $k \in \{0,\dots,K\}$, the distributions under $\mathbb{P}^k$ of $\varphi_{p,q}(Y(\iota(p,q)),V)$ and $\big(\mathbb{I}(S \leqslant p), \mathbb{I}\{q \leqslant B\}\big)$ coincide.*

Proof. A direct verification shows that, for all $(p,q) \in Q_6$ and $k \in [K]$, it holds that

$$\min\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(k)}}{\mathrm{d}\mathbb{P}^0_{Y(k)}}\right) = 1 - 2c_{\text{prob}} \cdot \varepsilon \leqslant \frac{\mathrm{d}\mathbb{P}^k_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}}{\mathrm{d}\mathbb{P}^0_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}} \leqslant 1 + 2c_{\text{prob}} \cdot \varepsilon = \max\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(k)}}{\mathrm{d}\mathbb{P}^0_{Y(k)}}\right)$$

and $\mathbb{P}^k_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})} \ll \mathbb{P}^0_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}$. For each $(p,q) \in Q_6$, by Theorem 44, there exists (and we fix)

$$\varphi_{p,q}\colon \{0,1\} \times [0,1] \to \{0,1\}^2$$

such that

$$\mathbb{P}^{\iota(p,q)}_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^{\iota(p,q)}_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})} \qquad \text{and} \qquad \mathbb{P}^0_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^0_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}\,.$$

Since for all $(p,q) \in Q_6$ and all $k \in [K]\setminus\{\iota(p,q)\}$, we have $\mathbb{P}^k_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})} = \mathbb{P}^0_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}$ (by Claim 6) and $\mathbb{P}^k_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^0_{\varphi_{p,q}(Y(\iota(p,q)),V)}$, then, for all $(p,q) \in Q_6$ and all $k \in \{0,\dots,K\}$, it holds that

$$\mathbb{P}^k_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^k_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}\,.$$

Moreover, since for all $(p,q) \in \mathcal{U}\setminus Q_6$ and for all $k \in \{0,\dots,K\}$, it holds that $\mathbb{P}^k_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})} = \mathbb{P}^0_{(\mathbb{I}(S\leqslant p),\mathbb{I}\{q\leqslant B\})}$ (by Claim 6), then, by Theorem 43, there exists (and we fix)

$$\widetilde{\varphi}_{p,q}\colon [0,1] \to \{0,1\}^2$$

such that, for all $k \in \{0, \dots, K\}$, it holds that

$$\mathbb{P}^k_{\widetilde{\varphi}_{p,q}(V)} = \mathbb{P}^k_{(\mathbb{I}(S \leqslant p), \mathbb{I}\{q \leqslant B\})} \,.$$

Defining for all $(p,q) \in \mathcal{U} \backslash Q_6$ and $(y,v) \in \{0,1\} \times [0,1]$, $\varphi_{p,q}(y,v) \coloneqq \widetilde{\varphi}_{p,q}(v)$, we obtain the result. □

For all $(p,q) \in \mathcal{U}$, fix a $\varphi_{p,q}$ as in Claim 7. Now, fix an arbitrary weakly-budget-balanced strategy α for the bilateral trade problem with realistic feedback. If needed, α has sequential access to the seeds $U_1, U_2, \dots$ for randomization purposes. Let $(P_1, Q_1), (P_2, Q_2), \dots$ be the sequence of prices posted by the strategy α observing the two-bit feedback $\big(\mathbb{I}\{S_t \leqslant P_t\}, \mathbb{I}\{Q_t \leqslant B_t\big)\}$ at round t. We now construct another strategy $\widetilde{\alpha}$ (based on α and the sequence of random seeds $V_1, V_2, \dots$) to solve this new problem in the following way:

- For each time $t \in [T]$, we use the algorithm α to select a pair $(\widetilde{P}_t, \widetilde{Q}_t) \in \mathcal{U}$, then play the action $\widetilde{I}_t \coloneqq \iota(\widetilde{P}_t, \widetilde{Q}_t) \in [2K]$.
- For each time $t \in [T]$, whenever the strategy α requests some feedback in $\{0,1\}^2$, we feed α with the feedback $\varphi_{\widetilde{P}_t, \widetilde{Q}_t}\big(Y_t(\widetilde{I}_t), V_t\big) \in \{0,1\}^2$.

By induction on t, Claim 7 implies that for all $k \in \{0, \dots, K\}$ and $t \in [T]$, we have

$$\mathbb{P}^k_{(\widetilde{P}_t, \widetilde{Q}_t)} = \mathbb{P}^k_{(P_t, Q_t)}$$

which, together with the fact that $\mathbb{P}^k_{(\widetilde{P}_t, \widetilde{Q}_t, Y_t)} = \mathbb{P}^k_{(\widetilde{P}_t, \widetilde{Q}_t)} \otimes \mathbb{P}^k_{Y_t}$ for all $k \in \{0, \dots, K\}$ and $t \in [T]$, yields

$$\begin{aligned}
R^k_T(\alpha) &\coloneqq T\mathbb{E}^k\big[\mathrm{gft}\big(v_k, (S,B)\big)\big] - \sum_{t=1}^T \mathbb{E}^k\big[\mathrm{gft}\big((P_t, Q_t), (S_t, B_t)\big)\big] \\
&\geqslant T\mathbb{E}^k\big[\rho(k+K, Y)\big] - \sum_{t=1}^T \mathbb{E}^k\Big[\rho\big(\iota(P_t, Q_t), Y_t\big)\Big] \\
&= T\mathbb{E}^k\big[\rho(k+K, Y)\big] - \sum_{t=1}^T \mathbb{E}^k\Big[\rho\big(\iota(\widetilde{P}_t, \widetilde{Q}_t), Y_t\big)\Big] \\
&= T\mathbb{E}^k\big[\rho(k+K, Y)\big] - \sum_{t=1}^T \mathbb{E}^k\big[\rho(\widetilde{I}_t, Y_t)\big] =: \widetilde{R}^k_T(\widetilde{\alpha}) \,,
\end{aligned}$$

where $R^k_T(\alpha)$ (resp., $\widetilde{R}^k_T(\widetilde{\alpha})$) is the regret suffered by the strategy α (resp., $\widetilde{\alpha}$) after T rounds of the bilateral trade problem with two-bit feedback (resp., the related problem on $2K$ actions) in the environment $\mathbb{P}^k$. Summing over $k \in [K]$ and dividing by K, this implies

$$\frac{1}{K}\sum_{k\in[K]} R^k_T(\alpha) \geqslant \frac{1}{K}\sum_{k\in[K]} \widetilde{R}^k_T(\widetilde{\alpha}) \geqslant \inf_{\bar{\alpha}\in\mathrm{Rand}} \frac{1}{K}\sum_{k\in[K]} \widetilde{R}^k_T(\bar{\alpha}) = \inf_{\bar{\alpha}\in\mathrm{Det}} \frac{1}{K}\sum_{k\in[K]} \widetilde{R}^k_T(\bar{\alpha}) \,,$$

where the first (resp., second) infimum is over the set Rand (resp., Det) all randomized (resp., deterministic) algorithms $\bar{\alpha}$ for the related problem on $2K$ actions, and the last standard equality is a straightforward consequence of the stochastic i.i.d. setting.

We now show that for any deterministic algorithm $\bar{\alpha}$ for the related problem on $2K$ actions, it either holds that $\frac{1}{K}\sum_{k\in[K]} \widetilde{R}^k_T(\bar{\alpha}) \geqslant \frac{1}{50^3} T^{3/4}$ or that $\widetilde{R}^0_T(\bar{\alpha}) \geqslant \frac{1}{50^3} T^{3/4}$. This, together with

the inequalities above will imply that there exists an $k \in \{0, \dots, K\}$ such that $R_T^k(\alpha) \geqslant \frac{1}{50^3} T^{3/4}$, concluding the proof. For any deterministic algorithm $\bar{\alpha}$ for the related problem on $2K$ actions, let $I_1^{\bar{\alpha}}, I_2^{\bar{\alpha}}, \dots$ be the actions played by $\bar{\alpha}$ on the basis of the sequential feedback $Z_1^{\bar{\alpha}}, Z_2^{\bar{\alpha}}, \dots$ and

$$N_t^{\bar{\alpha}} \coloneqq \sum_{i \in [K]} N_t^{\bar{\alpha}}(i) \ , \ M_t^{\bar{\alpha}} \coloneqq \sum_{i \in [K]} M_t^{\bar{\alpha}}(i) \ ,$$
$$\text{where } N_t^{\bar{\alpha}}(i) \coloneqq \sum_{s=1}^{t} \mathbb{I}\{I_s^{\bar{\alpha}} = i\} \ , \ M_t^{\bar{\alpha}}(i) \coloneqq \sum_{s=1}^{t} \mathbb{I}\{I_s^{\bar{\alpha}} = i + K\} \ .$$

Fix an arbitrary deterministic algorithm $\bar{\alpha}$ for the related problem on $2K$ actions. Then

$$\begin{aligned}\frac{1}{K} \sum_{k \in [K]} \widetilde{R}_T^k(\bar{\alpha}) &= \frac{1}{K} \sum_{k \in [K]} \Big(c_{\text{spike}} \cdot \varepsilon \cdot \mathbb{E}^k\big[T - M_T^{\bar{\alpha}}(k) - N_T^{\bar{\alpha}}\big] + (c_{\text{plat}} + c_{\text{spike}} \cdot \varepsilon) \cdot \mathbb{E}^k\big[N_T^{\bar{\alpha}}\big] \Big) \\ &\geqslant c_{\text{spike}} \cdot \varepsilon \left(T - \frac{1}{K} \sum_{k \in [K]} \mathbb{E}^k\big[M_T^{\bar{\alpha}}(k)\big] \right) \eqqcolon (\circ)\end{aligned}$$

Now, since for any $t \in [T]$ the action $I_t^{\bar{\alpha}} = \bar{\alpha}_t(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})$ selected by $\bar{\alpha}$ at round t is a deterministic function of $Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}}$, for each $k \in [K]$, we have

$$\begin{aligned}&\mathbb{E}^k\big[M_T^{\bar{\alpha}}(k)\big] - \mathbb{E}^0\big[M_T^{\bar{\alpha}}(k)\big] \\ &\quad= \sum_{t=2}^{T} \Big(\mathbb{P}^k\big[\bar{\alpha}_t(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}}) = k + K\big] - \mathbb{P}^0\big[\bar{\alpha}_t(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}}) = k + K\big] \Big) \\ &\quad= \sum_{t=2}^{T} \Big(\mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})}\big[\bar{\alpha}_t^{-1}(k + K)\big] - \mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})}\big[\bar{\alpha}_t^{-1}(k + K)\big] \Big) \\ &\quad\leqslant \sum_{t=2}^{T} \big\| \mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})} - \mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})} \big\|_\infty \leqslant \sum_{t=2}^{T} \big\| \mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})} - \mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_{t-1}^{\bar{\alpha}})} \big\|_{\text{TV}} \eqqcolon (\star)\end{aligned}$$

were we $\|\cdot\|_{\text{TV}}$ denotes the total variation norm. We will now prove that, for each $k \in [K]$ and $t \in [T]$, it holds that

$$\Big\| \mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})} - \mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})} \Big\|_{\text{TV}} \leqslant c_{\text{prob}} \cdot \varepsilon \cdot \sqrt{2\mathbb{E}[N_t^{\bar{\alpha}}(k)]} \tag{A.9}$$

By Pinsker's inequality and the chain rule for KL-divergence $\mathcal{D}_{\text{KL}}$, for each $k \in [K]$ and $t \in [T]$, we have

$$\begin{aligned}\Big\| \mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})} - \mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})} \Big\|_{\text{TV}} &\leqslant \sqrt{\frac{1}{2} \mathcal{D}_{\text{KL}}\big(\mathbb{P}^0_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})}, \mathbb{P}^k_{(Z_1^{\bar{\alpha}}, \dots, Z_t^{\bar{\alpha}})} \big)} \\ &\leqslant \sqrt{\frac{1}{2} \left(\mathcal{D}_{\text{KL}}\big(\mathbb{P}^0_{Z_1^{\bar{\alpha}}}, \mathbb{P}^k_{Z_1^{\bar{\alpha}}} \big) + \sum_{s=2}^{t} \mathbb{E}\Big[\mathcal{D}_{\text{KL}}\big(\mathbb{P}^0_{Z_s^{\bar{\alpha}} | Z_1^{\bar{\alpha}}, \dots, Z_{s-1}^{\bar{\alpha}}}, \mathbb{P}^k_{Z_s^{\bar{\alpha}} | Z_1^{\bar{\alpha}}, \dots, Z_{s-1}^{\bar{\alpha}}} \big) \Big] \right)} \eqqcolon (@)\end{aligned}$$

To upper bound (@), note first that, since $T \geqslant 8008$,

$$\frac{1}{2} \left(\ln \frac{1/2}{1/2 - c_{\text{prob}} \cdot \varepsilon} + \ln \frac{1/2}{1/2 + c_{\text{prob}} \cdot \varepsilon} \right) \leqslant 4 \cdot c_{\text{prob}}^2 \cdot \varepsilon^2$$

Then, since $\bar{\alpha}$ is a deterministic algorithm, $I_1^{\bar{\alpha}}$ is a fixed element of $[2K]$, which implies that, for all $k \in [K]$,

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1^{\bar{\alpha}}}, \mathbb{P}^k_{Z_1^{\bar{\alpha}}}\big) \\
&\quad = \left(\ln\left(\frac{\mathbb{P}^0[Y_1(k)=0]}{\mathbb{P}^k[Y_1(k)=0]}\right)\mathbb{P}^0[Y_1(k)=0] + \ln\left(\frac{\mathbb{P}^0[Y_1(k)=1]}{\mathbb{P}^k[Y_1(k)=1]}\right)\mathbb{P}^0[Y_1(k)=1]\right)\mathbb{I}\big\{I_1^{\bar{\alpha}}=k\big\} \\
&\quad = \frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}}\cdot\varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}}\cdot\varepsilon}\right)\cdot\mathbb{I}\{I_1^{\bar{\alpha}}=k\} \leqslant 4\cdot c_{\mathrm{prob}}^2\cdot\varepsilon^2\cdot\mathbb{P}^0[I_1^{\bar{\alpha}}=k]
\end{aligned}$$

Similarly, since $\bar{\alpha}$ is a deterministic algorithm, for all $s \geqslant 2$, the action $I_s^{\bar{\alpha}} = \bar{\alpha}_s(Z_1^{\bar{\alpha}}, \dots, Z_{s-1}^{\bar{\alpha}})$ selected by $\bar{\alpha}$ at time t a function of $Z_1^{\bar{\alpha}}, \dots, Z_{s-1}^{\bar{\alpha}}$ only, which implies, for all $k \in [K]$,

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s^{\bar{\alpha}}|Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}}, \mathbb{P}^k_{Z_s^{\bar{\alpha}}|Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}}\big) \\
&\quad = \mathbb{E}^0\left[\ln\left(\frac{\mathbb{P}^0[Z_s^{\bar{\alpha}}=0 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]}{\mathbb{P}^k[Z_s^{\bar{\alpha}}=0 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]}\right)\mathbb{P}^0[Z_s^{\bar{\alpha}}=0 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]\right. \\
&\qquad\qquad \left. + \ln\left(\frac{\mathbb{P}^0[Z_s^{\bar{\alpha}}=1 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]}{\mathbb{P}^k[Z_s^{\bar{\alpha}}=1 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]}\right)\mathbb{P}^0[Z_s^{\bar{\alpha}}=1 \mid Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}}]\right] \\
&\quad = \mathbb{E}^0\Bigg[\left(\ln\left(\frac{\mathbb{P}^0[Y_s(k)=0]}{\mathbb{P}^k[Y_s(k)=0]}\right)\mathbb{P}^0[Y_s(k)=0] + \ln\left(\frac{\mathbb{P}^0[Y_s(k)=1]}{\mathbb{P}^k[Y_s(k)=1]}\right)\mathbb{P}^0[Y_s(k)=1]\right) \\
&\qquad\qquad \times \mathbb{I}\big\{\bar{\alpha}_s(Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}})=k\big\}\Bigg] \\
&\quad = \frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}}\cdot\varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}}\cdot\varepsilon}\right)\mathbb{P}^0\big[\bar{\alpha}_s(Z_1^{\bar{\alpha}},\dots,Z_{s-1}^{\bar{\alpha}})=k\big] \\
&\quad \leqslant 4\cdot c_{\mathrm{prob}}^2\cdot\varepsilon^2\cdot\mathbb{P}^0[I_s^{\bar{\alpha}}=k]\,.
\end{aligned}$$

Plugging the two bounds in (@), we get, for all $k \in [K]$ and $t \in [T]$,

$$(@) \leqslant \sqrt{2\cdot c_{\mathrm{prob}}^2\cdot\varepsilon^2\cdot\sum_{s=1}^{t}\mathbb{P}^0[I_s^{\bar{\alpha}}=k]} \leqslant c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{2\mathbb{E}^0[N_t^{\bar{\alpha}}(k)]}$$

which prove claim (A.9). Therefore, we have, for any $k \in [K]$,

$$\mathbb{E}^k\big[M_T^{\bar{\alpha}}(k)\big] - \mathbb{E}^0\big[M_T^{\bar{\alpha}}(k)\big] \leqslant (\star) \leqslant \sum_{t=2}^{T} c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{2\mathbb{E}^0[N_{t-1}^{\bar{\alpha}}(k)]} \leqslant c_{\mathrm{prob}}\cdot\varepsilon\cdot T\cdot\sqrt{2\mathbb{E}^0[N_T^{\bar{\alpha}}(k)]}\,.$$

Rearranging, averaging, applying Jensen's inequality, and recalling that $\frac{1}{K} = \frac{1}{\lceil T^{1/4}\rceil} \leqslant \frac{1}{10}$, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k\in[K]}\mathbb{E}^k[M_T^{\bar{\alpha}}(k)] &\leqslant \frac{1}{K}\sum_{k\in[K]}\mathbb{E}^0[M_T^{\bar{\alpha}}(k)] + c_{\mathrm{prob}}\cdot\varepsilon\cdot T\cdot\sqrt{2\mathbb{E}^0\left[\frac{1}{K}\sum_{k\in[K]}N_T^{\bar{\alpha}}(k)\right]} \\
&= \frac{1}{K}\mathbb{E}^0[M_T^{\bar{\alpha}}] + c_{\mathrm{prob}}\cdot\varepsilon\cdot T\cdot\sqrt{\frac{2}{K}\mathbb{E}^0[N_T^{\bar{\alpha}}]} \leqslant \left(\frac{1}{10} + c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{\frac{2}{K}\mathbb{E}^0[N_T^{\bar{\alpha}}]}\right)\cdot T\,.
\end{aligned}$$

Substituting this inequality in (○), we obtain

$$(\circ) \geqslant c_{\text{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - c_{\text{prob}} \cdot \varepsilon \cdot \sqrt{\frac{2}{K} \mathbb{E}^0[N_T^{\bar{\alpha}}]} \right) \cdot T \geqslant c_{\text{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - \frac{c_{\text{prob}}}{2} \sqrt{\tau_{\bar{\alpha}}} \right) \cdot T \,,$$

where $\tau_{\bar{\alpha}} \coloneqq \frac{\mathbb{E}^0[N_T^{\bar{\alpha}}]}{\varepsilon T}$.

Now, if $\tau_{\bar{\alpha}} \leqslant \frac{1}{10}$, then, the previous inequality yields

$$\frac{1}{K} \sum_{k \in [K]} \widetilde{R}_T^k(\bar{\alpha}) \geqslant c_{\text{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - \frac{c_{\text{prob}}}{2} \sqrt{\tau_{\bar{\alpha}}} \right) \cdot T \geqslant \frac{1}{50^3} T^{3/4} \,.$$

If, on the other hand, it holds that $\tau_{\bar{\alpha}} > \frac{1}{10}$, then

$$\widetilde{R}_T^0(\bar{\alpha}) \geqslant c_{\text{plat}} \mathbb{E}^0[N_T^{\bar{\alpha}}] = c_{\text{plat}} \tau_{\bar{\alpha}} \varepsilon T > \frac{1}{50^3} T^{3/4} \,.$$

□

Appendix B

The Role of Transparency in Repeated First-Price Auctions with Unknown Valuations

B.1 Missing Details of the Proof of Theorem 20

In this section, we will complete the proof of Theorem 20, showing that the repeated first-price auctions with semi-transparent feedback (in the following, referred to as "our problem") are no easier than a K-armed bandit instance based on the probability measures $\mathbb{P}^1, \dots, \mathbb{P}^K$ introduced in Theorem 20. The structure of the proof is inspired by the proof of Theorem 17 in Appendix A.16, and leverages again the one-bit/two-scenarios inverse transformation representability result of Theorem 44.

The related bandit problem. The action space is $[K]$, where we recall that K was some arbitrarily fixed natural number. Let $Y, Y_1, Y_2, \dots$ be a sequence of $\{0,1\}^K$-valued random variables such that, for any $k \in \{0, 1, \dots, K\}$, the sequence is $\mathbb{P}^k$-i.i.d. and, for all $j \in [K]$

$$\mathbb{P}^k\big[Y(j) = 1\big] = \begin{cases} 1/2 & \text{if } j \neq k \\ 1/2 + 1/(6K) & \text{if } j = k \end{cases}$$

This sequence of latent random variables will determine the rewards of the actions. The reward function is

$$\rho\colon [K] \times \{0,1\} \to [0,1]\,, \qquad (i,y) \mapsto \frac{23 + 2y(i)}{192}$$

and the feedback received after playing an action I_t at time t is $Y_t(I_t)$ (which is equivalent to receiving the bandit feedback $\rho(I_t, Y_t)$ gathered at time t).

For any $k \in \{0, \dots, K\}$ and any $i \in [K]$ the expected reward is

$$\mathbb{E}^k\big[\rho(i,Y)\big] = \begin{cases} \dfrac{1}{8} & \text{if } i \neq k \\ \dfrac{1}{8} + \dfrac{\varepsilon}{144} & \text{if } i = k \end{cases}$$

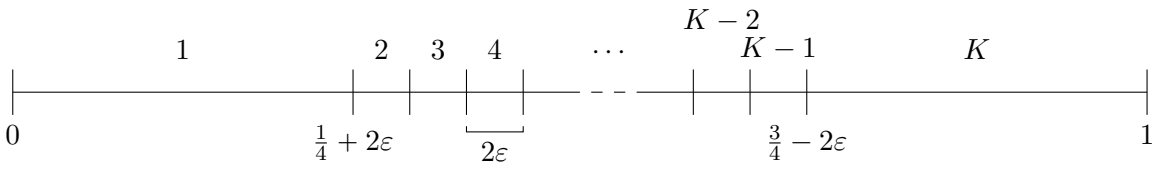


Figure B.1: A representation of the map ι through which the bids in the first-price auction problem are related to the K-arms of the bandit problem. The interval $[0,1]$ is partitioned in K disjoint intervals, the first and the last one of length $1/4 + 2\varepsilon$, and all the ones in between of length 2ε. ι maps each bid to the index of the interval to which it belongs.

Mapping our problem into this bandit problem. Assume that $K \geqslant 3$. We partition the interval $[0,1]$ in the following K disjoint regions: $J_1 = [0, w_1 + \varepsilon)$, $J_k = [w_k - \varepsilon, w_k + \varepsilon)$ (for all $k \in \{2, \dots, K-1\}$), and $J_K = [w_K - \varepsilon, 1]$. We define a function $\iota\colon [0,1] \to [K]$ that maps each point in the interval $[0,1]$ to one of the K arms by mapping each $b \in [0,1]$ to the unique $i \in [K]$ such that $b \in J_i$ (for a pictorial representation of the map ι, see Figure B.1).

Simulating the feedback. To lighten the notation, besides the already defined random functions $\psi_1, \psi_2, \dots$, define also:

$$\psi\colon [0,1] \to \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big)\,, \qquad b \mapsto \begin{cases} (V, \star) & \text{if } b \geqslant M \\ (\star, M) & \text{if } b < M \end{cases}$$

The next lemma shows that we can use the feedback observed in the bandit problem together with some independent noise to simulate exactly the feedback of our problem.

Lemma 21. *For each $b \in [0,1]$, there exists $\varphi_b\colon \{0,1\} \times [0,1] \to \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big)$ such that, if U' is a $[0,1]$-valued random variable such that, for each $k \in \{0, \dots, K\}$, the distribution U' with respect to $\mathbb{P}^k$ is a uniform on $[0,1]$ and U' is $\mathbb{P}^k$-independent of Y, then $\mathbb{P}^k_{\varphi_b(Y(\iota(b)),U')} = \mathbb{P}^k_{\psi(b)}$.*

Proof of Lemma 21. A direct verification shows that, for all $k \in [K]$ and all $b \in [0,1]$, $\mathbb{P}^k_{\psi(b)} \ll \mathbb{P}^0_{\psi(b)}$ (i.e., $\mathbb{P}^k_{\psi(b)}$ is absolutely continuous with respect to $\mathbb{P}^0_{\psi(b)}$) and the Radon-Nikodym derivative of the push-forward measure $\mathbb{P}^k_{\psi(b)}$ with respect to $\mathbb{P}^0_{\psi(b)}$ satisfies, for $\mathbb{P}^0_{\psi(b)}$-a.e. $(v,m) \in \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big)$,

$$\frac{\mathrm{d}\mathbb{P}^k_{\psi(b)}}{\mathrm{d}\mathbb{P}^0_{\psi(b)}}(v,m) = 1 + \varepsilon \cdot \frac{16}{9}\,(v-b)\operatorname{sgn}\left(v - \frac{15}{16}\right)\Lambda_{w_k,\varepsilon}(b)\mathbb{I}\left\{v \in \left[\frac{7}{8}, 1\right]\right\}$$

which implies, for $\mathbb{P}^0_{\psi(b)}$-a.e. $(v,m) \in \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big)$, that

$$\min\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(\iota(b))}}{\mathrm{d}\mathbb{P}^0_{Y(\iota(b))}}\right) = 1 - \frac{4}{3}\varepsilon \leqslant \frac{\mathrm{d}\mathbb{P}^k_{\psi(b)}}{\mathrm{d}\mathbb{P}^0_{\psi(b)}}(v,m) \leqslant 1 + \frac{4}{3}\varepsilon = \max\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(\iota(b))}}{\mathrm{d}\mathbb{P}^0_{Y(\iota(b))}}\right)$$

Thus, for each $b \in [0,1]$, by Theorem 44, there exists (and we fix)

$$\varphi_b\colon \{0,1\} \times [0,1] \to \big([0,1] \times \{\star\}\big) \cup \big(\{\star\} \times [0,1]\big)$$

such that

$$\mathbb{P}^{\iota(b)}_{\varphi_b(Y(\iota(b)),U')} = \mathbb{P}^{\iota(b)}_{\psi(b)} \qquad \text{and} \qquad \mathbb{P}^{0}_{\varphi_b(Y(\iota(b)),U')} = \mathbb{P}^{0}_{\psi(b)} \,.$$

Since for all $b \in [0,1]$ and all $k \in [K]\backslash\{\iota(b)\}$, we have $\mathbb{P}^{k}_{\psi(b)} = \mathbb{P}^{0}_{\psi(b)}$ (by Equation (3.1)) and $\mathbb{P}^{k}_{\varphi_b(Y(\iota(b)),U')} = \mathbb{P}^{0}_{\varphi_b(Y(\iota(b)),U')}$, then, for all $b \in [0,1]$ and all $k \in \{0,\dots,K\}$, it holds that

$$\mathbb{P}^{k}_{\varphi_b(Y(\iota(b)),U')} = \mathbb{P}^{k}_{\psi(b)} \,.$$

□

We now show that any algorithm α for our problem can be transformed into an algorithm $\widetilde{\alpha}$ to solve the bandit problem that suffers no-larger regret. To do so, we begin by formally explaining how algorithms for our problem work.

Functioning of an algorithm α for our problem A randomized algorithm α for our problem is a sequence of functions that take as input a sequence of random seeds $U_1, U_2, \dots$ and some feedback $Z_1, Z_2, \dots$ and generates bids B_t as described below. At time $t = 1$, α selects a bid B_1 as a deterministic function of U_1 and observes feedback $Z_1 = \psi_1(B_1)$. Inductively, for any $t \geqslant 2$, α selects a bid B_t as a deterministic function of $U_1, \dots, U_t, Z_1, \dots, Z_{t-1}$ (where $Z_s = \psi_s(B_s)$, for all $s \in [t-1]$). For all $k \in \{0,\dots,K\}$, the sequence of seeds is a $\mathbb{P}^k$-i.i.d. sequence of uniform random variables on $[0,1]$ that is $\mathbb{P}^k$-independent of $(V,M), (V_1,M_1), (V_2,M_2), \dots$.

Building $\widetilde{\alpha}$ from α We show now how to map α to an algorithm $\widetilde{\alpha}$ (that shares the same seeds for the randomization) for the bandit problem that suffers a worst-case regret that is no larger than that of α.

To do so, consider a sequence $U', U'_1, \dots$ of random variables that, for all $k \in \{0,\dots,K\}$ is a $\mathbb{P}^k$-i.i.d. sequence of uniforms on $[0,1]$ that $\widetilde{\alpha}$ can access as a further source of randomness. We will assume that, for all $k \in \{0,\dots,K\}$, the four sequences $Y, Y_1, \dots, (V,M), (V_1,M_1), \dots, U, U_1, \dots$,and $U', U'_1, \dots$ are independent of each other.

The algorithm $\widetilde{\alpha}$ acts as follows. At time 1, $\widetilde{\alpha}$ plays the arm $\widetilde{I}_1 = \iota(B'_t)$, where $B'_1 = B_1$ is the bid played by α at round $t = 1$ (chosen as a deterministic function of the random seed U_1). Then $\widetilde{\alpha}$ observes the bandit feedback $Y_1(\widetilde{I}_1)$ and feeds back to α the surrogate feedback $Z'_1 = \varphi_{B'_1}\big(Y_1(\widetilde{I}_1), U'_1\big)$. Then, inductively, for any time $t \geqslant 2$, assuming that $\widetilde{\alpha}$ played arms $\widetilde{I}_1, \dots, \widetilde{I}_{t-1}$ and fed back to α the surrogate feedback $Z'_1, \dots, Z'_{t-1}$, then

1. $\widetilde{\alpha}$ plays the arm $\widetilde{I}_t = \iota(B'_t)$, where B'_t is the bid played by α at round t (chosen as a deterministic function of the random seeds $U_1, \dots, U_t$ and past surrogate feedback $Z'_1, \dots, Z'_{t-1}$).

2. $\widetilde{\alpha}$ observes the bandit feedback $Y_t(\widetilde{I}_t)$ and feeds back to α the surrogate feedback $Z'_t = \varphi_{B'_t}\big(Y_t(\widetilde{I}_t), U'_t\big)$.

This way, we defined by induction the randomized algorithm $\widetilde{\alpha}$.

By induction on t, one can show that, if $B_1, B_2, \dots$ are the bids played by α on the basis of the feedback $Z_1 = \psi_1(B_1), Z_2 = \psi_2(B_2), \dots$, then, for all $k \in \{0,\dots,K\}$, we have

$$\mathbb{P}^{k}_{(B_t,Y_t)} = \mathbb{P}^{k}_{(B'_t,Y_t)}$$

which leads to

$$R_T^k(\alpha) = T \cdot \mathbb{E}^k\big[\mathrm{Util}(w_k)\big] - \sum_{t=1}^{T} \mathbb{E}^k\big[\mathrm{Util}_t(B_t)\big] \geqslant T \cdot \mathbb{E}^k\big[\rho(k,Y)\big] - \sum_{t=1}^{T} \mathbb{E}^k\Big[\rho\big(\iota(B_t), Y_t\big)\Big]$$
$$= T \cdot \mathbb{E}^k\big[\rho(k,Y)\big] - \sum_{t=1}^{T} \mathbb{E}^k\Big[\rho\big(\iota(B'_t), Y_t\big)\Big] = T \cdot \mathbb{E}^k\big[\rho(k,Y)\big] - \sum_{t=1}^{T} \mathbb{E}^k\Big[\rho\big(\widetilde{I}_t, Y_t\big)\Big] =: \widetilde{R}_T^k(\widetilde{\alpha}) \ ,$$

where $R_T^k(\alpha)$ is the regret of α in the environment determined by $\mathbb{P}^k$, while the last equality is just a definition. Now we are left to show only that for any algorithm $\widehat{\alpha}$ for the bandit problem which plays actions $I_1, I_2, \ldots$, there exists $k \in [K]$ such that

$$\widetilde{R}_T^k(\widehat{\alpha}) \coloneqq T \cdot \mathbb{E}^k\big[\rho(k,Y)\big] - \sum_{t=1}^{T} \mathbb{E}^k\Big[\rho\big(I_t, Y_t\big)\Big] = \Omega(T^{2/3})$$

(the first equality is a definition). Given that we are competing against a stochastic i.i.d. environments, it is sufficient to show this for deterministic algorithms $\widehat{\alpha}$ for the bandit problem.

Lemma 22. *Fix any deterministic algorithm $\widehat{\alpha}$ for the bandit problem on K actions, then there exists $k \in [K]$ such that $\widetilde{R}_T^k(\widehat{\alpha}) \geqslant \frac{3}{10^4} T^{2/3}$.*

Proof. For any deterministic algorithm $\widehat{\alpha}$ for the bandit problem on K actions, let $I_1, I_2, \ldots$ be the actions played by $\widehat{\alpha}$ on the basis of the sequential feedback received $Z_1, Z_2, \ldots$ and define $N_t(i)$ as the random variables counting the number of times the learning algorithm $\widehat{\alpha}$ plays action i, up to time t, for any $i \in [K]$ and any time $t \in [T]$:

$$N_t(i) = \sum_{s=1}^{t} \mathbb{I}\{I_s = i\}.$$

We relate the expected values of $N_T(k)$ under $\mathbb{P}^0$ and $\mathbb{P}^k$ as a function of the expected number of times the algorithm plays the corresponding actions k. This formalizes the intuition that to discriminate between the different $\mathbb{P}^k$ the learner needs to play exploring actions.

Claim 8. *The following inequality holds true for any $k \in [K]$:*

$$\mathbb{E}^k\big[N_T(k)\big] - \mathbb{E}^0\big[N_T(k)\big] \leqslant \frac{2}{3} \cdot \varepsilon \cdot T \cdot \sqrt{2\mathbb{E}^0[N_T(k)]}. \tag{B.1}$$

Proof of Claim 8. For any $t \in [T]$, the action $I_t = I_t(Z_1, \ldots, Z_{t-1})$ selected by $\widehat{\alpha}$ at round t is a deterministic function of $Z_1, \ldots, Z_{t-1}$, for each $k \in [K]$. In formula, we then have the following

$$\mathbb{E}^k\big[N_T(k)\big] - \mathbb{E}^0\big[N_T(k)\big] = \sum_{t=2}^{T} \Big(\mathbb{P}^k\big[I_t(Z_1, \ldots, Z_{t-1}) = k\big] - \mathbb{P}^0\big[I_t(Z_1, \ldots, Z_{t-1}) = k\big]\Big)$$
$$\leqslant \sum_{t=2}^{T} \big\|\mathbb{P}^k_{(Z_1,\ldots,Z_{t-1})} - \mathbb{P}^0_{(Z_1,\ldots,Z_{t-1})}\big\|_{\mathrm{TV}}, \tag{B.2}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm. We move now our attention towards bounding the total variation norm. To that end we use Pinsker's inequality and apply the chain rule for the KL divergence $\mathcal{D}_{\mathrm{KL}}$. For each $k \in [K]$ and $t \in [T]$ we have the following:

$$\left\|\mathbb{P}^0_{(Z_1,\dots,Z_t)} - \mathbb{P}^k_{(Z_1,\dots,Z_t)}\right\|_{\mathrm{TV}} \leqslant \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{(Z_1,\dots,Z_t)}, \mathbb{P}^k_{(Z_1,\dots,Z_t)}\big)}$$
$$\leqslant \sqrt{\frac{1}{2}\left(\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1}, \mathbb{P}^k_{Z_1}\big) + \sum_{s=2}^{t}\mathbb{E}\Big[\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s|Z_1,\dots,Z_{s-1}}, \mathbb{P}^k_{Z_s|Z_1,\dots,Z_{s-1}}\big)\Big]\right)} \tag{B.3}$$

We bound the two KL terms separately. $\widehat{\alpha}$ is a deterministic algorithm, thus I_1 is a fixed element of $[K]$, which implies that, for all $k \in [K]$,

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1}, \mathbb{P}^k_{Z_1}\big)\\
&\quad= \left(\ln\left(\frac{\mathbb{P}^0[Y_1(k)=0]}{\mathbb{P}^k[Y_1(k)=0]}\right)\mathbb{P}^0[Y_1(k)=0] + \ln\left(\frac{\mathbb{P}^0[Y_1(k)=1]}{\mathbb{P}^k[Y_1(k)=1]}\right)\mathbb{P}^0[Y_1(k)=1]\right)\mathbb{I}\big\{I_1=k\big\}\\
&\quad= \frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}}\cdot\varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}}\cdot\varepsilon}\right)\cdot\mathbb{I}\{I_1=k\}
\end{aligned} \tag{B.4}$$

Similarly, since $\widehat{\alpha}$ is a deterministic algorithm, for all $s \geqslant 2$, the action $I_s = I_s(Z_1,\dots,Z_{s-1})$ selected by $\widehat{\alpha}$ at time t is a function of $Z_1,\dots,Z_{s-1}$ only, which implies, for all $k \in [K]$,

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s|Z_1,\dots,Z_{s-1}}, \mathbb{P}^k_{Z_s|Z_1,\dots,Z_{s-1}}\big)\\
&\quad= \mathbb{E}^0\left[\ln\left(\frac{\mathbb{P}^0[Z_s=0 \mid Z_1,\dots,Z_{s-1}]}{\mathbb{P}^k[Z_s=0 \mid Z_1,\dots,Z_{s-1}]}\right)\mathbb{P}^0[Z_s=0 \mid Z_1,\dots,Z_{s-1}]\right.\\
&\qquad\left.+ \ln\left(\frac{\mathbb{P}^0[Z_s=1 \mid Z_1,\dots,Z_{s-1}]}{\mathbb{P}^k[Z_s=1 \mid Z_1,\dots,Z_{s-1}]}\right)\mathbb{P}^0[Z_s=1 \mid Z_1,\dots,Z_{s-1}]\right]\\
&\quad= \mathbb{E}^0\left[\left(\ln\left(\frac{\mathbb{P}^0[Y_s(k)=0]}{\mathbb{P}^k[Y_s(k)=0]}\right)\mathbb{P}^0[Y_s(k)=0] + \ln\left(\frac{\mathbb{P}^0[Y_s(k)=1]}{\mathbb{P}^k[Y_s(k)=1]}\right)\mathbb{P}^0[Y_s(k)=1]\right)\right.\\
&\qquad\left.\times\,\mathbb{I}\big\{I_s(Z_1,\dots,Z_{s-1})=k\big\}\right]\\
&\quad= \frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}}\cdot\varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}}\cdot\varepsilon}\right)\mathbb{P}^0\big[I_s(Z_1,\dots,Z_{s-1})=k\big]
\end{aligned} \tag{B.5}$$

Now, since $\varepsilon = \frac{1}{4K} \leqslant \frac{1}{4} \leqslant \frac{2}{3}$, the following useful inequality holds:

$$\frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}}\cdot\varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}}\cdot\varepsilon}\right) \leqslant 4\cdot(c_{\mathrm{prob}})^2\cdot\varepsilon^2. \tag{B.6}$$

We can combine the inequalities in Equation (B.4) and Equation (B.5) into Equation (B.3) and plug in the bound in to obtain:

$$\left\|\mathbb{P}^0_{(Z_1,\dots,Z_t)} - \mathbb{P}^k_{(Z_1,\dots,Z_t)}\right\|_{\mathrm{TV}} \leqslant c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{2\mathbb{E}[N_t(k)]}$$

Once we have this upper bound on the total variations of the random variables $(Z_1,\dots,Z_t)$ under $\mathbb{P}^0$ and $\mathbb{P}^k$ we can get back to the initial Equation (B.2) and obtain the desired bound via Jensen:

$$\mathbb{E}^k\big[N_T(k)\big] - \mathbb{E}^0\big[N_T(k)\big] \leqslant \sum_{t=2}^{T} c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{2\mathbb{E}^0[N_{t-1}(k)]} \leqslant c_{\mathrm{prob}}\cdot\varepsilon\cdot T\cdot\sqrt{2\mathbb{E}^0[N_T(k)]}.$$

□

Averaging the quantitative bounds in Claim 8 for all k in $[K]$, and applying Jensen's inequality, we get the following:

$$\frac{1}{K}\sum_{k\in[K]}\mathbb{E}^k[N_T(k)] \leqslant \frac{1}{K}\sum_{k\in[K]}\mathbb{E}^0[N_T(k)] + c_{\mathrm{prob}}\cdot\varepsilon\cdot T\cdot\sqrt{\frac{2}{K}\sum_{k\in[K]}\mathbb{E}^0\left[N_T(k)\right]}$$
$$= \left(\frac{1}{K} + c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{\frac{2T}{K}}\right)\cdot T\ . \tag{B.7}$$

Now, we have all the ingredients to lower bound the average regret suffered by $\widehat{\alpha}$. Note that every time a suboptimal arm is played the learner suffers (expected) instantaneous regret equal $\frac{1}{144}\cdot\varepsilon$. Then, recalling that $\varepsilon = 1/(4K)$ and setting $K = \lceil T^{1/3}\rceil$ we have, for all $T \geqslant 8$,

$$\frac{1}{K}\sum_{k\in[K]}\widetilde{R}_T^k(\widehat{\alpha}) = \frac{1}{K}\sum_{k\in[K]}\Big(c_{\mathrm{spike}}\cdot\varepsilon\cdot\mathbb{E}^k\big[T - N_T(k)\big]\Big) = c_{\mathrm{spike}}\cdot\varepsilon\left(T - \frac{1}{K}\sum_{k\in[K]}\mathbb{E}^k\big[N_T(k)\big]\right)$$
$$\geqslant c_{\mathrm{spike}}\cdot\varepsilon\cdot\left(1 - \frac{1}{K} - c_{\mathrm{prob}}\cdot\varepsilon\cdot\sqrt{\frac{2T}{K}}\right)\cdot T = c_{\mathrm{spike}}\cdot\frac{1}{4K}\cdot\left(1 - \frac{1}{K} - \frac{1}{6K}\cdot\sqrt{\frac{2T}{K}}\right)\cdot T$$
$$\geqslant \frac{1}{8\cdot 144}\Big(\frac{3-\sqrt{2}}{6}\Big)T^{2/3} \geqslant \frac{3}{10^4}T^{2/3}\ .$$

Therefore, for all $T \geqslant 8$, there exists $k \in [K]$ such that $\widetilde{R}_T^k(\widehat{\alpha}) \geqslant (3/10^4)\cdot T^{2/3}$, concluding the proof. □

B.2 Missing Proof of Proposition 2

Proof of Proposition 2. Let $\gamma > 0$. Notice that, for each $t \in \mathbb{N}$, it holds that $\sum_{y\geqslant M_t} p_t(y) \geqslant \gamma$. It follows, for each $x \in \mathcal{X}$ and $t \in \mathbb{N}$, that $\gamma\widehat{g}_t(x) \leqslant 1$, and hence

$$\exp(\gamma\widehat{g}_t(x)) \leqslant 1 + \gamma\widehat{g}_t(x) + (e-2)\gamma^2\big(\widehat{g}_t(x)\big)^2\ .$$

Then, for each $t \in \mathbb{N}$,

$$\frac{\|w_{t+1}\|_1}{\|w_t\|_1} = \sum_{x\in\mathcal{X}}\frac{w_t(x)}{\|w_t\|_1}\exp\big(\gamma\widehat{g}_t(x)\big) \leqslant 1 + \sum_{x\in\mathcal{X}}\frac{w_t(x)}{\|w_t\|_1}\Big(\gamma\widehat{g}_t(x) + (e-2)\gamma^2\big(\widehat{g}_t(x)\big)^2\Big)\ ,$$

which implies

$$\ln\left(\frac{\|w_{t+1}\|_1}{\|w_t\|_1}\right) \leqslant \sum_{x\in\mathcal{X}}\frac{w_t(x)}{\|w_t\|_1}\Big(\gamma\widehat{g}_t(x) + (e-2)\gamma^2\big(\widehat{g}_t(x)\big)^2\Big) \leqslant \frac{\gamma}{1-\gamma}\sum_{x\in\mathcal{X}}p_t(x)\Big(\widehat{g}_t(x) + (e-2)\gamma\big(\widehat{g}_t(x)\big)^2\Big).$$

Now, for each $t \in \mathbb{N}$, let $\mathcal{F}_t$ be the σ-algebra generated by p_t, V_t and M_t and denote by $\mathbb{E}_t := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. First, notice that, for each $t \in \mathbb{N}$ and each $x \in \mathcal{X}$

$$\mathbb{E}_t[\widehat{g}_t(x)] = \mathrm{Util}_t(x)\ ,$$

$$\mathbb{E}_t\left[\sum_{x\in\mathcal{X}} p_t(x)\widehat{g}_t(x)\right] = \mathbb{E}[\mathrm{Util}_t(B_t) \mid V_t, M_t]\ ,$$

and that

$$\mathbb{E}_t\left[\sum_{x\in\mathcal{X}} p_t(x)\big(\widehat{g}_t(x)\big)^2\right] \leqslant \mathbb{E}_t\left[\sum_{x\in\mathcal{X}} p_t(x)\frac{\mathbb{I}\{x \geqslant M_t\}\mathbb{I}\{M_t \leqslant B_t\}}{\big(\sum_{y\geqslant M_t} p_t(y)\big)^2}\right] = \mathbb{E}_t\left[\sum_{x\in\mathcal{X}} p_t(x)\frac{\mathbb{I}\{x \geqslant M_t\}}{\sum_{y\geqslant M_t} p_t(y)}\right] = 1\ .$$

It follows that, for each $x \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathrm{Util}_t(x)\right] - \ln\big(|\mathcal{X}|\big) &= \mathbb{E}\left[\sum_{t=1}^{T}\widehat{g}_t(x)\right] - \ln\big(|\mathcal{X}|\big) = \mathbb{E}\Big[\ln\big(w_{T+1}(x)\big)\Big] - \ln\big(|\mathcal{X}|\big) \\
&\leqslant \mathbb{E}\left[\ln\left(\frac{\|w_{T+1}\|_1}{\|w_1\|_1}\right)\right] = \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}_t\left[\ln\left(\frac{\|w_{t+1}\|_1}{\|w_t\|_1}\right)\right]\right] \\
&\leqslant \frac{\gamma}{1-\gamma}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathrm{Util}_t(B_t)\right] + (e-2)\gamma T\right)\ ,
\end{aligned}$$

which, after rearranging and upper bounding, yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathrm{Util}_t(x) - \sum_{t=1}^{T}\mathrm{Util}_t(B_t)\right] \leqslant \frac{\ln\left(|\mathcal{X}|\right)}{\gamma} + (e-1)\gamma T\ .$$

Selecting γ as in the statement of the theorem leads to the conclusion. □

B.3 Missing Details of the Proof of Theorem 22

Claim 4. *There exists two disjoint intervals I_+ and I_- in $[0,1]$ such that, for any $\varepsilon \in (0, \frac{1}{2})$ and any time t, the following inequalities hold:*

$$\max_{x\in[0,1]} \mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(x)] \geqslant \mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(b)] + \frac{1}{128}\varepsilon,\ \textit{for all } b \notin I_{\pm}$$

Proof. For any $\varepsilon \in (0, \frac{1}{2})$, the distributions $\mathbb{P}^{\pm\varepsilon}$ are such that, the set of all the bids that induce non-negative utility $\mathbb{E}^{\pm\varepsilon}[\mathrm{Util}_t(b)]$ is contained into two disjoint intervals $I_+ = [0, \frac{1}{8}]$ and $I_- = [\frac{1}{4}, 1]$*.

We consider separately the two cases $\mathbb{P}^{+\varepsilon}$ and $\mathbb{P}^{-\varepsilon}$. We start from the former. By simply looking at the definition (3.2), it is clear that $\mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(b)]$ is monotonically increasing in ε for any $b \in I_+$, on the contrary, it is monotonically decreasing for $b \in I_-$. We have the following:

$$\max_{b\in I_-} \mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(b)] \leqslant \max_{b\in I_-} \mathbb{E}^{0}[\mathrm{Util}_t(b)] = \tfrac{1}{128}.$$

On the other hand,

$$\max_{x\in[0,1]} \mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(\widehat{b})] \geqslant \mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(\tfrac{1}{16})] = \tfrac{1}{128}(1+\varepsilon) > \max_{b\in I_-} \mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(b)] + \tfrac{\varepsilon}{128}.$$

We consider now the other case, corresponding to $\mathbb{P}^{-\varepsilon}$. By the definition in Equation (3.2),

*The choice of I_+ and I_- is not tight.

$\mathbb{E}^{-\varepsilon}[\mathrm{Util}_t(b)]$ is monotonically increasing in its first argument for any $b \in I_-$, on the contrary, it is monotonically decreasing for $b \in I_+$. Similarly to the other case we have two steps. On the one hand, it holds that

$$\max_{b \in I_+} \mathbb{E}^{-\varepsilon}[\mathrm{Util}_t(b)] \leqslant \max_{b \in I_+} \mathbb{E}^{0}[\mathrm{Util}_t(b)] = \tfrac{1}{128},$$

while on the other hand it holds that

$$\max_{x \in [0,1]} \mathbb{E}^{-\varepsilon}[\mathrm{Util}_t(x)] \geqslant \mathbb{E}^{-\varepsilon}[\mathrm{Util}_t(\tfrac{7}{16})] = \tfrac{1}{128} + \varepsilon\tfrac{41}{128} > \max_{b \in I_-} \mathbb{E}^{+\varepsilon}[\mathrm{Util}_t(b)] + \tfrac{\varepsilon}{4}.$$

□

We need a preliminary result for the proof of Claim 5. Recall, we use the same random variable (V, M) to denote the highest competing bid/valuation pair drawn from the different probability distribution. When we change the underlying measure, we are changing its law. Consider now the push forward measures on $[0, 1]^2$ (with the Borel σ-algebra) induced by these three measures: $\mathbb{P}^0_{(V,M)}$, $\mathbb{P}^{+\varepsilon}_{(V,M)}$ and $\mathbb{P}^{-\varepsilon}_{(V,M)}$. With some simple calculations (similarly to what is done in, e.g., Appendix B of [169]) it is possible to bound the KL divergence:

Claim 9. *For any $\varepsilon \in (0, \frac{1}{2})$ the following inequality holds true:*

$$\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{+\varepsilon}_{(V,M)}, \mathbb{P}^0_{(V,M)}\right) = \mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{-\varepsilon}_{(V,M)}, \mathbb{P}^0_{(V,M)}\right) \leqslant 2\varepsilon^2$$

Proof. We simply apply the definition of $\mathcal{D}_{\mathrm{KL}}$ divergence for continuous random variables. We only do the calculations for $\mathbb{P}^{+\varepsilon}_{(V,M)}$, the other term is analogous:

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{+\varepsilon}_{(V,M)}, \mathbb{P}^0_{(V,M)}\right) &= \int_{Q_+ \cup Q_-} f^{+\varepsilon}(v,m) \ln \frac{f^{+\varepsilon}(v,m)}{f^0(v,m)} dm\, dv \\ &= \frac{1}{2}(1+\varepsilon)\ln(1+\varepsilon) + \frac{1}{2}(1-\varepsilon)\ln(1-\varepsilon) \leqslant 2\varepsilon^2,\end{aligned}$$

where the last inequality holds for any $\varepsilon \in (0, \frac{1}{2})$. □

Claim 5. *The following inequality hold:*

$$\frac{1}{2}\sum_{i=1,2} \mathbb{E}^i[N_i] \leqslant \frac{3}{4}T.$$

Proof. We have the following:

$$\begin{aligned}\mathbb{E}^i[N_i] - \mathbb{E}^0[N_i] &= \sum_{t=2}^{T} \mathbb{P}^i[B_t \in I_i] - \mathbb{P}^0[B_t \in I_i] \\ &\leqslant \sum_{t=2}^{T} ||\mathbb{P}^i_{(V_1,M_1),\ldots,(V_{t-1},M_{t-1})} - \mathbb{P}^0_{(V_1,M_1),\ldots,(V_{t-1},M_{t-1})}||_{\mathrm{TV}} && \text{(Total variation)} \\ &\leqslant \sum_{t=2}^{T} \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^i_{(V_1,M_1),\ldots,(V_{t-1},M_{t-1})}, \mathbb{P}^0_{(V_1,M_1),\ldots,(V_{t-1},M_{t-1})}\right)} \\ & && \text{(Pinsker's inequality)}\end{aligned}$$

$$\leqslant \sum_{t=2}^{T} \sqrt{\frac{t}{2} \mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{i}_{(V,M)}, \mathbb{P}^{0}_{(V,M)}\right)} \qquad ((V_1, M_1), \ldots, (V_{t-1}, M_{t-1}), \ldots \text{ are i.i.d.})$$

$$\leqslant \frac{1}{4\sqrt{T}} \sum_{t=2}^{T} \sqrt{t} \leqslant \frac{1}{4} T, \tag{B.8}$$

where in the last inequality we applied Claim 9 for our choice of $\varepsilon = 1/(4\sqrt{T})$. Note, $\mathbb{P}^{j}_{(V_1,M_1),\ldots,(M_t,V_t)}$ is the push-forward measure on $([0,1]^2)^t$ induced by t i.i.d. draws of (V, M) from distribution $\mathbb{P}^j$, $j \in \{0, 1, 2\}$. Averaging the result in Equation (B.8), we get the desired inequality:

$$\frac{1}{2} \sum_{i=1,2} \mathbb{E}^{i}\left[N_i\right] \leqslant \frac{1}{2} \sum_{i=1,2} \mathbb{E}^{0}\left[N_i\right] + \frac{T}{4} = \frac{3}{4} T.$$

□

Appendix C

Adaptive Maximization of Social Welfare

C.1 Commodity Taxation

In this appendix, we propose a generalization of our baseline model for optimal taxation to a model for commodity taxation with multiple goods $j \in \{1, \dots, k\}$ and continuous demand functions $y_i(x) \in [0,1]^k$, where $x \in [0,1]^k$ is a vector of tax rates. We again assume that there are no income effects. Our setup is a version of the classic Ramsey model [154].

In the following, we use $\langle x, y \rangle$ to denote the Euclidean inner product between x and y.

Setup At each time $i = 1, 2, \dots, T$, one individual arrives who is characterized by a utility function $u_i : [0,1]^k \to \mathbb{R}$. This individual is exposed to a tax vector $x_i \in [0,1]^k$, and makes a continuous consumption decision y_i. Public revenue is given by $\langle x_i, y_i \rangle$. Private utility is given by $u_i(y_i)$ plus the consumption of a numeraire good, which has a price normalized to 1 and enters utility additively. The individual consumption choice y_i costs $\langle x_i + p, y \rangle$, where p is the (exogenously given) vector of pre-tax prices. This cost of purchasing y_i reduces the consumption of the numeraire good. The optimal individual decision is therefore given by

$$y_i = G_i(x_i) = \operatorname*{arg\,max}_{y \in [0,1]^k} \left[u_i(y) - \langle x_i + p, y \rangle \right]. \tag{C.1}$$

The implied private welfare is

$$v_i(x) = v_0 + \max_{y \in [0,1]^k} \left[u_i(y) - \langle x + p, y \rangle \right],$$

where we have added a constant v_0, chosen such that $v_i(0) = 0$; this is just a normalization to simplify notation below.

We define social welfare as a weighted sum of public revenue and private welfare, with a weight λ for the latter. Social welfare for time period i, as a function of the tax vector x chosen by the learner, is therefore given by

$$U_i(x_i) = \underbrace{\langle x_i, y_i \rangle}_{\text{Public revenue}} + \lambda \cdot \underbrace{v_i(x_i)}_{\text{Private welfare}}. \tag{C.2}$$

After period i, we observe y_i. Nothing else is observed.

The regret can be defined analogously to what we have done in Section 4.2, and the goal of the learner is to obtain (optimal) sublinear regret rates.

We leave the search for a solution to this intriguing problem to future research. Here, we remark that the one-dimensional setting (i.e., when $k = 1$) can be easily solved by an adaptation of Algorithm 7 that leverages the envelope theorem [139, Theorem 2] to relate the feedback received and the derivative of the private welfare. However, how to adapt this algorithm to higher-dimensional problems remains unclear, and deeper insight may be necessary for a satisfying solution.

C.2 Proofs

In this appendix, we present detailed proofs of the theorems we discussed in the main body.

C.2.1 Theorem 26 (Stochastic Lower Bound)

We begin by presenting the proof of the stochastic $T^{2/3}$ lower bound.

Proof of Theorem 26.

Defining a family of distributions for v Recall that, for each $\varepsilon \in [-1, 1]$, the probability distribution μ^ε is defined as the probability measure supported on $(1/4, 1/2, 3/4, 1)$ with masses $\big(a, (1+\varepsilon)\cdot b, (1-\varepsilon)\cdot b, 1-a-2\cdot b\big)$, where

$$a \coloneqq \frac{(1-\lambda)\cdot(136-99\cdot\lambda)}{2\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)}\,, \qquad b \coloneqq \frac{1-\lambda}{2\cdot(24-17\cdot\lambda)}\,.$$

Furthermore, for each $\varepsilon \in [-1, 1]$, recall that $\mathbb{G}^\varepsilon$ and $\mathbb{U}^\varepsilon$ are respectively the demand function and the expected social welfare associated to μ^ε (see Figure 4.1 for an illustration). Let $v_1, v_2, \dots \in [0, 1]$ be the sequence of individual valuations. For each $\varepsilon \in [-1, 1]$, consider a distribution $\mathbb{P}^\varepsilon$ such that the individual valuations $v_1, v_2, \dots$ form a $\mathbb{P}^\varepsilon$-i.i.d. sequence (independent of the randomization used by the algorithm) with common distribution μ^ε.

Explicit lower bound on regret that will be proven Define

$$c_1 \coloneqq \frac{\lambda}{4}\cdot b\,, \quad c_2 \coloneqq \frac{1}{8}\cdot\frac{1-\lambda}{4-3\cdot\lambda}\,, \quad c_3 \coloneqq b\cdot\sqrt{\frac{2}{a\cdot(1-a-2\cdot b)}}\,.$$

We will prove that, for any randomized algorithm α and any time horizon $T \in \mathbb{N}$, there exists $\varepsilon \in [-1, 1]$ such that

$$R_T(\alpha, \mathbb{G}^\varepsilon) \geqslant C\cdot T^{2/3}\,,$$

where

$$\begin{aligned} C &\coloneqq \min\left(\frac{c_1^2\cdot c_3^2}{c_2},\ \frac{c_2}{2},\ \frac{1}{16}\cdot\sqrt[3]{\frac{c_1^2\cdot c_2}{c_3^2}}\right) \\ &= \min\left(\frac{\lambda^2\cdot(4-3\cdot\lambda)^3}{8\cdot(136-99\cdot\lambda)\cdot(26-19\cdot\lambda)},\ \frac{\lambda^{2/3}\cdot(1-\lambda)^{4/3}\cdot(136-99\cdot\lambda)^{1/3}\cdot(26-19\cdot\lambda)^{1/3}}{128\cdot(4-3\cdot\lambda)\cdot(24-17\cdot\lambda)^{4/3}}\right) > 0 \end{aligned} \tag{C.3}$$

Fix a randomized algorithm α to choose the policies $x_1, x_2, \dots$, and fix a time horizon $T \in \mathbb{N}$.

Number of mistakes and lower bound on regret We need to count the random number of times the algorithm has played in the regions $(1/2, 3/4]$, $[0, 1/2]$ and $(3/4, 1]$ up to time T. This can be done relying on the following random variables:

$$n_1 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{(1/2,3/4]}(x_i) \ , \qquad n_2 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{[0,1/2]}(x_i) \ , \qquad n_3 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{(3/4,1]}(x_i) \ .$$

Notice that since the intervals $(1/2, 3/4], [0, 1/2]$ and $(3/4, 1]$ form a partition of $[0, 1]$, we have that

$$n_1 + n_2 + n_3 = T \tag{C.4}$$

For each $\varepsilon \in [-1, 1]$, denote by $\mathbb{E}^\varepsilon$ the expectation taken with respect to the distribution $\mathbb{P}^\varepsilon$. Notice that, for each $\varepsilon \in [-1, 1]$, the expected regret when the underlying distribution is $\mathbb{P}^\varepsilon$ equals

$$R_T(\alpha, \mathbb{G}^\varepsilon) = T \cdot \sup_{x \in [0,1]} \mathbb{U}^\varepsilon(x) - \sum_{i=1}^{T} \mathbb{E}^\varepsilon \left[\mathbb{U}^\varepsilon(x_i) \right] \ . \tag{C.5}$$

Algebraic calculations show that, for each $\varepsilon \in [-1, 1]$

$$\max_{x \in (1/2,3/4]} \mathbb{U}^\varepsilon(x) = \mathbb{U}^\varepsilon(3/4) \ , \quad \max_{x \in [0,1/2]} \mathbb{U}^\varepsilon(x) = \mathbb{U}^\varepsilon(1/4) \ , \quad \max_{x \in (3/4,1]} \mathbb{U}^\varepsilon(x) = \mathbb{U}^\varepsilon(1) \ , \tag{C.6}$$

$$\text{and} \qquad \mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(1/4) = c_1 \cdot \varepsilon \ . \tag{C.7}$$

Further calculations show also that

$$\min_{\varepsilon \in [-1,1]} \min\big(\mathbb{U}^\varepsilon(1/4), \mathbb{U}^\varepsilon(1)\big) = \mathbb{U}^1(1/4) \ , \quad \max_{\varepsilon \in [-1,1]} \max_{x \in (1/2,3/4]} \mathbb{U}^\varepsilon(x) = \mathbb{U}^{-1}(3/4) \ , \tag{C.8}$$

$$\text{and} \qquad \mathbb{U}^1(1/4) - \mathbb{U}^{-1}(3/4) = c_2 \ . \tag{C.9}$$

Equations (C.6), (C.7), (C.8), and (C.9) imply that

$$\sup_{x \in [0,1]} \mathbb{U}^\varepsilon(x) = \mathbb{U}^\varepsilon(1) \ , \qquad \text{if } \varepsilon \in [0, 1] \ . \tag{C.10}$$

It follows that, if $\varepsilon \in [0, 1]$,

$$\begin{aligned}
R_T(\alpha, \mathbb{G}^\varepsilon) &\overset{(C.5)}{=} T \cdot \sup_{x \in [0,1]} \mathbb{U}^\varepsilon(x) - \sum_{i=1}^{T} \mathbb{E}^\varepsilon \left[\mathbb{U}^\varepsilon(x_i) \right] \\
&\overset{(C.10)}{=} T \cdot \mathbb{U}^\varepsilon(1) - \sum_{i=1}^{T} \mathbb{E}^\varepsilon \left[\mathbb{U}^\varepsilon(x_i) \cdot \big(\mathbb{I}_{(1/2,3/4]}(x_i) + \mathbb{I}_{[0,1/2]}(x_i) + \mathbb{I}_{(3/4,1]}(x_i) \big) \right] \\
&\overset{(C.6)}{\geqslant} T \cdot \mathbb{U}^\varepsilon(1) - \sum_{i=1}^{T} \mathbb{E}^\varepsilon \left[\mathbb{U}^\varepsilon(3/4) \cdot \mathbb{I}_{(1/2,3/4]}(x_i) + \mathbb{U}^\varepsilon(1/2) \cdot \mathbb{I}_{[0,1/2]}(x_i) + \mathbb{U}^\varepsilon(1) \cdot \mathbb{I}_{(3/4,1]}(x_i) \right] \\
&\overset{(C.4)}{=} \big(\mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(3/4) \big) \cdot \mathbb{E}^\varepsilon[n_1] + \big(\mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(1/4) \big) \cdot \mathbb{E}^\varepsilon[n_2] \\
&\overset{(C.8)}{\geqslant} \big(\mathbb{U}^1(1/4) - \mathbb{U}^{-1}(3/4) \big) \cdot \mathbb{E}^\varepsilon[n_1] + \big(\mathbb{U}^\varepsilon(1) - \mathbb{U}^\varepsilon(1/4) \big) \cdot \mathbb{E}^\varepsilon[n_2]
\end{aligned}$$

$$\overset{(C.9)}{=} c_2 \cdot \mathbb{E}^{\varepsilon}[n_1] + \left(\mathbb{U}^{\varepsilon}(1) - \mathbb{U}^{\varepsilon}(1/4)\right) \cdot \mathbb{E}^{\varepsilon}[n_2]$$
$$\overset{(C.7)}{=} c_2 \cdot \mathbb{E}^{\varepsilon}[n_1] + c_1 \cdot \varepsilon \cdot \mathbb{E}^{\varepsilon}[n_2] \tag{C.11}$$

Notice that inequality (C.11) quantifies how much regret the algorithm is going to suffer in terms of the expected number of times it plays in the wrong regions, when the demand function is $\mathbb{G}^{\varepsilon}$ and $\varepsilon > 0$.

In the same way inequality (C.11) was proven, we can prove that, if $\varepsilon \in [0, 1]$,

$$R_T(\alpha, \mathbb{G}^{-\varepsilon}) \geqslant c_2 \cdot \mathbb{E}^{-\varepsilon}[n_1] + c_1 \cdot \varepsilon \cdot \mathbb{E}^{-\varepsilon}[n_3] \geqslant c_1 \cdot \varepsilon \cdot \mathbb{E}^{-\varepsilon}[n_3] \,, \tag{C.12}$$

which again quantifies how much regret the algorithm is going to suffer in terms of the expected number of times it plays in the wrong regions, when the demand function is $\mathbb{G}^{-\varepsilon}$ and $\varepsilon > 0$.

Intuition for the remainder of the proof At high level, inequalities (C.11) and (C.12) tell us that, if $|\varepsilon|$ is not negligible, the algorithm has to play a substantially different number of times in the region $(3/4, 1]$, depending on the sign of ε, not to suffer significant regret when the demand function is $\mathbb{G}^{\varepsilon}$. The crucial idea is that the only way for the algorithm to present this different behavior is by playing in the only informative region about the sign of ε, i.e., the region $(1/2, 3/4]$. However, as shown in (C.11), selecting policies in this region comes at a cost in terms of regret. To relate quantitatively the number of times the algorithm has to play in this costly region with the difference in the expected number of times the algorithm selects policies in the region $(3/4, 1]$ is the last missing ingredient that we can obtain relying on information theoretic techniques. It can be proved (and a formal proof is provided at the end of the current proof) that, for each $\varepsilon \in [0, 1]$,

$$\mathbb{E}^{-\varepsilon}[n_3] \geqslant \mathbb{E}^{\varepsilon}[n_3] - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\mathbb{E}^{\varepsilon}[n_1]} \,. \tag{C.13}$$

Now, if the algorithm suffers low regret when $\varepsilon > 0$, then by (C.11) we have an upper bound on the number of times the algorithm plays in the region $(1/2, 3/4]$ and a lower bound on the number of times it plays in the region $(3/4, 1]$, whenever $\varepsilon > 0$. In turn, by (C.13), this gives a lower bound on the number of times the algorithm plays in the sub-optimal region $(3/4, 1]$ when $\varepsilon < 0$. Then, relying on (C.12), we have an explicit lower bound on how much regret the algorithm is going to suffer when $\varepsilon < 0$. We will now carry out this plan —and prove the theorem— as follows.

Low regret cannot be achieved for both positive and negative ε To get a contradiction, suppose that

$$\forall \varepsilon \in [-1, 1] \qquad R_T(\alpha, \mathbb{G}^{\varepsilon}) < C \cdot T^{2/3} \,. \tag{C.14}$$

It follows from (C.11) that, for each $\varepsilon \in [0, 1]$,

$$\mathbb{E}^{\varepsilon}[n_1] \overset{(C.11)}{\leqslant} \frac{R_T(\alpha, \mathbb{G}^{\varepsilon})}{c_2} \overset{(C.14)}{\leqslant} \frac{C}{c_2} \cdot T^{2/3} \,, \qquad \mathbb{E}^{\varepsilon}[n_2] \overset{(C.11)}{\leqslant} \frac{R_T(\alpha, \mathbb{G}^{\varepsilon})}{c_1 \cdot \varepsilon} \overset{(C.14)}{\leqslant} \frac{C}{c_1 \cdot \varepsilon} \cdot T^{2/3} \,. \tag{C.15}$$

This implies, relying also on (C.12) and (C.13), that for each $\varepsilon \in [0, 1]$ we have

$$R_T(\alpha, \mathbb{G}^{-\varepsilon}) \overset{(C.12)}{\geqslant} c_1 \cdot \varepsilon \cdot \mathbb{E}^{-\varepsilon}[n_3] \overset{(C.13)}{\geqslant} c_1 \cdot \varepsilon \cdot \left(\mathbb{E}^{\varepsilon}[n_3] - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\mathbb{E}^{\varepsilon}[n_1]}\right)$$

$$\begin{aligned}
&\overset{(C.4)}{=} c_1 \cdot \varepsilon \cdot \left(T - \mathbb{E}^{\varepsilon}[n_1] - \mathbb{E}^{\varepsilon}[n_2] - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\mathbb{E}^{\varepsilon}[n_1]}\right) \\
&\overset{(C.15)}{\geqslant} c_1 \cdot \varepsilon \cdot \left(T - \frac{C}{c_2} \cdot T^{2/3} - \frac{C}{c_1 \cdot \varepsilon} \cdot T^{2/3} - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\frac{C}{c_2} \cdot T^{2/3}}\right) \\
&= c_1 \cdot \varepsilon \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - \frac{C}{c_1 \cdot \varepsilon} \cdot T^{-1/3} - c_3 \cdot \varepsilon \cdot T^{1/3} \cdot \sqrt{\frac{C}{c_2}}\right) \cdot T \,. \qquad \text{(C.16)}
\end{aligned}$$

Pick $\varepsilon \coloneqq T^{-1/3} \cdot \sqrt{\frac{\sqrt{C \cdot c_2}}{c_1 \cdot c_3}}$. First, note that since $0 < C \overset{(C.3)}{\leqslant} \frac{c_1^2 \cdot c_3^2}{c_2}$ we have that $\varepsilon \in (0, 1]$. Plugging this value of ε in (C.16) leads to

$$\begin{aligned}
C \cdot T^{2/3} &\overset{(C.14)}{>} R_T(\alpha, \mathbb{G}^{-\varepsilon}) \\
&\overset{(C.16)}{\geqslant} \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - \frac{C}{c_2} \cdot T^{-1/3} - 2 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(C.3)}{\geqslant} \frac{1}{2} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot \left(1 - 4 \cdot \sqrt{\frac{c_3}{c_1 \cdot \sqrt{c_2}}} \cdot C^{3/4}\right) \cdot T^{2/3} \\
&\overset{(C.3)}{\geqslant} \frac{1}{4} \cdot \sqrt{\frac{\sqrt{C \cdot c_2} \cdot c_1}{c_3}} \cdot T^{2/3} \,, \qquad \text{(C.17)}
\end{aligned}$$

where the second to last inequality follows from $C \leqslant \frac{c_2}{2}$, while the last inequality follows from $C \leqslant \frac{1}{16} \sqrt[3]{\frac{c_1^2 \cdot c_2}{c_3^2}}$. Rearranging inequality (C.17) leads to the contradiction

$$C \overset{(C.17)}{>} \left(\frac{1}{4} \cdot \sqrt{\frac{c_1 \cdot \sqrt{c_2}}{c_3}}\right)^{4/3} = \frac{1}{8} \cdot \sqrt[3]{\frac{2 \cdot c_1^2 \cdot c_2}{c_3^2}} > \frac{1}{16} \cdot \sqrt[3]{\frac{c_1^2 \cdot c_2}{c_3^2}} \overset{(C.3)}{\geqslant} C \,.$$

Since (C.14) leads to a contradiction, it follows that there exists $\varepsilon \in [-1, 1]$ such that $R_T(\alpha, \mathbb{G}^{\varepsilon}) \geqslant C \cdot T^{2/3}$. Given that the time horizon T and the randomized algorithm were arbitrarily fixed, the theorem is proved. □

Claim (C.13) **(Relating choice probabilities for positive and negative ε)**

Proof of Claim (C.13).
Let $w_1, w_2, \cdots \in [0, 1]$ be the randomization seeds to be used by the algorithm. In the light of the Skorokhod representation theorem [185, Section 17.3], we may assume without (much) loss of generality that, for each $\varepsilon \in [-1, 1]$, these seeds form a sequence of $\mathbb{P}^{\varepsilon}$-i.i.d. $[0, 1]$-valued uniform random variables. In particular, this implies,

$$\mathbb{P}^{\varepsilon}_{(w_i)_{i \in \mathbb{N}}} = \mathbb{P}^{-\varepsilon}_{(w_i)_{i \in \mathbb{N}}} \,, \qquad \forall \varepsilon \in [0, 1] \,. \qquad \text{(C.18)}$$

Recall that a sequence of functions $\alpha \coloneqq (\alpha_i)_{i \in \mathbb{N}}$ is called a randomized algorithm if

$$\alpha_1 \colon [0, 1] \to [0, 1] \,, \qquad \forall i \in \mathbb{N}, \quad \alpha_{i+1} \colon [0, 1]^{i+1} \times \{0, 1\}^i \to [0, 1] \,.$$

The feedback function associated to our problem is

$$\varphi\colon [0,1] \times \{1/4, 1/2, 3/4, 1\} \to \{0,1\}\ , \qquad (x,v) \mapsto \mathbb{I}\{x \leqslant v\}\ .$$

Now, a randomized algorithm α generates a sequence of choices $x_1, x_2, \dots$ using the randomization seeds $w_1, w_2, \dots$ and the received feedback $z_1, z_2, \dots \in \{0,1\}$ in the following inductive way on $i \in \mathbb{N}$

$$\begin{aligned} x_1 &\coloneqq \alpha_1(w_1)\ , & z_1 &\coloneqq \varphi(x_1, v_1)\ , \\ x_{i+1} &\coloneqq \alpha_{i+1}(w_1, \dots, w_{i+1}, z_1, \dots, z_i)\ , & z_{i+1} &\coloneqq \varphi(x_{i+1}, v_{i+1})\ . \end{aligned}$$

For each $a \in [0,1]$, fix a binary representation $0.a_1a_2a_3\dots$ and define $\xi(a) \coloneqq 0.a_1a_3a_5\dots$ and $\zeta(a) \coloneqq 0.a_2a_4a_6\dots$. Notice that $\xi, \zeta\colon [0,1] \to [0,1]$ are independent with respect to the Lebesgue measure on $[0,1]$ and that their (common) distribution is a uniform on $[0,1]$. For each $x \in [0,1]$, define $\psi_x\colon [0,1] \to \{0,1\}, u \mapsto \mathbb{I}_{[0,1/4]}(x) + \mathbb{I}_{(1/4,1/2]}(x) \cdot \mathbb{I}_{[0,1-a]}(u) + \mathbb{I}_{(3/4,1]}(x) \cdot \mathbb{I}_{[0,1-a-2\cdot b]}(u)$. Define by induction on $i \in \mathbb{N}$ the following process

$$\begin{aligned} \widetilde{x}_1 &\coloneqq \alpha_1\big(\zeta(w_1)\big)\ , \\ \widetilde{z}_1 &\coloneqq \varphi\Big(\widetilde{x}_1, \psi_{\widetilde{x}_1}\big(\xi(w_1)\big)\Big)\ , \\ \widetilde{x}_{i+1} &\coloneqq \alpha_{i+1}\big(\zeta(w_1), \dots, \zeta(w_{i+1}), \widetilde{z}_1, \dots, \widetilde{z}_i\big), \\ \widetilde{z}_{i+1} &\coloneqq \begin{cases} \varphi(\widetilde{x}_{i+1}, v_{i+1}), & \widetilde{x}_{i+1} \in (1/2, 3/4] \\ \varphi\Big(\widetilde{x}_{i+1}, \psi_{\widetilde{x}_{i+1}}\big(\xi(w_{i+1})\big)\Big), & \text{otherwise.} \end{cases} \end{aligned}$$

Since, for each $\varepsilon \in [-1,1]$ and each $i \in \mathbb{N}$,

$$\mathbb{P}^\varepsilon[z_i = 1 \mid x_i] = \begin{cases} 1 & x_i \in [0, \frac{1}{4}] \\ 1-a & x_i \in (\frac{1}{4}, \frac{1}{2}] \\ 1-a-(1+\varepsilon)\cdot b & x_i \in (\frac{1}{2}, \frac{3}{4}] \\ 1-a-2\cdot b & x_i \in (\frac{3}{4}, 1] \end{cases}\ ,$$

$$\mathbb{P}^\varepsilon[\widetilde{z}_i = 1 \mid \widetilde{x}_i] = \begin{cases} 1 & \widetilde{x}_i \in [0, \frac{1}{4}] \\ 1-a & \widetilde{x}_i \in (\frac{1}{4}, \frac{1}{2}] \\ 1-a-(1+\varepsilon)\cdot b & \widetilde{x}_i \in (\frac{1}{2}, \frac{3}{4}] \\ 1-a-2\cdot b & \widetilde{x}_i \in (\frac{3}{4}, 1] \end{cases}$$

it follows that, for each $\varepsilon \in [-1,1]$ and each $i \in \mathbb{N}$, the random variable $\widetilde{x}_i$ has the same distribution as the random choice x_i made by the randomized algorithm α at time i when the underlying distribution is $\mathbb{P}^\varepsilon$, i.e.,

$$\mathbb{P}^\varepsilon_{\widetilde{x}_i} = \mathbb{P}^\varepsilon_{x_i}\ . \tag{C.19}$$

As with the process $x_1, x_2, \dots$, we have to count the number of times the process $\widetilde{x}_1, \widetilde{x}_2, \dots$ lands in the regions $(1/2, 3/4]$, $[0, 1/2]$ and $(3/4, 1]$ up to the time T. This can be done relying on the following

random variables

$$\widetilde{n}_1 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{(1/2,3/4]}(\widetilde{x}_i) \ , \qquad \widetilde{n}_2 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{[0,1/2]}(\widetilde{x}_i) \ , \qquad \widetilde{n}_3 \coloneqq \sum_{i=1}^{T} \mathbb{I}_{(3/4,1]}(\widetilde{x}_i) \ .$$

Since, for each $\varepsilon \in [-1,1]$ and each $j \in \{1,2,3\}$,

$$\mathbb{E}^{\varepsilon}\left[\widetilde{n}_j\right] = \sum_{i=1}^{T} \mathbb{P}^{\varepsilon}_{x_i}\left[(1/2, 3/4]\right] \overset{(C.19)}{=} \sum_{i=1}^{T} \mathbb{P}^{\varepsilon}_{\widetilde{x}_i}\left[(1/2, 3/4]\right] = \mathbb{E}^{\varepsilon}\left[n_j\right] \ ,$$

to prove the claim (C.13), it is enough to prove that, for each $\varepsilon \in [-1,1]$,

$$\mathbb{E}^{-\varepsilon}\left[\widetilde{n}_3\right] \geqslant \mathbb{E}^{\varepsilon}\left[\widetilde{n}_3\right] - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\mathbb{E}^{\varepsilon}\left[\widetilde{n}_1\right]} \ .$$

We first prove the result when the sequence of randomization seeds is fixed, i.e., we suppose first that $\bar{w}_1, \bar{w}_2, \ldots$ are such that $w_1 = \bar{w}_1, w_2 = \bar{w}_2, \ldots$. For each $\varepsilon \in [-1,1]$, we consider the associated probability distribution $\mathbb{Q}^{\varepsilon}$, defined as the conditional probability distribution $\mathbb{P}^{\varepsilon}[\cdot \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \ldots]$. For each $t \in \mathbb{N}$, let $I_t \coloneqq \left\{i \in \{1,\ldots,t\} \mid \widetilde{x}_i \in (1/2, 3/4]\right\}$, and for each $s \in \{1,\ldots,t\}$, let

$$Z_{t,s} \coloneqq \begin{cases} \varnothing & \text{if } s \notin I_t \ , \\ \mathbb{I}\left\{1/2 < v_s\right\} & \text{if } s \in I_t \ . \end{cases}$$

Notice that for each $t_1, t_2 \in \mathbb{N}$ and each $s \in \{1,\ldots,\min(t_1,t_2)\}$, we have that $Z_{t_1,s} = Z_{t_2,s}$. Then, for each $s \in \mathbb{N}$, it is well defined the random variable $Z_s \coloneqq Z_{t,s}$, where $t \in \mathbb{N}$ is any number $t \geqslant s$. Define, for each $t \in \mathbb{N}$, the random vector $\bar{Z}_t \coloneqq (Z_1, \ldots, Z_t)$. Notice that, given that the sequence of randomization seeds is fixed and that, for each $s \in \mathbb{N}$, we have that $v_s \in \{1/4, 1/2, 3/4, 1\}$ (hence, for each $x \in (1/2, 3/4]$, it holds that $\mathbb{I}\{1/2 < v_s\} = \mathbb{I}\{x = v_s\}$), the random vector $(\widetilde{x}_1, \ldots, \widetilde{x}_T)$ is measurable with respect to the σ-algebra generated by $\bar{Z}_{T-1}$. Hence, for each $\varepsilon \in [0,1]$ and each $i \in \{1,\ldots,T\}$, we can deduce from Pinsker's inequality (see, e.g., [175, Lemma 2.5]) that

$$\mathbb{Q}^{\varepsilon}\left[\widetilde{x}_i \in (3/4, 1]\right] \leqslant \mathbb{Q}^{-\varepsilon}\left[\widetilde{x}_i \in (3/4, 1]\right] + \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{\bar{Z}_{T-1}} \mid\mid \mathbb{Q}^{-\varepsilon}_{\bar{Z}_{T-1}}\right)} \ , \tag{C.20}$$

where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence. Now, for each $t \in \mathbb{N}$ and each $\varepsilon \in [0,1]$, by the chain rule for Kullback-Leibler divergence (see, e.g., [71, Theorem 2.5.3]), we have

$$\begin{aligned} &\mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{\bar{Z}_{t+1}} \mid\mid \mathbb{Q}^{-\varepsilon}_{\bar{Z}_{t+1}}\right) = \mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{(\bar{Z}_t, Z_{t+1})} \mid\mid \mathbb{Q}^{-\varepsilon}_{(\bar{Z}_t, Z_{t+1})}\right) \\ &= \mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{\bar{Z}_t} \mid\mid \mathbb{Q}^{-\varepsilon}_{\bar{Z}_t}\right) + \sum_{(\bar{z},z) \in \{\varnothing,0,1\}^t \times \{\varnothing,0,1\}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[Z_{t+1} = z \mid \bar{Z}_t = \bar{z}\right]}{\mathbb{Q}^{-\varepsilon}\left[Z_{t+1} = z \mid \bar{Z}_t = \bar{z}\right]}\right) \cdot \mathbb{Q}^{\varepsilon}\left[\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\right] . \end{aligned} \tag{C.21}$$

Notice that, for each $t \in \mathbb{N}$ and each $\varepsilon \in [0,1]$ we have

$$\sum_{(\bar{z},z) \in \{\varnothing,0,1\}^t \times \{\varnothing,0,1\}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[Z_{t+1} = z \mid \bar{Z}_t = \bar{z}\right]}{\mathbb{Q}^{-\varepsilon}\left[Z_{t+1} = z \mid \bar{Z}_t = \bar{z}\right]}\right) \cdot \mathbb{Q}^{\varepsilon}\left[\bar{Z}_t = \bar{z} \cap Z_{t+1} = z\right]$$

$$
\begin{aligned}
&= \sum_{\substack{(\bar{z},z)\in\{\varnothing,0,1\}^t\times\{\varnothing,0,1\}\\ t+1\in I_{t+1}}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[Z_{t+1}=z \mid \bar{Z}_t=\bar{z}\right]}{\mathbb{Q}^{-\varepsilon}\left[Z_{t+1}=z \mid \bar{Z}_t=\bar{z}\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\bar{Z}_t=\bar{z}\cap Z_{t+1}=z\right]\\
&= \left(\sum_{\substack{\bar{z}\in\{\varnothing,0,1\}^t\\ t+1\in I_{t+1}}} \mathbb{Q}^{\varepsilon}\left[\bar{Z}_t=\bar{z}\right]\right)\cdot \sum_{z\in\{0,1\}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}{\mathbb{Q}^{-\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]\\
&= \mathbb{Q}^{\varepsilon}\left[\widetilde{x}_{t+1}\in(1/2,3/4]\right]\cdot \sum_{z\in\{0,1\}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}{\mathbb{Q}^{-\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]\ . \qquad \text{(C.22)}
\end{aligned}
$$

Algebraic calculations show that, for each $t\in\mathbb{N}$ and each $\varepsilon\in[0,1]$,

$$
\begin{aligned}
&\sum_{z\in\{0,1\}} \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}{\mathbb{Q}^{-\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\mathbb{I}\left\{1/2<v_{t+1}\right\}=z\right]\\
&= \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[\frac{1}{2}<v_{t+1}\right]}{\mathbb{Q}^{-\varepsilon}\left[\frac{1}{2}<v_{t+1}\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\frac{1}{2}<v_{t+1}\right] + \log\left(\frac{\mathbb{Q}^{\varepsilon}\left[\frac{1}{2}\geqslant v_{t+1}\right]}{\mathbb{Q}^{-\varepsilon}\left[\frac{1}{2}\geqslant v_{t+1}\right]}\right)\cdot \mathbb{Q}^{\varepsilon}\left[\frac{1}{2}\geqslant v_{t+1}\right]\\
&= \log\left(\frac{1-a-(1+\varepsilon)\cdot b}{1-a-(1-\varepsilon)\cdot b}\right)\cdot\left(1-a-(1+\varepsilon)\cdot b\right) + \log\left(\frac{a+(1+\varepsilon)\cdot b}{a+(1-\varepsilon)\cdot b}\right)\cdot\left(a+(1+\varepsilon)\cdot b\right)\\
&\leqslant \frac{4\cdot b^2\cdot\varepsilon^2}{\left(1-a-(1-\varepsilon)\cdot b\right)\cdot\left(a+(1-\varepsilon)\cdot b\right)} \leqslant \frac{4\cdot b^2\cdot\varepsilon^2}{a\cdot(1-a-2b)} = 2\cdot c_3^2\cdot\varepsilon^2\ . \qquad \text{(C.23)}
\end{aligned}
$$

Putting (C.21), (C.22) and (C.23) together, we obtain that, for each $t\in\mathbb{N}$ and each $\varepsilon\in[0,1]$,

$$
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{\bar{Z}_{t+1}} \,||\, \mathbb{Q}^{-\varepsilon}_{\bar{Z}_{t+1}}\right) \leqslant \mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{Z_1} \,||\, \mathbb{Q}^{-\varepsilon}_{Z_1}\right) + 2\cdot c_3^2\cdot\varepsilon^2\cdot\sum_{s=1}^{t}\mathbb{Q}^{\varepsilon}\left[\widetilde{x}_{s+1}\in(1/2,3/4]\right]\ . \qquad \text{(C.24)}
$$

With the same technique used above, for each $\varepsilon\in[0,1]$, we can prove that

$$
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{Z_1} \,||\, \mathbb{Q}^{-\varepsilon}_{Z_1}\right) \leqslant 2\cdot c_3^2\cdot\varepsilon^2\cdot\mathbb{Q}^{\varepsilon}\left[\widetilde{x}_1\in(1/2,3/4]\right]\ . \qquad \text{(C.25)}
$$

For each $t\in\{1,\dots,T\}$, putting (C.24) and (C.25) together, we obtain

$$
\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\left(\mathbb{Q}^{\varepsilon}_{\bar{Z}_t} \,||\, \mathbb{Q}^{-\varepsilon}_{\bar{Z}_t}\right) &\overset{\text{(C.24)+(C.25)}}{\leqslant} 2\cdot c_3^2\cdot\varepsilon^2\cdot\sum_{s=1}^{t}\mathbb{Q}^{\varepsilon}\left[\widetilde{x}_s\in(1/2,3/4]\right]\\
&\leqslant 2\cdot c_3^2\cdot\varepsilon^2\cdot\mathbb{E}^{\varepsilon}\left[\widetilde{n}_1 \mid w_1=\bar{w}_1, w_2=\bar{w}_2,\dots\right]\ . \qquad \text{(C.26)}
\end{aligned}
$$

Now, (C.20) and (C.26) imply that, for each $\varepsilon\in[0,1]$ and each $i\in\{1,\dots,T\}$,

$$
\mathbb{Q}^{\varepsilon}\left[\widetilde{x}_i\in(3/4,1]\right] \leqslant \mathbb{Q}^{-\varepsilon}\left[\widetilde{x}_i\in(3/4,1]\right] + c_3\cdot\varepsilon\cdot\sqrt{\mathbb{E}^{\varepsilon}\left[\widetilde{n}_1 \mid w_1=\bar{w}_1, w_2=\bar{w}_2,\dots\right]}\ . \qquad \text{(C.27)}
$$

Taking the sum of (C.27) over $i\in\{1,\dots,T\}$, we obtain that for each $\varepsilon\in[0,1]$,

$$
\begin{aligned}
&\mathbb{E}^{-\varepsilon}\left[\widetilde{n}_3 \mid w_1=\bar{w}_1, w_2=\bar{w}_2,\dots\right]\\
&\qquad \geqslant \mathbb{E}^{\varepsilon}\left[\widetilde{n}_3 \mid w_1=\bar{w}_1, w_2=\bar{w}_2,\dots\right] - c_3\cdot\varepsilon\cdot T\cdot\sqrt{\mathbb{E}^{\varepsilon}\left[\widetilde{n}_1 \mid w_1=\bar{w}_1, w_2=\bar{w}_2,\dots\right]}\ . \qquad \text{(C.28)}
\end{aligned}
$$

Now, since the sequence $\bar{w}_1, \bar{w}_2, \dots$ of randomization seeds has been arbitrarily chosen, for each $\varepsilon \in [0,1]$, using the fact that $\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}} = \mathbb{P}^{-\varepsilon}_{(w_t)_{t\in\mathbb{N}}}$ and Jensen's inequality, we have that

$$\begin{aligned}
\mathbb{E}^{-\varepsilon}\left[\widetilde{n}_3\right] &= \int_{[0,1]^{\mathbb{N}}} \mathbb{E}^{-\varepsilon}\big[\widetilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big] \mathrm{d}\mathbb{P}^{-\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(C.18)}{=} \int_{[0,1]^{\mathbb{N}}} \mathbb{E}^{-\varepsilon}\big[\widetilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big] \mathrm{d}\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\overset{(C.28)}{\geqslant} \int_{[0,1]^{\mathbb{N}}} \mathbb{E}^{\varepsilon}\big[\widetilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big] \mathrm{d}\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \varepsilon \cdot T \cdot \int_{[0,1]^{\mathbb{N}}} \sqrt{\mathbb{E}^{\varepsilon}\big[\widetilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big]} \mathrm{d}\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
\text{(by Jensen)} \qquad &\geqslant \int_{[0,1]^{\mathbb{N}}} \mathbb{E}^{\varepsilon}\big[\widetilde{n}_3 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big] \mathrm{d}\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots) \\
&\quad - c_3 \cdot \varepsilon \cdot T \cdot \sqrt{\int_{[0,1]^{\mathbb{N}}} \mathbb{E}^{\varepsilon}\big[\widetilde{n}_1 \mid w_1 = \bar{w}_1, w_2 = \bar{w}_2, \dots\big] \mathrm{d}\mathbb{P}^{\varepsilon}_{(w_t)_{t\in\mathbb{N}}}(\bar{w}_1, \bar{w}_2, \dots)} \\
&= \mathbb{E}^{\varepsilon}\big[\widetilde{n}_3\big] - c_3 \cdot \varepsilon \cdot \sqrt{\mathbb{E}^{\varepsilon}\big[\widetilde{n}_1\big]}\,.
\end{aligned}$$

□

C.2.2 Theorem 27 (Adversarial Upper Bound)

The proof of this theorem builds upon the proof of Theorem 6.5 in [48]. Relative to this theorem, we need to additionally consider the discretization error introduced by Algorithm 7, and explicitly control the variance of estimated welfare.

Proof of Theorem 27.
Recall our notation $\boldsymbol{U}$ and $\boldsymbol{U}(x)$ for realized cumulative welfare, and for cumulative welfare for the counterfactual, fixed policy x. We further abbreviate $\boldsymbol{U}_{Tk} = \boldsymbol{U}(\widetilde{x}_k)$. Throughout this proof, the sequence $\{v_i\}_{i=1}^T$ is given and conditioned on in any expectations.

1. **Discretization**
   Recall that $U_i(x) = x \cdot \mathbb{I}\{x \leqslant v_i\} + \lambda \cdot \max(v_i - x, 0)$. Let

   $$\widetilde{v}_i \coloneqq \max_k \{\widetilde{x}_k : \widetilde{x}_k \leqslant v_i\}$$

   (this is v_i rounded down to the next gridpoint $\widetilde{x}_k$), and denote

   $$\begin{aligned}
   \widetilde{U}_i(x) &\coloneqq x \cdot \mathbb{I}\{x \leqslant v_i\} + \lambda \cdot \max(\widetilde{v}_i - x, 0), \\
   \widetilde{\boldsymbol{U}}_i(x) &\coloneqq \sum_{j \leqslant i} \widetilde{U}_j(x),
   \end{aligned}$$

   as well as $\widetilde{\boldsymbol{U}}_{ik} \coloneqq \widetilde{\boldsymbol{U}}_i(\widetilde{x}_k)$. Then it is immediate that $\widetilde{U}_i(x) \leqslant U_i(x)$,

   $$\sup_x |\widetilde{U}_i(x) - U_i(x)| \leqslant \frac{\lambda}{K},$$

and $\arg\max_x \widetilde{\boldsymbol{U}}_i(x) \in \{\widetilde{x}_1, \ldots, x_{K+1}\}$, and therefore

$$\max_k \widetilde{\boldsymbol{U}}_{ik} \geqslant \sup_x \boldsymbol{U}_i(x) - i \cdot \frac{\lambda}{K}$$

2. **Unbiasedness**
   At the end of period i, $\widehat{G}_k$ is an unbiased estimator of $\sum_{j \leqslant i} \mathbb{I}\{\widetilde{x}_k \leqslant v_j\}$ for all k. Therefore, $\mathbb{E}\left[\widehat{\boldsymbol{U}}_{ik}\right] = \widetilde{\boldsymbol{U}}_{ik}$ for all i and k.

3. **Upper bound on optimal welfare**
   Define $W_i \coloneqq \sum_k \exp(\eta \cdot \widehat{\boldsymbol{U}}_{ik})$, and $q_{ik} \coloneqq \exp(\eta \cdot \widehat{\boldsymbol{U}}_{ik})/W_i$.

   It is immediate that,

$$\mathbb{E}[\log W_T] \geqslant \eta \cdot \mathbb{E}[\max_k \widehat{\boldsymbol{U}}_{Tk}] \geqslant \eta \cdot \max_k \mathbb{E}[\widehat{\boldsymbol{U}}_{Tk}] = \eta \cdot \max_k \widetilde{\boldsymbol{U}}_{Tk}.$$

   Furthermore

$$\mathbb{E}[\log W_T] = \sum_{0 \leqslant i < T} \mathbb{E}\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right] + \log(W_0).$$

   Given our initialization of the algorithm, $\log(W_0) = \log(K+1)$.

4. **Lower bound on estimated welfare**
   Denote $\widehat{U}_{ik} \coloneqq \widetilde{x}_k \cdot \widehat{H}_k + \frac{\lambda}{K} \cdot \sum_{k'>k} \widehat{H}_{k'}$, where $\widehat{H}_k \coloneqq \frac{y_i}{p_{ik}} \cdot \mathbb{I}\{k_i = k\}$, so that $\widehat{\boldsymbol{U}}_{ik} = \sum_{j<i} \widehat{U}_{jk}$, and $\mathbb{E}[\widehat{U}_{jk}] = U_i(\widetilde{x}_k)$.

   By definition of W_i,

$$\log\left(\frac{W_{i+1}}{W_i}\right) = \log\left(\sum_k q_{ik} \cdot \exp(\eta \cdot \widehat{U}_{ik})\right).$$

   Since $p_k \geqslant \gamma/(K+1)$ for all k, $\widehat{U}_{ik} \in [0, (K+1)/\gamma]$ for all i and k, and therefore $\eta \cdot \widehat{U}_{ik} \leqslant (K+1)\cdot\eta/\gamma \leqslant 1$ (where the last inequality holds by assumption). Using $\exp(a) \leqslant 1 + a + (e-2)a^2$ for any $a \leqslant 1$ yields

$$\exp\left(\eta \widehat{U}_{ik}\right) \leqslant 1 + \eta \cdot \widehat{U}_{ik} + (e-2) \cdot \left(\eta \cdot \widehat{U}_{ik}\right)^2.$$

   Therefore,

$$\begin{aligned}\log\left(\frac{W_{i+1}}{W_i}\right) \leqslant& \log\left(\sum_k q_{ik} \cdot \left(1 + \eta \cdot \widehat{U}_{ik} + (e-2) \cdot \left(\eta \cdot \widehat{U}_{ik}\right)^2\right)\right) \\ \leqslant& \eta \cdot \sum_k q_{ik} \cdot \widehat{U}_{ik} + (e-2) \cdot \eta^2 \cdot \sum_k q_{ik} \cdot \widehat{U}_{ik}^2\end{aligned}$$

   The second inequality follows from $\log(1+x) \leqslant x$.

5. **Connecting the first order term to welfare**
   Note that, by definition, $q_{ik} = \left(p_{ik} - \frac{\gamma}{K+1}\right)/(1-\gamma)$. Therefore

$$\sum_k q_{ik} \cdot \widehat{U}_{ik} = \frac{1}{1-\gamma} \sum_k p_{ik} \cdot \widehat{U}_{ik} - \frac{\gamma}{(1-\gamma)(K+1)} \cdot \sum_k \widehat{U}_{ik},$$

and thus

$$\mathbb{E}\left[\sum_k q_{ik}\cdot\widehat{U}_{ik}\right]\leqslant\frac{1}{1-\gamma}\mathbb{E}\left[\widetilde{U}_i(x_i)\right],$$

where we have used the fact that $0\leqslant\widetilde{U}_k\leqslant 1$ for all k, given our definition of $\widetilde{U}$, and the fact that k_i is distributed according to p_{ik}, by construction.

6. **Bounding the second moment of estimated welfare**
It remains to bound the term $\mathbb{E}\left[\sum_k q_{ik}\cdot\widehat{U}_{ik}^2\right]$. As in the preceding item, we have

$$\sum_k q_{ik}\cdot\widehat{U}_{ik}^2\leqslant\frac{1}{1-\gamma}\sum_k p_{ik}\cdot\widehat{U}_{ik}^2.$$

We can rewrite

$$\widehat{U}_{ik}=\left(\widetilde{x}_k\cdot\mathbb{I}\{k_i=k\}+\tfrac{\lambda}{K}\cdot\mathbb{I}\{k_i>k\}\right)\cdot\frac{y_i}{p_{ik_i}}.$$

Bounding $y_i\leqslant 1$ immediately gives

$$\mathbb{E}_i\left[\widehat{U}_{ik}^2\right]\leqslant\frac{\widetilde{x}_k^2}{p_{ik}}+\left(\tfrac{\lambda}{K}\right)^2\cdot\sum_{k'>k}\frac{1}{p_{ik'}},$$

and therefore

$$\begin{aligned}\mathbb{E}_i\left[\sum_k p_{ik}\cdot\widehat{U}_{ik}^2\right]&\leqslant\sum_k\widetilde{x}_k^2+\left(\tfrac{\lambda}{K}\right)^2\cdot\sum_k\sum_{k'>k}\frac{p_{ik}}{p_{ik'}}\leqslant\sum_k\left(\tfrac{k}{K}\right)^2+\left(\tfrac{\lambda}{K}\right)^2\cdot\sum_k p_{ik}\sum_{k'\neq k}\tfrac{K+1}{\gamma}\\&=\tfrac{K(K+1)(2K+1)}{6K^2}+\tfrac{\lambda^2}{\gamma}\tfrac{K+1}{K}=\tfrac{K+1}{K}\cdot\left(\tfrac{2K+1}{6}+\tfrac{\lambda^2}{\gamma}\right).\end{aligned}$$

7. **Collecting inequalities**
Combining the preceding items, we get

$$\begin{aligned}&\eta\cdot\left(\sup_x\boldsymbol{U}(x)-T\cdot\frac{\lambda}{K}\right)\\&\leqslant\eta\cdot\max_k\widetilde{\boldsymbol{U}}_{Tk}\leqslant\mathbb{E}[\log W_T]&&\text{(Item 1)}\\&=\sum_{0\leqslant i<T}\mathbb{E}\left[\log\left(\frac{W_{i+1}}{W_i}\right)\right]+\log(K+1)&&\text{(Item 3)}\\&\leqslant\frac{\eta}{1-\gamma}\cdot\mathbb{E}\left[\widetilde{\boldsymbol{U}}\right]+(e-2)\cdot\frac{\eta^2}{1-\gamma}\sum_{1\leqslant i\leqslant T}\sum_k\mathbb{E}\left[p_{ik}\cdot\widehat{U}_{ik}^2\right]+\log(K+1)&&\text{(Item 4 and 5)}\\&\leqslant\frac{\eta}{1-\gamma}\cdot\mathbb{E}\left[\widetilde{\boldsymbol{U}}\right]+(e-2)\cdot\frac{\eta^2}{1-\gamma}T\cdot\tfrac{K+1}{K}\cdot\left(\tfrac{2K+1}{6}+\tfrac{\lambda^2}{\gamma}\right)+\log(K+1).&&\text{(Item 6)}\end{aligned}$$

Multiplying by $(1-\gamma)$ and dividing by η, adding $\gamma\sup_x\boldsymbol{U}(x)+T\frac{\lambda}{K}$ to both sides and subtracting

$\mathbb{E}\left[\widetilde{\boldsymbol{U}}\right]$, bounding $\sup_x \boldsymbol{U}(x) \leqslant T$, and $\mathbb{E}\left[\widetilde{\boldsymbol{U}}\right] \leqslant \mathbb{E}\left[\boldsymbol{U}\right]$ (from Item 1), yields

$$\sup_x \boldsymbol{U}(x) - \mathbb{E}\left[\boldsymbol{U}\right] \leqslant \left(\gamma + \eta \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2K+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \tfrac{\lambda}{K}\right) \cdot T + \tfrac{\log(K+1)}{\eta}. \tag{C.29}$$

This proves the first claim of the theorem.

8. **Optimizing tuning parameters**
Suppose now that we choose the tuning parameters as follows:

$$\gamma = c_1 \cdot \left(\tfrac{\log(T)}{T}\right)^{1/3}, \qquad \eta = c_2 \cdot \gamma^2, \qquad K = c_3/\gamma.$$

Plugging in we get

$$\begin{aligned}
&\sup_x \boldsymbol{U}(x) - \mathbb{E}\left[\boldsymbol{U}\right] \\
&\leqslant \left(\gamma + c_2 \cdot \gamma^2 \cdot (e-2)\tfrac{K+1}{K} \cdot \left(\tfrac{2c_3/\gamma+1}{6} + \tfrac{\lambda^2}{\gamma}\right) + \lambda \cdot \gamma/c_3\right) \cdot T + \tfrac{\log(K+1)}{c_2 \cdot \gamma^2} \\
&= \log(T)^{1/3} T^{2/3} \cdot \left(c_1 + (e-2)\tfrac{K+1}{K} \cdot c_1 c_2 \left(\tfrac{c_3}{3} + \lambda^2 + \tfrac{\gamma}{6}\right) + \lambda \tfrac{c_1}{c_3} + \frac{\log(T^{1/3}\log(T)^{-1/3} c_3/c_1 + 1)}{c_1^2 \log(T)}\right) \\
&= \log(T)^{1/3} T^{2/3} \cdot \left(c_1 + (e-2) \cdot c_1 c_2 \left(\tfrac{c_3}{3} + \lambda^2\right) + \lambda \tfrac{c_1}{c_3} + \frac{1}{3c_1^2} + o(1)\right).
\end{aligned}$$

The second claim of the theorem follows.

□

C.2.3 Theorem 28 (Lower Bound on Regret for the Concave Case)

We now present the $\sqrt{T}$ lower bound construction for the stochastic case where the (expected) utility function $\mathbb{U}$ is concave.

Proof of Theorem 28.

Defining a family of distributions for v Define $\bar{h} \coloneqq \frac{1-\sqrt{1-\lambda}}{2}$ and notice that $0 < \bar{h} < \frac{1}{2}$. Define $\bar{\eta} \coloneqq \left(\bar{h} \cdot (1-\bar{h})^{1-\lambda} \cdot (1-\lambda)\right)^{-1}$ and $\bar{\varepsilon} \coloneqq \frac{1}{2} \cdot \min(\bar{\eta}, \frac{2}{3} \cdot 2^{-\lambda})$. For each $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$ and each $x \in [0,1]$, define

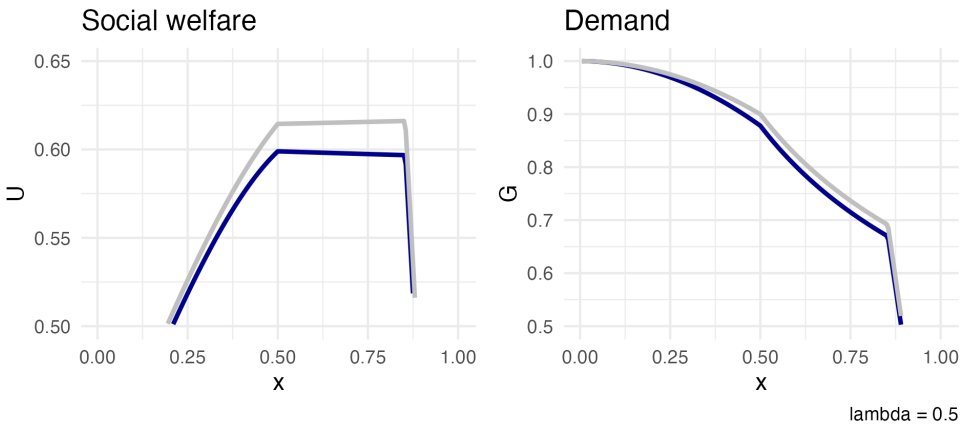
$$f^\varepsilon(x) \coloneqq \bar{c} \cdot \left(\left(2^{2-\lambda} - 8 \cdot \bar{h} \cdot \varepsilon\right) \cdot x \cdot \mathbb{I}_{[0,\frac{1}{2})}(x) + \frac{1}{x^{2-\lambda}} \cdot \mathbb{I}_{[\frac{1}{2}, 1-\bar{h}]}(x) + (\bar{\eta} + \varepsilon) \cdot \mathbb{I}_{(1-\bar{h},1]}(x)\right),$$

where $\bar{c}$ is such that $\int_0^1 f^0(x)\mathrm{d}x = 1$. For each $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, note that f^ε is a density function on $[0,1]$, i.e., a non-negative function whose integral is 1. For each $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, let μ^ε be the probability measure whose density is f^ε, and define $\mathbb{G}^\varepsilon$ and $\mathbb{U}^\varepsilon$ as the demand function and the expected social welfare associated to μ^ε, respectively (see Figure C.1 for an illustration).

Properties of $\mathbb{U}$ Define also $\bar{x} \coloneqq \frac{1}{2} \cdot \left(\frac{1}{2} + (1-\bar{h})\right) = \frac{3}{4} - \frac{\bar{h}}{2}$ and $\bar{m} \coloneqq \frac{1-\sqrt{1-\lambda}}{8} \cdot (1-\lambda)^{3/2}$. Notice that, for all $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, we have:

- $\mathbb{U}^\varepsilon$ is continuous and concave.

Figure C.1: Construction for proving the lower bound on regret for the concave case



- $\mathbb{U}^{\varepsilon}$ is strictly increasing in $[0, \frac{1}{2}]$, linear in $[\frac{1}{2}, 1-\bar{h}]$ with slope $(1-\lambda)\cdot\bar{h}\cdot\varepsilon$, and strictly decreasing on $[1-\bar{h}, 1]$, which in particular implies that the maximum of $\mathbb{U}^{\varepsilon}$ is at $1-\bar{h}$ if $\varepsilon > 0$, and at $\frac{1}{2}$ if $\varepsilon < 0$.
- If $\varepsilon > 0$, then $\mathbb{U}^{\varepsilon}(1-\bar{h}) - \max_{x\in[0,\bar{x}]} \mathbb{U}^{\varepsilon}(x) = \bar{m}\cdot|\varepsilon| = \mathbb{U}^{-\varepsilon}(\frac{1}{2}) - \max_{x\in[\bar{x},1]} \mathbb{U}^{-\varepsilon}(x)$.

Now, consider the sequence of individual valuations $v_1, v_2, \cdots \in [0,1]$, and assume that, for each $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, when the underlying distribution is $\mathbb{P}^{\varepsilon}$, this sequence is i.i.d. (independent of the randomization used by the algorithm) with common distribution μ^{ε}. The previous list of properties implies that, for each $\varepsilon \in (0, \bar{\varepsilon})$ (resp., $\varepsilon \in (-\bar{\varepsilon}, 0)$), when the underlying distribution is $\mathbb{P}^{\varepsilon}$, the expected instantaneous regret at time t is at least $\bar{m}\cdot|\varepsilon|$ if the learner plays in the region $\bar{I} \coloneqq [0, \bar{x}]$ (resp., in the region $\bar{J} \coloneqq (\bar{x}, 1]$). It follows that, in order not to suffer linear regret, the learner has to discriminate the sign of ε.

Intuition for the proof Now, the high-level idea is that in order to discriminate the sign of ε, due to information-theoretic arguments, the learner needs on the order of $\frac{1}{\varepsilon^2}$ observations. Therefore, for a number of periods on the order of $\frac{1}{\varepsilon^2}$, the algorithm is playing "in the dark", and thus suffers a regret on the order of $\min(\frac{\varepsilon}{\varepsilon^2}, \varepsilon\cdot T)$. Choosing ε on the order of $T^{-1/2}$, the algorithm ends to suffer a regret on the order of $\sqrt{T}$, when the underlying distribution is one between $\mathbb{P}^{\varepsilon}$ or $\mathbb{P}^{-\varepsilon}$.

Defining constants We now formalize this idea. Let

$$\gamma \coloneqq \left(\int_0^{1/2} \left(\frac{16\bar{h}}{2^{2-\lambda} - 8\bar{h}\bar{\varepsilon}} \right)^2 f^{-\bar{\varepsilon}}(x)\mathrm{d}x + \int_{1-\bar{h}}^{1} \left(\frac{2}{\bar{\eta} - \bar{\varepsilon}} \right)^2 f^{\bar{\varepsilon}}(x)\mathrm{d}x \right)^{1/2} > 0$$

Let $\bar{M} > 0$ such that $2\cdot\sqrt{\frac{\sqrt{2}}{3}\cdot\frac{\gamma\cdot\bar{M}}{\bar{m}}} = 1$. Let $M \in (0, \bar{M})$ such that

$$k \coloneqq \sqrt{\frac{\frac{M}{\bar{m}}}{\frac{\sqrt{2}}{3}\cdot\gamma}} < \bar{\varepsilon}\,.$$

From now on, fix a time horizon $T \in \mathbb{N}$ and let $\varepsilon \coloneqq \frac{k}{\sqrt{T}}$. In the following we use the notation $\mathbb{E}^{\varepsilon}$ (resp., $\mathbb{E}^{-\varepsilon}$) to denote the expectation with respect to the probability measure $\mathbb{P}^{\varepsilon}$ (resp., $\mathbb{P}^{-\varepsilon}$). Let $x_1, x_2, \dots$ be the policies chosen by the algorithm. Note that, since the algorithm bases its decision at time t only on the (partial) knowledge of $v_1, \dots, v_{t-1}$ and some independent randomization, there exists a (measurable) function $\varphi_t \colon [0,1]^{t-1} \to [0,1]$ such that

$$\mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\} \mid v_1, \dots, v_{t-1}\right] = \varphi_t(v_1, \dots, v_{t-1}) = \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\} \mid v_1, \dots, v_{t-1}\right].$$

Then, for each time t, it holds

$$\begin{aligned}\left|\mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right] - \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right]\right| &= \left|\mathbb{E}^{\varepsilon}\left[\varphi_t(v_1, \dots, v_{t-1})\right] - \mathbb{E}^{-\varepsilon}\left[\varphi_t(v_1, \dots, v_{t-1})\right]\right| \\ &\leqslant \left\| \bigotimes_{s=1}^{t-1} \mu^{\varepsilon} - \bigotimes_{s=1}^{t-1} \mu^{-\varepsilon} \right\|_{\mathrm{TV}} = (\star)\end{aligned}$$

Relating choice probabilities for positive and negative ε By Pinsker's inequality and the fact that the Kullback-Leibler divergence is upper bounded by the χ^2-divergence, it follows that

$$\begin{aligned}(\star) &\leqslant \sqrt{\frac{\mathcal{D}_{\mathrm{KL}}\left(\bigotimes_{s=1}^{t-1} \mu^{-\varepsilon}, \bigotimes_{s=1}^{t-1} \mu^{\varepsilon}\right)}{2}} = \sqrt{\frac{(t-1) \cdot \mathcal{D}_{\mathrm{KL}}\left(\mu^{-\varepsilon}, \mu^{\varepsilon}\right)}{2}} \leqslant \sqrt{\frac{(t-1) \cdot \mathcal{D}_{\chi^2}\left(\mu^{-\varepsilon}, \mu^{\varepsilon}\right)}{2}} \\ &= \sqrt{\frac{t-1}{2} \int_0^1 \left|\frac{f^{\varepsilon}(x)}{f^{-\varepsilon}(x)} - 1\right|^2 f^{-\varepsilon}(x)\mathrm{d}x} = (\star\star)\end{aligned}$$

Now, noticing that

$$\begin{aligned}\int_0^1 \left|\frac{f^{\varepsilon}(x)}{f^{-\varepsilon}(x)} - 1\right|^2 f^{-\varepsilon}(x)\mathrm{d}x &= \int_0^{1/2} \left(\frac{16\bar{h}\varepsilon}{2^{2-\lambda} + 8\bar{h}\varepsilon}\right)^2 f^{-\varepsilon}(x)\mathrm{d}x + \int_{1-\bar{h}}^1 \left(\frac{2\varepsilon}{\bar{\eta} - \varepsilon}\right)^2 f^{-\varepsilon}(x)\mathrm{d}x \\ &\leqslant \left(\int_0^{1/2} \left(\frac{16\bar{h}}{2^{2-\lambda} - 8\bar{h}\bar{\varepsilon}}\right)^2 f^{-\bar{\varepsilon}}(x)\mathrm{d}x + \int_{1-\bar{h}}^1 \left(\frac{2}{\bar{\eta} - \bar{\varepsilon}}\right)^2 f^{\bar{\varepsilon}}(x)\mathrm{d}x\right) \cdot \varepsilon^2 \\ &= \gamma^2 \cdot \varepsilon^2,\end{aligned}$$

it follows that

$$(\star\star) \leqslant \gamma \cdot \varepsilon \cdot \sqrt{\frac{t-1}{2}}.$$

Summing over $t = 1, 2, \dots, T$, we obtain

$$\left|\mathbb{E}^{\varepsilon}\left[\sum_{t=1}^{T} \mathbb{I}\left\{x_t \in \bar{I}\right\}\right] - \mathbb{E}^{-\varepsilon}\left[\sum_{t=1}^{T} \mathbb{I}\left\{x_t \in \bar{I}\right\}\right]\right| \leqslant \frac{\sqrt{2}}{3} \cdot \gamma \cdot \varepsilon \cdot T^{3/2} = \frac{\sqrt{2}}{3} \cdot \gamma \cdot k \cdot T.$$

Upper bound on regret for $\varepsilon > 0$ implies lower bound on regret for $-\varepsilon$. Now, suppose that in the scenario determined by $\mathbb{P}^{\varepsilon}$ the algorithm suffer a regret $R_T^{\varepsilon} \leqslant M \cdot \sqrt{T}$. Then

$$M \cdot \sqrt{T} \geqslant R_T^{\varepsilon} \geqslant \bar{m} \cdot \varepsilon \cdot \sum_{t=1}^{T} \mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right] = \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \sum_{t=1}^{T} \mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right].$$

and rearranging

$$\sum_{t=1}^{T} \mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right] \leqslant \frac{M \cdot T}{\bar{m} \cdot k}$$

It follows that the expected number of times the algorithm plays in the (correct) region I when the underlying scenario is determined by $\mathbb{P}^{-\varepsilon}$ is

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}(x_t \in \bar{I})\right] &= \left(\sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right] - \sum_{t=1}^{T} \mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right]\right) + \sum_{t=1}^{T} \mathbb{E}^{\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{I}\right\}\right] \\ &\leqslant \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k}\right) \cdot T\end{aligned}$$

The last inequality implies that the expected number of times that the algorithm plays in the (wrong) region $J = I^c$ when the underlying scenario is determined by $\mathbb{P}^{-\varepsilon}$ is lower bounded by

$$\sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in \bar{J}\right\}\right] = \sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \notin \bar{I}\right\}\right] \geqslant \left(1 - \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k}\right)\right) \cdot T\,,$$

which implies that the regret the algorithm suffers in the scenario determined by $\mathbb{P}^{-\varepsilon}$ is lower bounded by

$$\begin{aligned}R_T^{-\varepsilon} &\geqslant \bar{m} \cdot \varepsilon \cdot \sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in J\right\}\right] = \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \sum_{t=1}^{T} \mathbb{E}^{-\varepsilon}\left[\mathbb{I}\left\{x_t \in J\right\}\right] \\ &\geqslant \bar{m} \cdot \frac{k}{\sqrt{T}} \cdot \left(1 - \left(\frac{\sqrt{2}}{3} \cdot \gamma \cdot k + \frac{M}{\bar{m} \cdot k}\right)\right) \cdot T = \bar{m} \cdot k \cdot \left(1 - 2\sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}}\right) \cdot \sqrt{T}\,.\end{aligned}$$

Putting everything together, any algorithm suffers at least $\min\left(M, \bar{m} \cdot k \cdot \left(1 - 2 \cdot \sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}}\right)\right) \cdot \sqrt{T}$ regret, in at least one scenario between the ones determined by $\mathbb{P}^{\varepsilon}$ and $\mathbb{P}^{-\varepsilon}$. Recalling that our choice of M implies $1 - 2\sqrt{\frac{\sqrt{2} \cdot \gamma \cdot M}{3 \cdot \bar{m}}} > 0$, the conclusion follows. □

C.2.4 Theorem 29 (Stochastic Upper Bound on Regret of Dyadic Search for Social Welfare)

We now present a proof of the $\sqrt{T}$ upper bound on the regret achieved by Dyadic Search for Social Welfarein the stochastic case when the underlying (expected) utility function $\mathbb{U}$ is concave.

For the sake of simplicity, we assume that $\mathbb{U}$ admits a unique maximizer $x^\star \in [0, 1]$ (the other cases can be treated similarly and, actually, they ended up having better constants in the final regret guarantees).

For each epoch $\tau = 1, 2, \ldots$, we refer to the three current l (left), c (center) and r (right) points of the corresponding epoch τ using l_τ, c_τ and r_τ, respectively. For any time t, the epoch to which the time t belongs is denoted τ_t. The length of an interval J is denoted $|J|$, while the number of elements in a finite set A is denoted $\#A$.

Consider a family $(v_{x,i})_{x \in [0,1], i \in \mathbb{N}}$ of random variables such that, for each $x \in [0, 1]$, the sequence $(v_{x,i})_{i \in \mathbb{N}}$ is i.i.d. with the same distribution as $(v_i)_{i \in \mathbb{N}}$. With these random variables, we can define the auxiliary family $(y_{x,i})_{x \in [0,1], i \in \mathbb{N}} \coloneqq \left(\mathbb{I}\left\{x \leqslant v_{x,i}\right\}\right)_{x \in [0,1], i \in \mathbb{N}}$. We assume that, whenever we select a

policy $x \in [0,1]$ at time t, we observe $\mathbb{I}\{x \leqslant v_{x,n_t(x)}\}$ (recall that $n_t(x) = \sum_{s=1}^{t} \mathbb{I}\{x_s = x\}$) instead of $\mathbb{I}\{x \leqslant v_t\}$. This does not change anything in expectation, but will be useful in what follows.

The next lemma states that Algorithm 8 maintains confidence intervals containing the differences of the welfare function (among left, center and right points) with high probability.

Lemma 23 (Confidence intervals contain true welfare differences with high probability)**.** *There exists a constant $\widetilde{C} \in (0, 20]$ such that, for every time horizon T and any $\delta \in (0,1)$, if the learner runs Algorithm 8 with confidence parameter δ, then the probability of the event*

$$\mathcal{E} \coloneqq \bigcap_{t=1}^{T} \Big(\big\{ \mathbb{U}(c_{\tau_t}) - \mathbb{U}(l_{\tau_t}) \in J_t(l_{\tau_t}, c_{\tau_t}) \big\} \cap \big\{ \mathbb{U}(r_{\tau_t}) - \mathbb{U}(c_{\tau_t}) \in J_t(c_{\tau_t}, l_{\tau_t}) \big\} \cap \big\{ \mathbb{U}(r_{\tau_t}) - \mathbb{U}(l_{\tau_t}) \in J_t(r_{\tau_t}, l_{\tau_t}) \big\} \Big)$$

is lower bounded by $1 - \widetilde{C} \cdot T^2 \cdot \delta$.

Proof. For each $n \in \mathbb{N}$, let $\mathcal{D}_n \coloneqq \{k \cdot 2^{-n} \mid k \in \mathbb{Z}\}$, let $\mathcal{D}_n^\star \coloneqq \{x_{n,1}, \dots, x_{n,10}\} \subset \mathcal{D}_n$ such that

$$x_{n,1} < \cdots < x_{n,5} \leqslant x^\star \leqslant x_{n,6} < \cdots < x_{n,10}$$

and $x_{n,j+1} - x_{n,j} \leqslant 2^{-n}$, for all $j \in \{1, \dots, 9\}$. Define $\mathcal{D} \coloneqq \bigcup_{n=1}^{T} \mathcal{D}_n^\star \cap (0,1)$. Consider the following events

$$\mathcal{E}' \coloneqq \bigcap_{\substack{n,t \in \{1,\dots,T\} \\ j \in \{1,\dots,10\}}} \left\{ \left| \frac{1}{t} \sum_{s=1}^{t} y_{x_{n,j},s} - \mathbb{G}(x_{n,j}) \right| \leqslant \sqrt{\frac{1}{2t} \log\left(\frac{2}{\delta}\right)} \right\}$$

$$\mathcal{E}'' \coloneqq \bigcap_{\substack{n \in \{1,\dots,T\} \\ m \in \{1,\dots,\lfloor \log_2(T) \rfloor\} \\ j \in \{1,\dots,9\}}} \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j} + \frac{i}{2^{n+m}},1} - \frac{1}{x_{n,j+1} - x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \mathbb{G}(x) \mathrm{d}x \right| \leqslant \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} + \frac{2}{2^m} \right\}$$

and note that $\mathcal{E} \subset \mathcal{E}' \cap \mathcal{E}''$, since, in the event $\mathcal{E}' \cap \mathcal{E}''$, Algorithm 8 will query only points in $\mathcal{D}^\star$, given that it uses only a subset of the estimates in the definition of $\mathcal{E}'$ and $\mathcal{E}''$ to build its own estimates (in particular, due to the ties breaking rules, to estimate the integral terms it will only use the *first* query of the relevant dyadic points). Now, notice that for each $n \in \{1, \dots, n\}$, each $m \in \{1, \dots, \lfloor \log_2(T) \rfloor\}$ and each $j \in \{1, \dots, 9\}$ we have

$$\begin{aligned}
&\left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j} + \frac{i}{2^{n+m}},1} - \frac{1}{x_{n,j+1} - x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \mathbb{G}(x) \mathrm{d}x \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} + \frac{2}{2^m} \right\} \\
&\quad \subset \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j} + \frac{i}{2^{n+m}},1} - \frac{1}{2^m} \sum_{i=1}^{2^m-1} \mathbb{G}\left(x_{n,j} + \frac{i}{2^{n+m}} \right) \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} \right\} \\
&\qquad \cup \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} \mathbb{G}\left(x_{n,j} + \frac{i}{2^{n+m}} \right) - \frac{1}{x_{n,j+1} - x_{n,j}} \cdot \int_{x_{n,j}}^{x_{n,j+1}} \mathbb{G}(x) \mathrm{d}x \right| > \frac{2}{2^m} \right\} \\
&\quad = \left\{ \left| \frac{1}{2^m} \sum_{i=1}^{2^m-1} y_{x_{n,j} + \frac{i}{2^{n+m}},1} - \frac{1}{2^m} \sum_{i=1}^{2^m-1} \mathbb{G}\left(x_{n,j} + \frac{i}{2^{n+m}} \right) \right| > \sqrt{\frac{1}{2 \cdot 2^m} \log\left(\frac{2}{\delta}\right)} \right\}
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
&\left|\frac{1}{2^m}\sum_{i=1}^{2^m-1}\mathbb{G}\left(x_{n,j}+\frac{i}{2^{n+m}}\right)-\frac{1}{x_{n,j+1}-x_{n,j}}\cdot\int_{x_{n,j}}^{x_{n,j+1}}\mathbb{G}(x)\mathrm{d}x\right| \\
&\leqslant \sum_{i=1}^{2^m-1}\int_{x_{n,j}+\frac{i-1}{2^{n+m}}}^{x_{n,j}+\frac{i}{2^{n+m}}}\left(\mathbb{G}(x)-\mathbb{G}\left(x_{n,j}+\frac{i}{2^{n+m}}\right)\right)\mathrm{d}x+\frac{1}{2^m} \\
&\leqslant \sum_{i=1}^{2^m-1}\int_{x_{n,j}+\frac{i-1}{2^{n+m}}}^{x_{n,j}+\frac{i}{2^{n+m}}}\left(\mathbb{G}\left(x_{n,j}+\frac{i-1}{2^{n+m}}\right)-\mathbb{G}\left(x_{n,j}+\frac{i}{2^{n+m}}\right)\right)\mathrm{d}x+\frac{1}{2^m} \\
&\leqslant \frac{1}{2^m}\cdot\left(\mathbb{G}\left(x_{n,j}\right)-\mathbb{G}\left(x_{n,j+1}\right)\right)+\frac{1}{2^m}\leqslant\frac{2}{2^m}
\end{aligned}$$

By De Morgan's laws, a union bound and Hoeffding's inequality, we have $\mathbb{P}\big[\mathcal{E}^c\big] \leqslant \mathbb{P}\big[(\mathcal{E}')^c\big] + \mathbb{P}\big[(\mathcal{E}'')^c\big] \leqslant 20\cdot T^2\cdot\delta$. □

The following lemma establishes the rate of shrinking of the length of the confidence intervals as the length of an epoch increases.

Lemma 24 (Confidence intervals shrink with epoch length)**.** *For any $\delta\in(0,1)$, if the learner runs Algorithm 8 with confidence parameter δ then, for any time t,*

$$\max\left(\left|J_t(l_{\tau_t},c_{\tau_t})\right|,\left|J_t(c_{\tau_t},r_{\tau_t})\right|,\left|J_t(l_{\tau_t},r_{\tau_t})\right|\right)\leqslant\frac{\widetilde{c}_\delta}{\sqrt{t-t_{\tau_t-1}}}\,, \tag{C.30}$$

whenever $t-t_{\tau_t-1}\geqslant\widetilde{n}$, where $\widetilde{n}=10$ and $\widetilde{c}_\delta=72\cdot\sqrt{10}\cdot\Big(\sqrt{2\log(2/\delta)}+4\Big)$.

We break the proof of Lemma 24 in several steps. Let $d_1,d_2,d_3,d_4,d_5>0$ be constants. For each $k\in\{1,2,3\}$, define

$$f_k:\{0,1,2,\dots\}\to[0,+\infty],\qquad n\mapsto\frac{d_k}{\sqrt{n}}$$

and for each $k\in\{4,5\}$ define

$$f_k:\{0,1,2,\dots\}\to[0,+\infty],\qquad n\mapsto\frac{d_4}{\sqrt{2^{\lfloor\log_2(n+1)\rfloor}-1}}+\frac{d_5}{2^{\lfloor\log_2(n+1)\rfloor}}\,,$$

with the usual convention that $a/0=+\infty$, for any $a>0$. Suppose that $m_1(0),m_2(0),m_3(0),m_4(0),m_5(0)\in\{0,1,2,\dots\}$ and consider the following algorithm.

Algorithm 12 Index selection

```
for s = 1, 2, ... do
    Let k_s = min( argmax_{k∈[5]} f_k(m_k(s−1)) )
    m_{k_s}(s) = m_{k_s}(s−1) + 1
    for i ∈ [5]\{k_s} do
        m_i(s) = m_i(s−1)
```

The following lemma holds.

Lemma 25. *Consider Algorithm 12 and the notation defined therein. For each $s\in\mathbb{N}$ there exists an index $i\in[5]$ for which $m_i(s)\geqslant\lceil s/5\rceil$*

Proof. Let $s \in \mathbb{N}$ and suppose by contradiction that for each $k \in [5]$ it holds that $m_k(s) < s/5$. Then

$$s \leqslant \sum_{k=1}^{5} m_k(s) \leqslant 5 \cdot \max_{k \in [5]} m_k(s) < 5 \cdot \frac{s}{5} = s \ ,$$

which is a contradiction. It follows that there exists $k \in [5]$ for which $m_k(s) \geqslant s/5$, which also implies $m_k(s) \geqslant \lceil s/5 \rceil$. Given that s was arbitrarily chosen, the conclusion follows. □

Notice that, for each $n \in \{0, 1, 2, \dots\}$, we have

$$\frac{d_4}{\sqrt{n}} \leqslant \frac{d_4}{\sqrt{2^{\lfloor \log_2(n+1) \rfloor} - 1}} \leqslant \frac{2d_4}{\sqrt{n}}$$

and

$$0 \leqslant \frac{d_5}{\sqrt{2^{\lfloor \log_2(n+1) \rfloor}}} \leqslant \frac{2d_5}{n} \ ,$$

which implies that, for each $k \in [5]$ and each $n \in \{0, 1, 2, \dots\}$

$$\frac{d_k}{\sqrt{n}} \leqslant f_k(n) \leqslant \frac{D_k}{\sqrt{n}}$$

where $D_1 = d_1, D_2 = d_2, D_3 = d_3, D_4 = D_5 = 2(d_4 + d_5)$.

The following lemma holds.

Lemma 26. *Consider Algorithm 12 and the notation defined therein. For any $i, j \in [5]$ and any $s \in \mathbb{N}$ it holds*

$$m_i(s) \geqslant \left(\frac{d_i}{D_j}\right)^2 (m_j(s) - 1) \ .$$

Proof. Let $i, j \in [5]$. Suppose by contradiction that the conclusion does not hold. Then there exists a smallest $s \in \{0, 1, 2, \dots\}$ for which

$$m_i(s) < \left(\frac{d_i}{D_j}\right)^2 (m_j(s) - 1) \ ,$$

which we call s_0. Notice that $s_0 \neq 0$. Then, the fact that

$$m_i(s_0 - 1) \geqslant \left(\frac{d_i}{D_j}\right)^2 (m_j(s_0 - 1) - 1) \ ,$$

implies that at time s_0 the algorithm selected $k_{s_0} = j$, which in turn implies that $m_i(s_0 - 1) = m_i(s_0)$ and $m_j(s_0 - 1) = m_j(s_0) - 1$. It follows that

$$\left(\frac{d_i}{D_j}\right)^2 m_j(s_0 - 1) = \left(\frac{d_i}{D_j}\right)^2 \big(m_j(s_0) - 1\big) > m_i(s_0) = m_i(s_0 - 1) \ ,$$

Rearranging, we get

$$m_j(s_0 - 1) > \left(\frac{D_j}{d_i}\right)^2 m_i(s_0 - 1) \ .$$

from which it follows that

$$f_j\big(m_j(s_0-1)\big) \leqslant \frac{D_j}{\sqrt{m_j(s_0-1)}} < \frac{d_i}{\sqrt{m_i(s_0-1)}} \leqslant f_i\big(m_i(t_0-1)\big)\,.$$

This last inequality implies that at time s_0 the algorithm should have chosen the index i and not the index j, which is a contradiction. □

Combining the last two lemmas we can prove the following result.

Lemma 27. *Consider Algorithm 12 and the notation defined therein. Then, for any $s \geqslant 5$ it holds that*

$$\max_{k\in[5]} f_k\big(m_k(s)\big) \leqslant \frac{D}{\sqrt{s-5}}$$

where $D = \sqrt{5}\cdot\big(\max_{j\in[5]} D_j\big)\cdot\big(\max_{k\in[5]}\frac{D_k}{d_k}\big)$.

Proof. Let $s \geqslant 5$. Pick $j \in [5]$ such that $m_j(s) \geqslant \lceil s/5\rceil$ (which does exist by Lemma 25). Then, by Lemma 26

$$\begin{aligned}\max_{k\in[5]} f_k\big(m_k(s)\big) &\leqslant \max_{k\in[5]} \frac{D_k}{\sqrt{m_k(s)}} \leqslant \max_{k\in[5]} \frac{D_k}{\sqrt{\left(\frac{d_k}{D_j}\right)^2\big(m_j(s)-1\big)}} \\ &= D_j\cdot\max_{k\in[5]}\left(\frac{D_k}{d_k}\right)\frac{1}{\sqrt{m_j(s)-1}} \leqslant D_j\cdot\max_{k\in[5]}\left(\frac{D_k}{d_k}\right)\frac{1}{\sqrt{\lceil s/5\rceil-1}} \leqslant \frac{D}{\sqrt{s-5}}\,.\end{aligned}$$

□

We are now ready for the proof of Lemma 24.

Proof of Lemma 24. It is enough to notice that Algorithm 8 with confidence parameter $\delta\in(0,1)$ relies, inside each epoch, on the same routine given by Algorithm 12 with $d_1 = l\cdot\sqrt{\frac{\log(2/\delta)}{2}}, d_2 = c\cdot\sqrt{\frac{\log(2/\delta)}{2}}, d_3 = r\cdot\sqrt{\frac{\log(2/\delta)}{2}}, d_4 = \lambda\cdot(c-l)\cdot\sqrt{\frac{\log(2/\delta)}{2}}, d_5 = 2\cdot\lambda\cdot(c-l)$, with the convention that l correspond to 1, c corresponds to 2, r corresponds to 3, (l,c) corresponds to 4 and (c,r) corresponds to 5, the correspondence between times is given by $s = t - t_{\tau_t-1}$, and, for each $s\in\{0,1,2,\dots\}$, $m_1(s) = n_{s+t_{\tau_t-1}}(l)$, $m_2(s) = n_{s+t_{\tau_t-1}}(c)$, $m_3(s) = n_{s+t_{\tau_t-1}}(r)$, $m_4(s) = \sum_{i\leqslant s+t_{\tau_t-1}}\mathbb{I}\{x_i\in(l,c)\}$, $m_5(s) = \sum_{i\leqslant s+t_{\tau_t-1}}\mathbb{I}\{x_i\in(c,r)\}$. With these conventions, in Lemma 27 we have that $D \leqslant 9\cdot\sqrt{5}\cdot\left(\sqrt{2\log(2/\delta)}+4\right)$ and, for example (the other cases can be proved analogously)

$$\begin{aligned}\big|J_t(l_{\tau_t}, r_{\tau_t})\big| &\leqslant 2\cdot\big(\Gamma_t(r)+\Gamma_t(l)+\Gamma_t(l,c)+\Gamma_t(c,r)\big) \leqslant 2\cdot 4\cdot\max_{k\in[5]} f_k\big(m_k(s)\big) \\ &\leqslant 8\cdot\frac{D}{\sqrt{t-t_{\tau_t-1}-5}} \leqslant \frac{\widetilde{c}_\delta}{\sqrt{2}}\cdot\frac{1}{\sqrt{t-t_{\tau_t-1}-5}} \leqslant \frac{\widetilde{c}_\delta}{\sqrt{t-t_{\tau_t-1}}}\end{aligned}$$

where in the last inequality we used the fact that $t - t_{\tau_t-1} \geqslant 10$. □

Lemma 23 and Lemma 24 allow us to prove Theorem 29, which closely follows the proof given in [23].

Proof of Theorem 29. Define τ_T as the last epoch, $t_0 = 0$ and (if not already defined) $t_{\tau_T} = T$.

Due to Lemma 23, we may (and do!) assume that for each $t \in \{1, \dots, T\}$ it holds

$$\big(\mathbb{U}(c_{\tau_t}) - \mathbb{U}(l_{\tau_t}) \in J_t(l_{\tau_t}, c_{\tau_t})\big) \wedge \big(\mathbb{U}(r_{\tau_t}) - \mathbb{U}(c_{\tau_t}) \in J_t(c_{\tau_t}, l_{\tau_t})\big) \wedge \big(\mathbb{U}(r_{\tau_t}) - \mathbb{U}(l_{\tau_t}) \in J_t(r_{\tau_t}, l_{\tau_t})\big) \,.$$

This is because, given our choice $\delta = \frac{1}{T^{5/2}}$, assuming these conditions costs us in the expected regret a further additive term which is no greater than $T \cdot \widetilde{C} \cdot T^2 \cdot \delta = \widetilde{C} \cdot \sqrt{T}$.

Under these assumptions, notice that for each $\tau \in [\tau_T]$ we have that $x^\star \in I_\tau$. In fact, if the confidence intervals are guaranteed to contain the corresponding differences in the expected welfare, every time Algorithm 8 shrinks the active interval is because all the discarded points are guaranteed to be suboptimal.

For each epoch $\tau \in \{1, \dots, \tau_T\}$, define

$$B_\tau \coloneqq (t_\tau - 1) - t_{\tau-1} \,.$$

Now, for each epoch $\tau \in \{1, \dots, \tau_T\}$ if $B_\tau \geqslant \widetilde{n}$, then

$$\max_{x \in [l_\tau, r_\tau]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \leqslant 2 \cdot \widetilde{c}_\delta \cdot \sqrt{\frac{1}{B_\tau}} \,.$$

In fact, assume that $x^\star > r_\tau$ (the other cases have similar proofs). Then, leveraging concavity, and recalling that $\inf\big(J_{t_\tau - 1}(l_\tau, r_\tau)\big) < 0$ and that $x^\star \in I_\tau$ (which implies $\frac{x^\star - l_\tau}{r_\tau - l_\tau} \leqslant 2$), we have

$$\begin{aligned}
\max_{x \in [l_\tau, r_\tau]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) = \mathbb{U}(x^\star) - \mathbb{U}(l_\tau) &= \frac{\mathbb{U}(x^\star) - \mathbb{U}(r_\tau)}{x^\star - r_\tau}(x^\star - r_\tau) + \mathbb{U}(r_\tau) - \mathbb{U}(l_\tau) \\
&\leqslant \frac{\mathbb{U}(r_\tau) - \mathbb{U}(l_\tau)}{r_\tau - l_\tau}(x^\star - r_\tau) + \mathbb{U}(r_\tau) - \mathbb{U}(l_\tau) = \frac{x^\star - l_\tau}{r_\tau - l_\tau} \cdot \big(\mathbb{U}(r_\tau) - \mathbb{U}(l_\tau)\big) \\
&\leqslant 2 \cdot \big(\mathbb{U}(r_\tau) - \mathbb{U}(l_\tau)\big) \leqslant 2 \cdot \sup(J_{t_\tau - 1}(l_\tau, r_\tau)) \leqslant 2 \cdot |J_{t_\tau - 1}(l_\tau, r_\tau)| \\
&\leqslant 2 \cdot \widetilde{c}_\delta \cdot \sqrt{\frac{1}{B_\tau}} \,,
\end{aligned}$$

where the final inequality follows by Lemma 24.

Let $\tau^\star$ be the first epoch from which it holds $x^\star \in [l_\tau, r_\tau]$. If $\tau^\star \geqslant 2$, then for each $\tau \in \{2, \dots, \tau^\star - 1\}$ it holds that

$$\max_{x \in [l_\tau, r_\tau]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \leqslant \frac{3}{4} \cdot \max_{x \in [l_{\tau-1}, r_{\tau-1}]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \,.$$

In fact, either for all $\tau \in \{1, \dots, \tau^\star - 1\}$ it holds that $r_\tau < x^\star$, or for all $\tau \in \{1, \dots, \tau^\star - 1\}$ it holds that $l_\tau > x^\star$. In the first case, for all $\tau \in \{1, \dots, \tau^\star - 1\}$, leveraging concavity and recalling that $x^\star \in I_\tau$ (which implies $\frac{x^\star - l_\tau}{x^\star - l_{\tau-1}} \leqslant \frac{3}{4}$), we have

$$\begin{aligned}
\max_{x \in [l_\tau, r_\tau]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) = \mathbb{U}(x^\star) - \mathbb{U}(l_\tau) &= \frac{\mathbb{U}(x^\star) - \mathbb{U}(l_\tau)}{x^\star - l_\tau} \cdot (x^\star - l_\tau) \\
&\leqslant \frac{\mathbb{U}(x^\star) - \mathbb{U}(l_{\tau-1})}{x^\star - l_{\tau-1}} \cdot (x^\star - l_\tau) \\
\leqslant \frac{3}{4} \cdot \big(\mathbb{U}(x^\star) - \mathbb{U}(l_{\tau-1})\big) & \\
&= \frac{3}{4} \cdot \max_{x \in [l_{\tau-1}, r_{\tau-1}]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \,,
\end{aligned}$$

while the second case can be deduced analogously.

For each $m \in \mathbb{N}$, let $A_m \coloneqq \{x \in (0,1) : \exists k \in \{1,\dots,2^m-1\}, x = k/2^m\}$ be the dyadic mesh in $(0,1)$ of index m. For any epoch $\tau \in \mathbb{N}$, let $m_\tau \coloneqq -\log_2(c_\tau - l_\tau)$ be the index of the dyadic mesh in $(0,1)$ at epoch τ of Algorithm 8 (note that $m_\tau \geqslant 2$ for all $\tau \in \mathbb{N}$ because Algorithm 8 begins with a step-size of $1/4$).

Let $m^\star \coloneqq \min\{m \in \mathbb{N} : \#\big(A_m \cap (0,x^\star]\big) \geqslant 4 \text{ and } \#\big(A_m \cap [x^\star,1)\big) \geqslant 4\}$ be the smallest index of the dyadic mesh in $(0,1)$ such that there are at least 4 points of the dyadic mesh in $(0,1)$ to the right and to the left of $x^\star$. For each $m \geqslant m^\star$ let $x_1^m < x_2^m < x_3^m < x_4^m \leqslant x^\star$ be the four points of $A_m \cap (0,x^\star]$ closest to $x^\star$ and $x^\star \leqslant x_5^m < x_6^m < x_7^m < x_8^m$ be the four points of $A_m \cap [x^\star,1)$ closest to $x^\star$. Observe that, for all epochs $\tau \geqslant \tau^\star + 3$, Algorithm 8 selects policies only in the closed interval $[x_1^{m_\tau}, x_8^{m_\tau}]$. Observe further that, for each $m \geqslant m^\star + 1$, it holds

$$\max_{x \in [x_1^m, x_8^m]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \leqslant \frac{4}{7} \cdot \max_{x \in [x_1^{m-1}, x_8^{m-1}]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \,.$$

In fact, either $\max_{x \in [x_1^m, x_8^m]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) = \mathbb{U}(x^\star) - \mathbb{U}(x_1^m)$ or $\max_{x \in [x_1^m, x_8^m]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) = \mathbb{U}(x^\star) - \mathbb{U}(x_8^m)$. In the first case, leveraging concavity and observing that $\frac{x^\star - x_1^m}{x^\star - x_1^{m-1}} \leqslant \frac{4}{7}$, we have

$$\begin{aligned}
\max_{x \in [x_1^m, x_8^m]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) = \mathbb{U}(x^\star) - \mathbb{U}(x_1^m) &= \frac{\mathbb{U}(x^\star) - \mathbb{U}(x_1^m)}{x^\star - x_1^m} \cdot (x^\star - x_1^m) \\
&\leqslant \frac{\mathbb{U}(x^\star) - \mathbb{U}(x_1^{m-1})}{x^\star - x_1^{m-1}} \cdot (x^\star - x_1^m) \leqslant \frac{4}{7} \cdot \big(\mathbb{U}(x^\star) - \mathbb{U}(x_1^{m-1})\big) \\
&\leqslant \frac{4}{7} \cdot \max_{x \in [x_1^{m-1}, x_8^{m-1}]} \big(\mathbb{U}(x^\star) - \mathbb{U}(x)\big) \,.
\end{aligned}$$

The second case can be worked out similarly.

Define $\tau^\# \coloneqq \big\lfloor 4 + 2\log_{4/3}(\sqrt{T}) \big\rfloor$ so that

$$\left(\frac{3}{4}\right)^{\left\lfloor \frac{\tau^\# - 1}{2} \right\rfloor} = \left(\frac{3}{4}\right)^{\left\lfloor \frac{\left\lfloor 4 + 2\log_{4/3}(\sqrt{T}) \right\rfloor - 1}{2} \right\rfloor} \leqslant \left(\frac{3}{4}\right)^{\log_{4/3}(\sqrt{T})} = \frac{1}{\sqrt{T}} \,.$$

Assume that $\tau^\# < \tau^\star$ and $\tau^\star + 2 + \tau^\# < \tau_T$ (the other cases can be treated analogously, omitting terms which are not there anymore). Then, the expected regret can be decomposed as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big) &= \sum_{\tau=1}^{\tau^\#} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big) + \sum_{\tau=\tau^\#+1}^{\tau^\star-1} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big) \\
&+ \sum_{\tau=\tau^\star}^{\tau^\star+2} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big) + \sum_{\tau=\tau^\star+3}^{\tau^\star+2+\tau^\#} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big) + \sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star) - \mathbb{U}(x_t)\big).
\end{aligned}$$

We analyze these five terms individually.

For the first one, we further split the sum into two terms, depending on whether or not $B_\tau \coloneqq t_\tau - 1 - t_{\tau-1} \geqslant \widetilde{n}$. Recalling that for each $\tau \in \{1,\dots,\tau_T\}$ and for each $t \in \{t_{\tau-1}+1,\dots,t_\tau\}$

Algorithm 8 selects the policy x_t in the closed interval $[l_\tau, r_\tau]$, we have that

$$\begin{aligned}\sum_{\substack{\tau=1\\B_\tau\geqslant\widetilde{n}}}^{\tau^\#}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big) &\leqslant \sum_{\substack{\tau=1\\B_\tau\geqslant\widetilde{n}}}^{\tau^\#}(B_\tau+1)\cdot\max_{x\in[l_\tau,r_\tau]}\big(\mathbb{U}(x^\star)-\mathbb{U}(x)\big)\\ &\leqslant \sum_{\substack{\tau=1\\B_\tau\geqslant\widetilde{n}}}^{\tau^\#}(B_\tau+1)\cdot 2\cdot\widetilde{c}_\delta\cdot\sqrt{\frac{\log(2/\delta)}{B_\tau}}\\ &\leqslant 4\cdot\widetilde{c}_\delta\cdot\sum_{\substack{\tau=1\\B_\tau\geqslant\widetilde{n}}}^{\tau^\#}\sqrt{B_\tau}\leqslant 4\cdot\widetilde{c}_\delta\cdot\tau^\#\cdot\sqrt{T}.\end{aligned}$$

On the other hand, we also have that

$$\sum_{\substack{\tau=1\\B_\tau\leqslant(\widetilde{n}-1)}}^{\tau^\#}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big)\leqslant(\widetilde{n}-1)\sum_{\tau=0}^{\infty}\big(3/4\big)^\tau=4\cdot(\widetilde{n}-1).$$

Thus, the first term is upper bounded by $4\cdot\widetilde{c}_\delta\cdot\tau^\#\cdot\sqrt{T}+4\cdot(\widetilde{n}-1)$.

For the second term, leveraging the definition of $\tau^\#$, we obtain

$$\begin{aligned}\sum_{\tau=\tau^\#+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big)&\leqslant\sum_{\tau=\tau^\#+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(3/4\big)^{\tau-1}\leqslant\big(3/4\big)^{\tau^\#-1}\cdot\sum_{\tau=\tau^\#+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}1\\ &\leqslant\big(3/4\big)^{\left\lfloor\frac{\tau^\#-1}{2}\right\rfloor}\cdot\sum_{\tau=\tau^\#+1}^{\tau^\star-1}\sum_{t=t_{\tau-1}+1}^{t_\tau}1\leqslant\sqrt{T}.\end{aligned}$$

For the third term, we further split the sum into two terms, depending on whether or not $B_\tau\geqslant\widetilde{n}$. Proceeding exactly as for the first term, we obtain

$$\sum_{\tau=\tau^\star}^{\tau^\star+2}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big)\leqslant 3\cdot 4\cdot\widetilde{c}_\delta\cdot\sqrt{T}+3\cdot(\widetilde{n}-1).$$

For the fourth term, we split again the sum into two terms, depending on whether or not $B_\tau\geqslant\widetilde{n}$. If $B_\tau\geqslant\widetilde{n}$, proceeding exactly as for the corresponding part of the first term, we obtain

$$\sum_{\substack{\tau=\tau^\star+3\\B_\tau\geqslant\widetilde{n}}}^{\tau^\star+2+\tau^\#}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big)\leqslant 4\cdot\widetilde{c}_\delta\cdot\tau^\#\cdot\sqrt{T}.$$

Instead, if $B_\tau\leqslant(\widetilde{n}-1)$, we get

$$\begin{aligned}\sum_{\substack{\tau=\tau^\star+3\\B_\tau\leqslant(\widetilde{n}-1)}}^{\tau^\star+2+\tau^\#}\sum_{t=t_{\tau-1}+1}^{t_\tau}\big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big)&\leqslant(\widetilde{n}-1)\cdot\sum_{\substack{\tau=\tau^\star+3\\B_\tau\leqslant(\widetilde{n}-1)}}^{\tau^\star+2+\tau^\#}\max_{x\in[l_\tau,r_\tau]}\big(\mathbb{U}(x^\star)-\mathbb{U}(x)\big)\\ &\leqslant(\widetilde{n}-1)\cdot\sum_{\substack{\tau=\tau^\star+3\\B_\tau\leqslant(\widetilde{n}-1)}}^{\tau^\star+2+\tau^\#}\max_{x\in\left[x_1^{m_\tau},x_8^{m_\tau}\right]}\big(\mathbb{U}(x^\star)-\mathbb{U}(x)\big)\end{aligned}$$

$$\leqslant 2 \cdot (\widetilde{n}-1) \cdot \sum_{\tau=0}^{\infty} \big(4\!/\!7\big)^{\tau}$$
$$\leqslant \frac{14}{3} \cdot (\widetilde{n}-1)\,.$$

For the last term, we have

$$\sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(\mathbb{U}(x^\star)-\mathbb{U}(x_t)\big) \leqslant \sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} \max_{x\in\left[x_1^{m_\tau},x_8^{m_\tau}\right]} \big(\mathbb{U}(x^\star)-\mathbb{U}(x)\big)$$
$$\leqslant \sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} \big(4\!/\!7\big)^{\left\lfloor \frac{\tau-(\tau^\star+3)-1}{2} \right\rfloor}$$
$$\leqslant \big(3\!/\!4\big)^{\left\lfloor \frac{\tau^\#-1}{2} \right\rfloor} \sum_{\tau=\tau^\star+3+\tau^\#}^{\tau_T} \sum_{t=t_{\tau-1}+1}^{t_\tau} 1 \leqslant \sqrt{T}\,.$$

Putting everything together, and recalling the definition of $\tau^\#$, the conclusion follows. □

C.2.5 Theorem 30 (Upper Bound on Regret of Tempered Exp3 for Optimal Income Taxation)

Proof of Theorem 30.
We prove this result by *reduction to our baseline model*, as analyzed in Section 4.3. Assume that $\mathcal{W}=\{w^1,\dots,w^H\}$ with $0=w^1<w^2<\cdots<w^H\leqslant 1$. For each tax bracket $[w^h,w^{h+1})$, *Tempered Exp3 for Optimal Income Taxation* α essentially reduces to a separate instance of *Tempered Exp3 for Social Welfare*. Denote

$$U_i^h\big(\mathbf{x}(\cdot)\big)=U_i\big(\mathbf{x}(\cdot)\big)\cdot\mathbb{I}\Big\{\lfloor w_i\rfloor=w^h\Big\}, \qquad \boldsymbol{U}_i^h\big(\mathbf{x}(\cdot)\big)=\sum_{j\leqslant i}U_j^h\big(\mathbf{x}(\cdot)\big),$$
$$\boldsymbol{U}_i^h=\sum_{j\leqslant i}U_j^h\big(\mathbf{x}_j(\cdot)\big), \qquad T^h=\sum_{i\leqslant T}\mathbb{I}\Big\{\lfloor w_i\rfloor=w^h\Big\},$$
$$\mathcal{R}_T^h=\sup_{\mathbf{x}(\cdot)\in\mathcal{X}_{\mathcal{W}}}\mathbb{E}\Big[\boldsymbol{U}_T^h\big(\mathbf{x}(\cdot)\big)-\boldsymbol{U}_T^h\Big].$$

It is immediate that

$$R_T\big(\alpha,(v_i,w_i)_{i=1}^T\big)=\sum_h\mathcal{R}_T^h, \text{ and } T=\sum_h T^h.$$

Assume for a moment that the *upper bound* on the regret of Theorem 27 (with λ replaced by 1) *applies to each instance* (tax bracket) h, separately. That is, assume that

$$\mathcal{R}_T^h\leqslant\Big(\gamma+\eta\cdot(e-2)\tfrac{K+1}{K}\cdot\Big(\tfrac{2K+1}{6}+\tfrac{1}{\gamma}\Big)+\tfrac{1}{K}\Big)\cdot T^h+\tfrac{\log(K+1)}{\eta}.$$

Then it follows that

$$R_T\big(\alpha,(v_i,w_i)_{i=1}^T\big)\leqslant\Big(\gamma+\eta\cdot(e-2)\tfrac{K+1}{K}\cdot\Big(\tfrac{2K+1}{6}+\tfrac{1}{\gamma}\Big)+\tfrac{1}{K}\Big)\cdot T+\tfrac{H\cdot\log(K+1)}{\eta},$$

and the claims of Theorem 30 are immediate.

It remains to show that indeed the upper bound on regret of Theorem 27 applies to each instance (tax bracket) h. For any given pair of sequences $\{v_i\}_{i=1}^T, \{w_i\}_{i=1}^T$, consider the subsequence of observations i for which $\lfloor w_i \rfloor = w^h$. Along this subsequence, the policy choice reduces to the choice of a tax rate $x_i = \mathbf{x}_i(w^h) \in \mathcal{X}$, and the algorithm *Tempered Exp3 for Optimal Income Taxation* reduces to an instance of the algorithm *Tempered Exp3 for Social Welfare*, with the following *modifications*:

1. Estimated demand $\widehat{G}_i(x, w^h)$ is multiplied by an additional factor $w_i \in [0, 1]$.
2. Estimated social welfare $\widehat{\boldsymbol{U}}_{i+1}(x, w^h)$ is updated with a term for private welfare that includes a time-varying welfare weight $\omega(w_i) \leqslant 1$, rather than a fixed weight λ.

We need to verify that, with these modifications, the following key claims in the proof of Theorem 27 continue to hold:

1. Unbiasedness: $\widehat{\boldsymbol{U}}_i(x, w^h)$ is an unbiased estimator of $\widetilde{\boldsymbol{U}}_i(x, w^h)$, for a suitably discretized version of cumulative social welfare. (Step 2 of the original proof.) In the present setting, discretization requires substituting $\widetilde{v}_i$ for v_i, where $\widetilde{v}_i = \min\{x \in \mathcal{X} : \ w_i(1-x) \geqslant v_i\}$.
2. Bounded support: $\widehat{U}_i(x, w^h) < \frac{K+1}{\gamma}$, where

$$\widehat{U}_i(x, w^h) = x \cdot \widehat{G}_i(x, w^h) + \frac{\omega(w_i)}{K} \cdot \sum_{x' \in \mathcal{X}, x' > x} \widehat{G}_i(x', w^h).$$

   (Step 4 of the original proof.)

3. Bounded second moment of $\widehat{U}_i(x, w^h)$:

$$\mathbb{E}_i\left[\widehat{U}_i(x, w^h)^2\right] \leqslant \frac{x^2}{p_i(x|w^h)} + \left(\frac{1}{K}\right)^2 \cdot \sum_{x' \in \mathcal{X}, x' > x} \frac{1}{p_i(x'|w^h)},$$

   (Step 6 of the original proof.)

Unbiasedness follows as before. To show bounded support, as well as the bound on the second moment, note that we can rewrite

$$\widehat{U}_i(x, w^h) = \left(x \cdot \mathbb{I}(x_i = x) + \frac{\omega(w_i)}{K} \cdot \mathbb{I}(x_i > x)\right) \cdot \frac{y_i \cdot w_i}{p_i(x_i|w^h)}.$$

Recall that x, $\omega(w_i)$, w_i, and y_i are all bounded above by 1, and that $p_i(x|w^h) \geqslant \frac{\gamma}{K+1}$. Bounded support and the bound on the second moment follow. The remaining steps of the proof are as before. □

Appendix D

Nonstochastic Bandits with Composite Anonymous Feedback

D.1 Comments on the Preliminary Version

Chapter 5 is based on [56], which is an extended and improved version of a preliminary paper that appeared as [53]. In addition to [53], in [56] we provide an analysis of the stability of FTRL with Tsallis Entropy. As a consequence, we are able to shave a logarithmic factor in the regret with respect to the stated guarantees in the preliminary version.

More importantly, there were two *critical* issues in [53].

Firstly, in Cesa-Bianchi et al. [53, last line of Eq. (11)], we find the inequality

$$\mathbb{E}\left[\sum_{s=0}^{d-1}\sum_{i:p_{t-s}(i)>p_{t-d+1}(i)}(p_{t-s}(i)-p_{t-d+1}(i))\right]\leqslant\xi$$

but, whenever an update occurred at round $t-d+2$, the same term is summed $\Theta(d)$ times (rather than 1), leading to a bound of order $\Theta(d\xi)$ rather than the claimed $\Theta(\xi)$, and a consequently looser upper bound for the performance of the wrapper.

Secondly, in Cesa-Bianchi et al. [53, first line of Eq. (10)], we find the inequality

$$\mathbb{E}\left[\sum_{t\in\mathcal{U},t\geqslant 2d-2}\Delta_t^k\right]\geqslant q\big(1-q(2d-1)\big)\sum_{t=2d-2}^{T}\mathbb{E}[\Delta_t^k]$$

which would follow from the provided discussion on $\mathbb{P}'\left[\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right]$ *if* Δ_t^k *were non-negative*, but this is not necessarily the case.

In [56] we propose a different wrapper and patch both things up, as described below.

When analyzing the term corresponding to the first issue in [53], we take a different route. We begin with a change of variables and never upper bound the losses $\ell_{t-s}^{(s)}$ with 1, relying instead on Lemma 12 to obtain the correct dependence on d. This can be seen in the upper bound of (I) in the proof of Theorem 31.

For the second issue, the problem with the original wrapper in [53] is to disentangle Δ_t^k from the random variable $\mathbb{I}\{t \text{ is an update round}\}$. There is, however, an easier way to get around this

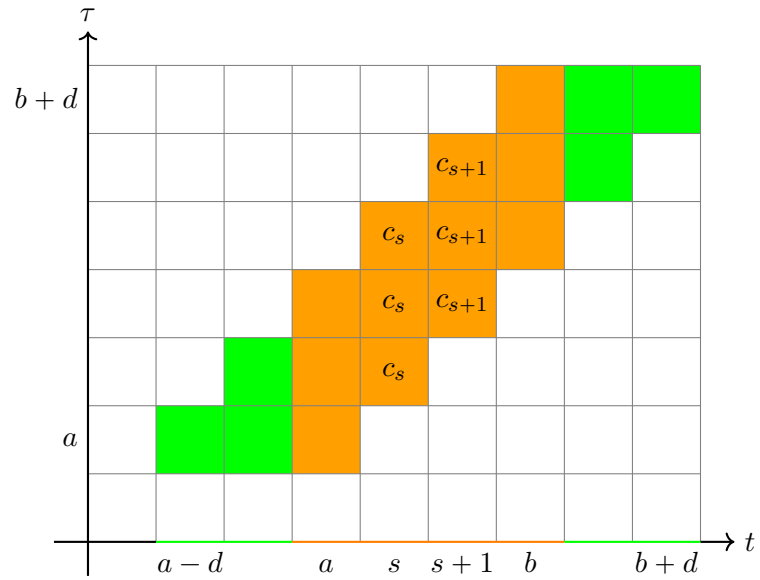


Figure D.1: On the right-hand side of the equation in Lemma 28, we are summing each row of the squares in the picture. On the left hand side, we are summing the columns. Each column s contains the same constant c_s in each component.

roadblock, taking a slightly different route and relying on a different definition of draw, stay, and update rounds (à la [75], see Definition 2 and the subsequent discussion). This greatly simplifies the analysis as can be seen in the upper bound of (II) in the proof of Theorem 31.

D.2 An Accountants' Lemma

The next elementary lemma can be proved straightforwardly by swapping the order of the sums (see Figure D.1).

Lemma 28. *If $(c_t)_{t\in\mathbb{Z}} \subset \mathbb{R}$, $a, b \in \mathbb{Z}$ are such that $a \leqslant b$ and $d \geqslant 0$ then*

$$\sum_{t=a-d}^{a-1} (t-a+d+1)c_t + (d+1)\sum_{t=a}^{b} c_t + \sum_{t=b+1}^{b+d} (b+d+1-t)c_t = \sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t \,.$$

D.3 Stability of FTRL with Tsallis Entropy

In this section, we prove a key stability property of FTRL with Tsallis entropy that could be of independent interest. A general technique [164, Lemma 2.10] to do so for the FTRL family of algorithms is to show that the regularizers are μ-strongly convex with respect to the desired norm (in our case, the ℓ^1-norm). To the best of our knowledge, the existing results in this direction (e.g., Reem et al. 156, Section 7.3) lead to a (probably loose) upper bound on $1/\mu$ of order K, which in turn yields a suboptimal dependence on K in the stability of FTRL with Tsallis entropy. To obtain the correct dependence on K in the regret in Theorem 4, stability of order $\sqrt{K}$ or better is required instead. If one wanted to follow this path, it would therefore be required to prove the tighter upper bound $1/\mu \lesssim \sqrt{K}$, which seems non-trivial. Instead, we take a different route skipping this middle step and controlling the stability of FTRL with Tsallis entropy directly.

We begin by showing that, when $\eta > 0$, there exists a unique solution of the optimization problem that defines $\boldsymbol{q}_n$ in Algorithm 11 and provide a formula for it in terms of the corresponding Lagrange multiplier λ. An analogous result was stated in [193, Section 3.3] for the related algorithm Tsallis-INF.

Lemma 29. *Let $\boldsymbol{w} \in \mathbb{R}^K$ and $\eta > 0$. Then*

$$\exists! \lambda_{\boldsymbol{w},\eta} < \min_{i \in [K]} \boldsymbol{w}(i) \ , \quad \sum_{i \in [K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda_{\boldsymbol{w},\eta}\big)^2} = 1 \ .$$

Furthermore, defining for all $i \in [K]$,

$$\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}(i) \coloneqq \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda_{\boldsymbol{w},\eta}\big)^2} \ ,$$

we have that $\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}$ is the unique global minimizer of the function:

$$\Delta_K \to \mathbb{R} \ , \quad \boldsymbol{q} \mapsto \sum_{i \in [K]} \boldsymbol{w}(i)\boldsymbol{q}(i) - 2\eta \sum_{i \in [K]} \sqrt{\boldsymbol{q}(i)} \ .$$

Proof. Define the two auxiliary functions

$$f_{\boldsymbol{w},\eta} \colon [0,\infty)^K \to \mathbb{R} \ , \qquad \boldsymbol{q} \mapsto \sum_{i \in [K]} \boldsymbol{w}(i)\boldsymbol{q}(i) - 2\eta \sum_{i \in [K]} \sqrt{\boldsymbol{q}(i)}$$
$$\varphi \colon [0,\infty)^K \to \mathbb{R} \ , \qquad \boldsymbol{q} \mapsto \sum_{i \in [K]} \boldsymbol{q}(i)$$

Note that $f_{\boldsymbol{w},\eta}$ is strictly convex and continuous on $[0,\infty)^K$ and differentiable on $(0,\infty)^K$. Thus, by the Lagrange multiplier theorem, a point $\boldsymbol{q} \in (0,\infty)^K$ is the unique global minimizer of $f_{\boldsymbol{w},\eta}$ on the simplex Δ_K if and only if

$$\varphi(\boldsymbol{q}) = 1 \qquad \text{and} \qquad \exists \lambda \in \mathbb{R}, \ \nabla f_{\boldsymbol{w},\eta}(\boldsymbol{q}) = \lambda \nabla \varphi(\boldsymbol{q}) \tag{D.1}$$

A direct verification shows that a pair $(\boldsymbol{q}, \lambda) \in (0,\infty)^K \times \mathbb{R}$ satisfies condition (D.1) if and only if

$$\lambda < \min_{i \in [K]} \boldsymbol{w}(i) \ , \quad \sum_{i \in [K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2} = 1 \ , \quad \text{and} \quad \forall i \in [K] \ , \ \boldsymbol{q}(i) = \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2} \tag{D.2}$$

Note that, letting $m \coloneqq \min_{i \in [K]} \boldsymbol{w}(i)$, the function

$$g \colon (-\infty, m) \to \mathbb{R} \ , \quad \lambda \mapsto \sum_{i \in [K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2}$$

is continuous, strictly increasing, and it satisfies

$$\lim_{\lambda \to -\infty} g(\lambda) = 0 \qquad \text{and} \qquad \lim_{\lambda \to m^-} g(\lambda) = +\infty$$

hence, there exists a unique $\lambda_{\boldsymbol{w},\eta} \in (-\infty, m)$ such that $g(\lambda_{\boldsymbol{w},\eta}) = 1$. Therefore, $\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}$ is the unique

global minimizer of $f_{\boldsymbol{w},\eta}$ on the simplex Δ_K. □

This lemma controls the variation of the Lagrange multipliers corresponding to two points that vary by quantity δ only in a single coordinate.

Lemma 30. *Let $\boldsymbol{w} \in \mathbb{R}^K$ such that $\boldsymbol{w}(1) \leqslant \cdots \leqslant \boldsymbol{w}(K)$ and $\eta > 0$. Then, for all $\delta > 0$ and $j \in [K]$,*

$$\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta} - \lambda_{\boldsymbol{w},\eta} \leqslant \frac{1}{j}\delta \,,$$

where $\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta}$ and $\lambda_{\boldsymbol{w},\eta}$ are defined as in Lemma 29 and $\boldsymbol{e}_1, \dots, \boldsymbol{e}_K$ is the canonical basis of $\mathbb{R}^K$.

Proof. Define $\mathcal{X} \coloneqq \big\{\big(\boldsymbol{u}(1), \dots, \boldsymbol{u}(K), \lambda\big) \in \mathbb{R}^K \times \mathbb{R} \mid \lambda < \min_{i\in[K]} \boldsymbol{u}(i)\big\}$ and

$$\psi \colon \mathcal{X} \to \mathbb{R} \,, \quad \big(\boldsymbol{u}(1), \dots, \boldsymbol{u}(K), \lambda\big) \mapsto \sum_{i\in[K]} \frac{\eta^2}{\big(\boldsymbol{u}(i) - \lambda\big)^2}$$
$$\Lambda \colon \mathbb{R}^K \to \mathbb{R} \,, \quad \boldsymbol{u} \mapsto \lambda_{\boldsymbol{u},\eta}$$

where $\lambda_{\boldsymbol{u},\eta}$ is defined as in Lemma 29. By the implicit function theorem, we have that Λ is infinitely differentiable and for all $\boldsymbol{u} \in \mathbb{R}^K$,

$$\begin{aligned} D_j\Lambda(\boldsymbol{u}) &= -\frac{D_j\psi\big(\boldsymbol{u}(1), \dots, \boldsymbol{u}(K), \lambda_{\boldsymbol{u},\eta}\big)}{D_{K+1}\psi\big(\boldsymbol{u}(1), \dots, \boldsymbol{u}(K), \lambda_{\boldsymbol{u},\eta}\big)} \\ &= -\frac{-2\frac{\eta^2}{(\boldsymbol{u}(j)-\lambda_{\boldsymbol{u},\eta})^3}}{2\sum_{i\in[K]} \frac{\eta^2}{(\boldsymbol{u}(i)-\lambda_{\boldsymbol{u},\eta})^3}} = \frac{1}{1 + \sum_{i\in[K], i\neq j} \left(\frac{\boldsymbol{u}(j)-\lambda_{\boldsymbol{u},\eta}}{\boldsymbol{u}(i)-\lambda_{\boldsymbol{u},\eta}}\right)^3} \,, \end{aligned}$$

where we denoted the partial derivative with respect to the j-th coordinate by D_j. Now, by the fundamental theorem of calculus,

$$\begin{aligned} &\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta} - \lambda_{\boldsymbol{w},\eta} = \Lambda(\boldsymbol{w} + \delta\boldsymbol{e}_j) - \Lambda(\boldsymbol{w}) = \delta \int_0^1 D_j\Lambda(\boldsymbol{w} + s\delta\boldsymbol{e}_j)\,\mathrm{d}s \\ &= \delta \int_0^1 \frac{1}{1 + \sum_{i\in[K], i\leqslant j-1} \left(\frac{\boldsymbol{w}(j)+s\delta-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3 + \sum_{i\in[K], i\geqslant j+1} \left(\frac{\boldsymbol{w}(j)+s\delta-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3} \,\mathrm{d}s \\ &\leqslant \delta \int_0^1 \frac{1}{1 + \sum_{i\in[K], i\leqslant j-1} \left(\frac{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3} \,\mathrm{d}s = \frac{1}{j}\delta \,. \end{aligned}$$

□

We can finally prove the stability of FTRL with $\frac{1}{2}$-Tsallis entropy.

Proof of Theorem 32. Consider an arbitrary sequence of losses $(\ell_m)_{m\in\mathbb{N}} \subset [0,1]$. Fix any $\eta > 0$. For each $\boldsymbol{w} \in \mathbb{R}^K$, let $\lambda_{\boldsymbol{w}} \coloneqq \lambda_{\boldsymbol{w},\eta}$, where $\lambda_{\boldsymbol{w},\eta}$ is defined as in Lemma 29. Let $\mathcal{F}_0$ be the trivial σ-algebra (containing only the sample space and the empty set) and for all $n \in \mathbb{N}$, $\mathcal{F}_n \coloneqq \sigma(J_1, \dots J_n)$. Note that, for each $n \in \mathbb{N}$, we have that $\widehat{L}_{n-1}$ is $\mathcal{F}_{n-1}$-measurable and, as a consequence of Lemma 29, for

all $i \in [K]$,

$$\boldsymbol{q}_n(i) = \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2} \,.$$

Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_K$ be the canonical basis of $\mathbb{R}^K$ and fix any $n \in \mathbb{N}$. Define for each $j \in [K]$, $\widehat{\ell}_{n,j} \coloneqq \frac{\ell_n(j)}{\boldsymbol{q}_n(j)} \boldsymbol{e}_j$ and note that $\widehat{\ell}_n = \widehat{\ell}_{n,J_n}$. To make the notation more compact, we also let $\mathbb{E}_n[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{n-1}]$. Then

$$\begin{aligned}
\mathbb{E}_n\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+\right] &= \mathbb{E}_n\left[\sum_{i\in[K]} \left(\frac{\eta^2}{\big(\widehat{L}_n(i) - \lambda_{\widehat{L}_n}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+\right] \\
&= \mathbb{E}_n\left[\sum_{i\in[K]} \left(\frac{\eta^2}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,J_n}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,J_n}}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+\right] \\
&= \sum_{j\in[K]} \mathbb{E}_n\left[\mathbb{I}\{J_n = j\} \sum_{i\in[K]} \left(\frac{\eta^2}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+\right] \\
&= \sum_{j\in[K]} \mathbb{E}_n\big[\mathbb{I}\{J_n = j\}\big] \sum_{i\in[K]} \left(\frac{\eta^2}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+ \\
&= \sum_{j\in[K]} \boldsymbol{q}_n(j) \sum_{i\in[K]} \left(\frac{\eta^2}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+ = (\star) \,.
\end{aligned}$$

Now, note that, for each $j \in [K]$:

- $\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} \geqslant \lambda_{\widehat{L}_{n-1}}$ (because, as we show in the proof of Lemma 30, the directional derivatives of $\boldsymbol{u} \mapsto \lambda_{\boldsymbol{u}}$ are positive).
- For each $i \in [K]\backslash\{j\}$, $\frac{\eta^2}{\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}} \leqslant \frac{\eta^2}{\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}} = \frac{\eta^2}{\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$ (by the previous point).
- $\frac{\eta^2}{\widehat{L}_{n-1}(j) - \lambda_{\widehat{L}_{n-1}}} \geqslant \frac{\eta^2}{\widehat{L}_{n-1}(j) + \widehat{\ell}_{n,j}(j) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$ (by the previous point and the fact that the following equalities hold $\sum_{i\in[K]} \frac{\eta^2}{\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}} = 1 = \sum_{i\in[K]} \frac{\eta^2}{\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$).

It follows that

$$\begin{aligned}
(\star) &= \sum_{j\in[K]} \boldsymbol{q}_n(j) \sum_{i\in[K], i\neq j} \left(\frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\right) \\
&= \eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j) \sum_{i\in[K], i\neq j} \frac{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2 - \big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2 \big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2} \\
&= \eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j) (\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \sum_{i\in[K], i\neq j} \frac{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big) + \big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2 \big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2} \\
&\leqslant 2\eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j) (\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \sum_{i\in[K], i\neq j} \frac{1}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^3}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant 2\eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \left(\sum_{i\in[K],i\neq j} \frac{1}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} \right)^{3/2} \\
&= 2\eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \left(\sum_{i\in[K],i\neq j} \frac{1}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} \right)^{3/2} \\
&\leqslant 2\eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \left(\sum_{i\in[K]} \frac{1}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} \right)^{3/2} \\
&= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) = (\star\star) \,.
\end{aligned}
$$

Now, let σ be a random permutation of $[K]$ such that $\widehat{L}_{n-1}\big(\sigma(1)\big) \leqslant \cdots \leqslant \widehat{L}_{n-1}\big(\sigma(K)\big)$. Then, using Lemma 30, we have

$$
\begin{aligned}
(\star\star) &= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,\sigma(j)}} - \lambda_{\widehat{L}_{n-1}}) \\
&= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big) \left(\lambda_{\widehat{L}_{n-1}+\frac{\ell_n(\sigma(j))}{\boldsymbol{q}_n(\sigma(j))}\boldsymbol{e}_{\sigma(j)}} - \lambda_{\widehat{L}_{n-1}} \right) \\
&\leqslant \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big) \frac{1}{j} \frac{\ell_n(\sigma(j))}{\boldsymbol{q}_n(\sigma(j))} \leqslant \frac{2}{\eta} \sum_{j\in[K]} \frac{1}{j} \leqslant 2\frac{1+\ln K}{\eta} \,.
\end{aligned}
$$

In conclusion:

$$
\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \mid \mathcal{F}_{n-1} \right] \leqslant 2\frac{1+\ln K}{\eta} \,.
$$

It follows that:

$$
\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \mid \mathcal{F}_{n-1} \right] \right] \leqslant 2\frac{1+\ln K}{\eta} \,.
$$

Being n and $(\ell_m)_{m\in\mathbb{N}}$ arbitrary, the result follows. □

D.4 Stability of Exp3

In this section, we prove the stability of Exp3.

Lemma 31. *Exp3 with learning rate η is ξ-stable with $\xi = \eta$.*

Proof. Consider an arbitrary sequence of losses $(\ell_n)_{n\in\mathbb{N}} \subset [0,1]$. In this case, stability holds pointwise (for all realizations of the actions $J_1, J_2, \ldots$ played by Exp3 on the sequence of losses $(\ell_n)_{n\in\mathbb{N}}$) rather that in expectation. From [51, Lemma 1] we have, for any round $n \in \mathbb{N}$ and all arms $i \in [K]$,

$$
\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i) \leqslant \eta\, \boldsymbol{q}_{n+1}(i) \sum_{j=1}^{K} \boldsymbol{q}_n(j)\widehat{\ell}_n(j)
$$

where $\widehat{\ell}_n(j) = \frac{\ell_n(j)\mathbb{I}\{J_n=j\}}{\boldsymbol{q}_n(j)}$, $\boldsymbol{q}_n(j) = \boldsymbol{w}_n(j)/\sum_{k=1}^{K} \boldsymbol{w}_n(k)$, and if $n = 1$, $\boldsymbol{w}_n(k) = 1$ while, if $n \geqslant 2$,

$\boldsymbol{w}_n(k)$ is defined inductively by $\boldsymbol{w}_n(k) = \boldsymbol{q}_{n-1}(k)e^{-\eta\widehat{\ell}_{n-1}(k)}$. Hence we can write, for any $n \in \mathbb{N}$,

$$\begin{aligned}\sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big) &\leqslant \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \eta\, \boldsymbol{q}_{n+1}(i) \sum_{j=1}^{K} \boldsymbol{q}_n(j)\widehat{\ell}_n(j) \\ &= \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \eta\, \boldsymbol{q}_{n+1}(i)\ell_n(J_n) \leqslant \eta \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \boldsymbol{q}_{n+1}(i) \leqslant \eta\,.\end{aligned}$$

Being $(\ell_n)_{n\in\mathbb{N}}$ arbitrary, the result follows. □

D.5 Lower Bound (Missing Proofs)

We begin this section by showing a reduction mapping each algorithm for bandits with composite anonymous feedback to one for linear bandits with a better or equal regret.

Proof of Lemma 13. Fix an instance $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{T/(d+1)}$ of a linear bandit problem and use it to construct an instance of the d-delayed bandit setting with loss components

$$\ell_t^{(s)}(i) = \begin{cases} \boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{e}_i & \text{if } t+s = 0 \pmod{d+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_K$ are the elements of the canonical basis of $\mathbb{R}^K$. These components define the following composite loss incurred by any algorithm A_d playing actions $I_1, I_2, \dots$

$$\ell_t^\circ(I_{t-d}, \dots, I_t) = \sum_{s=0}^{d} \ell_{t-s}^{(s)}(I_{t-s}) = \begin{cases} (d+1)\,\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t & \text{if } t = 0 \pmod{d+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{q}_t$ is defined from $I_{t-d}, \dots, I_t \in [K]$ as follows

$$\boldsymbol{q}_t(j) = \frac{1}{d+1} \sum_{s=t-d}^{t} \mathbb{I}\{I_s = j\} \qquad j \in [K]. \tag{D.3}$$

Note that $\boldsymbol{q}_t(i)$ is the fraction of times action i was played by A_d in the last $d+1$ rounds. Given the algorithm A_d, we define the algorithm A for playing linear bandits on the loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{T/(d+1)}$ as follows. If $t \neq 0 \pmod{d+1}$, then A skips the round. On the other hand, when $t = 0 \pmod{d+1}$, A performs action $\boldsymbol{q}_t$ defined in (D.3), observes the loss $\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t$, and returns to A_d the composite loss $\ell_t^\circ(I_{t-d}, \dots, I_t)$. Essentially, A_d observes a nonzero composite loss only every d time steps, when $t = 0 \pmod{d+1}$. When this happens, the composite loss of A_d is $d\,\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t$, which is $d+1$ times the loss of A.

Now it is enough to note that, using (5.3),

$$\min_{k=1,\dots,K} \sum_{t=1}^{T} \ell_t^\circ(k, \dots, k) = \min_{k=1,\dots,K} (d+1) \sum_{s=1}^{T/(d+1)} \boldsymbol{\ell}_s^\top \boldsymbol{e}_k = \min_{\boldsymbol{q}\in\Delta_K} (d+1) \sum_{s=1}^{T/(d+1)} \boldsymbol{\ell}_s^\top \boldsymbol{q}\,.$$

This concludes the proof. □

We remark that our lower bound construction relies crucially on the power of the adversary to

plan the assignment of delays: losses of order d are revealed only on T/d time steps, leading to a multiplicative dependence on d. This stacking effect is not possible in settings like the ones studied in [149], where delays are drawn i.i.d. over rounds and are, therefore, *independently spread* across time steps.

We now prove a lower bound for linear bandits.

Proof of Lemma 14. The statement is essentially proven in [166, Theorem 5], where the author shows a $\Omega\big(\sqrt{K/T}\big)$ lower bound on the error of *bandit linear optimization* in the probability simplex.* As explained in [166, Section 1.1], (cumulative) regret lower bounds for linear bandits can be obtained by multiplying the lower bounds on bandit linear optimization error by T. A possible issue is that the proof in [166, Theorem 5] uses unbounded Gaussian losses. However, in [166, Appendix B] it is shown how lower bounds for Gaussian losses can be converted into lower bounds for losses in $[-1, 1]$ at the cost of a $1/\sqrt{\ln T}$ factor in the regret. Finally, note that our setting requires losses in $[0, 1]$, but this is not an issue either because we are in a linear setting, and thus we can add the $(1, \dots, 1)$ constant vector to all loss vectors without affecting the regret. □

*It is worth stressing that the lower bound in [166] is based on stochastic i.i.d. generation of losses, hence it does not violate our assumption about the obliviousness of the adversary.