No-Regret Learning in Bilateral Trade via Global Budget Balance[*]

Martino Bernasconi
Bocconi University
Milan, Italy
martino.bernasconi@unibocconi.it

Matteo Castiglioni
Politecnico di Milano
Milan, Italy
matteo.castiglioni@polimi.it

Andrea Celli
Bocconi University
Milan, Italy
andrea.celli2@unibocconi.it

Federico Fusco
Sapienza University of Rome
Rome, Italy
fuscof@diag.uniroma1.it

ABSTRACT

Bilateral trade models the problem of intermediating between two rational agents — a seller and a buyer — both characterized by a private valuation for an item they want to trade. We study the online learning version of the problem, in which at each time step a new seller and buyer arrive and the learner has to set prices for them without any knowledge about their (adversarially generated) valuations.

In this setting, known impossibility results rule out the existence of no-regret algorithms when budget balanced has to be enforced at each time step. In this paper, we introduce the notion of *global budget balance*, which only requires the learner to fulfill budget balance over the entire time horizon. Under this natural relaxation, we provide the first no-regret algorithms for adversarial bilateral trade under various feedback models. First, we show that in the full-feedback model, the learner can guarantee $\tilde{O}(\sqrt{T})$ regret against the best fixed prices in hindsight, and that this bound is optimal up to poly-logarithmic terms. Second, we provide a learning algorithm guaranteeing a $\tilde{O}(T^{3/4})$ regret upper bound with one-bit feedback, which we complement with a $\Omega(T^{5/7})$ lower bound that holds even in the two-bit feedback model. Finally, we introduce and analyze an alternative benchmark that is provably stronger than the best fixed prices in hindsight and is inspired by the literature on bandits with knapsacks.

CCS CONCEPTS

• **Theory of computation** → **Algorithmic game theory**; • **Computing methodologies** → *Online learning settings.*

KEYWORDS

Online Learning, Bilateral Trade, Budget Balance, Partial Feedback

ACM Reference Format:
Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. 2024. No-Regret Learning in Bilateral Trade via Global Budget Balance. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24), June 24–28, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3618260.3649653

1 INTRODUCTION

Bilateral trade is a classic economic problem where two agents — a seller and a buyer — are interested in trading a good. Both agents are characterized by a private valuation for the item, and their goal is to maximize their own utility. Solving this problem requires the design of a mechanism that intermediates between the two parties, facilitating the trade. Ideally, the mechanism should maximize efficiency (*i.e.,* trade whenever the buyer's valuation exceeds the seller's one) while ensuring that agents behave according to their true preferences (*incentive compatibility*), and that the utility for participating in the mechanism of each agent is non-negative (*individual rationality*). These properties ensure favorable outcomes for the agents, yet they do not guarantee the economic viability of the mechanism. To see this, consider the following mechanism $\mathcal{M}$. $\mathcal{M}$ asks the agents for their valuations, s for the seller and b for the buyer, and makes the trade happen if it is convenient (*i.e.,* if $s \leq b$). In case of a trade, $\mathcal{M}$ then charges s to the buyer and pays b to the buyer. It is not hard to see that $\mathcal{M}$ enforces incentive compatibility and individual rationality, and is efficient by design. However, it exhibits the major drawback of allowing the intermediary to incur a net loss when $b > s$. To avoid such situations, a crucial constraint in bilateral trade is *budget balance*, which restricts the mechanism from subsidizing the agents.

As highlighted by the above example, an incentive compatible mechanism maximizing efficiency for bilateral trade may not be budget balanced. This phenomenon was first observed by Vickrey [49]; subsequently Myerson and Satterthwaite [43], provided a more general impossibility result by showing the existence of instances where a fully efficient mechanism that satisfies incentive compatibility, individual rationality, and budget balance does not exist. This result holds even when probabilistic information on the agents' valuations is available. To circumvent these impossibility results, the extensive subsequent research primarily focuses on finding approximately efficient mechanisms in the Bayesian setting. There, various incentive compatible mechanisms exist that give a constant-factor approximation to the social welfare (see, *e.g.,* Blumrosen and Dobzinski [11], Kang et al. [34], while more recent works also consider the harder problem of approximating the gain from trade [12, 14, 25, 28, 41]. While the Bayesian assumption of having perfect knowledge about the underlying distributions of valuations

[*] The full version can be found on the arXiv [8].

STOC '24, June 24–28, 2024, Vancouver, BC, Canada
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0383-6/24/06
https://doi.org/10.1145/3618260.3649653

is, in some sense, necessary for extracting meaningful approximations to the social welfare [27], it is important to observe that this assumption is oftentimes unrealistic.

Following the recent line of work initiated by Cesa-Bianchi et al. [18], we study this fundamental mechanism design problem through the lens of regret minimization in a repeated setting where at each time t, a new seller/buyer pair arrives. The seller arriving at time t has a private valuation s_t representing the lowest price they are willing to accept for the item. Analogously, the buyer has a private valuation b_t representing the highest price they are willing to pay for the item. The learner, without any knowledge about the private valuations at the current time t, posts two (possibly randomized) prices: p_t to the seller and q_t to the buyer. A trade happens when both agents agree to trade, i.e., when $s_t \le p_t$ and $q_t \le b_t$. After posting (p_t, q_t), the learner observes some feedback about the transaction, and is awarded the *gain from trade*:

$$\mathrm{GFT}_t(p,q) = \mathbb{I}\{s_t \le p\}\mathbb{I}\{q \le b_t\}(b_t - s_t).$$

The goal of the learner is to maximize the overall gain from trade or, equivalently, minimize the regret with respect to the best price in hindsight. Prior research has investigated the impact of different budget balance notions on the problem's learnability. When the mechanism is constrained to enforce per-round *strong budget balance* (*i.e., $p_t = q_t$* at each time step t), it is possible to attain sublinear regret only when the sequence of valuations is drawn i.i.d. from some fixed unknown distribution, and the learner has either full feedback, or some stringent assumptions regarding the sequence of valuations are enforced. Specifically, in partial feedback regime, valuations have to be drawn i.i.d. from a smooth distribution, independently for the seller and the buyer [18, 20]. If the learner is only required to enforce (step-wise) *weak budget balance* (*i.e., $p_t \le q_t$* for each t), then Azar et al. [4] provide a learning algorithm achieving sublinear 2-regret when the sequence of valuation is generated by an oblivious adversary.[1] They also show that this result is tight: no algorithm can achieve sublinear $(2-\varepsilon)$-regret in the adversarial case, for any constant $\varepsilon > 0$. In an attempt to overcome this barrier, Cesa-Bianchi et al. [19] show that sublinear regret can be achieved beyond the i.i.d. stochastic setting, under the assumption that the adversary is constrained to choose randomized (possibly non-stationary) sequences of valuations that are not "too concentrated" (*i.e.,* under a σ-smooth adversary model). Inspired by the positive results obtained in the literature by transitioning from strong to weak budget balance, we investigate the following natural open question:

Is it possible to achieve sublinear regret against an oblivious adversary in the repeated bilateral trade problem under a realistic notion of budget balance?

We answer this question positively by introducing *global budget balance*, where the learner is required to maintain budget balance only "overall". The idea behind global budget balance is to allow the learner to reinvest the profit gained in previous rounds (obtained by posting a lower price for the seller compared to the buyer), with the constraint that the learner cannot subsidize the market *over the whole time horizon.* Formally, a learning algorithm that posts prices $(p_1,q_1),(p_2,q_2),\dots$ is global budget balanced if the following inequality holds almost surely: $\sum_{t=1}^{T} \textsc{Profit}_t(p_t,q_t) \ge 0$. The profit $\textsc{Profit}_t(p_t,q_t) = \mathbb{I}\{s_t \le p_t\}\mathbb{I}\{q_t \le b_t\}(q_t - p_t)$ is nonnegative when $p_t \le q_t$, and may drop below zero only by posting prices that are not step-wise budget balanced, *i.e., $p_t > q_t$*. We argue that this constraint is more realistic than the restrictive notions of per-round budget balance. For instance, in contexts like ride-hailing platforms (such as Uber and Lyft), the platform might opt to forego some short-term profit to enhance other metrics, like the overall welfare of the system.

1.1 Overview of Our Results

We report here an overview of our results, we also refer to Table 1 for a comparison with the state of the art. In this paper we introduce the notion of global budget balance for the repeated bilateral trade problem, and provide the following results in terms of regret with respect to the best fixed price in hindsight in the adversarial case:

- In the full feedback model, when the learner observes seller and buyer valuations after posting prices, we design a learning algorithm characterized by a $\tilde{O}(T^{1/2})$ regret upper bound (Theorem 4.2). We also prove that no learning algorithm can improve this bound by more than a poly-log T factor (Theorem 4.3).
- In the *one-bit feedback* model, where the learner can observe only whether the trade happened or not, we show that it is possible to guarantee a $\tilde{O}(T^{3/4})$ regret upper bound (Theorem 5.4). Then, we provide an $\Omega(T^{\frac{5}{7}\approx 0.714})$ lower bound, which holds even in the *two-bit feedback* model, where the learner can observe which agent accepted and who declined the offered prices (Theorem 5.5).

These results demonstrate how the notion of global budget balance enables online learnability, allowing us to provide the first no-regret algorithms for repeated bilateral trade within an oblivious adversary framework, in contrast to the per-round approaches considered in previous works. Furthermore, the regret rates separate full feedback and the two partial feedback models (one or two bits). In partial feedback, the surprising lower bound of $\Omega(T^{5/7})$, together with the $O(T^{3/4})$ upper bound, mark a clear separation between this problem and other partial feedback models (*e.g.,* partial monitoring [7] and online learning with feedback graph [2], where the minimax regret have been characterized to fall in one of three admissible rates: $\sqrt{T}$, $T^{2/3}$ and T). This separation had already been hinted at in the special case of σ-smooth adversary by Cesa-Bianchi et al. [19].

Finally, inspired by work on bandits with knapsacks (see Section 1.3 for detailed references), we introduce a stronger learning benchmark: the best fixed feasible distribution over prices. Such benchmark is allowed to post prices that are not per-round budget balanced, but is global budget balanced in "expectation".

- We show that there exists a constant $\varepsilon_0 > 0$ such that it is impossible to achieve sublinear α-regret against this benchmark for any $\alpha \in [1, 1+\varepsilon_0)$ (Theorem 6.2).
- We prove that the best feasible distribution over prices collects at most twice the gain from trade extracted by the best

[1] The α-regret measures the difference between the gain from trade of the best fixed price in hindsight and α times that of the algorithm (see e.g., Kakade et al. [33]).

Table 1: Comparison of prior results on bilateral trade. The positive result for a stochastic adversary in the partial feedback, marked with an asterisk ($*$), holds under the assumption that the seller and buyer valuations are drawn independently from smooth distributions. All the bounds in the second row (Azar et al. [4]), marked with a dagger (†), apply to 2-regret.

	Type of Adversary	Budget Balance	Regret Upper Bounds	Regret Lower Bounds
Cesa-Bianchi et al. [18]	stochastic setting	strong	• Full: $\tilde{O}(T^{1/2})$ • Partial: $\tilde{O}(T^{2/3})^*$	• Full: $\Omega(T^{1/2})$ • Partial: $\Omega(T^{2/3})$
Azar et al. [4]	adversarial setting	weak	• Full: $\tilde{O}(T^{1/2})^\dagger$ • Partial: $\tilde{O}(T^{3/4})^\dagger$	• Full: $\Omega(T^{1/2})^\dagger$ • Partial: $\Omega(T^{2/3})^\dagger$
Cesa-Bianchi et al. [19]	σ-smooth adversary	weak	• Full: $\tilde{O}(T^{1/2})$ • Partial: $\tilde{O}(T^{3/4})$	• Full: $\Omega(T^{1/2})$ • Partial: $\Omega(T^{3/4})$
This paper	adversarial setting	**global**	• Full: $\tilde{O}(T^{1/2})$ • Partial: $\tilde{O}(T^{3/4})$	• Full: $\Omega(T^{1/2})$ • Partial: $\Omega(T^{5/7 \approx 0.714})$

fixed price in hindsight (Theorem 6.3). This implies the existence of algorithms with sublinear 2-regret against this new benchmark.

- We show that the multiplicative gap of 2 between the gain from trade attainable by the two different benchmarks is tight (Theorem 6.5).

First, we observe that the task of learning the best feasible distribution over prices is reminiscent of the problem of bandits with knapsacks in the presence of replenishment [9, 37, 47]. In contrast to previous work, we consider the more challenging adversarial setting and provide learning algorithms with a competitive ratio that is an absolute constant. In the adversarial bandits with knapsacks literature, the only setting where sublinear $\Theta(1)$-regret can be achieved is when the available budget is $\Omega(T)$ [15], while in general the competitive ratio is $O(\log T)$ [31]. Second, the tight multiplicative gap of 2 between the two benchmarks suggests that to design a better learning algorithm with sublinear α-regret with respect to the best feasible distribution (for $\alpha \in (1+\varepsilon_0, 2)$), a more direct approach is needed.

1.2 Challenges and Techniques

The key aspects that distinguish bilateral trade from standard online learning models with full or bandit feedback can be identified in two main features: the action space and the challenging partial feedback structure. The applicability of previous results to our model is significantly limited due to adversarial input sequences and the need to handle the global budget balance constraint effectively.

Action space. The action space is continuous and bidimensional (prices belong to $[0,1]^2$), and neither the gain from trade nor the profit functions are continuous in the prices posted. This makes it challenging to discretize the space with a finite grid G such that the best prices in G perform similarly to the best prices in $[0,1]^2$, and such that grid G is small enough that it is possible to learn in an online way its best pair of prices. In the absence of any probabilistic or smoothness assumption on the adversary, we cannot rely on a "smoothing trick" to induce regularity on the expected gain from trade, as in previous works [19].

Partial Feedback. Partial feedback models for bilateral trade are inherently challenging. The one-bit feedback model only informs the learner on whether the trade happened or not, which is significantly less informative than the traditional bandit feedback model, since the learner cannot even reconstruct the gain from trade received for the specific prices it posted. For example, if the learner posts price $1/2$ to both agents, and they accept the trade, there is no way of distinguishing between the case in which the gain from trade is constant (*e.g.,* valuations are $(0,1)$) from the case in which the gain from trade is arbitrarily small (*e.g.,* valuations are $(1/2-\varepsilon, 1/2+\varepsilon)$ for some small ε). On the other hand, if one of the two agents rejects the trade, then the learner can only infer loose bounds on the valuations.

Gain from Trade vs. Profit trade-off. Global budget balance requires that the cumulative sum of profits at the end of the time horizon must be greater than or equal to 0. Therefore, the learner has to maximize its cumulative gain from trade, while accumulating enough profit to enforce global budget balance. Balancing this trade-off is a complex task due to the different nature of the two objectives: gain from trade is maximized by setting identical prices for both agents, whereas profit is maximized by selecting prices that are "far from each other". To see this, consider an instance where valuations are either $(s_t, b_t) = (0,1)$ or $(s_t, b_t) = (1/2-\varepsilon, 1/2+\varepsilon)$ with equal probability, for some small $\varepsilon > 0$. To achieve maximum expected profit, the learner would always set the price at 0 for the seller and 1 for the buyer. On the other hand, to maximize the expected gain from trade, the learner would always offer $1/2$ to both agents.

Our Two-Phase Approach. Our learning algorithms follow a two-phase approach, initially focusing on maximizing profit through a carefully designed multiplicative grid F_K of candidate prices and then switching to maximizing gain from trade on a different (additive) grid H_K of non-budget-balanced prices. At a high level, the first phase is used to collect budget, which can be subsequently reinvested in the second phase. This poses several challenges due to the non-stationary nature of the adversary. The pairs of prices in H_K, which are not per-round budget balanced, enable the algorithm to circumvent the negative results that hinder discretization in scenarios with per-round budget balance (see, *e.g.,* , the "needle in a haystack" phenomenon in Theorem 7 of Cesa-Bianchi et al.

[20]). The multiplicative nature of the grid F_K is crucial in ensuring that the gain from trade accrued by the algorithm during the first phase does not yield too much regret. This last result is surprising since, in the first phase, the learning algorithm is maximizing profit, an objective that is inherently orthogonal to the gain from trade. Finally, the scarcity of feedback in the one-bit feedback model is addressed via a carefully designed estimation technique that allows the learner to estimate the gain from trade in one point of the grid H_K posting two different prices. In contrast to the technique by Azar et al. [4], our procedure is "asymmetric" in how it deals with the seller and buyer, and it provides biased estimates.

Lower bounds. Besides the typical challenges in proving lower bounds for repeated bilateral trade with respect to the best fixed price in hindsight, in our model the agent is allowed to post prices that are not per-round budget balanced (*i.e.,* it may be the case that $p_t > q_t$). This considerably complicates the construction of the hard instances, as any algorithm could sacrifice temporarily some profit by posting prices with $p_t > q_t$ to extract a large gain from trade (that the fixed price benchmark may not be able to obtain). To deter this kind of behavior, we incorporate into the hard instances certain unfavorable trade opportunities that dissuade the learner from setting prices that are not budget balanced. This additional complication comes at some cost: in the partial (two-bit) feedback model we recover a lower bound of $\Omega(T^{5/7})$, whereas the corresponding lower bound by Cesa-Bianchi et al. [19] is $\Omega(T^{3/4})$.

1.3 Further Related Works

Online Learning and Economics. Regret minimization techniques have found applications across different domains motivated by economics, with the goal of overcoming unrealistic assumptions. For example, they have been applied to one-sided pricing [29, 36], auctions [6, 21, 22, 24, 38, 42, 44, 50], contract design [26, 30, 52], brokerage [13], and Bayesian persuasion [10, 16, 17, 53].

Partial feedback. Repeated bilateral trade naturally involves challenges due to partial feedback. Therefore, our work aligns with the research that explores online learning with feedback models beyond the conventional full feedback and bandit models. Our one- and two-bit feedback models share similarities with *graph-structured feedback* [2] and with the *partial monitoring* framework [7, 23].

Bandits with knapsacks. Another related line of work is that of online learning under long-term constraints. Some works study the case of static constraints and develop projection-free algorithms with sublinear regret and constraint violations [32, 39], while others study the case of time-varying constraints [40, 48, 51]. Badanidiyuru et al. [5] introduced and solved the (stochastic) bandits with knapsacks (BwK) framework, in which they consider bandit feedback and stochastic objective and cost functions. In this model, the learner's objective is to maximize utility while guaranteeing that, for each of the m available resources, cumulative costs are below a certain budget B. Other optimal algorithms for stochastic BwK were proposed by Agrawal and Devanur [1], Immorlica et al. [31]. The setting with adversarial inputs was first studied in Immorlica et al. [31], where the baseline considered is the best fixed distribution over arms. Achieving no-regret is not possible under this baseline and, therefore, they provide no-α-regret guarantees for their algorithm. If we denote by ρ the per-iteration budget of the learner, the best-known guarantees on the competitive ratio α are $1/\rho$ in the case in which $B = \Omega(T)$ [15], and $O(\log m \log T)$ in the general case [35]. When considering a benchmark similar to the adversarial BwK scenario, we show that our algorithm ensures a $\alpha = 2$ guarantee. Kumar and Kleinberg [37] recently proposed a generalization of the stochastic BwK model in which resource consumption can be non-monotonic; that is, resources can be replenished or renewed over time. Our model also admits replenishment. It should be noted that, in our setting, directly utilizing techniques from BwK is not feasible due to the complex continuous action space and the limited availability of feedback, which is less informative compared to traditional bandit feedback.

Learning Protocol of Repeated Bilateral Trade

```
Initial budget $B_0 = 0$
for $t = 1, 2, \dots$ do
    The adversary privately chooses $(s_t, b_t)$ in $[0, 1]^2$
    The learner posts prices $(p_t, q_t) \in [0, 1]^2$ such that
      $p_t - q_t \leq B_t$
    The learner receives a (hidden) reward
      $\mathrm{GFT}_t(p_t, q_t) \in [-1, 1]$
    The budget of the learner is updated
      $B_t \leftarrow B_{t-1} + \textsc{Profit}_t(p_t, q_t)$
    Feedback $z_t$ is revealed to the learner
```

2 REPEATED BILATERAL TRADE

We study repeated bilateral trade problem in an online learning setting, where the learner has to enforce global budget balance and the sequence of valuations is generated by an oblivious adversary.

The learning protocol. The learner repeatedly interacts with the environment according to the following protocol (see also pseudocode). At each time step t, a new pair of buyer and seller arrives, characterized by valuations $b_t \in [0, 1]$ and $s_t \in [0, 1]$, respectively. Without knowing s_t and b_t, the learner posts two prices: $p_t \in [0, 1]$ to the seller, and $q_t \in [0, 1]$ to the buyer. If both the seller and the buyer accept (*i.e.,* $s_t \leq p_t$ and $q_t \leq b_t$), then the learner is awarded the gain from trade

$$\mathrm{GFT}_t(p_t, q_t) = \mathbb{I}\{s_t \leq p_t\}\mathbb{I}\{q_t \leq b_t\}(b_t - s_t),$$

that corresponds to the increase in social welfare generated by the trade. To simplify the notation, we omit the second argument of GFT_t (and of $\textsc{Profit}_t$) when the same price is posted to both agents. After posting the prices, the learner does *not* observe directly the gain from trade or the valuations, but receives some feedback z_t.

Global budget balance. For each time step t, the notion of *profit* of the learner is naturally defined: if the agents accept prices p_t and q_t, then the learner receives a net profit of $q_t - p_t \in [-1, 1]$. Unlike the case of the gain from trade, the learner naturally knows its profit at the end of each time step, as it sets the prices and always observes whether the trade occurred. The learner maintains a budget B_t, which is initially 0 ($B_0 = 0$) and is updated at each time step according to the profit generated or consumed: $B_t \leftarrow B_{t-1} + \textsc{Profit}_t(p_t, q_t)$. We restrict the learner to enforce a *global budget balance* property which states that the final budget B_T has to be non-negative with probability 1. In practice, we require the learner to always post prices p_t, q_t such that $(p_t - q_t) \le B_{t-1}$.[2]

Feedback models. In this paper, we study three feedback models, that we list here in increasing order of intricacy:

- *Full feedback*: at the end of each round, the agents reveal their valuations (*i.e.,* $z_t = (s_t, b_t)$).
- *Two-bit feedback*: the agents only reveal their willingness to accept the prices offered by the learner (*i.e.,* z_t is composed by the two bits $(\mathbb{I}\{s_t \le p_t\}, \mathbb{I}\{q_t \le b_t\})$)
- *One-bit feedback*: the learner only observes whether the trade happened or not (*i.e.,* $z_t = \mathbb{I}\{s_t \le p_t\} \cdot \mathbb{I}\{q_t \le b_t\}$).

These feedback models are not only interesting from the theoretical learning perspective, but they are also well motivated in terms of practical applications. The full-feedback model can be used to describe sealed-bid-type auctions, while the two partial feedback settings (one- and two-bit) enforce the desirable property (for the agents) of revealing a minimal amount of information to the learner.

Regret with respect to the best fixed price. The goal is to maximize the total gain from trade on a fixed and known time horizon T while enforcing the global budget balance condition. Following the literature on repeated bilateral trade [18], we measure the performance of a learning algorithm in terms of its regret with respect to the best fixed price(s) in hindsight. For any learning algorithm $\mathcal{A}$ and sequence of valuations $\mathcal{S} = \{(s_t, b_t)\}_{t=1}^T$ we define:

$$R_T(\mathcal{A}, \mathcal{S}) = \max_{\substack{(p,q)\in[0,1]^2\\ p\le q}} \sum_{t=1}^T \mathrm{GFT}_t(p,q) - \mathbb{E}\left[\sum_{t=1}^T \mathrm{GFT}_t(p_t,q_t)\right], \quad (1)$$

where the sequence $\mathcal{S}$ induces the GFT_t functions and the expectation is with respect to (possibly) randomized prices p_t and q_t generated by the learning algorithm $\mathcal{A}$. One simple property that follows immediately by definition is that, for any sequence of valuations, there exists a fixed pair of identical prices that maximizes the gain from trade. This means that the notion of "best price in hindsight" is well defined, and confirms the intuition that posting two different prices only helps during learning, but does not impact the maximization of gain from trade in hindsight. Finally, we define the regret of an algorithm $\mathcal{A}$ (without the dependence on a specific sequence of valuations) as its worst-case performance: $R_T(\mathcal{A}) = \sup_{\mathcal{S}} R_T(\mathcal{A}, \mathcal{S})$, where the sup is over the set of all the possible sequences of T pairs of valuations.

A stronger benchmark: the best feasible distribution over prices. In this paper we also introduce a new (stronger) benchmark for the study of repeated bilateral trade: the best fixed budget-feasible distribution over prices. This benchmark captures the flexibility of the global budget balance condition, and it arises naturally from the literature on *bandits with knapsacks.* Before proceeding with the definition, let $\Delta([0,1]^2)$ be the family of all the probability measures over the measurable space $([0,1]^2, \mathcal{B}([0,1]^2))$, where $\mathcal{B}$ denotes the Borel σ-algebra.

Definition 2.1 (Best feasible distribution)**.** For any sequence $\mathcal{S}$ of seller's and buyer's valuations, we define the best fixed budget-feasible distribution over prices as the solution of:

$$\sup_{\gamma} \mathbb{E}_\gamma\left[\sum_{t=1}^T \mathrm{GFT}_t(p,q)\right] \text{ s.t. } \mathbb{E}_\gamma\left[\sum_{t=1}^T \textsc{Profit}_t(p,q)\right] \ge 0, \quad (2)$$

where $\mathbb{E}_\gamma$ denotes that the expectation is with respect to prices (p,q) sampled according to γ.

3 PRICE DISCRETIZATIONS AND TWO-PHASE ALGORITHM

In this section we present our two-phase meta algorithm, preceeded by two key results on how to discretize the price space in a way that ensures certain essential properties about profit and gain from trade. First, in Section 3.1 we prove that the gain from trade of the best fixed price in hindsight is close to that of the best pair of (non-budget-balanced) prices on a suitable "additive" grid. Second, in Section 3.2 we construct an hybrid "multiplicative-additive" grid in which each interval of a one-dimensional additive grid is further divided into sub-intervals with geometrically decreasing length. This grid has the surprising property that the profit of the best fixed pair of prices on it is close to the gain from trade generated by the best fixed price in the $[0,1]$ interval, up to a poly-logarithmic multiplicative factor. Finally, we introduce our two-phase learning via the meta-algorithm GFT-Max.

3.1 Additive Grid for Gain from Trade

For any integer K, we denote by $G_K = \{0, 1/K, 2/K, \dots, 1 - 1/K, 1\}$ the *K-uniform grid* over $[0,1]$. Similarly, we denote with $H_K = \{(i{+}1/K, i/K) : i \in \{0,1,\dots,K-1\}\}$ the set of pairs formed by contiguous points in the K-uniform grid such that the first element of the pair is greater than the second. This latter grid can be proved to enjoy the desirable property of well-approximating the gain from trade of the best fixed price, while violating the global budget balance condition by a small amount. The argument behind the approximation guarantee is simple: if p^* is the best fixed price in hindsight, then the pair of prices $((i{+}1)/K, i/K)$ such that p^* belongs to the interval $[i/K, (i{+}1)/K]$ are nearly as good as p^*. We have the following result; its proof, and all the missing ones in the rest of the paper, can be found in the full version [8].

Proposition 3.1. *For any K and sequence of valuations, we have:*

$$\max_{p\in[0,1]} \sum_{t=1}^T \mathit{GFT}_t(p) \le \max_{(p,q)\in H_K} \sum_{t=1}^T \mathit{GFT}_t(p,q) + \frac{T}{K}.$$

For any $(p,q) \in H_K$, total profit $\sum_{t=1}^T \textsc{Profit}_t(p,q)$ is at least $-T/K$.

A simple calculation shows that $\mathrm{GFT}_t((i{+}1)/K, i/K)$ is bounded by the sum of $\mathrm{GFT}_t(i/K)$ and $\mathrm{GFT}_t((i{+}1)/K)$. Therefore, we obtain the following known result as a Corollary to Proposition 3.1.

[2] In fact, this condition is not just sufficient, but also necessary. Indeed, if $p_t - q_t > B_t$, the adversary might select valuations (s_t, b_t) such that $\textsc{Profit}_t(p_t, q_t) < -B_{t-1}$ and thus $B_t < 0$. After that, the adversary might select valuations $(s_\tau, b_\tau) = (0,0)$ for all $\tau \ge t+1$, thereby forcing $B_T = B_t < 0$.

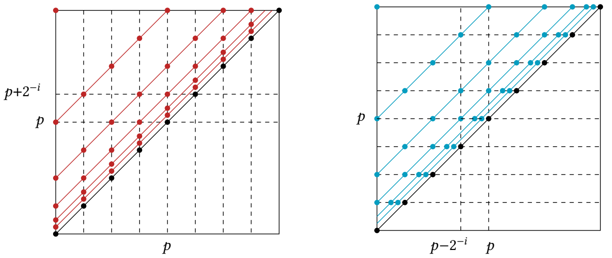

Figure 1: F_K^+ (left) and F_K^- (right) for $K = 8$, $T = 32$, $i = 3$.

Corollary 3.2 (Claim 1 of Azar et al. [4]). *For any K and sequence of valuations, we have:*

$$\max_{p\in[0,1]} \sum_{t=1}^{T} GFT_t(p) \le 2 \cdot \max_{p\in G_K} \sum_{t=1}^{T} GFT_t(p) + \frac{T}{K}.$$

3.2 Multiplicative Grid for Profit

For any K, we construct the two-dimensional grid F_K starting from the points on the one-dimensional grid G_K. For each $p \in G_K$, we add to F_K points of the form $(p - 2^{-i}, p)$ and $(p, p + 2^{-i})$, for $i = 0, 1, \ldots, \lceil \log T \rceil$ so that they define intervals of geometrically decreasing length to the left and upward of (p, p). Formally, we define F_K as the union of F_K^- and F_K^+ (see also Figure 1):

$$F_K^- = \left\{(p - 2^{-i}, p) : p \in G_K \text{ and } i \in \{0, 1, \ldots, \lceil \log T \rceil\}\right\} \cap [0, 1]^2,$$
$$F_K^+ = \left\{(p, p + 2^{-i}) : p \in G_K \text{ and } i \in \{0, 1, \ldots, \lceil \log T \rceil\}\right\} \cap [0, 1]^2.$$

The additive-multiplicative nature of F_K endows it with two crucial properties: (i) its cardinality is $O(K \log T)$ an thus only depends linearly in K and (ii) the profit of the best prices in F_K is at least a $O(\log T)$ fraction of the GFT at the best fixed price in $[0, 1]$, up to an additive factor of $O(T/K)$.

Proposition 3.3. *For any K and sequence of valuations, we have:*

$$\max_{p\in[0,1]} \sum_{t=1}^{T} GFT_t(p) \le 12 \log T \cdot \max_{(p,q)\in F_K} \sum_{t=1}^{T} \textsc{Profit}_t(p, q) + \frac{5T}{K}.$$

3.3 Our Two-Phase Meta-Algorithm: GFT-Max

We describe our two-phase learning approach by presenting the meta-algorithm GFT-Max. For details, we refer to the pseudocode. The algorithm takes in input a budget threshold β and an integer K (which induces the two grids F_K and H_K) and employs two regret minimizers—$\mathcal{A}_{\text{P}}$ for the profit and $\mathcal{A}_{\text{G}}$ for the gain from trade—as internal routines. In the first phase (Line 1), the algorithm uses function Profit-Max to maximize profit until the collected budget reaches a given threshold β. This is achieved by running a regret minimizer $\mathcal{A}_{\text{P}}$ over the set F_K of pairs of prices (see Section 3.2) using profit as objective. Then, in the second phase (from Line 2 onward), the algorithm exploits a regret minimizer $\mathcal{A}_{\text{G}}$ to maximize the gain from trade over the grid H_K, whose prices which are "almost budget-balanced" and consume only a small fraction of the previously acquired budget (see Proposition 3.1). In Section 4 and Section 5 we provide regret upper bounds for this meta-algorithm in the full and one-bit feedback model, respectively. The budget threshold β, the regret minimizers, and the grid parameter K are tuned according to the specific case considered.

Algorithm 1: GFT-Max

```
   Input: • budget threshold $\beta$
          • integer $K$ and price-grids $F_K$ and $H_K$
          • regret minimizers $\mathcal{A}_P$ and $\mathcal{A}_G$
1  Run Profit-Max $(\beta, F_K, \mathcal{A}_P)$                  /* Phase I */
2  if Profit-Max terminated at time step $\tau < T$ then
3      Initialize $\mathcal{A}_G$ on $H_K$                    /* Phase II */
4      for $t = \tau + 1, 2, \ldots, T$ do
5          Receive from $\mathcal{A}_G$ the prices $(p_t, q_t)$
6          Post prices $(p_t, q_t)$ and observe feedback $z_t$
7          Feed feedback $z_t$ to $\mathcal{A}_G$
8  function Profit-Max $(\beta, F_K, \mathcal{A}_P)$
       Input: • budget threshold $\beta$
              • grid $F_K$ of pairs of prices
              • regret minimizer $\mathcal{A}_P$
9      Initialize $\mathcal{A}_P$ on $|F_K|$ actions, one for each $(\hat{p}, \hat{q}) \in F_K$,
         and set $B_0 \leftarrow 0$
10     for $t = 1, 2, \ldots, T$ do
11         Receive from $\mathcal{A}_P$ the prices $(p_t, q_t)$
12         Post prices $(p_t, q_t)$ and observe feedback $z_t$
13         Feed feedback $z_t$ to $\mathcal{A}_P$
14         Update $B_t \leftarrow B_{t-1} + \textsc{Profit}_t(p_t, q_t)$
15         if $B_t \ge \beta$ then Terminate the algorithm
```

4 FULL FEEDBACK

We start by studying the *full feedback* input model where the agents reveal their valuations (s_t, b_t) at the end of each time step t. Here, the learner has counterfactual information regarding all the prices they could have posted, *independently* of the pair of prices actually posted at time t. In Section 4.1, we first present a two-phase learning algorithm (GFT-Max) which guarantees $\tilde{O}(\sqrt{T})$ regret with respect to the best fixed price in hindsight. In Section 4.2 we complement this result by proving that this is tight, up to poly-logarithmic terms.

4.1 $\tilde{O}(\sqrt{T})$ Upper Bound with Full Feedback

We start the analysis by looking at the first phase of GFT-Max, Profit-Max (reported as a function in the pseudocode of GFT-Max). We employ the Hedge algorithm (see, *e.g.,* Section 5.3 of Slivkins [46]) as the regret minimizer $\mathcal{A}_{\text{P}}$, which is used on the action space of the prices in F_K. As a first step, we note that the gain from trade of any fixed price in the first phase (which terminates at the stopping time τ) is not too large.

Lemma 4.1. *Consider Profit-Max with budget threshold β, grid F_K, and learning algorithm Hedge as $\mathcal{A}_{\text{P}}$. Then, with probability at least $1 - 1/T$, we have*

$$\max_{p\in[0,1]} \sum_{t=1}^{\tau} GFT_t(p) \le 8(\beta + 1) \log T + \frac{5T}{K} + 32 \log T \sqrt{T \log(T|F_K|)}.$$

Lemma 4.1 helps us bounding the regret of GFT-Max up to the (random) time step τ, when the algorithm switches from profit to

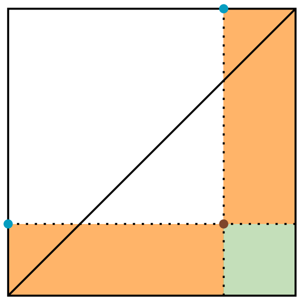

Figure 2: Partition of $[0,1]^2$ as in the proof of Theorem 4.3.

gain from trade maximization. Setting $\beta = \sqrt{T}$ and $K = \sqrt{T}$, and using Hedge as regret minimizer also in the second phase, yields the following result.

Theorem 4.2. *Consider the repeated bilateral trade problem in the full feedback model. There exists a learning algorithm $\mathcal{A}$ that respects global budget balance and whose regret with respect to the best fixed price in hindsight verifies $R_T(\mathcal{A}) \le 92 \log^{3/2}(T)\sqrt{T}$.*

4.2 $\Omega(\sqrt{T})$ Lower Bound with Full Feedback

We present a lower bound that shows how the regret rate in Theorem 4.2 is optimal up to poly-logarithmic factors. The lower bound is based on the following stochastic sequence: at each time step t the pair (s_t, b_t) is drawn uniformly at random between 3 pairs of valuations: $(0, 1/4)$, $(3/4, 1)$ and $(3/4, 1/4)$. These three points naturally partition the $[0,1]^2$ square into four regions (see Figure 2). Crucially, prices in the $[3/4, 1] \times [0, 1/3]$ region (green in Figure 2) incur in negative expected gain from trade, while prices in the $[0, 3/4) \times (1/3, 1]$ region (white in Figure 2) miss all trades. Therefore, the only reasonable option for any learner is to post prices in the two remaining regions (orange in Figure 2), with an expected gain from trade of $1/12$. This allows for a reduction to an expert problem with 2 available actions (one for each of the two orange regions). This construction highlights a key difficulty if compared to lower bounds for per-round budget balanced algorithms: we need to disincentivize the learner from choosing non budget balanced prices below the diagonal. We have the following result.

Theorem 4.3. *Consider the repeated bilateral trade problem in the full feedback model. Any learning algorithm that satisfies global budget balance suffers at least $\Omega(\sqrt{T})$ regret with respect to the best fixed price in hindsight.*

5 PARTIAL FEEDBACK

In this section, we study the more challenging partial feedback models. In Section 5.1, we provide a positive result for the case of one-bit feedback ($z_t = \mathbb{I}\{s_t \le p_t\} \cdot \mathbb{I}\{q_t \le b_t\}$), where the learner only observes whether the trade happened or not. In particular, we show that GFT-Max, with a suitable initialization, achieves a regret of the order $\tilde{O}(T^{3/4})$. Differently from the full-information setting, the design of a no-regret algorithm for the gain from trade (*i.e.,* $\mathcal{A}_G$) is particularly challenging as we need to build an estimator for the gain from trade by only playing non-budget balanced prices in H_K.

In Section 5.2 we complement the regret upper bound by proving that every algorithm has regret at least $\Omega(T^{5/7})$, even with two-bit feedback ($z_t = (\mathbb{I}\{s_t \le p_t\}, \mathbb{I}\{q_t \le b_t\})$), *i.e.,* where each agent separately reveal their willingness to accept the prices posted. One of the main challenges posed by such a lower bound resides in handling non-budget balanced prices, as any algorithm could temporarily sacrifice some profit while collecting large GFT.

5.1 $\tilde{O}(T^{3/4})$ Upper Bound with One-Bit Feedback

We show how to employ GFT-Max with a suitable choice of parameters β and K, and regret minimizers $\mathcal{A}_P$ and $\mathcal{A}_G$ to achieve the desired regret bound. Section 5.1.1 presents a regret-minimizing algorithm that can be employed as $\mathcal{A}_P$, while Section 5.1.2 provides a suitable regret minimizer to be employed as $\mathcal{A}_G$. Finally, in Section 5.1.3, we present the final regret upper bound.

5.1.1 *Regret Minimizer for Profit under Partial Feedback.*

As in the full-information setting, we exploit Profit-Max to maximize the profit until the accrued budget is at least a given threshold β. In particular, we instantiate the subroutine Profit-Max with EXP3.P [3] as regret minimizer $\mathcal{A}_P$ and grid F_K. The following lemma shows that the gain from trade of any fixed price p in the first phase is small enough up to the stopping time τ that terminates the first phase.

Lemma 5.1. *Consider Profit-Max with budget threshold β, grid F_K, and learning algorithm EXP3.P as $\mathcal{A}_P$. Then with probability at least $1 - 1/T$, we have that $\max_{p\in[0,1]} \sum_{t=1}^{\tau} GFT_t(p)$ is at most $8(\beta+1)\log T + \frac{5T}{K} + 256 \log T \sqrt{|F_K| T \log(|F_K| T)}$.*

5.1.2 *Regret Minimizer for Gain from Trade under Partial Feedback.*

A crucial ingredient we need is an estimation procedure capable of extracting quantitative information from the gain from trade, having only access to one bit of feedback. More precisely, we need an estimation procedure of the gain from trade function $H_K \ni (p,q) \mapsto \mathrm{GFT}_t(p,q)$. A similar challenge is faced in Azar et al. [4], where the action set consists of a discretization of a single price (*i.e.,* their estimation procedure posts p to both seller and buyer). However, in our scenario, such symmetry no longer applies. Here, we must consider the grid H_K, which employs distinct prices for the seller and the buyer ($p + 1/K$ and p, respectively). Thus, our estimation procedure GFT-Est has an asymmetric structure (see the pseudocode, in particular Lines 17 and 20).

First, GFT-Est draws a sample from a Bernoulli distribution with parameter $(pK+1)/(K+1)$ (Line 15). If the result is 1, it posts price p to the buyer, and the seller receives a price drawn uniformly at random from $[0, p + 1/K]$ (Line 17). Otherwise, if the result is 0, GFT-Est posts price p to the seller, and the buyer's price is drawn uniformly at random from $[p, 1]$. We denote the final estimate at t by $\widehat{\mathrm{GFT}}_t(p + 1/K, p)$ (Line 20). Overall, our estimator has a small bias, as formalized in the following Lemma.

Lemma 5.2. *For every $(p + 1/K, p) \in H_K$, the random variable $\widehat{GFT}_t(p + 1/K, p)$ is an $1/K$-biased estimate of $GFT_t(p + 1/K, p)$,* i.e., $\left|GFT_t\left(p + \frac{1}{K}, p\right) - \mathbb{E}\left[\widehat{GFT}_t\left(p + \frac{1}{K}, p\right)\right]\right| \le \frac{2}{K}$.

Given the estimation procedure GFT-Est, it is possible to turn any no-regret algorithm for the full-feedback setting into a regret

Algorithm 2: Block-Dec

```
Input: • Number of rounds $T$ and number of blocks $N$
       • Set of prices $H_K$
Initialize Hedge over action space $H_K$ and time horizon $N$
Initialize random mappings $h_j$ for all $j \in \{0, \dots, N-1\}$
$\mathcal{B}_j \leftarrow \{j\frac{T}{N}+1, \dots, (j+1)\frac{T}{N}\}$ for all $j \in \{0, \dots, N-1\}$
for $j \in \{0, \dots, N-1\}$ do
    Receive from $\mathcal{A}$ the distribution over pair of prices $\boldsymbol{x}_j$
    for $t \in \mathcal{B}_j$ do
        if $t \notin S_j$ then
            Play $(p,q) \sim \boldsymbol{x}_j$ and observe $\mathbb{I}\{s_t \le p \wedge q \le b_t\}$
        else
            Select prices $(p,q)$ such that $h_j(p,q) = t$
            Compute $\widehat{\mathrm{GFT}}_t(p,q)$ through GFT-Est
            $\hat{\boldsymbol{r}}_j(p,q) \leftarrow \widehat{\mathrm{GFT}}_t(p,q)$
    Update $\mathcal{A}$ with reward vector $\hat{\boldsymbol{r}}_j$

function GFT-Est
    Input: prices $(p + 1/K, p) \in H_K$
    Sample $Z$ from a Bernoulli with parameter $\frac{pK+1}{K+1}$
    if $Z = 1$ then
        Post price $(\tilde{p}, p)$, with $\tilde{p} \sim U[0, p + 1/K]$
        $\widehat{\mathrm{GFT}}_t(p + 1/K, p) \leftarrow \mathbb{I}\{s_t \le \tilde{p}\}\mathbb{I}\{p \le b_t\}$
    else
        Post price $(p, \tilde{p})$, with $\tilde{p} \sim U[p, 1]$
        $\widehat{\mathrm{GFT}}_t(p + 1/K, p) \leftarrow \mathbb{I}\{s_t \le p\}\mathbb{I}\{\tilde{p} \le b_t\}$
    return $\widehat{\mathrm{GFT}}_t(p + 1/K, p)$
```

minimizer for the partial feedback setting by the standard block decomposition technique (see, e.g., Chapter 4 of Nisan et al. [45]). The procedure, which we call Block-Dec is described in the pseudocode. We assume to employ Hedge as the full-feedback regret minimizer $\mathcal{A}$. The algorithm works by subdividing the time horizon T into N blocks, and the same randomized strategy is played for the entire block except for a single step, uniformly distributed in the block, in which the estimation procedure GFT-Est is run. Then, the reward computed by GFT-Est is fed to Hedge as the estimated reward for that block. Differently from the standard analysis, we have that the reward computed by GFT-Est is not unbiased but has a small $O(1/K)$ bias which does not hinder the overall guarantees of the algorithm, which are presented in the following statement.

Lemma 5.3. *Block-Dec with $K = T^{1/4}$ and $N = T^{1/2}$ guarantees:*

$$\sup_{(p,q)\in H_K} \sum_{t=1}^{T} GFT_t(p,q) - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{(p,q)\sim \boldsymbol{x}_t} [GFT_t(p,q)] \le \frac{5}{2} T^{3/4}\sqrt{\log(T)}.$$

5.1.3 Putting Everything Together. GFT-Max with the two regret minimizers described in Sections 5.1.1 and 5.1.2 guarantees a $O(T^{3/4})$ bound on the regret.

Theorem 5.4. *Consider the repeated bilateral trade problem in the one-bit feedback model. There exists a learning algorithm $\mathcal{A}$ that respects global budget balance and whose regret with respect to the best fixed price in hindsight verifies:*

$$R_T(\mathcal{A}) \le 1282 \cdot T^{3/4} \log^2 T.$$

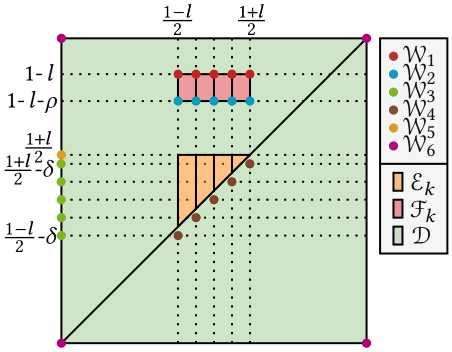

Figure 3: Representation of the support of the distributions μ_k and their partitions.

5.2 $\Omega(T^{5/7})$ Lower Bound with Two-Bit Feedback

In this section, we provide a lower bound for learning the best price against any oblivious adversary, with global budget balance constraints and two-bit feedback. Our construction builds upon the one by Cesa-Bianchi et al. [19], but exhibits two key differences. First, we are not constrained to use smooth value distributions. This allows us to simplify the construction, avoiding the reduction to multi-armed bandits with feedback graphs. Second, we only require algorithms to be globally budget balanced (instead of per-round weakly budget balanced); looser budget balance constraints enhance the capabilities of the learning algorithm. All in all, we derive a lower bound that is slightly looser $T^{5/7} \approx T^{0.714}$ compared to the $\Omega(T^{3/4})$. We further elaborate on this comparison at the end of the Section.

Theorem 5.5. *Consider the problem of repeated bilateral trade in the two-bit feedback model. Any learning algorithm that satisfies global budget balance suffers regret at least $\Omega(T^{5/7})$.*

The rest of the Section is devoted to the proof of Theorem 5.5; for the missing details, we refer to the full version. Our lower bound construction is based on N stochastic distributions of valuations that are close with respect to some statistical measure of distance and ensure that any pair of prices that reveals information on the underlying instance is highly suboptimal in terms of GFT (*i.e.*, gathering information is "costly"). We proceed in 5 steps.

i) Building a set of hard instances. We introduce the N hard instances of the bilateral trade problem. Our goal is to show that any learning algorithm suffers $\Omega(T^{5/7})$ regret in at least one of the N instances. We define a distribution $\mu_k \in \Delta([0,1]^2)$ of valuations (s,b) over $[0,1]^2$ for each $k \in \{0, \dots, N-1\}$, where we have $N-1$ "perturbed" distributions corresponding to indices $k \in \{1, \dots, N-1\}$, and a "base" distribution corresponding to $k = 0$.

Let $\ell = 1/12$, $\Delta = \ell/(N-1)$, $\rho = 1/32$ and $\delta = \Delta/2$. Then, for any instance $k \in \{0, \dots, N-1\}$, the distributions μ_k are supported on the same set $\mathcal{W}$ of finitely many valuations. We describe the set $\mathcal{W}$ by partitioning it into six different sets. An illustration of the set of valuations can be found in Figure 3. First, we define the two sets $\mathcal{W}_1 = \{w_1^i = (1-\ell/2 + i\Delta, 1-\ell) : i = 0, \dots, N-1\}$ and

$\mathcal{W}_2 = \left\{w_2^i = (\nicefrac{1-\ell}{2} + i\Delta, 1 - \ell - \rho) : i = 0, \dots, N-1\right\}$. These valuations are "balanced out" by the following N valuations: $\mathcal{W}_3 = \left\{w_3^i = (0, \nicefrac{1-\ell}{2} - \delta + i\Delta) : \; i = 0, \dots, N-1\right\}$. Moreover, we define $\mathcal{W}_4 = \left\{w_4^i = (\nicefrac{1-\ell}{2} + i\Delta, \nicefrac{1-\ell}{2} - \delta + i\Delta) : i = 0, \dots, N-1\right\}$, and a single valuation belonging to $\mathcal{W}_5 = \{(0, \nicefrac{1-\ell}{2})\}$. We conclude by defining the set of the "extremal" valuations $\mathcal{W}_6 = \{0,1\}^2$.

We assign different probabilities to the valuations in each set $\mathcal{W}_j$ depending on the instance. In particular, for any instance $k \in \{1, \dots, N-1\}$ with distribution μ_k, we have that

$$\mu_k(w_j^i) = \frac{1}{64N^2} = \gamma_1 \quad \forall j \in \{1,2\}, i \notin \{k, k+1\}, \tag{3}$$

while we perturb by ε the probability of the following valuations:

$$\mu_k(w_1^k) = \gamma_1 + \varepsilon, \; \mu_k(w_1^{k+1}) = \gamma_1 - \varepsilon, \tag{4}$$
$$\mu_k(w_2^k) = \gamma_1 - \varepsilon, \; \mu_k(w_2^{k+1}) = \gamma_1 + \varepsilon.$$

Conversely, for the base instance μ_0, we place equal probability $\mu_0(w) = \gamma_1$ on all the valuations $w \in \mathcal{W}_1 \cup \mathcal{W}_2$, and hence all these valuations have the same probability. For each instance $k \in \{0, \dots, N-1\}$ with distribution μ_k, the probability of valuations w_3^i, with $i \in \{0, \dots, N-1\}$, is set as

$$\mu_k(w_3^i) = \gamma_1 \cdot \frac{1 - \ell - \rho - 2i\Delta}{\nicefrac{1-\ell}{2} - \delta + i\Delta} \in (0, 2\gamma_1).$$

Let $\gamma_3^{\text{tot}} = \sum_{w \in \mathcal{W}_3} \mu_k(w) < 2\gamma_1 N$ be the total probability assigned to valuations in $\mathcal{W}_3$. Moreover, for any instance $k \geq 1$ with distribution μ_k, we assign to every trade in $\mathcal{W}_4$ probability $\gamma_4 = 4\gamma_1(13N - 14)$, *i.e.,*

$$\mu_k(w) = 4\gamma_1(13N - 14) \quad \forall w \in \mathcal{W}_4. \tag{5}$$

Then, for any instance $k \geq 1$ with distribution μ_k, we assign probability γ_5 to the single valuation in $\mathcal{W}_5$, *i.e.,* $\mu_k(0, (1-\ell)/2) = \gamma_5 = 1/64$. Finally, all the remaining probability is equally divided into the 4 extremal trades in $\mathcal{W}_6$, *i.e.,*

$$\mu_k(w) = \frac{1 - \left(2\gamma_1 N + \gamma_3^{\text{tot}} + 4\gamma_1 N(13N - 14) + \gamma_5\right)}{4} = \gamma_6 \, \forall w \in \mathcal{W}_6.$$

Now, we define $\mathcal{G}_{\mathcal{W}}$ as the grid generated by such valuations. Formally $\mathcal{G}_{\mathcal{W}}^s = \{s : \exists (s, \cdot) \in \mathcal{W}\}$, $\mathcal{G}_{\mathcal{W}}^b = \{b : \exists (\cdot, b) \in \mathcal{W}\}$, and $\mathcal{G}_{\mathcal{W}} = \left\{(s,b) : s \in \mathcal{G}_{\mathcal{W}}^s \text{ and } b \in \mathcal{G}_{\mathcal{W}}^b\right\}$. Thus, $\mathcal{G}_{\mathcal{W}}^s$ and $\mathcal{G}_{\mathcal{W}}^b$ represent the projections of $\mathcal{G}_{\mathcal{W}}$ onto its first (seller) and second (buyer) component, respectively. **ii) Analysis of the gain from trade.** By construction, we can restrict our attention to consider algorithms that play only prices in $\mathcal{G}_{\mathcal{W}}$, without loss of generality. Consider in fact, any instance k and any randomized algorithm $\mathcal{A}$. One can easily prove that there exists an equivalent algorithm (in terms of both feedback, profit, and GFT), that only has distribution supported on the grid $\mathcal{G}_{\mathcal{W}}$ generated by the valuations $\mathcal{W}$.

Next, for any $p \in \mathcal{G}_{\mathcal{W}}^s$, we characterize the value of posting the pair of prices $(p, p+\delta)$ under the valuation's distribution μ_k, with $k \in \{0, \dots, N-1\}$. Note that posting the pair $(p, p+\delta) \in \mathcal{G}_{\mathcal{W}}$ under any instance μ_k, is equivalent to posting a single price $p \in \mathcal{G}_{\mathcal{W}}^s$ to both the seller and the buyer, with the only difference that $(p,p) \notin \mathcal{G}_{\mathcal{W}}$, while $(p, p+\delta) \in \mathcal{G}_{\mathcal{W}}$. Then, for any $p \in \mathcal{G}_{\mathcal{W}}^s$, we relate the GFT obtained by posting a pair $(p, p+\delta)$ under valuations sampled from μ_k, with $k \in \{1, \dots, N-1\}$, and under the base distribution μ_0. For every $k \in \{0, \dots, N-1\}$, let $\mathbb{E}_k$ and $\mathbb{P}_k$ be expectation and probability measure under instance μ_k, respectively. Direct calculations shows that, for all $p \in \mathcal{G}_{\mathcal{W}}^s$ and $k \in \{1, \dots, N-1\}$, it holds that $\mathbb{E}_k[\text{GFT}(p, p+\delta, s, b)] = \mathbb{E}_0[\text{GFT}(p, p+\delta, s, b)] + \rho\varepsilon\mathbb{I}\left\{p = p_k^*\right\}$, where $\text{GFT}(p, p+\delta, s, b)$ is simply the gain from trade when the prices posted are $(p, p+\delta)$ and valuations (s,b), and $p_k^* = \frac{1-\ell}{2} + k\Delta$. Moreover, for all $p \in \mathcal{G}_{\mathcal{W}}^s$ it holds that $\mathbb{E}_0[\text{GFT}(p, p+\delta, s, b)]$ is

$$\begin{cases} c_1 = \gamma_5 \frac{1+\ell}{2} + \mu_0(0,1) + \gamma_1 \frac{77}{96} N & \text{if } p \in [0, \frac{1+\ell}{2}] \\ c_2 = \mu_0(0,1) + \gamma_1 \frac{77}{96} N & \text{if } p \in (\frac{1+\ell}{2}, 1 - \ell - c] \\ c_3 = \mu_0(0,1) + \gamma_1 \frac{5}{12} N & \text{if } p \in (1 - \ell - c, 1 - \ell] \\ c_4 = \mu_0(0,1) & \text{if } p \in (1 - \ell, 1] \end{cases}$$

From these calculations, we show that in an instance $k \geq 1$ the pair that maximizes the expected gain from trade is $(p_k^*, p_k^* + \delta)$.

Lemma 5.6. *For any instance $k \in \{1, \dots, N-1\}$, we have that:*

$$\max_{(p,q),\, p \leq q} \mathbb{E}_k[\mathit{GFT}(p, q, s, b)] = \mathbb{E}_k[\mathit{GFT}(p_k^*, p_k^* + \delta, s, b)] = c_1 + \rho \cdot \varepsilon.$$

The previous lemma characterizes the optimal fixed budget balanced strategy. Then, we show that all the strategies that are *not* budget balanced are dominated. Indeed, one of the main challenges of our reduction is that, in general, a globally budget balanced algorithm could get a larger GFT by temporarily sacrificing some profit and posting prices (p,q) with $q < p$. In the following lemma we show that our instances are built so that these strategies are dominated and thus can be discarded. Intuitively, every tuple of prices p, q that tries to gain higher GFT than the one obtained by playing on the diagonal must win also trades in $\mathcal{W}_4$. Then, since trades in $\mathcal{W}_4$ have negative GFT and happen with sufficiently high probability γ_4, we have that posting prices $q < p$ is dominated.

Lemma 5.7. *For every pair of posted prices $(p,q) \in \mathcal{G}_{\mathcal{W}} \cap \{(p,q) \in [0,1]^2 \,|\, p < q\}$, $(p', q') \in \mathcal{G}_{\mathcal{W}} \cap \{(p,q) \in [0,1]^2 \,|\, p \geq q\}$, and instance $k \in \{0, \dots, N-1\}$, we have that $\mathbb{E}_k[\mathit{GFT}(p, q, s, b)] \leq \mathbb{E}_k[\mathit{GFT}(p', q', s, b)] \leq c_1 + \rho\,\varepsilon\,\mathbb{I}\left\{(p', q') = (p_k^*, p_k^* + \delta)\right\}$.*

We complete this section by showing that also strategies that propose a high price to the buyer are dominated in every instance. In particular, we show that when the algorithm places prices (p,q) with $q > (1+\ell)/2$, it looses a constant GFT with respect to choosing a smaller q. This is because the learner cannot induce the trade $\mathcal{W}_5$ which guarantees expected GFT of $\Theta(\gamma_5)$. Formally,

Lemma 5.8. *For any instance k, $p \in \mathcal{W}_{\mathcal{G}}^s \cap [(1-\ell)/2, (1+\ell)/2]$, and $q \in ((1+\ell)/2, 1] \cap \mathcal{G}_{\mathcal{W}}^b$ we have $\mathbb{E}_k[\mathit{GFT}(p, p+\delta)] \geq \mathbb{E}_k[\mathit{GFT}(p, q)] + \frac{\gamma_5}{3}$.*

Intuitively, this lemma shows that exploring is costly. Indeed, as we show shortly, the algorithm must post $q \geq (1+\ell)/2$ to gain information on the instance.

iii) Analysis of the feedback. Now, we show that for any instance μ_k and any posted prices (p,q), the distribution of the two-bit feedback is independent on the instance almost everywhere. Specifically, the feedback distribution depends on the instance k only within a "small" and instance-dependent region of prices. For every instance $k \geq 1$, let $\mathcal{F}_k = [\nicefrac{1-\ell}{2} + (k-1)\Delta, \nicefrac{1-\ell}{2} + k\Delta) \times (1 - \ell - c, 1 - \ell]$. Then, the feedback is independent from the instance for each pair outside the sets $\mathcal{F}_k$ (see Claim 2 of Cesa-Bianchi et al. [19] for a similar result).

Lemma 5.9. *For all $(p, q) \in [0, 1]^2 \setminus \bigcup_{k' \in \{1,\dots,N-1\}} \mathcal{F}_{k'}$ it holds that for all $z \in \{0, 1\}^2$ and $\forall j, k \in \{0, \dots, N-1\}$ we have that $\mathbb{P}_k[(\mathbb{I}\{s \le p\}, \mathbb{I}\{q \le b\}) = z] = \mathbb{P}_j[(\mathbb{I}\{s \le p\}, \mathbb{I}\{q \le b\}) = z]$.*

iv) Price regions. We partition $[0, 1]^2$ in the following regions, also depicted in Figure 3.

- *Exploration regions*. We have the $N-1$ regions $\mathcal{F}_k$. These are the regions where the probability of observing a certain two-bit feedback depends on the instance μ_k.
- *Exploitation regions*. We define the regions $\mathcal{E}_k$ for any $k \ge 1$ as
$$\mathcal{E}_k = \left\{(p, q) \middle| \; q \ge p, \; q \le \tfrac{1+\ell}{2}, \; p \in \left[\tfrac{1-\ell}{2} + (k-1)\Delta, \tfrac{1-\ell}{2} + k\Delta\right)\right\}.$$
All these regions are such that the GFT collected by posting $(p, q) \in \mathcal{E}_k$ is close (and smaller than or equal to) to the optimal GFT, *i.e.,* the one obtained by posting $(p^*_k, p^*_k + \delta)$.
- *Dominated regions*. We define $\mathcal{D}$ as the remaining set of possible valuations $\mathcal{D} = [0, 1]^2 \setminus (\cup_k(\mathcal{F}_k \cup \mathcal{E}_k))$. It's easy to verify that $(p, q) \in \mathcal{D}$ obtain a GFT that is at most c_1.

Next, we define the number of times an algorithm plays in the exploration, exploitation and dominated regions, which are $\mathcal{N}_k = \sum_{t\in[T]} \mathbb{I}\{(p_t, q_t) \in \mathcal{F}_k\}$, $\mathcal{M}_k = \sum_{t\in[T]} \mathbb{I}\{(p_t, q_t) \in \mathcal{E}_k\}$, and $\mathcal{O} = \sum_{t\in[T]} \mathbb{I}\{(p_t, q_t) \in \mathcal{D}\}$, respectively. Then, we can upperbound the gain from trade of an algorithm $\mathcal{A}$ by considering only the number of plays in each region. In particular:

- **Cost of exploration:** the GFT collected by posting prices in $\mathcal{F}_j$ is at most c_2 for all j (Lemma 5.8);
- **Exploitation:** the GFT collected by posting prices in $\mathcal{E}_j$ is at most $c_1 + \rho \cdot \varepsilon \mathbb{I}\{j = k\}$ (Lemma 5.7);
- **Cost of domination:** the GFT collected by posting prices in $\mathcal{D}$ is at most c_1 (Lemma 5.7).

Formally, these observations lead to the following upper bound.

Lemma 5.10. *Let $\{(p_t, q_t)\}_{t\in[T]}$ be the sequences of prices posted by any algorithm $\mathcal{A}$. Then*
$$\sum_{t=1}^{T} \mathbb{E}_k[GFT(p_t, q_t, s, b)] \le \mathbb{E}_k\left[\rho\varepsilon\mathcal{M}_k + \sum_{k=1}^{N-1}(c_1\mathcal{M}_j + c_2\mathcal{N}_j + c_1\mathcal{O})\right].$$

v) Relating the algorithm behavior on different instances. Now we relate the number of exploitation rounds $\mathcal{M}_k$ in different instances. This difference depends on the probability measures $\mathbb{P}_k$ and $\mathbb{P}_0$ through the Pinsker's inequality on a suitably defined multinomial random variable that encodes the four possible feedbacks.

Lemma 5.11. *For all $k \in \{1, \dots, N-1\}$ we have that $\mathbb{E}_k[\mathcal{M}_k] - \mathbb{E}_0[\mathcal{M}_k] \le T\varepsilon\sqrt{2\mathbb{E}_0[\mathcal{N}_k]/\gamma_6}$.*

vi) Lower bounding the regret.

We define the expected regret under instance k as:
$$R^k_T = \max_{(p,q)\in[0,1]^2, p \ge q} \mathbb{E}_k\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p, q) - \sum_{t=1}^{T} \mathrm{GFT}_t(p_t, q_t)\right].$$

Then, combining all the previous results leads to the following lemma which gives a lower bound in terms of ε, N, and T.

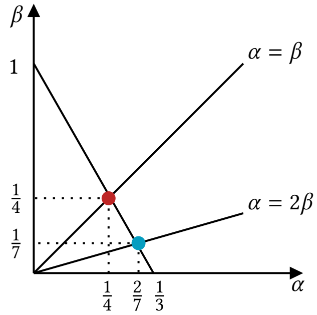

Figure 4: Order of α and β reachable by Cesa-Bianchi et al. [19] (red) and this work (blue).

Lemma 5.12. *There is an instance $k \in \{0, \dots, N-1\}$ and an absolute constant $c \in (0, 1)$ such that $R^k_T \ge c \cdot \min\left(N\varepsilon^{-2}, \varepsilon T\right)$.*

By using Lemma 5.12 we can readily conclude the proof of Theorem 5.5 as follows. Let $\varepsilon = T^{-\alpha}$ and $N = T^{\beta}$, with $\alpha, \beta > 0$. Now we simply have to optimize over the choice of parameters α and β. In doing so, we need to take into account the additional constraints necessary to have well defined instance distribution μ_k. In particular, we have that $\varepsilon \le \gamma_1$ from Equation (4), and $2N\gamma_1 < 1$ from Equation (3). Moreover, we also need to impose $\gamma_4 N < 1$ by Equation (5). Since $\gamma_4 = 4\gamma_1(14N - 13)$, this also implies that $\gamma_1 < \frac{1}{4N(13N-14)} < \frac{1}{N^2}$ for $N > 2$. Therefore, the constraint $\varepsilon < \gamma_1$ implies: $\varepsilon = T^{-\alpha} \le 1/T^{2\beta} = 1/N^2$ which yields that $\alpha \ge 2\beta$. Note that this dominates the constraint $\varepsilon < 1/N$ (or equivalently written as $\alpha \ge \beta$) that would have been implied by Equation (4) alone.

The lower bound of is maximized when α and β are solution to $\max_{\alpha\ge0} (1-\alpha)$ subject to $\alpha \ge 2\beta$, $1 - \alpha = \beta + 2\alpha$, which gives $\alpha = 2/7$ and $\beta = 1/7$. This implies a lower bound $\Omega(T^{5/7})$.

Connection with the $\Omega(T^{3/4})$ lower bound of Cesa-Bianchi et al. [19]. While our result and the one of Cesa-Bianchi et al. [19] build on a similar constructions (at least conceptually), we obtain a weaker lower bound. The main reason is that the learner in Cesa-Bianchi et al. [19] is weak budget balanced, while in our work the learner has only a global budget balance constraint. To preclude this option to the learner, we penalize the GFT of prices in the lower triangle by adding the set of valuations $\mathcal{W}_4$. If γ_4 is large enough w.r.t. γ_1, then posting prices in the lower triangle is dominated. In particular, we must choose $\gamma_4 = \Theta(\gamma_1 N)$ as we prove in Lemma 5.7. Once we prove that the lower triangle is dominated, we can conceptually reduce our problem to the one of Cesa-Bianchi et al. [19]. However, the choice of $\gamma_4 = \Theta(\gamma_1 N)$ imposes the additional constraint $\alpha \ge 2\beta$, which is not needed in the original construction. Hence, they can set $\alpha = \beta = 1/4$, and get a bound of $\Omega(T^{3/4})$. This difference is depicted in Figure 4.

6 BEST FEASIBLE DISTRIBUTION OF PRICES

In this section, we analyse the regret with respect to the best fixed distribution over prices which satisfies global budget balance *on average*. First, we present a negative result that clearly separates this new benchmark from the best fixed price in hindsight: in Theorem 6.2, we prove that it is impossible to achieve sublinear $(1+\varepsilon)$-regret with respect to the best feasible distribution, even in the full feedback setting. On the positive side, we show that the two benchmarks are only a multiplicative factor 2 apart (Theorem 6.3). This implies that any learning algorithm that exhibits sublinear regret with respect to the best fixed price in hindsight automatically achieves sublinear 2-regret with respect to the best feasible distribution. Finally, we complement this positive result by proving that this multiplicative gap of 2 is tight (Theorem 6.5).

6.1 Linear Lower Bound

The best feasible distribution has a crucial advantage with respect to any budget balanced learner: it has the possibility to "run some deficit" in a preliminary phase of the sequence as it knows it will be possible to extract enough profit to ensure global budget balance in some later stages. For instance, consider a half-sequence where (s_t, b_t) is either $(0, 1/3)$ or $(2/3, 1)$, for $t \le T/2$. Any learning algorithm has to enforce budget balance at time $T/2$ (to be protected about the possibility that $(s_t, b_t) = 0$ for all future t), while the randomized benchmark, which knows the future, may run a deficit and collect more gain from trade by posting the budget unbalanced prices $(2/3, 1/3)$ with some probability. Inspired by this example, we state the following Lemma.

Lemma 6.1. *For any algorithm $\mathcal{A}$ that enforces global budget balance, there exists a deterministic sequence of valuations $\mathcal{S}_1$ with the following properties: (i) the expected gain from trade of $\mathcal{A}$ is at most $T/9$; (ii) the valuations (s_t, b_t) are either $(0, 1/3)$ or $(2/3, 1)$ for all $t \le T/2$; (iii) the valuations (s_t, b_t) are equal to $(0, 0)$ for all $t > T/2$.*

The lemma is crucial in proving the impossibility result in the following Theorem, which holds even under full feedback.

Theorem 6.2. *Fix any constant $\alpha \in [1, 36/35)$, and any globally budget balanced learning algorithm $\mathcal{A}$ with full-feedback. Then there exists a sequence of valuations such that*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{(p,q)\sim\gamma^*} GFT_t(p,q) - \alpha \cdot \sum_{t=1}^{T} \mathbb{E}[GFT_t(p_t, q_t)] \ge \frac{5}{18}\left(\frac{36}{35} - \alpha\right) T,$$

where distribution γ^* is the optimal feasible distribution.

6.2 Comparison of the Two Benchmarks

Surprisingly, it holds that the performance of the optimal fixed price is to not far from that of optimal global budget balanced distribution.

Theorem 6.3. *Denote with p^*, resp. γ^*, the best fixed price, resp. the best feasible distribution. Then, for any sequence of valuations:*

$$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\gamma^*} GFT_t(p,q) \le 2 \sum_{t=1}^{T} GFT_t(p^*).$$

As a corollary, we have that any algorithm that achieves sublinear regret with respect to the best fixed price also guarantees sublinear 2-regret with respect to the best feasible prices distribution.

Corollary 6.4. *Let $\mathcal{A}$ be a learning algorithm for the repeated bilateral trade problem which guarantees an upper bound of $f(T)$ on the regret with respect to the best fixed price in hindsight. Then, the 2-regret of $\mathcal{A}$ with respect to the best budget feasible distribution over prices is at most $f(T)$.*

Surprisingly, the factor 2 between the two benchmarks is optimal. This implies that the analysis of the performance of the algorithms in Corollary 6.4 is essentially tight.

Theorem 6.5. *For any $\varepsilon > 0$, there exists a sequence of valuations such that*

$$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\gamma^*} GFT_t(p,q) \ge (2-\varepsilon) \sum_{t=1}^{T} GFT_t(p^*),$$

where p^* and γ^* are the best fixed price and global budget balanced distribution, respectively.

7 FINAL REMARKS AND OPEN PROBLEMS

In this paper we introduce the notion of global budget balance in the repeated bilateral trade problem. With this notion, we show for the first time that it is possible to achieve sublinear regret with respect to the best fixed price in hindsight, without relying on any additional assumption. In the full feedback model we prove that the minimax regret rate of the learning problem is $\tilde{\Theta}(\sqrt{T})$, while in the partial feedback models, we provide an upper bound on the regret of order $\tilde{O}(T^{3/4})$, which is complemented with a $\Omega(T^{5/7})$ lower bound. Our regret results proves a clear separation between the two feedback models, but leave an open gap between the $T^{5/7}$ and $T^{3/4}$ rates in partial feedback.

Inspired by Bandits with Knapsack, we formulated a new benchmark: the best feasible distribution over prices. Against this harder benchmark we prove that it is possible to achieve sublinear 2-regret, while no algorithm can achieve sublinear $(1+\varepsilon_0)$-regret. We leave as an open question the characterization of the optimal competitive ratio $\alpha \in [1+\varepsilon_0, 2]$ obtainable against this benchmark.

ACKNOWLEDGMENTS

MB, MC, AC, FF are partially supported by the FAIR (Future Artificial Intelligence Research) project PE0000013, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence). FF is also partially supported by ERC Advanced Grant 788893 AMDROMA "Algorithmic and Mechanism Design Research in Online Markets", and PNRR MUR project IR0000013-SoBigData.it. MC is also partially supported by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237). AC is partially supported by MUR - PRIN 2022 project 2022R45NBB funded by the NextGenerationEU program.

REFERENCES

[1] Shipra Agrawal and Nikhil R. Devanur. 2019. Bandits with Global Convex Constraints and Objective. *Oper. Res.* 67, 5 (2019), 1486–1502. https://doi.org/10.1287/opre.2019.1840

[2] Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. 2017. Nonstochastic Multi-Armed Bandits with Graph-Structured Feedback. *SIAM J. Comput.* 46, 6 (2017), 1785–1826. https://doi.org/10.1137/140989455

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* 32, 1 (2002), 48–77. https://doi.org/10.1137/S0097539701398375

[4] Yossi Azar, Amos Fiat, and Federico Fusco. 2022. An $alpha$-regret analysis of Adversarial Bilateral Trade. In *NeurIPS*.

[5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2018. Bandits with knapsacks. *J. ACM* 65, 3 (2018), 1–55. https://doi.org/10.1145/3164539

[6] Santiago R. Balseiro and Yonatan Gur. 2019. Learning in Repeated Auctions with Budgets: Regret Minimization and Equilibrium. *Manag. Sci.* 65, 9 (2019), 3952–3968. https://doi.org/10.1287/mnsc.2018.3174

[7] Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. 2014. Partial Monitoring - Classification, Regret Bounds, and Algorithms. *Math. Oper. Res.* 39, 4 (2014), 967–997. https://doi.org/10.1287/moor.2014.0663

[8] Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. 2023. No-Regret Learning in Bilateral Trade via Global Budget Balance. *CoRR* abs/2310.12370 (2023).

[9] Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. 2024. Bandits with Replenishable Knapsacks: the Best of both Worlds. In *ICLR*.

[10] Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. 2023. Optimal Rates and Efficient Algorithms for Online Bayesian Persuasion. In *ICML (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 2164–2183.

[11] Liad Blumrosen and Shahar Dobzinski. 2014. Reallocation mechanisms. In *EC*. ACM, 617. https://doi.org/10.1145/2600057.2602843

[12] Liad Blumrosen and Yehonatan Mizrahi. 2016. Approximating Gains-from-Trade in Bilateral Trading. In *WINE (Lecture Notes in Computer Science, Vol. 10123)*. Springer, 400–413. https://doi.org/10.1007/978-3-662-54110-4_28

[13] Nataša Bolić, Tommaso Cesari, and Roberto Colomboni. 2024. An Online Learning Theory of Brokerage. *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems* (2024).

[14] Johannes Brustle, Yang Cai, Fa Wu, and Mingfei Zhao. 2017. Approximating gains from trade in two-sided markets via simple mechanisms. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. 589–590. https://doi.org/10.1145/3033274.3085148

[15] Matteo Castiglioni, Andrea Celli, and Christian Kroer. 2022. Online Learning with Knapsacks: the Best of Both Worlds. In *ICML (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 2767–2783.

[16] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. 2020. Online Bayesian Persuasion. In *NeurIPS*.

[17] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. 2023. Regret minimization in online Bayesian persuasion: Handling adversarial receiver's types under full and partial feedback models. *Artif. Intell.* 314 (2023), 103821.

[18] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. 2021. A Regret Analysis of Bilateral Trade. In *EC*. ACM, 289–309. https://doi.org/10.1145/3465456.3467645

[19] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. 2023. Repeated Bilateral Trade Against a Smoothed Adversary. In *COLT (Proceedings of Machine Learning Research, Vol. 195)*. PMLR, 1095–1130.

[20] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. 2024. Bilateral trade: A regret minimization perspective. *Mathematics of Operations Research* 49, 1 (2024), 171–203. https://doi.org/10.1287/moor.2023.1351

[21] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. 2024. The Role of Transparency in Repeated First-Price Auctions with Unknown Valuations. In *STOC*. ACM.

[22] Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2015. Regret Minimization for Reserve Prices in Second-Price Auctions. *IEEE Trans. Inf. Theory* 61, 1 (2015), 549–564. https://doi.org/10.1109/TIT.2014.2365772

[23] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. 2006. Regret Minimization Under Partial Monitoring. *Math. Oper. Res.* 31, 3 (2006), 562–580. https://doi.org/moor.1060.0206

[24] Constantinos Daskalakis and Vasilis Syrgkanis. 2022. Learning in auctions: Regret is hard, envy is easy. *Games Econ. Behav.* 134 (2022), 308–343.

[25] Yuan Deng, Jieming Mao, Balasubramanian Sivan, and Kangning Wang. 2022. Approximately efficient bilateral trade. In *STOC*. ACM, 718–721. https://doi.org/10.1145/3519935.3520054

[26] Paul Duetting, Guru Guruganesh, Jon Schneider, and Joshua Ruizhi Wang. 2023. Optimal No-Regret Learning for One-Sided Lipschitz Functions. In *ICML (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 8836–8850.

[27] Paul Dütting, Federico Fusco, Philip Lazos, Stefano Leonardi, and Rebecca Reiffenhäuser. 2021. Efficient two-sided markets with limited information. In *STOC*. ACM, 1452–1465. https://doi.org/10.1145/3406325.3451076

[28] Yumou Fei. 2022. Improved approximation to first-best gains-from-trade. In *International Conference on Web and Internet Economics*. Springer, 204–218. https://doi.org/10.1007/978-3-031-22832-2_12

[29] Michal Feldman, Tomer Koren, Roi Livni, Yishay Mansour, and Aviv Zohar. 2016. Online pricing with strategic and patient buyers. *Advances in Neural Information Processing Systems* 29 (2016).

[30] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. 2016. Adaptive Contract Design for Crowdsourcing Markets: Bandit Algorithms for Repeated Principal-Agent Problems. *J. Artif. Intell. Res.* 55 (2016), 317–359.

[31] Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. 2022. Adversarial bandits with knapsacks. *J. ACM* 69, 6 (2022), 1–47. https://doi.org/10.1145/3557045

[32] Rodolphe Jenatton, Jim C. Huang, and Cédric Archambeau. 2016. Adaptive Algorithms for Online Convex Optimization with Long-term Constraints. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 402–411.

[33] Sham M. Kakade, Adam Tauman Kalai, and Katrina Ligett. 2009. Playing Games with Approximation Algorithms. *SIAM J. Comput.* 39, 3 (2009), 1088–1106. https://doi.org/10.1137/070701704

[34] Zi Yang Kang, Francisco Pernice, and Jan Vondrák. 2022. Fixed-Price Approximations in Bilateral Trade. In *SODA*. SIAM, 2964–2985. https://doi.org/10.1137/1.9781611977073.115

[35] Thomas Kesselheim and Sahil Singla. 2020. Online Learning with Vector Costs and Bandits with Knapsacks. In *COLT (Proceedings of Machine Learning Research, Vol. 125)*. PMLR, 2286–2305.

[36] Robert D. Kleinberg and Frank Thomson Leighton. 2003. The Value of Knowing a Demand Curve: Bounds on Regret for Online Posted-Price Auctions. In *FOCS*. IEEE Computer Society, 594–605. https://doi.org/10.1109/SFCS.2003.1238232

[37] Raunak Kumar and Robert Kleinberg. 2022. Non-monotonic Resource Utilization in the Bandits with Knapsacks Problem. In *NeurIPS*.

[38] Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. 2016. Learning and Efficiency in Games with Dynamic Population. In *SODA*. SIAM, 120–129. https://doi.org/10.1137/1.9781611977554.ch17

[39] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. 2012. Trading regret for efficiency: online convex optimization with long term constraints. *J. Mach. Learn. Res.* 13 (2012), 2503–2528.

[40] Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. 2009. Online Learning with Sample Path Constraints. *J. Mach. Learn. Res.* 10 (2009), 569–590.

[41] R Preston McAfee. 2008. The gains from trade under fixed price mechanisms. *Applied economics research bulletin* 1, 1 (2008), 1–10.

[42] Jamie Morgenstern and Tim Roughgarden. 2015. On the Pseudo-Dimension of Nearly Optimal Auctions. In *NIPS*. 136–144.

[43] Roger B Myerson and Mark A Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of economic theory* 29, 2 (1983), 265–281.

[44] Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, and Vianney Perchet. 2022. Learning in Repeated Auctions. *Found. Trends Mach. Learn.* 15, 3 (2022), 176–334. https://doi.org/10.1561/2200000077

[45] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani (Eds.). 2007. *Algorithmic Game Theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511800481

[46] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *Found. Trends Mach. Learn.* 12, 1-2 (2019), 1–286. https://doi.org/10.1561/2200000068

[47] Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. 2023. Contextual Bandits with Packing and Covering Constraints: A Modular Lagrangian Approach via Regression. In *COLT (Proceedings of Machine Learning Research, Vol. 195)*. PMLR, 4633–4656.

[48] Wen Sun, Debadeepta Dey, and Ashish Kapoor. 2017. Safety-Aware Algorithms for Adversarial Contextual Bandit. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 3280–3288.

[49] William Vickrey. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance* 16, 1 (1961), 8–37.

[50] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. 2016. Online learning in repeated auctions. In *COLT (JMLR Workshop and Conference Proceedings, Vol. 49)*. JMLR.org, 1562–1583.

[51] Hao Yu, Michael J. Neely, and Xiaohan Wei. 2017. Online Convex Optimization with Stochastic Constraints. In *NIPS*. 1428–1438.

[52] Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I. Jordan. 2023. The Sample Complexity of Online Contract Design. In *EC*. ACM, 1188. https://doi.org/10.1145/3580507.3597673

[53] You Zu, Krishnamurthy Iyer, and Haifeng Xu. 2021. Learning to Persuade on the Fly: Robustness Against Ignorance. In *EC*. ACM, 927–928. https://doi.org/10.1145/3465456.3467593