POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Online Bilateral Trade: Learning the Optimal Policy in a Stochastic Environment

Laurea Magistrale in Computer Science and Engineering

Mattia Piccinato, 10717274

Advisor:
M. Castiglioni

Co-advisors:
A. Lunghi, A. Marchesi

Academic year:
2024-2025

Abstract: Bilateral trade models the interaction between a seller and a buyer with private valuations. Within the online learning framework introduced by [9], mechanisms post fixed prices sequentially and are evaluated against the best fixed price in hindsight. More recently, [4] introduce a new notion of budget balance, together with a stronger benchmark: the best randomized pricing strategy over *price pairs* that satisfies global budget balance in expectation. While they prove that learning against such baseline is impossible in adversarial environments, this work shows that the distributional benchmark is attainable in stochastic environment, introducing a globally budget-balanced learning algorithm that achieves $\widetilde{O}(T^{3/4})$ pseudo-regret against such distributional benchmark.

Key-words: bilateral trade; online learning; regret minimization; global budget balance; stochastic environment; realistic feedback; distributional benchmark; gain from trade

1. Introduction

1.1. Background and Motivation

Bilateral trade models the problem of intermediating between two self-interested agents: a seller, who is willing to transfer a good only if the offered price meets at least their private valuation, and a buyer, who is willing to acquire the good only if the price does not exceed their private valuation.
Despite its apparent simplicity, a classical impossibility theorem by [20] shows that, under mild assumptions, no mechanism can simultaneously satisfy efficiency, incentive compatibility, individual rationality, and budget balance. This result holds even in Bayesian settings where the distributions of valuations are known to the mechanism designer.
As a consequence, a large body of subsequent work has focused on designing approximately efficient mechanisms — though still requiring strong informational assumptions such as complete knowledge of prior distributions. See [2, 5, 12–15, 17] for recent advances on constant-factor approximations of gain from trade.
More recently, a different line of research has addressed these limitations from a new online learning point of view. See [1, 4, 8, 9, 11, 19] for a selection of representative works. In particular, [9] introduced a regret-minimization framework for bilateral trade, in which the mechanism repeatedly interacts with i.i.d. seller-buyer pairs by posting prices without prior knowledge of the valuation distributions, and performance is measured via regret with respect to the best fixed price in hindsight. Within this framework, efficiency is again pursued approximately, as performance is assessed via regret against an appropriate efficiency benchmark, incentive compatibility and individual rationality are inherently satisfied by the posted-price model, since agents can only accept or reject the offered prices, while budget balance is explicitly modeled as a constraint that must be enforced by the learning algorithm.
Building on this paradigm, the role of the budget constraint was revisited in subsequent work, as illustrated in the following table.

Environment	Assumptions	Budget	Feedback	Benchmark	Regret
Stochastic (i.i.d.)	–	SBB	Full	Best fixed price	$\widetilde{O}(\sqrt{T})$
Stochastic (i.i.d.)	iv + bd	SBB	Realistic	Best fixed price	$\widetilde{O}(T^{2/3})$
Stochastic (i.i.d.)	bd	WBB	Realistic	Best fixed price	$\widetilde{O}(T^{3/4})$
Adversarial	–	GBB	Full	Best fixed price	$\widetilde{O}(\sqrt{T})$
Adversarial	–	GBB	Realistic	Best fixed price	$\widetilde{O}(T^{3/4})$

Table 1: Known asymptotical regret upper bounds.

In particular, [9] showed that under realistic feedback — where the learner only observes the agents' acceptance decisions — any mechanism that posts a single price to both agents incurs linear regret as soon as the independence assumption between seller and buyer valuations (iv) is dropped. This motivated relaxing per-round budget balance to weak budget balance – which allows the mechanism to post two different prices to seller and buyer while requiring non-negative profit at each round, restoring learnability in stochastic environments under bounded density assumption (bd) at the price of a slower regret rate: under realistic feedback, the best achievable regret against the best fixed feasible price then becomes $\widetilde{O}(T^{3/4})$. Meanwhile, to address the limitations of adversarial environments, [4] introduced global budget balance – which requires budget balance only in aggregate over the entire horizon – allowing temporary deficits as long as cumulative profit remains non-negative; under this notion, sublinear regret becomes attainable against adversarial sequences under partial feedback, with regret $\widetilde{O}(T^{3/4})$ with respect to the best fixed price in hindsight. See [11, 19] for related developments on budget-feasible learning guarantees under global budget balance and partial feedback.
However, competing only with the best fixed price — corresponding to posting identical prices to seller and buyer — does not fully exploit the additional flexibility enabled by global budget balance: once the mechanism is allowed to post arbitrary pairs of prices, it becomes natural to benchmark performance against the best distribution over price pairs that is globally budget balanced in expectation. See [3, 7, 16] for examples of distributional benchmarks in online learning with expectation constraints. While [4] show that this distributional benchmark is unattainable with sublinear regret in adversarial environments — even under full feedback — it remains open whether it can be approached in stochastic settings.
The present work addresses this question by studying the stochastic i.i.d. environment under realistic feedback — by proposing a budget-balanced learning algorithm based on a novel estimator of the gain from trade tailored to this feedback model — and shows that a pseudo-regret bound of $\widetilde{O}(T^{3/4})$ can be achieved against the distributional benchmark.

1.2. Overview of Results

The main goal of this work is to determine whether the new distributional benchmark is attainable in a stochastic environment under realistic feedback, and to provide an algorithm that achieves sublinear regret in this setting. The main result is an upper bound of $\widetilde{O}(T^{3/4})$ on the pseudo-regret in this setting, under the assumption of bounded density of the valuations. Notably, this rate matches the canonical $\widetilde{O}(T^{3/4})$ regret previously established for learning the best fixed feasible price under weak budget balance, despite the distributional benchmark being strictly stronger.
The presentation begins with the full feedback model: this section does not address the full feedback problem per se, but rather serves as an entry point to the core ideas underlying the algorithm developed for the realistic feedback setting, with particular focus on the gain from trade optimization step. The discussion then moves to the bandit model, in which only the realized gain from trade of the chosen action is observed; this intermediate case retains analytical simplicity by laying out some key computations on which the final algorithm is built — again, with particular attention to the gain from trade optimization. Finally, the realistic feedback model is analyzed in the subsequent section, where the $\widetilde{O}(T^{3/4})$ pseudo-regret bound under global budget balance is established.
In order to obtain the aforesaid result, the bounded joint density assumption on the valuations is essential for learnability under realistic feedback: without it — as shown by [9] — realistic feedback admits instances in which every algorithm incurs linear regret, since any discrete approximation of the action space fails to capture the optimal price.

1.3. Key Concepts

Here the central notions that underlie the approach are highlighted.

Online learning algorithm. At its core, an online learning algorithm repeatedly interacts with an environment by selecting an action at each round, observing some feedback, and updating its strategy accordingly. See [10] for background on regret minimization in repeated decision-making. The algorithm's objective is to maximize a performance metric — such as the cumulative gain from trade — leveraging as a source of information the feedback associated with the chosen actions. This paradigm provides the formal framework in which the problem of bilateral trade is dealt with under budget constraints.

Stochastic environment. The realizations (s_t, b_t) of the valuations (S_t, B_t) of seller and buyer at round t are drawn i.i.d. from an unknown joint distribution $\mathcal{D}$ over $[0,1]^2$. The stochastic i.i.d. setting differs from adversarial models in that it allows the learner to exploit statistical regularities of $\mathcal{D}$ in order to converge to near-optimal performance. This statistical assumption is key to making the distributional benchmark attainable.

Action space. At each round, the mechanism chooses a pair of prices $(p, q) \in [0,1]^2$, where p is offered to the seller and q to the buyer. Unlike the restricted one-dimensional setting of [9] — where the mechanism is limited to choose a fixed price p, or, equivalently, a pair of prices (p, q) on the diagonal $p = q$ — this formulation involves operating in a two-dimensional action space, increasing the learning complexity but enabling deeper exploration of the action space. Notably, any implementable procedure must in practice operate on a finite set of price pairs, which requires approximating the continuum by a discrete grid.

Regularity and discretization. To solve a problem defined on a continuous space, any learning algorithm must discretize it onto a grid. See [6, 9, 18, 21] for examples of bandit optimization over continuous action spaces, where regularity controls discretization error. In particular, as proven in [9], the absence of a bounded density assumption on the valuations (S_t, B_t), the realistic feedback model admits hard instances in which no learning algorithm can recover meaningful information about the valuation distribution. As a consequence, assuming some regularity — such as assuming that the valuations (S, B) admit density $f_{S,B}$ bounded by $\|f_{S,B}\|_\infty \leq \sigma$ — is strictly necessary.

Objective functions. Two performance measures play a central role in the present work, namely the gain from trade and the profit

$$\begin{aligned} \mathrm{GFT}(p, q, s, b) &= (b - s) \cdot \mathbb{I}[\, s \leq p \,] \cdot \mathbb{I}[\, q \leq b \,] \\ \mathrm{Pro}(p, q, s, b) &= (q - p) \cdot \mathbb{I}[\, s \leq p \,] \cdot \mathbb{I}[\, q \leq b \,] \end{aligned}$$

These two quantities are intrinsically linked by a fundamental trade-off: gain from trade is maximized when the two prices (p, q) are close to the diagonal $p = q$, while profit increases with the spread $q - p$.

In bilateral trade, the primary objective is to maximize gain from trade, since it differs from social welfare only by the additive term s_t — which is independent of the posted prices and hence irrelevant for regret minimization; however, accumulating profit is essential for enabling more effective learning strategies while enforcing budget balance.
Here, it is convenient to introduce notation $\mathrm{GFT}(p,q) = \mathbb{E}[\mathrm{GFT}(p,q,S,B)]$ and $\mathrm{Pro}(p,q) = \mathbb{E}[\mathrm{Pro}(p,q,S,B)]$ since any online learning algorithm ultimately reasons about these expected quantities rather than their per-round realizations.

Partial feedback. In real settings, the learner does not directly observe the agents' valuations — as in full feedback — but only limited acceptance information. Two standard partial-feedback models are: *bandit feedback*, where only the realized gain from trade for the posted price pair is observed at each round t; and *realistic feedback*, which hides the obtained gain from trade and only reports the acceptance events. The realistic feedback model is further divided into two variants: *two-bit feedback*, in which the learner observes the seller's and the buyer's acceptance decisions separately, and *one-bit feedback*, in which only the occurrence of the trade is revealed.

Distributional benchmark. Up to date, previous analyses compare the learner against the best fixed budget balanced price pair in hindsight [4, 9]. In contrast, the new distributional benchmark considers the best feasible distribution γ over $[0,1]^2$ that is globally budget balanced in expectation. This comparator is surely not weaker, since the best fixed price is included as a degenerate distribution supported on a single pair of prices on the diagonal $p = q$, and also strictly stronger, even if [4] prove that its gain from trade is at most twice that of the best fixed price. To illustrate the additional power of the distributional benchmark, consider the following simple stochastic instance.
Example. Let the joint distribution of (S,B) assign probability $1/3$ to each of the valuation pairs $(\frac{3}{5},1)$, $(0,1)$, and $(0,\frac{2}{5})$. Consider the fixed price pair $(p,q) = (\frac{3}{5},\frac{2}{5})$: this pair trades with probability one and yields a high gain from trade, but incurs a negative profit whenever trade occurs, hence it is infeasible under any per-round budget constraint. On the other hand, the price pair $(p,q) = (0,1)$ yields zero gain from trade with probability $2/3$, but produces with probability $1/3$ a large positive profit when $(S,B) = (0,1)$. A randomized policy that plays $(0,1)$ with small probability and $(\frac{3}{5},\frac{2}{5})$ otherwise is globally budget balanced in expectation and achieves strictly higher expected gain from trade than any fixed price pair.

Three-Phase approach. The algorithm developed in this work follows a three-phase structure. In the Budget Accumulation phase (Phase I), it collects the budget required to support future deficit-incurring actions. The Pure Exploration phase (Phase II) actively probes the action space to construct accurate gain from trade estimates based on realistic feedback. Finally, in the Optimization phase (Phase III), the algorithm solves an optimistic linear program over a discretized grid of price pairs, using these estimates to guide online learning under the limited information provided by realistic feedback.

2. Repeated Bilateral Trade

2.1. Framework

Bilateral trade models the interaction between a seller with valuation s and a buyer with valuation b, where a trade occurs only if the posted prices satisfy $s \leq p$ and $q \leq b$. The present work considers the problem of repeated bilateral trade in an online learning setting under a stochastic i.i.d. model.

Stochastic environment. Throughout this work, *stochastic* refers to the assumption that the sequence of valuation pairs $(S_t, B_t)_{t=1}^T$ is generated i.i.d. from a fixed but unknown joint distribution $\mathcal{D}$ over a bounded support (here $[0,1]^2$). This assumption does not impose any parametric form: $\mathcal{D}$ may be arbitrary, possibly even supported on finitely many points (discrete), or admitting a density (continuous), or a mixture of both. Moreover, no independence between seller and buyer valuations is required: within each round, S_t and B_t may be arbitrarily dependent (correlated), and only independence across time is assumed.

The learning protocol. At each round $t = 1, 2, \ldots, T$, a new seller-buyer pair arrives with private valuations $(s_t, b_t) \in [0,1]^2$, drawn independently and identically from an unknown joint distribution $\mathcal{D}$. The learner posts a price $p_t \in [0,1]$ to the seller and a price $q_t \in [0,1]$ to the buyer. A trade occurs if and only if both agents accept their respective prices, namely $\mathbb{I}\{s_t \leq p_t\}$ and $\mathbb{I}\{q_t \leq b_t\}$, regardless of the relative order between p_t and q_t. The mechanism is awarded with gain from trade $\mathrm{GFT}(p_t, q_t, s_t, b_t)$.

Global budget balance. During the execution, the learner maintains a cumulative budget BB_t, initialized as $BB_0 = 0$ and updated after each round as

$$BB_t = BB_{t-1} + \mathrm{Pro}(p_t, q_t, s_t, b_t)$$

To avoid notational overload, the cumulative budget is denoted by BB_t, reserving B for the buyer valuation. Notice that global budget balance is said to hold if and only if $BB_T \geq 0$. In the algorithm considered in this work, global budget balance is enforced by design.

Feedback structure. The present work considers the following feedback models:

- *Full feedback:* The learner transparently observes (s_t, b_t) at each round t, so it is possible to compute the $\mathrm{GFT}(p, q, s_t, b_t)$ for all (p, q).
- *Bandit feedback:* The learner directly observes the reward $\mathrm{GFT}(p, q, s_t, b_t)$ only for the played couple (p_t, q_t) at each round t.
- *One-bit feedback:* The learner only observes the trade indicator $z_t = \mathbb{I}\{s_t \leq p_t\} \cdot \mathbb{I}\{q_t \leq b_t\}$ only for the played couple (p_t, q_t) at each round t.

This work focuses on the one-bit feedback model, in which the learner observes only whether both agents accept the posted price. For readability, all results are stated by explicitly referring to the two acceptance bits separately as in the two-bit feedback model; however, the analysis extends verbatim to the strictly weaker one-bit setting, since the two indicators are never exploited separately and only appear through their product.

Benchmark. Let $\Delta([0,1]^2)$ denote the set of probability measures over $[0,1]^2$. The most natural benchmark in this setting is the best fixed distribution γ^* in $\Delta([0,1]^2)$ that satisfies global budget balance in expectation. The benchmark value is defined as

$$\begin{aligned} \mathrm{OPT} := \sup_{\gamma \in \Delta([0,1]^2)} & \mathbb{E}_{(p,q)\sim\gamma}[\mathrm{GFT}(p,q)] \\ \text{s.t.} \quad & \mathbb{E}_{(p,q)\sim\gamma}[\mathrm{Pro}(p,q)] \geq 0 \end{aligned}$$

where the expectation is taken over the randomized price pairs $(p, q) \sim \gamma$, while the expected $\mathrm{GFT}(p, q)$ is taken over valuations $(s, b) \sim \mathcal{D}$.

Regret with respect to the benchmark. Given a learning algorithm $\mathcal{A}$, its pseudo-regret is defined as

$$\mathcal{R}_T(\mathcal{A}) = T \cdot \mathrm{OPT} - \mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T} \mathrm{GFT}(p_t, q_t)\right]$$

where the outer expectation is taken over both the algorithm's randomization and the i.i.d. valuation sequence drawn from $\mathcal{D}$. Given the stochastic nature of the environment, this work focuses on pseudo-regret rather than

the general definition of (realized) regret, in line with the existing literature. In the following, the *pseudo-regret* will just be referred to as *regret*.

2.2. Further Assumptions

A central technical requirement in this work is a mild regularity condition on the distribution of valuations of the buyer and the seller. In the realistic feedback model, the joint law of (S, B) will be assumed to admit a density $f_{S,B}$ bounded by a constant σ, or, equivalently, it will be assumed that $\|f_{S,B}\|_\infty \leq \sigma$.
This bounded density hypothesis is not merely convenient but necessary for learnability under partial feedback. As shown in [9], if valuations concentrate on sets of arbitrarily small measure, any discretization of the action space fails to capture the optimal prices, and every algorithm suffers linear regret under realistic feedback. Intuitively, acceptance events convey too little information to distinguish between neighbouring prices, making the reconstruction of gain from trade signals impossible. Formally, the bounded density condition guarantees that the expected gain from trade $\mathrm{GFT}(p, q)$ is 4σ-Lipschitz on $[0, 1]^2$, which in turn implies that the discretized grid approximates the continuous distributional benchmark with error $O(1/K)$, where K determines the resolution of the price grid.
This regularity assumption is rather natural in economic environments: it excludes highly pathological valuation distributions but accommodates all continuously distributed preferences typically considered in applied models. As such, it strikes a balance between analytical tractability and modelling realism, while remaining substantially weaker than requiring smoothness or parametric structure.

3. Price Discretization and Algorithm

This section presents the algorithmic pipeline in detail, together with the discrete grids that support its action space. First, the discrete sets of price pairs that serve as supports for the learning algorithm are described. Then, the algorithmic pipeline is introduced.

3.1. Additive Grid for Gain From Trade

During the second (Pure Exploration) and third (Optimization) phases, the continuous price space $[0,1]^2$ is approximated with a homogeneous two-dimensional grid, namely

$$\mathcal{G}_K \;=\; \left\{ \frac{i}{K-1} : i = 0, \dots, K-1 \right\}^2$$

Thus, $|\mathcal{G}_K| = K^2$ and the grid resolution is $1/(K-1)$.
Now, let $\eta_K \geq 0$ be a (small) discretization slack to be specified later. From this point on, the reader should mind the difference between the benchmark γ^* and the *benchmark on the additive grid* γ^*_K with value

$$\mathrm{OPT}_K \;=\; \sup_{\gamma \in \Delta(\mathcal{G}_K)} \sum_{(p,q) \in \mathcal{G}_K} [\, \gamma(p,q)\, \mathrm{GFT}(p,q) \,]$$

$$\text{s.t.} \quad \sum_{(p,q) \in \mathcal{G}_K} [\, \gamma(p,q)\, \mathrm{Pro}(p,q) \,] \geq -\eta_K$$

From this point on, γ_K will be referred to as the *benchmark on the grid*.
The reader should notice that, under the bounded density assumption, the optimal continuous policy can outperform its nearest neighbour projection onto the grid only by a vanishing amount. However, nearest neighbour projections need not preserve feasibility, which would prevent a clean decomposition of regret into a learning term with respect to OPT_K and a discretization error. Instead, allowing a vanishing slack $-\eta_K$ restores feasibility and ensures that OPT_K dominates any nearest neighbour projection in terms of GFT.

3.2. Multiplicative Grid for Profit

Consider the two following sets of pairs of prices

$$\mathcal{F}_K^- \;=\; \left\{ \big(p - 2^{-a},\, p\big) \in [0,1]^2 :\; p = \frac{i}{K-1},\; i = 0, \dots, K-1,\; a = 0, \dots, \lfloor \log T \rfloor \right\}$$

$$\mathcal{F}_K^+ \;=\; \left\{ \big(p,\, p + 2^{-a}\big) \in [0,1]^2 :\; p = \frac{i}{K-1},\; i = 0, \dots, K-1,\; a = 0, \dots, \lfloor \log T \rfloor \right\}$$

The first phase (Budget Accumulation) runs on their union

$$\mathcal{F}_K \;=\; \mathcal{F}_K^- \cup \mathcal{F}_K^+$$

In contrast to the full additive grid, which has $|\mathcal{G}_K| = \Theta(K^2)$, the multiplicative grid is a sparse band above the diagonal containing $|\mathcal{F}_K| = \Theta(K \log(T))$ points overall. The reader should also notice that the grid does not include points lying below the diagonal. Consequently, the multiplicative grid is contained in the upper triangle — where $q > p$ and thus both the per-round profit and the per-round gain from trade are certainly non-negative.

3.3. Three-Phase Learning Algorithm

The algorithm takes as input a budget threshold β, a discretization parameter K, and a probe batch size m for the Pure Exploration phase, which will be described later. From this point on, the three components of the procedure are referred to as Phase I, Phase II, and Phase III.
The algorithmic structure is designed so that Phase I and Phase III are shared across realistic, bandit, and full feedback models, which therefore rely on the same underlying pipeline. In contrast, Phase II is unnecessary under full and bandit feedback, since the gain from trade is directly observed for the chosen action. The algorithm for realistic feedback can thus be interpreted as an extension of this common structure, obtained by inserting Phase II between Phase I and Phase III.

The algorithm *GFT-Max* for full and bandit feedback is defined as the execution of Phase I and Phase III only:

Algorithm 1 GFT-Max

Require: threshold $\beta > 0$, grid size $K \in \mathbb{N}$, horizon T, probe batch m

define $\mathcal{F}_K$, $\mathcal{G}_K$, $\mathcal{A}_P$, $\mathcal{A}_G$ and initialize $BB \leftarrow 0$ and $t \leftarrow 1$
while $t \leq T$ **and** $BB < \beta$ **do** ▷ Phase I
  $(p_t, q_t) \leftarrow \mathcal{A}_P$
  play (p_t, q_t), observe feedback z_t; update $\mathcal{A}_P$ with z_t
  $BB \leftarrow BB + \mathrm{Pro}_t$; $t \leftarrow t + 1$
end while
while $t \leq T$ **and** $BB > 0$ **do** ▷ Phase III
  $(p_t, q_t) \leftarrow \mathcal{A}_G$
  play (p_t, q_t), observe feedback z_t; update $\mathcal{A}_G$ with z_t
  $BB \leftarrow BB + \mathrm{Pro}_t$
  $t \leftarrow t + 1$
end while
return sequence $\{(p_t, q_t)\}_{t=1}^{T}$

The algorithm *GFT-Max 2.0* corresponds to the full three-phase procedure for the realistic feedback model:

Algorithm 2 GFT-Max 2.0

Require: threshold $\beta > 0$, grid size $K \in \mathbb{N}$, horizon T, probe batch m

define $\mathcal{F}_K$, $\mathcal{G}_K$, $\mathcal{A}_P$, $\mathcal{A}_G$ and initialize $BB \leftarrow 0$ and $t \leftarrow 1$
while $t \leq T$ **and** $BB < \beta$ **do** ▷ Phase I
  $(p_t, q_t) \leftarrow \mathcal{A}_P$
  play (p_t, q_t), observe feedback z_t; update $\mathcal{A}_P$ with z_t
  $BB \leftarrow BB + \mathrm{Pro}_t$; $t \leftarrow t + 1$
end while
if realistic feedback **and** $t + 2mK \leq T$ **then** ▷ Phase II
  for each $p \in \{0, \frac{1}{K-1}, \ldots, 1\}$ **do**
    for $r = 1$ **to** m **do**
      play a probing round (p, U_r); observe z_t
      $BB \leftarrow BB + \mathrm{Pro}_t$
      $t \leftarrow t + 1$; **if** $t > T$ **break**
    end for
  end for
  for each $q \in \{0, \frac{1}{K-1}, \ldots, 1\}$ **do**
    for $r = 1$ **to** m **do**
      play a probing round (V_r, q); observe z_t
      $BB \leftarrow BB + \mathrm{Pro}_t$
      $t \leftarrow t + 1$; **if** $t > T$ **break**
    end for
  end for
  compute $\widehat{R}(p, q)$, $\widehat{L}(p, q)$ on $\mathcal{G}_K$
  set $\overline{\mathrm{GFT}}_t(p, q) \leftarrow \widehat{R}(p, q) + \widehat{L}(p, q) + \overline{\mathrm{Pro}_t}(p, q)$
end if
while $t \leq T$ **and** $BB > 0$ **do** ▷ Phase III
  $(p_t, q_t) \leftarrow \mathcal{A}_G$
  play (p_t, q_t), observe feedback z_t; update $\mathcal{A}_G$ with z_t
  $BB \leftarrow BB + \mathrm{Pro}_t$
  $t \leftarrow t + 1$
end while
return sequence $\{(p_t, q_t)\}_{t=1}^{T}$

In this algorithm, Phase I accumulates profit on the multiplicative grid $\mathcal{F}_K$ until the desired budget β is reached. Phase II may consume up to $2mK$ units of budget in the worst case; therefore, the accumulated budget must satisfy $\beta \geq 2mK + \beta_{\text{III}}$, where β_{III} denotes the budget required to safely execute Phase III, to be defined later. Phase III (Optimization) is executed only as long as the budget BB remains positive and is halted otherwise. As a consequence, the entire procedure is globally budget balanced by construction, since the cumulative budget is never negative at termination.

3.4. Following Analysis

As anticipated, the next section considers the full-feedback model, introducing the discretized formulation of the problem, the use of empirical estimators, and the structure of the optimistic linear program. This setting is not meant to provide a solution to the full-feedback scenario, but rather to serve as an introduction to the underlying mechanics of gain from trade optimization on a finite grid. The main focus of this section is to lay out the computational structure underlying Phase III (Optimization).
The subsequent section turns to the bandit model, which retains the essential optimization components while clarifying the role of estimates derived solely from the played actions in the required computations. Bandit feedback is not commonly considered in bilateral trade and — just like the previous — serves primarily as a warm-up for the reader. Again, the calculations developed in this section are mainly devoted to Phase III (Optimization).
Finally, in the section on realistic feedback, Phase II (Pure Exploration) becomes necessary to reconstruct the gain from trade signal from limited acceptance information before applying the same bandit-style optimization in Phase III (Optimization) introduced in the previous section. This section also addresses Phase I (Budget Accumulation) and Phase II (Pure Exploration).
The progression from full to bandit and finally to realistic is not merely sequential but conceptual: the full feedback model establishes the analytical foundation of the method and familiarizes the reader with the framework, the bandit model isolates some core calculations that are subsequently reused, and the realistic feedback model integrates these components to complete the presentation of the three-phase structure.

4. Full Feedback

In this preliminary section, the full feedback model is analyzed, in which agents reveal their valuations (s_t, b_t) at the end of each round t. The purpose of this section is not to provide a complete solution to the full feedback problem, but rather to introduce — in a simplified setting — the optimization structure and the linear program that will later be employed under realistic feedback. Here — as much as in the bandit case — the algorithm follows the two-phase template *GFT-Max*, since the gain from trade can be directly computed from the observed valuations. The focus of the analysis is therefore on the behaviour of the algorithm during Phase III.

Define the empirical means

$$\overline{\mathrm{GFT}_t}(p,q) \;:=\; \frac{1}{t}\sum_{\tau=1}^{t} \mathrm{GFT}(p,q,s_\tau,b_\tau)$$

$$\overline{\mathrm{Pro}_t}(p,q) \;:=\; \frac{1}{t}\sum_{\tau=1}^{t} \mathrm{Pro}(p,q,s_\tau,b_\tau)$$

Evidently, the expected value of the two random functionals $\overline{\mathrm{GFT}}(p,q)$ and $\overline{\mathrm{Pro}}(p,q)$ equal the expected values $\mathrm{GFT}(p,q)$ and $\mathrm{Pro}(p,q)$ respectively.
During Phase I, UCB is employed as regret minimizer $\mathcal{A}_P$ to maximize $\overline{\mathrm{Pro}_t}(p,q)$ over the multiplicative grid $\mathcal{F}_K$ and gather profit. Then, during the second and last phase (Phase III, Optimization) the algorithm $\mathcal{A}_G$ defines the following linear program over the additive grid $\mathcal{G}_K$ at every round t, based on the two following optimistic estimators

$$\max_{\gamma_t \in \Delta(\mathcal{G}_K)} \quad \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, (\overline{\mathrm{GFT}_t}(p,q) + \phi_t) \,\right]$$

$$\text{s.t.} \quad \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, (\overline{\mathrm{Pro}_t}(p,q) + \phi_t) \,\right] \geq -\,\eta_K$$

where $\phi_t : \mathbb{N} \to \mathbb{R}$ and $\eta_K \geq 0$. Then, the mechanism samples $(p_t, q_t) \sim \gamma_t$ from the solution distribution and plays that pair.
It is important to note that an optimistic estimation of GFT enables lower regret (as in classic UCB), while an optimistic estimation of profit enables deeper exploration near the diagonal $p = q$, relying on having accumulated sufficient budget during Phase I.

4.1. Optimization Phase: Regret over the Grid

The following lemma provides a uniform bound on the estimation error with respect to OPT_K, along with a margin for the maximum budget needed to play the linear program solution, during the Optimization.

Lemma 4.1. *With probability at least* $1-\delta$

$$\begin{cases} \displaystyle\sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, \mathrm{GFT}(p,q) \,\right] \;\geq\; T\,\mathrm{OPT}_K \;-\; 4\sqrt{2\,T\log(\frac{4K^2T}{\delta})} \\ \displaystyle\sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, \mathrm{Pro}(p,q) \,\right] \;\geq\; -\,T\,\eta_K \;-\; 4\sqrt{2\,T\log(\frac{4K^2T}{\delta})} \end{cases}$$

Proof. Step 1: Calibration of ϕ_t. For each pair $(p,q) \in \mathcal{G}_K$ and round $t \in \{1,\dots,T\}$, since $\mathrm{GFT}(p,q,s_t,b_t) \in [-1,1]$ and $\mathrm{Pro}(p,q,s_t,b_t) \in [-1,1]$, applying Hoeffding for any $\phi_t > 0$ yields the events

$$E^{\mathrm{GFT}}_{t,(p,q)} \;:=\; \left|\, \overline{\mathrm{GFT}_t}(p,q) - \mathrm{GFT}(p,q) \,\right| \;\leq\; \phi_t$$

$$E^{\mathrm{Pro}}_{t,(p,q)} \;:=\; \left|\, \overline{\mathrm{Pro}_t}(p,q) - \mathrm{Pro}(p,q) \,\right| \;\leq\; \phi_t$$

each holding with probability at least $1 - 2\exp\left\{-\frac{t\,\phi_t^2}{2}\right\}$.

Define the clean event $\mathcal{E}_\delta$ as the simultaneous intersection over all times t and all grid pairs (p,q), as

$$\mathcal{E}_\delta \;:=\; \bigcap_{t=1}^{T} \bigcap_{(p,q)\in\mathcal{G}_K} \left(E^{\mathrm{GFT}}_{t,(p,q)} \;\cap\; E^{\mathrm{Pro}}_{t,(p,q)} \right)$$

By applying the union bound

$$\begin{aligned}\mathrm{P}[\,\mathcal{E}_\delta^c\,] \;&\leq\; \sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K}\Big[\,\mathrm{P}[\,(E^{\mathrm{GFT}}_{t,(p,q)}\,)^c\,]\;+\;\mathrm{P}[\,(E^{\mathrm{Pro}}_{t,(p,q)})^c\,]\,\Big]\\ &\leq\; 4K^2\sum_{t=1}^{T}\exp\!\left\{-\frac{t\,\phi_t^2}{2}\right\}\end{aligned}$$

To guarantee $\mathrm{P}(\mathcal{E}_\delta)\geq 1-\delta$, require for all $t\in[T]$

$$\exp\!\left\{-\frac{t\,\phi_t^2}{2}\right\}\leq\frac{\delta}{4K^2T}\quad\Longleftrightarrow\quad\phi_t\;=\;\sqrt{\frac{2}{t}\,\log(\frac{4K^2T}{\delta})}$$

Step 2: Computation of the Gain From Trade of the program solution. On the clean event $\mathcal{E}_\delta$, $\gamma_k^\star$ is provably a feasible solution for the program, since

$$\sum_{(p,q)}\,[\,\gamma_k^\star(p,q)(\overline{\mathrm{Pro}_t}(p,q)+\phi_t)\,]\;\geq\;\sum_{(p,q)}\,[\,\gamma_k^\star(p,q)\;\mathrm{Pro}(p,q)\,]\;\geq\;-\eta_K$$

Using the symmetry of the Hoeffding bound, the optimality of γ_t for the linear program and $\sum_{(p,q)}\gamma_t(p,q)=1$

$$\begin{aligned}\sum_{(p,q)}\,[\,\gamma_t(p,q)\;\mathrm{GFT}(p,q)\,]\;&\geq\;\sum_{(p,q)}\,[\,\gamma_t(p,q)(\overline{\mathrm{GFT}_t}(p,q)-\phi_t)\,]\\ &=\;\sum_{(p,q)}\,[\,\gamma_t(p,q)(\overline{\mathrm{GFT}_t}(p,q)+\phi_t)\,]\;-\;2\phi_t\\ &\geq\;\sum_{(p,q)}\,[\,\gamma_k^\star(p,q)(\overline{\mathrm{GFT}_t}(p,q)+\phi_t)\,]\;-\;2\phi_t\\ &\geq\;\sum_{(p,q)}\,[\,\gamma_k^\star(p,q)\;\mathrm{GFT}(p,q)\,]\;-\;2\phi_t\\ &=\;\mathrm{OPT}_K-2\phi_t\end{aligned}$$

Finally, summing the GFT inequality over all T rounds gives

$$\begin{aligned}\sum_{t=1}^{T}\sum_{(p,q)}\,[\,\gamma_t(p,q)\;\mathrm{GFT}(p,q)\,]\;&\geq\;T\;\mathrm{OPT}_K\;-\;2\sum_{t=1}^{T}\phi_t\\ &\geq\;T\;\mathrm{OPT}_K\;-\;2\sum_{t=1}^{T}\sqrt{\frac{2}{t}\,\log(\tfrac{4K^2T}{\delta})}\\ &\geq\;T\;\mathrm{OPT}_K\;-\;2\,\sqrt{2\log(\tfrac{4K^2T}{\delta})}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\\ &\geq\;T\;\mathrm{OPT}_K\;-\;4\,\sqrt{2\,T\log(\tfrac{4K^2T}{\delta})}\end{aligned}$$

Step 3: The program solution is GBB. The second and last statement of the lemma is proven, similarly to what was done in the previous step

$$\begin{aligned}\sum_{(p,q)}\,[\,\gamma_t(p,q)\;\mathrm{Pro}(p,q)\,]\;&\geq\;\sum_{(p,q)}\,[\,\gamma_t(p,q)(\overline{\mathrm{Pro}_t}(p,q)-\phi_t)\,]\\ &=\;\sum_{(p,q)}\,[\,\gamma_t(p,q)(\overline{\mathrm{Pro}_t}(p,q)+\phi_t)\,]\;-\;2\phi_t\\ &\geq\;-\,\eta_K\;-\;2\phi_t\end{aligned}$$

Summing the profit inequality over all T rounds gives

$$\sum_{t=1}^{T} \sum_{(p,q)} [\, \gamma_t(p,q)\, \mathrm{Pro}(p,q)\,] \;\geq\; -\,T\,\eta_K \;-\; 2\sum_{t=1}^{T} \phi_t$$

$$\geq\; -\,T\,\eta_K \;-\; 4\sqrt{2T \log\!\left(\tfrac{4K^2T}{\delta}\right)}$$

□

The following corollary proves that the regret of the algorithm with respect to γ_K^* during the Optimization phase — assuming sufficient budget to run — is sublinear in the clean event.

Corollary 4.1. *For $\delta = \frac{1}{T}$*

$$T\,\mathrm{OPT}_K - \mathbb{E}\big[\, \sum_{t=1}^{T} \sum_{(p,q)} [\, \gamma_t(p,q)\, \mathrm{GFT}(p,q)\,]\, \big] \;=\; O(\, \sqrt{T \log(KT)}\,)$$

Proof.

$$T\,\mathrm{OPT}_K - \mathbb{E}\big[\, \sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} [\, \gamma_t(p,q)\, \mathrm{GFT}(p,q)\,]\, \big]$$

$$\leq\; T\,\mathrm{OPT}_K \;-\; [\, (1-\delta)\, (\, T\,\mathrm{OPT}_K - 4\sqrt{2\,T \log(\frac{4K^2T}{\delta})}\,)\,]$$

$$=\; \delta\, T\,\mathrm{OPT}_K \;+\; 4\,(1-\delta)\sqrt{2\,T \log(\frac{4K^2T}{\delta})}$$

$$=\; \mathrm{OPT}_K \;+\; 8\,(1-\tfrac{1}{T})\,\sqrt{T \log(2KT)}$$

$$\leq\; 1 \;+\; 8\,\sqrt{T \log(2KT)}$$

$$=\; O(\, \sqrt{T \log(KT)}\,)$$

□

This section served as a simple illustrative case showing how to tackle the maximization problem over a discretized grid assuming sufficient budget, and provides the conceptual foundation for the following sections — where the learner no longer observes full information but only a partial feedback.

5. Bandit Feedback

In this section, the bandit input model is analyzed, where at the end of each round t the learner only observes the feedback generated by the played pair (p_t, q_t). In this setting, the learner cannot directly evaluate $\mathrm{GFT}(p,q)$ and $\mathrm{Pro}(p,q)$ for unplayed pairs, and therefore relies on unbiased single-round estimators constructed from the bandit signal, as the estimation process is now based on the counts of played price pairs, while following the same two-phase template *GFT-Max*. As noted earlier, bandit feedback is not a standard model in the bilateral-trade literature, but it allows isolating optimization components that characterize the regret incurred by the algorithm during Phase III and will later be reused under realistic feedback.

Let $N_t(p,q)$ be the number of updates collected for pair (p,q) up to the beginning of round t. Define the unbiased empirical means

$$\overline{\mathrm{GFT}_t}(p,q) \;:=\; \frac{1}{N_t(p,q)} \sum_{\tau=1}^{t} \left[\, \mathbb{I}\{(p_\tau,q_\tau)=(p,q)\} \cdot \mathrm{GFT}(p,q,s_\tau,b_\tau) \,\right]$$

$$\overline{\mathrm{Pro}_t}(p,q) \;:=\; \frac{1}{N_t(p,q)} \sum_{\tau=1}^{t} \left[\, \mathbb{I}\{(p_\tau,q_\tau)=(p,q)\} \cdot \mathrm{Pro}(p,q,s_\tau,b_\tau) \,\right]$$

In Phase I, UCB is employed as regret minimizer $\mathcal{A}_P$. Then, during the second and last phase (Phase III, Optimization) the algorithm $\mathcal{A}_G$ uses optimistic estimates to compute a distribution over grid prices by solving the linear program

$$\begin{aligned} \max_{\gamma_t \in \Delta(\mathcal{G}_K)} \quad & \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, (\overline{\mathrm{GFT}_t}(p,q) + \phi_t(p,q)) \,\right] \\ \text{s.t.} \quad & \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, (\overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q)) \,\right] \geq -\,\eta_K \end{aligned}$$

where $\phi_t : \mathbb{N} \to \mathbb{R}$ and $\eta_K \geq 0$. Then, the mechanism samples $(p_t, q_t) \sim \gamma_t$ from the solution distribution and plays that pair.

5.1. Optimization Phase: Regret Upper Bound over the Grid

As in the previous section, the following lemma provides a bound on the estimation error with respect to OPT_K, along with a margin for the maximum budget needed to play the linear program solution.

Lemma 5.1. *With probability at least* $1-\delta$

$$\begin{cases} \displaystyle\sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, \mathrm{GFT}(p,q) \,\right] \;\geq\; T\, \mathrm{OPT}_K \;-\; 4K\,\sqrt{T}\,\log(\tfrac{8T}{\delta}) \\ \displaystyle\sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} \left[\, \gamma_t(p,q)\, \mathrm{Pro}(p,q) \,\right] \;\geq\; -\,T\,\eta_K \;-\; 4K\,\sqrt{T}\,\log(\tfrac{8T}{\delta}) \end{cases}$$

using for each pair (p,q) *the confidence bound* $\phi_t(p,q) \;=\; \sqrt{\frac{2}{N_t(p,q)}\,\log(\frac{8T}{\delta})}$.

Proof. Step 1: Calibration of ϕ_t. For each round $t \in \{1,\dots,T\}$, since $\mathrm{GFT}(p,q,s_t,b_t) \in [-1,1]$ and $\mathrm{Pro}(p,q,s_t,b_t) \in [-1,1]$, applying Hoeffding yields the events

$$\begin{aligned} E_t^{\mathrm{GFT}} \;&:=\; \left|\, \overline{\mathrm{GFT}_t}(p_t,q_t) - \mathrm{GFT}(p_t,q_t) \,\right| \;\leq\; \phi_t(p_t,q_t) \\ E_t^{\mathrm{Pro}} \;&:=\; \left|\, \overline{\mathrm{Pro}_t}(p_t,q_t) - \mathrm{Pro}(p_t,q_t) \,\right| \;\leq\; \phi_t(p_t,q_t) \end{aligned}$$

each holding with probability at least $1 - 2\exp\left\{-\frac{N_t(p,q)\;\phi_t(p,q)^2}{2}\right\}$.

Unlike in the previous section, since the optimistic estimator is updated for only one pair (p,q) at each time t, it suffices to apply a union bound over such t events to conclude that $\overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q) \geq \mathrm{Pro}_t(p,q) \quad \forall\,(p,q),\, t$, a property to be needed later.

Thus, define the clean event $\mathcal{E}_{\frac{\delta}{2}}$ as the simultaneous intersection over all t rounds

$$\mathcal{E}_{\frac{\delta}{2}} \;:=\; \bigcap_{t=1}^{T} \;\big(\; E_t^{\mathrm{GFT}} \;\cap\; E_t^{\mathrm{Pro}} \;\big)$$

By applying the union bound

$$\begin{aligned}
\mathrm{P}[\; \mathcal{E}_{\frac{\delta}{2}}^{c} \;] \;&\leq\; \sum_{t=1}^{T} \;\big[\; \mathrm{P}[\; (E_t^{\mathrm{GFT}})^{c} \;] \;+\; \mathrm{P}[\; (E_t^{\mathrm{Pro}})^{c} \;] \;\big] \\
&\leq\; 4 \sum_{t=1}^{T} \exp\!\left\{ -\frac{N_t(p,q)\,\phi_t^2(p,q)}{2} \right\}
\end{aligned}$$

To guarantee $\mathrm{P}(\mathcal{E}_{\frac{\delta}{2}}) \geq 1 - \frac{\delta}{2}$, require for all $t \in [T]$

$$\exp\!\left\{ -\frac{N_t(p,q)\;\phi_t(p,q)^2}{2} \right\} \leq \frac{\delta}{2} \cdot \frac{1}{4T} \quad\Longleftrightarrow\quad \phi_t(p,q) \;=\; \sqrt{\frac{2}{N_t(p,q)}\;\log(\frac{8T}{\delta})}$$

Step 2: Computation of the Gain From Trade of the program solution. On the clean event, as stated before, feasibility of $\gamma_K^{\star}$ for the optimistic linear program holds, since

$$\sum_{(p,q)} \;[\; \gamma_K^{\star}(p,q)(\overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q)) \;] \;\geq\; \sum_{(p,q)} \;[\; \gamma_K^{\star}(p,q)\;\mathrm{Pro}(p,q) \;] \;\geq\; -\,\eta_K$$

Hence, using the symmetry of the bound and the optimality of γ_t for the linear program

$$\begin{aligned}
\sum_{(p,q)} \;[\; \gamma_t(p,q)\;\mathrm{GFT}(p,q) \;] \;&\geq\; \sum_{(p,q)} \;[\; \gamma_t(p,q)(\overline{\mathrm{GFT}_t}(p,q) - \phi_t(p,q)) \;] \\
&=\; \sum_{(p,q)} \;[\; \gamma_t(p,q)(\overline{\mathrm{GFT}_t}(p,q) + \phi_t(p,q)) \;] \;-\; 2 \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\phi_t(p,q) \;] \\
&\geq\; \sum_{(p,q)} \;[\; \gamma_K^{\star}(p,q)(\overline{\mathrm{GFT}_t}(p,q) + \phi_t(p,q)) \;] \;-\; 2 \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\phi_t(p,q) \;] \\
&\geq\; \sum_{(p,q)} \;[\; \gamma_K^{\star}(p,q)\;\mathrm{GFT}(p,q) \;] - 2 \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\phi_t(p,q) \;] \\
&=\; \mathrm{OPT}_K \;-\; 2 \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\phi_t(p,q) \;]
\end{aligned}$$

Summing over $t = 1, \ldots, T$

$$\sum_{t=1}^{T} \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\mathrm{GFT}(p,q) \;] \;\geq\; T\;\mathrm{OPT}_K \;-\; 2 \sum_{t=1}^{T} \sum_{(p,q)} \;[\; \gamma_t(p,q)\;\phi_t(p,q) \;]$$

Let arms be indexed by $i = 1, \ldots, A$. Applying the Azuma-Hoeffding inequality gives that, with probability at least $1 - \frac{\delta}{2}$, the variable component of the nested sum on the right can be bounded as

$$\sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{G}_K|} \;[\; \frac{\gamma_t(i)}{\sqrt{N_t(i)}} \;] \;\leq\; \sum_{t=1}^{T} \;[\; \frac{1}{\sqrt{N_t(i_t)}} \;] \;+\; \sqrt{2T\log(\tfrac{2}{\delta})}$$

Also, it is possible to say that, for a fixed arm i, the count $N_t(i)$ increases by one each time i is played.

Leveraging the new indexing of the arms, it possible to rewrite the sum over the T rounds and apply Jensen's inequality to obtain that

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{\sqrt{N_t(i_t)}} \;&=\; \sum_{i=1}^{|\mathcal{G}_K|} \sum_{j=1}^{N_T(i)} \frac{1}{\sqrt{j}} \;\leq\; 2 \sum_{i=1}^{|\mathcal{G}_K|} \sqrt{N_{T+1}(i)} \\
&\leq\; 2K\sqrt{T}
\end{aligned}$$

Therefore, such variable component of the nested sum mentioned before can be bounded as

$$\sum_{t=1}^{T}\sum_{i=1}^{|\mathcal{G}_K|}\Big[\ \frac{\gamma_t(i)}{\sqrt{N_t(i)}}\ \Big]\ \leq\ 2K\sqrt{T}\ +\sqrt{2T\log(\tfrac{2}{\delta})}$$

Now, the overall bound for the nested sum can be written by multiplying by $\sqrt{2\log(\frac{8T}{\delta})}$

$$\begin{aligned}\sum_{t=1}^{T}\sum_{(p,q)}\big[\ \gamma_t(p,q)\ \phi_t(p,q)\ \big] &= \sqrt{2\log(\tfrac{8T}{\delta})}\ \sum_{t=1}^{T}\sum_{i=1}^{|\mathcal{G}_K|}\Big[\ \frac{\gamma_t(i)}{\sqrt{N_t(i)}}\ \Big]\\ &\leq\ \sqrt{2\log(\tfrac{8T}{\delta})}\ (\ 2K\sqrt{T}\ +\ \sqrt{2T\log(\tfrac{2}{\delta})}\ )\\ &\leq\ 2K\sqrt{2T}\ \log(\tfrac{8T}{\delta})\ +\ \sqrt{2\log(\tfrac{8T}{\delta})}\ \sqrt{2T\log(\tfrac{2}{\delta})}\\ &\leq\ 2K\sqrt{2T}\ \log(\tfrac{8T}{\delta})\ +\ 2\sqrt{T}\ \log(\tfrac{8T}{\delta})\\ &\leq\ 4K\ \sqrt{T}\ \log(\tfrac{8T}{\delta})\end{aligned}$$

Finally, if the clean event and the Azuma-Hoeffding bound happen to be simultaneously true, which occurs with probability at least $(1-\frac{\delta}{2})-\frac{\delta}{2}=1-\delta$, then the first claim holds

$$\sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K}\big[\ \gamma_t(p,q)\ \mathrm{GFT}(p,q)\ \big]\ \geq\ T\ \mathrm{OPT}_K\ -\ 4K\ \sqrt{T}\ \log(\tfrac{8T}{\delta})$$

Step 3: The program solution is GBB. If the clean event $\mathcal{E}_{\frac{\delta}{2}}$ and Azuma-Hoeffding bound hold simultaneously, which occurs with probability $1-\delta$, then it holds that for all (p,q)

$$\mathrm{Pro}(p,q)\ \geq\ \overline{\mathrm{Pro}_t}(p,q)\ -\ \phi_t(p,q)$$

Thus, it follows that

$$\begin{aligned}\sum_{(p,q)}\big[\ \gamma_t(p,q)\ \mathrm{Pro}(p,q)\ \big]\ &\geq\ \sum_{(p,q)}\big[\ \gamma_t(p,q)(\overline{\mathrm{Pro}_t}(p,q)-\phi_t(p,q))\ \big]\\ &=\ \sum_{(p,q)}\big[\ \gamma_t(p,q)(\overline{\mathrm{Pro}_t}(p,q)+\phi_t(p,q))\ \big]\ -2\ \sum_{(p,q)}\big[\ \gamma_t(p,q)\ \phi_t(p,q)\ \big]\\ &\geq\ -\eta_K\ -\ 2\ \sum_{(p,q)}\big[\ \gamma_t(p,q)\ \phi_t(p,q)\ \big]\end{aligned}$$

where the last inequality uses feasibility of γ_t for the optimistic constraint. Summing over t, similarly to what was done in the previous step, gives

$$\sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K}\big[\ \gamma_t(p,q)\ \mathrm{Pro}(p,q)\ \big]\ \geq\ -\,T\,\eta_K\ -\ 4K\ \sqrt{T}\ \log(\tfrac{8T}{\delta})$$

proving the second claim.

□

The following corollary proves that the regret of the algorithm with respect to γ_K^* (benchmark on the grid) during the Optimization phase — assuming sufficient budget to run — is sublinear in the clean event.

Corollary 5.1. *For* $\delta=\frac{1}{T}$

$$T\ \mathrm{OPT}_K-\mathbb{E}\Big[\ \sum_{t=1}^{T}\sum_{(p,q)}\big[\ \gamma_t(p,q)\ \mathrm{GFT}(p,q)\ \big]\ \Big]\ =\ O(\ K\sqrt{T}\log(T)\ )$$

Proof.

$$\begin{aligned}&T\ \mathrm{OPT}_K\ -\ \mathbb{E}\Big[\ \sum_{t=1}^{T}\sum_{(p,q)\in\mathcal{G}_K}\big[\ \gamma_t(p,q)\ \mathrm{GFT}(p,q)\ \big]\ \Big]\\ &\leq\ T\ \mathrm{OPT}_K\ -\big[\ (1-\delta)\ (\ T\ \mathrm{OPT}_K\ -\ 4K\ \sqrt{T}\ \log(\tfrac{8T}{\delta})\ )\ \big]\end{aligned}$$

$$
\begin{aligned}
&= \delta\, T\, \mathrm{OPT}_K \;+\; 4K\,(1-\delta)\,\sqrt{T}\,\log(\tfrac{8T}{\delta}) \\
&= \mathrm{OPT}_K \;+\; 8K\,(1-\tfrac{1}{T})\,\sqrt{T}\,\log(\sqrt{8}\,T) \\
&\leq\; 1 \;+\; 8K\,\sqrt{T}\log(\sqrt{8}\,T) \\
&= O(\; K\sqrt{T}\,\log(T)\;)
\end{aligned}
$$

□

The bandit model captures the scenario in which the learner observes only the realized gain from trade for the chosen action, and must therefore rely on unbiased empirical estimates to feed the linear program used in the Optimization phase (Phase III).

In the realistic feedback setting, after Phase II (Pure Exploration) the learner constructs an estimator of the gain from trade based on acceptance information, which can then be used in exactly the same way as under bandit feedback, thereby reusing the computations developed in this section. This is the reason for including the bandit model here, even though it has not been previously studied in the bilateral trade literature. Moreover, under realistic feedback the profit signal is of bandit type, so similar ideas as in the analysis carried out in this section directly apply to profit under the realistic feedback model.

6. Realistic Feedback

This section addresses the realistic feedback $\mathbb{I}\{s_t \le p_t\},\ \mathbb{I}\{q_t \le b_t\}$, that is the feedback model of interest. Here, the regret guarantees stated in the introduction are established.

Here, the optimization ideas developed in the full feedback setting and the bandit-style analysis of the previous section will be combined, and adapted to the more limited information structure where only acceptance decisions are observed. In particular, the algorithm pipeline introduced earlier is now instantiated in its three-phase form *GFT-Max 2.0*: Phase I (Budget Accumulation) runs bandit UCB as regret minimizer $\mathcal{A}_p$ on the multiplicative grid $\mathcal{F}_K$ to gather sufficient budget for the subsequent phases to run; Phase II (Pure Exploration) uses a probe-based procedure to construct estimates of the gain from trade on the additive grid $\mathcal{G}_K$; finally, in Phase III (Optimization) the algorithm $\mathcal{A}_G$ solves an optimistic linear program over $\mathcal{G}_K$ at each round t as in the previous. In particular, Phase I must also compensate the worst-case budget consumption of Phase II, which may be as large as $2mK$ since probing rounds can yield profit as low as -1.

6.1. Estimator of the Gain From Trade

This chapter constructs an unbiased estimator of $\mathrm{GFT}(p,q)$ tailored to this setting, to be built through Phase II and then used during Phase III. The adopted strategy generalizes the gain from trade decomposition introduced by [9] for the diagonal case $p = q$, extending it to asymmetric price pairs (p,q). The key step is to include a third term $(p-s)$ in their decomposition

$$\mathrm{GFT}(p,q,s,b) \;=\; \mathbb{I}\{s \le p\} \cdot \mathbb{I}\{q \le b\} \cdot [(b-q) + (q-p) + (p-s)]$$

Recall that $\mathrm{GFT}(p,q) = \mathbb{E}[\mathrm{GFT}(p,q,s,b)]$. Based on their idea, the following lemma provides an unbiased estimator of the gain from trade.

Lemma 6.1. *Let (S,B) denote the two valuations, drawn from an unknown distribution on $[0,1]^2$. Let $U \sim \mathrm{Unif}[0,1]$ and $V \sim \mathrm{Unif}[0,1]$ independent of (S,B).*

Then, the expected value of the gain from trade can be decomposed as

$$\begin{aligned}\mathrm{GFT}(p,q) \;=\; \mathbb{E}\,[\,&\mathbb{I}\{q \le U \le B\} \cdot \mathbb{I}\{S \le p\} \\ &+ \mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le V \le p\} \\ &+ \mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le p\} \cdot (q-p)\,]\end{aligned}$$

Thus, the following quantity is an unbiased estimator of $\mathrm{GFT}(p,q)$

$$\begin{aligned}\mathrm{G}(p,q,s,b) := \mathbb{E}\,[\,&\mathbb{I}\{q \le U \le B\} \cdot \mathbb{I}\{S \le p\} \\ &+ \mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le V \le p\} \\ &+ \mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le p\} \cdot (q-p) \mid S = s, B = b\,]\end{aligned}$$

Proof. Right term. By independence and the uniformity of U, given the realizations (s,b)

$$\begin{aligned}\mathbb{E}[\,\mathbb{I}\{q \le U \le B\} \cdot \mathbb{I}\{S \le p\} \mid S = s, B = b\,] &= \mathbb{I}\{s \le p\} \cdot \mathbb{P}(q \le U \le b) \\ &= \mathbb{I}\{s \le p\} \cdot \mathbb{I}\{q \le b\} \cdot (b-q)\end{aligned}$$

Left term. Similarly, by independence and the uniformity of V

$$\begin{aligned}\mathbb{E}[\,\mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le V \le p\} \mid S = s, B = b\,] &= \mathbb{I}\{q \le b\} \cdot \mathbb{P}(s \le V \le p) \\ &= \mathbb{I}\{q \le b\} \cdot \mathbb{I}\{s \le p\} \cdot (p-s)\end{aligned}$$

Profit term. The third component is deterministic given (s,b), hence

$$\mathbb{E}[\,\mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le p\} \cdot (q-p) \mid S = s, B = b\,] = \mathbb{I}\{q \le b\} \cdot \mathbb{I}\{s \le p\} \cdot (q-p)$$

Summing up. Adding the three conditional contributions yields

$$\begin{aligned}\mathrm{G}(p,q,s,b) &= \mathbb{I}\{q \le b\} \cdot \mathbb{I}\{s \le p\} \cdot [(b-q) + (p-s) + (q-p)] \\ &= \mathbb{I}\{q \le b\} \cdot \mathbb{I}\{s \le p\} \cdot (b-s)\end{aligned}$$

Finally, by the law of total expectation

$$\mathrm{GFT}(p,q) = \mathbb{E}[\, \mathbb{I}\{q \le B\} \cdot \mathbb{I}\{S \le p\} \cdot (B - S)\,] = \mathbb{E}[\, \mathrm{G}(p,q,S,B)\,]$$

□

The estimator is implemented through a short pure-exploration stage, Phase II, as probe variables $U, V \sim \mathrm{Unif}[0,1]$ are drawn to estimate the right and left geometric components of the decomposition.

Since each component requires a different probe, they are estimated in separate batches: for each seller price p, run m rounds posting p to the seller while drawing a buyer probe $U \sim \mathrm{Unif}[0,1]$; for each buyer price q, run m rounds posting q to the buyer while drawing a seller probe $V \sim \mathrm{Unif}[0,1]$. Formally

$$\widehat{R}(p,q) \;=\; \frac{1}{m}\sum_{r=1}^{m} \mathbb{I}\{s_r \le p\} \cdot \mathbb{I}\{q \le U_r \le b_r\}$$

$$\widehat{L}(p,q) \;=\; \frac{1}{m}\sum_{r=1}^{m} \mathbb{I}\{q \le b_r\} \cdot \mathbb{I}\{s_r \le V_r \le p\}$$

over two independent batches of m rounds each.

The reader should notice that if probes were drawn as in [9] as $U \sim \mathrm{Unif}[q,1]$ and $V \sim \mathrm{Unif}[0,p]$, then samples would not be transferable, implying a cost of $\Theta(K^2)$ operations. In contrast, sharing probes across pairs (p,q) with the same p and q reduces the exploration cost down to $2 \cdot m \cdot K = \Theta(Km)$.

6.2. Phase III: Regret over the Grid

Having constructed an estimator of the gain from trade for realistic feedback, this chapter turns to the analysis of Phase III, where the learner optimizes over the discretized grid $\mathcal{G}_K$.

For the profit, the bandit feedback estimator can be re-used

$$\overline{\mathrm{Pro}_t}(p,q) := \frac{1}{N_t(p,q)} \sum_{\tau=1}^{t} \mathbb{I}\{(p_\tau, q_\tau) = (p,q)\} \cdot \mathrm{Pro}(p,q,s_\tau,b_\tau)$$

and the corresponding estimator of gain from trade is defined as

$$\overline{\mathrm{GFT}}_t(p,q) := \widehat{R}(p,q) + \widehat{L}(p,q) + \overline{\mathrm{Pro}_t}(p,q)$$

At each round of Phase III, the algorithm $\mathcal{A}_G$ solves a linear program that replaces the unknown quantities with the following optimistic estimates

$$\begin{aligned} \max_{\gamma_t \in \Delta(\mathcal{G}_K)} \quad & \sum_{(p,q)\in\mathcal{G}_K} \Big[\, \gamma_t(p,q)\, (\widehat{R}(p,q) + \widehat{L}(p,q) + \overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q)) \,\Big] \\ \text{s.t.} \quad & \sum_{(p,q)\in\mathcal{G}_K} [\, \gamma_t(p,q)\, (\overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q))\,] \ge -\,\eta_K \end{aligned}$$

where $\phi_t : \mathbb{N} \to \mathbb{R}$ and $\eta_K \ge 0$.

As in the previous sections, the following lemma provides a uniform bound on the estimation error with respect to OPT_K, together with a margin accounting for the maximum budget required to execute the linear-program solution. The analysis extends the previous reasoning by integrating it with the new estimator of GFT.

Lemma 6.2. *With probability at least* $1-\delta$

$$\begin{cases} \displaystyle\sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} [\, \gamma_t(p,q)\, \mathrm{GFT}(p,q)\,] \;\ge\; T\,\mathrm{OPT}_K \;-\; 4K\sqrt{T}\,\log(\tfrac{6T}{\delta}) - \;4T\varepsilon \\ \displaystyle\sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} [\, \gamma_t(p,q)\, \mathrm{Pro}(p,q)\,] \;\ge\; -\,T\,\eta_K \;-\; 4K\,\sqrt{T}\,\log(\tfrac{6T}{\delta}) \end{cases}$$

where $\phi_t(p,q) = \sqrt{\frac{2}{N_t(p,q)}\,\log(\frac{6T}{\delta})}$ *and* $\varepsilon = \sqrt{\frac{1}{2m}\log(\frac{12K^2}{\delta})}$.

Proof. Step 1: Calibration of the clean event. As in the previous section, it is sufficient to define only t events. Since $\mathrm{Pro}(p,q,s_t,b_t) \in [-1,1]$, Hoeffding's inequality and a union bound over all rounds t yields that the bandit events

$$E_t^{\mathrm{Pro}} \;:=\; \big|\, \overline{\mathrm{Pro}_t}(p_t,q_t) - \mathrm{Pro}(p_t,q_t) \,\big| \;\le\; \phi_t(p_t,q_t)$$

hold simultaneously with probability at least $1-\frac{\delta}{3}$ by using the following value for $\phi_t(p,q)$

$$\exp\!\left\{-\frac{N_t(p,q)\;\phi_t(p,q)^2}{2}\right\} \;\le\; \frac{\delta}{3}\cdot\frac{1}{2T} \quad\Longleftrightarrow\quad \phi_t(p,q) \;=\; \sqrt{\frac{2}{N_t(p,q)}\;\log(\frac{6T}{\delta})}$$

Concerning the left and right components of the estimator, Hoeffding's inequality and a union bound over all the couples $(p,q) \in \mathcal{G}_K$ for both of them yields that the following events

$$E_{(p,q)}^{R} \;:=\; \Big|\, \widehat{R}(p,q) - R(p,q) \,\Big| \;\le\; \varepsilon$$

$$E_{(p,q)}^{L} \;:=\; \Big|\, \widehat{L}(p,q) - L(p,q) \Big| \;\le\; \varepsilon$$

hold simultaneously with probability at least $1-\frac{\delta}{3}$ by using the following value for ε. As before, it follows that

$$\exp\!\big\{-2m\cdot\varepsilon^2\big\} \;\le\; \frac{\delta}{3}\cdot\frac{1}{4K^2} \quad\Longleftrightarrow\quad \varepsilon \;=\; \sqrt{\frac{1}{2m}\;\log(\frac{12K^2}{\delta})}$$

Step 2: Computation of the Gain From Trade of the program solution. As in the previous section, on the clean event, feasibility of $\gamma_K^\star$ for the optimistic linear program holds since

$$\sum_{(p,q)} [\; \gamma_K^\star(p,q)(\overline{\mathrm{Pro}_t}(p,q) + \phi_t(p,q)) \;] \;\ge\; \sum_{(p,q)} [\; \gamma_K^\star(p,q)\;\mathrm{Pro}(p,q) \;] \;\ge\; -\,\eta_K$$

Using the symmetry of the bounds and the optimality of γ_t for the linear program

$$\begin{aligned}
&\sum_{(p,q)} [\; \gamma_t(p,q)\;\mathrm{GFT}(p,q) \;] \;\ge\\
&\ge\; \sum_{(p,q)} [\; \gamma_t(p,q)\cdot(\widehat{R}(p,q)-\varepsilon) \;] + \sum_{(p,q)} [\; \gamma_t(p,q)\cdot(\widehat{L}(p,q)\;-\varepsilon) \;] \;+\; \sum_{(p,q)} [\; \gamma_t(p,q)\cdot(\overline{\mathrm{Pro}_t}(p,q) - \phi_t(p,q)) \;]\\
&=\; \sum_{(p,q)} [\; \gamma_t(p,q)\cdot(\widehat{R}(p,q)+\varepsilon+\widehat{L}(p,q)+\varepsilon+\overline{\mathrm{Pro}_t}(p,q)\;+\phi_t(p,q)) \;] \;-\; 2\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] \;-\; 4\varepsilon\\
&\ge\; \sum_{(p,q)} [\; \gamma_K^\star(p,q)\cdot(\widehat{R}(p,q)+\widehat{L}(p,q)+2\varepsilon+\overline{\mathrm{Pro}_t}(p,q)\;+\phi_t(p,q)) \;] \;-\; 2\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] \;-\; 4\varepsilon\\
&\ge\; \sum_{(p,q)} [\; \gamma_K^\star(p,q)\cdot(R(p,q)+L(p,q)+\mathrm{Pro}(p,q)) \;] \;-\; 2\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] -\; 4\varepsilon\\
&=\; \mathrm{OPT}_K \;-\; 2\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] -\; 4\varepsilon
\end{aligned}$$

Summing over $t = 1,\ldots,T$

$$\sum_{t=1}^{T}\sum_{(p,q)} [\; \gamma_t(p,q)\;\mathrm{GFT}(p,q) \;] \;\ge\; T\;\mathrm{OPT}_K \;-\; 2\sum_{t=1}^{T}\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] \;-\; 4T\varepsilon$$

To upper bound the middle term, index the arms by $i = 1,\ldots,|\mathcal{G}_K|$ and apply the Azuma-Hoeffding bound as in the previous section. With probability at least $1-\frac{\delta}{3}$

$$\sum_{t=1}^{T}\sum_{(p,q)} [\; \gamma_t(p,q)\;\phi_t(p,q) \;] = \sqrt{2\log(\tfrac{6T}{\delta})}\;\sum_{t=1}^{T}\sum_{i=1}^{|\mathcal{G}_K|}\frac{\gamma_t(i)}{\sqrt{N_t(i)}} \;\le\; 4K\;\sqrt{T}\;\log(\tfrac{6T}{\delta})$$

Now, let $\mathcal{E}_\delta := E^{\mathrm{Pro}} \cap E^R \cap E^L \cap E^{\mathrm{Azuma}}$. Therefore, on $\mathcal{E}_\delta$

$$\sum_{t=1}^{T}\sum_{(p,q)} [\; \gamma_t(p,q)\;\mathrm{GFT}(p,q) \;] \;\ge\; T\;\mathrm{OPT}_K \;-\; 4K\sqrt{T}\;\log(\tfrac{6T}{\delta}) -\; 4T\varepsilon$$

Step 3: The program solution is GBB. As before, by symmetry of Hoeffding, $\text{Pro}(p,q) \geq \overline{\text{Pro}_t}(p,q) - \phi_t(p,q)$ for every (p,q) and t. Hence, using feasibility of γ_t

$$\sum_{(p,q)} [\, \gamma_t(p,q)\, \text{Pro}(p,q)\,] \;\geq\; -\,\eta_K \;-\; 2 \sum_{(p,q)} [\, \gamma_t(p,q)\, \phi_t(p,q)\,]$$

Summing over $t = 1, \ldots, T$, similarly to what was done previously

$$\sum_{t=1}^{T} \sum_{(p,q)} [\, \gamma_t(p,q)\, \text{Pro}(p,q)\,] \;\geq\; -\,T\,\eta_K \;-\; 4K\,\sqrt{T}\,\log(\tfrac{6T}{\delta})$$

□

The following corollary proves that the regret of the algorithm with respect to γ_K^* (benchmark on the grid) during the Phase III — assuming sufficient budget to run — is sublinear in the clean event.

Corollary 6.1. *For* $\delta = \frac{1}{T}$

$$T\,\text{OPT}_K - \mathbb{E}\big[\, \sum_{t=1}^{T} \sum_{(p,q)} [\, \gamma_t(p,q)\, \text{GFT}(p,q)\,]\, \big] \;=\; O(\, K\sqrt{T}\,\log(T) \;+\; T\sqrt{\tfrac{1}{m}\log(K^2T)}\,)$$

Proof.

$$\begin{aligned}
& T\,\text{OPT}_K - \mathbb{E}\big[\, \sum_{t=1}^{T} \sum_{(p,q)\in\mathcal{G}_K} [\, \gamma_t(p,q)\, \text{GFT}(p,q)\,]\, \big] \\
& \leq\; T\,\text{OPT}_K \;-\; [\, (1-\delta)\, (\, T\,\text{OPT}_K - 4K\,\sqrt{T}\,\log(\tfrac{6T}{\delta}) \;-\; 4T\varepsilon\,)\,] \\
& =\; \delta\,T\,\text{OPT}_K \;+\; 4K\,(1-\delta)\,\sqrt{T}\,\log(\tfrac{6T}{\delta}) \;+\; 4T\varepsilon\,(1-\delta) \\
& =\; \text{OPT}_K \;+\; 8K\,(1-\tfrac{1}{T})\,\sqrt{T}\,\log(\sqrt{6}\,T) \;+\; 4T\,(1-\tfrac{1}{T})\sqrt{\tfrac{1}{2m}\,\log(12K^2T)} \\
& \leq\; 1 \;+\; 8K\,\sqrt{T}\log(\sqrt{6}\,T) \;+\; 4T\sqrt{\tfrac{1}{2m}\,\log(12K^2T)} \\
& =\; O(\, K\sqrt{T}\,\log(T) \;+\; T\sqrt{\tfrac{1}{m}\log(K^2T)}\,)
\end{aligned}$$

□

The reader should notice that m is a tunable parameter that can be made arbitrarily large, at the expense of allocating more rounds in Phase II to estimate the left and right components of the gain from trade estimator. The resulting trade-off is analyzed in the next chapter.

6.3. Phases II and III: Regret over the Grid

The overall regret over the grid in Phases II and III is bounded by the following function of time

$$2Km \;+\; 8K\sqrt{T}\,\log(\sqrt{6}\,T) \;+\; 4T\sqrt{\tfrac{1}{2m}\,\log(12K^2T)}$$

where the first term corresponds to the regret due to Phase II, while the two other terms refer to the Phase III, as shown in the previous chapter.
The following lemma shows that there exist a choice for m and K such that the regret over the grid during Phases II and III is sublinear.

Lemma 6.3. *The pseudo-regret of Phases II and III with respect to a benchmark* γ_K^* *(on the grid) is at most*

$$\mathcal{R}_T(\mathcal{A}^{II}) \;=\; \widetilde{O}(\, T^{3/4}\,)$$

obtained by choosing $K = T^{1/4}$ *and* $m = \sqrt{T}$.

Proof. Starting from the bound, up to logarithmic factors, the regret can be expressed in the form

$$\mathcal{R}_T(\mathcal{A}^{\mathrm{II}}) \;=\; \widetilde{O}(\; Km \;+\; K\sqrt{T} \;+\; \frac{T}{\sqrt{m}}\;)$$

Let K and m be functions of the horizon T

$$K \;=\; T^{\alpha} \qquad \text{and} \qquad m \;=\; T^{\beta}$$

for some exponents $\alpha, \beta > 0$. Substituting into the three terms yields the following asymptotic behaviours

$$Km \asymp T^{\alpha+\beta} \qquad K\sqrt{T} \asymp T^{\alpha+1/2} \qquad \frac{T}{\sqrt{m}} \asymp T^{1-\beta/2}$$

Since the largest exponent dominates the asymptotic behaviour, minimizing the regret exponent is equivalent to solving the min-max problem

$$\min_{\alpha,\,\beta>0} \; \max\{\alpha+\beta,\; \alpha+\tfrac{1}{2},\; 1-\tfrac{\beta}{2}\}$$

This can be written in epigraph form as a linear program with affine objective and constraints. In such a formulation, any optimal solution must occur at a vertex of the feasible polyhedron, where at least two of the constraints are active.
Consequently, the values of α and β for which two of the exponents coincide are obtained by solving

$$\alpha+\beta = \alpha+\tfrac{1}{2} \qquad \text{and} \qquad \alpha+\beta = 1-\tfrac{\beta}{2}$$

Solving the system yields $\alpha = \frac{1}{4}$ and $\beta = \frac{1}{2}$, thus plugging them into K and m

$$K = T^{1/4} \qquad \text{and} \qquad m = \sqrt{T}$$

Substituting these values back into the regret expression gives

$$\begin{aligned}\mathcal{R}_T(\mathcal{A}^{\mathrm{II}}) \;&=\; \widetilde{O}(\; T^{1/4}\,\sqrt{T} \;+\; T^{1/4}\sqrt{T} \;+\; \frac{T}{\sqrt{\sqrt{T}}}\;)\\ &=\; \widetilde{O}(\; T^{3/4} \;+\; T^{3/4} \;+\; T^{3/4}\;)\\ &=\; \widetilde{O}(\; T^{3/4}\;)\end{aligned}$$

Proving that by choosing $K = T^{1/4}$ and $m = \sqrt{T}$, it is possible to obtain the best possible upper bound of $O(T^{3/4})$.

□

More into detail, by replacing the optimal values of K and m in the expression, the regret over the grid during Phases II and III becomes

$$\mathcal{R}_T(\mathcal{A}^{II}) \;=\; 2\,T^{3/4} \;+\; 8\,T^{3/4}\log(\sqrt{6}\,T) \;+\; 4\,T^{3/4}\sqrt{\log(12\sqrt{T^3})}$$

6.4. Phase I: Profit Maximization on the Multiplicative Grid

Phase I is used to accumulate enough budget BB_t for the subsequent phases and to enable trading over the entire horizon T through phases II and III. In this phase, the learner employs bandit UCB as the profit optimizer $\mathcal{A}_P$, maximizing profit over the multiplicative grid $\mathcal{F}_K$.
Recall that the cumulative budget is initialized as $BB_0 = 0$ and evolves according to

$$BB_t \;=\; BB_{t-1} + \mathrm{Pro}(p_t, q_t, s_t, b_t)$$

and that Phase I stops at the (random) hitting time

$$\tau \;:=\; \min\Big\{t \in \{1,\dots,T\} : \sum_{u=1}^{t} \mathrm{Pro}(p_u, q_u, s_u, b_u) \;\geq\; \beta\Big\}$$

for a threshold $\beta > 0$ to be specified.
The following lemma establishes the bound on the regret incurred by the algorithm during Phase I while accumulating budget to be used during phases II and III.

Lemma 6.4. *The regret due to Phase I with respect to the benchmark γ^* is*

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;= \widetilde{O}(T^{3/4})$$

when requiring to collect a budget $\beta = \widetilde{O}(T^{3/4})$

Proof. Recall that

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;=\; \tau \cdot \mathrm{OPT} \;-\; \mathbb{E}[\, \sum_{t=1}^{\tau} \mathrm{GFT}(p_t, q_t)\,]$$

Since Phase I only plays points in $\mathcal{F}_K$ — which lie weakly above the diagonal — the per-round gain from trade is non-negative, implying

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;\leq\; \tau \cdot \mathrm{OPT}$$

By Theorem 6.3 of [4], the gain from trade achieved by the two-dimensional benchmark — that is, the best feasible distribution — is at most twice that of the one-dimensional benchmark - that is, the optimal fixed price in hindsight. Applied on the first τ rounds

$$\tau \cdot \mathrm{OPT} \;\leq\; 2 \cdot \max_{p\in[0,1]} \sum_{t=1}^{\tau} \mathrm{GFT}(p)$$

Also, by Proposition 3.3 of [4], the gain from trade accumulated by the optimal fixed price in hindsight is

$$\max_{p\in[0,1]} \sum_{t=1}^{\tau} \mathrm{GFT}(p) \;\leq\; 12\log(\tau) \;\cdot \max_{(p,q)\in\mathcal{F}_K} \sum_{t=1}^{\tau} \mathrm{Pro}(p,q) \;+\; \frac{5\tau}{K}$$

Moreover, using UCB as profit optimizer $\mathcal{A}_P$ on the arm set $\mathcal{F}_K$ over τ rounds guarantees that, with probability $1-\delta$

$$\max_{(p,q)\in\mathcal{F}_K} \sum_{t=1}^{\tau} \mathrm{Pro}(p,q) \;\leq\; \sum_{t=1}^{\tau} \mathrm{Pro}(p_t,q_t) \;+\; 4\,\sqrt{\tau\,K\,\log(K)\,\log(\tfrac{2\tau K\log(K)}{\delta})}$$

Let $\mathcal{E}_\delta$ be the event that the above inequality holds. Then, since the pseudo-regret is defined in expectation

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;=\; \mathbb{E}[\,\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \mid \mathcal{E}_\delta\,]\cdot(1-\delta) \;+\; \mathbb{E}[\,\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \mid \mathcal{E}_\delta^c\,]\cdot\delta$$

On $\mathcal{E}_\delta$, the following deterministic chain of inequalities holds

$$\begin{aligned}
\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;&\leq\; \tau \;\cdot\; \mathrm{OPT} \\
&\leq\; 2 \;\cdot \max_{p\in[0,1]} \sum_{t=1}^{\tau} \mathrm{GFT}(p) \\
&\leq\; 24\log(\tau) \;\cdot \max_{(p,q)\in\mathcal{F}_K} \sum_{t=1}^{\tau} \mathrm{Pro}(p,q) \;+\; \frac{10\tau}{K} \\
&\leq\; 24\log(\tau) \;\cdot\; (\sum_{t=1}^{\tau} \mathrm{Pro}(p_t,q_t) + 4\sqrt{\tau\,K\,\log(K)\,\log(\tfrac{2\tau K\log(K)}{\delta})}\,) \;+\; \frac{10\tau}{K} \\
&\leq\; 24\log(T) \;\cdot\; [\,(\beta+1) + 4\,\sqrt{TK\log(K)\log(\tfrac{2TK\log(K)}{\delta})}\,] \;+\; \frac{10T}{K}
\end{aligned}$$

On the failure event $\mathcal{E}_\delta^c$, the regret of $\mathcal{A}_p$ in terms of profit is at most T; therefore, choosing $\delta = 1/T$ makes the additive δT term negligible. Setting $K = T^{1/4}$ and choosing $\beta = \widetilde{O}(T^{3/4})$ gives

$$\begin{aligned}
\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;&=\; O(\,\log(T) \;\cdot\; [\,T^{3/4} \;+\; T^{1/2}\,T^{1/4}\,\sqrt{\log(T^{5/4})}\,] \;+\; T^{3/4}\,) \\
&=\; \widetilde{O}(T^{3/4})
\end{aligned}$$

which completes the proof. □

The bandit UCB profit maximization on $\mathcal{F}_K$ guarantees that the budget accumulated during Phase I is sufficient to sustain both subsequent phases. In particular, the total budget threshold can be decomposed as $\beta \geq 2mK + \beta_{\mathrm{III}}$, where the first term upper bounds the worst-case budget consumption incurred during Phase II (Pure Exploration), and β_{III} denotes the budget required to safely execute Phase III (Optimization), as characterized in the previous chapter. Since both terms scale as $\widetilde{O}(T^{3/4})$ under the chosen parameters, Phase I indeed needs to accumulate a budget of order $\widetilde{O}(T^{3/4})$, thus implying a contribution to the overall regret of the same order.
The next chapter combines this bound with the regret incurred in Phases II and III and, ultimately, the discretization error induced by the additive grid, yielding the final $\widetilde{O}(T^{3/4})$ pseudo-regret guarantee against the distributional benchmark.

6.5. Discretization Error and Total Regret Upper Bound

In this chapter, the regret guarantees derived for Phases I, II and III are consolidated by adding the discretization term due to restricting prices to a homogeneous grid. To do so, the joint law of (S, B) is assumed to have bounded density, a standard regularity condition that excludes atoms and implies that $\mathrm{GFT}(p, q)$ is 4σ-Lipschitz on $[0, 1]^2$, thereby enabling a controlled approximation error on the grid. The necessity of such regularity for learnability is underscored by Theorems 4.3-4.4 in [9], which shows linear regret in its absence.

Lemma 6.5. *Assume that (S, B) admit a density $f_{S,B}$ bounded by $\|f_{S,B}\|_\infty \leq \sigma$. Then $\mathrm{GFT}(p, q)$ is 4σ-Lipschitz on $[0, 1]^2$ and $\mathrm{Pro}(p, q)$ is $(4\sigma + 2)$-Lipschitz on $[0, 1]^2$. Consequently, using $\eta_K = \frac{8\,(\sigma+1)}{K-1}$, then*

$$\mathrm{OPT} - \mathrm{OPT}_K \;\leq\; \frac{4\,\sigma}{K-1}$$

Proof. Fix $(p, q) \in [0, 1]^2$ and let $(p_K, q_K) \in \mathcal{G}_K$ be a nearest neighbour in ℓ_∞, rounding each coordinate to the closest grid value. Then

$$|\, p - p_K \,| \;\leq\; \frac{1}{2\,(K-1)}$$

$$|\, q - q_K \,| \;\leq\; \frac{1}{2\,(K-1)}$$

hence $|p - p_K| + |q - q_K| \leq \frac{1}{K-1}$.

Step 1: Gain from trade is Lipschitz. As in Lemma 4.1 of [9], bounded joint density implies $|\mathrm{GFT}(p, q) - \mathrm{GFT}(p', q')| \;\leq\; 4\,\sigma \cdot (|p - p'| + |q - q'|)$. Therefore

$$\mathrm{GFT}(p_K, q_K) \;\geq\; \mathrm{GFT}(p, q) - \frac{4\,\sigma}{K-1}$$

Step 2: Profit is Lipschitz. Write $\mathrm{Pro}(p, q) = (q - p) \cdot \Pr[S \leq p,\, B \geq q]$. The map $(p, q) \mapsto \Pr[S \leq p,\, B \geq q]$ is 2σ-Lipschitz in ℓ_1. Multiplying by $(q - p) \in [-1, 1]$ and accounting for the additional variation of $(q - p)$ itself yields a bound of the form $|\mathrm{Pro}(p, q) - \mathrm{Pro}(p', q')| \leq (4\sigma + 2)\,(|p - p'| + |q - q'|)$. Thus

$$\mathrm{Pro}(p_K, q_K) \;\geq\; \mathrm{Pro}(p, q) - \frac{(4\sigma + 2)}{K-1} \;\geq\; \mathrm{Pro}(p, q) - \eta_K$$

Step 3: From point-wise rounding to distributions. Let $\gamma^\star$ be an optimal feasible distribution for OPT. Define $\gamma_K^\star$ as the push-forward of $\gamma^\star$ under coordinate-wise rounding to $\mathcal{G}_K$. By linearity of expectations and the point-wise bounds above

$$\mathbb{E}_{\gamma_K^\star}[\mathrm{GFT}] \;\geq\; \mathbb{E}_{\gamma^\star}[\mathrm{GFT}] - \frac{4\,\sigma}{K-1} \;=\; \mathrm{OPT} - \frac{4\,\sigma}{K-1}$$

$$\mathbb{E}_{\gamma_K^\star}[\mathrm{Pro}] \;\geq\; \mathbb{E}_{\gamma^\star}[\mathrm{Pro}] - \eta_K \;\geq\; -\,\eta_K$$

Thus, $\gamma_K^\star$ is feasible for the slacked grid benchmark defining OPT_K. Therefore

$$\mathrm{OPT}_K \;\geq\; \mathbb{E}_{\gamma_K^\star}[\mathrm{GFT}] \;\geq\; \mathrm{OPT} - \frac{4\,\sigma}{K-1}$$

Which implies $\mathrm{OPT} - \mathrm{OPT}_K \leq \frac{4\,\sigma}{K-1}$. □

Recall that the vanishing slack defined previously is set to $\eta_K = \frac{8\,(\sigma+1)}{K-1}$. The next theorem provides a final result for the regret of the proposed algorithm.

Theorem 6.1. *Run the three-phase algorithm in the stochastic i.i.d. model with realistic feedback, using*

$$K \;=\; T^{1/4}$$

$$m \;=\; \sqrt{T}$$

$$\beta \;=\; 2mK \;+\; 8\,K\sqrt{T}\,\log(\sqrt{6}\,T) \;+\; T\cdot\eta_K \;+\; 1$$

Then, under bounded density assumption, the pseudo-regret against the distributional benchmark satisfies

$$\mathcal{R}_T \;=\; \widetilde{O}(T^{3/4})$$

Proof. Decompose the regret as

$$\mathcal{R}_\tau(\mathcal{A}^I) \;+\; \mathcal{R}_\tau(\mathcal{A}^{II}) \;+\; T\cdot(\mathrm{OPT}-\mathrm{OPT}_K)$$

By the Phase I lemma proved above, with this choice of β

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{I}}) \;=\; \widetilde{O}(T^{3/4})$$

From the Phase II (realistic feedback) bound, plugging $K = T^{1/4}$ and $m = \sqrt{T}$ also gives

$$\mathcal{R}_\tau(\mathcal{A}^{\mathrm{II}}) \;=\; \widetilde{O}(T^{3/4})$$

Finally, assuming bounded density gives

$$T\cdot(\mathrm{OPT}-\mathrm{OPT}_K) \;\le\; T\cdot\frac{4\,\sigma}{K-1} \;=\; O(\sigma\cdot T^{3/4})$$

Summing the three terms up gives

$$\mathcal{R}_T = \widetilde{O}(T^{3/4})$$

□

The desired result of this work was thereby obtained.

6.6. Recap

To summarize, this work proposes an algorithm based on a separation between budget accumulation and optimization, which mainly relies on a high-probability analysis of the problem on a discretized action space, that is subsequently lifted back to the continuous setting.

The profit accumulated in Phase I exceeds the budget required to execute both Phase II (Pure Exploration) and Phase III (Optimization) with very high probability. Consequently, with very high probability, Phase III can be run without exhausting the budget, and the regret with respect to the benchmark on the grid remains bounded as derived above.

On the other hand, in the complementary event, the accumulated budget may be insufficient and the algorithm either stops trading or reverts to a trivial safe policy, incurring a worst-case additional regret of at most T. However, by choosing $\delta = 1/T$ for the probability of the failure event, this worst-case contribution becomes negligible in expectation, while the impact of δ on the regret bounds for Phase III remains only logarithmic.

Finally, the guarantees with respect to the continuous benchmark follow by adding the discretization error to the regret on the grid, which is controlled under the bounded density assumption.

7. Conclusion

This work studies repeated bilateral trade under global budget balance in the stochastic i.i.d. setting and establishes that competing against the distributional benchmark — the best distribution over price pairs that is globally budget balanced in expectation — is achievable with pseudo-regret $\widetilde{O}(T^{3/4})$ under realistic feedback, showing that in stochastic environments the distributional benchmark is in fact attainable. Importantly, this shows that — in the stochastic setting — competing against the distributional benchmark under global budget balance is no harder, in terms of regret rates, than learning the best fixed feasible price under weak budget balance.

The algorithmic contribution is a three-phase design that separates profit accumulation, estimator construction, and gain from trade optimization. Phase I operates on a multiplicative grid $\mathcal{F}_K$ to gather budget; Phase II constructs transferable estimators of the geometric components of GFT from acceptance data; Phase III solves an optimistic linear program over an additive grid $\mathcal{G}_K$, using optimism both in reward and profit to enable feasibility while exploring high-GFT regions near the diagonal.

On the technical side, the analysis proceeds through full and bandit feedback, thereby isolating the optimization and estimation ingredients reused then in the realistic feedback. Under a bounded density assumption on (S, B), gain from trade is 4σ-Lipschitz, implying a discretization error of order $O(\sigma \cdot T^{3/4})$. Combining each bound to the regret of Phases I, II and III on the grid along with the discretization term yields the overall $\widetilde{O}(T^{3/4})$ pseudo-regret, while maintaining global budget balance in expectation.

References

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

[2] Moshe Babaioff, Kira Goldner, and Yannai A Gonczarowski. Bulow-klemperer-style results for welfare maximization in two-sided markets. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2452–2471. SIAM, 2020.

[3] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), March 2018.

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

[5] Liad Blumrosen and Yehonatan Mizrahi. Approximating gains-from-trade in bilateral trading. In *Web and Internet Economics: 12th International Conference, WINE 2016, Montreal, Canada, December 11-14, 2016, Proceedings 12*, pages 400–413. Springer, 2016.

[6] Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

[7] Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022.

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

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

[10] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, USA, 2006.

[11] Houshuang Chen, Yaonan Jin, Pinyan Lu, and Chihao Zhang. Tight regret bounds for fixed-price bilateral trade. *arXiv preprint arXiv:2504.04349*, 2025.

[12] Riccardo Colini-Baldeschi, Paul Goldberg, Bart de Keijzer, Stefano Leonardi, and Stefano Turchetta. Fixed price approximability of the optimal gain from trade. In *Web and Internet Economics: 13th International Conference, WINE 2017, Bangalore, India, Proceedings 13*, pages 146–160. Springer, 2017.

[13] Riccardo Colini-Baldeschi, Paul W Goldberg, Bart de Keijzer, Stefano Leonardi, Tim Roughgarden, and Stefano Turchetta. Approximately efficient two-sided combinatorial auctions. *ACM Transactions on Economics and Computation (TEAC)*, 8(1):1–29, 2020.

[14] Riccardo Colini-Baldeschi, Bart de Keijzer, Stefano Leonardi, and Stefano Turchetta. Approximately efficient double auctions with strong budget balance. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 1424–1443. SIAM, 2016.

[15] Yuan Deng, Jieming Mao, Balasubramanian Sivan, and Kangning Wang. Approximately efficient bilateral trade. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 718–721, 2022.

[16] Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), November 2022.

[17] Zi Yang Kang and Jan Vondrák. Fixed-price approximations to optimal efficiency in bilateral trade. *Available at SSRN 3460336*, 2019.

[18] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

[19] Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. Better regret rates in bilateral trade via sublinear budget violation. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2026.

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

[21] Aleksandrs Slivkins. Multi-armed bandits on implicit metric spaces. *Advances in Neural Information Processing Systems*, 24, 2011.

Acknowledgements

I thank my advisor *Prof. M. Castiglioni* for the opportunity to contribute with his lab to ongoing research in our field, and I particularly thank *Anna* for her generous availability and continuous assistance throughout the development of this work. I thank *mamma* and *papà* for their support (and their patience) throughout these years, my brother *Mirco* for being the best sibling I could ask for, and I thank all of the rest of my *family* for their unconditional love. I thank *all of my friends* for brightening these years, by sharing with me both the hard and rewarding moments of the exam periods, and I particularly thank *Matteo* for being my partner in crime throughout this journey; last but not least, I thank *Angelica*, for always being by my side. I thank *Portugal* for teaching me a new language, and I thank *Brazil* for *actually* teaching me that language; in general, I thank both, for welcoming me and for all the beautiful people I had the chance to meet. I would like to thank everyone who allowed me to experience *meaningful emotions*, and I particularly thank *Lucia* for her solid contribution to my strong academic performance. I thank *Building 3 and Building 11* for hosting countless hours of study, and *Politecnico di Milano* for providing them, even though finding a seat was not always *that* easy. Lastly, I would like to thank *myself* — I think I did a good job.