POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Computer Science and Engineering

Author: Mattia Piccinato

Advisor: Prof. M. Castiglioni

Co-advisors: A. Lunghi, A. Marchesi

Academic Year: 2024-2025

1. Introduction

Bilateral trade models the interaction between two rational agents: a seller and a buyer. Each agent has a private valuation for an indivisible good, denoted by $s \in [0, 1]$ for the seller and $b \in [0, 1]$ for the buyer.

Given that a trade occurs if and only if the mechanism posts a pair of prices (p, q) such that both agents accept — namely if $s \leq p$ and $q \leq b$ — the goal of the mechanism is to maximize the social welfare — or, equivalently, the associated gain from trade.

Despite the simplicity of this setting, the classical impossibility result of [3] shows that no mechanism can simultaneously satisfy efficiency, incentive compatibility, individual rationality, and strong budget balance under fairly general assumptions.

In the online learning framework introduced in [2], a mechanism interacts with a sequence of T seller-buyer pairs with valuations $(S_t, B_t)_{t=1}^T$ and posts prices sequentially. Here, performance is evaluated through regret with respect to an appropriate benchmark, thereby providing a principled way to address the problem without any prior knowledge of the valuation distributions.

In particular, the present work studies repeated bilateral trade in a stochastic i.i.d. environment under realistic feedback — thus only observing acceptance events — using the *global budget balance* notion introduced in [1].

This work proves that the problem of learning against the *globally budget-balanced distributional benchmark* — also introduced in [1] - is attainable in this setting. Moreover, an online learning algorithm is proposed, achieving pseudo-regret $\widetilde{O}(T^{3/4})$ under a bounded density assumption on valuations.

Key Results

In the stochastic i.i.d. setting under realistic feedback, learning against the globally budget-balanced distributional benchmark is possible with sublinear pseudo-regret. In particular, a three-phase online learning algorithm is proposed, achieving $\widetilde{O}(T^{3/4})$ pseudo-regret while maintaining global budget balance.

2. Problem Setting

Stochastic environment. At each round $t = 1, \dots, T$, a seller–buyer pair arrives with private valuations $(S_t, B_t) \in [0,1]^2$ drawn i.i.d. from an unknown joint distribution $\mathcal{D}$. No independence between S_t and B_t within a round is assumed; only independence across time is required.

Mechanism and objectives. At round t, the mechanism posts a seller price $p_t \in [0,1]$ and a buyer price $q_t \in [0,1]$. A trade occurs if and only if both agents accept:

$$\mathbb{1}\{S_t \le p_t\} \cdot \mathbb{1}\{q_t \le B_t\}$$

The two key per-round quantities are gain from trade and profit:

$$\begin{aligned} \mathrm{GFT}(p,q,s,b) &= (b-s)\,\mathbb{1}\{s \le p\}\,\mathbb{1}\{q \le b\} \\ \mathrm{Pro}(p,q,s,b) &= (q-p)\,\mathbb{1}\{s \le p\}\,\mathbb{1}\{q \le b\} \end{aligned}$$

Gain from trade is the relevant efficiency proxy in this setting — equivalent to social welfare for this problem — while profit is the accounting quantity used to enforce budget balance and to enable exploration.

Global budget balance. Given an instance of the problem, let the cumulative budget be initialized as $BB_0 = 0$ and updated by

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

A run is *globally budget balanced* if $BB_T \ge 0$. The algorithm considered in this work enforces $BB_t \ge 0$ throughout the execution, and therefore satisfies global budget balance.

Realistic feedback.pr The observed feedback consists of acceptance information only: after posting (p_t, q_t), the mechanism only observes the one-bit indicator $\mathbb{1}\{S_t \le p_t\} \cdot \mathbb{1}\{q_t \le B_t\}$. For readability, the feedback is always expressed as the product of two acceptance indicators, although only this product is observed.

Distributional benchmark. Under global budget balance, randomization over price pairs can strictly improve performance, because occasional profitable actions may subsidize frequent deficit-incurring actions with higher gain from trade. Accordingly, the benchmark is the best distribution over price pairs that is budget balanced *in expectation*. Let $\Delta([0,1]^2)$ denote the set of distributions on $[0,1]^2$ and define the expected quantities

$$\begin{aligned} \mathrm{GFT}(p,q) &:= \mathrm{E}_{(S,B)\sim\mathcal{D}}[\mathrm{GFT}(p,q,S,B)] \\ \mathrm{Pro}(p,q) &:= \mathrm{E}_{(S,B)\sim\mathcal{D}}[\mathrm{Pro}(p,q,S,B)] \end{aligned}$$

Then the benchmark value is

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

The distributional benchmark is strictly richer than the best fixed price. The learning algorithm must therefore approximate not only the best action, but the best feasible distribution over actions.

Pseudo-regret. The goal is to minimize pseudo-regret with respect to the benchmark:

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

where the expectation is over the algorithm randomization and the i.i.d. valuation sequence.

3. Methodology

The mechanism operates on a continuous action space $[0,1]^2$. Since no algorithm can optimize directly over a continuum, the price space must be discretized while controlling the induced approximation error. As shown in [2], without suitable regularity assumptions, discretization under realistic feedback leads to linear regret.

Regularity assumption. Throughout the analysis, the joint law of (S, B) is assumed to admit a bounded density: $\|f_{S,B}\|_\infty \le \sigma$. Under this assumption, the expected gain from trade is Lipschitz in the posted prices. This regularity property guarantees that a sufficiently fine grid approximates the continuous benchmark with vanishing error.

Three-phase algorithmic structure. The proposed algorithm separates budget management and gain from trade optimization through a three-phase procedure:

- Phase I (Budget Accumulation). The algorithm first accumulates profit in order to finance potentially deficit-incurring price pairs used later.

- Phase II (Pure Exploration). Under realistic feedback, additional exploration is required to construct estimators of the gain from trade.
- Phase III (Optimization). The algorithm then optimizes gain from trade subject to the global budget constraint.

Each phase operates on a different discretization of the price space, tailored to its specific objective.

Additive grid. To approximate the continuous action space, define a uniform grid

$$\mathcal{G}_K = \left\{0, \frac{1}{K-1}, \dots, 1\right\}^2, \qquad |\mathcal{G}_K| = K^2.$$

Phases II and III optimize over distributions supported on $\mathcal{G}_K$. A small feasibility slack $\eta_K \geq 0$ is introduced in the profit constraint to ensure that the discretized benchmark remains feasible after rounding.

Multiplicative grid. Phase I requires a discretization specifically designed to maximize profit efficiently. Starting from the one-dimensional additive grid underlying $\mathcal{G}_K$, for each base price p the algorithm considers price pairs obtained by shifting p by geometrically decreasing amounts of the form 2^{-a}, for integers a up to order $\log T$. This construction produces a sparse band of price pairs above the diagonal $q = p$, where the buyer-seller spread follows a dyadic scale. The resulting set contains only $\Theta(K \log T)$ pairs, substantially fewer than the K^2 elements of the additive grid, and guarantees non-negative per-round profit whenever trade occurs.
This structure enables efficient budget accumulation without incurring excessive learning complexity.

Warm-up feedback models.
For expositional clarity, the analysis develops progressively through simpler feedback models before addressing realistic feedback. In particular, Phase III is first studied in the full-feedback and bandit-feedback settings, which serve as intermediate steps, where the optimization component of the algorithm can be analyzed without the additional difficulties induced by acceptance-only observations.
These warm-up cases isolate the core regret and discretization arguments that are later combined with the estimator construction required under realistic feedback.

4. Decomposition of GFT

A central ingredient is an unbiased estimator of $\mathrm{GFT}(p, q)$ under realistic feedback, which reconstructs gain from trade using only acceptance information without observing the valuations (S, B), while leveraging probe sharing across many grid pairs for better efficiency.
The gain from trade admits the decomposition $\mathbb{1}\{s \leq p\}\,\mathbb{1}\{q \leq b\}\,[(b-q)+(q-p)+(p-s)]$, extending the decomposition lemma in [2].
The term $(q-p)$ is always observable. The remaining terms can be recovered in expectation through independent uniform probes. Let $U, V \sim \mathrm{Unif}[0,1]$ be independent of (S, B). Then $\mathrm{GFT}(p, q)$ becomes

$$\begin{aligned}\mathrm{GFT}(p,q) = \mathrm{E}\Big[\ & \mathbb{1}\{S \leq p\}\,\mathbb{1}\{q \leq U \leq B\} \\ & + \mathbb{1}\{q \leq B\}\,\mathbb{1}\{S \leq V \leq p\} \\ & + \mathbb{1}\{S \leq p\}\,\mathbb{1}\{q \leq B\}\,(q-p)\ \Big]\end{aligned}$$

Consequently, the quantity inside the expectation provides an unbiased estimator of $\mathrm{GFT}(p, q)$ that depends only on acceptance events and the sampled probes.
Each indicator is evaluable from acceptance events when the mechanism controls the probe variable (by posting U or V as a price to one side).

Probe sharing. Phase II uses two probe batches: (i) for each seller price p on the grid, m rounds are executed with seller price fixed at p and buyer probe $U \sim \mathrm{Unif}[0,1]$; (ii) for each buyer price q on the grid, m rounds are executed with buyer price fixed at q and seller probe $V \sim \mathrm{Unif}[0,1]$. This produces estimators of the first two components that can be reused for all pairs sharing the same p or q, reducing exploration to $\Theta(K)$ rounds rather than $\Theta(K^2)$.

5. Optimistic Linear Program

Phase III implements an optimism-driven selection rule over distributions on the discretized action space. For each $(p, q) \in \mathcal{G}_K$, the algorithm maintains an estimate $\widehat{g}_t(p, q)$ of $\mathrm{GFT}(p, q)$ and an estimate $\widehat{\pi}_t(p, q)$ of $\mathrm{Pro}(p, q)$, together with

a confidence bonus $\phi_t(p,q)$ shrinking with the number of samples collected for (p,q).
At each time t in Phase III, a distribution $\gamma_t \in \Delta(\mathcal{G}_K)$ is computed by solving the optimistic program

$$\max_{\gamma_t \in \Delta(\mathcal{G}_K)} \sum_{(p,q)\in\mathcal{G}_K} \gamma_t(p,q)\left(\widehat{g}_t(p,q)+\phi_t(p,q)\right)$$

$$\text{s.t.} \sum_{(p,q)\in\mathcal{G}_K} \gamma_t(p,q)\left(\widehat{\pi}_t(p,q)+\phi_t(p,q)\right) \geq -\eta_K$$

The confidence bonus $\phi_t(p,q)$ is calibrated for a confidence level $\delta \in (0,1)$ so that, with probability at least $1-\delta$, the estimates $(\widehat{g}_t(p,q), \widehat{\pi}_t(p,q))$ are simultaneously accurate for all t and all $(p,q) \in \mathcal{G}_K$. In particular, $\delta = \frac{1}{T}$ ensures that the high-probability regret bound holds except on an event of probability $O(1/T)$; since regret is trivially bounded by T, the contribution of the failure event to the expected regret is at most $T \cdot \delta = O(1)$, and is therefore absorbed into the asymptotical rate.
The confidence bonus appears in both the objective and the constraint. Optimism in the objective encourages exploration as in classic UCB, while optimism in the profit constraint allows the algorithm to sample pairs near the diagonal $p \approx q$ — typically providing the highest GFT — by leveraging the budget accumulated in the initial phase. This dual use of optimism ensures that the linear program remains feasible with very high probability, enabling a high-probability analysis that is subsequently translated into the final regret bound through appropriate parameter tuning.

6. Parameter Tuning

Parameters. Recall that the integer T denotes the time horizon and the integer K is the discretization resolution. The variable σ controls the Lipschitz regularity of GFT$(\cdot)$, while the integer m is the number of probe rounds executed per one-dimensional grid point to estimate the GFT during Phase II, while scalar β is the target budget accumulated in Phase I and subsequently used to subsidize deficit-incurring actions in the subsequent phases.

Tuning. The main question addressed is whether the benchmark is learnable under realistic feedback in a stochastic i.i.d. environment. This question is non-trivial because [1] shows that competing with the same benchmark is impossible with sublinear regret in adversarial environments. The stochastic i.i.d. assumption enables concentration and uniform estimation, making the benchmark attainable.
The regret contributions of the three algorithmic phases scale, up to logarithmic factors, as follows: Phase I (budget accumulation) incurs $\widetilde{O}(K\sqrt{T})$ regret, while Phase II contributes $\widetilde{O}(Km)$, and Phase III adds $\widetilde{O}(K\sqrt{T}+T/\sqrt{m})$. These terms account for learning the best feasible distribution over price pairs on the grid. Then, under bounded density $\|f_{S,B}\|_\infty \leq \sigma$, the expected gain from trade is Lipschitz in prices; thus, restricting policies to the grid $\mathcal{G}_K$ introduces a discretization gap of order $O(\sigma \cdot T/K)$. Balancing these terms amounts to minimizing

$$\widetilde{O}\left(Km + K\sqrt{T} + \frac{T}{\sqrt{m}} + \frac{T}{K}\right)$$

As mentioned before, the confidence bonuses are calibrated at level $\delta \in (0,1)$ so that the contribution of the failure event to the expected regret is at most $T\delta = O(1)$, and is therefore absorbed into the asymptotical rate.
Finally, choosing the following values for the parameters $K = T^{1/4}$, $m = T^{1/2}$ and $\beta = \widetilde{O}(T^{3/4})$ yields the final pseudo-regret rate

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

This matches the canonical $\widetilde{O}(T^{3/4})$ rate previously associated with realistic feedback and budget constraints in related formulations [2], despite the benchmark being strictly stronger because it allows arbitrary randomization over two-dimensional price pairs. Throughout the execution, the cumulative budget is never allowed to become negative; therefore, global budget balance holds.

7. Comments on Learnability

The contrast between adversarial and stochastic environments is central to this work. Bernasconi et al. [1] prove that sublinear regret against the globally budget-balanced distributional benchmark is impossible under adversarial valuations. These results show that this limitation stems from the lack of statistical regularity rather

than from the benchmark itself. In stochastic environments, concentration enables uniform estimation and makes the stronger distributional benchmark attainable with sublinear regret, thus highlighting a structural separation between adversarial and stochastic regimes.

8. Conclusions

This work studies repeated bilateral trade under global budget balance in the stochastic i.i.d. setting and establishes that learning against a strong *distributional* benchmark is attainable under realistic feedback. The benchmark is the best distribution over price pairs that is globally budget balanced in expectation, and it is strictly more powerful than the best fixed price policy. While this benchmark is provably unattainable in adversarial environments [1], the stochastic assumption enables a learning algorithm with sublinear pseudo-regret.

The algorithmic contribution is a three-phase design that separates (i) budget accumulation, (ii) estimator construction under acceptance-only feedback, and (iii) optimistic constrained optimization over a discretized price space. A key technical ingredient is an unbiased gain from trade estimator that generalizes the decomposition of [2] to arbitrary pairs (p, q) and supports probe sharing, reducing the cost of exploration. Under a bounded density assumption on the valuation distribution and an appropriate choice of discretization and exploration parameters, the resulting pseudo-regret scales as $\widetilde{O}(T^{3/4})$ while maintaining global budget balance.

This shows that learning against a stronger distributional benchmark in stochastic environments does not worsen achievable regret rates relative to classical guarantees.

References

[1] 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*, 2024.

[2] 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*, 2021.

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