REGRET MINIMIZATION IN BILATERAL TRADE WITH PERTURBED MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the problem of maximizing Gain from Trade (GFT) in repeated buyer-seller exchanges subject to global budget balance constraints. While this problem is well-understood in purely adversarial and stochastic settings, these environments exhibit a sharp dichotomy: adversarial environments allow for no-regret learning against the best fixed-price mechanism, whereas stochastic environments allow for no-regret learning against the best distribution over prices that is budget balanced in expectation. This gap is significant, as policies balanced in expectation can increase the GFT by a multiplicative factor of two. In this work, we bridge these extremes by studying perturbed markets, where an underlying stochastic distribution is subject to an adversarial corruption C. For a smooth unperturbed distribution, we design an algorithm that adaptively scales with the level of corruption, achieving an $\tilde{\mathcal{O}}(T^{3/4})+\mathcal{O}(C\log(T))$ regret bound against the best budget-balanced distribution over prices. Simultaneously, our algorithm maintains the worst-case $\tilde{\mathcal{O}}(T^{3/4})$ regret bound relative to a per-round budget-balanced baseline, ensuring optimality even in fully adversarial environments.

1 INTRODUCTION

The bilateral trade problem stands as a cornerstone of mechanism design, modeling the exchange of a good between a buyer and a seller through an intermediary Myerson and Satterthwaite (1983). While classic results primarily address the Bayesian setting (Blumrosen and Dobzinski, 2014; Kang et al., 2022; McAfee, 2008; Blumrosen and Mizrahi, 2016; Brustle et al., 2017; Deng et al., 2022; Fei, 2022), starting from (Cesa-Bianchi et al., 2021), recent interest has shifted toward the repeated version of this problem, analyzed through the lens of online learning.

In the online version of the bilateral trade problem, a buyer and a seller with valuations b_t and s_t, respectively, arrive at each time step $t \in [T]$, with the learner implementing a mechanism that posts prices (p_t, q_t) to the seller and the buyer, respectively. A trade occurs if both agents find the transaction profitable, *i.e.*, $s_t \leq p_t$ and $b_t \geq q_t$. The goal of the intermediary, *i.e.*, the learner, is to maximize the cumulative Gain from Trade (GFT):

$$\sum_{t=1}^{T} \mathsf{GFT}_t(p_t, q_t) = \sum_{t=1}^{T} (b_t - s_t)\mathbb{I}(s_t \leq p_t, b_t \geq q_t),$$

subject to some budget balance constraints related to not subsidizing the market.

Early works focused on mechanisms that satisfy budget balance—either strong or weak—on a per-round basis. Under these constraints, the learner is restricted to ensuring that its revenue, *i.e.*, $(q_t - p_t)\mathbb{I}(s_t \leq p_t, b_t \geq q_t)$ is either exactly zero or non-negative in every single interaction. We provide an extensive discussion on these works in Section A. As noted by subsequent works, these constraints might be unnecessarily restrictive. A more flexible paradigm was introduced by Bernasconi et al. (2024b), which proposed the concept of Global Budget Balance (GBB). Under GBB, the mechanism is allowed to subsidize certain trades (incurring negative revenue) as long as these are offset by profits in other rounds. This cumulative approach ensures that, by the end of the time horizon, the protocol remains economically viable for the intermediary while significantly improving the GFT.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

While the primary reason why Bernasconi et al. (2024b) introduces GBB constraints is driven by facilitating learning, it also shows that this can induce a large (in particular a 2 multiplicative factor) increase in the achievable GFT. Under traditional per-round constraints, the best the intermediary could do is to learn the best fixed price in hindsight. In contrast, GBB allows the learner to balance revenue across rounds, making the best fixed distribution over prices (budget-balanced in expectation) a more ambitious and natural benchmark. This distributional benchmark was first explored by Bernasconi et al. (2024b), which established that such a benchmark is unlearnable in adversarial environments and sublinear regret is impossible.

Motivated by this negative result, Lunghi et al. (2026) shows that in stochastic environments with sufficient regularity, sublinear regret of $\tilde{\mathcal{O}}(T^{3/4})$ is attainable. These contrasting results raise the following question:

Can we design a "best-of-both-worlds" learning algorithm that effectively interpolates between the stochastic and adversarial regimes?

We address this question by designing algorithms that operate optimally in "perturbed markets", which introduce a total corruption C to an underlying stochastic environment. This framework allows us to bridge the gap between the two regimes: when $C = 0$, we recover the purely stochastic setting, whereas $C = T$ captures the adversarial one.

1.1 Our Results and Techniques

We design an algorithm that adapts to the nature of the environment, achieving the best-attainable guarantees across stochastic, perturbed, and adversarial settings. We evaluate our learner against two benchmarks: (i) the fixed-price benchmark $\mathsf{Opt}^{\mathrm{F}}$, which is the cumulative GFT of the best fixed price; (ii) the distributional benchmark $\mathsf{Opt}^{\mathrm{D}}$, which is the cumulative GFT of the best GBB distribution.

Our algorithm ensures global budget balance, while providing the following regret guarantee:

$$\sum_{t=1}^{T} \mathsf{GFT}_t(p_t, q_t) \geq \max\left\{\mathsf{Opt}^{\mathrm{D}} - \mathcal{O}(C \log T),\, \mathsf{Opt}^{\mathrm{F}}\right\} - \tilde{\mathcal{O}}(T^{3/4}).$$

This matches the optimal stochastic rates, remains robust to sublinear corruption C, and recovers the worst-case adversarial guarantees (see Table 1 for a schematic comparison).

Primal-Dual Methods and Bilateral Trade To the best of our knowledge, this work presents the first primal-dual framework applied to bilateral trade. At its core lies the tension between maximizing the GFT and maintaining a non-negative cumulative revenue. We resolve this by alternating between two specialized procedures: (i) a primal-dual method that handles the constrained optimization of GFT and (ii) a revenue maximization algorithm, based on the one of Bernasconi et al. (2024b), invoked to replenish the budget when a violation arises.

Bypassing the Slater Condition In constrained online problems, primal-dual methods usually yield optimal rates only under Slater condition—which in our context would require the existence of a policy that generates strictly positive revenue at every round. This condition is clearly violated in bilateral trade. We demonstrate that this requirement can be replaced by a significantly weaker condition: a bounded ratio between the optimal GFT and the optimal revenue (which is at most $\mathcal{O}(\log T)$). This allows us to relate the regret incurred during the revenue maximization phase directly to the collected revenue. We then link this revenue—which is in turn related to the violation of the primal-dual algorithm—to its negative regret. This "surplus" gain from trade generated during budget-violating rounds compensates for the regret suffered while replenishing the budget.

We also mention that one non-trivial component needed by our primal-dual method is the design of a novel primal regret minimizer that optimizes a linear combination of GFT and revenue.

Why Perturbed Markets? We define regret relative to the corruption of a stochastic distribution. While alternative definitions of non-stationarity exist (*e.g.*, total variation between consecutive rounds), these are too stringent for this problem. As shown by the lower bounds in (Bernasconi et al., 2024b), even a single change in the environment over the horizon can force linear regret

Table 1: Comparison with state of the art.

	Baseline	Assumptions	Total Variation C	Regret
Bernasconi et al. (2024b)	$\mathsf{Opt}^{\mathrm{F}}$	None	$C \in [0, T]$	$\tilde{\mathcal{O}}(T^{3/4})$
Bernasconi et al. (2024b)	$\mathsf{Opt}^{\mathrm{D}}$	None	$C \in [0, T]$	$\Omega(T)$
Lunghi et al. (2026)	$\mathsf{Opt}^{\mathrm{D}}$	σ-smooth	$C = 0$	$\tilde{\mathcal{O}}(T^{3/4})$
Our work	$\mathsf{Opt}^{\mathrm{D}}$	σ-smooth	$C \in [0, T]$	$\tilde{\mathcal{O}}(T^{3/4} + C)$
Our work	$\mathsf{Opt}^{\mathrm{F}}$	None	$C \in [0, T]$	$\tilde{\mathcal{O}}(T^{3/4})$

against a distributional benchmark. This suggests that perturbed markets are the right tool to analyze "nearly-stochastic" environments.

Relation to Bandit With Knapsacks Our approach is conceptually related to "best-of-both-worlds" algorithms for bandits with knapsacks and more general constraints, where primal-dual methods are highly effective Balseiro et al. (2020); Fikioris and Tardos (2023); Bernasconi et al. (2025); Slivkins et al. (2023); Castiglioni et al. (2023; 2022). However, this literature typically relies on the Slater condition to achieve optimal bounds. Without it, it is generally impossible to achieve a constant fraction of the optimal reward. Our results can be viewed as a Slater-free result for bilateral trade, achieved by leveraging strict satisfiability in a global sense. In particular, we employ the $\log(T)$ upper bound on the ratio of reward to negative violation that holds globally. The closest to our work is (Bernasconi et al., 2024a), which uses a similar dual analysis to bound the primal violation as a function of its negative regret.

2 Preliminaries

We study repeated bilateral trade through the lens of online learning. Over $T \in \mathbb{N}$ rounds (with $T \geq 2$), a sequence of seller-buyer pairs arrives. In each round t, the seller and buyer hold private valuations, denoted by $s_t \in [0, 1]$ and $b_t \in [0, 1]$, respectively. These valuations are sampled independently across rounds from a joint time-dependent probability distribution $\mathcal{L}_t$ with support in $[0, 1]^2$, chosen by an oblivious adversary.

Learning Protocol At each round t, the learner posts two prices: $p_t \in [0, 1]$ to the seller and $q_t \in [0, 1]$ to the buyer. A trade occurs if both parties accept—that is, if $s_t \leq p_t$ and $b_t \geq q_t$. When a trade occurs, the seller receives p_t and the buyer pays q_t; otherwise no transaction takes place.

To evaluate the performance of the mechanism, we consider two metrics: *revenue*, which represents the profit or loss of the mechanism, and the *Gain from Trade* (GFT), which serves as a proxy for the social welfare generated. Formally, for valuations $(s, b) \in [0, 1]^2$ and prices $(p, q) \in [0, 1]^2$, we define:

$$\mathsf{GFT}(p, q, s, b) := (b - s)\mathbb{I}(b \geq q, s \leq p),$$
$$\mathsf{Rev}(p, q, s, b) := (q - p)\mathbb{I}(b \geq q, s \leq p).$$

For brevity, we define $\mathsf{GFT}_t(p, q) := \mathsf{GFT}(p, q, s_t, b_t)$ and $\mathsf{Rev}_t(p, q) := \mathsf{Rev}(p, q, s_t, b_t)$.

Feedback The learner does not observe the private valuations (s_t, b_t), either before or after posting prices; instead, only partial information is revealed through the feedback received after each round. We focus on the most restrictive setting, **one-bit feedback**. Under this model, the learner only observes whether the trade occurred, *i.e.*, the realization of the indicator $\mathbb{I}(b_t \geq q_t, s_t \leq p_t)$.

Budget Balance We consider the most relaxed notion of budget balance, which is called *Global Budget Balance* (GBB). An algorithm is said to be GBB if its cumulative revenue is non-negative, *i.e.*, $\sum_{t=1}^{T} \mathsf{Rev}_t(p_t, q_t) \geq 0$. This constraint must hold almost surely regardless of the algorithm's internal randomization and the realizations of the environment.

Baselines We consider two standard benchmarks in the literature. The first is the best fixed price, which is inherently budget-balanced on a per-round basis and represents the optimal performance achievable in adversarial environments (Bernasconi et al., 2024b). Formally, we define:

$$\mathsf{Opt}^{\mathrm{F}} := \max_{p\in[0,1]} \sum_{t=1}^{T} \mathsf{GFT}_t(p,p). \tag{1}$$

The second baseline is the best feasible distribution over price pairs that is budget-balanced in expectation. This serves as the target benchmark in stochastic settings. Formally, we define:

$$\begin{aligned} \mathsf{Opt}^{\mathrm{D}} := \max_{\pi\in\Delta([0,1]^2)} \quad & \sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] \\ \text{s.t.} \quad & \sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq 0, \end{aligned} \tag{2}$$

where $\Delta([0,1]^2)$ represents the space of all the well-defined distributions over $[0,1]^2$.

Environment We study *perturbed markets* that interpolate between stochastic and adversarial settings. Let $\Delta([0,1]^2)$ be the set of distributions supported on $[0,1]^2$. The environment is characterized by a probability distribution $\mathcal{P} \in \Delta([0,1]^2)$, representing the private valuations distribution in the unperturbed market. Moreover, there is a perturbation parameter C, which is defined as the cumulative total variation distance between the sequence of distributions $\{\mathcal{L}_t\}_{t=1}^T$ chosen by the adversary and the unperturbed stationary distribution $\mathcal{P}$:

$$C := \sum_{t=1}^{T} \|\mathcal{L}_t - \mathcal{P}\|_{\mathrm{TV}}.$$

Note that $C = 0$ in a purely stochastic environment where all $\mathcal{L}_t$ are equal to $\mathcal{P}$, while C can be as large as T in the worst case (the adversarial setting).

When dealing with the distributional benchmark $\mathsf{Opt}^{\mathrm{D}}$, as shown by Lunghi et al. (2026), it is necessary to impose some additional regularity conditions on the environment. In particular, we assume the unperturbed distribution $\mathcal{P}$ to be σ-smoothed:

Definition 1. *Let X support a uniform probability measure ν, and let $\sigma \geq 1$. A probability measure μ on X is σ-smooth if $\mu(A) \leq \sigma\nu(A)$ for every measurable $A \subseteq X$.*

3 Main Result and Proof Plan

We design a best-of-both-worlds algorithm with optimal guarantees in both stochastic and adversarial settings. Furthermore, its performance degrades gracefully as a function of the perturbation.

Theorem 1. *Suppose the unperturbed distribution $\mathcal{P}$ is σ-smooth for a constant σ. There exists a GBB algorithm that, with probability $1 - \mathcal{O}(1/T)$, guarantees:*

$$\sum_{t=1}^{T} \mathsf{GFT}_t(p_t,q_t) \geq \max\{\mathsf{Opt}^{\mathrm{D}} - \mathcal{O}(C\log(T)), \mathsf{Opt}^{\mathrm{F}}\} - \tilde{\mathcal{O}}(T^{3/4}).$$

The guarantee against $\mathsf{Opt}^{\mathrm{F}}$ holds without the smoothness assumption.

Two considerations are in order. First, a linear dependency on C is expected in this context, while we leave it as an open problem to determine whether the multiplicative $\log T$ factor is strictly necessary. Second, our results recover the optimal regret rates of Lunghi et al. (2026) and Bernasconi et al. (2024b) for the stochastic and adversarial regimes, respectively. It is worth noting that the result in Bernasconi et al. (2024b) holds without the smoothness assumption. This is also true for our algorithm: the smoothed adversary assumption is only required when competing against the stronger $\mathsf{Opt}^{\mathrm{D}}$ baseline, where this assumption is required Lunghi et al. (2026).

The remainder of this paper is organized as follows. Section 4 shows how and when we can restrict to a finite grid of price pairs. Section 5 presents the main algorithm, while Section 6 presents our primal-dual algorithm. In Section 7, we present the primal regret minimizer. Finally, in Section 8, we combine all the results to prove Theorem 1.

4 From Continuous to Discrete Decision Space

We begin by examining the conditions under which the continuous price space can be effectively approximated by a discrete grid. In bilateral trade, both the objective and the budget balance constraint are ill-behaved, *e.g.*, they lack continuity. Hence, as already clear by the first works on the topic, see, *e.g.*, (Cesa-Bianchi et al., 2024a), methods working on the continuum are ineffective, leaving grid-based approaches as the only feasible strategy. However, without some form of regularity, grids are often ineffective. This was already noted by Cesa-Bianchi et al. (2021; 2023) regarding the fixed-price benchmark when the learner is constrained by per-round budget balance prices. Lunghi et al. (2026) showed a similar result holds for the distributional benchmark, even when the learner can employ GBB policies. In particular, both works show that it is impossible to achieve sublinear regret, even in a completely stochastic environment ($C = 0$ in our setting).

Both works circumvent these impossibility results by assuming σ-smoothness (see Definition 1). Notice that this regularity assumption is required only when algorithm and baseline have comparable power. Indeed, Bernasconi et al. (2024b) shows that it is possible to achieve no regret with respect to the best fixed price using GBB policies, without assumptions. This is true also for our algorithm. Therefore, we follow the approach of Lunghi et al. (2026) and show that a uniform grid

$$\mathcal{G}_K \coloneqq \left\{ \left(\frac{i}{K}, \frac{j}{K} \right) : i, j \in \{0, \dots, K\} \right\}$$

provides an approximation with error $\tilde{\mathcal{O}}(T/K + C)$ when the unperturbed distribution $\mathcal{P}$ is σ-smooth.

Let $\Delta(\mathcal{G}_K)$ be the space of all possible distributions over pair of prices on the grid $\mathcal{G}_K$. Define:

$$\mathsf{Opt}_K^{\mathrm{F}} = \begin{cases} \max\limits_{\pi \in \Delta(\mathcal{G}_K)} \sum\limits_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] \\ \quad \text{s.t.} \quad \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq -\frac{1}{K} \quad \forall t \in [T], \end{cases} \tag{3}$$

$$\mathsf{Opt}_K^{\mathrm{D}} = \begin{cases} \max\limits_{\pi \in \Delta(\mathcal{G}_K)} \sum\limits_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] \\ \quad \text{s.t.} \quad \sum\limits_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq 0. \end{cases} \tag{4}$$

Intuitively, $\mathsf{Opt}_K^{\mathrm{D}}$ requires prices to be budget balanced in expectation over the rounds, similarly to GBB, while $\mathsf{Opt}_K^{\mathrm{F}}$ requires the distribution to be almost budget balanced at each round. Then:

Lemma 1. *With probability at least* $1 - 1/T$*, it holds:* $R_{0,K} = \mathsf{Opt}^{\mathrm{F}} - \mathsf{Opt}_K^{\mathrm{F}} \leq \widetilde{\mathcal{O}}\left(\frac{T}{K} + \sqrt{T \log(TK)}\right)$.

Lemma 2. *If* $\mathcal{P}$ *is* σ*-smooth, it holds:* $R_K = \mathsf{Opt}^{\mathrm{D}} - \mathsf{Opt}_K^{\mathrm{D}} \leq \tilde{\mathcal{O}}\left(\frac{T}{K} + C\right)$*, where the notation* $\mathcal{O}(\cdot)$ *contains the linear dependence on* σ*, which we assume constant.*

As promised, we need the σ-smoothness assumption only when dealing with the distributional benchmark. Similarly to Bernasconi et al. (2024b), we remove the assumption for per-round budget balance prices thanks to GBB. These results allow us to restrict our focus to a discrete uniform grid.

5 A Primal-Dual Approach to Bilateral Trade

In this section, we introduce our novel algorithm, whose pseudocode is detailed in Algorithm 1. The algorithm alternates between two procedures: Primal-Dual and Rev-Max. At the core of our approach is a high-level strategy inspired by Bernasconi et al. (2024b), which involves alternating between a revenue maximization phase—designed to generate positive profit and recover lost budget as quickly as possible—and a policy that might violate the GBB constraint. Unlike Bernasconi et al. (2024b), however, we cannot precompute the violations. Hence, we must continuously alternate between these two policies. Both procedures are initialized once and retain their state between calls. Each call resumes one round of the selected procedure; only that procedure receives the feedback and updates its state. In the pseudocode of each procedure, t counts its own active rounds, up to the common horizon T.

It will be useful to define the set of rounds $\mathcal{T}_1 \subseteq [T] := \{1, \dots, T\}$ in which Algorithm 1 chooses to follow procedure Rev-Max and $\mathcal{T}_2 \subseteq [T]$ as the set of rounds in which Algorithm 1 chooses to follow procedure Primal-Dual, such that $\mathcal{T}_1 \cup \mathcal{T}_2 = [T]$.

Algorithm 1 No-Regret Algorithm

Input: T, δ
Initialize both procedures; $B_0 \leftarrow 0$
for $t = 1, \dots, T$ **do**
  if $B_{t-1} < 1$ **then**
    Run one round of Rev-Max
  else
    Run one round of Primal-Dual
  $B_t \leftarrow B_{t-1} + \mathsf{Rev}_t(p_t, q_t)$

The main component of our algorithm is a completely novel primal-dual algorithm that seeks to maximize the GFT while satisfying the budget constraint. The pseudocode of the algorithm is presented in Algorithm 2, while its analysis is deferred to Section 6. To the best of our knowledge, this is the first primal-dual approach applied to bilateral trade. Although we cannot show that the algorithm satisfies the constraint, we show that whenever the algorithm violates the budget, it incurs *negative regret*. This negative regret provides a buffer, allowing the learner to "lose" rounds during the Rev-Max phase.

Instead, the procedure Rev-Max is specifically designed to recover the lost budget, while partially mitigating the loss in terms of GFT while doing so. We use the same procedure of Bernasconi et al. (2024b) which consists in employing an algorithm of revenue maximization over a grid of prices of logarithmic nature. Bernasconi et al. (2024b) shows the following result.

Algorithm 2 Primal-Dual

Input: T, δ
Initialize Primal as in Section 8
$\lambda_1 \leftarrow 0$
$M \leftarrow 16 \log(T)$
$\eta \leftarrow 1/\sqrt{T}$
for $t = 1, \dots, T$ **do**
  Post $(p_t, q_t) \leftarrow$ Primal
  Feed $\mathbb{I}(s_t \le p_t, b_t \ge q_t)$ and λ_t to Primal
  $\lambda_{t+1} \leftarrow \Pi_{[0,M]}\left(\lambda_t - \eta \mathsf{Rev}_t(p_t, q_t)\right)$

Lemma 3 (Essentially (Bernasconi et al., 2024b))**.** *There exists an algorithm* Rev-Max*, posting only $p_t \le q_t$, that guarantees with probability at least $1 - 1/T$:*

$$\sum_{t \in \mathcal{T}_1} \mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] \le 16 \log T \cdot \left(\sum_{t \in \mathcal{T}_1} \mathsf{Rev}_t(p_t, q_t) + 1 \right) + \tilde{\mathcal{O}}(T^{3/4})$$

for all $\pi \in \Delta([0,1]^2)$ such that $\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}\,(p,q,s,b)] \ge 0$.

Since Rev-Max has nonnegative revenue and every primal round starts with budget at least 1 and loses at most 1, Algorithm 1 maintains $B_t \ge 0$ deterministically.

Interestingly, this result is related to a relaxed Slater condition. Standard primal-dual methods Bernasconi et al. (2025); Castiglioni et al. (2023; 2022); Balseiro et al. (2020) for settings with adversarial constraints require a positive and sufficiently large parameter

$$\rho = \max_{\pi \in \Delta([0,1]^2)} \min_{t \in [T]} \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)].$$

In our setting the Slater parameter ρ can be zero, so a strictly positive lower bound cannot be assumed. Despite that, our setting satisfies a weaker condition in which the relationship between the cumulative GFT and the accumulated budget has a global nature and does not require a strict per-round assumption.

6 Regret Minimization via Primal-Dual Framework

We frame the online bilateral trade problem as an online constrained optimization problem, which we Lagrangify obtaining the reward function:

$$L_t(\pi, \lambda) = \mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] + \lambda \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)].$$

The algorithm works by instantiating two distinct regret minimizers: (i) a primal procedure that optimizes the reward function L_t without explicit budget balance constraints and (ii) a dual procedure that updates the multipliers to penalize constraint violations.

While the primal procedure can be any algorithm that achieves no-regret with high probability, the dual multipliers must be selected by an algorithm that guarantees weakly-adaptive regret minimization. Given two integers a, b, we define $[a, b] = \{a, a+1, \dots, b-1, b\}$. Specifically, the dual algorithm must work on the domain $[0, M]$, where $M = 16 \log T$, and ensure no-regret over all possible sub-intervals $[t_1, t_2] \subseteq [T]$. The intervals are in the dual procedure's active-round index (equivalently, use zero gradients on $\mathcal{T}_1$). This can be achieved by Online Gradient Descent, which provides the following guarantees (see, *e.g.*, (Hazan, 2016)):

$$\sum_{t=t_1}^{t_2} (\lambda_t - \lambda) \mathsf{Rev}_t(p_t, q_t) \leq R_T^D = \tilde{\mathcal{O}}(\sqrt{T}) \quad \forall \lambda \in [0, M], [t_1, t_2] \subseteq [T]. \tag{5}$$

Instead, designing a primal regret minimizer is a nontrivial task that we address in Section 7. For now, we assume that with probability at least $1 - \mathcal{O}(1/T)$, the primal minimizer satisfies:

$$\sum_{t=1}^{T} \left(\mathbb{E}_{(p,q)\sim\pi} [\mathsf{GFT}_t(p, q) + \lambda_t \mathsf{Rev}_t(p, q)] \right) - \sum_{t=1}^{T} \left(\mathsf{GFT}_t(p_t, q_t) + \lambda_t \mathsf{Rev}_t(p_t, q_t) \right) \leq R_T^P. \tag{6}$$

Notice that we parametrize the results by R_T^P and discuss the specific form of the regret bound R_T^P in the following.

Reordering the terms of Equation (6), and considering that the primal regret minimizer is active on the subset of rounds $\mathcal{T}_2$ we get that, given a $\pi \in \Delta(\mathcal{G}_K)$, with probability at least $1 - \mathcal{O}(1/T)$:

$$\begin{aligned} &\sum_{t \in \mathcal{T}_2} \mathbb{E}_{(p,q)\sim\pi} [\mathsf{GFT}_t(p, q)] - \sum_{t \in \mathcal{T}_2} \mathsf{GFT}_t(p_t, q_t) \\ &\qquad \leq R_T^P + \sum_{t \in \mathcal{T}_2} \lambda_t \mathsf{Rev}_t(p_t, q_t) - \sum_{t \in \mathcal{T}_2} \lambda_t \mathbb{E}_{(p,q)\sim\pi} [\mathsf{Rev}_t(p, q)]. \end{aligned} \tag{7}$$

In the following sections, we bound the term $\sum_{t \in \mathcal{T}_2} \lambda_t \mathsf{Rev}_t(p_t, q_t)$ exploiting the adaptivity of the dual regret minimizer and the term $-\sum_{t=1}^{T} \lambda_t \mathbb{E}_{(p,q)\sim\pi} [\mathsf{Rev}_t(p, q)]$ exploiting the definition of C.

In the following sections, we will study separately the two regret terms on the right that depend on the revenue.

6.1 Adaptivity of the Dual Update

In this section, we bound the term $\sum_{t \in \mathcal{T}_2} \lambda_t \mathsf{Rev}_t(p_t, q_t)$ by exploiting the weak adaptivity of the dual regret minimizer. This term relates the budget balance violations of the Primal-Dual procedure to its accumulated negative regret. Crucially, we require any budget violation to be "sufficiently profitable" to offset the regret incurred while replenishing the budget via the Rev-Max procedure.

This idea is conceptually similar to the approach employed by Bernasconi et al. (2024a) for the related problem of bandits with knapsacks. We split $\mathcal{T}_2$ into two intervals around the last activation round τ of Rev-Max. By construction, the budget at time τ is close to zero. Formally, $B_\tau = \sum_{t \in \mathcal{T}_1 \cap [\tau]} \mathsf{Rev}_t(p_t, q_t) + \sum_{t \in \mathcal{T}_2 \cap [\tau]} \mathsf{Rev}_t(p_t, q_t) \in [0, 2]$.

Applying the dual regret guarantee in the interval $\mathcal{T}_2 \cap [\tau]$ with respect to $\lambda = M$, we can state:

$$\sum_{t \in \mathcal{T}_2 \cap [\tau]} \lambda_t \mathsf{Rev}_t(p_t, q_t) \leq M \left(\sum_{t \in \mathcal{T}_2 \cap [\tau]} \mathsf{Rev}_t(p_t, q_t) \right) + R_T^D.$$

For the remaining rounds in the second interval $\mathcal{T}_2 \cap [\tau + 1, T]$, the term remains small because the dual algorithm guarantees no regret with respect to $\lambda = 0$:

$$\sum_{t \in \mathcal{T}_2 \cap [\tau+1, T]} \lambda_t \mathsf{Rev}_t(p_t, q_t) \leq 0 + R_T^D.$$

This leads to the following result, which employs the almost equivalence of the negative revenue in $\mathcal{T}_2 \cap [\tau]$ and the positive revenue in $\mathcal{T}_1$.

Lemma 4. *Algorithm 1 guarantees:*

$$\sum_{t\in\mathcal{T}_2}\lambda_t\mathsf{Rev}_t(p_t,q_t)\le M\left(-\sum_{t\in\mathcal{T}_1}\mathsf{Rev}_t(p_t,q_t)+2\right)+2R_T^D$$

Notice that the simpler approach that does not employ weak adaptivity is doomed to fail. If, for example, the Rev-Max algorithm is forced to balance a temporary loss in the budget of the Primal-Dual algorithm—which the Primal-Dual would have recovered on its own later on—it generates a large sum $\sum_{t\in\mathcal{T}_1}\mathsf{Rev}_t(p_t,q_t)+\sum_{t\in\mathcal{T}_2}\mathsf{Rev}_t(p_t,q_t)$, while the individual term $\sum_{t\in\mathcal{T}_1}\mathsf{Rev}_t(p_t,q_t)$ (intuitively related to the regret in rounds $\mathcal{T}_1$) also remains large.

6.2 Dependence on C

We bound the comparator-revenue term in Equation (7) separately for the two benchmarks. For a solution of Program 3, the per-round expected-revenue lower bound $-1/K$ yields an MT/K bound up to concentration. For a solution of Program 4, budget balance holds only over the full horizon, and we use the perturbation C to control the effect of time-varying multipliers.

Lemma 5. *Let π be a solution to Program 3 , then with probability at least $1-1/T$:*

$$-\sum_{t\in\mathcal{T}_2}\lambda_t\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)]\le M\cdot\left(\frac{T}{K}\right)+\tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

Moreover, let π be a solution to Program 4, then with probability at least $1-1/T$:

$$-\sum_{t\in\mathcal{T}_2}\lambda_t\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)]\le 2M\cdot C+\tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

7 Primal Algorithm and the Observability Challenge

In this section, we develop a primal regret minimizer designed to satisfy the requirements of Equation (6) with an optimal regret bound R_T^P. The algorithm works over a set of price pairs $\mathcal{A}$—specifically, we will apply the result for the grid $\mathcal{G}_K$—and provides guarantees that scale linearly with the maximum value attained by the Lagrangian multipliers $M=16\log(T)$.

A fundamental challenge in designing learning algorithms for bilateral trade is the observability challenge. Unlike standard learning settings, feedback here is significantly more limited: even basic bandit feedback is missing. To overcome this limited observability, we leverage the decomposition introduced by Lunghi et al. (2026). We partition the GFT into three components: a "bandit" term (the revenue) and two "global" terms that can be estimated independently by probing each side of the market.

Algorithm 3 Primal Algorithm

Input: $\mathcal{A},\eta^P,T,\gamma,\alpha,M$
$w_1(p,q)\leftarrow 1$, $\hat{\pi}_1(p,q)\leftarrow 1/|\mathcal{A}|\quad\forall(p,q)\in\mathcal{A}$
for $t=1,\dots,T$ **do**
  Sample $(\hat{p}_t,\hat{q}_t)\sim\hat{\pi}_t$
  Sample $H_t\in\{0,1,2\}$, such that
    $\mathbb{P}(H_t=0)=1-\alpha$
    $\mathbb{P}(H_t=1)=\mathbb{P}(H_t=2)=\alpha/2$
  Post prices

$$(p_t,q_t)=\begin{cases}(\hat{p}_t,\hat{q}_t) & \text{if } H_t=0\\ (U_t\sim\mathcal{U}([0,1]),\hat{q}_t) & \text{if } H_t=1\\ (\hat{p}_t,V_t\sim\mathcal{U}([0,1])) & \text{if } H_t=2\end{cases}$$

  Receive feedback $\mathbb{I}(s_t\le p_t,b_t\ge q_t)$ and Lagrangian multiplier λ_t
  Build loss feedback $\tilde{\ell}_t(p,q)\ \forall(p,q)\in\mathcal{A}$ as in Equation (8)
  $w_{t+1}(p,q)\leftarrow w_t(p,q)\exp\left(-\eta^P\tilde{\ell}_t(p,q)\right)\ \forall(p,q)\in\mathcal{A}$
  $\hat{\pi}_{t+1}(p,q)\leftarrow\frac{w_{t+1}(p,q)}{\sum_{(p',q')\in\mathcal{A}}w_{t+1}(p',q')}\quad\forall(p,q)\in\mathcal{A}$

Lemma 6 (Lemma 8.1, Lunghi et al. (2026))**.** *Let $U,V\sim\mathcal{U}([0,1])$ be independent, where $\mathcal{U}([0,1])$ is the uniform distribution over $[0,1]$. For fixed valuations (s_t,b_t), the GFT of prices (p,q) can be decomposed as:*

$$\mathsf{GFT}_t(p,q)=\underbrace{L_t(p,q)}_{\textit{Seller Term}}+\underbrace{R_t(p,q)}_{\textit{Buyer Term}}+\underbrace{\mathsf{Rev}_t(p,q)}_{\textit{Bandit Term}},$$

where $L_t(p,q) \coloneqq \mathbb{E}_U[\mathbb{I}(s_t \le U \le p, q \le b_t)]$ *and* $R_t(p,q) \coloneqq \mathbb{E}_V[\mathbb{I}(s_t \le p, q \le V \le b_t)]$.

Let $\mathcal{A}_s \coloneqq \{p : (p,q) \in \mathcal{A}\}$ and $\mathcal{A}_b \coloneqq \{q : (p,q) \in \mathcal{A}\}$ be the set of unique seller and buyer prices, respectively, induced by $\mathcal{A}$. We will show that $L_t(p,q)$ can be estimated simultaneously for all $p \in \mathcal{A}_s$ whenever q is fixed, and $R_t(p,q)$ can be estimated for all $q \in \mathcal{A}_b$ whenever p is fixed. This, however, requires a "full exploration" strategy, which we invoke with probability α. By sharing feedback across actions that share a common price, we significantly reduce the estimator variance and hence the number of rounds dedicated to exploration.

Formally, our algorithm employs an EXP3.IX-style update with loss estimator $\tilde{\ell}_t(p,q)$ that employs importance sampling combined with an implicit exploration bias Kocák et al. (2014); Neu (2015).

$$\tilde{\ell}_t(p,q) = \begin{cases} \dfrac{1 - \mathbb{I}(s_t \le U_t \le p, b_t \ge q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} & \text{if } q = \hat{q}_t \text{ and } H_t = 1 \\ \dfrac{1 - \mathbb{I}(s_t \le p, b_t \ge V_t \ge q)}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}} \hat{\pi}_t(p,q') + \gamma} & \text{if } p = \hat{p}_t \text{ and } H_t = 2 \\ \dfrac{(1+\lambda_t)\,(1 - (q_t - p_t)\mathbb{I}(s_t \le \hat{p}_t, b_t \ge \hat{q}_t))}{(1-\alpha)\hat{\pi}_t(p,q) + \gamma} & \text{if } (p,q) = (\hat{p}_t, \hat{q}_t) \text{ and } H_t = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

For this choice of $\tilde{\ell}_t$ and learning algorithm we can guarantee the following:

Theorem 2. *Let* $|\mathcal{A}| \ge 2$, $0 < \alpha \le 1/2$, *and let* $0 \le \lambda_t \le M$ *be predictable multipliers with a deterministic bound* $M \ge 1$. *For the choices of* η^P *and* γ *given in the proof, Algorithm 3 ensures, simultaneously for all* $\pi \in \Delta(\mathcal{A})$ *and* $\tau \in [T]$, *with probability at least* $1 - \mathcal{O}(1/T)$*:*

$$\sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\left(\mathsf{GFT}_t(p,q) + \lambda_t \mathsf{Rev}_t(p,q)\right) - \sum_{t=1}^{\tau}\left(\mathsf{GFT}_t(p_t,q_t) + \lambda_t \mathsf{Rev}_t(p_t,q_t)\right) \le \tilde{\mathcal{O}}\left(M\left(\sqrt{|\mathcal{A}|T} + \frac{|\mathcal{A}_s| + |\mathcal{A}_b|}{\sqrt{|\mathcal{A}|}}\frac{\sqrt{T}}{\alpha} + \alpha \cdot T\right)\right).$$

8 Putting Everything Together

Finally, we can prove Theorem 1. While the full proof is deferred to the appendix, we provide a sketch here. Set $K = T^{1/4}, \alpha = T^{-1/4}, \mathcal{A} = \mathcal{G}_K, M = 16\log(T)$. We consider two cases:

Distributional Benchmark With σ-Smooth Adversary Let π be the solution of Program 2, by Lemma 3 with probability at least $1 - \mathcal{O}(1/T)$

$$\sum_{t\in\mathcal{T}_1} \mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] - \sum_{t\in\mathcal{T}_1} \mathsf{GFT}_t(p_t,q_t) \le M \sum_{t\in\mathcal{T}_1} \mathsf{Rev}_t(p_t,q_t) + \tilde{\mathcal{O}}(T^{3/4}).$$

By Lemma 5, Lemma 4, and Lemma 2 with probability at least $1 - \mathcal{O}(1/T)$:

$$\sum_{t\in\mathcal{T}_2}\left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] - \mathsf{GFT}_t(p_t,q_t)\right) \le \tilde{\mathcal{O}}(T^{3/4}) - M\sum_{t\in\mathcal{T}_1} \mathsf{Rev}_t(p_t,q_t) + \mathcal{O}(\log(T)C).$$

Hence, with probability at least $1 - \mathcal{O}(1/T)$:

$$\mathsf{Opt}^{\mathrm{D}} - \sum_{t\in[T]}\left(\mathsf{GFT}_t(p_t,q_t)\right) \le \tilde{\mathcal{O}}\left(T^{3/4}\right) + \left(\log(T)C\right).$$

Fixed-Price Benchmark If we use a solution of Program 1 as a benchmark, we can replace Lemma 2 with Lemma 1, and instead of Equation (4) we can use Equation (3) to get:

$$\mathsf{Opt}^{\mathrm{F}} - \sum_{t\in[T]} \mathsf{GFT}_t(p_t,q_t) \le \tilde{\mathcal{O}}\left(T^{3/4}\right).$$

AI USE STATEMENT

Generative AI tools were used for editing the manuscript, including checking the exposition, notation, and consistency of the proof presentation.

REFERENCES

Yossi Azar, Amos Fiat, and Federico Fusco. An α-regret analysis of adversarial bilateral trade. *Artificial Intelligence*, 337:104231, 2024.

Moshe Babaioff, Amitai Frey, and Noam Nisan. Learning to maximize gains from trade in small markets. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 195–195, 2024.

François Bachoc, Nicolò Cesa-Bianchi, Tom Cesari, and Roberto Colomboni. Fair online bilateral trade. *Advances in Neural Information Processing Systems*, 37:37241–37263, 2024.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A contextual online learning theory of brokerage. In *Forty-second International Conference on Machine Learning*, 2025a.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A tight regret analysis of non-parametric repeated contextual brokerage. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025b. URL `https://openreview.net/forum?id=U2RgUAySB6`.

Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pages 613–628. PMLR, 2020.

M Bernasconi, M Castiglioni, A Celli, F Fusco, et al. Bandits with replenishable knapsacks: the best of both worlds. In *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR, 2024a.

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, 2024b.

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization. In *International Conference on Machine Learning*, pages 3877–3898. PMLR, 2025.

Liad Blumrosen and Shahar Dobzinski. Reallocation mechanisms. In *EC*, page 617. ACM, 2014. doi: 10.1145/2600057.2602843.

Liad Blumrosen and Yehonatan Mizrahi. Approximating gains-from-trade in bilateral trading. In *WINE*, volume 10123 of *Lecture Notes in Computer Science*, pages 400–413. Springer, 2016. doi: 10.1007/978-3-662-54110-4_28.

Natasa Bolić, Tommaso Cesari, and Roberto Colomboni. An online learning theory of brokerage. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 216–224, 2024. ISBN 9798400704864.

Johannes Brustle, Yang Cai, Fa Wu, and Mingfei Zhao. Approximating gains from trade in two-sided markets via simple mechanisms. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 589–590, 2017.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35:33589–33602, 2022.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning under budget and roi constraints via weak adaptivity. *arXiv preprint arXiv:2302.01203*, 2023.

Matteo Castiglioni, Anna Lunghi, and Alberto Marchesi. The sample complexity of uniform approximation for multi-dimensional cdfs and fixed-price mechanisms. *arXiv preprint arXiv:2602.10868*, 2026.

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.

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

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

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, 2024b.

Tommaso Cesari and Roberto Colomboni. An online learning theory of trading-volume maximization. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Emanuele Coccia, Martino Bernasconi, and Andrea Celli. Nonparametric contextual online bilateral trade. In *The Fourteenth International Conference on Learning Representations*.

Romain Cosson, Federico Fusco, Anupam Gupta, Stefano Leonardi, Renato Paes Leme, and Matteo Russo. Contextual online bilateral trade. *arXiv preprint arXiv:2602.12903*, 2026.

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

Simone Di Gregorio, Paul Dütting, Federico Fusco, and Chris Schwiegelshohn. Nearly tight regret bounds for profit maximization in bilateral trade. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1570–1594, 2025. doi: 10.1109/FOCS63196.2025.00083.

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

Giannis Fikioris and Éva Tardos. Approximately stationary bandits with knapsacks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3758–3782. PMLR, 2023.

Solenne Gaucher, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Vianney Perchet. Feature-based online bilateral trade. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xnF2U0ro7b.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

Zi Yang Kang, Francisco Pernice, and Jan Vondrák. Fixed-price approximations in bilateral trade. In *SODA*, pages 2964–2985. SIAM, 2022. doi: 10.1137/1.9781611977073.115.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. *Better Regret Rates in Bilateral Trade via Sublinear Budget Violation*, pages 6494–6536. doi: 10.1137/1.9781611978971.233. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978971.233.

Anna Lunghi, Matteo Castiglioni, and Alberto Marchesi. Online two-sided markets: Many buyers enhance learning. *arXiv preprint arXiv:2503.01529*, 2025.

Anna Lunghi, Mattia Piccinato, Matteo Castiglioni, and Alberto Marchesi. A stronger benchmark for online bilateral trade: From fixed prices to distributions. *arXiv preprint arXiv:2602.05681*, 2026.

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

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

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4633–4656. PMLR, 2023.

A Additional Related Works

The study of GFT maximization in online bilateral trade was initiated by Cesa-Bianchi et al. (2024a), who show that under one-bit feedback the problem is in general unlearnable. However, they showed that Strong Budget Balanced (SBB) mechanisms are learnable when the seller and buyer valuation distributions are independent and admit bounded density. Subsequent works shifted toward adversarial environments. Azar et al. (2024) showed that no-2-regret is achievable, while Cesa-Bianchi et al. (2024b) introduced the smoothed-adversary assumption to online bilateral trade, proving that no-regret can be achieved under this assumption.

Bernasconi et al. (2024b) and Chen et al. (2025) showed how GBB can be used to bypass the smoothed-adversary assumption. Despite this, these works mainly focus on competing with the best SBB fixed price. The most relevant result to our work is a lower bound from Bernasconi et al. (2024b) showing that no-regret with respect to the best GBB distribution is unachievable in adversarial settings. More recently, Lunghi et al. analyze the trade-off between the regret and the GBB constraint violation. More closely related to our work is Lunghi et al. (2026), which showed that it is possible to achieve no regret with respect to the distributional benchmark in stochastic setting under a smoothness assumption.

Other related lines of work focus on revenue maximization Castiglioni et al. (2026); Di Gregorio et al. (2025). Finally, we mention that many works extended bilateral trade to more complex problems such as contextual settings (Gaucher et al., 2025; Cosson et al., 2026; Coccia et al.), fairer objectives (Bachoc et al., 2024), multiple buyers (Babaioff et al., 2024; Lunghi et al., 2025), and settings with symmetric roles (Bolić et al., 2024; Cesari and Colomboni, 2025; Bachoc et al., 2025a;b).

B Omitted Proofs From Section 4

To prove that the approximation on the grid $\mathcal{G}_K$ is a good approximation of the optimum on the continuum, we will proceed in two steps: first, we show that there exists a distribution on the grid that is slightly unfeasible but achieves the optimum GFT up to an additive factor $\mathcal{O}(\frac{T}{K} + C)$, then we show that we can interpolate between this policy and a highly profitable policy to build a policy that is feasible and maintains the same order distance from the optimum.

This procedure is highly similar to the one proposed by Lunghi et al. (2026), slightly adapted to the non-stationarity of the environment, for what concern the best distribution.

Lemma 7. *Given the sequence of distributions $\{\mathcal{L}_t\}_{t=1}^T$, for all $\gamma \in \Delta([0,1]^2)$ satisfying $\sum_{t=1}^T \mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq 0$, there exists a $\hat{\gamma} \in \Delta(\mathcal{G}_K)$ that satisfies*

$$\sum_{t=1}^T \mathbb{E}_{(p,q)\sim\hat{\gamma},(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] \geq \sum_{t=1}^T \mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] - \mathcal{O}\left(\frac{\sigma T}{K} + C\right),$$

and

$$\sum_{t=1}^T \mathbb{E}_{(p,q)\sim\hat{\gamma},(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq -\mathcal{O}\left(\frac{\sigma T}{K} + C\right).$$

Proof. Let $\mathcal{P}$ be a σ-smooth distribution . Let $(p,q) \in [0,1]^2$, and let $(\Pi_K^s(p), \Pi_K^b(q))$ denote the nearest point in the grid $\mathcal{G}_K$ such that $\Pi_K^s(p) \geq p$ and $\Pi_K^b(q) \leq q$.

We decompose the gain-from-trade as follows:

$$\begin{aligned}
\mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{GFT}(p,q,s,b)] &= \mathbb{E}_{(s,b)\sim\mathcal{P}}\big[(b-s)\mathbb{I}(s \leq p,\, b \geq q)\big] \\
&= \mathbb{E}_{(s,b)\sim\mathcal{P}}\big[(b-s)\mathbb{I}(s \leq \Pi_K^s(p),\, b \geq \Pi_K^b(q))\big] \\
&\quad - \mathbb{E}_{(s,b)\sim\mathcal{P}}\big[(b-s)\mathbb{I}(p < s \leq \Pi_K^s(p),\, b \geq q)\big] \\
&\quad - \mathbb{E}_{(s,b)\sim\mathcal{P}}\big[(b-s)\mathbb{I}(s \leq p,\, \Pi_K^b(q) \leq b < q)\big] \\
&\quad - \mathbb{E}_{(s,b)\sim\mathcal{P}}\big[(b-s)\mathbb{I}(p < s \leq \Pi_K^s(p),\, \Pi_K^b(q) \leq b < q)\big].
\end{aligned}$$

Using $|b-s| \le 1$ and σ-smoothness, we obtain

$$\begin{aligned}\mathbb{E}_{(s,b)\sim\mathcal{P}}[GFT(p,q,s,b)] &\le \mathbb{E}\big[(b-s)\mathbb{I}(s \le \Pi_K^s(p),\, b \ge \Pi_K^b(q))\big] \\ &\quad + \mathbb{P}(p < s \le \Pi_K^s(p),\, \Pi_K^b(q) \le b) + \mathbb{P}(s \le p,\, \Pi_K^b(q) \le b < q) \\ &\le \mathbb{E}_{(s,b)\sim\mathcal{P}}[GFT(\Pi_K^s(p), \Pi_K^b(q), s, b)] + \frac{2\sigma}{K}.\end{aligned}$$

Given any feasible distribution $\gamma \in \Delta([0,1]^2)$, define the discretized distribution $\hat{\gamma}_K$ by

$$\hat{\gamma}_K(p,q) = \mathbb{P}_{(x,y)\sim\gamma}\left((\Pi_K^s(x), \Pi_K^b(y)) = (p,q)\right) \quad \forall (p,q) \in \mathcal{G}_K$$

Then,

$$\begin{aligned}\mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{P}}[GFT(p,q,s,b)] &\le \mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{P}}\left[GFT(\Pi_K^s(p), \Pi_K^b(q), s, b) + \frac{2\sigma}{K}\right] \\ &\le \mathbb{E}_{(p,q)\sim\hat{\gamma}_K,(s,b)\sim\mathcal{P}}[GFT(p,q,s,b)] + \frac{2\sigma}{K}.\end{aligned}$$

A similar argument yields

$$\begin{aligned}\mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)] &\le \mathbb{E}_{(s,b)\sim\mathcal{P}}[\left((\Pi_K^s(p) - p) + (q - \Pi_K^b(q))\right)\mathbb{I}(s \le p, b \ge q)] \\ &\quad + \mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(\Pi_K^s(p), \Pi_K^b(q), s, b)] \\ &\quad + \mathbb{P}_{(s,b)\sim\mathcal{P}}(p < s \le \Pi_K^s(p),\, \Pi_K^b(q) \le b) \\ &\quad + \mathbb{P}_{(s,b)\sim\mathcal{P}}(s \le p,\, \Pi_K^b(q) \le b < q) \\ &\le \mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(\Pi_K^s(p), \Pi_K^b(q), s, b)] + \frac{2\sigma+2}{K},\end{aligned}$$

and therefore

$$\mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)] \le \mathbb{E}_{(p,q)\sim\hat{\gamma}_K,(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)] + \mathcal{O}\left(\frac{\sigma}{K}\right).$$

Then, if we consider $\overline{\mathcal{L}}$ to be the average over all rounds $[T]$ of the distributions $\{\mathcal{L}_t\}_t$, so that

$$\|\overline{\mathcal{L}} - \mathcal{P}\|_{\mathrm{TV}} \le \frac{C}{T}$$

for all γ

$$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\gamma,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \le \sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\hat{\gamma}_K,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] + \mathcal{O}\left(\frac{T\sigma}{K} + C\right).$$

The same transfer from $\mathcal{P}$ to $\overline{\mathcal{L}}$ gives the claimed GFT bound, with an additional error at most $4C$, since $|\mathsf{GFT}| \le 1$. Finally, feasibility of γ gives:

$$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\hat{\gamma}_K,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \ge -\mathcal{O}\left(\frac{T\sigma}{K} + C\right).$$

□

Lemma 1. *With probability at least* $1 - 1/T$*, it holds:* $R_{0,K} = \mathsf{Opt}^{\mathrm{F}} - \mathsf{Opt}_K^{\mathrm{F}} \le \widetilde{\mathcal{O}}\left(\frac{T}{K} + \sqrt{T\log(TK)}\right)$.

Proof. Let p^* attain $\mathsf{Opt}^{\mathrm{F}}$, and round it outward to $p^\dagger = \Pi_K^s(p^*)$ and $q^\dagger = \Pi_K^b(p^*)$. Then $0 \le p^\dagger - q^\dagger \le 1/K$, so this pair is feasible for Program 3. Every trade at p^* is preserved; any additional trade with negative GFT loses at most $1/K$. Thus, pathwise,

$$\sum_{t=1}^{T} \mathsf{GFT}_t(p^\dagger, q^\dagger) \ge \mathsf{Opt}^{\mathrm{F}} - T/K.$$

Hoeffding's inequality and a union bound over the $(K+1)^2$ grid pairs give, with probability at least $1 - 1/T$,

$$\max_{(p,q)\in\mathcal{G}_K} \left| \sum_{t=1}^{T} \left(\mathsf{GFT}_t(p,q) - \mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{GFT}(p,q,s,b)] \right) \right| \leq \sqrt{2T\log(2(K+1)^2T)}.$$

This uniform event also holds at the data-dependent pair $(p^\dagger, q^\dagger)$. Its expected GFT is at most $\mathsf{Opt}_K^{\mathrm{F}}$, proving the claim. □

Lemma 2. *If $\mathcal{P}$ is σ-smooth, it holds: $R_K = \mathsf{Opt}^{\mathrm{D}} - \mathsf{Opt}_K^{\mathrm{D}} \leq \tilde{\mathcal{O}}\left(\frac{T}{K} + C\right)$, where the notation $\mathcal{O}(\cdot)$ contains the linear dependence on σ, which we assume constant.*

Proof Sketch. In general the proof is identical to Lemma 5.2 of Lunghi et al. (2026), where in this case we distinguish two scenarios.

- There exist $\phi \in \Delta(\mathcal{G}_K)$ such that

  $$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\phi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq \frac{2\sigma T}{K} + 2C.$$

  In this case we can build a well-define mixture distribution between $\hat{\gamma}_k$ defined by Lemma 7 and ϕ, that satisfies the claim.

- It does *not* exist $\phi \in \Delta(\mathcal{G}_K)$ such that

  $$\sum_{t=1}^{T} \mathbb{E}_{(p,q)\sim\phi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq \frac{2\sigma T}{K} + 2C.$$

  In this case, using the fact that maximum revenue achievable and the optimum achievable GFT are linked up to a $\mathcal{O}(\log(T))$ multiplicative factor, any feasible distribution cannot be worse than the optimum of $\tilde{\mathcal{O}}(\frac{T}{K} + C)$. Indeed, by Bernasconi et al. (2024b)

  $$\begin{aligned} OPT^D &\leq 2OPT^F \\ &\leq 2\left(8\log(T)\left(\max_{(p,q)\in\mathcal{G}_K}\sum_t \mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathrm{Rev}(p,q,s,b)] + \mathcal{O}\left(\frac{T\sigma}{K}\right)\right) + \frac{5T}{K}\right), \end{aligned}$$

  where we used the fact the revenue of any point in the continuum $(p,q) \in [0,1]^2$ such that $q \geq p$ can be approximated by a point on our uniform grid G_K, losing at most a factor of $\mathcal{O}(\frac{\sigma}{K})$ under the bounded density assumption.

  Therefore if it does not exist $\phi \in \Delta(G_K)$ such that $\sum_t \mathbb{E}_{(p,q)\sim\phi,(s,b)\sim\mathcal{L}_t}[\mathrm{Rev}(p,q,s,b)] \geq \frac{2\sigma T}{K} + 2C$,

  this implies that

  $$\max_{(p,q)\in G_K}\sum_t \mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathrm{Rev}(p,q,s,b)] \leq \frac{2\sigma T}{K} + 2C.$$

  Therefore, we can conclude:

  $$OPT^D \leq \tilde{\mathcal{O}}\left(\frac{\sigma T}{K} + C\right),$$

  and thus, $R_K \leq \tilde{\mathcal{O}}\left(\frac{\sigma T}{K} + C\right)$.

□

C OMITTED PROOFS FROM SECTION 6

Lemma 4. *Algorithm 1 guarantees:*

$$\sum_{t\in\mathcal{T}_2}\lambda_t\textsf{Rev}_t(p_t,q_t)\le M\left(-\sum_{t\in\mathcal{T}_1}\textsf{Rev}_t(p_t,q_t)+2\right)+2R_T^D$$

Proof. The proof of this lemma uses the alternating strategy of algorithm 1 between procedures and the weakly adaptive nature of the dual algorithm.

In particular, let τ be the last round in which Rev-Max is activated, i.e. $\tau=\max\{t:t\in\mathcal{T}_1\}$. Since $\tau\in\mathcal{T}_1$ this means that $B_\tau\le B_{\tau-1}+1\le 2$, while $B_\tau\ge 0$ by budget balance. The lower bound $B_\tau\ge 1$ need not hold when $\tau=T$. These considerations imply the following:

$$\sum_{t\in\mathcal{T}_2\cap[\tau]}\textsf{Rev}_t(p_t,q_t)+\sum_{t\in\mathcal{T}_1\cap[\tau]}\textsf{Rev}_t(p_t,q_t)\in[0,2]$$

and

$$\sum_{t\in\mathcal{T}_1\cap[\tau]}\textsf{Rev}_t(p_t,q_t)=\sum_{t\in\mathcal{T}_1}\textsf{Rev}_t(p_t,q_t).$$

Therefore, we can use the dual regret guarantees in the following way:

$$\sum_{t\in\mathcal{T}_2\cap[\tau]}\lambda_t\textsf{Rev}_t(p_t,q_t)\le M\left(\sum_{t\in\mathcal{T}_2\cap[\tau]}\textsf{Rev}_t(p_t,q_t)\right)+R_T^D$$

and

$$\sum_{t\in\mathcal{T}_2\cap[\tau+1,T]}\lambda_t\textsf{Rev}_t(p_t,q_t)\le 0\cdot\left(\sum_{t\in\mathcal{T}_2\cap[\tau+1,T]}\textsf{Rev}_t(p_t,q_t)\right)+R_T^D=R_T^D.$$

To conclude,

$$\begin{aligned}\sum_{t\in\mathcal{T}_2}\lambda_t\textsf{Rev}_t(p_t,q_t)&\le\sum_{t\in\mathcal{T}_2\cap[\tau]}\lambda_t\textsf{Rev}_t(p_t,q_t)+\sum_{t\in\mathcal{T}_2\cap[\tau+1,T]}\lambda_t\textsf{Rev}_t(p_t,q_t)\\&\le M\sum_{t\in\mathcal{T}_2\cap[\tau]}\textsf{Rev}_t(p_t,q_t)+2R_T^D\\&\le -M\sum_{t\in\mathcal{T}_1}\textsf{Rev}_t(p_t,q_t)+2M+2R_T^D.\end{aligned}$$

□

Lemma 5. *Let π be a solution to Program 3 , then with probability at least $1-1/T$:*

$$-\sum_{t\in\mathcal{T}_2}\lambda_t\mathbb{E}_{(p,q)\sim\pi}[\textsf{Rev}_t(p,q)]\le M\cdot\left(\frac{T}{K}\right)+\tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

Moreover, let π be a solution to Program 4, then with probability at least $1-1/T$:

$$-\sum_{t\in\mathcal{T}_2}\lambda_t\mathbb{E}_{(p,q)\sim\pi}[\textsf{Rev}_t(p,q)]\le 2M\cdot C+\tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

Proof. We prove the more general statement for predictable $w_t\in[0,M]$, taking $w_t=\lambda_t\mathbb{I}(t\in\mathcal{T}_2)$ for the displayed bounds. Define $\overline{\textsf{Rev}}(p,q)=\mathbb{E}_{(s,b)\sim\mathcal{P}}[\textsf{Rev}(p,q,s,b)]$. We split the proof into two cases.

Case 1: π Solution of Opt_K^F If π is a solution Opt_K^F, then by definition $\mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] \geq \frac{-1}{K}$, for all $t \in [T]$. Therefore, by applying Azuma-Hoeffding inequality, with probability at least $1 - 1/T$

$$\begin{aligned}\sum_{t=1}^{T} w_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] &= \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]\right) \\ &\quad + \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi,(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]\right) \\ &\geq -M \cdot \tilde{\mathcal{O}}\left(\sqrt{T}\right) - MT\frac{1}{K}.\end{aligned}$$

Case 2: π Solution of Opt_K^D

$$\begin{aligned}&\sum_{t=1}^{T} w_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] \\ &\quad= \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)] + \mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)]\right) \\ &\quad= \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)]\right) + \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)]\right) \\ &\quad= \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi}[\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]]\right) \\ &\quad+ \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]] - \mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)]\right) \\ &\quad+ \sum_{t=1}^{T} w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\overline{\mathsf{Rev}}(p,q)]\right)\end{aligned}$$

Additionally, by definition of C

$$\sum_{t=1}^{T} \left| \mathbb{E}_{(p,q)\sim\pi}\left[\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)] - \mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)]\right] \right| \leq C$$

and by definition of feasibility of π

$$\sum_{t\in[T]} \left(\mathbb{E}_{(p,q)\sim\pi}\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]\right) \geq 0.$$

We can use C also to bound the following quantity

$$\left| \sum_{t\in[T]} \left(\mathbb{E}_{(p,q)\sim\pi}\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]\right) - \sum_{t\in[T]} \left(\mathbb{E}_{(p,q)\sim\pi}\mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)]\right) \right| \leq C,$$

which means that

$$\left(\mathbb{E}_{(p,q)\sim\pi}\mathbb{E}_{(s,b)\sim\mathcal{P}}[\mathsf{Rev}(p,q,s,b)]\right) \geq \frac{-C}{T}.$$

Since w_t is $\mathcal{F}_{t-1}$-measurable and (s_t, b_t) is drawn independently from $\mathcal{L}_t$, the centered revenue term has conditional mean zero.

$$\begin{aligned}&\mathbb{E}[w_t \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi}[\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]]\right) |\mathcal{F}_{t-1}] \\ &\quad= \mathbb{E}[w_t|\mathcal{F}_{t-1}]\mathbb{E}[\left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - \mathbb{E}_{(p,q)\sim\pi}[\mathbb{E}_{(s,b)\sim\mathcal{L}_t}[\mathsf{Rev}(p,q,s,b)]]\right) |\mathcal{F}_{t-1}].\end{aligned}$$

Then, by applying Azuma-Hoeffding with probability at least $1 - 1/T$

$$\sum_{t=1}^{T} w_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] \geq -2M \cdot C - M \cdot \tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

□

D Omitted Proof from Section 7

In this section, we provide the proof of Theorem 2. Our analysis adapts the implicit-exploration framework of Neu (2015), and shares some similarities with the literature of multi-armed bandits with side information in the definition of the estimators (e.g. Kocák et al. (2014)). In this section we provide some auxiliary results related to the primal algorithm loss estimators and their second moment. Finally, we will use these results to prove Theorem 2.

D.1 Primal Loss Estimators and Their Guarantees

As a first step, we analyze the primal loss vector and its estimators.

We employ as true primal loss $\ell_t(p,q) = \underbrace{(1 - L_t(p,q))}_{L_t^-(p,q)} + \underbrace{(1 - R_t(p,q))}_{R_t^-(p,q)} + (\lambda_t + 1)\underbrace{(1 - \mathsf{Rev}_t(p,q))}_{\mathsf{Rev}_t^-(p,q)}$.
This choice has the benefit of guaranteeing that each of its component is greater or equal to zero and the introduced bias with respect to the Lagrangian problem does not affect the general regret.

For each component, we can define two estimators, one unbiased and based on importance sampling principle, indicated by a $\hat{\cdot}$, and one identical but for the addition on the denominator of the implicit exploration parameter γ, indicated by $\tilde{\cdot}$.

More formally

$$\tilde{L}_t^-(p,q) = \frac{1 - \mathbb{I}(s_t \leq U_t \leq p, b_t \geq q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma}\mathbb{I}(q = \hat{q}_t, H_t = 1)$$

$$\hat{L}_t^-(p,q) = \frac{1 - \mathbb{I}(s_t \leq U_t \leq p, b_t \geq q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)}\mathbb{I}(q = \hat{q}_t, H_t = 1)$$

$$\tilde{R}_t^-(p,q) = \frac{1 - \mathbb{I}(s_t \leq p, b_t \geq V_t \geq q)}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}} \hat{\pi}_t(p,q') + \gamma}\mathbb{I}(p = p_t, H_t = 2)$$

$$\hat{R}_t^-(p,q) = \frac{1 - \mathbb{I}(s_t \leq p, b_t \geq V_t \geq q)}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}} \hat{\pi}_t(p,q')}\mathbb{I}(p = p_t, H_t = 2)$$

$$\widetilde{\mathsf{Rev}}_t^-(p,q) = \frac{1 - (q_t - p_t)\mathbb{I}(s_t \leq \hat{p}_t, b_t \geq \hat{q}_t)}{(1-\alpha)\hat{\pi}_t(p,q) + \gamma}\mathbb{I}((p,q) = (\hat{p}_t, \hat{q}_t), H_t = 0)$$

$$\widehat{\mathsf{Rev}}_t^-(p,q) = \frac{1 - (q_t - p_t)\mathbb{I}(s_t \leq \hat{p}_t, b_t \geq \hat{q}_t)}{(1-\alpha)\hat{\pi}_t(p,q)}\mathbb{I}((p,q) = (\hat{p}_t, \hat{q}_t), H_t = 0).$$

$$\hat{\ell}_t(p,q) = \hat{L}_t^-(p,q) + \hat{R}_t^-(p,q) + (\lambda_t + 1)\widehat{\mathsf{Rev}}_t^-(p,q)$$

and finally

$$\tilde{\ell}_t(p,q) = \tilde{L}_t^-(p,q) + \tilde{R}_t^-(p,q) + (\lambda_t + 1)\widetilde{\mathsf{Rev}}_t^-(p,q).$$

First, we establish that the estimator $\hat{\ell}_t(p,q)$ is unbiased with respect to our Lagrangian objective $\ell_t(p,q)$.

Lemma 8 (Unbiasedness of $\hat{\ell}_t$)**.** *Let $\hat{\ell}_t(p,q)$ be defined as:*

$$\hat{\ell}_t(p,q) = \hat{L}_t(p,q)^- + \hat{R}_t(p,q)^- + (\lambda_t + 1)\widehat{\mathsf{Rev}}_t(p,q)^-.$$

Then, $\mathbb{E}[\hat{\ell}_t(p,q) \mid \mathcal{F}_{t-1}, s_t, b_t] = \ell_t(p,q)$.

Proof. All expectations in this proof are conditional on the history $\mathcal{F}_{t-1}$ and the current valuations (s_t, b_t); thus λ_t, $\hat{\pi}_t$, and ℓ_t are fixed. By the law of total expectation, $\mathbb{E}[\hat{\ell}_t(p,q)] = \sum_{k=0}^{2} \mathbb{P}(H_t = k)\mathbb{E}[\hat{\ell}_t(p,q) \mid H_t = k]$. We analyze the components:

Case $H_t = 0$ **(Bandit Component):**

$$\begin{aligned}\mathbb{P}(H_t = 0)\mathbb{E}[\hat{\ell}_t \mid H_t = 0] &= (1-\alpha)\mathbb{E}\left[\frac{(\lambda_t+1)(1-\mathsf{Rev}_t(\hat{p}_t,\hat{q}_t))}{(1-\alpha)\hat{\pi}_t(p,q)}\mathbb{I}((p,q)=(\hat{p}_t,\hat{q}_t))\right] \\ &= (\lambda_t+1)(1-\mathbb{E}[\mathsf{Rev}_t(p,q)]).\end{aligned}$$

Case $H_t = 1, 2$ **(Marginal Components):** For $H_t = 1$ (exploring the seller side while fixing the buyer's price $\hat{q}_t$):

$$\begin{aligned}\mathbb{P}(H_t = 1)\mathbb{E}[\hat{\ell}_t \mid H_t = 1] &= \frac{\alpha}{2}\mathbb{E}\left[\frac{1-\mathbb{I}(s_t \le U_t \le p, b_t \ge q)}{\frac{\alpha}{2}\sum_{p'}\hat{\pi}_t(p',q)}\mathbb{I}(q=\hat{q}_t)\right] \\ &= (1-\mathbb{E}[L_t(p,q)]).\end{aligned}$$

Similarly, for $H_t = 2$ (exploring the buyer side):

$$\mathbb{P}(H_t = 2)\mathbb{E}[\hat{\ell}_t \mid H_t = 2] = (1-\mathbb{E}[R_t(p,q)]).$$

Summing these

$$\mathbb{E}[\hat{\ell}_t(p,q)] = (\lambda_t+1)(1-\mathsf{Rev}_t(p,q)) + (1-L_t(p,q)) + (1-R_t(p,q)) = \ell_t(p,q).$$

□

Let us define these new quantities :

$$\tilde{L}_t^-(q) = \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}}\hat{\pi}_t(p',q)+\gamma}\mathbb{I}(q=\hat{q}_t, H_t=1)$$

so that $\tilde{L}_t^-(p,q) \le \tilde{L}_t^-(q)$, and similarly,

$$\tilde{R}_t^-(p) = \frac{1}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}}\hat{\pi}_t(p,q')+\gamma}\mathbb{I}(p=\hat{p}_t, H_t=2)$$

so that $\tilde{R}_t^-(p,q) \le \tilde{R}_t^-(p)$.

Define $L_t^-(q) = R_t^-(p) = 1$, and let $\hat{L}_t^-(q), \hat{R}_t^-(p)$ be the same envelopes with $\gamma = 0$. These quantities give the following result.

Lemma 9.

$$\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\left(\hat{\ell}_t(p,q)-\tilde{\ell}_t(p,q)\right) \le \gamma\left(\frac{2}{\alpha}\sum_{q\in\mathcal{A}_b}\widetilde{L}_t^-(q) + \frac{2}{\alpha}\sum_{p\in\mathcal{A}_s}\widetilde{R}_t^-(p) + \frac{(1+\lambda_t)}{1-\alpha}\sum_{(p,q)\in\mathcal{A}}\widetilde{\mathsf{Rev}}_t^-(p,q)\right)$$

Proof. We prove the desired inequality by focusing on each component of $\hat{\ell}$ and $\tilde{\ell}$.

Therefore,

$$\begin{aligned}&\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\left(\hat{L}_t^-(p,q)-\tilde{L}_t^-(p,q)\right) \\ &\le \sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\cdot \\ &\quad\cdot\left(\frac{1-\mathbb{I}(s_t\le U_t\le p, b_t\ge q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}}\hat{\pi}_t(p',q)}\mathbb{I}(q=\hat{q}_t,H_t=1) - \frac{1-\mathbb{I}(s_t\le U_t\le p, b_t\ge q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}}\hat{\pi}_t(p',q)+\gamma}\mathbb{I}(q=\hat{q}_t,H_t=1)\right) \\ &\le \sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\left(1-\mathbb{I}(s_t\le U_t\le p, b_t\ge q)\right)\mathbb{I}(q=\hat{q}_t,H_t=1)\cdot\end{aligned}$$

$$\cdot \underbrace{\left(\frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)} - \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \right)}_{(\star)}$$

and $(\star)$ can be further bounded as

$$\begin{aligned}
(\star) &= \left(\frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)} - \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \right) \\
&= \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)} \left(\frac{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma - \frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \right) \\
&\leq \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)} \left(\frac{\gamma}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \right)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\hat{L}_t^-(p,q) - \tilde{L}_t^-(p,q) \right) \\
&\leq \gamma \sum_{(p,q)\in\mathcal{A}} \frac{\hat{\pi}_t(p,q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q)} \tilde{L}_t^-(p,q) \leq \gamma \frac{2}{\alpha} \sum_{q\in\mathcal{A}_b} \tilde{L}_t^-(q).
\end{aligned}$$

With analogous analysis

$$\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\hat{R}_t^-(p,q) - \tilde{R}_t^-(p,q) \right) \leq \gamma \frac{2}{\alpha} \sum_{p\in\mathcal{A}_s} \tilde{R}_t^-(p),$$

$$\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\widehat{\mathsf{Rev}}_t^-(p,q) - \widetilde{\mathsf{Rev}}_t^-(p,q) \right) \leq \gamma \frac{1}{1-\alpha} \sum_{(p,q)\in\mathcal{A}} \widetilde{\mathsf{Rev}}_t^-(p,q).$$

Finally, recalling the definition of $\hat{\ell}_t(p,q)$ and $\tilde{\ell}_t(p,q)$ we conclude the proof. □

Then, we can use Lemma 1 from Neu (2015), on each component, obtaining the following result.

Lemma 10. *Let $\gamma > 0$ and let $\alpha_t^1(q), \alpha_t^2(p), \alpha_t^3(p,q)$ be predictable nonnegative weights, each at most 2γ. Then, with probability at least $1 - 1/T$, for all $\tau \in [T]$*

$$\sum_{t=1}^{\tau} \sum_{q\in\mathcal{A}_b} \alpha_t^1(q)(\tilde{L}_t^-(q) - L_t^-(q)) \leq \log(3T^2),$$

$$\sum_{t=1}^{\tau} \sum_{p\in\mathcal{A}_s} \alpha_t^2(p)(\tilde{R}_t^-(p) - R_t^-(p)) \leq \log(3T^2),$$

$$\sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \alpha_t^3(p,q)(\widetilde{\mathsf{Rev}}_t^-(p,q) - \mathsf{Rev}_t^-(p,q)) \leq 2\log(3T^2).$$

Proof. The envelopes estimate the constant losses $L_t^-(q) = R_t^-(p) = 1$, while $\mathsf{Rev}_t^-(p,q)/2 \in [0,1]$. Apply Lemma 1 of Neu (2015) to each observation channel (adding a dummy outcome when that channel is inactive), normalize the revenue loss by two, and take a union bound over the three channels and all $\tau \in [T]$. □

In addition we give the following result.

Lemma 11. *Fix $(p,q) \in \mathcal{A}$. Let $\gamma > 0$, let $0 \le \lambda_t \le M$ be predictable, and let $\alpha_t \in [0, 2\gamma]$ be predictable. With probability at least $1 - 1/T$, simultaneously for all $\tau \in [T]$,*

$$\sum_{t=1}^{\tau} \alpha_t \big(\tilde{\ell}_t(p,q) - \ell_t(p,q)\big) \le (4 + 2M) \log(3T^2).$$

Proof. Apply the implicit-exploration argument of Lemma 1 of Neu (2015) separately to the three observation events for the fixed pair. The seller and buyer numerators lie in $[0,1]$ and have conditional means $L_t^-(p,q)$ and $R_t^-(p,q)$. Normalize the revenue numerator by two and use weights $\alpha_t(1 + \lambda_t)/(M+1) \le 2\gamma$. The three resulting bounds are $\log(3T^2)$, $\log(3T^2)$, and $2(M+1)\log(3T^2)$, after a union bound over the channels and prefixes. □

In addition, similar to Neu (2015) we have the following Corollary.

Corollary 1. *For all $(p,q) \in \mathcal{A}$ and $\tau \in [T]$, it holds simultaneously with probability at least $1 - 1/T$:*

$$\sum_{t=1}^{\tau} \Big(\widetilde{\ell}_t(p,q) - \ell_t(p,q)\Big) \le \frac{(4 + 2M) \ln\big(3|\mathcal{A}|T^2\big)}{2\gamma},$$

where $M \ge \max_{t\in[T]} |\lambda_t|$.

D.2 Bounding the Second Moment of the Estimators

The regret of EXP3.IX depends on the second moment of the estimators $\{\tilde{\ell}_t\}_t$, so in this section we will do exactly that.

Let's recall the definition of $\tilde{L}_t^-(q), \tilde{R}_t^-(p)$ as

$$\tilde{L}_t^-(q) = \frac{1}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \mathbb{I}(q = \hat{q}_t, H_t = 1)$$

and

$$\tilde{R}_t^-(p) = \frac{1}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}} \hat{\pi}_t(p,q') + \gamma} \mathbb{I}(p = \hat{p}_t, H_t = 2).$$

Lemma 12. *For all $t \in [T]$, the second moment of the biased estimators satisfies:*

$$\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \Big(\tilde{\ell}_t(p,q)\Big)^2 \le \left(\frac{6}{\alpha}\sum_{q\in\mathcal{A}_b} \tilde{L}_t^-(q) + \frac{6}{\alpha}\sum_{p\in\mathcal{A}_s} \tilde{R}_t^-(p) + \frac{6(\lambda_t+1)^2}{1-\alpha}\sum_{(p,q)\in\mathcal{A}} \widetilde{\mathsf{Rev}}_t^-(p,q)\right).$$

Proof. Applying the Cauchy-Schwarz inequality, $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we bound the squared terms of the estimator. For the revenue term, use $(\mathsf{Rev}_t^-)^2 \le 2\mathsf{Rev}_t^-$; the other numerators are at most one.

$$\begin{aligned}
\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \Big(\tilde{\ell}_t(p,q)\Big)^2 \le{}& 3\sum_{q\in\mathcal{A}_b}\sum_{p:(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\frac{1 - \mathbb{I}(s_t \le U_t \le p, b_t \ge q)}{\frac{\alpha}{2}\sum_{p':(p',q)\in\mathcal{A}} \hat{\pi}_t(p',q) + \gamma} \mathbb{I}(q = \hat{q}_t, H_t = 1)\right)^2 \\
&+ 3\sum_{p\in\mathcal{A}_s}\sum_{q:(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\frac{1 - \mathbb{I}(s_t \le p, b_t \ge V_t \ge q)}{\frac{\alpha}{2}\sum_{q':(p,q')\in\mathcal{A}} \hat{\pi}_t(p,q') + \gamma} \mathbb{I}(p = \hat{p}_t, H_t = 2)\right)^2 \\
&+ 3(\lambda_t+1)^2 \sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q) \left(\frac{1 - (q_t - p_t)\mathbb{I}(s_t \le \hat{p}_t, b_t \ge \hat{q}_t)}{(1-\alpha)\hat{\pi}_t(p,q) + \gamma} \mathbb{I}((p,q) = (\hat{p}_t, \hat{q}_t), H_t = 0)\right)^2 \\
\le{}& \frac{6}{\alpha}\sum_{q\in\mathcal{A}_b} \widetilde{L}_t^-(q) + \frac{6}{\alpha}\sum_{p\in\mathcal{A}_s} \widetilde{R}_t^-(p) + \frac{6(\lambda_t+1)^2}{1-\alpha}\sum_{(p,q)\in\mathcal{A}} \widetilde{\mathsf{Rev}}_t^-(p,q).
\end{aligned}$$

□

D.3 Proof of Theorem 2

We are ready to prove Theorem 2.

Theorem 2. *Let $|\mathcal{A}| \geq 2$, $0 < \alpha \leq 1/2$, and let $0 \leq \lambda_t \leq M$ be predictable multipliers with a deterministic bound $M \geq 1$. For the choices of η^P and γ given in the proof, Algorithm 3 ensures, simultaneously for all $\pi \in \Delta(\mathcal{A})$ and $\tau \in [T]$, with probability at least $1 - \mathcal{O}(1/T)$:*

$$\sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\left(\mathsf{GFT}_t(p,q) + \lambda_t \mathsf{Rev}_t(p,q)\right) - \sum_{t=1}^{\tau}\left(\mathsf{GFT}_t(p_t,q_t) + \lambda_t \mathsf{Rev}_t(p_t,q_t)\right) \leq \tilde{\mathcal{O}}\left(M\left(\sqrt{|\mathcal{A}|T} + \frac{|\mathcal{A}_s| + |\mathcal{A}_b|}{\sqrt{|\mathcal{A}|}}\frac{\sqrt{T}}{\alpha} + \alpha \cdot T\right)\right).$$

Proof. We use $0 < \alpha \leq 1/2$, so $(1-\alpha)^{-1} \leq 2$. Let π_t denote the actual distribution of the posted prices on $[0,1]^2$, including the continuous exploration draws. First, we observe that the regret with respect to the lagrangian function is equivalent to the regret with respect to the ℓ_t loss functions.

$$\begin{aligned}
&\sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\left(\mathsf{GFT}_t(p,q) + \lambda_t \mathsf{Rev}_t(p,q)\right) - \sum_{t=1}^{\tau}\left(\mathsf{GFT}_t(p_t,q_t) + \lambda_t \mathsf{Rev}_t(p_t,q_t)\right) \\
&\quad= \sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\left(L_t(p,q) + R_t(p,q) + (\lambda_t+1)\mathsf{Rev}_t(p,q)\right) \\
&\qquad - \sum_{t=1}^{\tau}\left(L_t(p_t,q_t) + R_t(p_t,q_t) + (\lambda_t+1)\mathsf{Rev}_t(p_t,q_t)\right) \\
&\quad= -\sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\left((1-L_t(p,q)) + (1-R_t(p,q)) + (\lambda_t+1)(1-\mathsf{Rev}_t(p,q))\right) \\
&\qquad + \sum_{t=1}^{\tau} \sum_{(p,q)\in\mathcal{A}} \pi(p,q)(3+\lambda_t) \\
&\qquad + \sum_{t=1}^{\tau}\left((1-L_t(p_t,q_t)) + (1-R_t(p_t,q_t)) + (\lambda_t+1)(1-\mathsf{Rev}_t(p_t,q_t))\right) - \sum_{t=1}^{\tau}(3+\lambda_t) \\
&\quad= \sum_{t=1}^{\tau}\left(\ell_t(p_t,q_t) - \sum_{(p,q)\in\mathcal{A}} \pi(p,q)\ell_t(p,q)\right).
\end{aligned}$$

Hence, we decompose the regret with respect to the loss functions ℓ_t in manageable parts.

For all $(p^\dagger, q^\dagger) \in \mathcal{A}$, for all $\tau \in [T]$

$$\begin{aligned}
\sum_{t=1}^{\tau}\left(\ell_t(p_t,q_t) - \ell_t(p^\dagger,q^\dagger)\right) \leq{}& \underbrace{\sum_{t=1}^{\tau}\left(\ell_t(p_t,q_t) - \mathbb{E}_{(p,q)\sim\pi_t}[\ell_t(p,q)]\right)}_{(1)} \\
&+ \underbrace{\sum_{t=1}^{\tau}\left(\mathbb{E}_{(p,q)\sim\pi_t}[\ell_t(p,q)] - \sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q)\ell_t(p,q)\right)}_{(2)} \\
&+ \underbrace{\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q)\ell_t(p,q) - \sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q)\tilde{\ell}_t(p,q)\right)}_{(3)}
\end{aligned}$$

$$+\underbrace{\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\tilde{\ell}_t(p,q)-\tilde{\ell}_t(p^\dagger,q^\dagger)\right)}_{(4)}$$

$$+\underbrace{\sum_{t=1}^{\tau}\left(\tilde{\ell}_t(p^\dagger,q^\dagger)-\ell_t(p^\dagger,q^\dagger)\right)}_{(5)}.$$

First Term The first term can be bounded by applying Azuma-Hoeffding inequality, and with probability at least $1-1/T$, simultaneously over all prefixes

$$(1)\le\tilde{\mathcal{O}}(M\sqrt{T}).$$

Second Term The second term can be bounded by definition of π_t and $\hat{\pi}_t$. Indeed,

$$\begin{aligned}(2)&=\sum_{t=1}^{\tau}\left(\mathbb{E}_{(p,q)\sim\pi_t}[\ell_t(p,q)]-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\ell_t(p,q)\right)\\&\le\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}(1-\alpha)\hat{\pi}_t(p,q)\ell_t(p,q)-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\ell_t(p,q)\right)+\tau\alpha(2M+4)\le\tau\alpha(2M+4),\end{aligned}$$

since $0\le\ell_t(p,q)\le 2M+4$ for all $t\in[T]$ and $(p,q)\in[0,1]^2$.

Third term To bound the third term we can divide this term in two, employing the unbiased loss estimators $\{\hat{\ell}_t\}$.

$$\begin{aligned}(3)=&\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\ell_t(p,q)-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\hat{\ell}_t(p,q)\right)\\&+\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\hat{\ell}_t(p,q)-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\tilde{\ell}_t(p,q)\right),\end{aligned}$$

of which the second term is bounded by Lemma 9 with probability at least $1-1/T$ as

$$\begin{aligned}&\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\hat{\ell}_t(p,q)-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\tilde{\ell}_t(p,q)\right)\\&\quad\le\sum_{t=1}^{\tau}\gamma\left(\frac{2}{\alpha}\sum_{q\in\mathcal{A}_b}\widetilde{L}_t^-(q)+\frac{2}{\alpha}\sum_{p\in\mathcal{A}_s}\widetilde{R}_t^-(p)+\frac{(1+\lambda_t)}{1-\alpha}\sum_{(p,q)\in\mathcal{A}}\widetilde{\mathsf{Rev}}_t^-(p,q)\right)\\&\quad\le\sum_{t=1}^{\tau}\gamma\left(\frac{2}{\alpha}\sum_{q\in\mathcal{A}_b}L_t^-(q)+\frac{2}{\alpha}\sum_{p\in\mathcal{A}_s}R_t^-(p)+\frac{(1+\lambda_t)}{1-\alpha}\sum_{(p,q)\in\mathcal{A}}\mathsf{Rev}_t^-(p,q)\right)\\&\qquad+\left(\frac{2}{\alpha}+\frac{M+1}{1-\alpha}\right)\log(3T^2)\\&\quad\le\tau\gamma\left(\frac{2(|\mathcal{A}_b|+|\mathcal{A}_s|)}{\alpha}+\frac{2(M+1)}{1-\alpha}|\mathcal{A}|\right)+\left(\frac{2}{\alpha}+\frac{M+1}{1-\alpha}\right)\log(3T^2),\end{aligned}$$

where the second inequality follows from Lemma 10, using weights 2γ for each marginal envelope and $2\gamma(1+\lambda_t)/(M+1)$ for revenue, then multiplying the resulting bounds by the corresponding factors. Then we can apply Azuma-Hoeffding inequality to the first term and with probability at least $1-1/T$

$$\sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\ell_t(p,q)-\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\hat{\ell}_t(p,q)\right)\le\tilde{\mathcal{O}}\left(\frac{M}{\alpha}\sqrt{T}\right),$$

because the centered term is a martingale difference by Lemma 8, and, since only one observation channel is active,

$$0 \leq \sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q)\hat{\ell}_t(p,q) \leq \frac{2}{\alpha} + \frac{2(M+1)}{1-\alpha}.$$

Fourth term The fourth term is regret term generated by EXP3.

$$\begin{aligned}
(4) &= \sum_{t=1}^{\tau}\left(\sum_{(p,q)\in\mathcal{A}} \hat{\pi}_t(p,q)\tilde{\ell}_t(p,q) - \tilde{\ell}_t(p^\dagger,q^\dagger)\right) \leq \frac{\log|\mathcal{A}|}{\eta^P} + \sum_{t=1}^{\tau}\frac{\eta^P}{2}\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\tilde{\ell}_t(p,q)^2 \\
&\leq \frac{\log|\mathcal{A}|}{\eta^P} + \sum_{t=1}^{\tau}\frac{\eta^P}{2}\left(\frac{6}{\alpha}\sum_{q\in\mathcal{A}_b}\tilde{L}_t^-(q) + \frac{6}{\alpha}\sum_{p\in\mathcal{A}_s}\tilde{R}_t^-(p) + \frac{6(\lambda_t+1)^2}{1-\alpha}\sum_{a\in\mathcal{A}}\widetilde{\mathsf{Rev}}_t^-(a)\right) \\
&\leq \frac{\log|\mathcal{A}|}{\eta^P} + \sum_{t=1}^{\tau}\frac{\eta^P}{2}\left(\frac{6}{\alpha}\sum_{q\in\mathcal{A}_b}L_t^-(q) + \frac{6}{\alpha}\sum_{p\in\mathcal{A}_s}R_t^-(p) + \frac{6(\lambda_t+1)^2}{1-\alpha}\sum_{a\in\mathcal{A}}\mathsf{Rev}_t^-(a)\right) \\
&+ \frac{\eta^P}{2\gamma}\left(\frac{6}{\alpha} + \frac{6(M+1)^2}{1-\alpha}\right)\log(3T^2) \\
&\leq \frac{\log|\mathcal{A}|}{\eta^P} + \frac{\tau\cdot\eta^P}{2}\left(\frac{6|\mathcal{A}_b|}{\alpha} + \frac{6|\mathcal{A}_s|}{\alpha} + \frac{12(M+1)^2}{1-\alpha}|\mathcal{A}|\right) \\
&+ \frac{\eta^P}{2\gamma}\left(\frac{6}{\alpha} + \frac{6(M+1)^2}{1-\alpha}\right)\log(3T^2),
\end{aligned}$$

where the second inequality holds by Lemma 12:

$$\sum_{(p,q)\in\mathcal{A}}\hat{\pi}_t(p,q)\left(\tilde{\ell}_t(p,q)\right)^2 \leq \left(\frac{6}{\alpha}\sum_{q\in\mathcal{A}_b}\tilde{L}_t^-(q) + \frac{6}{\alpha}\sum_{p\in\mathcal{A}_s}\tilde{R}_t^-(p) + \frac{6(\lambda_t+1)^2}{1-\alpha}\sum_{a\in\mathcal{A}}\widetilde{\mathsf{Rev}}_t^-(a)\right).$$

The third inequality follows from Lemma 10, using weights 2γ for each marginal envelope and $2\gamma(1+\lambda_t)^2/(M+1)^2$ for revenue.

Fifth term The fifth and last component is a direct application of Corollary 1. With probability at least $1-1/T$

$$(5) \leq \frac{(4+2M)\ln\left(3|\mathcal{A}|T^2\right)}{2\gamma}.$$

Conclusion Finally, with $\eta^P = \frac{1}{M}\sqrt{\frac{\log|\mathcal{A}|}{|\mathcal{A}|T}}$ and $\gamma = \frac{M\eta^P}{2}$, by union bound over the events, we have with probability at least $1-\mathcal{O}(1/T)$ for all $\pi\in\Delta(\mathcal{A})$, for all $\tau\in[T]$

$$\sum_{t=1}^{\tau}\left(\ell_t(p_t,q_t) - \sum_{(p,q)\in\mathcal{A}}\pi(p,q)\ell_t(p,q)\right) \leq \tilde{\mathcal{O}}\left(M\left(\sqrt{|\mathcal{A}|T} + \frac{|\mathcal{A}_s|+|\mathcal{A}_b|}{\sqrt{|\mathcal{A}|}}\frac{\sqrt{T}}{\alpha} + \alpha\cdot T\right)\right).$$

□

E Proof of Theorem 1

To prove Theorem 1 we will decompose the regret of Algorithm 1 in its components, and then we will show that there exists a choice of parameters such that Theorem 1 is verified.

Lemma 13. *With probability at least $1-\mathcal{O}(1/T)$, the regret of Algorithm 1 satisfies:*

$$\mathit{Opt}^{\mathrm{D}} - \sum_{t=1}^{T}\mathit{GFT}_t(p_t,q_t) \leq R_T^P + 2R_T^D + 2M\cdot C + \tilde{\mathcal{O}}(T^{3/4}) + R_K$$

and

$$\mathsf{Opt}^{\mathrm{F}} - \sum_{t=1}^{T} \mathsf{GFT}_t(p_t, q_t) \le R_T^P + 2R_T^D + MT/K + \tilde{\mathcal{O}}(T^{3/4}) + R_{0,K}.$$

where $M = 16\log(T)$ is the bound on the dual variables, R_T^P and R_T^D are the primal and dual regrets respectively, R_K and $R_{0,K}$ are the regret generated by the approximation.

Proof. The following argument also covers $\mathcal{T}_2 = \emptyset$, with empty sums interpreted as zero.

Let $\pi \in \Delta(\mathcal{A})$, we have with probability at least $1 - \mathcal{O}(1/T)$ Equation (7)

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_2} \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)]\right) - \sum_{t\in\mathcal{T}_2} \left(\mathsf{GFT}_t(p_t,q_t)\right) \\
&\qquad \le R_T^P - \sum_{t\in\mathcal{T}_2} \lambda_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] + \sum_{t\in\mathcal{T}_2} \lambda_t \mathsf{Rev}_t(p_t,q_t) \\
&\qquad \le R_T^P - \sum_{t\in\mathcal{T}_2} \lambda_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] - M\left(\sum_{t\in\mathcal{T}_1} \mathsf{Rev}_t(p_t,q_t)\right) + 2M + 2R_T^D,
\end{aligned}$$

where the last inequality employs Lemma 4.

For the revenue maximization rounds, Lemma 3 and the nonnegative GFT of Rev-Max give, with probability at least $1 - 1/T$,

$$\begin{aligned}
\sum_{t\in\mathcal{T}_1} \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] - \mathsf{GFT}_t(p_t,q_t)\right) \le M \sum_{t\in\mathcal{T}_1} \mathsf{Rev}_t(p_t,q_t) \\
- \sum_{t\in\mathcal{T}_1} \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] + \tilde{\mathcal{O}}(T^{3/4}).
\end{aligned}$$

Set $w_t = \mathbb{I}(t \in \mathcal{T}_1) + \lambda_t \mathbb{I}(t \in \mathcal{T}_2)$. These weights are predictable and belong to $[0, M]$. Adding the two phase bounds cancels the learner's revenue terms and yields

$$\sum_{t=1}^{T} \left(\mathbb{E}_{(p,q)\sim\pi}[\mathsf{GFT}_t(p,q)] - \mathsf{GFT}_t(p_t,q_t)\right) \le R_T^P + 2R_T^D - \sum_{t=1}^{T} w_t \mathbb{E}_{(p,q)\sim\pi}[\mathsf{Rev}_t(p,q)] + \tilde{\mathcal{O}}(T^{3/4}).$$

Choose π to solve Program 4 or Program 3. Both optimizers depend only on the obliviously chosen distributions. Thus, for each optimizer, Hoeffding's inequality changes its cumulative realized GFT to its expected value at a cost $\tilde{\mathcal{O}}(\sqrt{T})$ with probability at least $1 - 1/T$. Apply the general version of Lemma 5 to w_t, and then the appropriate grid approximation lemma. A union bound gives

$$\mathsf{Opt}^{\mathrm{D}} - \sum_{t=1}^{T} \mathsf{GFT}_t(p_t, q_t) \le R_T^P + 2R_T^D + 2MC + \tilde{\mathcal{O}}(T^{3/4}) + R_K,$$

$$\mathsf{Opt}^{\mathrm{F}} - \sum_{t=1}^{T} \mathsf{GFT}_t(p_t, q_t) \le R_T^P + 2R_T^D + MT/K + \tilde{\mathcal{O}}(T^{3/4}) + R_{0,K}.$$

□

To conclude the proof of Theorem 1, it is sufficient to recall that with probability at least $1 - \mathcal{O}(1/T)$

- $R_T^P = \tilde{\mathcal{O}}\left(M\left(\alpha T + (K+1)\sqrt{T} + \frac{\sqrt{T}}{\alpha}\right)\right)$
- $R_T^D = \mathcal{O}(M^2\sqrt{T}) = \tilde{\mathcal{O}}(\sqrt{T})$
- $M = \mathcal{O}(\log(T))$
- $R_K = \tilde{\mathcal{O}}(\frac{T}{K} + C)$ if the unperturbed distribution $\mathcal{P}$ is σ-smooth
- $R_{0,K} = \tilde{\mathcal{O}}(\frac{T}{K} + \sqrt{T})$.

Hence, by setting $K = \lceil T^{1/4} \rceil$ and $\alpha = \min\{1/2, T^{-1/4}\}$, we get the desired results.