Journal of Machine Learning Research 25 (2024) 1-36 Submitted 12/23; Published 7/24

Regret Analysis of Bilateral Trade with a Smoothed Adversary

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano and Politecnico di Milano,
Milan, Italy

Tommaso Cesari TCESARI@UOTTAWA.CA
University of Ottawa,
Ottawa, Canada

Roberto Colomboni ROBERTO.COLOMBONI@UNIMI.IT
Università degli Studi di Milano and Politecnico di Milano,
Milan, Italy

Federico Fusco FUSCOF@DIAG.UNIROMA1.IT
Sapienza Università di Roma,
Rome, Italy

Stefano Leonardi LEONARDI@DIAG.IT
Sapienza Università di Roma,
Rome, Italy

Editor: Shipra Agrawal

Abstract

We study repeated bilateral trade where an adaptive σ-smooth adversary generates the valuations of sellers and buyers. We completely characterize the regret regimes for fixed-price mechanisms under different feedback models in the two cases where the learner can post the same or different prices to buyers and sellers. We begin by showing that, in the full-feedback scenario, the minimax regret after T rounds is of order $\sqrt{T}$. Under partial feedback, any algorithm that has to post the same price to buyers and sellers suffers worst-case linear regret. However, when the learner can post two different prices at each round, we design an algorithm enjoying regret of order $T^{3/4}$, ignoring log factors. We prove that this rate is optimal by presenting a surprising $T^{3/4}$ lower bound, which is the paper's main technical contribution.

Keywords: two-sided markets, online learning, regret minimization, smoothed analysis

1. Introduction

In the bilateral trade problem, two strategic agents—a seller and a buyer—wish to trade some good. They both privately hold a personal valuation for it and strive to maximize their quasi-linear utility. The solution to the problem consists of designing a mechanism that intermediates between the two parties to make the trade happen. In general, an ideal mechanism for the bilateral trade problem would optimize the efficiency, i.e., the gain in social welfare resulting from trading the item from seller to buyer, while enforcing incentive compatibility (IC) and individual rationality (IR). The assumption that makes a two-sided mechanism design more complex than its one-sided counterpart is budget balance (BB): the mechanism cannot subsidize the market. Unfortunately, as Vickrey (1961) observed in his seminal work, the optimal incentive compatible mechanism maximizing social welfare for bilateral trade may not be budget balanced. A more general result due to Myerson and Satterthwaite (1983) shows some problem instances where a fully efficient mechanism for bilateral trade that satisfies IC, IR, and BB does not exist. This impossibility result holds even if prior information on the buyer and seller's valuations is available and the truthful notion is relaxed to Bayesian incentive compatibility. To circumvent this obstacle, the subsequent vast body of work primarily aims at *approximately* maximize expected efficiency (where the expectation is with respect to the valuations' randomness). There are many incentive compatible, individually rational, and budget balanced mechanisms that give a constant approximation to the social welfare (see, e.g., Blumrosen and Dobzinski, 2014; Dütting et al., 2021), and more recently to the more challenging problem of approximating the gain from trade (Deng et al., 2022). Although in some sense necessary—without any information on the priors there is no way to extract any meaningful approximation to the social welfare (Dütting et al., 2021)—the Bayesian assumption of perfect knowledge of the valuations' underlying distributions is unrealistic.

Following recent work (Cesa-Bianchi et al., 2021; Azar et al., 2022; Cesa-Bianchi et al., 2024), we study this fundamental mechanism design problem in an online learning setting where at each time t, a new seller/buyer pair arrives. The seller has a private valuation $s_t \in [0,1]$ representing the smallest price they are willing to accept in order to trade. Similarly, the buyer has a private value $b_t \in [0,1]$ representing the highest price they would pay for the item. We assume an adversary generates both valuations. Independently, the learner posts two (possibly randomized) prices: $p_t \in [0,1]$ to the seller and $q_t \in [0,1]$ to the buyer. We require budget balance: it must hold that $p_t \le q_t$ for all t or, equivalently, that the pair (p_t, q_t) belongs to the upper triangle $\mathcal{U} = \big\{(x,y) \in [0,1]^2 \mid x \le y\big\}$. A trade happens if and only if both agents agree to trade, i.e., when $s_t \le p_t$ and $q_t \le b_t$. When this is the case, the learner observes some feedback z_t and is awarded the gain from trade, which measures the increase in the agents' social welfare at the end of each time t:

$$\mathrm{GFT}_t(p,q) = \Big(\underbrace{(b_t - q)}_{\text{buyer's utility}} + \underbrace{(p - s_t)}_{\text{seller's utility}} \Big) \cdot \mathbb{I}\{s_t \le p \le q \le b_t\}^*.$$

When the two prices p and q are equal, we omit one of the arguments to simplify the notation. When we want to stress the dependence on the valuations, we use the notation $\mathrm{GFT}(p,q,s_t,b_t)$ instead of $\mathrm{GFT}_t(p,q)$; moreover, we omit the dependence in t when we refer to the generic gain from trade function. We consider the following learning protocol (σ-smoothness is formally defined below).

Learning protocol for sequential bilateral trade against a σ-smooth adversary

```
for time $t = 1, 2, \dots$ do
    The adversary privately chooses the $\sigma$-smooth distribution of a r.v. $(S_t, B_t)$ on $[0,1]^2$
    Seller and buyer valuations $(s_t, b_t)$ are drawn from $(S_t, B_t)$
    The learner posts prices $(p_t, q_t) \in \mathcal{U}$
    The learner receives a (hidden) reward $\mathrm{GFT}_t(p_t, q_t) \in [0,1]$
    Feedback $z_t$ is revealed to the learner
```

[*] Other works consider the similar definition $(b_t - s_t) \cdot \mathbb{I}\{s_t \le p \le q \le b_t\}$. Our results translate with minimal effort to this definition.

The regret of a learning algorithm $\mathcal{A}$ against an adversary $\mathcal{S}$ generating the sequence of random pairs (S_t, B_t) is defined by:

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

We use P_t, Q_t to stress that the prices are possibly randomized, with the convention that uppercase letters refer to random variables and the corresponding lowercase letters to their realizations. The expectation in the previous formula is then with respect to the internal randomization of the learning algorithm and the adversary. The regret $R_T(\mathcal{A})$ of a learning algorithm $\mathcal{A}$ is defined as its performance against the hardest adversary, i.e., as the supremum over all adversaries $\mathcal{S}$ (in a certain class we describe in the next paragraph) of $R_T(\mathcal{A}, \mathcal{S})$. Our goal is to study the minimax regret $R_T^\star$, which measures the performance of the best algorithm against the worst possible adversary, i.e., the infimum over all algorithms $\mathcal{A}$ of $R_T(\mathcal{A})$. The set of learning algorithms we allow varies with the settings we consider, i.e., with how many prices are posted and what feedback is available—see below.

Smoothed analysis of algorithms, initially introduced by Spielman and Teng (2004) and later formalized for online learning by Rakhlin et al. (2011) and Haghtalab et al. (2020), is an approach to the analysis of algorithms in which the instances at every round are generated from a distribution that is not too concentrated. Recent works on the smoothed analysis of online learning algorithms include Haghtalab et al. (2020), Haghtalab et al. (2022), and Block et al. (2022)—see Section 1.3 for additional related works.

In this work, we consider a (stochastic) smoothed valuation-generating model that, in the limit, recovers the adversarial regime. This is a natural choice for the bilateral trade problem, where algorithms with sublinear regret only exist for the stochastic i.i.d. setting (with additional assumptions) and where the adversarial model is known to be intractable (Cesa-Bianchi et al., 2024). At each time step t, a pair of valuations (s_t, b_t) is sampled according to the random variable (S_t, B_t), whose distribution is chosen by the adversary. Our adversary is adaptive because the distribution of (S_t, B_t) may depend on the past realizations of the valuations and the past internal randomization of the algorithm. We focus on σ-smoothed adversaries, where the distributions of (S_t, B_t) are not too concentrated, according to the following notion.

Definition 1 (Haghtalab et al. (2021)) *Let X be a domain supporting a uniform distribution ν. A measure μ on X is said to be σ-smooth if for all measurable subsets $A \subseteq X$, we have $\mu(A) \leq \frac{\nu(A)}{\sigma}$.*

We say that a random variable is σ-smooth if its distribution is σ-smooth. We consider two families of learning algorithms, corresponding to two ways of being budget balanced:

- **Single-price mechanisms**. If we want to enforce a stricter notion of budget balance, namely strong budget balance, the mechanism is neither allowed to subsidize nor extract revenue from the system. This is modeled by imposing $p_t = q_t$, for all t.
- **Two-price mechanisms**. If a budget balanced algorithm enforces (weak) budget balance, then two different prices can be posted, p_t to the seller and q_t to the buyer, as long as $p_t \leq q_t$ at each time step. Namely, we only require that trades are never subsidized; the mechanism can still make a profit.

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

	Full Feedback	Two-bit Feedback	One-bit Feedback
Single Price	$\widetilde{O}(\sqrt{T})$ Theorem 2	$\Omega(T)$	$\Omega(T)$
Two Prices	$\Omega(\sqrt{T})$	$\Omega(T^{3/4})$ Theorem 4	$\widetilde{O}(T^{3/4})$ Theorem 5

Table 1: Overview of the regret regimes against a σ-smooth adversary. The lower bound for the full feedback model is from Cesa-Bianchi et al. (2024, Thm. 3.3), the one for single price with two-bit feedback is from Theorem 5 in the same paper. Our classification identifies three minimax regret regimes: $\sqrt{T}$ (green), $T^{3/4}$ (orange), and T (red).

We consider three natural types of feedback models presented in increasing order of difficulty for the learner. The last two are partial feedback models that enjoy the desirable property of requiring only a minimal amount of information from the agents:

- **Full feedback.** $z_t = (s_t, b_t)$: The learner observes both seller and buyer valuations. This model corresponds to a direct revelation mechanism.
- **Two-bit feedback.** $z_t = \big(\mathbb{I}\{s_t \le p_t\}, \mathbb{I}\{q_t \le b_t\}\big)$: The learner observes separately if the two agents accept the prices offered to each of them.
- **One-bit feedback.** $z_t = \mathbb{I}\{s_t \le p_t \le q_t \le b_t\}$: The learner only observes whether or not the trade occurs. This is arguably the minimal feedback the learner could get.

We remark that by Observation 1, the only reason to post two distinct prices in a given round is to get information. This implies that, in the full feedback model, there is no reason to do that, as all the relevant information is revealed anyway.

1.1 Overview of Our Results

We characterize (up to logarithmic factors) the dependence in the time horizon of the minimax regret regimes for the online learning version of the bilateral trade problem against an adaptive σ-smooth adversary for various feedback models and notions of budget balance, as outlined in Table 1. We prove the following results:

- For the full feedback model, we analyze a continuous version of Hedge, posting a single price at each time step and enjoying a $O(\sqrt{T \ln T})$ bound on the regret (Theorem 2). By Cesa-Bianchi et al. (2024, Theorem 3.3), this rate is optimal up to logarithmic factors.
- For the one-bit feedback model, we design the Blind-Exp3 algorithm, posting two prices at each time step and enjoying a $\widetilde{O}(T^{3/4})$ bound on the regret (Theorem 5). The same rate was already obtained by the Scouting Blindits algorithm in Cesa-Bianchi et al. (2024), but only under the additional assumptions that the adversary chooses the seller/buyer valuations according to an i.i.d. process. In this work, we drop this assumption and show that smoothness alone is the crucial property enabling sublinear regret.
- We prove that, surprisingly, the $T^{3/4}$ rate is optimal up to logarithmic terms (Theorem 4), even if the adversary is forced to choose valuations according an i.i.d. process and the learner has access to the more informative two-bit feedback. Notably, our lower bound closes –in an unexpected way– an open problem in Cesa-Bianchi et al. (2024).
- We prove that no algorithm can achieve worst-case sublinear regret when the platform is forced to post a single price but receives partial feedback (one or two bits), even in the case where

the seller/buyer evaluations are σ-smooth, independent of each other, and form an independent sequence (Theorem 3). This complements a result in Cesa-Bianchi et al. (2024, Theorem 5), where the same lower bound was proven for an i.i.d. smoothed adversary.

We highlight three salient qualitative features of our results. First, we construct a (surprising) lower bound of order $T^{3/4}$ for the minimax regret of the problem with partial feedback where the learner is allowed to post two prices. This lower bound, which is also our main technical contribution, is strictly worse that the $T^{2/3}$ rate that can be obtained with access to bandit feedback,[†] and substantially departs from the rates $\sqrt{T}, T^{2/3}, T$ that can be found in the two most closely related partial feedback models in the literature: online learning with feedback graphs (Alon et al., 2017) and partial monitoring (Bartók et al., 2014). Second, we introduce the first sublinear-regret learning algorithm for the partial feedback version of the bilateral trade problem beyond the (strict) stochastic i.i.d. assumption on the valuations. Third, our results imply that, from the online learning perspective, there is no difference between receiving one or two bits of feedback when two prices can be posted. This is in agreement, and extends beyond the i.i.d. case, what was already noted in Cesa-Bianchi et al. (2024, Section 7) for the smoothed i.i.d. case. This is also in stark contrast with what happens in the stochastic case: if only one price can be posted, then $O(T^{2/3})$ regret is achievable when the learner has access to two-bit feedback and S, B are independent and smooth. On the other hand, one bit of feedback is not enough to obtain sublinear regret—see Cesa-Bianchi et al. (2024, Sections 5 and 8).

1.2 Technical Challenges and Our Techniques

The repeated bilateral trade problem is characterized by two key features that set it apart from the standard model of online learning with full or bandit feedback: the nature of the action space and the partial feedback structure. Both these features need to be taken into account to construct the $T^{3/4}$ lower bound, which is the main technical endeavor of this work.

The action space & the smooth adversary. The action space of the bilateral trade problem is continuous (the prices live in a subset of $[0,1]^2$), while the gain from trade is discontinuous. This entails that, without any smoothness assumptions on the distributions, the problem turns out to be utterly intractable in the standard adversarial setting—see the "needle in a haystack" phenomenon in Cesa-Bianchi et al. (2024, Theorem 6) and Azar et al. (2022, Theorem 3). We show that the σ-smoothness induces regularity on the expected gain from trade (Lemma 1), which in turn allows us to prove a key discretization result (Claim 1). In the full feedback model, we actually prove something stronger: a continuous version of the Hedge algorithm directly exhibits sublinear regret with respect to the best *continuous* price, without resorting to a finite grid of candidate prices (Theorem 2). We expand on this technique, which may be of general and independent interest, in Appendix A.

Partial feedback. The main peculiarity of the bilateral trade problem lies in the partial feedback models that are naturally associated with it. Receiving only information about the relative ordering of the prices posted and the realized valuations does not allow the learner to directly reconstruct the gain from trade received at each time step. For instance, if the learner posts the same price 0.5

[†] Although our decision space is two-dimensional, one can see that, in a bandit feedback (in which the learner observes the gain from trade at each time step) with a smooth adversary, a regret of order $T^{2/3}$ can be obtained by running an optimal bandit algorithm (e.g., MOSS of Audibert and Bubeck 2009, whose upper bound on the regret is of order $\sqrt{KT}$) on a discretization of $K = \Theta(T^{1/3})$ equally spaced prices on the diagonal $\{(p,q) \in \mathcal{U} \mid p = q\}$. Similar results appeared, e.g., in Kleinberg (2004); Auer et al. (2007). This is not in contradiction with our $T^{3/4}$ lower bound because the feedback model considered there is less informative.

to both agents and they both accept, there is no way of assessing whether its gain from trade is constant (e.g., $(s, b) = (0, 1)$) or arbitrarily small (e.g., $s = 0.5 - \varepsilon$ and $b = 0.5 + \varepsilon$). Conversely, if one of the two agents rejects the price posted, the learner can only infer loose bounds on the lost trade opportunity. The key technical tool to address this challenge is given by a one-bit estimation technique that exploits the possibility of posting *two* prices to estimate the gain from trade it would have achieved by posting *one* single price to both agents (Cesa-Bianchi et al., 2024; Azar et al., 2022). This tool, together with our discretization result (Lemma 1) are behind our Blind-Exp3 algorithm achieving a $T^{3/4}$ regret.

Our $T^{3/4}$ lower bound. At a (very) high level, we show that bilateral trade with partial feedback contains instances that are closely related to instances of online learning with feedback graphs (Alon et al., 2015). The corresponding feedback graph G_K is over $2K$ actions: K of them are "exploring" and the others are "exploiting". Exploring actions are costly and reveal feedback on the corresponding exploiting actions. One of the exploiting actions is optimal, but none of them returns any feedback. We build "hard" instances so that any algorithm is forced to spend a long time playing each one of the many exploring actions. By selecting optimally the number of arms in the reduction and the difference in reward between exploiting actions, we obtain the $T^{3/4}$ rate. This proof sketch hides many technical challenges; crucially, we need to carefully design σ-smooth distributions of the adversary with the desired properties. This presents two problems: on the one hand, the gains from trade achievable at different prices are related (while in usual lower bound constructions for online learning with feedback graphs, the rewards can be chosen independently, Alon et al. 2015); on the other hand, the embedding needs to preserve the feedback structure, which is significantly different from the standard bandit or expert feedback and requires novel and subtle arguments.

1.3 Additional Related Work

Further applications of smoothed analysis to online learning problems include the works by Block and Simchowitz (2022), Block et al. (2023), Cesa-Bianchi et al. (2024), Aggarwal et al. (2024), and Durvasula et al. (2023). Sachs et al. (2022) study a related stochastic adversary in the more general online convex optimization setting; however, they do not insist on the smoothness of the distributions.

In online learning settings with partial feedback, like the one we study here, smoothed analysis has been primarily applied to linear contextual bandits (Kannan et al., 2018; Raghavan et al., 2020; Sivakumar et al., 2020, 2022), where contexts are drawn from smooth distributions. However, the focus of those works has been on improving regret bounds specifically for the greedy algorithm, whose worst-case regret is linear. Although the smoothed adversary causes the expected gain from trade to be Lipschitz, the best possible regret rates for the partial feedback models considered here are provably worse than those achievable with bandit feedback. To the best of our knowledge, bilateral trade with a smoothed adversary was previously studied only by Cesa-Bianchi et al. (2021) in the two-bit feedback model.Another line of work considers regret bounds parameterized by variations of losses across time and other related measures of smoothness (Hazan and Kale, 2010; Chiang et al., 2012; Steinhardt and Liang, 2014). See also Chen et al. (2021) for recent results in this area.

The minimax regret of online learning with partial feedback is rather well understood when the learner selects actions from a finite set—see, e.g., the vast literature on feedback graphs and the recent work by Lattimore (2022) on partial monitoring. General analyses of settings with infinitely many actions sets are mostly limited to bandit feedback (Kleinberg et al., 2019).

Conference Version and Follow-up Work. A preliminary version of this paper appeared in the Conference on Learning Theory (Cesa-Bianchi et al., 2023). In follow-up work, Bernasconi et al. (2024) showed how to achieve sublinear regret in adversarial repeated bilateral trade by allowing the learning algorithm to enforce a global notion of budget balance. Recently, many economically-motivated problems, related to bilateral trade, have been studied from the (online) learning perspective. For example, fair bilateral trade (Bachoc et al., 2024a), double auctions with one seller and two buyers (Babaioff et al., 2024), brokerage (Bolic et al., 2024; Bachoc et al., 2024b), contextual bilateral trade (Gaucher et al., 2024), and trading volume maximization (Cesari and Colomboni, 2024).

2. Warm-up: One-Price Setting

In this section, we present our discretization error result (sharpening by a constant the bound in Cesa-Bianchi et al. 2024) and present our results in the single-price setting.

2.1 Regret due to Discretization

Our first theoretical result concerns the study of how discretization impacts the regret against σ-smooth adversaries. Although the gain from trade is, in general, discontinuous, its expectation is $1/\sigma$-Lipschitz, thus opening the way to discretization methods, as formalized by the following result.

Lemma 1 (Lipschitzness) *Let (S, B) be a σ-smooth random variable on $[0,1]^2$, then the induced expected gain from trade* GFT *is $1/\sigma$-Lipschitz:*

$$|\mathbb{E}\left[\mathrm{GFT}(y) - \mathrm{GFT}(x)\right]| \leq \frac{1}{\sigma}|y-x|, \quad \forall x, y \in [0,1]. \tag{1}$$

Proof Let $x > y$ be any two prices in $[0,1]$, and U and V, two independent uniform random variables in $[0,1]$, we have the following chain of inequalities:

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

Note that in the second to last inequality we used the smoothness of (S, B). ■

This regularity result implies that the definition of regret we are considering is well posed, as there always exists a single price maximizing the gain from trade in hindsight. To see this, consider any choice of the sequence of σ-smooth distributions of the adversary; by Observation 1, we know that we only need to focus on one single price, and from Lemma 1 that the total gain from trade is Lipschitz and therefore continuous on $[0,1]$. We prove now that for any fixed grid of prices G in $[0,1]$, it is possible to relate the gain from trade of the best price in G with that of the best fixed price in $[0,1]$, paying a discretization error that depends on the smoothness parameter and the coarseness of the grid. To this end, for any finite grid G, we define the parameter $\delta(G)$ as follows:

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

Claim 1 (Discretization error) *Let G be any finite grid of prices in $[0,1]$, then for any sequence of σ-smooth distributions $\mathcal{S} = (S_1, B_1), \dots, (S_T, B_T)$, we have the following:*

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

Proof Let p^* be the best fixed price in hindsight in $[0,1]$ with respect to the sequence $\mathcal{S}$. We have two cases. If $p^* \in G$, then there is nothing to prove. If this is not the case, then there exists $p_G \in G$, such that $|p^* - p_G| \le \delta(G)$. We have the following:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p^*)\right] - \max_{p\in G} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p)\right] \\
&\quad\le \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p^*)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}_t(p_G)\right] \\
&\quad\le \frac{|p^* - p_Q|}{\sigma} \le \frac{\delta(G)}{\sigma},
\end{aligned}$$

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

2.2 Posting a Single Price in Full Information

In the full feedback model, the learner observes a realization $z_t = (s_t, b_t)$ of (S_t, B_t) at the end of each round t, allowing for a reconstruction of the gain from trade that would have been achieved by any other pair of prices. We show that running Hedge (Freund and Schapire, 1997) on the continuum of arms/prices in $[0,1]$ gives a regret rate that is optimal in T and exponentially better in the smoothness parameter compared to the direct "discretization + discrete Hedge" approach[‡]. Our algorithm, Continuous-Price Hedge, is a version of the classic Hedge algorithm played on a continuum of prices where, at time t, a price p_t is drawn according to the *continuous* distribution μ_t with density f_t defined on $[0,1]$ as follows:

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

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

While continuous versions of Hedge have already been studied, we are the first to provide positive results under the assumption that *expected* rewards are Lipschitz. Previous work (Maillard and Munos, 2010; Krichene et al., 2015) assumes Lipschitzness of the rewards for *any realization*. The latter assumption is, however, not applicable for gain from trade, which is discontinuous and not even one-sided Lipschitz in general. This seemingly small difference –from a rewards family that is

[‡] We refer to the conference version for further details (Cesa-Bianchi et al., 2023, Theorem 2).

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

realization-wise Lipshitz to one that is regular only in expectation– entails significant technical issues in the analysis that we bypass by proving two general results that we believe are of independent interest: a log-exp analogous of Minkowski's integral inequality (Lemma 3 in Appendix B) and a generalized freezing lemma (Lemma 5 in Appendix C). Given the technical nature of the arguments, we postpone the proof of these results to the Appendices, and we report here the statement of our result.

Theorem 2 *Consider the problem of repeated bilateral trade against a σ-smooth adaptive adversary in the full feedback model, for any $\sigma \in (0,1]$. If we run* Continuous-Price Hedge *with learning rate $\eta \in (0,1)$, then, for each time horizon $T \in \mathbb{N}$, we have that*

$$R_T(\text{Continuous-Price Hedge}) \le \frac{1}{\eta} \ln\left(\frac{\eta T \max(\frac{1}{\sigma}, 2)}{1 - e^{-\eta T}}\right) + (e-2)\eta T \; .$$

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

$$R_T(\text{Continuous-Price Hedge}) \le \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln\big(\max(\frac{1}{\sigma}, 2)\big)}{\ln(2T)}\right) \; .$$

Besides this specific result for gain from trade, in Appendix A, we prove a general version of this Theorem, namely Theorem 6, that holds for any situation where the expected reward function is Lipshitz, without requiring realization-wise regularity.

The bound in Theorem 2 is optimal in the time horizon (Cesa-Bianchi et al., 2024, Theorem 3.3) up to logarithmic terms, and exhibits an extremely mild dependence on $1/\sigma$ —disappearing completely if T is larger than $1/\sigma$— without requiring the knowledge of σ to tune the parameter learning rate η. This dependence in the smoothness parameter is exponentially better than the one achievable by directly combining Claim 1 with Hedge on a finite set of candidates. Indeed, this latter, simpler approach yields a regret bound of $O(\sqrt{T \log T / \sigma})$.[§] In regimes where σ is small, e.g., $1/\sigma = T^\alpha$, with $\alpha \ge 1$, the latter bound guarantees are vacuous, while Theorem 2 maintains a near-optimal $\sqrt{T}$ regret, only paying α multiplicatively.

2.3 Posting a Single Price in Partial Information

Cesa-Bianchi et al. (2024) proved that sublinear regret is achievable with one price and partial information in the stochastic i.i.d. case, when seller and buyer distributions are smooth and independent

[§] We refer to the conference version (Cesa-Bianchi et al., 2023) for further details

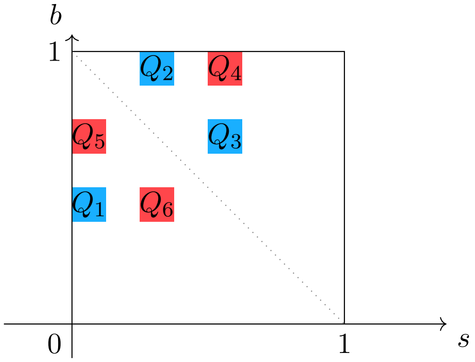

Figure 1: The squares $Q_1, \dots, Q_6$ appearing in the proof of Theorem 3.

of each other. They also showed that removing either the smoothness assumption or the mutual independence of S and B leads to linear lower bounds. They did not, however, investigate whether the i.i.d. assumption could be lifted in a setting other than the classic adversarial one while still achieving sublinear regret. In contrast to the full information scenario above (and the one with two prices and partial feedback that we discuss later), we give a negative answer to this question.

Theorem 3 *Consider the problem of repeated bilateral trade against a σ-smooth adversary in the two-bit feedback model, for any $\sigma \le \frac{1}{64}$. Then any learning algorithm that posts a single price per time step suffers at least $\frac{T}{24}$ regret, even if $S_1, B_1, S_2, \dots$ is an independent family of random variables.*

Proof Consider the following six squares, depicted in Figure 1:

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

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

From Cesa-Bianchi et al. (2024, Theorem 4.3), we know that any learning algorithm $\mathcal{A}$ that can only post one price p_t suffers linear regret against at least one of the following i.i.d. instance: the adversary chooses at the beginning of time either the red or the blue distribution and extracts valuations from it i.i.d. over the rounds. In formula:

$$\max_{\text{color}\in\{\text{blue,red}\}} \left(\max_{p\in[0,1]} \sum_{t=1}^{T} \mathbb{E}^{\text{color}}\big[\mathrm{GFT}_t(p) - \mathrm{GFT}_t(p_t)\big] \right) \ge \frac{1}{24} T. \tag{2}$$

We cannot use directly this construction for our result, as seller and buyer valuations are not independent in the blue and red distributions. However, we can exploit the non i.i.d. structure of the smooth adversary, to generate an equivalent random sequence of smooth distributions such that each one of them has *independent* seller and buyer valuations.

Consider the following family F of $1/64$-smooth oblivious adversaries: each $\mathcal{S}$ of them is characterized by a color red or blue, and a sequence $\{i_t\}$ of T indices, where red adversaries have $i_t \in \{4,5,6\}$ and blue adversaries have $i_t \in \{1,2,3\}$. We denote with F^{red} the set of all such adversaries and with F^{blue} the blue ones. Any $\mathcal{S}$ in the sequence generates the valuations as follows: (S_t, B_t) is drawn independently and uniformly at random from Q_{i_t}. Note that any $\mathcal{S} \in F$ enjoys the property that the distribution chosen at each time step has independent seller and buyer. We argue that any learning algorithm $\mathcal{A}$ suffers linear regret against at least one of these adversaries. In formula:

$$
\begin{aligned}
R_T(\mathcal{A}) &\geq \max_{\mathcal{S}\in F}\left[\max_{p\in[0,1]}\left(\sum_{t=1}^{T}\mathbb{E}^{i_t}\left[\mathrm{GFT}_t(p)-\mathrm{GFT}_t(p_t)\right]\right)\right]\\
&= \max_{\text{color}\in\{\text{red},\text{blue}\}}\ \max_{\mathcal{S}\in F^{\text{color}}}\left[\max_{p\in[0,1]}\left(\sum_{t=1}^{T}\mathbb{E}^{i_t}\left[\mathrm{GFT}_t(p)-\mathrm{GFT}_t(p_t)\right]\right)\right]\\
&\geq \max_{\text{color}\in\{\text{red},\text{blue}\}}\left[\max_{p\in[0,1]}\left(\sum_{t=1}^{T}\mathbb{E}^{\text{color}}\left[\mathrm{GFT}_t(p)-\mathrm{GFT}_t(p_t)\right]\right)\right]. \qquad (3)
\end{aligned}
$$

Note that the i_t are the indices induced by $\mathcal{S}$. The previous inequality, combined with Equation (2) concludes the proof. The only delicate step we need to clarify is the last inequality in Equation (3). To this end, fix any color, let's say red (same argument holds for blue). The regret of $\mathcal{A}$ against the worst sequence in F^{red} is at least the expected regret of $\mathcal{A}$ against a randomized adversary that is obtained by drawing u.a.r. $\mathcal{S}$ from F^{red} (note that the adversaries in F^{red} are oblivious). Now, the crucial argument is that the sequence of valuations (S_t, B_t) obtained by choosing u.a.r. an adversary $\mathcal{S}$ from F^{red} follows the exact same distribution as drawing (S_t, B_t) i.i.d. from the red distribution. In fact, the valuations at different steps are independent and every square has the same probability of being chosen at each time step. ■

3. A $T^{3/4}$ Lower Bound: Two Bits and Two Prices

In this section, we present the main contribution of this paper: an unexpected and intriguing lower bound of order $T^{3/4}$. This result has two notable implications. First, it provides a formalization to the intuition that *partial* feedback (both one and two-bit models) is strictly less informative than the *bandit* feedback, being the regret of the latter of order at most $T^{2/3}$. Second, noting that the hard instances used in the proof are i.i.d., we solve an open problem in Cesa-Bianchi et al. (2024), where it was erroneously conjectured that the correct minimax rate was $T^{2/3}$.

We prove this result —formally stated in Theorem 4— in Section 3.2. Preliminarily, in Section 3.1 we introduce the hard family of adversaries used in Section 3.2, while Section 3.3 is devoted to the formal derivation of a technical passage of the proof of Theorem 4, which is postponed there for the sake of readability.

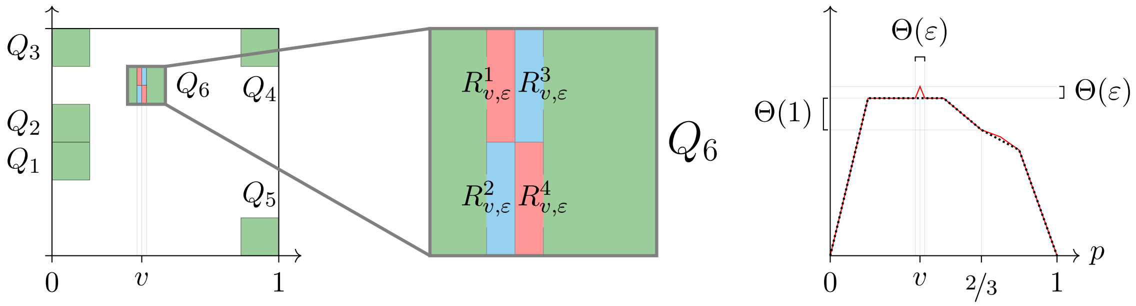

Figure 2: Left/center: The six squares $Q_1, \dots, Q_6$ (in green) are the support of the base density f, and the four rectangles $R^1_{v,\varepsilon}, \dots, R^4_{v,\varepsilon}$ (in red and blue) inside Q_6 are the regions where the density is perturbed with $g_{v,\varepsilon}$. Right: The corresponding qualitative plots of $p \mapsto \mathbb{E}[\mathrm{GFT}(p,S,B)]$ (black, dotted) and $p \mapsto \mathbb{E}^{v,\varepsilon}[\mathrm{GFT}(p,S,B)]$ (red, solid).

3.1 A hard family of adversaries

We construct a "hard" family of i.i.d. σ-smooth adversaries for the repeated bilateral trade learning problem. Under each such adversary the valuations (S_t, B_t) are drawn i.i.d. from a fixed distribution, for this reason, each adversary is identified by a probability measure over $[0,1]^2$, according to which the random valuations are drawn. These probability measures are absolutely continuous with respect to the Lebesgue measure and are obtained by suitable perturbations over a base distribution f, whose support is given by the union of the six squares $Q_1, \dots, Q_6$ (see Figure 2, left):

$$Q_1 = \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{1}{3}, \tfrac{1}{2}\right) , \qquad Q_2 = \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{1}{2}, \tfrac{2}{3}\right] , \qquad Q_3 = \left[0, \tfrac{1}{6}\right] \times \left[\tfrac{5}{6}, 1\right] ,$$
$$Q_4 = \left[\tfrac{5}{6}, 1\right] \times \left[\tfrac{5}{6}, 1\right] , \qquad Q_5 = \left[\tfrac{5}{6}, 1\right] \times \left[0, \tfrac{1}{6}\right] , \qquad Q_6 = \left[\tfrac{1}{3}, \tfrac{1}{2}\right] \times \left[\tfrac{2}{3}, \tfrac{5}{6}\right] .$$

The base probability density function f is defined for all $(x,y) \in [0,1]^2$ by

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

where a is set to $2 \cdot \ln(27/16)$ for normalization. Each perturbations is parametrized by a center v and a scale ε, with $(v,\varepsilon) \in \Xi = \left\{ (v,\varepsilon) \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right) \times \left(0, \tfrac{1}{12}\right) \mid \tfrac{1}{3} + \varepsilon \le v \le \tfrac{1}{2} - \varepsilon \right\}$, and has support on four disjoint rectangles (Figure 2, left):

$$R^1_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\tfrac{3}{4}, \tfrac{5}{6}\right] , \qquad R^2_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\tfrac{2}{3}, \tfrac{3}{4}\right) ,$$
$$R^3_{v,\varepsilon} = [v, v+\varepsilon) \times \left[\tfrac{3}{4}, \tfrac{5}{6}\right] , \qquad R^4_{v,\varepsilon} = [v, v+\varepsilon) \times \left[\tfrac{2}{3}, \tfrac{3}{4}\right) .$$

The $R^i_{v,\varepsilon}$ rectangles are included in Q_6 and are the support of the corresponding perturbation $g_{v,\varepsilon}$ defined for all $(x,y) \in [0,1]^2$ by

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

The perturbed density functions are obtained by summing together the base probability density function f and one of the perturbations $g_{v,\varepsilon}$. Formally, for all $(v,\varepsilon) \in \Xi$, we let $f_{v,\varepsilon} = f + g_{v,\varepsilon}$.

Let $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$, for all $(v,\varepsilon) \in \Xi$) be a probability measure such that the sequence of seller/buyer evaluations (S,B), (S_1,B_1), (S_2,B_2), $\dots$ is i.i.d. and the distribution of (S,B) has density f (resp., $f_{v,\varepsilon}$) with respect to the Lebesgue measure. We denote the expectation with respect to $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$, for all $(v,\varepsilon) \in \Xi$) by $\mathbb{E}$ (resp., $\mathbb{E}^{v,\varepsilon}$). In Claim 2, we formally prove that $\mathbb{P}_{(S,B)}$ (resp., $\mathbb{P}^{v,\varepsilon}_{(S,B)}$, for all $(v,\varepsilon) \in \Xi$) is $1/9$-smooth. Therefore, each adversary corresponding to these distributions is σ-smooth for any $\sigma \le 1/9$.

Claim 2 (Smoothness) *$\mathbb{P}_{(S,B)}$ and $\mathbb{P}^{v,\varepsilon}_{(S,B)}$ are $1/9$-smooth, for all $(v,\varepsilon) \in \Xi$.*

Proof To prove this result, it is enough to argue that the probability density functions of (S,B) are uniformly upper bounded by 9 under both $\mathbb{P}$ and $\mathbb{P}^{v,\varepsilon}$. We start by analyzing f; in particular, note that for all $(x,y) \in Q_1$, it holds that

$$\frac{5-6(y+x)}{6(y-x)} = 1 + \frac{5-12y}{6(y-x)} \le 1 + \frac{5-12\frac{1}{3}}{6\left(\frac{1}{3}-\frac{1}{6}\right)} = 2 \le 2a, \tag{4}$$

where in the first inequality we used that the expression is monotonically increasing for $x \in [0, 1/6]$ and decreasing for $y \in [1/3, 1/2)$, while in the last inequality we applied the definition of the normalization parameter $a = 2\ln(27/16)$. Combining the definition of f with the inequalities in Equation (4), we observe that for all $(x,y) \in [0,1]^2$, it holds that:

$$f(x,y) \le \frac{72a}{1+8a} \le 8.04. \tag{5}$$

We have then just proved the part of the statement concering $\mathbb{P}_{(S,B)}$. We now move our attention to $\mathbb{P}^{v,\varepsilon}_{(S,B)}$ and observe that $g_{v,\varepsilon}$ is supported in Q_6, for all $(v,\varepsilon) \in \Xi$. The probability density function $f_{v,\varepsilon}$ is then different from f only in Q_6, where we have the following uniform upper bound

$$f_{v,\varepsilon}(x,y) = \frac{72}{1+8a} \le 7.69.$$

This concludes the proof. ■

Expected gain from trade. The choice of distributions $f_{v,\varepsilon}$ is due to the specific structure of the expected gain from trade and feedback they induce. We start analyzing the former. For each $(v,\varepsilon) \in \Xi$, and $p \in [0,1]$, it is easy to argue, by linearity, that

$$\begin{aligned}\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,S,B)\big] &= \mathbb{E}\big[\mathrm{GFT}(p,S,B)\big] + \int_{[0,p]\times[p,1]} (y-x) g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \mathbb{E}\big[\mathrm{GFT}(p,S,B)\big] + \tfrac{\varepsilon}{864(1+8a)} \cdot \Lambda_{v,\varepsilon}(p) + \tfrac{\varepsilon^2}{72(1+8a)} \cdot \Lambda_{\frac{3}{4},\frac{1}{12}}(p)\ ,\end{aligned} \tag{6}$$

where $\Lambda_{u,r}$ is the tent map centered at u with radius r, $\Lambda_{u,r}(x) = \left(1 - \frac{|x-u|}{r}\right)^+$. Equation (6) nicely decomposes the expected gain from trade in a fixed term that depends only on the base distribution, a perturbation term centered in v and a second order $\Theta(\varepsilon^2)$ term. As we have an explicit formula for the probability density function of (S,B) under $\mathbb{P}$, we can express analytically the expected gain from trade $\mathbb{E}\left[\mathrm{GFT}(p,S,B)\right]$:

$$\mathbb{E}\big[\mathrm{GFT}(p,S,B)\big] = \frac{1}{6(1+8a)} \cdot \begin{cases} 3p\big(5+29a-6(1+3a)p\big) & \text{if } p \in \left[0, \frac{1}{6}\right] \\ 2+13a & \text{if } p \in \left(\frac{1}{6}, \frac{1}{2}\right] \\ -18ap^2 + 3ap + 2(1+8a) & \text{if } p \in \left(\frac{1}{2}, \frac{2}{3}\right] \\ -18p^2 + 15p + 10a & \text{if } p \in \left(\frac{2}{3}, \frac{5}{6}\right] \\ 72ap(1-p) & \text{if } p \in \left(\frac{5}{6}, 1\right] \end{cases} \tag{7}$$

To have a qualitative understanding of Equation (7) we refer to Figure 2 (dotted black plot on the right): it is clear that the maximum is attained in the plateau, for $p \in [\frac{1}{6}, \frac{1}{2}]$. We highlight that we constructed the base distribution with the explicit goal of having this plateau of maximizers, whose existence is crucial to the lower bound construction. Furthermore, for each $(v, \varepsilon) \in \Xi$, price v is the unique maximizer of the perturbed expected gain from trade $\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p, S, B)\big]$, which is increasing on $\big[0, \frac{1}{6}\big]$, constant on $\big[\frac{1}{6}, v - \varepsilon\big]$, has a symmetric spike on $[v - \varepsilon, v + \varepsilon]$, becomes constant again on $\big[v + \varepsilon, \frac{1}{2}\big]$, and decreases on $\big[\frac{1}{2}, 1\big]$ (Figure 2, red plot on the right). Recalling that, as noted in Observation 1, the expected gain from trade is maximized on the diagonal $\big\{(p, p) \mid p \in [0, 1]\big\}$, we obtain that the expected gain from trade under $\mathbb{E}^{v,\varepsilon}$ is maximized by posting (v, v); it holds:

$$\max_{(p,q)\in\mathcal{U}} \mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p, q, S, B)\big] = \mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(v, S, B)\big],$$

where we denote with $\mathcal{U}$ the upper right triangle of the $[0, 1]^2$ squares, which corresponds to the set of budget balanced prices that the learner can post.

Two-bit feedback. We move our attention to the description of the distribution of the 2-bit feedback $\big(\mathbb{I}\big\{S \le p\big\}, \mathbb{I}\big\{q \le B\big\}\big)$. It is the same regardless of the underlying perturbed probability measure unless the learner selects a pair of prices (p, q) in one of the four rectangles $R^j_{v,\varepsilon}$ where the perturbations occur. We denote with $R_{v,\varepsilon}$ the union of the four rectangles $R^j_{v,\varepsilon}$. For the sake of simplicity, we use the random variable Z to denote $\big(\mathbb{I}\big\{S \le p\big\}, \mathbb{I}\big\{q \le B\big\}\big)$.

Claim 3 *Fix any $(v, \varepsilon) \in \Xi$, $(p, q) \in \mathcal{U} \setminus R_{v,\varepsilon}$, and let $Z = (\mathbb{I}\{S \le p\}, \mathbb{I}\{q \le B\})$. Then Z follows the same distribution both under $\mathbb{P}$ and $\mathbb{P}^{v,\varepsilon}$. Formally, the following holds*

$$\mathbb{P}^{v,\varepsilon}\Big[Z = (i, j)\Big] = \mathbb{P}\Big[Z = (i, j)\Big] \quad \forall\, (i, j) \in \{0, 1\}^2.$$

Proof For each $(v, \varepsilon) \in \Xi$, and each $(p, q) \in \mathcal{U}$, the distribution under $\mathbb{P}^{v,\varepsilon}$ of the 2-bit feedback Z is given by:

- $\mathbb{P}\left[Z = (0, 0)\right] = \mathbb{P}^{v,\varepsilon}\big[S > p \cap B < q\big] = \int_p^1 \int_0^q f(x, y)\,\mathrm{d}x\mathrm{d}y + \int_p^1 \int_0^q g_{v,\varepsilon}(x, y)\,\mathrm{d}x\mathrm{d}y.$
- $\mathbb{P}\left[Z = (0, 1)\right] = \mathbb{P}^{v,\varepsilon}\big[S > p \cap B \ge q\big] = \int_p^1 \int_q^1 f(x, y)\,\mathrm{d}x\mathrm{d}y + \int_p^1 \int_q^1 g_{v,\varepsilon}(x, y)\,\mathrm{d}x\mathrm{d}y.$
- $\mathbb{P}\left[Z = (1, 0)\right] = \mathbb{P}^{v,\varepsilon}\big[S \le p \cap B < q\big] = \int_0^p \int_0^q f(x, y)\,\mathrm{d}x\mathrm{d}y + \int_0^p \int_0^q g_{v,\varepsilon}(x, y)\,\mathrm{d}x\mathrm{d}y.$
- $\mathbb{P}\left[Z = (1, 1)\right] = \mathbb{P}^{v,\varepsilon}\big[S \le p \cap B \ge q\big] = \int_0^p \int_q^1 f(x, y)\,\mathrm{d}x\mathrm{d}y + \int_0^p \int_q^1 g_{v,\varepsilon}(x, y)\,\mathrm{d}x\mathrm{d}y.$

By symmetry, all integrals of $g_{v,\varepsilon}$ in the previous formulae vanish if (p, q) does not belong to $R_{v,\varepsilon}$, so that the only non-zero contribution is the one of f, which is shared by all distributions. ■

The cost of exploration and of suboptimality. The family of adversaries has two crucial features: first, prices that are ε-far from the optimal one yield $\Theta(\varepsilon)$ instantaneous regret; second, the learner is forced to post prices in a suboptimal region Q_6 to locate the actual perturbation (see Claim 3), incurring in constant instantaneous regret. We formalize these two properties in the following claims. By the analytic expression of the expected gain from trade it is easy to derive a bound on the cost of posting prices far from the optimal one.

Claim 4 (Cost of suboptimality) *Fix any perturbation pair $(v, \varepsilon) \in \Xi$ and let (p, q) be any price not in $([v-\varepsilon, v+\varepsilon] \times [1/3, 2/3]) \cap \mathcal{U}$, then the following holds:*

$$\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(v, S, B)\big] - \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] \geq \frac{1}{10^4}\varepsilon.$$

Proof Consider any $(p, q) \notin ([v-\varepsilon, v+\varepsilon] \times [1/3, 2/3]) \cap \mathcal{U}$, we divide the analysis in two cases, depending on whether $p \in [2/3, 5/6]$ or not. If $p \notin [2/3, 5/6]$, then:

$$\begin{aligned}
\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] &\leq \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, S, B)\big] && \text{(Posting different prices is suboptimal)}\\
&= \mathbb{E}\big[\text{GFT}(p, S, B)\big] && \text{(By Equation (6), as } p \notin [v-\varepsilon, v+\varepsilon] \cup [2/3, 5/6])\\
&\leq \mathbb{E}\big[\text{GFT}(v, S, B)\big]. && (v \text{ is optimal under } \mathbb{P}\text{, see Equation (7))}
\end{aligned}$$

Plugging the above inequality in the left-hand side of the statement of the Claim, together with the gain from trade decomposition of Equation (6) yields:

$$\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(v, S, B)\big] - \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] \geq \tfrac{\varepsilon}{864(1+8a)} \geq \tfrac{\varepsilon}{10^4},$$

where the last inequality follows from the definition of the normalization parameter a. Consider now the other case, i.e. $p \in [2/3, 5/6]$, we have

$$\begin{aligned}
\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] &\leq \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, S, B)\big] && \text{(Posting different prices is suboptimal)}\\
&\leq \mathbb{E}\big[\text{GFT}(p, S, B)\big] + \tfrac{\varepsilon^2}{72(1+8a)} && \text{(By Equation (6), as } p \in [2/3, 5/6])\\
&\leq \mathbb{E}\big[\text{GFT}(2/3, S, B)\big] + \tfrac{\varepsilon^2}{72(1+8a)}. && (\mathbb{E}\left[\text{GFT}(\cdot)\right] \text{ is decreasing in } [2/3, 5/6])\\
&= \tfrac{10a+2}{6(1+8a)} + \tfrac{\varepsilon^2}{72(1+8a)}. && \text{(By Equation (7))}
\end{aligned}$$

We substitute this inequality in the left-hand side of the statement of the Claim:

$$\begin{aligned}
\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(v, S, B)\big] - \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] &\geq \tfrac{13a+2}{6(1+8a)} + \tfrac{\varepsilon}{864(1+8a)} - \tfrac{10a+2}{6(1+8a)} - \tfrac{\varepsilon^2}{72(1+8a)}\\
&\geq \tfrac{3a}{6(1+8a)} \geq \tfrac{\varepsilon}{10^4}. && (\text{As } \varepsilon \leq 1/12)
\end{aligned}$$

This concludes the proof. ■

To bound the istantaneous regret when posting prices in the exploration region Q_6 we need to resort once again to the analytic expression of the gain from trade in Equations (6) and (7).

Claim 5 (Cost of exploration) *Fix any perturbation pair $(v, \varepsilon) \in \Xi$ and let $(p, q) \in Q_6$, the following holds:*

$$\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(v, S, B)\big] - \mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] \geq \frac{1}{20}.$$

Proof Fix any perturbation pair $(v, \varepsilon) \in \Xi$ and consider any pair of prices $(p, q) \in Q_6$. The expected gain from trade corresponding to (p, q) is dominated by the one attainable by posting $(\frac{1}{2}, \frac{2}{3})$ (each pair of valuations (s, b) such that a trade happens for (p, q) also yields a trade under $(\frac{1}{2}, \frac{2}{3})$). Thus the following holds:

$$\mathbb{E}^{v,\varepsilon}\big[\text{GFT}(p, q, S, B)\big] \leq \mathbb{E}^{v,\varepsilon}\left[\text{GFT}\left(\tfrac{1}{2}, \tfrac{2}{3}, S, B\right)\right] \leq \mathbb{E}^{v,\varepsilon}\left[\text{GFT}\left(\tfrac{2}{3}, S, B\right)\right].$$

On the other hand, posting v is at least as good as posting $1/2$:

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(v,S,B)\big] \ge \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\tfrac{1}{2},S,B\right)\right] \ .$$

Putting the two inequalities together we get the claimed bound:

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(v,S,B) - \mathrm{GFT}(p,q,S,B)\big] \ge \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\tfrac{1}{2},S,B\right) - \mathrm{GFT}\left(\tfrac{2}{3},S,B\right)\right] = \frac{a}{2+16a}.$$

The statement follows by plugging in the value of the normalization parameter $a = 2 \cdot \ln(27/16)$. ■

3.2 The $T^{3/4}$ Lower Bound

The family of adversaries we introduced are the crucial ingredient for our lower bound. Set $K = \lceil T^{1/4} \rceil$, $\varepsilon = 1/(12K)$ and, for each $i \in \{1,\dots,K\}$, let $v_i = 1/3 + (2i-1)\varepsilon$ be a candidate center. For the sake of convenience, for each $i \in [K]$ denote $\mathbb{P}^{v_i,\varepsilon}$ by $\mathbb{P}^i$ and the corresponding expectation by $\mathbb{E}^i$, and similarly, denote $\mathbb{P}$ by $\mathbb{P}^0$ and the corresponding expectation by $\mathbb{E}^0$.

Price regions. The family of measures $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$ naturally partitions the $[0,1]^2$ square into price regions characterized by similar feedback and similar expected gain from trade:

- *The expolaration regions:* the square Q_6 contains the K disjoint supports of the perturbations. Denote with R_i^j ($i \in [K]$) the support of the perturbations characterizing $\mathbb{P}^i$ ($R^j_{(v_i,\varepsilon)}$ for our choice of ε) and with R_i the union for $j = 1,\dots,4$ of the R_i^j. Recall, posting prices in R_i^j is the only way to discriminate between $\mathbb{P}^i$ and the other distributions (Claim 3), but induces $\Omega(1)$ instantaneous regret (Claim 5).
- *The (exploitation) candidate regions:* for each $i \in [K]$, let O_i be the trapezoid induced by the intersection of $[v_i - \varepsilon, v_i + \varepsilon] \times [1/3, 2/3)$ with $\mathcal{U}$. The learner gets no information by posting prices there, but each O_i contains (v_i, v_i) which is optimal under $\mathbb{P}^i$ and guarantees $\Theta(\varepsilon)$ regret under $\mathbb{P}^j$ for $j \neq i$ (Claim 4).

We have all the ingredients for proving the main result of the paper. We constructed a family of K i.i.d. σ-smooth adversaries for the repeated bilateral trade problem, each characterized by a probability measure $\mathbb{P}^i$ where the valuations (S,B) are sampled from. Under each $\mathbb{P}^i$ the expected gain from trade is maximized in a different pair of prices (v_i, v_i). Every time the learner posts a price that is $\Omega(\varepsilon)$ far from the optimal v_i it suffers instantaneous regret that is $\Omega(\varepsilon)$ (Claim 4). To identify the optimal v_i, the learner needs to identify the actual perturbation. There are $K = \Theta(1/\varepsilon)$ different possible perturbations and, due to the feedback structure, the learner needs to probe separately the K disjoint exploration regions in Q_6 ($\Omega(1/\varepsilon^2)$ times each) to identify the actual perturbation it is playing against. Recall, posting prices in the suboptimal region Q_6 leads to a constant instantaneous regret (Claim 5). All in all any learner suffers a regret of order $\Omega\big(\min\big(K/\varepsilon^2, \varepsilon T\big)\big) = \Omega(T^{3/4})$, given our choices of K and ε. We formalize this intuition in the following theorem.

Theorem 4 *Consider the problem of repeated bilateral trade against a σ-smooth adversary in the two-bit feedback model, for any $\sigma \le \frac{1}{9}$. If $T \ge 6562$, then any learning algorithm $\mathcal{A}$ that posts two prices per time step suffers at least a regret of*

$$R_T(\mathcal{A}) \ge \frac{1}{10^6} T^{3/4} \ .$$

Proof We prove the lower bound via Yao's Principle: there exists a randomized family of adversaries that induces any deterministic algorithm $\mathcal{A}$ to suffer $\Omega(T^{3/4})$ regret. Let K and ε as above, ($K = \lceil T^{1/4} \rceil$, and $\varepsilon = 1/(12K)$) and consider the $K+1$ adversaries corresponding to the probability measures $\mathbb{P}^0$, $\mathbb{P}^1, \dots \mathbb{P}^K$ described at the beginning of the Section. A suitable mixture over these adversaries is the randomized family we apply Yao's Principle on.

For any fixed deterministic algorithm $\mathcal{A}$, we introduce some notation. Let (P_1, Q_1), $(P_2, Q_2) \dots$ be the prices played by $\mathcal{A}$ on the basis of the sequential feedback received $Z_1, Z_2, \dots$. For any $i \in [K]$, define $N_t(i)$ as the random variables counting the number of times the learning algorithm $\mathcal{A}$ plays in the exploration region R_i; similarly, $M_t(i)$ counts the number of times that $\mathcal{A}$ plays in candidate region O_i:

$$N_t(i) = \sum_{s=1}^{t} \mathbb{I}\{(P_s, Q_s) \in R_i\}, \;\; M_t(i) = \sum_{s=1}^{t} \mathbb{I}\{(P_s, Q_s) \in O_i\}.$$

Using these variables, we can define the N_t and M_t as the counters of how many times exploring, respectively exploiting, actions have been played up to time t, for any $t \in [T]$:

$$N_t = \sum_{i \in [K]} N_t(i) \;, \; M_t = \sum_{i \in [K]} M_t(i) \;.$$

In the following Claim, whose proof is deferred to Section 3.3, we relate the expected values of $M_T(i)$ under $\mathbb{P}^0$ and $\mathbb{P}^i$ as a function of the expected number of times the algorithm plays the corresponding exploring actions, i.e., $N_T(i)$. This formalizes the intuition that to discriminate between the different $\mathbb{P}^i$ the learner needs to play exploring actions.

Claim 6 *The following inequality holds true for any $i \in [K]$:*

$$\mathbb{E}^i\big[M_T(i)\big] - \mathbb{E}^0\big[M_T(i)\big] \le 2\varepsilon T \cdot \sqrt{\mathbb{E}^0[N_T(i)]}.$$

We are now ready to bound directly the performance of the learner against the adversaries. Consider any $\mathbb{P}^i$, for $i \in [K]$; algorithm $\mathcal{A}$ suffers $\Theta(1)$ instantaneous regret when it posts prices in the exploration region Q_6 (we count these events with N_T) and suffers at least $\Theta(\varepsilon)$ instantaneous regret when posts prices that are nor in Q_6, nor in O_i, which contains the optimal price v_i (we count these events with $T - N_T - M_T(i)$). All in all, we have the following lower bound on the regret suffered by $\mathcal{A}$:

$$\begin{aligned}
\mathbb{E}^i\left[R_T(\mathcal{A})\right] &\ge \underbrace{\tfrac{\varepsilon}{10^4}}_{\text{Claim 4}} \mathbb{E}^i\left[T - N_T - M_T(i)\right] + \underbrace{\tfrac{1}{20}}_{\text{Claim 5}} \mathbb{E}^i\left[N_T\right] \\
&\ge \tfrac{\varepsilon}{10^4}\left(T - \mathbb{E}^0\left[M_T(i)\right] - 2\varepsilon T\sqrt{\mathbb{E}^0\left[N_T(i)\right]}\right). && \text{(by Claim 6)}
\end{aligned}$$

Averaging with respect to $\mathbb{P}^i$, $i = 1, \dots, K$, we get:

$$\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}^i\left[R_T(\mathcal{A})\right] &\ge \frac{\varepsilon}{10^4}\left(T - \frac{\mathbb{E}^0\left[M_T\right]}{K} - 2\varepsilon T\sqrt{\frac{\mathbb{E}^0\left[N_T\right]}{K}}\right) && \text{(by Jensen Inequality)} \\
&\ge \frac{\varepsilon}{10^4}\left(\frac{9}{10} - 2\varepsilon\sqrt{\frac{\mathbb{E}^0\left[N_T(i)\right]}{K}}\right)T && (T \ge 6562 \implies K \ge 10) \\
&\ge \frac{1}{10^7}\left(27 - 5\sqrt{\frac{\mathbb{E}^0\left[N_T\right]}{T^{3/4}}}\right)T^{3/4}, && (8)
\end{aligned}$$

where the last inequality follows by the definition of ε and K. We can quantify the regret suffered by $\mathcal{A}$ in a simpler way: every time $\mathcal{A}$ plays in the exploration region Q_6 it suffers constant regret:

$$\mathbb{E}^0\left[R_T(\mathcal{A})\right] \geq \tfrac{1}{20}\mathbb{E}^0\left[N_T\right]. \tag{9}$$

Consider now the randomized family of adversaries generated as follows: with probability $1/2$, the sequence of valuations is drawn i.i.d. according to $\mathbb{P}^0$, while with the remaining probability one of the K probability measures $\mathbb{P}^i$ is chosen, uniformly at random. We conclude by Yao's principle, showing that any deterministic algorithm $\mathcal{A}$ suffers $\Omega(T^{3/4})$ regret:

$$\begin{aligned}
R_T^* &\geq \frac{1}{2}\mathbb{E}^0\left[R_T(\mathcal{A})\right] + \frac{1}{2K}\sum_{i=1}^{K}\mathbb{E}^i\left[R_T(\mathcal{A})\right] && \text{(by Yao's principle)}\\
&\geq \frac{1}{2}\left[\frac{1}{20}\mathbb{E}^0\left[N_T\right] + \frac{T^{3/4}}{10^7}\left(27 - 5\sqrt{\frac{\mathbb{E}^0\left[N_T\right]}{T^{3/4}}}\right)\right] && \text{(by Equations (8) and (9))}\\
&\geq \frac{1}{10^6}T^{3/4},
\end{aligned}$$

where the last inequality holds for any possible value of $\mathbb{E}^0\left[N_T\right]$. ■

We conclude with an observation about our result and analysis. Our main goal is to provide an asymptotic lower bound to the minimax regret; as such, we do not optimize the constant multiplying the leading term but often prefer looser bounds to favor readability.

3.3 The Final Ingredient: Claim 6

This Section is devoted to the proof of the technical Claim 6, which quantifies the different behaviour of any deterministic algorithm against the two adversaries $\mathbb{P}^i$ and $\mathbb{P}^0$, in terms of the number of times that the exploring actions are played.

Claim 6 *The following inequality holds true for any $i \in [K]$:*

$$\mathbb{E}^i\big[M_T(i)\big] - \mathbb{E}^0\big[M_T(i)\big] \leq 2\varepsilon T \cdot \sqrt{\mathbb{E}^0[N_T(i)]}.$$

Proof For any $t \in [T]$, the action (P_t, Q_t) selected by $\mathcal{A}$ at round t (and therefore to wich region (P_t, Q_t) belongs to) is a deterministic function of $Z_1, \dots, Z_{t-1}$. In formula, we then have the following

$$\begin{aligned}
\mathbb{E}^i\big[M_T(i)\big] - \mathbb{E}^0\big[M_T(i)\big] &= \sum_{t=2}^{T}\Big(\mathbb{P}^i\big[(P_t, Q_t) \in O_i\big] - \mathbb{P}^0\big[(P_t, Q_t) \in O_i\big]\Big)\\
&\leq \sum_{t=2}^{T}\big\|\mathbb{P}^i_{(Z_1,\dots,Z_{t-1})} - \mathbb{P}^0_{(Z_1,\dots,Z_{t-1})}\big\|_{\text{TV}},
\end{aligned} \tag{10}$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation norm, and $\mathbb{P}^i_{(Z_1,\dots,Z_t)}$ denotes the push-forward measure over $\{0,1\}^{2t}$ induced by the feedback variables when the valuations are drawn according to $\mathbb{P}^i$. We move now our attention towards bounding the total variation norm. To that end we use the Pinsker's inequality and apply the chain rule for the KL divergence $\mathcal{D}_{\mathrm{KL}}$. For each $i \in [K]$ and $t \in [T]$ we have the following:

$$\begin{aligned}\big\|\mathbb{P}^0_{(Z_1,\dots,Z_t)} - \mathbb{P}^i_{(Z_1,\dots,Z_t)}\big\|_{\mathrm{TV}} &\le \sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{(Z_1,\dots,Z_t)},\, \mathbb{P}^i_{(Z_1,\dots,Z_t)}\big)} \\ &= \sqrt{\frac{1}{2}\left(\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1},\, \mathbb{P}^i_{Z_1}\big) + \sum_{s=2}^{t} \mathbb{E}^0\Big[\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s|Z_1,\dots,Z_{s-1}},\, \mathbb{P}^i_{Z_s|Z_1,\dots,Z_{s-1}}\big)\Big]\right)}. \end{aligned} \tag{11}$$

We bound the two types of KL terms separately; starting from the one relative to the first time step.

The pair of prices (P_1, Q_1) is deterministic and, by Claim 3, the KL divergence between the feedback observed against $\mathbb{P}^0$ and $\mathbb{P}^i$ is non-zero if and only if (P_1, Q_1) belongs to the support of the perturbation defining $\mathbb{P}^i$, that we called R_i. We have the following:

$$\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1}, \mathbb{P}^i_{Z_1}\big) = \sum_{z \in \{0,1\}^2} \log\left(\frac{\mathbb{P}^0\left[Z_1 = z\right]}{\mathbb{P}^i\left[Z_1 = z\right]}\right) \mathbb{P}^0\left[Z_1 = z\right] \mathbb{I}\{(P_1, Q_1) \in R_i\}. \tag{12}$$

Focus on the terms of the form $\mathbb{P}^i\left[Z_1 = z\right]$. To get a better understanding of them, we introduce four rectangles that represent the regions of the $[0,1]^2$ square which correspond to the four possible feedback $z \in \{0,1\}^2$ received by the learner:

$$Q_{1,1} = [0, P_1] \times [Q_1, 1], Q_{0,1} = (P_1, 1] \times [Q_1, 1], Q_{1,0} = [0, P_1] \times [0, Q_1), Q_{0,0} = (P_1, 1] \times [0, Q_1).$$

By definition of $\mathbb{P}^i$ and of the rectangles Q_z, we have that $\mathbb{P}^i\left[Z_1 = z\right] = \mathbb{P}^0\left[Z_1 = z\right] + \Delta_z$, where $\Delta_z = \frac{36}{1+8a}\left(|(R_i^1 \cup R_i^4) \cap Q_z| - |(R_i^2 \cup R_i^3) \cap Q_z|\right)$ and we denoted with $|\cdot|$ the area. Now, the function $x \rightarrow x \log {}^{x}\!/\!{}_{x+a}$ is monotonically decreasing in its domain; moreover, the probabilities $\mathbb{P}^0\left[Z_1 = z\right]$ are at least $1/6$ for any $z \in \{0,1\}^2$, when (P_1, Q_1) is in R_i. Applying these considerations to Equation (12) we have that

$$\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1}, \mathbb{P}^i_{Z_1}\big) \le \frac{1}{6} \sum_{z \in \{0,1\}^2} \log\left(\frac{1/6}{1/6 + \Delta_z}\right) \mathbb{I}\{(P_1, Q_1) \in R_i\}. \tag{13}$$

A first, crucial, consideration on the Δ_z is that $\Delta_{0,0}$ and $\Delta_{1,1}$ are non-negative, while the remaining terms are non-positive. This is due to the definition of these terms as difference between two areas; for $z = (0,0)$ and $(1,1)$ it holds that $|(R_i^1 \cup R_i^4) \cap Q_z| \ge |(R_i^2 \cup R_i^3) \cap Q_z|$; for the other two z the converse inequality holds. This means that

$$\Delta_{0,0} \cdot \Delta_{1,1} \ge 0, \quad \Delta_{1,0} \cdot \Delta_{0,1} \ge 0. \tag{14}$$

Consider now the two sums of terms with the same sign. The absolute value of $\Delta_{0,0} + \Delta_{1,1}$ is mazimized when (P_1, Q_1) is equal to $(v, 3/4)$ (i.e., at the center of the R^i rectangle); the same holds for $\Delta_{1,0} + \Delta_{0,1}$, so

$$\max\{|\Delta_{0,0} + \Delta_{1,1}|, |\Delta_{1,0} + \Delta_{0,1}|\} = \tfrac{6}{1+8a}\varepsilon \le \tfrac{2}{3}\varepsilon. \tag{15}$$

Finally, by definition of the Q_z and that of the perturbation rectangles, it holds that the Δ_z terms sum up to 0 :

$$\Delta_{0,0} + \Delta_{1,1} + \Delta_{1,0} + \Delta_{0,1} = 0. \tag{16}$$

We get back to Equation (13) and apply the simple inequalities we just proved, together with the fact that the function $x \to \log {}^{1}\!/\!{(1+6x)}$ is monotonically non-increasing:

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1}, \mathbb{P}^i_{Z_1}\big) \\
&\quad \le \frac{1}{6}\left(\log \frac{1}{1+6(\Delta_{0,0}+\Delta_{1,1})} + \log \frac{1}{1+6(\Delta_{1,0}+\Delta_{0,1})}\right)\mathbb{I}\{(P_1,Q_1)\in R_i\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Equation (14))} \\
&\quad = \frac{1}{6}\left(\log \frac{1}{1-36|\Delta_{0,0}+\Delta_{1,1}|\cdot|\Delta_{1,0}+\Delta_{0,1}|}\right)\mathbb{I}\{(P_1,Q_1)\in R_i\} \qquad \text{(by Equation (16))} \\
&\quad \le \frac{1}{6}\left(\log \frac{1}{1-16\varepsilon^2}\right)\mathbb{I}\{(P_1,Q_1)\in R_i\} \qquad \text{(by Equation (15))} \\
&\quad \le 3\varepsilon^2 \mathbb{I}\{(P_1,Q_1)\in R_i\}, \qquad (17)
\end{aligned}$$

where the last inequality can be verified analytically and holds for any $\varepsilon \in (0, {}^{1}\!/\!{10})$, and $\varepsilon = {}^{1}\!/\!{12\lceil T^{1/4}\rceil} \le {}^{1}\!/\!{10}$.

The other terms in Equation (11) can be handled similarly: $\mathcal{A}$ is a deterministic algorithm, thus for any time s and fixed feedback history $Z_1, \dots Z_{s-1}$, it holds that (P_s, Q_s) is a fixed element of $[0,1]^2$:

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}(\mathbb{P}^0_{Z_s|Z_1,\dots,Z_{s-1}}, \mathbb{P}^i_{Z_s|Z_1,\dots,Z_{s-1}}) \\
&= \sum_{z\in\{0,1\}^2} \log \tfrac{\mathbb{P}^0[Z_s=z|Z_1,\dots,Z_{s-1}]}{\mathbb{P}^i[Z_s=z|Z_1,\dots,Z_{s-1}]} \mathbb{P}^0\left[Z_1 = z \mid Z_1,\dots,Z_{s-1}\right] \mathbb{I}\{(P_s,Q_s)\in R_i \mid Z_1,\dots,Z_{s-1}\} \\
&= \sum_{z\in\{0,1\}^2} \log \tfrac{\mathbb{P}^0[Z_s=z|(P_s,Q_s)\in R_i]}{\mathbb{P}^i[Z_s=z|(P_s,Q_s)\in R_i]} \mathbb{P}^0\left[Z_s = z \mid (P_s,Q_s)\in R_i\right] \mathbb{I}\{(P_s,Q_s)\in R_i \mid Z_1,\dots,Z_{s-1}\}.
\end{aligned}$$

The same calculations we carried over for $s = 1$ can be repeated for the generic s, yielding the same bound of $3\varepsilon^2$:

$$\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s|Z_{1:s_1}}, \mathbb{P}^i_{Z_s|Z_1,\dots,Z_{s-1}}\big) \le 3\varepsilon^2 \mathbb{I}\{(P_s,Q_s)\in R_i \mid Z_1,\dots,Z_{s-1}\} \qquad (18)$$

Plugging Equation (17) and Equation (18) into Equation (11), we get the desired bound on the total variation distance:

$$\big\|\mathbb{P}^0_{(Z_1,\dots,Z_t)} - \mathbb{P}^k_{(Z_1,\dots,Z_t)}\big\|_{\mathrm{TV}} \le 2\varepsilon\sqrt{\mathbb{E}^0\left[N_{t-1}(i)\right]} \le 2\varepsilon\sqrt{\mathbb{E}^0\left[N_T(i)\right]}. \qquad (19)$$

Plugging Equation (19) into Equation (10) yields the Claim. ■

4. A $T^{3/4}$ Upper Bound: One Bit and Two Prices

In this section, we introduce our algorithm, Blind-Exp3, for the one-bit feedback setting against a σ-smooth adaptive adversary that achieves a bound on the regret of order $T^{3/4}$, up to logarithmic terms. A key technique that we use is a Monte Carlo estimation procedure $\widehat{\mathrm{GFT}}$ (see pseudocode for details) that allows us to estimate the expected gain from trade $\mathbb{E}\big[\mathrm{GFT}(p, S_t, B_t)\big]$ of a price p, by posting *two* different prices $(\hat{p}, \hat{q})$ and receiving *one* bit of feedback.

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

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

Learning algorithm with 1-bit feedback and two prices: Blind-Exp3

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

Lemma 2 (Lemma 1 of Azar et al. (2022)) *Fix any agents' valuations $(s, b) \in [0, 1]^2$. For any price $p \in [0, 1]$, it holds that $\widehat{\mathrm{GFT}}(p)$ is an unbiased estimator of $\mathrm{GFT}(p)$, i.e., $\mathbb{E}\left[\widehat{\mathrm{GFT}}(p)\right] = \mathrm{GFT}(p)$, where the expectation is with respect to the randomness of the estimation procedure.*

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

Theorem 5 *Consider the problem of repeated bilateral trade against a σ-smooth adaptive adversary in the one-bit feedback model, for any $\sigma \in (0, 1]$. If we run* Blind-Exp3 *with exploration rate $\gamma \in (0, 1)$, learning rate $\eta > 0$, and the uniform K-grid G such that $\frac{2\eta K}{\gamma} \le 1$, then, for each time horizon $T \in \mathbb{N}$, we have that*

$$R_T(\text{Blind-Exp3}) \le \frac{\ln K}{\eta} + \left(\gamma + \eta \frac{K}{\gamma} + \frac{1}{\sigma K}\right) T.$$

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

Proof The analysis of Blind-Exp3 needs to carefully take into account many sources of randomness: the internal randomness of the algorithm, of the estimation procedures and of the σ-smooth distributions of the adversary. Note, moreover, that the adversary is non-oblivious, so the choice of the distribution (S_t, B_t) depends on all the realizations of the past randomization. Fix any exploration rate $\gamma \in (0,1)$, learning rate $\eta > 0$ and number of grid points $K \in \mathbb{N}$ such that $2\eta K/\gamma \leq 1$. Fix also any time horizon $T \in \mathbb{N}$. In the following, we use the random variables (P_t, Q_t) to denote the randomized prices posted by the algorithm at time t.

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

$$\begin{aligned}
\ln\left(\frac{W_{T+1}}{W_1}\right) &= \ln\left(\prod_{t=1}^{T}\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T}\ln\left(\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T}\ln\left(\sum_{i\in[K]}\pi_t(i)\exp\left(\eta\hat{r}_t(i)\right)\right)\\
&\leq \sum_{t=1}^{T}\ln\left(1 + \eta\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i) + \eta^2\sum_{i\in[K]}\pi_t(i)\big(\hat{r}_t(i)\big)^2\right)\\
&\leq \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i) + \eta^2\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\big(\hat{r}_t(i)\big)^2 \qquad\qquad (\text{using } \hat{r}_t(i) \leq \tfrac{K}{\gamma})\\
&\leq \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i)\left[1 + \eta\frac{K}{\gamma}\right]. \qquad (20)
\end{aligned}$$

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

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

Inequality 20 is the pivot of our analysis, as we construct upper and lower bounds to its two extremes. We start from its first term, take the expectation with respect to the whole randomness of the process and consider any price g_i in the grid G:

$$\begin{aligned}
\mathbb{E}\left[\ln\left(\frac{W_{T+1}}{W_1}\right)\right] &= \mathbb{E}\left[\ln\left(W_{T+1}\right)\right] - \ln K \geq \mathbb{E}\left[\ln\left(w_{T+1}(i)\right)\right] - \ln K\\
&= \eta\sum_{t=1}^{T}\mathbb{E}\left[\hat{r}_t(i)\right] - \ln K = \eta\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{GFT}_t(g_i)\right] - \ln K. \qquad (21)
\end{aligned}$$

The only delicate passage of the previous formula is the last equality, where we used that $\mathbb{E}\left[\hat{r}_t(i)\right] = \mathbb{E}\left[\mathrm{GFT}_t(g_i)\right]$. To see why the latter holds, consider the filtration $\{\mathcal{F}_t\}_t$ relative to the story of the process: $\mathcal{F}_t$ is the σ-algebra generated by all the random variables involved in the process up to time

t (excluded). Moreover, let $\mathcal{E}_t^i$ be the event that at round t the coin toss results in Heads and the price selected u.a.r. for exploration is g_i. We have the following:

$$\begin{aligned}
\mathbb{E}\left[\hat{r}_t(i) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\hat{r}_t(i) \mid \mathcal{F}_t\right] && \hat{r}_t(i) = \mathbb{I}\{\mathcal{E}_t^i\}\hat{r}_t(i)\\
&= \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\mathbb{E}\left[\hat{r}_t(i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Law of total exp.}\\
&= \frac{K}{\gamma}\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\mathbb{E}\left[\widehat{\mathrm{GFT}}_t(g_i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Def. of } \hat{r}_t(i)\\
&= \frac{K}{\gamma}\mathbb{P}[\mathcal{E}_t^i \mid \mathcal{F}_t]\mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right] && \text{Lemma 2 and } (S_t, B_t) \text{ indep. of } \mathcal{E}_t^i\\
&= \mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right].
\end{aligned}$$

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

Let's go back to Equation (20) and focus on the last term. Conditioning with respect to $\mathcal{F}_t$:

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

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

$$\mathbb{E}\left[\mathrm{GFT}_t(P_t, Q_t)\right] \geq (1-\gamma)\sum_{i\in[K]}\mathbb{E}\left[\pi_t(i)\mathrm{GFT}_t(g_i)\right] = (1-\gamma)\sum_{i\in[K]}\mathbb{E}\left[\pi_t(i)\hat{r}_t(i)\right], \tag{22}$$

where the first inequality follows from the fact that with probability $1-\gamma$ the learner at time t chooses exploitation and thus posts a price in the grid G according to distribution π_t. We can plug Equation (21) and Equation (22) into Equation (20) to obtain the following:

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

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

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

The argument so far holds for any adaptive adversary $\mathcal{S}$ and any choice of price on the grid g_i. This, together with the discretization result Claim 1 gives the desired bound:

$$R_T(\text{Blind-Exp3}) \leq \frac{\ln K}{\eta} + \left(\gamma + \eta\frac{K}{\gamma} + \frac{1}{\sigma K}\right)T.$$

■

The dependence of the regret rate of Blind-Exp3 on the (unknown) smoothness parameter is possibly suboptimal, given the lower bound in Theorem 4, and exponentially worse than the one provided by Continuous-Price Hedge in full feedback. Continuous-Price Hedge crucially relies on the full feedback received by the learner to work, that is used to compute a weight for the counterfactual performance *of each price* p. In partial feedback, it is not clear how to estimate these (uncountably many) weights, thus making it hard to extend such continuous method beyond full feedback.

5. Conclusions and Open Problems

In this paper, we initiated the study of σ-smooth adversaries in online learning for pricing problems. Focusing on the repeated bilateral trade problem, we proved that a single bit of feedback is sufficient to achieve sublinear regret, pushing the boundary of learnability beyond the i.i.d. setting. We hope that the smoothed adversarial approach will find more applications to learning pricing strategies that cannot otherwise be efficiently learned in the adversarial model under partial feedback.

The surprising minimax regret regime of $T^{3/4}$ surpasses the $\sqrt{T}$ vs. $T^{2/3}$ dichotomy observed in other partial feedback models (e.g., partial monitoring and feedback graph), and motivates the intriguing question of whether techniques based on the generalized information ratio (Lattimore and Szepesvári, 2019) could be used to define a unified algorithmic tool in our framework and, more generally, to analyze online problems in digital markets.

Finally, we remark that there is still a gap in our understanding of the precise impact of the smoothness parameter on the minimax regret rate for the partial information feedback model. We leave this as an open question for future research.

Acknowledgments

NCB, RC, FF, and SL are partially supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

NCB and RC are also partially supported by the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), funded by the NextGenerationEU program within the PNRR scheme, and the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability).

FF and SL are also partially supported by ERC Advanced Grant 788893 AMDROMA and PNRR MUR project IR0000013-SoBigData.it.

TC gratefully acknowledges the support of the University of Ottawa through grant GR002837 (Start-Up Funds) and that of the Natural Sciences and Engineering Research Council of Canada (NSERC) through grants RGPIN-2023-03688 (Discovery Grants Program) and DGECR2023-00208 (Discovery Grants Program, DGECR - Discovery Launch Supplement).

References

Gagan Aggarwal, Ashwinkumar Badanidiyuru, Paul Dütting, and Federico Fusco. Selling joint ads: A regret minimization perspective. In *EC*. ACM, 2024.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 23–35. JMLR.org, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 454–468. Springer, 2007.

Yossi Azar, Amos Fiat, and Federico Fusco. An α-regret analysis of adversarial bilateral trade. In *NeurIPS*, 2022.

Moshe Babaioff, Amitai Frey, and Noam Nisan. Learning to maximize gains from trade in small markets. In *EC*. ACM, 2024.

François Bachoc, Nicolò Cesa-Bianchi, Tommaso Cesari, and Roberto Colomboni. Fair online bilateral trade. *arXiv preprint arXiv:2405.13919*, 2024a.

François Bachoc, Tommaso Cesari, and Roberto Colomboni. A contextual online learning theory of brokerage. *arXiv preprint arXiv:2407.01566*, 2024b.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. No-regret learning in bilateral trade via global budget balance. In *STOC*, pages 247–258. ACM, 2024.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons: New York, 1995.

Adam Block and Max Simchowitz. Efficient and near-optimal smoothed online learning for generalized linear functions. In *NeurIPS*, 2022.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 1716–1786. PMLR, 2022.

Adam Block, Max Simchowitz, and Russ Tedrake. Smoothed online learning for prediction in piecewise affine systems. *arXiv preprint arXiv:2301.11187*, 2023.

Liad Blumrosen and Shahar Dobzinski. Reallocation mechanisms. In *EC*, page 617. ACM, 2014.

Natasa Bolic, Tommaso Cesari, and Roberto Colomboni. An online learning theory of brokerage. In *AAMAS*, pages 216–224. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *EC*, pages 289–309. ACM, 2021.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Repeated bilateral trade against a smoothed adversary. In *COLT*, volume 195 of *Proceedings of Machine Learning Research*, 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, 2024.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. In *STOC*, pages 225–236. ACM, 2024.

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

Tommaso Cesari and Roberto Colomboni. Trading volume maximization with online learning. *arXiv preprint arXiv:2405.13102*, 2024.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 1216–1259. PMLR, 2021.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT*, volume 23 of *JMLR Proceedings*, pages 6.1–6.20. JMLR.org, 2012.

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

Naveen Durvasula, Nika Haghtalab, and Manolis Zampetakis. Smoothed analysis of online non-parametric auctions. In *EC*, pages 540–560. ACM, 2023.

Paul Dütting, Federico Fusco, Philip Lazos, Stefano Leonardi, and Rebecca Reiffenhäuser. Efficient two-sided markets with limited information. In *STOC*, pages 1452–1465. ACM, 2021.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Solenne Gaucher, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Vianney Perchet. Feature-based online bilateral trade. *arXiv preprint arXiv:2405.18183*, 2024.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *NeurIPS*, 2020.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *FOCS*, pages 942–953. IEEE, 2021.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for smoothed adversaries. In *NeurIPS*, 2022.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80:165–188, 2010.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *NeurIPS*, pages 2231–2241, 2018.

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

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS*, pages 697–704, 2004.

Walid Krichene, Maximilian Balandat, Claire J. Tomlin, and Alexandre M. Bayen. The hedge algorithm on a continuum. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 824–832. JMLR.org, 2015.

Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 1547–1575. PMLR, 2022.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 2111–2139. PMLR, 2019.

Odalric-Ambrym Maillard and Rémi Munos. Online learning in adversarial lipschitz environments. In *ECML/PKDD (2)*, volume 6322 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2010.

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

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. Greedy algorithm almost dominates in smoothed contextual bandits. *arXiv preprint arXiv:2005.10624*, 2020.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, pages 1764–1772, 2011.

Sarah Sachs, Hédi Hadiji, Tim van Erven, and Cristóbal Guzmán. Between stochastic and adversarial online convex optimization: Improved regret bounds via smoothness. In *NeurIPS*, 2022.

Vidyashankar Sivakumar, Zhiwei Steven Wu, and Arindam Banerjee. Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9026–9035. PMLR, 2020.

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 20253–20277. PMLR, 2022.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1593–1601. JMLR.org, 2014.

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

Appendix A. An Improved Analysis of Continuous Hedge

In what follows, we denote with $\mathcal{B}_{[0,1]}$, respectively $\mathcal{B}_{[0,+\infty]}$, the Borel σ-algebra of $[0,1]$, respectively $[0,+\infty]$, while $\mathcal{B}$ stands for the Borel σ-algebra of $\mathbb{R}$. For any any measurable function $g\colon [0,1] \to \mathbb{R}$, we denote with $\|g\|_1$ the integral with respect the Lebesgue measure of $|g|$ on $[0,1]$.

The following result implies directly theoretical guarantees for Hedge. We state the theorem in an abstract way to highlight that its claims are really about the properties of some stochastic processes rather than specific online learning protocols.

Theorem 6 *Let $(\mathcal{Y},\mathcal{E}_{\mathcal{Y}})$ be a measurable space. Let $\rho\colon [0,1]\times\mathcal{Y}\to[0,1]$ be a $(\mathcal{E}_{\mathcal{Y}}\otimes\mathcal{B}_{[0,1]})/\mathcal{B}_{[0,1]}$-measurable function. Let $(X_t,Y_t)_{t\in\mathbb{N}}$ be a $[0,1]\times\mathcal{Y}$-valued stochastic process. For any $t\in\mathbb{N}$, let $\mathcal{H}_t=\sigma(X_1,Y_1,\dots,X_{t-1},Y_{t-1})$ be the σ-algebra generated by the history up to the end of time $t-1$ (with the understanding that $\mathcal{H}_1=\sigma\big(\{\varnothing\}\big)$). Let $M\ge 2$ and $\eta\in(0,1)$. Assume that:*

- *For any $t\in\mathbb{N}$, the conditional law $\mathbb{P}_{X_t\mid\mathcal{H}_t}$ of X_t given $\mathcal{H}_t$ admits as a density (w.r.t. the Lebesgue measure on $[0,1]$) the (random) function $f_t(\cdot)=\frac{\sum_{s=1}^{t-1}\exp\big(\eta\rho(\cdot,Y_s)\big)}{\int_{[0,1]}\sum_{s=1}^{t-1}\exp\big(\eta\rho(x,Y_s)\big)\mathrm{d}x}$ (for $t=1$, $f_1=\mathbb{I}_{[0,1]}$) .*
- *For any $t\in\mathbb{N}$, the two random variables X_t and Y_t are conditionally independent given $\mathcal{H}_t$.*
- *For any $t\in\mathbb{N}$, the function $[0,1]\to[0,1],\ x\mapsto\mathbb{E}\big[\rho(x,Y_t)\big]$ is M-Lipschitz.*

Then, for any $T\in\mathbb{N}$,

$$\max_{x\in[0,1]}\mathbb{E}\left[\sum_{t=1}^{T}\rho(x,Y_t)\right]-\mathbb{E}\left[\sum_{t=1}^{T}\rho(X_t,Y_t)\right]\le\frac{1}{\eta}\ln\left(\frac{\eta TM}{1-e^{-\eta T}}\right)+(e-2)\eta T\ .$$

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

$$\max_{x\in[0,1]}\mathbb{E}\left[\sum_{t=1}^{T}\rho(x,Y_t)\right]-\mathbb{E}\left[\sum_{t=1}^{T}\rho(X_t,Y_t)\right]\le\sqrt{(e-2)T\ln(2T)}\cdot\left(\frac{5}{2}+\frac{\ln(M)}{\ln(2T)}\right)\ .$$

Proof Define $W_1(x)=1$ for all $x\in[0,1]$ and, for each $t\in\mathbb{N}$, define by induction $W_{t+1}(\cdot)=W_t(\cdot)\exp(\eta\rho(\cdot,Y_t))$. Then, denoting for any measurable function $g\colon[0,1]\to\mathbb{R}$, the integral with respect the Lebesgue measure of $|g|$ on $[0,1]$ by $\|g\|_1$, we have

$$\begin{aligned}
\ln\big(\|W_{T+1}\|_1\big) &= \ln\left(\prod_{t=1}^{T}\frac{\|W_{t+1}\|_1}{\|W_t\|_1}\right) = \sum_{t=1}^{T}\ln\left(\int_{[0,1]}\exp\big(\eta\rho(x,Y_t)\big)f_t(x)\,\mathrm{d}x\right) \\
&\le \sum_{t=1}^{T}\ln\left(\int_{[0,1]}\Big(1+\eta\rho(x,Y_t)+(e-2)\eta^2\big(\rho(x,Y_t)\big)^2\Big)f_t(x)\,\mathrm{d}x\right) \\
&= \sum_{t=1}^{T}\ln\left(1+\int_{[0,1]}\Big(\eta\rho(x,Y_t)+(e-2)\eta^2\big(\rho(x,Y_t)\big)^2\Big)f_t(x)\,\mathrm{d}x\right) \\
&\le \eta\sum_{t=1}^{T}\int_{[0,1]}\rho(x,Y_t)f_t(x)\,\mathrm{d}x+(e-2)\eta^2\sum_{t=1}^{T}\int_{[0,1]}\big(\rho(x,Y_t)\big)^2 f_t(x)\,\mathrm{d}x \\
&\le \eta\sum_{t=1}^{T}\int_{[0,1]}\rho(x,Y_t)f_t(x)\,\mathrm{d}x+(e-2)\eta^2 T \\
&= \eta\sum_{t=1}^{T}\mathbb{E}\big[\rho(X_t,Y_t)\mid\sigma(Y_t,\mathcal{H}_t)\big]+(e-2)\eta^2 T\ ,
\end{aligned}$$

where the last equality follows from the generalized freezing lemma (Lemma 5) noticing that, for each $t\in[T]$, Φ_t defined for each Borel subset $A\subset[0,1]$ via $\Phi_t[A]=\int_A f_t(x)\,\mathrm{d}x$ is a regular conditional probability for $\mathbb{P}_{X_t|\mathcal{H}_t}$ and $\int_{[0,1]}\rho(x,Y_t)f_t(x)\,\mathrm{d}x=\int_{[0,1]}\rho(x,Y_t)\,\mathrm{d}\Phi_t(x)$. Hence, using the tower rule,

$$\mathbb{E}\left[\ln\big(\|W_{T+1}\|_1\big)\right]\le\eta\mathbb{E}\left[\sum_{t=1}^{T}\rho(X_t,Y_t)\right]+(e-2)\eta^2 T\ .$$

On the other hand, let $x^\star\in[0,1]$ be a point belonging to $\operatorname{argmax}_{x\in[0,1]}\sum_{t=1}^{T}\mathbb{E}\big[\rho(x,Y_t)\big]$, which does exist due to the fact that this last sum, as a function of x, is MT-Lipschitz (hence continuous on the compact set $[0,1]$). Then, for any $x\in[0,1]$,

$$\sum_{t=1}^{T}\mathbb{E}\big[\rho(x^\star,Y_t)\big]-\sum_{t=1}^{T}\mathbb{E}\big[\rho(x,Y_t)\big]\le T\min\big(1,M|x-x^\star|\big)\ . \tag{23}$$

Let X be a uniform random variable on $[0,1]$ independent of $Y_1,\dots,Y_T$. It follows that

$$\begin{aligned}
\mathbb{E}\Big[\ln\big(\|W_{T+1}\|_1\big)\Big] &= \mathbb{E}\left[\ln\left(\int_{[0,1]}\exp\left(\eta\sum_{t=1}^T\rho(x,Y_t)\right)\,\mathrm{d}x\right)\right]\\
&= \mathbb{E}\left[\ln\mathbb{E}\left[\exp\left(\eta\sum_{t=1}^T\rho(X,Y_t)\right)\mid X\right]\right]\\
&\geq \ln\mathbb{E}\left[\exp\left(\mathbb{E}\left[\eta\sum_{t=1}^T\rho(X,Y_t)\mid(Y_1,\dots,Y_T)\right]\right)\right]\\
&= \ln\left(\int_{[0,1]}\exp\left(\mathbb{E}\left[\eta\sum_{t=1}^T\rho(x,Y_t)\right]\right)\,\mathrm{d}x\right)\\
&= \eta\sum_{t=1}^T\mathbb{E}\big[\rho(x^\star,Y_t)\big]+\ln\left(\int_{[0,1]}\exp\left(\eta\left(\sum_{t=1}^T\mathbb{E}\big[\rho(x,Y_t)\big]-\sum_{t=1}^T\mathbb{E}\big[\rho(x^\star,Y_t)\big]\right)\right)\,\mathrm{d}x\right)\\
&\geq \eta\sum_{t=1}^T\mathbb{E}\big[\rho(x^\star,Y_t)\big]+\ln\left(\int_{[0,1]}\exp\left(-\eta T\min(1,M|x-x^\star|)\right)\,\mathrm{d}x\right)=(\star)\ ,
\end{aligned}$$

where

- the second and the third equalities follow from the Freezing Lemma (see Lemma 4 in the appendix).
- the first inequality follows from the log-exp analogous of Minkowski's integral inequality, in the form of Corollary 2, with $(\mathcal{V},\mathcal{E}_\mathcal{V})=\big([0,1],\mathcal{B}_{[0,1]}\big)$, $(\mathcal{W},\mathcal{E}_\mathcal{W})=(\mathcal{Y}^T,\otimes^T\mathcal{E}_\mathcal{Y})$, $V=X$, $W=(Y_1,\dots,Y_T)$, and $g\colon[0,1]\times\mathcal{Y}^T\to[0,+\infty]$, $\big(x,(y_1,\dots,y_T)\big)\mapsto\eta\sum_{t=1}^T\rho(x,y_t)$.
- the last inequality follows from Eq. (23).

Now, if $x^\star\leq\frac{1}{2}$, then, for any $x\in\big[x^\star,x^\star+\frac{1}{M}\big]$ we have that

$$\min(1,M|x-x^\star|)=M|x-x^\star|$$

and then, recalling that $M\geq 2$,

$$\begin{aligned}
(\star)&\geq\eta\sum_{t=1}^T\mathbb{E}\big[\rho(x^\star,Y_t)\big]+\ln\left(\int_{\left[x^\star,x^\star+\frac{1}{M}\right]}\exp\left(-\eta T\min(1,M|x-x^\star|)\right)\,\mathrm{d}x\right)\\
&=\eta\sum_{t=1}^T\mathbb{E}\big[\rho(x^\star,Y_t)\big]+\ln\left(\frac{1-\exp\left(-\eta T\right)}{\eta TM}\right).
\end{aligned}$$

The case $x^\star>\frac{1}{2}$ can be worked out analogously obtaining the same result. In any case, putting everything together, we get

$$\eta\mathbb{E}\left[\sum_{t=1}^T\rho(X_t,Y_t)\right]+(e-2)\eta^2T\geq\eta\mathbb{E}\left[\sum_{t=1}^T\rho(x^\star,Y_t)\right]+\ln\left(\frac{1-\exp\left(-\eta T\right)}{\eta TM}\right)$$

```
Online Protocol: $\mathcal{X}$-Armed Experts
Instance parameters: Known action space $\mathcal{X}$, unknown environment's action space $\mathcal{Y}$, unknown reward function $\rho\colon \mathcal{X} \times \mathcal{Y} \to [0,1]$
    for time $t = 1, 2, \ldots$ do
        The environment secretly selects an action $Y_t \in \mathcal{Y}$ (possibly at random)
        The learner secretly selects an action $X_t \in \mathcal{X}$ (possibly at random)
        The learner gains reward $\rho(X_t, Y_t)$
        $X_t$ is revealed to the environment and $G_t(\cdot) = \rho(\cdot, Y_t)$ is revealed to the learner
```

```
Learning algorithm with full feedback: Hedge for $[0,1]$-Armed Experts
Input: $\eta \in (0,1)$
Initialization: Initialize $W_1(x) = 1$, for all $x \in [0,1]$
    for time $t = 1, 2, \ldots$ do
        Play $X_t \sim \mu_t$, where $\mu_t$ is a distribution with density defined, for all $x \in [0,1]$, by $f_t(x) = \frac{W_t(x)}{\lVert W_t \rVert_1}$
        Update $W_{t+1}(x) = W_t(x) \cdot \exp(\eta G_t(x))$, for each $x \in [0,1]$
```

which, dividing by η and rearranging, becomes

$$\mathbb{E}\left[\sum_{t=1}^{T} \rho(x^\star, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \le \frac{1}{\eta} \ln\left(\frac{\eta T M}{1 - e^{-\eta T}}\right) + \eta(e-2)T$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T} \rho(x^\star, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \le \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln(M)}{\ln(2T)}\right) .$$

■

In the same spirit of the previous theorem, we now obtain an immediate corollary that provides theoretical guarantees for Hedge run for $[0,1]$-armed experts (see the general online protocol of $\mathcal{X}$-armed experts and the corresponding definition of Hedge when $\mathcal{X} = [0,1]$) with Lipschitz *expected* rewards.

Corollary 1 *If there exists $M \ge 2$ such that, for all $t \in \mathbb{N}$, $x \mapsto \mathbb{E}[G_t(x)]$ is an M-Lipschitz function, then, for any time horizon $T \in \mathbb{N}$, the regret of Hedge for $[0,1]$-Armed Experts run with parameter $\eta \in (0,1)$ is*[¶]

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \le \frac{1}{\eta} \ln\left(\frac{\eta T M}{1 - e^{-\eta T}}\right) + (e-2)\eta T .$$

[¶] Formally, we are assuming that $(\mathcal{Y}, \mathcal{E}_{\mathcal{Y}})$ is a measurable space; for all $t \in \mathbb{N}$, Y_t is chosen in a measurable way as a function of the information available to the environment at the beginning of time t, including its possible randomization; and ρ is a $(\mathcal{B}_{[0,1]} \otimes \mathcal{E}_{\mathcal{Y}})/\mathcal{B}_{[0,1]}$-measurable function.

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

$$\max_{x\in[0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \rho(x, Y_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \rho(X_t, Y_t)\right] \le \sqrt{(e-2)T\ln(2T)} \cdot \left(\frac{5}{2} + \frac{\ln(M)}{\ln(2T)}\right) .$$

We remark that Hedge achieves an extremely mild dependence on M —disappearing completely if T is larger than M— without requiring the knowledge of M to tune the parameter η.

Finally, we highlight a key feature of our Theorem 6 and Corollary 1: they only assume that *expected* rewards are Lipschitz. This is in contrast with the classic assumption that the rewards themselves are Lipschitz. This seemingly small difference entails significant technical issues in the analysis that we bypassed by proving two general lemmas (a log-exp analogous of Minkowski's integral inequality, Lemma 3, and a generalized freezing lemma Lemma 5) that we believe are of independent interest. Besides the novelty of the techniques, having results in settings where rewards are only required to be Lipschitz in expectation unlocks the possibility of using Hedge in problems like bilateral trade, where the reward functions are not even continuous.

Appendix B. A Log-Exp Minkovski's Integral Inequality

In this section, we prove a log-exp analogous to Minkowski's integral inequality. In its original form, Minkowski's inequality states that

$$\int_{\mathcal{V}} \left(\int_{\mathcal{W}} \big(g(v,w)\big)^p \,\mathrm{d}\mu_{\mathcal{W}}(w)\right)^{1/p} \mathrm{d}\mu_{\mathcal{V}}(v) \ge \left(\int_{\mathcal{W}} \left(\int_{\mathcal{V}} g(v,w)\,\mathrm{d}\mu_{\mathcal{V}}(v)\right)^{p} \mathrm{d}\mu_{\mathcal{W}}(w)\right)^{1/p} ,$$

where $p \ge 1$, $(\mathcal{V}, \mathcal{E}_{\mathcal{V}}, \mu_{\mathcal{V}})$ and $(\mathcal{W}, \mathcal{E}_{\mathcal{W}}, \mu_{\mathcal{W}})$ are two σ-finite measure spaces[‖] and $g \colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$ is a measurable function.

We now prove a log-exp analogous of Minkowski's Integral Inequality. To the best of our knowledge, the following result has not been previously presented in the literature, and we believe it may be of independent interest.

We recall that $\mathcal{B}_{[0,+\infty]}$ denotes the Borel σ-algebra of $[0, +\infty]$.

Lemma 3 (Log-Exp Minkowski's Integral Inequality) *Let* $(\mathcal{V}, \mathcal{E}_{\mathcal{V}}, \mu_{\mathcal{V}})$ *and* $(\mathcal{W}, \mathcal{E}_{\mathcal{W}}, \mu_{\mathcal{W}})$ *be two* σ*-finite measure spaces such that* $\mu_{\mathcal{V}}[\mathcal{V}] \neq 0 \neq \mu_{\mathcal{W}}[\mathcal{W}]$*. Let* $g \colon \mathcal{V} \times \mathcal{W} \to [0, +\infty]$ *be a* $(\mathcal{E}_{\mathcal{V}} \otimes \mathcal{E}_{\mathcal{Y}})/\mathcal{B}_{[0,+\infty]}$ *measurable function. Then (with the understanding that* $0 \cdot \infty = 0$*):*

$$\int_{\mathcal{V}} \ln\left(\int_{\mathcal{W}} \exp\big(g(v,w)\big)\,\mathrm{d}\mu_{\mathcal{W}}(w)\right) \mathrm{d}\mu_{\mathcal{V}}(v) \ge \mu_{\mathcal{V}}[\mathcal{V}] \ln\left(\int_{\mathcal{W}} \exp\left(\int_{\mathcal{V}} g(v,w)\,\mathrm{d}\mu_{\mathcal{V}}(v)\right) \mathrm{d}\mu_{\mathcal{W}}(w)\right) .$$

Proof Assume first that both $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{W}}$ are finite measures. Let $L^\infty(\mathcal{W})$ be the set of bounded $\mathcal{E}_{\mathcal{W}}/\mathcal{B}$-measurable functions. Define

$$\Phi \colon L^\infty(\mathcal{W}) \to \mathbb{R} \qquad f \mapsto \ln \int_{\mathcal{W}} \exp\big(f(w)\big)\,\mathrm{d}\mu_{\mathcal{W}}(w).$$

[‖] We recall that a measure space $(\mathcal{A}, \mathcal{E}_{\mathcal{A}}, \mu_{\mathcal{A}})$ is σ-finite if there exist a countable family $A_1, A_2, \dots \in \mathcal{E}_{\mathcal{A}}$ such that $\mu_{\mathcal{A}}(A_k) < +\infty$ for all $k \in \mathbb{N}$ and $\bigcup_{k\in\mathbb{N}} A_k = \mathcal{A}$.

Notice that Φ is convex. In fact, for any $f_1, f_2 \in L^\infty(\mathcal{W})$ and any $\lambda \in (0,1)$, we have

$$\begin{aligned}
\Phi\big((1-\lambda)f_1 + \lambda f_2\big) &= \ln \int_{\mathcal{W}} \exp\big((1-\lambda)f_1(w) + \lambda f_2(w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w) \\
&= \ln \int_{\mathcal{W}} \Big(\exp\big(f_1(w)\big)\Big)^{1-\lambda} \Big(\exp\big(f_2(w)\big)\Big)^{\lambda} \,\mathrm{d}\mu_{\mathcal{W}}(w) \\
&\le \ln \left(\left(\int_{\mathcal{W}} \exp\big(f_1(w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w) \right)^{1-\lambda} \left(\int_{\mathcal{W}} \exp\left(f_2(w)\right) \,\mathrm{d}\mu_{\mathcal{W}}(w) \right)^{\lambda} \right) \\
&= (1-\lambda) \ln \left(\int_{\mathcal{W}} \exp\big(f_1(w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w) \right) + \lambda \ln \left(\int_{\mathcal{W}} \exp\left(f_2(w)\right) \,\mathrm{d}\mu_{\mathcal{W}}(w) \right) \\
&= (1-\lambda)\Phi(f_1) + \lambda \Phi(f_2) \;,
\end{aligned}$$

where the inequality follows from Hölder inequality with $p = \frac{1}{1-\lambda}$ and $q = \frac{1}{\lambda}$, the monotonicity of the integral, and the fact that $\ln$ is monotonically increasing. Now, notice that Φ is differentiable from the Banach space $(L^\infty(\mathcal{W}), \|\cdot\|_\infty)$ to $\mathbb{R}$ (where $\|f\|_\infty = \sup_{w \in \mathcal{W}} |f(w)|$), and for each $f \in L^\infty(\mathcal{W})$ the differential of Φ at any $f \in L^\infty(\mathcal{W})$ satisfies

$$\mathrm{d}\Phi(f)(h) = \frac{\int_{\mathcal{W}} \exp\big(f(w)\big) h(w) \,\mathrm{d}\mu_{\mathcal{W}}(w)}{\int_{\mathcal{W}} \exp\big(f(w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w)} \;, \qquad \text{for each } h \in L^\infty(\mathcal{W}) \;.$$

The convexity and the differentiability of Φ together implies that for any $f_1, f_2 \in L^\infty(\mathcal{W})$ it holds that

$$\Phi(f_1) \ge \Phi(f_2) + \mathrm{d}\Phi(f_2)(f_1 - f_2) \;.$$

Now, if $g \in L^\infty(\mathcal{V} \times \mathcal{W})$ (i.e., if g is bounded and $(\mathcal{E}_{\mathcal{V}} \otimes \mathcal{E}_{\mathcal{W}})/\mathcal{B}_{[0,+\infty]}$ measurable), define

$$G \colon \mathcal{V} \to L^\infty(\mathcal{W}) \;, \qquad v \mapsto g(v, \cdot) \;,$$

and define also

$$f_2(\cdot) = \int_{\mathcal{V}} g(v', \cdot) \,\mathrm{d}\mu_{\mathcal{V}}(v') \in L^\infty(\mathcal{W}) \;.$$

It follows that, for any $v \in \mathcal{V}$,

$$\begin{aligned}
\ln \int_{\mathcal{W}} \exp\big(g(v,w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w) &= \ln \int_{\mathcal{W}} \exp\big(G(v)(w)\big) \,\mathrm{d}\mu_{\mathcal{W}}(w) = \Phi\big(G(v)\big) \\
&\ge \Phi(f_2) + \mathrm{d}\Phi(f_2)\big(G(v) - f_2\big) \\
&= \ln \left(\int_{\mathcal{W}} \exp\left(\int_{\mathcal{V}} g(v', w) \,\mathrm{d}\mu_{\mathcal{V}}(v') \right) \mathrm{d}\mu_{\mathcal{W}}(w) \right) \\
&\quad + \frac{\int_{\mathcal{W}} \Big(\exp\big(\int_{\mathcal{V}} g(v', w) \,\mathrm{d}\mu_{\mathcal{V}}(v')\big) \left(g(v,w) - \int_{\mathcal{V}} g(v', w) \,\mathrm{d}\mu_{\mathcal{V}}(v') \right) \Big) \,\mathrm{d}\mu_{\mathcal{W}}(w)}{\int_{\mathcal{W}} \exp\big(\int_{\mathcal{V}} g(v', w) \,\mathrm{d}\mu_{\mathcal{V}}(v')\big) \,\mathrm{d}\mu_{\mathcal{W}}(w)} \;.
\end{aligned}$$

Given that this last inequality holds for any $v \in \mathcal{V}$, we can integrate both sides with respect to $\mathrm{d}\mu_{\mathcal{V}}(v)$ and get

$$\begin{aligned}
&\int_{\mathcal{V}} \ln\left(\int_{\mathcal{W}} \exp\bigl(g(v,w)\bigr)\,\mathrm{d}\mu_{\mathcal{W}}(w)\right)\mathrm{d}\mu_{\mathcal{V}}(v)\\
&\quad\geq \mu_{\mathcal{V}}[\mathcal{V}]\ln\left(\int_{\mathcal{W}}\exp\left(\int_{\mathcal{V}} g(v',w)\,\mathrm{d}\mu_{\mathcal{V}}(v')\right)\mathrm{d}\mu_{\mathcal{W}}(w)\right)\\
&\qquad+\int_{\mathcal{V}}\frac{\int_{\mathcal{W}}\Bigl(\exp\bigl(\int_{\mathcal{V}} g(v',w)\,\mathrm{d}\mu_{\mathcal{V}}(v')\bigr)\,\bigl(g(v,w)-\int_{\mathcal{V}} g(v',w)\,\mathrm{d}\mu_{\mathcal{V}}(v')\bigr)\Bigr)\,\mathrm{d}\mu_{\mathcal{W}}(w)}{\int_{\mathcal{W}}\exp\bigl(\int_{\mathcal{V}} g(v',w)\,\mathrm{d}\mu_{\mathcal{V}}(v')\bigr)\,\mathrm{d}\mu_{\mathcal{W}}(w)}\,\mathrm{d}\mu_{\mathcal{V}}(v)\\
&\quad= \mu_{\mathcal{V}}[\mathcal{V}]\ln\left(\int_{\mathcal{W}}\exp\left(\int_{\mathcal{V}} g(v',w)\,\mathrm{d}\mu_{\mathcal{V}}(v')\right)\mathrm{d}\mu_{\mathcal{W}}(w)\right),
\end{aligned}$$

where the last equality follows from Fubini's theorem. Notice that we have proved the theorem under the assumption that $g \in L^\infty(\mathcal{V}\times\mathcal{W})$ and that $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{W}}$ are finite measures.

Now, if $g \notin L^\infty(\mathcal{V}\times\mathcal{W})$ but $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{W}}$ are finite, given that $g \geq 0$, we can find a sequence $(g_n)_{n\in\mathbb{N}} \subset L^\infty(\mathcal{V}\times\mathcal{W})$ such that $g_n \uparrow g$ pointwise, and obtain the conclusion from the monotone convergence theorem. If $\mu_{\mathcal{V}}[\mathcal{V}] = +\infty$ but $\mu_{\mathcal{W}}$ is finite, given that $\mu_{\mathcal{V}}$ is σ-finite, we can find a sequence $A_1 \subset A_2 \subset \dots$ such that $\bigcup_{n\in\mathbb{N}} A_n = \mathcal{V}$ and, for each $n\in\mathbb{N}$ it holds that $A_n \in \mathcal{E}_{\mathcal{V}}$ and $\mu_{\mathcal{V}}[A_n] < +\infty$ and apply the theorem to the restriction of $\mu_{\mathcal{V}}$ to A_n and let $n\to\infty$ to obtain the conclusion via the monotone convergence theorem. Finally, if $\mu_{\mathcal{W}}[\mathcal{W}] = +\infty$, given that $\mu_{\mathcal{W}}$ is σ-finite, we can find a sequence $B_1 \subset B_2 \subset \dots$ such that $\bigcup_{n\in\mathbb{N}} B_n = \mathcal{W}$ and, for each $n\in\mathbb{N}$ it holds that $B_n \in \mathcal{E}_{\mathcal{W}}$ and $\mu_{\mathcal{W}}[B_n] < +\infty$ and apply the theorem to the restriction of $\mu_{\mathcal{W}}$ to B_n and let $n\to\infty$ to obtain the conclusion via the monotone convergence theorem again. ■

As an immediate corollary of the previous lemma, we get the following.

Corollary 2 (Log-Exp Minkowski's Integral Inequality for probability measures) *Let $(\mathcal{V},\mathcal{E}_{\mathcal{V}})$ and $(\mathcal{W},\mathcal{E}_{\mathcal{W}})$ be two measurable spaces and let $g\colon \mathcal{V}\times\mathcal{W}\to[0,+\infty]$ be a $\mathcal{E}_{\mathcal{V}}\otimes\mathcal{E}_{\mathcal{W}}/\mathcal{B}_{[0,+\infty]}$-measurable function. Assume that V and W are an $\mathcal{V}$-valued and a $\mathcal{W}$-valued random variables, respectively, independent of each other. Then*

$$\mathbb{E}\Bigl[\ln\mathbb{E}\bigl[\exp\bigl(g(V,W)\bigr)\mid V\bigr]\Bigr] \geq \ln\mathbb{E}\Bigl[\exp\Bigl(\mathbb{E}\bigl[g(V,W)\mid W\bigr]\Bigr)\Bigr].$$

Appendix C. A Generalized Freezing Lemma

The classic "freezing lemma" (see, e.g., Cesari and Colomboni 2021, Lemma 8) states that the conditional expectation of a measurable function of two independent random variables given one of them can be computed as an expectation only with respect to the other random variable followed by a composition with the random variable in the conditioning.

Lemma 4 (The *freezing lemma*) *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. Let $(\mathcal{V},\mathcal{F}_{\mathcal{V}})$ and $(\mathcal{W},\mathcal{F}_{\mathcal{W}})$ be two measurable spaces. Let $f\colon\mathcal{V}\times\mathcal{W}\to[0,+\infty]$, $V:\Omega\to\mathcal{V}$, $W:\Omega\to\mathcal{W}$ be three measurable functions. If V and W are $\mathbb{P}$-independent, then*

$$\mathbb{E}\bigl[f(V,W)\mid V\bigr] = \Bigl[\mathbb{E}\bigl[f(v,W)\bigr]\Bigr]_{v=V}. \tag{24}$$

$\mathbb{P}$*-almost surely, where the right hand side is the composition*

$$\Big[\mathbb{E}\big[f(v,W)\big]\Big]_{v=V} = \Big(v \mapsto \mathbb{E}\big[f(v,W)\big]\Big) \circ V \ .$$

The freezing lemma is extremely useful in derivations as it allows one to isolate the random parts that are being averaged while keeping the others fixed. Unfortunately, the freezing lemma does not cover the case where the expectations are replaced with conditional expectation on some σ-algebra, which is often the case in online learning, where expectations and probabilities are typically intended as conditional on the history up to the present time. This problem cannot be solved by simply replacing expectations with conditional expectations everywhere because of the fact that versions of conditional expectations remain as such if changed on a probability-zero event, making the naive extension to the right-hand side of Eq. (24) not even well-defined. To aid us in giving a sound statement of such a generalization of the freezing lemma, we begin by recalling the definition of regular conditional probability.

Definition 7 (Regular conditional probability) *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. Let $(\mathcal{X},\mathcal{E}_\mathcal{X})$ be a measurable space. Let $X\colon \Omega \to \mathcal{X}$ be a $\mathcal{F}/\mathcal{E}_\mathcal{X}$-measurable. Let $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. We say that $\Phi\colon \mathcal{E}_\mathcal{X} \to [0,1]^\Omega$ is a regular conditional probability for $\mathbb{P}_{X\mid\mathcal{H}}$ if:*

- *For each $A \in \mathcal{E}_\mathcal{X}$, the function $\omega \mapsto \Phi[A](\omega)$ is $\mathcal{H}/\mathcal{B}_{[0,1]}$-measurable.*
- *For each $\omega \in \Omega$, the function $A \mapsto \Phi[A](\omega)$ is a probability measure.*
- *For each $A \in \mathcal{E}_\mathcal{X}$ and each $H \in \mathcal{H}$, it holds that $\mathbb{P}\big[H \cap \{X \in A\}\big] = \mathbb{E}\big[\mathbb{I}_H \Phi[A]\big]$.*

Notice that the first and the third bullet imply that $\Phi[A] = \mathbb{E}[\mathbb{I}_{X\in A} \mid \mathcal{H}]$ for each $A \in \mathcal{E}_\mathcal{X}$.

We can now state and prove a generalized version of the freezing lemma, which we believe may be of independent interest.

We recall that $\mathcal{B}_{[0,+\infty]}$ denotes the Borel σ-algebra of $[0,+\infty]$.

Lemma 5 (Generalized Freezing Lemma) *Let $(\mathcal{X},\mathcal{E}_\mathcal{X})$ and $(\mathcal{Y},\mathcal{E}_\mathcal{Y})$ be two measurable spaces. Let $g\colon \mathcal{X}\times\mathcal{Y} \to [0,\infty]$ be a $(\mathcal{E}_\mathcal{X}\otimes\mathcal{E}_\mathcal{Y})/\mathcal{B}_{[0,+\infty]}$- measurable function. Let $(\Omega,\mathcal{E},\mathbb{P})$ be a probability space and $\mathcal{F},\mathcal{G},\mathcal{H}$ be three sub-σ-algebras of $\mathcal{E}$. Let $X : \Omega \to \mathcal{X}$ be a $\mathcal{F}/\mathcal{E}_\mathcal{X}$-measurable random variable. Let $Y : \Omega \to \mathcal{Y}$ be a $\mathcal{G}/\mathcal{E}_\mathcal{Y}$-measurable random variable. Assume that $\mathcal{F}$ and $\mathcal{G}$ are $\mathbb{P}$-conditionally independent given $\mathcal{H}$. Assume that Φ is a regular conditional probability for $\mathbb{P}_{X\mid\mathcal{H}}$. Then*

$$\int_\mathcal{X} g(x,Y)\,\mathrm{d}\Phi(x) = \mathbb{E}\big[g(X,Y) \mid \sigma(\mathcal{G},\mathcal{H})\big] \ .$$

Proof First, notice that the random variable $\int_\mathcal{X} g\big(x,Y\big)\,\mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G},\mathcal{H})$-measurable. In fact, if $A \in \mathcal{E}_\mathcal{X}$ and $B \in \mathcal{E}_\mathcal{Y}$ we have

$$\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y)\,\mathrm{d}\Phi(x) = \Phi[A]\mathbb{I}_B(Y),$$

which implies that $\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y)\,\mathrm{d}\Phi(x)$, as a product of a $\mathcal{H}$-measurable function and a $\mathcal{G}$-measurable function is $\sigma(\mathcal{G},\mathcal{H})$-measurable. Now, consider the family

$$\mathcal{C} = \left\{ C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y} \mid \int_\mathcal{X} \mathbb{I}_C(x,Y)\,\mathrm{d}\Phi(x) \text{ is } \sigma(\mathcal{G},\mathcal{H})\text{-measurable} \right\} .$$

Notice that $\mathcal{X} \times \mathcal{Y} \in \mathcal{C}$, that $\mathcal{C}$ is closed under complementation and that if $(C_n)_{n\in\mathbb{N}} \subset \mathcal{C}$ is such that $C_1 \subset C_2 \subset \dots$ then $\bigcup_{n\in\mathbb{N}} C_n \in \mathcal{C}$. Hence, $\mathcal{C}$ is a λ-system which contains the π-system $\mathcal{D} = \{C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y} \mid \exists A \in \mathcal{E}_\mathcal{X}, \exists B \in \mathcal{E}_\mathcal{Y}, C = A \times B\}$. Hence, by the π-λ theorem (Billingsley, 1995, Theorem 3.2) it holds that $\sigma(\mathcal{D}) \subset \mathcal{C}$, and since $\sigma(\mathcal{D}) = \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ it holds that $\mathcal{C} = \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$. It follows that for each $C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ the random variable $\int_\mathcal{X} \mathbb{I}_C(x, Y) \,\mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable. By pointwise monotone increasing approximation via $\mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$-measurable simple functions[**], we get that the random variable $\int_\mathcal{X} g\bigl(x, Y\bigr) \,\mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable.

Now, pick $A \in \mathcal{E}_\mathcal{X}, B \in \mathcal{E}_\mathcal{Y}, G \in \mathcal{G}$ and $H \in \mathcal{H}$. Notice that

$$\begin{aligned}
\mathbb{E}\left[\int_\mathcal{X} \mathbb{I}_A(x)\mathbb{I}_B(Y) \,\mathrm{d}\Phi(x)\mathbb{I}_{G\cap H}\right] &= \mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\Phi[A]\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)} \mid \mathcal{H}\right] \Phi[A]\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)} \mid \mathcal{H}\right] \mathbb{E}[\mathbb{I}_{X\in A} \mid \mathcal{H}]\mathbb{I}_H\right] \\
(\mathcal{F} \text{ and } \mathcal{G} \text{ are conditionally independent given } \mathcal{H}) \qquad &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\mathbb{I}_{X\in A} \mid \mathcal{H}\right] \mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{I}_{G\cap(Y\in B)}\mathbb{I}_{X\in A}\mathbb{I}_H\right] \\
&= \mathbb{E}\left[\mathbb{I}_A(X)\mathbb{I}_B(Y)\mathbb{I}_{G\cap H}\right] \ .
\end{aligned}$$

Applying twice a π-λ argument as done above, we can prove that for each $C \in \mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$ and each $K \in \sigma(\mathcal{G}, \mathcal{H})$, it holds that

$$\mathbb{E}\left[\int_\mathcal{X} \mathbb{I}_C(x, Y) \,\mathrm{d}\Phi(x)\mathbb{I}_K\right] = \mathbb{E}\left[\mathbb{I}_C(X, Y)\mathbb{I}_K\right] \ .$$

Applying again a pointwise monotone approximation argument using $\mathcal{E}_\mathcal{X} \otimes \mathcal{E}_\mathcal{Y}$-measurable simple functions, we can prove that for each $K \in \sigma(\mathcal{G}, \mathcal{H})$ it holds that

$$\mathbb{E}\left[\int_\mathcal{X} g(x, Y) \,\mathrm{d}\Phi(x)\mathbb{I}_K\right] = \mathbb{E}\left[g(X, Y)\mathbb{I}_K\right] \ .$$

Given that we have already proved that the random variable $\int_\mathcal{X} g(x, Y) \,\mathrm{d}\Phi(x)$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable, the conclusion follows. ■

[**] We recall that simple functions are linear combinations of indicator functions.