Proceedings of Machine Learning Research vol 195:1–36, 2023 36th Annual Conference on Learning Theory

Repeated Bilateral Trade Against a Smoothed Adversary

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

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

Roberto Colomboni ROBERTO.COLOMBONI@UNIMI.IT
Università degli Studi di Milano, Milano, Italy and Istituto Italiano di Tecnologia, Genova, Italy

Federico Fusco FUSCOF@DIAG.UNIROMA1.IT
Stefano Leonardi LEONARDI@DIAG.IT
Sapienza Università di Roma, Roma, Italy

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study repeated bilateral trade where an adaptive σ-smooth adversary generates the valuations of sellers and buyers. We provide a complete characterization of the regret regimes for fixed-price mechanisms under different feedback models in the two cases where the learner can post either the same or different prices to buyers and sellers. We begin by showing that the minimax regret after T rounds is of order $\sqrt{T}$ in the full-feedback scenario. 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 main technical contribution of the paper.

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 respective quasi-linear utility. The solution to the problem consists in 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 that there are 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 considers the Bayesian version of the problem, where agents' valuations are drawn from some distribution and the efficiency is evaluated in expectation with respect to the valuations' randomness. There are many incentive compatible 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 harder 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., 2023), 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 will pay for the item. We assume both valuations are generated by an adversary. 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 $\mathcal{U} \coloneqq \big\{(x,y) \in [0,1]^2 \mid x \le y\big\}$. 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 at time t:

$$\mathrm{GFT}_t(p,q) \coloneqq \big((b_t - q) + (p - s_t)\big) \cdot \mathbb{I}\{s_t \le p \le q \le b_t\}^{*}.$$

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 $\mathrm{GFT}_t(p,q)$. We consider the following learning protocol (the definition of σ-smoothness is recalled below).

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

for time $t = 1, 2, \ldots$ **do**
- The adversary privately chooses the σ-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

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:

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

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 of 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

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

class we define 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 different settings we consider, i.e., with how many prices are posted and what feedback is available—see below.

Smoothed analysis of algorithms, originally 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., 2023). 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 we require that the mechanism 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; i.e., 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.*

We consider three natural types of feedback models, 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. (By Observation 1, in this model, there is no reason to post two distinct prices, as all the relevant information is revealed anyway.)
- **Two-bit feedback.** $z_t = \big(\mathbb{I}\{s_t \leq p_t\}, \mathbb{I}\{q_t \leq b_t\}\big)$: The learner observes separately if the two agents accept the prices offered to each of them.

	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 6

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. (2023, 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).

- **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.

1.1. Overview of 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 design the Price-Hedge algorithm, 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. (2023, 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 6). The same rate was already obtained by the Scouting Blindits algorithm in Cesa-Bianchi et al. (2023), but only under the additional assumption 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 an open problem in (Cesa-Bianchi et al., 2023, Section 7).
- We prove that no algorithm can achieve worst-case sublinear regret when the platform is allowed 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. (2023, 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

[†]Although our decision space is two-dimensional, one can see that, in a bandit feedback with a smooth adversary, a regret of order $T^{2/3}$ can be obtained by running an optimal bandit algorithm (e.g., MOSS Audibert and Bubeck 2009, whose upper bound on the regret is of order $\sqrt{KT}$) on a discretization of $K = \Theta(T^{1/3})$ equispaced prices on the diagonal $\{(p, q) \in \mathcal{U} \mid p = q\}$. Similar results appeared, e.g., in Kleinberg (2004); Auer et al. (2007).

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. Finally, 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. (2023, Section 8) for the smoothed i.i.d. case, and it is in stark contrast with what happens in the stochastic case when only one price can be posted.

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 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. (2023, Theorem 6) and Azar et al. (2022, Theorem 3). We show that the σ-smoothness induces regularity on the expected gain from trade (Lemma 7, in Appendix A). This in turn allows us to prove a key discretization result (Claim 1).

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 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., 2023; 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", see Figure 2, left. 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 then 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: we need to carefully design σ-smooth distributions of the adversary that we can map into instances of online learning with feedback graphs that achieve their lower bound. 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. To address the second challenge, we prove a general information-theoretic result (Theorem 10, in Appendix E.1) that may be of independent interest for further lower-bounds constructions in related problems.

1.3. Additional related works

Further applications of smoothed analysis to online learning problems include the works by Block and Simchowitz (2022) and Block 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. (2023) 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)

2. Warm-up: one-price setting

In this section, we present our discretization error result (sharpening by constant factors the bound in Cesa-Bianchi et al. 2023) and present our results in the single-price setting.

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 (see Lemma 7 in Appendix A), thus opening the way to discretization methods, as formalized by the following result.

Claim 1 (Discretization error) *Let G be any finite grid of prices in $[0,1]$ and let $\delta(G)$ be the largest distance of a point in $[0,1]$ to G, i.e., $\delta(G) := \max_{p\in[0,1]} \min_{g\in G} |p-g|$, then for any sequence of σ-smooth distributions $\mathcal{S} = (S_1,B_1),\dots,(S_T,B_T)$, we have the following:*

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

Learning algorithm with full feedback : Price-Hedge

Input: time horizon T, Hedge algorithm $\mathcal{A}$, grid of prices G, with $|G| = K$
Initialization: Initialize $\mathcal{A}$ on time horizon T with K actions, one for each $p \in G$
for time $t = 1, 2, \dots$ **do**
  Receive from $\mathcal{A}$ the price $p_t \sim P_t \in G$
  Post price p_t to the agents and receive feedback $z_t = (s_t, b_t)$
  Feed to $\mathcal{A}$ the rewards $\mathrm{GFT}_t(p) = (b_t - s_t)\mathbb{I}\{s_t \le p \le b_t\}$, for all $p \in G$

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. Thus, they are able to reconstruct the gain from trade of any other pair of prices. By Claim 1, we can therefore run our favorite learning algorithm for (non-oblivious adversarial) online learning with expert advice on a discrete set of prices. For example, using Hedge (Freund and Schapire, 1997) we obtain the Price-Hedge algorithm, whose regret is controlled by the following theorem (for a proof, see Appendix B).

Theorem 2 *Consider the problem of repeated bilateral trade against a σ-smooth adversary in the full feedback model, for any $\sigma \in (0, 1]$. Then the regret of Price-Hedge, run using the uniform K-grid G on $[0, 1]$, for $K \ge 2$, satisfies:*

$$R_T(\textit{Price-Hedge}) \le 2\sqrt{T \ln K} + \frac{T}{\sigma K}.$$

In particular, if $T \ge 4$, tuning $K = \lfloor \sqrt{T} \rfloor$, the bound becomes: $R_T(\textit{Price-Hedge}) \le \frac{4}{\sigma} \cdot \sqrt{T \ln T}$.

We note here that the upper bound we achieved in Theorem 2 is tight in the time horizon, up to logarithmic factors. This follows from the fact that the distribution used in the $\Omega(\sqrt{T})$ lower bound in (Cesa-Bianchi et al., 2023, Theorem 3.3) is σ-smooth, for $\sigma \le 1/4$.

Posting a single price in partial information. Cesa-Bianchi et al. (2023) 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 of each other. They also showed that removing either the smoothness assumption or the 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. The proof of the following result is deferred to Appendix C.

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 $\big(S_t, B_t\big)_{t \ge 1}$ is an independent family of random variables, and S_t is independent of B_t for each t.*

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 lower bound of order $T^{3/4}$. This result has two notable implications. First, it provides a formalization to the intuition

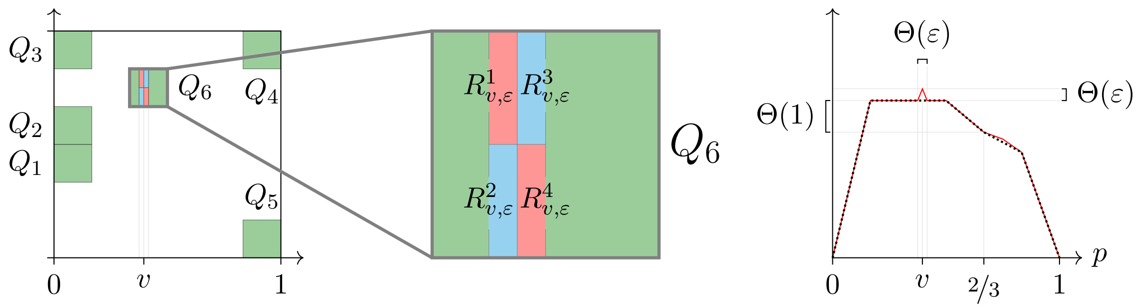

Figure 1: Left/center: The six squares $Q_1, \ldots, Q_6$ (in green) are the support of the base density f, and the four rectangles $R^1_{v,\varepsilon}, \ldots, R^4_{v,\varepsilon}$ (in red and blue) inside Q_6 are the regions where the density is perturbed with $g_{v,\varepsilon}$. Right: The corresponding 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).

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 constructed in the proof of Theorem 4 are i.i.d., we solve an open problem in Cesa-Bianchi et al. (2023), disproving their conjecture that the correct minimax rate is $T^{2/3}$.

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 $T \ge 8008$, then any learning algorithm $\mathcal{A}$ posting two prices per time step suffers at least a regret of*

$$R_T(\mathcal{A}) \ge \frac{1}{50^3} T^{3/4} .$$

The rest of the section is devoted to sketching the proof of the theorem (for a full proof, see Appendix F). The sketch is divided into three steps: first, we construct a hard instance of the repeated bilateral trade problem; then, we present a related problem on a discrete set of actions that preserves the relevant features of the original problem while allowing for an easier analysis of the regret; finally, we show how the minimax regret of the second problem leads to a $T^{3/4}$ regret for bilateral trade.

3.1. The construction of a hard family of adversaries

Here, we construct the family of σ-smooth adversaries for the repeated bilateral trade learning problem that we use to prove the lower bound. We consider i.i.d. adversaries: i.e., the valuations (S_t, B_t) are drawn i.i.d. according to a fixed distribution, obliviously of the actions of the learner.[‡] We build this family of distributions by suitable perturbations over a base distribution, whose support is given by the union of the six squares $Q_1, \ldots, Q_6$ (see Figure 1, left). The squares are obtained by translating $[0, 1/6]^2$, respectively, by $(0, \frac{1}{3}), (0, \frac{1}{2}), (0, \frac{5}{6}), (\frac{5}{6}, \frac{5}{6}), (\frac{5}{6}, 0), (\frac{1}{2}, \frac{2}{3})$. Letting $a \coloneqq 2 \ln(27/16)$, the probability density function f of the base distribution is

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

The perturbations to this base distribution are parametrized by two terms: a translation $v \in \left(\frac{1}{3}, \frac{1}{2}\right)$

[‡]This assumption makes our result even stronger: restricting the adversary can never make the lower bound bigger.

and a scale $\varepsilon \in \left(0, \frac{1}{12}\right)$ such that $\frac{1}{3} + \varepsilon \le v \le \frac{1}{2} - \varepsilon$. We denote the set of these parameters by Ξ. Each perturbed distribution has density $f_{v,\varepsilon} := f + g_{v,\varepsilon}$, where $g_{v,\varepsilon}$ is defined as follows:

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

and the rectangles $R^i_{v,\varepsilon}$ (see Figure 1, left/center) have the following analytic expression: $R^1_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\frac{3}{4}, \frac{5}{6}\right]$, $R^2_{v,\varepsilon} = [v-\varepsilon, v) \times \left[\frac{2}{3}, \frac{3}{4}\right)$, $R^3_{v,\varepsilon} = [v, v+\varepsilon] \times \left[\frac{3}{4}, \frac{5}{6}\right]$, $R^4_{v,\varepsilon} = [v, v+\varepsilon] \times \left[\frac{2}{3}, \frac{3}{4}\right)$. Note that the rectangles $R^i_{v,\varepsilon}$ are included in Q_6 for all $i \in [4]$ and $(v,\varepsilon) \in \Xi$.

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 probability density function f (resp., $f_{v,\varepsilon}$). We denote the expectation with respect to $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$) by $\mathbb{E}$ (resp., $\mathbb{E}^{v,\varepsilon}$). Note that the distribution of (S,B) with respect to $\mathbb{P}$ (resp., $\mathbb{P}^{v,\varepsilon}$), for all $(v,\varepsilon) \in \Xi$) is σ-smooth, for all $\sigma \le 1/9$. Given the explicit form for the base distribution, we can compute the corresponding expected value of the gain from trade $\mathbb{E}\big[\mathrm{GFT}(p,S,B)\big]$ obtained by posting price $p \in [0,1]$ to both agents, when (S,B) is drawn from the base distribution. The analytic expression of $\mathbb{E}\big[\mathrm{GFT}(p,S,B)\big]$ can be found in Appendix F (Equation (8)), and a plot is reported in Figure 1 (right, dotted black). What is relevant to our argument is that the function $p \mapsto \mathbb{E}\big[\mathrm{GFT}(p,S,B)\big]$ is continuous, maximized at every point of the plateau region $\left[\frac{1}{6}, \frac{1}{2}\right]$, and its value at $\frac{2}{3}$ is bounded away from the maximum. We can explicitly compute the expected gain from trade $\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,S,B)\big]$ obtainable by posting any price $p \in [0,1]$ to both agents, when (S,B) is drawn from the distribution with perturbation parameters v and ε. We have the following:

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,S,B)\big] = \mathbb{E}\big[\mathrm{GFT}(p,S,B)\big] + \frac{1}{864(1+8a)}\big(\varepsilon \cdot \Lambda_{v,\varepsilon}(p) + 12\varepsilon^2 \cdot \Lambda_{\frac{3}{4},\frac{1}{12}}(p)\big)$$

where $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as $\Lambda_{u,r}(x) = (1 - |x-u|/r)^+$. Thus, for each $(v,\varepsilon) \in \Xi$, the plot of $\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(v,S,B)\big]$ coincides with that of $\mathbb{E}\big[\mathrm{GFT}(v,S,B)\big]$ up to two small deviations (around v and $3/4$), and it is maximized at v (see Figure 1, right).

We now focus our attention on the feedback received by a learner that posts prices (p,q), when the underlying distribution corresponds to perturbations parameters $(v,\varepsilon) \in \Xi$.

Claim 2 *Fix any $(v,\varepsilon) \in \Xi$, $(p,q) \in \mathcal{U} \setminus \bigcup_{i\in[4]} R^i_{v,\varepsilon}$, and let $Z := \big(\mathbb{I}\{S \le p\}, \mathbb{I}\{q \le B\}\big)$. Then Z follows the same distribution under both $\mathbb{P}$ and $\mathbb{P}^{v,\varepsilon}$.*

Proof Here we consider only the event $\big\{Z = (0,0)\big\}$; for a full proof, see Claim 4 in Appendix F.

$$\mathbb{P}^{v,\varepsilon}\big[Z = (0,0)\big] = \mathbb{P}^0\big[Z = (0,0)\big] + \int_{(p,1]\times[0,q)} g_{v,\varepsilon}(x,y)\mathrm{d}x\mathrm{d}y \ .$$

If (p,q) is not in $R^{v,\varepsilon}$, by symmetry, the integral term is 0. ■

Claim 2 implies that if the learner wants to locate $v \in \left[\frac{1}{3} + \varepsilon, \frac{1}{2} - \varepsilon\right]$ observing samples of the two-bit feedback Z drawn according to the distribution $\mathbb{P}^{v,\varepsilon}$, they have to post prices in the region Q_6. However, in doing so, they suffer constant instantaneous regret. Indeed, a direct verification shows that for any $(v,\varepsilon) \in \Xi$ and all $(p,q) \in Q_6$,

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

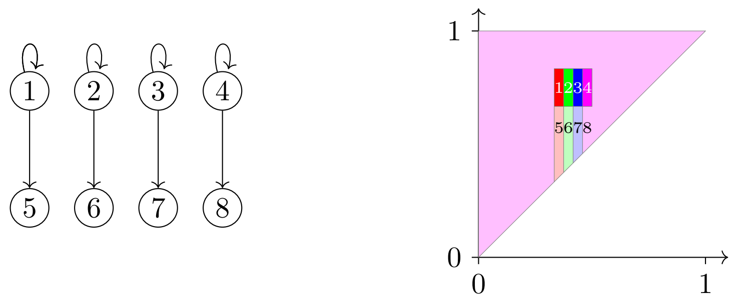

Figure 2: Left: The feedback graph of multi-apple tasting for $K = 4$. Right: The map ι.

So far, we built a family of i.i.d. adversaries for our bilateral trade problem such that the optimal pair of prices belongs to $\mathscr{D}_{\text{opt}} \coloneqq \big\{(p,q) \in \mathcal{U} \mid p = q \in [1/3, 1/2]\big\}$, but, when the underlying distribution is determined by one of the probability measures $\mathbb{P}^{v,\varepsilon}$, in order not to suffer regret $\Omega(\varepsilon T)$, the learner has to detect an ε-spike inside $\mathscr{D}_{\text{opt}}$. As observed in Claim 2, this can only be accomplished by posting prices in Q_6, which, as shown above, has an instantaneous regret of order $\Omega(1)$. The missing piece is now to quantify how long the learner can be forced to spend time posting prices in Q_6. To this end, we build a reduction from a simplified *online learning with feedback graph* problem on $2K$ arms that highlights the underlying structure of our problem. Our goal is to show that for any algorithm $\mathcal{A}$ for the repeated bilateral trade problem there exists an algorithm $\widetilde{\mathcal{A}}$ for the new problem such that the regret suffered by the latter is a lower bound on the regret suffered by the former.

3.2. The multi-apple tasting problem

In this section, we introduce an auxiliary online learning problem on a discrete set of actions that we call *multi-apple tasting*: it will be easier to analyze than our original bilateral trade problem while still capturing its difficulties. The multi-apple tasting problem has the following form: there are $2K$ actions, the first K are called the *exploration* arms, while the others are the *exploitation* arms. Playing an exploitation arm yields no feedback, while an exploration arm i gives information about the performance of the corresponding exploitation arm $i + K$. The reader familiar with the notion of online learning with directed feedback graphs (Alon et al., 2015) will recognize that the feedback model described here corresponds to the simple (weakly observable) feedback graph in Figure 2 (left).

The rewards. We now describe the random rewards of $K+1$ instances of the multi-apple tasting problem associated to $K+1$ probability measures $\mathbb{P}^0, \dots, \mathbb{P}^K$. Set c_{prob} to be $7/(2a)$ and consider the i.i.d. sequence of random vectors $Y, Y_1, Y_2, \dots, Y_T$ such that $Y \in \{0,1\}^{2K}$ and, for each $k \in \{0, \dots, K\}$ and $i \in [K]$, it holds that $Y(i+K) = Y_1(i+K) = \dots = Y_T(i+K) = 0$ and

$$\mathbb{P}^k\big[Y(i) = 1\big] = \begin{cases} \frac{1}{2} & \text{if } i \in [K]\backslash\{k\} \\ \frac{1}{2} + c_{\text{prob}} \cdot \varepsilon & \text{if } i = k \end{cases}$$

The random vectors $Y_1, Y_2, \dots, Y_T$ control the rewards the learner gets in this new problem. Formally, a learner playing action $i \in [2K]$ at time t gets reward $\rho_t(i) \coloneqq \rho(i, Y_t)$ where

$$\rho(i, y) \coloneqq \begin{cases} 0 & \text{if } j \in [K] \\ c_{\text{plat}} + \frac{c_{\text{spike}}}{c_{\text{prob}}} \cdot \big(y(j-K) - \frac{1}{2}\big) & \text{otherwise} \end{cases}$$

$c_{\text{plat}} \coloneqq \frac{a}{2(1+8a)}$, $c_{\text{spike}} \coloneqq \frac{1}{6(1+8a)} \cdot \frac{1}{144}$, and, for any $i \in [K]$, we denoted i-th component of y by $y(i)$. Observe that for all $k \in \{0, \dots, K\}$ and $i \in \{K+1, \dots, 2K\}$, we have

$$\mathbb{E}^k\big[\rho(i, Y)\big] = \begin{cases} c_{\text{plat}} & \text{if } k \neq i - K \\ c_{\text{plat}} + c_{\text{spike}} \cdot \varepsilon & \text{otherwise} \end{cases}$$

The feedback. The learner in multi-apple tasting receives two types of feedback. If they play action $i \geq K+1$ (an exploitation arm) at time t, then they receive no feedback (modeled by $Y_t(i) = 0$). If instead, they play action $i \leq K$ (an exploration arm), they receive feedback $Y_t(i)$. This feedback structure describes an instance of online learning with feedback graphs, where the underlying graph is the one in Figure 2 (left). The rewards incurred by the exploring arms are fixed and known irregardless of the action played, while the only way to learn the expected value of $\rho_t(i)$ for $i > K$ is to play the corresponding exploring action $i - K$.

The minimax regret. Leveraging a standard information-theoretic argument, it can be proved that any algorithm for the multi-apple tasting problem has to suffer a regret of order at least $\Omega\big(\min(\frac{K}{\varepsilon^2}, \varepsilon T)\big)$ on at least one of the instances induced by $\mathbb{P}^0, \dots, \mathbb{P}^K$. Intuitively, in order to prevent losing εT, the learner has to play each one of the K exploring arms at least $\Omega(1/\varepsilon^2)$ times.

3.3. Relating the two problems

We have described multi-apple tasting, and $K+1$ distributions to generate the sequence of rewards for it. We now show how to simulate any distribution of the feedback in instances $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$ of the bilateral trade problem using the random variables Y (and some extra random seeds). Let $K = \lceil T^{1/4} \rceil$ and $\varepsilon = \frac{1}{2K}$, and consider the baseline instance and the K perturbed instances of the repeated bilateral trade problem above, each corresponding to $(v_k, \frac{\varepsilon}{6})$ for $v_k = \frac{1}{3} + (2k-1)\frac{\varepsilon}{6}$ and $k \in [K]$. For each one of these instances, we construct an instance of multi-apple tasting that can be used to *simulate* it.

As a first step, we explain how to associate each pair of prices in the upper triangle (i.e., the set of actions in the bilateral trade problem) to one of the $2K$ actions in the feedback graph problem. We partition the upper triangle $\mathcal{U}$ of the unit square $[0,1]^2$ into $2K$ subsets, each corresponding to areas of "similar" behavior:

- $J_k \coloneqq [v_k - \frac{\varepsilon}{6}, v_k + \frac{\varepsilon}{6}) \times [\frac{2}{3}, \frac{5}{6}]$, $\forall k \in [K-1]$, and $J_K \coloneqq [v_K - \frac{\varepsilon}{6}, v_K + \frac{\varepsilon}{6}] \times [\frac{2}{3}, \frac{5}{6}]$.
- $J_{k+K} \coloneqq \{(p,q) \in \mathcal{U} \mid v_k - \frac{\varepsilon}{6} \leq p < v_k + \frac{\varepsilon}{6} \text{ and } q < \frac{2}{3}\}$, $\forall k \in [K-1]$, and $J_{2K} \coloneqq \mathcal{U} \setminus \bigcup_{k=1}^{2K-1} J_k$.

Given the partition, we can introduce the map ι which associates each $(p,q) \in \mathcal{U}$ with the unique $i \in [2K]$ such that $(p,q) \in J_i$ (see Figure 2, right, for a pictorial representation of ι). Then, we introduce an i.i.d. sequence $V, V_1, V_2, \dots, V_T$ of uniform random variables in $[0,1]$, independent of the sequence of Ys. Both the Y and the V sequences are independent of the sequence of valuations $(S_1, B_1), (S_2, B_2), \dots, (S_T, B_T)$.

The next claim is the core of our reduction: it can be proved by applying our novel information-theoretic result (Theorem 10, Appendix E). To do it, one can verify that, for all $k \in [K]$, the Radon-Nikodym derivative of the distribution of the feedback $\big(\mathbb{I}(S \leq p), \mathbb{I}\{q \leq B\}\big)$ under $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$ with respect to its distribution under $\mathbb{P}$ is bounded from above (resp., below) by the maximum (resp., minimum) of the Radon-Nikodym derivative of the distribution of $Y(\iota(p,q))$ under $\mathbb{P}^k$ with respect to its distribution under $\mathbb{P}^0$. For a proof, see Claim 5 in Appendix F.

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 $\mathrm{GFT}(p)$

Claim 3 *For any $(p, q) \in \mathcal{U}$ there exists a function $\varphi_{p,q} : \{0, 1\} \times [0, 1] \to \{0, 1\}^2$ such that, for all $k \in [K]$, the distribution of $\varphi_{p,q}(Y(\iota(p, q)), V)$ under $\mathbb{P}^0$ (resp., $\mathbb{P}^k$, for all $k \in [K]$) is the same as that of $\big(\mathbb{I}(S \le p), \mathbb{I}\{q \le B\}\big)$ under $\mathbb{P}$ (resp., $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$).*

We now proceed as follows. Let $\mathcal{A}$ be any algorithm for the original bilateral trade problem. We show how to simulate its behavior over the instances $\mathbb{P}$ and $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$, for $k \in [K]$, using an algorithm $\widetilde{\mathcal{A}}$ for multi-apple tasting (together with the sequence of random seeds $V_1, V_2, \dots, V_T$) over the distributions $\mathbb{P}^0$ and $\mathbb{P}^k$, for $k \in [K]$. When algorithm $\mathcal{A}$ chooses prices $(p_t, q_t) \in \mathcal{U}$ at time t, then $\widetilde{\mathcal{A}}$ plays the action $\iota(p_t, q_t) \in [2K]$, receives reward $\rho_t(\iota(p_t, q_t))$ and observes the feedback $Y_t(\iota(p_t, q_t))$. Algorithm $\mathcal{A}$ is then fed the feedback $\varphi_{p_t,q_t}\big(Y_t(\iota(p_t, q_t)), V_t\big) \in \{0, 1\}^2$ which it uses to select its new action (p_{t+1}, q_{t+1}). Crucially, leveraging Claim 3 and the structure of the rewards in two problems, one can prove that the regret $R^0_T(\mathcal{A})$ (resp., $R^k_T(\mathcal{A})$, for any $k \in [K]$) that algorithm $\mathcal{A}$ suffers under probability $\mathbb{P}$ (resp., $\mathbb{P}^{v_k, \frac{\varepsilon}{6}}$) in the repeated bilateral trade problem is at least the regret $\widetilde{R}^0_T(\widetilde{\mathcal{A}})$ (resp., $\widetilde{R}^k_T(\widetilde{\mathcal{A}})$) that algorithm $\widetilde{\mathcal{A}}$ suffers under probability $\mathbb{P}^0$ (resp., $\mathbb{P}^k$) in the multi-apple tasting problem. Finally, the proof can be concluded by putting together the lower bound $\Omega\big(\min(\frac{K}{\varepsilon^2}, \varepsilon T)\big)$ for the multi-apple tasting problem with our choices of K and ε to obtain that the minimax regret for the bilateral trade problem is at least of order $\Omega(T^{3/4})$.

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.

Lemma 5 (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., $\mathbb{E}\Big[\widehat{\mathrm{GFT}}(p)\Big] = \mathrm{GFT}(p)$, 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, in Appendix D 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

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 $W_1 \coloneqq K$
for time $t = 1, 2, \dots$ **do**
  Let $\pi_t(i) \coloneqq \frac{w_t(i)}{W_t}$ for all $i \in [K]$
  Toss a biased coin with probability γ of Heads
  if Tails **then** ▷ Exploitation step
    Post price p_t drawn according to distribution π_t and set $\hat{r}_t(i) \coloneqq 0$ 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 $\hat{r}_t(I_t) \coloneqq \frac{K}{\gamma} \cdot \widehat{\mathrm{GFT}}_t(g_{I_t})$ and $\hat{r}_t(j) \coloneqq 0$ for all $j \neq I_t$.
  Let $w_{t+1}(i) \coloneqq w_t(i) \cdot \exp\big(\eta \hat{r}_t(i)\big)$ for all $i \in [K]$ ▷ Exponential weight update
  Let $W_{t+1} \coloneqq \sum_{p_i \in G} w_{t+1}(i)$

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 6 *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 $\frac{2\eta K}{\gamma} \leq 1$, then, for each time horizon $T \in \mathbb{N}$, we have that*

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

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 $R_T(\textit{Blind-Exp3}) \leq 2\left(\frac{1}{\sigma} + (\ln T)^{1/3}\right) \cdot T^{3/4}$.

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.

Acknowledgments

NCB, RC, FF, and SL are partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and 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 EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence). FF and SL are also partially supported by ERC Advanced Grant 788893 AMDROMA and PNRR MUR project IR0000013-SoBigData.it.

References

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 J. Comput.*, 46(6): 1785–1826, 2017.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory (COLT)*, 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.

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

Richard F. Bass. *Real analysis for graduate students*. Createspace Ind Pub, 2013.

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 *ACM Conference on Economics and Computation (EC)*, page 617, 2014.

Nicolò Cesa-Bianchi, Tommaso R. 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. Bilateral trade: A regret minimization perspective. *Mathematics of Operations Research*, 0(0), 2023.

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.

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.

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.

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.

Takesaki Masamichi. *Theory of Operator Algebras I*. Springer New York, NY, 1979.

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. Missing proofs from Section 2: Lipschitzess and discretization

Lemma 7 (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:*

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

Proof Let $x > y$ be any two prices in $[0,1]$, we have the following:

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

Note that in the second to last inequality we used the assumption on the smoothness of (S, B) and we introduced U and V, two independent uniform random variables in $[0,1]$. ∎

Claim 1 (Discretization error) *Let G be any finite grid of prices in $[0,1]$ and let $\delta(G)$ be the largest distance of a point in $[0,1]$ to G,* i.e.*, $\delta(G) := \max_{p \in [0,1]} \min_{g \in G} |p - g|$, then for any sequence of σ-smooth distributions $\mathcal{S} = (S_1, B_1), \ldots, (S_T, B_T)$, we have the following:*

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

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

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

where, in the second to last inequality, we used the Lipschitz property of the expected gain from trade as in Lemma 7. ∎

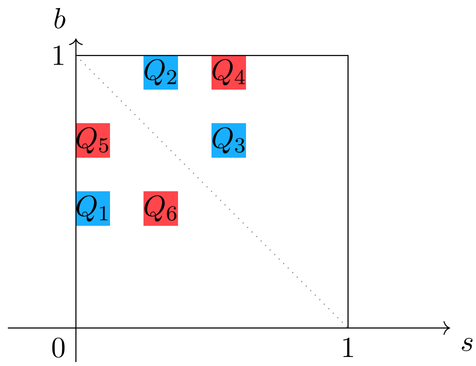

Figure 3: The squares $Q_1, \dots, Q_6$ appearing in the proof of Theorem 3.

Appendix B. Missing proofs from Section 2: Full feedback

Theorem 2 *Consider the problem of repeated bilateral trade against a σ-smooth adversary in the full feedback model, for any $\sigma \in (0, 1]$. Then the regret of Price-Hedge, run using the uniform K-grid G on $[0, 1]$, for $K \geq 2$, satisfies:*

$$R_T(\textit{Price-Hedge}) \leq 2\sqrt{T \ln K} + \frac{T}{\sigma K}.$$

In particular, if $T \geq 4$, tuning $K = \lfloor \sqrt{T} \rfloor$, the bound becomes: $R_T(\textit{Price-Hedge}) \leq \frac{4}{\sigma} \cdot \sqrt{T \ln T}$.

Proof We show how running Price-Hedge for the right choices grid of prices G yields the desired result. As $\mathcal{A}$, we choose the Hedge algorithm for full information feedback (Freund and Schapire, 1997), while for any fixed $K \geq 2$, we consider the uniform grid G on $[0, 1]$ of the positive integer multiples of $\frac{1}{K}$: $G = \left\{\frac{1}{K}, \frac{2}{K}, \dots, 1\right\}$. For any σ-smooth adversary $\mathcal{S}$, we have the following:

$$\begin{aligned}
R_T(\text{Price-Hedge}, \mathcal{S}) &= \max_{p \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \text{GFT}_t(p) - \sum_{t=1}^{T} \text{GFT}_t(P_t)\right] \pm \max_{p \in G} \mathbb{E}\left[\sum_{t=1}^{T} \text{GFT}_t(p)\right] \\
&\leq \max_{p \in G} \mathbb{E}\left[\sum_{t=1}^{T} \text{GFT}_t(p) - \sum_{t=1}^{T} \text{GFT}_t(P_t)\right] + \frac{T}{\sigma K} \\
&\leq 2\sqrt{T \ln K} + \frac{T}{\sigma K}.
\end{aligned}$$

Note that, in the first inequality we used Claim 1, which holds for any (possibly adaptive) sequence of σ-smooth random variables, while in the second inequality, we used the well-known bound on the regret of Hedge—see, e.g., Arora et al. (2012, Theorem 2.5) with $\eta = \frac{1}{\sqrt{T}}$. We remark that the last bound holds for any (possibly adversarial) realizations of the agents' valuations, in expectation with respect to the internal randomness of the algorithm. ■

Appendix C. Missing proofs from Section 2: Linear lower bound

Theorem 3 *Consider the problem of repeated bilateral trade against a σ-smooth adversary in the two-bit feedback model, for any $\sigma \leq \frac{1}{64}$. Then any learning algorithm that posts a single price per* *time step suffers at least $\frac{T}{24}$ regret, even if $\big(S_t, B_t\big)_{t\geq 1}$ is an independent family of random variables, and S_t is independent of B_t for each t.*

Proof Consider the following six squares, depicted in Figure 3:

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

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 $\frac{1}{64}$ smooth.

From Cesa-Bianchi et al. (2023, 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:

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

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:

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

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. ■

Appendix D. Missing proofs from Section 4

Theorem 6 *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*

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

In particular, if $T \ge 16$, 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 $R_T(\textit{Blind-Exp3}) \le 2\left(\frac{1}{\sigma} + (\ln T)^{1/3}\right) \cdot T^{3/4}$.

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 $\frac{2\eta K}{\gamma} \le 1$. 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:

$$\begin{aligned}
\ln\left(\frac{W_{T+1}}{W_1}\right) &= \ln\left(\prod_{t=1}^{T}\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T}\ln\left(\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T}\ln\left(\sum_{i\in[K]}\pi_t(i)\exp\left(\eta\hat{r}_t(i)\right)\right)\\
&\le \sum_{t=1}^{T}\ln\left(1 + \eta\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i) + \eta^2\sum_{i\in[K]}\pi_t(i)\big(\hat{r}_t(i)\big)^2\right)\\
&\le \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i) + \eta^2\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\big(\hat{r}_t(i)\big)^2 \qquad \text{(using } \hat{r}_t(i) \le \tfrac{K}{\gamma}\text{)}\\
&\le \eta\sum_{t=1}^{T}\sum_{i\in[K]}\pi_t(i)\hat{r}_t(i)\left[1 + \eta\frac{K}{\gamma}\right]. \qquad (4)
\end{aligned}$$

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 ($\frac{2\eta K}{\gamma} \leq 1$) 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 4 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:

$$\begin{aligned}\mathbb{E}\left[\ln\left(\frac{W_{T+1}}{W_1}\right)\right] &= \mathbb{E}\left[\ln\left(W_{T+1}\right)\right] - \ln K \geq \mathbb{E}\left[\ln\left(w_{T+1}(i)\right)\right] - \ln K \\ &= \eta \sum_{t=1}^{T} \mathbb{E}\left[\hat{r}_t(i)\right] - \ln K = \eta \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{GFT}_t(g_i)\right] - \ln K. \end{aligned} \tag{5}$$

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:

$$\begin{aligned}
\mathbb{E}\left[\hat{r}_t(i) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\hat{r}_t(i) \mid \mathcal{F}_t\right] && \hat{r}_t(i) = \mathbb{I}\{\mathcal{E}_t^i\}\hat{r}_t(i)\\
&= \mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\mathbb{E}\left[\hat{r}_t(i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Law of total exp.}\\
&= \frac{K}{\gamma}\mathbb{E}\left[\mathbb{I}\{\mathcal{E}_t^i\}\mathbb{E}\left[\widehat{\mathrm{GFT}}_t(g_i) \mid \mathcal{F}_t, \mathcal{E}_t^i\right] \mid \mathcal{F}_t\right] && \text{Def. of } \hat{r}_t(i)\\
&= \frac{K}{\gamma}\mathbb{P}[\mathcal{E}_t^i \mid \mathcal{F}_t]\mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right] && \text{Lemma 5 and } (S_t, B_t) \text{ indep. of } \mathcal{E}_t^i\\
&= \mathbb{E}\left[\mathrm{GFT}_t(g_i) \mid \mathcal{F}_t\right]
\end{aligned}$$

For the final step, note that, conditioned on $\mathcal{F}_t$, the event $\mathcal{E}_t^i$ has probability $\frac{\gamma}{K}$: 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 (4) 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:

$$\mathbb{E}\left[\mathrm{GFT}_t(P_t, Q_t)\right] \geq (1-\gamma) \sum_{i \in [K]} \mathbb{E}\left[\pi_t(i)\mathrm{GFT}_t(g_i)\right] = (1-\gamma) \sum_{i \in [K]} \mathbb{E}\left[\pi_t(i)\hat{r}_t(i)\right], \tag{6}$$

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 (5) and Equation (6) into Equation (4) 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 $\frac{1-\gamma}{\eta}$, 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:

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

■

Appendix E. One-bit/two-scenarios inverse-transformation representability

We recall that given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(\Omega, \mathcal{F})$, we say that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and we write $\mathbb{Q} \ll \mathbb{P}$ if for all $E \in \mathcal{F}$ such that $\mathbb{P}[E] = 0$, it holds that $\mathbb{Q}[E] = 0$. Moreover, if $\mathbb{Q} \ll \mathbb{P}$, the Radon-Nikodym theorem states that there exists a density (called Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ and denoted by) $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \colon \Omega \to [0, \infty)$ such that, for all $E \in \mathcal{F}$, it holds that

$$\mathbb{Q}[E] = \int_E \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega) \,\mathrm{d}\mathbb{P}(\omega) \ .$$

For a reference of the previous result, see (Bass, 2013, Theorem 13.4).

Moreover, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ is a measurable space, and X is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, we denote by $\mathbb{P}_X$ the push-forward measure of $\mathbb{P}$ by X, i.e., the probability measure defined on $\mathcal{F}_{\mathcal{X}}$ by $\mathbb{P}_X[F] := \mathbb{P}[X \in F]$, for all $F \in \mathcal{F}_{\mathcal{X}}$.

If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are two measurable spaces, we denote by $\mathcal{F} \otimes \mathcal{F}'$ the σ-algebra of subsets of $\Omega \times \Omega'$ generated by the collection of subsets of the form $F \times F'$, where $F \in \mathcal{F}$ and $F' \in \mathcal{F}'$. If $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega', \mathcal{F}', \mathbb{P}')$ are two probability spaces, we denote the product measure of $\mathbb{P}$ and $\mathbb{P}'$ by $\mathbb{P} \otimes \mathbb{P}'$, i.e., $\mathbb{P} \otimes \mathbb{P}'$ is the unique probability measure defined on $\mathcal{F} \otimes \mathcal{F}'$ which satisfies $(\mathbb{P} \otimes \mathbb{P}')[F \times F'] = \mathbb{P}[F]\mathbb{P}'[F']$, for all $E \in \mathcal{F}$ and $E' \in \mathcal{F}'$.

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ are measurable spaces, X is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, and Y is a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, we denote the conditional probability of X given Y by $\mathbb{P}_{X|Y}$, i.e., $\mathbb{P}_{X|Y}[E] = \mathbb{P}[X \in E \mid Y]$, for each $E \in \mathcal{F}_{\mathcal{X}}$. In this case, for each $E \in \mathcal{F}_{\mathcal{X}}$, we recall that $\mathbb{P}_{X|Y}[E]$ is a $\sigma(Y)$-measurable random variable. Furthermore, if X' is another random variable from $(\Omega, \mathcal{F})$ to some measurable space $(\mathcal{X}', \mathcal{F}_{\mathcal{X}'})$, f and g are two real-valued bounded measurable functions (respectively from $(\mathcal{X} \otimes \mathcal{Y}, \mathcal{F}_{\mathcal{X}} \otimes \mathcal{F}_{\mathcal{Y}})$ to the reals and from $(\mathcal{X}' \otimes \mathcal{Y}, \mathcal{F}_{\mathcal{X}'} \otimes \mathcal{F}_{\mathcal{Y}})$ to the reals), and both $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{X}, \mathcal{F}_{\mathcal{X}'})$ are measurable spaces that arise from considering the Borel subsets of separable and complete metric space $(\mathcal{X}, d)$ and $(\mathcal{X}', d')$ respectively, it holds that

$$\mathbb{E}\big[f(X,Y)g(X',Y) \mid Y\big] = \mathbb{E}\big[f(X,Y) \mid Y\big] \cdot \mathbb{E}\big[g(X',Y) \mid Y\big]$$

whenever

$$\mathbb{P}_{(X,X')|Y} = \mathbb{P}_{X|Y} \otimes \mathbb{P}_{X'|Y} \ .$$

$$
\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "{(X,U)}"] \arrow[dr, "Y"'] & \big(\{0,1\} \times [0,1], 2^{\{0,1\}} \otimes \mathcal{B}\}\big) \arrow[d, "\varphi"] \\
& (\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})
\end{tikzcd}
$$

Figure 4: Pictorial representation of Theorem 10. The way to interpret it is not event by event but in probability: the probability of a measurable set in $\mathcal{F}_Y$ can be computed in Ω equivalently via the pullback of Y, or of $\varphi \circ (X, U)$.

E.1. Our inverse-transformation result

In this section, we present a theorem that extends, in spirit, the classic inverse transformation method. This result that can be of independent interest for replacing a type of feedback with another of better quality in lower-bound constructions based on reductions to simpler games.

Definition 8 (Inverse-transformation representability) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{B}$ be the Borel σ-algebra of $[0,1]$. We say that $\mathbb{P}$ is* inverse-transformation-representable *if there exists a measurable function ψ from $\big([0,1], \mathcal{B}\big)$ to $(\Omega, \mathcal{F})$ such that*[§] $\mathbb{P} = \mathbb{L}_\psi$.

The following theorem is a simple consequence of (Masamichi, 1979, Corollary A.11), and shows "inverse-transformation representability in separable and complete metric spaces".

Theorem 9 *Suppose that $(\mathcal{Y}, d)$ is a separable and complete metric space, with $\mathcal{F}_{\mathcal{Y}}$ as the Borel σ-algebra of $(\mathcal{Y}, d)$. Then any probability measure defined on $\mathcal{F}_{\mathcal{Y}}$ is inverse-transformation-representable.*

We are now ready to state the main theorem of this section. When we are uncertain about the underlying probability according to which some samples are drawn, and the uncertainty is between two probability measure $\mathbb{P}$ and $\mathbb{Q}$, the theorem provides a characterization under which we can simulate a random variable Y using some independent random seed U and having access to a 1-bit random variable X. This theorem can be of independent interest as a tool for lower bound reductions in online learning problems, as we used for example in Theorem 4. It establishes "One-bit/two-scenarios inverse-transformation representability in separable and complete metric spaces".

Theorem 10 *Suppose that $(\mathcal{Y}, d)$ is a separable and complete metric space with $\mathcal{F}_{\mathcal{Y}}$ as the Borel σ-algebra of $(\mathcal{Y}, d)$. Let $(\Omega, \mathcal{F})$ be a measurable space, X a random variable from $(\Omega, \mathcal{F})$ to $\big(\{0,1\}, 2^{\{0,1\}}\big)$, Y a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$, and U random variable from $(\Omega, \mathcal{F})$ to $\big([0,1], \mathcal{B}\big)$, where $\mathcal{B}$ is the Borel σ-algebra of $[0,1]$. Suppose that $\mathbb{P}, \mathbb{Q}$ are probability measures defined on $\mathcal{F}$, and $p \in (0,1)$, $q \in [0,1]$ are such that:*

- $\mathbb{P}[X = 1] = p$ *and* $\mathbb{Q}[X = 1] = q$.
- *U is a uniform random variable on $[0,1]$ both under $\mathbb{P}$ and $\mathbb{Q}$, i.e., we have that $\mathbb{P}_U = \mathbb{L} = \mathbb{Q}_U$.*
- *U is independent of X both under $\mathbb{P}$ and $\mathbb{Q}$, i.e., $\mathbb{P}_{(X,U)} = \mathbb{P}_X \otimes \mathbb{P}_U$ and $\mathbb{Q}_{(X,U)} = \mathbb{Q}_X \otimes \mathbb{Q}_U$.*

[§]We recall that $\mathbb{L}$ is the Lebesgue measure on $\mathcal{B}$.

Then, the following are equivalent:

1. *There exists a measurable function φ from $\big(\{0,1\}\times[0,1], 2^{\{0,1\}}\otimes\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$ such that*

$$\mathbb{P}_Y = \mathbb{P}_{\varphi(X,U)} \qquad \text{and} \qquad \mathbb{Q}_Y = \mathbb{Q}_{\varphi(X,U)} \ .$$

2. *$\mathbb{Q}_Y \ll \mathbb{P}_Y$, and $\mathbb{P}_Y$-almost-surely it holds that*

$$\min \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \le \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} \le \max \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \ .$$

Proof We divide the proof in two parts, depending on whether or not $p = q$.

Assume first that $p \neq q$. In this case, we will prove the chain of equivalencies

$$\text{Item } 1 \iff \text{Item a} \iff \text{Item b} \iff \text{Item c} \iff \text{Item } 2 \ ,$$

where Item a, Item b, and Item c are the following propositions:

a) There exists two probability measures μ_0 and μ_1 over $\mathcal{F}_\mathcal{Y}$ such that

$$\mathbb{P}_Y = (1-p)\mu_0 + p\mu_1 \qquad \text{and} \qquad \mathbb{Q}_Y = (1-q)\mu_0 + q\mu_1 \ .$$

b) $\frac{q}{q-p}\mathbb{P}_Y - \frac{p}{q-p}\mathbb{Q}_Y \ge 0$ and $\frac{1-p}{q-p}\mathbb{P}_Y - \frac{1-q}{q-p}\mathbb{Q}_Y \ge 0$.

c) $\mathbb{Q}_Y \ll \mathbb{P}_Y$ and $\min\big(\frac{q}{p}, \frac{1-q}{1-p}\big) \le \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} \le \max\big(\frac{q}{p}, \frac{1-q}{1-p}\big)$ for all $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] > 0$.

We begin by proving that Item 1 is equivalent to Item a. Assume Item 1. Define $\mu_0 \coloneqq \mathbb{P}_{\varphi(0,U)}$ and $\mu_1 \coloneqq \mathbb{P}_{\varphi(1,U)}$. Since U is uniform under both under $\mathbb{P}$ and $\mathbb{Q}$, it also holds that $\mu_0 = \mathbb{Q}_{\varphi(0,U)}$ and $\mu_1 = \mathbb{Q}_{\varphi(1,U)}$. Thus

$$\begin{aligned}
\mathbb{P}_Y &= \mathbb{P}_{\varphi(X,U)} = (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-p)\mu_0 + p\mu_1 \\
\mathbb{Q}_Y &= \mathbb{Q}_{\varphi(X,U)} = (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = (1-q)\mu_0 + q\mu_1 \ ,
\end{aligned}$$

where we used that fact that X and U are independent both under $\mathbb{P}$ and $\mathbb{Q}$ and that $\mathbb{P}[X=1] = p$, $\mathbb{Q}[X=1] = q$. This proves Item a.

Vice versa, assume Item a. By Theorem 9, we can find two measurable functions ψ_0, ψ_1 from $\big([0,1],\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$ such that $\mu_0 = \mathbb{L}_{\psi_0}$ and $\mu_1 = \mathbb{L}_{\psi_1}$ and define

$$\varphi(x,u) \coloneqq \begin{cases} \psi_0(u) & \text{if } x = 0 \\ \psi_1(u) & \text{if } x = 1 \end{cases}$$

for all $x \in \{0,1\}$ and $u \in [0,1]$. Then φ is a measurable function from $\big(\{0,1\}\times[0,1], 2^{\{0,1\}}\otimes\mathcal{B}\big)$ to $(\mathcal{Y},\mathcal{F}_\mathcal{Y})$, and since X is independent of U and U is uniform on $[0,1]$ both under $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}
\mathbb{P}_{\varphi(X,U)} &= (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-p)\mathbb{P}_{\psi_0(U)} + p\mathbb{P}_{\psi_1(U)} \\
&= (1-p)\mathbb{L}_{\psi_0} + p\mathbb{L}_{\psi_1} = (1-p)\mu_0 + p\mu_1 = \mathbb{P}_Y \\
\mathbb{Q}_{\varphi(X,U)} &= (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = (1-q)\mathbb{Q}_{\psi_0(U)} + q\mathbb{Q}_{\psi_1(U)} \\
&= (1-q)\mathbb{L}_{\psi_0} + q\mathbb{L}_{\psi_1} = (1-q)\mu_0 + q\mu_1 = \mathbb{Q}_Y
\end{aligned}$$

This proves Item 1 and in turn yields that Item 1 is equivalent to Item a.

We now prove that Item a is equivalent to Item b. Assume Item a. Then, for each $A \in \mathcal{F}_\mathcal{Y}$ we have that the pair $\big(\mu_0[A], \mu_1[A]\big)$ is the (only) solution of the linear system

$$\begin{cases} (1-p)x_0 + px_1 &= \mathbb{P}_Y[A] \\ (1-q)x_0 + qx_1 &= \mathbb{Q}_Y[A] \end{cases}$$

in the two variables (x_0, x_1), which implies

$$\mu_0[A] = \frac{q}{q-p}\mathbb{P}_Y[A] - \frac{p}{q-p}\mathbb{Q}_Y[A] \qquad \text{and} \qquad \mu_1[A] = \frac{1-p}{q-p}\mathbb{Q}_Y[A] - \frac{1-q}{q-p}\mathbb{P}_Y[A] \; .$$

Since μ_0 and μ_1 are (non-negative) measures, this implies Item b.

Vice versa, assume Item b. Define

$$\mu_0 \coloneqq \frac{q}{q-p}\mathbb{P}_Y - \frac{p}{q-p}\mathbb{Q}_Y \qquad \text{and} \qquad \mu_1 \coloneqq \frac{1-p}{q-p}\mathbb{Q}_Y - \frac{1-q}{q-p}\mathbb{P}_Y \; .$$

Since μ_0 and μ_1 are a linear combination of measures, they are signed measures and, by Item b, actually, they are (non-negative) measures. The fact that they are also probability measures follows trivially from $\mathbb{P}_Y[\mathcal{Y}] = 1 = \mathbb{Q}_Y[\mathcal{Y}]$. Now, a direct verification shows that $\mathbb{P}_Y = (1-p)\mu_0 + p\mu_1$ and $\mathbb{Q}_Y = (1-q)\mu_0 + q\mu_1$, i.e., that Item a holds. We have then proved that Item a is equivalent to Item b.

We now prove that Item b is equivalent to Item c. Firstly, note that by elementary linear-algebra (dividing by $\widetilde{p}$ and solving by $\widetilde{q}/\widetilde{p}$ the linear system of inequalities), for each $\widetilde{q} \in [0,1]$ and $\widetilde{p} \in (0,1]$, the following equivalence holds

$$\begin{cases} \dfrac{q}{q-p}\widetilde{p} - \dfrac{p}{q-p}\widetilde{q} \geq 0 \\ \dfrac{1-p}{q-p}\widetilde{q} - \dfrac{1-q}{q-p}\widetilde{p} \geq 0 \end{cases} \iff \min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \leq \frac{\widetilde{q}}{\widetilde{p}} \leq \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \tag{7}$$

Assume Item b. Note that if $p < q$ (resp., $q < p$), then if $A \in \mathcal{F}_\mathcal{Y}$ is such that $\mathbb{P}_Y[A] = 0$, the first (resp., second) inequality in Item b implies that also $\mathbb{Q}_Y[A] = 0$, which in turn yields $\mathbb{Q}_Y \ll \mathbb{P}_Y$. Furthermore, for each $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] \neq 0$, the equivalence in (7) with $\widetilde{p} \coloneqq \mathbb{P}_Y[A]$ and $\widetilde{q} \coloneqq \mathbb{Q}_Y[A]$ implies that

$$\min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \leq \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} \leq \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right)$$

which yields Item c.

Vice versa, assume Item c. Note that Item b holds

- For all $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] = 0$, because in this case also $\mathbb{Q}_Y[A] = 0$
- For all $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] \neq 0$, by the equivalence in (7) with $\widetilde{p} \coloneqq \mathbb{P}_Y[A]$ and $\widetilde{q} \coloneqq \mathbb{Q}_Y[A]$

This proves that Item b and Item c are equivalent.

We now prove that Item c is equivalent to Item 2. Assume Item c. Assume by contradiction that Item 2 does not hold. Then, there exists $A \in \mathcal{F}_\mathcal{Y}$ such that $\mathbb{P}_Y[A] > 0$ such that either for all $y \in A$ it holds that $\max\big(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\big) < \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}(y)$ or it holds that $\min\big(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\big) > \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}(y)$. In the first case

$$\max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) = \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right) \geq \frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} = \frac{1}{\mathbb{P}_Y[A]}\int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y}\,\mathrm{d}\mathbb{P}_Y > \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \; ,$$

yielding the contradiction we were seeking. The second case yields a contradiction in an analogous manner.

Vice versa, assume Item 2. Then, if $A \in \mathcal{F}_{\mathcal{Y}}$ is such that $\mathbb{P}_Y[A] > 0$, notice that

$$\min\left(\frac{q}{p}, \frac{1-q}{1-p}\right) = \min\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \le \frac{1}{\mathbb{P}_Y[A]} \int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} \,\mathrm{d}\mathbb{P}_Y \le \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) = \max\left(\frac{q}{p}, \frac{1-q}{1-p}\right)$$

which together with

$$\frac{\mathbb{Q}_Y[A]}{\mathbb{P}_Y[A]} = \frac{1}{\mathbb{P}_Y[A]} \int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} \,\mathrm{d}\mathbb{P}_Y$$

(since $\mathbb{Q}_Y \ll \mathbb{P}_Y$), implies Item c. This proves that Item c and Item 2 are equivalent and shows in turn that Item 1 is equivalent to Item 2 whenever $p \neq q$.

Assume now that $p = q$. Assume Item 1. Since X is independent of U and U is uniform on $[0, 1]$ both under $\mathbb{P}$ and $\mathbb{Q}$, we get

$$\mathbb{P}_Y = \mathbb{P}_{\varphi(X,U)} = (1-p)\mathbb{P}_{\varphi(0,U)} + p\mathbb{P}_{\varphi(1,U)} = (1-q)\mathbb{Q}_{\varphi(0,U)} + q\mathbb{Q}_{\varphi(1,U)} = \mathbb{Q}_{\varphi(X,U)} = \mathbb{Q}_Y \ .$$

Hence, in particular $\mathbb{Q}_Y \ll \mathbb{P}_Y$ and $\frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} = 1$ $\mathbb{P}_Y$-almost-surely, which, together with the fact

$$\min\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \le 1 \le \max\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right)$$

implies Item 2.

Vice versa, assume Item 2. Fix a measurable function ψ from $\big([0,1], \mathcal{B}\big)$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ such that $\mathbb{P}_Y = \mathbb{L}_\psi$ (whose existence is guaranteed by Theorem 9). Let $\varphi(x,u) := \psi(u)$ for all $x \in \{0,1\}$ and $u \in [0,1]$. Being U uniform both under $\mathbb{P}$ and $\mathbb{Q}$, we get that $\mathbb{P}_{\varphi(X,U)} = \mathbb{P}_{\psi(U)} = \mathbb{L}_\psi = \mathbb{Q}_{\psi(U)} = \mathbb{Q}_{\varphi(X,U)}$. Moreover, since $p = q$, we have that $\min \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} = 1 = \max \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}$, which, together with Item 2, yields that, for any $A \in \mathcal{F}_{\mathcal{Y}}$,

$$\mathbb{Q}_Y[A] = \int_A \frac{\mathrm{d}\mathbb{Q}_Y}{\mathrm{d}\mathbb{P}_Y} \,\mathrm{d}\mathbb{P}_Y = \int_A 1 \,\mathrm{d}\mathbb{P}_Y = \mathbb{P}_Y[A] \ ,$$

thus $\mathbb{P}_Y = \mathbb{Q}_Y$. Putting everything together, since we proved that all distributions $\mathbb{P}_{\varphi(X,U)}$, $\mathbb{P}_Y$, $\mathbb{Q}_{\varphi(X,U)}$, $\mathbb{Q}_Y$ are equal to each other, we obtain Item 1, concluding the proof. ■

Appendix F. Missing proofs from Section 3

This section is devoted to proving the main result of the paper: under the two-bit feedback model, every learner suffers at least $\Omega(T^{3/4})$ regret, even if it is allowed to post two different prices, one to the seller and one (larger) to the buyer.

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 $T \ge 8008$, then any learning algorithm $\mathcal{A}$ posting two prices per time step suffers at least a regret of*

$$R_T(\mathcal{A}) \ge \frac{1}{50^3} T^{3/4} \ .$$

Proof We prove this result in several steps: we begin by constructing a hard instance of the learning problem, then we present a related (easier) learning problem and, finally, we show that the minimax regret of the latter (and therefore, the former) is at least $\Omega(T^{3/4})$.

The construction of a hard family of adversaries

Fix any $\sigma \in (0, 1/9]$ and $T \geq 8008$. Since the regret against an i.i.d. adversary is entirely characterized by the distribution the adversary uses to draw seller/buyer valuations, we will (and it is equivalent to) model the adversary's behavior with probability measures rather than strategies $\mathcal{S}$. More precisely, we will model the adversary with a single sequence of seller/buyer valuations $(S, B), (S_1, B_1), \dots$ whose distribution we can change by changing the underlying probability measure. For any strategy $\mathcal{A}$ of the learner, we will find an underlying probability measure such that the process $(S, B), (S_1, B_1), \dots$ is σ-smooth, i.i.d., independent of the player's randomization, and it satisfies

$$\max_{p \in [0,1]} T\mathbb{E}\big[\mathrm{GFT}(p, p, S, B)\big] - \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GFT}(P_t, Q_t, S_t, B_t)\right] \geq \frac{1}{50^3} T^{3/4} .$$

Let $a \coloneqq 2 \cdot \ln(27/16)$. Define the six disjoint squares (Figure 1, left)

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

Fix the base probability density function $f \colon [0,1]^2 \to [0, \infty)$ defined for all $(x, y) \in [0,1]^2$ by

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

We define a set of perturbations of f parameterized by the elements of

$$\Xi \coloneqq \left\{ (v, \varepsilon) \in \left(\tfrac{1}{3}, \tfrac{1}{2}\right) \times \left(0, \tfrac{1}{12}\right) \mid \tfrac{1}{3} + \varepsilon \leq v \leq \tfrac{1}{2} - \varepsilon \right\} .$$

For all $(v, \varepsilon) \in \Xi$, define the four disjoint rectangles (Figure 1, left)

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

and the corresponding perturbation $g_{v,\varepsilon} \colon [0,1]^2 \to \mathbb{R}$ defined for all $(x, y) \in [0,1]^2$ by

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

Note that the rectangles $R^i_{v,\varepsilon}$ are included in Q_6 for all $i \in [4]$ and $(v, \varepsilon) \in \Xi$. We define perturbed density functions by summing together the base probability density function f and one of the perturbations above. Formally, for all $(v, \varepsilon) \in \Xi$, we let

$$f_{v,\varepsilon} \coloneqq f + g_{v,\varepsilon} .$$

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., $\mathbb{E}^{v,\varepsilon}$). Note that $\mathbb{P}_{(S,B)}$ (resp., $\mathbb{P}^{v,\varepsilon}_{(S,B)}$, for all $(v,\varepsilon) \in \Xi$) is $1/9$-smooth (hence, it is σ-smooth). Note also that, for each $(v,\varepsilon) \in \Xi$, and $p \in [0,1]$,

$$\begin{aligned}\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,p,S,B)\big] &= \mathbb{E}\big[\mathrm{GFT}(p,p,S,B)\big] + \int_{[0,p]\times[p,1]} (y-x) g_{v,\varepsilon}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \mathbb{E}\big[\mathrm{GFT}(p,p,S,B)\big] + \frac{1}{6(1+8a)} \cdot \frac{\varepsilon}{144} \cdot \Lambda_{v,\varepsilon}(p) + \frac{1}{6(1+8a)} \cdot \frac{\varepsilon^2}{12} \cdot \Lambda_{\frac{3}{4},\frac{1}{12}}(p) \ ,\end{aligned}$$

where, for each $u \in \mathbb{R}$ and each $r > 0$, $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as

$$\Lambda_{u,r} \colon \mathbb{R} \to \mathbb{R} \ , \qquad x \mapsto \left(1 - \frac{|x-u|}{r}\right)^+ \ .$$

A direct computation shows that, for each $p \in [0,1]$

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

from which it can be seen that the function $p \mapsto \mathbb{E}\big[\mathrm{GFT}(p,p,S,B)\big]$ is continuous and maximized at every point of the plateau region $\big[\frac{1}{6},\frac{1}{2}\big]$ (Figure 1, right). Putting everything together, we see that, for each $(v,\varepsilon) \in \Xi$, the point v is the unique maximizer of the perturbed function $p \mapsto \mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,p,S,B)\big]$, 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 $\big[\frac{1}{2},1\big]$. Given that, regardless which is the underlying distribution, the expected gain from trade is maximized on the diagonal $\{(p,q) \in [0,1]^2 \mid p = q\}$, it follows that for each $(v,\varepsilon) \in \Xi$,

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

where we recall that $\mathcal{U}$ is the upper triangle.

Now, we show that the distribution of the 2-bit feedback $\big(\mathbb{I}\{S \le p\}, \mathbb{I}\{q \le B\}\big)$ 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 where the perturbations occur.

Claim 4 *For all $(v,\varepsilon) \in \Xi$, $(p,q) \in \mathcal{U} \setminus \bigcup_{k\in[4]} R^k_{v,\varepsilon}$, and $(i,j) \in \{0,1\}^2$, it holds*

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \le p\},\, \mathbb{I}\{q \le B\}\big) = (i,j)\Big] = \mathbb{P}\Big[\big(\mathbb{I}\{S \le p\},\, \mathbb{I}\{q \le B\}\big) = (i,j)\Big] \ .$$

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 $\big(\mathbb{I}\{S \le p\}, \mathbb{I}\{q \le B\}\big)$ is given, for all $(i,j) \in \{0,1\}^2$, by

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \le p\},\, \mathbb{I}\{q \le B\}\big) = (i,j)\Big] = \begin{cases} \mathbb{P}^{v,\varepsilon}\big[S > p \cap B < q\big] & \text{if } (i,j) = (0,0) \\ \mathbb{P}^{v,\varepsilon}\big[S > p \cap B \ge q\big] & \text{if } (i,j) = (0,1) \\ \mathbb{P}^{v,\varepsilon}\big[S \le p \cap B < q\big] & \text{if } (i,j) = (1,0) \\ \mathbb{P}^{v,\varepsilon}\big[S \le p \cap B \ge q\big] & \text{if } (i,j) = (1,1) \end{cases}$$

$$= \begin{cases} \int_{(p,1]\times[0,q)} f(x,y)\mathrm{d}x\mathrm{d}y + \int_{(p,1]\times[0,q)} g_{v,\varepsilon}(x,y)\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (0,0) \\ \int_{(p,1]\times[q,1]} f(x,y)\mathrm{d}x\mathrm{d}y + \int_{(p,1]\times[q,1]} g_{v,\varepsilon}(x,y)\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (0,1) \\ \int_{[0,p]\times[0,q)} f(x,y)\mathrm{d}x\mathrm{d}y + \int_{[0,p]\times[0,q)} g_{v,\varepsilon}(x,y)\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (1,0) \\ \int_{[0,p]\times[q,1]} f(x,y)\mathrm{d}x\mathrm{d}y + \int_{[0,p]\times[q,1]} g_{v,\varepsilon}(x,y)\mathrm{d}x\mathrm{d}y & \text{if } (i,j) = (1,1) \end{cases}$$

and noting that, by symmetry, all integrals of $g_{v,\varepsilon}$ in the previous formula vanish if (p,q) does not belong to one of the four rectangles $R^1_{v,\varepsilon}, R^2_{v,\varepsilon}, R^3_{v,\varepsilon}, R^4_{v,\varepsilon}$, we get that $(p,q) \notin R^1_{v,\varepsilon} \cup R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon} \cup R^4_{v,\varepsilon}$ implies

$$\mathbb{P}^{v,\varepsilon}\Big[\big(\mathbb{I}\{S \le p\},\, \mathbb{I}\{q \le B\}\big) = (i,j)\Big] = \mathbb{P}\Big[\big(\mathbb{I}\{S \le p\},\, \mathbb{I}\{q \le B\}\big) = (i,j)\Big] \,.$$

■

It follows that, for any fixed $\varepsilon \in \big(0, \frac{1}{12}\big)$, if the learner wants to locate $v \in \big[\frac{1}{3}+\varepsilon, \frac{1}{2}-\varepsilon\big]$ observing samples of the 2-bit feedback drawn according to the distribution $\mathbb{P}^{v,\varepsilon}$, since $R^1_{v,\varepsilon} \cup R^2_{v,\varepsilon} \cup R^3_{v,\varepsilon} \cup R^4_{v,\varepsilon} \subset Q_6$, they have to post prices in the region Q_6. However, note that for each $(v,\varepsilon) \in \Xi$ and $(p,q) \in Q_6$

$$\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,q,S,B)\big] \le \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\frac{1}{2},\frac{2}{3},S,B\right)\right] \le \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\frac{2}{3},\frac{2}{3},S,B\right)\right]$$

while posting prices (p',p') for p' belonging to the potentially optimal region $\big[\frac{1}{3},\frac{1}{2}\big]$ would return

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

Hence, for each $(v,\varepsilon) \in \Xi$, each $p' \in \big[\frac{1}{3},\frac{1}{2}\big]$ and each $(p,q) \in Q_6$, we have

$$\begin{aligned} &\mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p',p',S,B)\big] - \mathbb{E}^{v,\varepsilon}\big[\mathrm{GFT}(p,q,S,B)\big] \\ &\ge \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\frac{1}{2},\frac{1}{2},S,B\right)\right] - \mathbb{E}^{v,\varepsilon}\left[\mathrm{GFT}\left(\frac{2}{3},\frac{2}{3},S,B\right)\right] = \frac{a}{2(1+8a)} \in [0.05, 0.06] = \Theta(1) \end{aligned}$$

which means that the learner suffers an instantaneous regret of order $\Theta(1)$ when trying to locate where the perturbation occurs.

Define $K := \lceil T^{1/4} \rceil$ and $\varepsilon := \frac{1}{2K}$. For each $k \in \{0,\dots,K\}$, define $v_k := \frac{1}{3} + (2k-1)\frac{\varepsilon}{6}$. For the sake of convenience, for each $k \in [K]$ denote $\mathbb{P}^{v_k,\frac{\varepsilon}{6}}$ by $\mathbb{P}^k$ and the corresponding expectation by $\mathbb{E}^k$, and similarly, denote $\mathbb{P}$ by $\mathbb{P}^0$ and the corresponding expectation by $\mathbb{E}^0$.

Interlude

Before proceeding further, let's recap what we have obtained so far and where we plan to go. At a high level, we built a problem in which we know in advance the region where the optimal pair of prices belongs (i.e., the diagonal $\{(p,q) \in [0,1]^2 \mid p = q \in [\frac{1}{3}, \frac{1}{2}]\}$), but, when the underlying scenario is determined by the probability measure $\mathbb{P}^k$ for some $k \in [K]$, in order not to suffer regret $\Omega(\varepsilon T)$, the learner has to detect inside this potentially optimal region where a spike of height (and base) $\Theta(\varepsilon)$ in the reward occurs. This last task can be accomplished only by locating where the perturbation in the base probability measure occurs, which, given the feedback structure, can only be done by playing in the costly region Q_6, suffering instantaneous regret of order $\Omega(1)$ whenever doing so. However, the region Q_6 can be further partitioned into $\Theta(\frac{1}{\varepsilon})$ disjoint rectangles where these perturbations can occur, and again, given the feedback structure, this implies that each of these rectangles deserves its own dedicated exploration. To better highlight this underlying structure, we will show that the bilateral trade problem is no easier than a simplified problem (that we call multi-apple tasting) where the learner can play $2K$ actions, which we may identify with the set $[2K]$, and where the instances we consider are determined by the probability measures $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$. Each (exploring) action $i \in [K]$ gives zero reward (and corresponds to one of the $\Theta(\frac{1}{\varepsilon})$ rectangles inside the region Q_6), but, if played at time $t \in \mathbb{N}$, it reveals the realization of a Bernoulli random variable $Y_t(i)$ which is, up to a rescaling and a shifting, the reward of the corresponding (exploiting) action $i + K$ at time t. (The reader familiar with the notion of online learning with directed feedback graphs Alon et al. 2015 can see that the feedback model described here corresponds to the weakly observable feedback graph in Figure 2, left). The biases of these Bernoullis depend on which is the underlying probability measure among $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$. Specifically, for each $i \in [K]$, each $k \in \{0, \dots, K\}$, and each $t \in \mathbb{N}$, the bias of $Y_t(i)$ under $\mathbb{P}^k$ is $\frac{1}{2}$ if $i \neq k$, while it is $\frac{1}{2} + \Theta(\varepsilon)$ if $i = k$. This way, the exploiting actions $K+1, \dots, 2K$ (which correspond to the regions where the spike in the expected gain from trade can occur) have an expected reward of order $\Omega(1)$ regardless of the underlying probability measure, so that the potentially optimal arm is among them. The catch is that no informative feedback is revealed by these K exploiting actions, and only one of them is optimal when the underlying probability measure is one among $\mathbb{P}^1, \dots, \mathbb{P}^K$. Specifically, the arm $i + K$ is the only optimal action when the underlying probability measure is $\mathbb{P}^i$, having an expected reward that is $\Theta(\varepsilon)$ higher that the other potentially optimal actions. Therefore, since spotting the Bernoulli random variable with bias $\frac{1}{2} + \Theta(\varepsilon)$ among the other $K - 1$ unbiased Bernoullis requires playing the K exploring actions $\Theta\big(\frac{1}{\varepsilon^2}\big)$ times each, any algorithm for this new problem (and hence, for the bilateral trade problem) should suffer a regret of order $\Omega\big(\min\big(\frac{K}{\varepsilon^2}, \varepsilon T\big)\big) = \Omega(T^{3/4})$ in at least one scenario among $\mathbb{P}^0, \mathbb{P}^1, \dots, \mathbb{P}^K$, given our choices of K and ε. We will now formalize this idea.

The multi-apple tasting problem

We now described the multi-apple tasting problem on $2K$ arms.

Pick a sequence of $\{0,1\}^{2K}$-valued random variables $Y, Y_1, \dots, Y_T$ and a sequence of $[0,1]$-valued random variables $U, U_1, \dots, U_T, V, V_1, \dots, V_T$ such that:

- For each $k \in \{0, \dots, K\}$ the sequence $Y, Y_1, \dots, Y_T$ is $\mathbb{P}^k$-i.i.d.

- Letting $c_{\text{prob}} \coloneqq \frac{7}{2a}$, for each $k \in \{0,\dots,K\}$ and each $i \in [K]$ we have that $Y(i+K) = Y_1(i+K) = \dots = Y_T(i+K) = 0$ and

$$\mathbb{P}^k[Y(i)=1] = \begin{cases} \frac{1}{2} & \text{if } i \in [K]\backslash\{k\} \\ \frac{1}{2} + c_{\text{prob}} \cdot \varepsilon & \text{if } i = k \end{cases}$$

- For each $k \in \{0,\dots,K\}$ the sequence $V, V_1, \dots, V_T$ is $\mathbb{P}^k$-i.i.d. and $\mathbb{P}^k_V = \mathbb{L}$.
- For each $k \in \{0,\dots,K\}$, we have

$$\mathbb{P}^k_{\big(((S,B),(S_1,B_1),\dots,(S_T,B_T)),(U,U_1,\dots U_T),(Y,Y_1,\dots Y_T),(V,V_1,\dots V_T)\big)} \\ = \mathbb{P}^k_{((S,B),(S_1,B_1),\dots,(S_T,B_T))} \otimes \mathbb{P}^k_{(U,U_1,\dots U_T)} \otimes \mathbb{P}^k_{(Y,Y_1,\dots Y_T)} \otimes \mathbb{P}^k_{(V,V_1,\dots V_T)}$$

The multi-apple tasting problem proceeds as follows. At each time $t \in [T]$, the player can play any action i in the set $[2K]$, receiving no feedback if $i \geq K+1$ (modeled by $Y(i) = Y_1(i) = \dots = Y_T(i) = 0$) and feedback $Y_t(i)$ if $i \in [K]$, obtaining in any case (but not observing) a reward $\rho(i, Y_t)$, where letting $c_{\text{plat}} \coloneqq \frac{a}{2(1+8a)}$ and $c_{\text{spike}} \coloneqq \frac{1}{6(1+8a)} \cdot \frac{1}{144}$,

$$\rho \colon [2K] \times \{0,1\}^{2K} \to \mathbb{R} \ , \qquad (j,y) \mapsto \begin{cases} 0 & \text{if } j \in [K] \\ c_{\text{plat}} + \frac{c_{\text{spike}}}{c_{\text{prob}}} \cdot \big(y(j-K) - \frac{1}{2}\big) & \text{otherwise} \end{cases}$$

Observe that for all $k \in \{0,\dots,K\}$ and $i \in \{K+1,\dots,2K\}$, we have

$$\mathbb{E}^k\big[\rho(i,Y)\big] = \begin{cases} c_{\text{plat}} & \text{if } k \neq i-K \\ c_{\text{plat}} + c_{\text{spike}} \cdot \varepsilon & \text{otherwise} \end{cases}$$

Relating the two problems

To map the bilateral trade problem into the multi-apple tasting problem, we first partition the upper triangle $\mathcal{U}$ in the following $2K$ disjoint regions:

- $\forall k \in [K-1]$, $J_k \coloneqq [v_k - \frac{\varepsilon}{6}, v_k + \frac{\varepsilon}{6}) \times [\frac{2}{3}, \frac{5}{6}]$
- $J_K \coloneqq [v_K - \frac{\varepsilon}{6}, v_K + \frac{\varepsilon}{6}] \times [\frac{2}{3}, \frac{5}{6}]$
- $\forall k \in [K-1]$, $J_{k+K} \coloneqq \{(p,q) \in \mathcal{U} \mid v_k - \frac{\varepsilon}{6} \leq p < v_k + \frac{\varepsilon}{6}$ and $q < \frac{2}{3}\}$
- $J_{2K} \coloneqq \mathcal{U} \backslash \bigcup_{k=1}^{2K-1} J_k$

Define $\iota \colon \mathcal{U} \to [2K]$ as the map that associates to each $(p,q) \in \mathcal{U}$ the unique $i \in [2K]$ such that $(p,q) \in J_i$ (Figure 2, right).

Claim 5 *For any $(p,q) \in \mathcal{U}$ there exists a function $\varphi_{p,q} : \{0,1\} \times [0,1] \to \{0,1\}^2$ such that, for all $k \in \{0,\dots,K\}$, the distributions under $\mathbb{P}^k$ of $\varphi_{p,q}(Y(\iota(p,q)),V)$ and $\big(\mathbb{I}(S \leq p), \mathbb{I}\{q \leq B\}\big)$ coincide.*

Proof A direct verification shows that, for all $(p,q) \in Q_6$ and $k \in [K]$, it holds that

$$\min\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(k)}}{\mathrm{d}\mathbb{P}^0_{Y(k)}}\right) = 1 - 2c_{\text{prob}} \cdot \varepsilon \leq \frac{\mathrm{d}\mathbb{P}^k_{(\mathbb{I}(S\leq p),\mathbb{I}\{q\leq B\})}}{\mathrm{d}\mathbb{P}^0_{(\mathbb{I}(S\leq p),\mathbb{I}\{q\leq B\})}} \leq 1 + 2c_{\text{prob}} \cdot \varepsilon = \max\left(\frac{\mathrm{d}\mathbb{P}^k_{Y(k)}}{\mathrm{d}\mathbb{P}^0_{Y(k)}}\right)$$

and $\mathbb{P}^k_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})} \ll \mathbb{P}^0_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}$. For each $(p,q)\in Q_6$, by Theorem 10, there exists (and we fix)

$$\varphi_{p,q}\colon \{0,1\}\times[0,1]\to\{0,1\}^2$$

such that

$$\mathbb{P}^{\iota(p,q)}_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^{\iota(p,q)}_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})} \qquad \text{and} \qquad \mathbb{P}^{0}_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^{0}_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}\,.$$

Since for all $(p,q)\in Q_6$ and all $k\in[K]\setminus\big\{\iota(p,q)\big\}$, we have $\mathbb{P}^k_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})} = \mathbb{P}^0_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}$ (by Claim 4) and $\mathbb{P}^k_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^0_{\varphi_{p,q}(Y(\iota(p,q)),V)}$, then, for all $(p,q)\in Q_6$ and all $k\in\{0,\dots,K\}$, it holds that

$$\mathbb{P}^{k}_{\varphi_{p,q}(Y(\iota(p,q)),V)} = \mathbb{P}^{k}_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}\,.$$

Moreover, since for all $(p,q)\in\mathcal{U}\setminus Q_6$ and for all $k\in\{0,\dots,K\}$, it holds that $\mathbb{P}^k_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})} = \mathbb{P}^0_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}$ (by Claim 4), then, by Theorem 9, there exists (and we fix)

$$\widetilde{\varphi}_{p,q}\colon [0,1]\to\{0,1\}^2$$

such that, for all $k\in\{0,\dots,K\}$, it holds that

$$\mathbb{P}^{k}_{\widetilde{\varphi}_{p,q}(V)} = \mathbb{P}^{k}_{(\mathbb{I}(S\le p),\mathbb{I}\{q\le B\})}\,.$$

Defining for all $(p,q)\in\mathcal{U}\setminus Q_6$ and $(y,v)\in\{0,1\}\times[0,1]$, $\varphi_{p,q}(y,v)\coloneqq\widetilde{\varphi}_{p,q}(v)$, we obtain the result. ■

For all $(p,q)\in\mathcal{U}$, fix a $\varphi_{p,q}$ as in Claim 5. Now, fix an arbitrary weakly-budget-balanced algorithm $\mathcal{A}$ for the bilateral trade problem with two-bit feedback. If needed, $\mathcal{A}$ has sequential access to the seeds $U_1,U_2,\dots$ for randomization purposes. Let $(P_1,Q_1),(P_2,Q_2),\dots$ be the sequence of prices posted by the algorithm $\mathcal{A}$ observing the two-bit feedback $\big(\mathbb{I}\{S_t\le P_t\},\mathbb{I}\{Q_t\le B_t\}\big)$ at round t. We now construct an algorithm $\widetilde{\mathcal{A}}$ (based on $\mathcal{A}$ and the sequence of random seeds $V_1,V_2,\dots$) to solve this new problem in the following way:

- For each time $t\in[T]$, we use the algorithm $\mathcal{A}$ to select a pair $(\widetilde{P}_t,\widetilde{Q}_t)\in\mathcal{U}$, then play the action $\widetilde{I}_t\coloneqq\iota(\widetilde{P}_t,\widetilde{Q}_t)\in[2K]$.
- For each time $t\in[T]$, whenever the algorithm $\mathcal{A}$ requests some feedback in $\{0,1\}^2$, we feed $\mathcal{A}$ with the feedback $\varphi_{\widetilde{P}_t,\widetilde{Q}_t}\big(Y_t(\widetilde{I}_t),V_t\big)\in\{0,1\}^2$.

By induction on t, Claim 5 implies that for all $k\in\{0,\dots,K\}$ and $t\in[T]$, we have

$$\mathbb{P}^k_{(\widetilde{P}_t,\widetilde{Q}_t)} = \mathbb{P}^k_{(P_t,Q_t)}$$

which, together with the fact that $\mathbb{P}^k_{(\widetilde{P}_t,\widetilde{Q}_t,Y_t)} = \mathbb{P}^k_{(\widetilde{P}_t,\widetilde{Q}_t)} \otimes \mathbb{P}^k_{Y_t}$ for all $k \in \{0,\dots,K\}$ and $t \in [T]$, yields

$$
\begin{aligned}
R_T^k(\mathcal{A}) &\coloneqq T\mathbb{E}^k\big[\mathrm{GFT}(v_k,v_k,S,B)\big] - \sum_{t=1}^T \mathbb{E}^k\big[\mathrm{GFT}(P_t,Q_t,S_t,B_t)\big] \\
&\ge T\mathbb{E}^k\big[\rho(k+K,Y)\big] - \sum_{t=1}^T \mathbb{E}^k\Big[\rho\big(\iota(P_t,Q_t),Y_t\big)\Big] \\
&= T\mathbb{E}^k\big[\rho(k+K,Y)\big] - \sum_{t=1}^T \mathbb{E}^k\Big[\rho\big(\iota(\widetilde{P}_t,\widetilde{Q}_t),Y_t\big)\Big] \\
&= T\mathbb{E}^k\big[\rho(k+K,Y)\big] - \sum_{t=1}^T \mathbb{E}^k\big[\rho(\widetilde{I}_t,Y_t)\big] \eqqcolon \widetilde{R}_T^k(\widetilde{\mathcal{A}}) \ ,
\end{aligned}
$$

where $R_T^k(\mathcal{A})$ (resp., $\widetilde{R}_T^k(\widetilde{\mathcal{A}})$) is the regret suffered by the algorithm $\mathcal{A}$ (resp., $\widetilde{\mathcal{A}}$) after T rounds of the bilateral trade problem with two-bit feedback (resp., the related problem on $2K$ actions) in the scenario $\mathbb{P}^k$. Summing over $k \in [K]$ and dividing by K, this implies

$$
\frac{1}{K}\sum_{k\in[K]} R_T^k(\mathcal{A}) \ge \frac{1}{K}\sum_{k\in[K]} \widetilde{R}_T^k(\widetilde{\mathcal{A}}) \ge \inf_{\bar{\mathcal{A}}\in\mathrm{Rand}} \frac{1}{K}\sum_{k\in[K]} \widetilde{R}_T^k(\bar{\mathcal{A}}) = \inf_{\bar{\mathcal{A}}\in\mathrm{Det}} \frac{1}{K}\sum_{k\in[K]} \widetilde{R}_T^k(\bar{\mathcal{A}}) \ ,
$$

where the first (resp., second) infimum is over the set Rand (resp., Det) all randomized (resp., deterministic) algorithms $\bar{\mathcal{A}}$ for the related problem on $2K$ actions, and the last standard equality is a straightforward consequence of the stochastic i.i.d. setting.

We now show that for any deterministic algorithm $\bar{\mathcal{A}}$ for the related problem on $2K$ actions, it either holds that $\frac{1}{K}\sum_{k\in[K]} \widetilde{R}_T^k(\bar{\mathcal{A}}) \ge \frac{1}{50^3}T^{3/4}$ or that $\widetilde{R}_T^0(\bar{\mathcal{A}}) \ge \frac{1}{50^3}T^{3/4}$. This, together with the inequalities above will imply that there exists an $k \in \{0,\dots,K\}$ such that $R_T^k(\mathcal{A}) \ge \frac{1}{50^3}T^{3/4}$, concluding the proof. For any deterministic algorithm $\bar{\mathcal{A}}$ for the related problem on $2K$ actions, let $I_1^{\bar{\mathcal{A}}}, I_2^{\bar{\mathcal{A}}},\dots$ be the actions played by $\bar{\mathcal{A}}$ on the basis of the sequential feedback $Z_1^{\bar{\mathcal{A}}}, Z_2^{\bar{\mathcal{A}}},\dots$ and

$$
N_t^{\bar{\mathcal{A}}} \coloneqq \sum_{i\in[K]} N_t^{\bar{\mathcal{A}}}(i) \ , \ M_t^{\bar{\mathcal{A}}} \coloneqq \sum_{i\in[K]} M_t^{\bar{\mathcal{A}}}(i) \ ,
$$

$$
\text{where } N_t^{\bar{\mathcal{A}}}(i) \coloneqq \sum_{s=1}^t \mathbb{I}\{I_s^{\bar{\mathcal{A}}} = i\} \ , \ M_t^{\bar{\mathcal{A}}}(i) \coloneqq \sum_{s=1}^t \mathbb{I}\{I_s^{\bar{\mathcal{A}}} = i+K\} \ .
$$

Fix an arbitrary deterministic algorithm $\bar{\mathcal{A}}$ for the related problem on $2K$ actions. Then

$$
\begin{aligned}
\frac{1}{K}\sum_{k\in[K]} \widetilde{R}_T^k(\bar{\mathcal{A}}) &= \frac{1}{K}\sum_{k\in[K]} \Big(c_{\mathrm{spike}}\cdot\varepsilon\cdot\mathbb{E}^k\big[T - M_T^{\bar{\mathcal{A}}}(k) - N_T^{\bar{\mathcal{A}}}\big] + (c_{\mathrm{plat}} + c_{\mathrm{spike}}\cdot\varepsilon)\cdot\mathbb{E}^k\big[N_T^{\bar{\mathcal{A}}}\big] \Big) \\
&\ge c_{\mathrm{spike}}\cdot\varepsilon\left(T - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}^k\big[M_T^{\bar{\mathcal{A}}}(k)\big] \right) \eqqcolon (\circ)
\end{aligned}
$$

Now, since for any $t \in [T]$ the action $I_t^{\bar{\mathcal{A}}} = \bar{\mathcal{A}}_t(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})$ selected by $\bar{\mathcal{A}}$ at round t is a deterministic function of $Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}}$, for each $k \in [K]$, we have

$$\begin{aligned}
&\mathbb{E}^k\big[M_T^{\bar{\mathcal{A}}}(k)\big] - \mathbb{E}^0\big[M_T^{\bar{\mathcal{A}}}(k)\big] \\
&\quad = \sum_{t=2}^{T}\Big(\mathbb{P}^k\big[\bar{\mathcal{A}}_t(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}}) = k+K\big] - \mathbb{P}^0\big[\bar{\mathcal{A}}_t(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}}) = k+K\big]\Big) \\
&\quad = \sum_{t=2}^{T}\Big(\mathbb{P}^k_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})}\big[\bar{\mathcal{A}}_t^{-1}(k+K)\big] - \mathbb{P}^0_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})}\big[\bar{\mathcal{A}}_t^{-1}(k+K)\big]\Big) \\
&\quad \le \sum_{t=2}^{T}\big\|\mathbb{P}^k_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})} - \mathbb{P}^0_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})}\big\|_\infty \le \sum_{t=2}^{T}\big\|\mathbb{P}^k_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})} - \mathbb{P}^0_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{t-1}^{\bar{\mathcal{A}}})}\big\|_{\mathrm{TV}} =: (\star)
\end{aligned}$$

were we $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm. We will now prove that, for each $k \in [K]$ and $t \in [T]$, it holds that

$$\Big\|\mathbb{P}^0_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_t^{\bar{\mathcal{A}}})} - \mathbb{P}^k_{(Z_1^{\bar{\mathcal{A}}}, \dots, Z_t^{\bar{\mathcal{A}}})}\Big\|_{\mathrm{TV}} \le c_{\mathrm{prob}} \cdot \varepsilon \cdot \sqrt{2\mathbb{E}[N_t^{\bar{\mathcal{A}}}(k)]} \tag{9}$$

By Pinsker's inequality and the chain rule for KL-divergence $\mathcal{D}_{\mathrm{KL}}$, for each $k \in [K]$ and $t \in [T]$, we have

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

To upper bound $(@)$, note first that, since $T \ge 8008$,

$$\frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}} \cdot \varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}} \cdot \varepsilon}\right) \le 4 \cdot c_{\mathrm{prob}}^2 \cdot \varepsilon^2$$

Then, since $\bar{\mathcal{A}}$ is a deterministic algorithm, $I_1^{\bar{\mathcal{A}}}$ is a fixed element of $[2K]$, which implies that, for all $k \in [K]$,

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_1^{\bar{\mathcal{A}}}}, \mathbb{P}^k_{Z_1^{\bar{\mathcal{A}}}}\big) \\
&\quad = \left(\ln\left(\frac{\mathbb{P}^0[Y_1(k) = 0]}{\mathbb{P}^k[Y_1(k) = 0]}\right)\mathbb{P}^0[Y_1(k) = 0] + \ln\left(\frac{\mathbb{P}^0[Y_1(k) = 1]}{\mathbb{P}^k[Y_1(k) = 1]}\right)\mathbb{P}^0[Y_1(k) = 1]\right)\mathbb{I}\big\{I_1^{\bar{\mathcal{A}}} = k\big\} \\
&\quad = \frac{1}{2}\left(\ln\frac{1/2}{1/2 - c_{\mathrm{prob}} \cdot \varepsilon} + \ln\frac{1/2}{1/2 + c_{\mathrm{prob}} \cdot \varepsilon}\right) \cdot \mathbb{I}\{I_1^{\bar{\mathcal{A}}} = k\} \le 4 \cdot c_{\mathrm{prob}}^2 \cdot \varepsilon^2 \cdot \mathbb{P}^0[I_1^{\bar{\mathcal{A}}} = k]
\end{aligned}$$

Similarly, since $\bar{\mathcal{A}}$ is a deterministic algorithm, for all $s \geq 2$, the action $I_s^{\bar{\mathcal{A}}} = \bar{\mathcal{A}}_s(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}})$ selected by $\bar{\mathcal{A}}$ at time t a function of $Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}$ only, which implies, for all $k \in [K]$,

$$
\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^0_{Z_s^{\bar{\mathcal{A}}} \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}}, \mathbb{P}^k_{Z_s^{\bar{\mathcal{A}}} \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}}\big) \\
&\quad = \mathbb{E}^0 \Bigg[\ln \left(\frac{\mathbb{P}^0[Z_s^{\bar{\mathcal{A}}} = 0 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}]}{\mathbb{P}^k[Z_s^{\bar{\mathcal{A}}} = 0 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}]} \right) \mathbb{P}^0[Z_s^{\bar{\mathcal{A}}} = 0 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}] \\
&\qquad\quad + \ln \left(\frac{\mathbb{P}^0[Z_s^{\bar{\mathcal{A}}} = 1 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}]}{\mathbb{P}^k[Z_s^{\bar{\mathcal{A}}} = 1 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}]} \right) \mathbb{P}^0[Z_s^{\bar{\mathcal{A}}} = 1 \mid Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}] \Bigg] \\
&\quad = \mathbb{E}^0 \Bigg[\left(\ln \left(\frac{\mathbb{P}^0[Y_s(k) = 0]}{\mathbb{P}^k[Y_s(k) = 0]} \right) \mathbb{P}^0[Y_s(k) = 0] + \ln \left(\frac{\mathbb{P}^0[Y_s(k) = 1]}{\mathbb{P}^k[Y_s(k) = 1]} \right) \mathbb{P}^0[Y_s(k) = 1] \right) \\
&\qquad\quad \times \mathbb{I}\big\{ \bar{\mathcal{A}}_s(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}) = k \big\} \Bigg] \\
&\quad = \frac{1}{2} \left(\ln \frac{1/2}{1/2 - c_{\mathrm{prob}} \cdot \varepsilon} + \ln \frac{1/2}{1/2 + c_{\mathrm{prob}} \cdot \varepsilon} \right) \mathbb{P}^0 \big[\bar{\mathcal{A}}_s(Z_1^{\bar{\mathcal{A}}}, \dots, Z_{s-1}^{\bar{\mathcal{A}}}) = k \big] \\
&\quad \leq 4 \cdot c_{\mathrm{prob}}^2 \cdot \varepsilon^2 \cdot \mathbb{P}^0[I_s^{\bar{\mathcal{A}}} = k] \, .
\end{aligned}
$$

Plugging the two bounds in (@), we get, for all $k \in [K]$ and $t \in [T]$,

$$
(@) \leq \sqrt{2 \cdot c_{\mathrm{prob}}^2 \cdot \varepsilon^2 \cdot \sum_{s=1}^{t} \mathbb{P}^0[I_s^{\bar{\mathcal{A}}} = k]} \leq c_{\mathrm{prob}} \cdot \varepsilon \cdot \sqrt{2\mathbb{E}^0[N_t^{\bar{\mathcal{A}}}(k)]}
$$

which prove claim (9). Therefore, we have, for any $k \in [K]$,

$$
\mathbb{E}^k \big[M_T^{\bar{\mathcal{A}}}(k) \big] - \mathbb{E}^0 \big[M_T^{\bar{\mathcal{A}}}(k) \big] \leq (\star) \leq \sum_{t=2}^{T} c_{\mathrm{prob}} \cdot \varepsilon \cdot \sqrt{2\mathbb{E}^0[N_{t-1}^{\bar{\mathcal{A}}}(k)]} \leq c_{\mathrm{prob}} \cdot \varepsilon \cdot T \cdot \sqrt{2\mathbb{E}^0[N_T^{\bar{\mathcal{A}}}(k)]} \, .
$$

Rearranging, averaging, applying Jensen's inequality, and recalling that $\frac{1}{K} = \frac{1}{\lceil T^{1/4} \rceil} \leq \frac{1}{10}$, we obtain

$$
\begin{aligned}
\frac{1}{K} \sum_{k \in [K]} \mathbb{E}^k[M_T^{\bar{\mathcal{A}}}(k)] &\leq \frac{1}{K} \sum_{k \in [K]} \mathbb{E}^0[M_T^{\bar{\mathcal{A}}}(k)] + c_{\mathrm{prob}} \cdot \varepsilon \cdot T \cdot \sqrt{2\mathbb{E}^0 \left[\frac{1}{K} \sum_{k \in [K]} N_T^{\bar{\mathcal{A}}}(k) \right]} \\
&= \frac{1}{K} \mathbb{E}^0[M_T^{\bar{\mathcal{A}}}] + c_{\mathrm{prob}} \cdot \varepsilon \cdot T \cdot \sqrt{\frac{2}{K} \mathbb{E}^0[N_T^{\bar{\mathcal{A}}}]} \leq \left(\frac{1}{10} + c_{\mathrm{prob}} \cdot \varepsilon \cdot \sqrt{\frac{2}{K} \mathbb{E}^0[N_T^{\bar{\mathcal{A}}}]} \right) \cdot T \, .
\end{aligned}
$$

Substituting this inequality in $(\circ)$, we obtain

$$
(\circ) \geq c_{\mathrm{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - c_{\mathrm{prob}} \cdot \varepsilon \cdot \sqrt{\frac{2}{K} \mathbb{E}^0[N_T^{\bar{\mathcal{A}}}]} \right) \cdot T \geq c_{\mathrm{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - \frac{c_{\mathrm{prob}}}{2} \sqrt{\tau_{\bar{\mathcal{A}}}} \right) \cdot T \, ,
$$

where $\tau_{\bar{\mathcal{A}}} \coloneqq \frac{\mathbb{E}^0[N_T^{\bar{\mathcal{A}}}]}{\varepsilon T}$.

Now, if $\tau_{\bar{\mathcal{A}}} \leq \frac{1}{10}$, then, the previous inequality yields

$$\frac{1}{K} \sum_{k \in [K]} \widetilde{R}_T^k(\bar{\mathcal{A}}) \geq c_{\text{spike}} \cdot \varepsilon \cdot \left(\frac{9}{10} - \frac{c_{\text{prob}}}{2} \sqrt{\tau_{\bar{\mathcal{A}}}} \right) \cdot T \geq \frac{1}{50^3} T^{3/4} \ .$$

If, on the other hand, it holds that $\tau_{\bar{\mathcal{A}}} > \frac{1}{10}$, then

$$\widetilde{R}_T^0(\bar{\mathcal{A}}) \geq c_{\text{plat}} \mathbb{E}^0[N_T^{\bar{\mathcal{A}}}] = c_{\text{plat}} \tau_{\bar{\mathcal{A}}} \varepsilon T > \frac{1}{50^3} T^{3/4} \ .$$

■