The Role of Transparency in Repeated First-Price Auctions with Unknown Valuations*

Nicolo Cesa-Bianchi
University of Milan &
Politecnico di Milano
Milan, Italy
cesa.bianchi@gmail.com

Tommaso Cesari
University of Ottawa
Ottawa, Canada
tommaso.cesari@gmail.com

Roberto Colomboni
University of Milan &
Politecnico di Milano
Milan, Italy
robertocolomboni@gmail.com

Federico Fusco
Sapienza University of Rome
Rome, Italy
fuscof@diag.uniroma1.it

Stefano Leonardi
Sapienza University of Rome
Rome, Italy
leonardi@diag.uniroma1.it

ABSTRACT

We study the problem of regret minimization for a single bidder in a sequence of first-price auctions where the bidder discovers the item's value only if the auction is won. Our main contribution is a complete characterization, up to logarithmic factors, of the minimax regret in terms of the auction's *transparency*, which controls the amount of information on competing bids disclosed by the auctioneer at the end of each auction. Our results hold under different assumptions (stochastic, adversarial, and their smoothed variants) on the environment generating the bidder's valuations and competing bids. These minimax rates reveal how the interplay between transparency and the nature of the environment affects how fast one can learn to bid optimally in first-price auctions.

CCS CONCEPTS

• **Theory of computation** → **Algorithmic game theory**; • **Computing methodologies** → *Online learning settings*.

KEYWORDS

Online Learning, First-Price Auction, Transparency

1 INTRODUCTION

The online advertising market has recently transitioned from second to first-price auctions. A remarkable example is Google Ad-Sense's move at the end of 2021 [40], following the switch made by

*The full version of this paper can be found in Cesa-Bianchi et al. [16].

Google AdManager and AdMob. Earlier examples include OpenX, AppNexus, Index Exchange, and Rubicon [37]. To increase transparency in first-price auctions, some platforms (like AdManager) have a single bidding session for each available impression (unified bidding) and require all partners to share and receive bid data. After the first-price auction closes, bidders receive the minimum bid price that would have won them the impression [11]. In practice, advertisers face two main sources of uncertainty in the bidding phase: they ignore the value of the competing bids and, crucially, ignore the actual value of the impression they are bidding on. Indeed, clicks and conversion rates—which are only measured *after* the auction is won and the ad is displayed—can vary wildly over time or highly correlate with competing bids. We remark that ignoring the value of the impression strongly affects the bidder's utility: it may lead to overbidding for an impression of low value or, conversely, underbidding and losing a valuable one. To cope with this uncertainty, advertisers rely on auto-bidders that use the feedback provided in the auctions to learn good bidding strategies. We study the learning problem faced by a single bidder within the framework of regret minimization according to the following protocol:

Online Bidding Protocol

for $t = 1, 2, \ldots, T$ **do**
    Valuation V_t and competing bid M_t are privately generated
    The learner posts a bid B_t and receives utility $\text{Util}_t(B_t)$:

$$\text{Util}_t(B_t) = (V_t - B_t)\mathbb{I}\{B_t \geq M_t\}$$

    The learner observes some feedback Z_t

The bidder has no initial information on the environment and seeks to learn the relevant features of the problem on the fly. The performance of a learning strategy for the bidder—also referred to as the learner—is measured in terms of the difference in total utility with respect to the best fixed bid. This difference is called *regret*, and the main goal is to design strategies with asymptotically vanishing time-averaged regret with respect to the best fixed-bid strategy or, equivalently, regret sublinear in the time horizon. In this work, we are specifically interested in understanding how the "transparency" of the auctions—i.e., the amount of information on competing bids disclosed by the auctioneer *after* the auction takes place—affects the learning process. There is a clear tension regarding transparency:

Nicolo Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi

Table 1: Summary of our results. Rows correspond to feedback models (F=Full, T=Transparent, S=Semi-Transparent, and B=Bandit) while columns to environments. The minimax regret of every problem falls in one of the following three regimes: $\widetilde{\Theta}(\sqrt{T})$ (green), $\widetilde{\Theta}(T^{2/3})$ (yellow) and $\widetilde{\Theta}(T)$ (red).

	Stochastic i.i.d.		Adversarial	
	Smooth	General	Smooth	General
F	Th.5: $\Omega(\sqrt{T})$			Th.8: $\Omega(T)$
T		Th.4: $O(\sqrt{T})$	Th.7: $\widetilde{O}(\sqrt{T})$	
S	Th.3: $\Omega(T^{2/3})$	Th.2: $\widetilde{O}(T^{2/3})$		
B		Th.1: $\Omega(T)$	Th.6: $O(T^{2/3})$	

on the one hand, bidders want to receive as much information as possible about the environment to learn the competitor's bidding strategies while revealing as little as possible about their (private) bids. On the other hand, the platform may not want to publicly reveal its revenue (i.e., the winning bid). Our investigation addresses both sides of the "transparency dilemma". Our algorithmic results provide bidders with a toolbox of learning strategies to (optimally) exploit the various degrees of transparency, while the tightness of our results fully characterizes the impact of transparency on learnability. This complete picture allows platforms to make an informed decision in choosing their level of transparency, as it is in their interest to create a thriving environment for advertisers.

To model the level of transparency, we distinguish four natural types of feedback Z_t, specifying the conditions under which the highest competing bid M_t and the bidder's valuation V_t are revealed to the bidder after each round t. In the transparent feedback setting, M_t is always observed after the auction is concluded, while V_t is only known if the auction is won, i.e., when $B_t \geq M_t$. In the semi-transparent setting, M_t is only observed when the auction is lost. In other words, in the semi-transparent setting, the platform publicly reveals only the winning bid, whereas in the transparent setting, the platform reveals all bids. We also consider two extreme settings that provide two natural learning benchmarks: full feedback (M_t and V_t are always observed irrespective of the auction's outcome) and bandit feedback (M_t is never observed while V_t is only observed by the winning bidder). Note that the learner can compute the value of the utility $\text{Util}_t(B_t)$ at time t with any type of feedback, including bandit feedback. In this paper, we characterize the learner's minimax regret not only with respect to the degree of transparency of the auction but also with respect to the nature of the process generating the sequence of pairs (V_t, M_t). In particular, we consider four types of environments: stochastic i.i.d., adversarial, and their smooth versions (see Section 1.3 for a discussion about smoothness, and Section 2 for the formal definition).

1.1 Overview of Our Results

We report here an overview of our results (see also Table 1). For simplicity, we often hide the logarithmic factors with the $\widetilde{O}$ notation.

Stochastic i.i.d. settings.

- In both the full and transparent feedback models, the minimax regret is of order $\sqrt{T}$ (Theorems 4 and 5), and adding the smoothness requirement leaves this rate unchanged.

- In the semi-transparent feedback model, the minimax regret is of order $T^{2/3}$ (Theorems 2 and 3). Also in this case, adding the smoothness requirement leaves this rate unchanged.

- In the bandit feedback model, smoothness is crucial for sublinear regret (Theorem 1). In particular, smoothness implies a minimax regret of $T^{2/3}$ (this is obtained by combining the upper bound in Theorem 6 and the lower bound in Theorem 3).

Adversarial settings.

- Without smoothness, sublinear regret cannot be achieved, even with full feedback (Theorem 8).

- In both the full and transparent feedback model, the minimax regret in a smooth environment is of order $\sqrt{T}$ (combining the lower bound in Theorem 5 and the upper bound in Theorem 7).

- Both with semi-transparent and bandit feedback, the minimax regret in a smooth environment is of order $T^{2/3}$ (combining the lower bound in Theorem 3 and the upper bound in Theorem 6).

Interestingly, the minimax regret rates for first-price auctions mirror the allowed regret regimes in finite partial monitoring games [8] and online learning with feedback graphs [2]. This is somehow surprising, as it has been shown in Lattimore [33] that games with continuous outcome/action spaces allow for a much larger set of regret rates—see also Bernasconi et al. [10], Bolić et al. [13], Cesa-Bianchi et al. [15].

Table 1 reveals some interesting properties of the learnability of the problem: full feedback and transparent feedback are essentially equivalent, while semi-transparent feedback and bandit feedback differ only in the stochastic i.i.d. setting. Qualitatively, this tells the platform that disclosing all bids (instead of only the winning one) drastically improves the learnability of the problem (green vs. yellow entries in Table 1). Besides, revealing at least the winning bid avoids some pathological behavior (yellow entries vs. red entry for the general i.i.d. environment with bandit feedback). Moreover, while smoothness is key for learning in the adversarial setting, in the stochastic case smoothness is only relevant for bandit feedback.
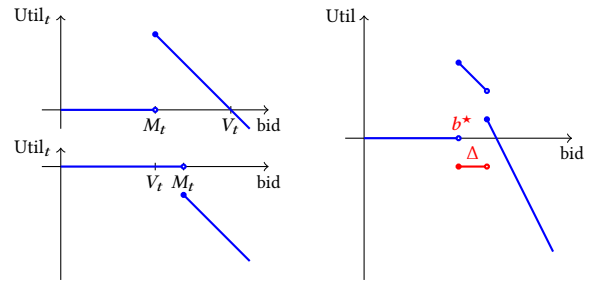

Figure 1: The utility function is generally neither Lipschitz nor continuous. If $M_t \leq V_t$ (top left plot), then Util_t is upper-semi continuous and one-sided Lipschitz; conversely, if $M_t \geq V_t$ (bottom left plot), then Util_t is still one-sided Lipschitz—from the other side—and lower-semi continuous. Summing up the two types of utilities results in a total utility that may be neither one-sided Lipschitz nor semi-continuous (right plot, where the two utility functions of the other two plots are summed up. There, $b^\star$ is the optimal bid and Δ is the neighborhood of $b^\star$ where the total utility is "good enough").

1.2 Technical Challenges

The utility function. The utilities $\text{Util}_t(b) = (V_t - b)\mathbb{I}\{M_t \leq b\}$ are defined over a continuous decision space $[0, 1]$ and are neither Lipschitz nor continuous, see Figure 1. Actually, even weaker properties, i.e., that the expected cumulative reward $b \mapsto \sum_{t \in [T]} \mathbb{E}[\text{Util}_t(b)]$ is one-sided Lipschitz or semi-continuous, do not hold in general. We address this problem by developing techniques designed to control the approximation error incurred when discretizing the bidding space. This is a non-trivial problem without regularity assumption, as the neighborhood of the optimal bid where the total utility is "good enough" can be arbitrarily small in general (see the red interval Δ in the rightmost plot of Figure 1). In the stochastic i.i.d. setting, the approximation error is controlled by building a sample-based *non-uniform* grid of candidate bids, which can be of independent interest. This allows us to estimate the distribution of the competing bids uniformly over the subintervals of $[0, 1]$. In the adversarial setting, instead, we use the smoothness assumption to guarantee that the expected utility is Lipschitz. In this case, the approximation error is controlled using a uniform grid with an appropriate grid-size (Lemma 4).

The feedback models. Our feedback models interpolate between bandit (only the bidder's utility is observed) and full feedback (V_t and M_t are always observed). In the stochastic i.i.d. case, the different levels of transparency are crucial to the process of building the non-uniform grids used to control the discretization error. In the adversarial case, when there are only K allowed bids, the optimal rates are of order $\sqrt{T \ln K}$ and $\sqrt{KT}$ under full and bandit feedback, respectively. While the semi-transparent feedback is not enough to improve on the bandit rate, the transparent one can be exploited via a more sophisticated approach. To this end, we design an algorithm, Exp3.FPA, enjoying the full feedback regret rate of order $\sqrt{T \ln K}$ while only relying on the weaker transparent feedback.

Lower bounds. The linear lower bounds (Theorems 1 and 8) exploit a "needle in a haystack" phenomenon, where there is a hidden optimal bid $b^\star$ in the $[0, 1]$ interval and the learner has no way of finding $b^\star$ using the feedback it has access to. This is indeed the case in the non-smooth adversarial full-feedback setting and in the non-smooth i.i.d. bandit setting. To prove the remaining lower bounds, we design careful embeddings of known hard instances into our framework. In particular, in Theorem 5 we embed the hard instance for prediction with two experts and in Theorem 3 the hard instance for $K = \Theta(T^{1/3})$ bandits.

1.3 Related Work

Transparency in first-price auctions. The role of transparency in repeated first-price auctions has been investigated by Bergemann and Hörner [9], but mostly from a game-theoretic viewpoint. In particular, they study the impact of the feedback policy on the bidders' strategy and show how disclosing the bids at the end of each round affects the equilibria of a bidding game with infinite horizon. In contrast, we want to characterize the impact of different amounts of feedback (or degrees of transparency) on the learner's regret, which is measured against the optimal fixed bid in hindsight.

Auctions with unknown valuations. Although the problem of regret minimization in first-price auctions has been studied before, only a few papers consider the natural setting of unknown valuations. Feng et al. [23] introduce a general framework for the study of regret in auctions where a bidder's valuation is only observed when the auction is won. In the special case of first-price auctions, their setting is equivalent to our transparent feedback when the sequence of pairs (V_t, M_t) is adversarially generated. Following a parameterization introduced by Weed et al. [39], Feng et al. [23] provide a $O(\sqrt{T \ln \max\{\Delta_0^{-1}, T\}})$ regret bound, where $\Delta_0 = \min_{t < t'} |M_t - M_{t'}|$ is controlled by the environment. In the stochastic i.i.d. case, their results translate into *distribution-dependent* guarantees that do not translate into a worst-case sublinear bound (we obtain a $\sqrt{T}$ rate). In the adversarial case, their guarantees are still linear in the worst-case (we obtain $\sqrt{T}$ bounds by leveraging the smoothness assumption). Achddou et al. [1] consider a stochastic i.i.d. setting with the additional assumption that V_t and M_t are independent. Their main result is a bidding algorithm with *distribution-dependent* regret rates (of order $T^{1/3+\varepsilon}$ or $\sqrt{T}$, depending on the assumptions on the underlying distribution) in the transparent setting. Again, this result is not comparable to ours because of the independence assumption and the distribution-dependent rates (which do not allow to recover our minimax rates). Other works consider regret minimization in repeated second-price auctions with unknown valuations. Dikkala and Tardos [19] investigate a repeated bidding setting, but do not consider regret minimization. Weed et al. [39] derive regret bounds for the case when M_t are adversarially generated, while V_t are stochastically or adversarially generated and the feedback is transparent.

First-price auctions with known valuations. Considerably more works study first price auctions when the valuation V_t is known to the bidder at the beginning of each round t. Note that these results are not directly comparable to ours. Balseiro et al. [7] look at the case when the V_t are adversarial and the M_t are either stochastic i.i.d. or adversarial. In the bandit feedback case (when M_t is never observed), they show that the minimax regret is $\widetilde{\Theta}(T^{2/3})$ in the stochastic case and $\widetilde{\Theta}(T^{3/4})$ in the adversarial case. Han et al. [29] prove a $\widetilde{O}(\sqrt{T})$ regret bound in the semi-transparent setting (M_t observed only when the auction is lost) with adversarial valuations and stochastic bids. Han et al. [28] focus on the adversarial case, when V_t and M_t are both generated adversarially. They prove a $\widetilde{O}(\sqrt{T})$ regret bound in the full feedback setting (M_t always observed) when the regret is defined with respect to all Lipschitz shading policies. This setup is extended in Zhang et al. [42] where the authors consider the case in which the bidder is provided access to hints before each auction. Zhang et al. [41] also studied the full information feedback setting and design a space-efficient variant of the algorithm proposed by Han et al. [28]. Badanidiyuru et al. [5] introduce a contextual model in which V_t is adversarial and $M_t = \langle \theta, x_t \rangle + \varepsilon_t$ where $x_t \in \mathbb{R}^d$ is contextual information available at the beginning of each round t, $\theta \in \mathbb{R}^d$ is an unknown parameter, and ε_t is drawn from an unknown log-concave distribution. They study regret in bandit and full feedback settings.

Dynamics in first-price auctions. A different thread of research is concerned with the convergence property of the regret minimization dynamics in first-price auctions (or, more specifically, with the learning dynamics of mean-based regret minimization algorithms). Feldman et al. [22] show that with continuous bid levels, coarse-correlated equilibria exist whose revenue is below the second price. Feng et al. [24] prove that regret minimizing bidders converge to a Bayesian Nash equilibrium in a first-price auctions when bidder values are drawn i.i.d. from a uniform distribution on $[0, 1]$. Kolumbus and Nisan [32] show that if two bidders with finitely many bid values converge, then the equilibrium revenue of the bidder with the highest valuation is the second price. Deng et al. [18] characterize the equilibria of the learning dynamics depending on the number of bidders with the highest valuation. Their characterization is for both time-average and last-iterate convergence.

Smoothed adversary. Smoothed analysis of algorithms, originally introduced by Spielman and Teng [38] and later formalized for online learning by Rakhlin et al. [35] and Haghtalab et al. [25], is a known 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 Block et al. [12], Bolić et al. [13], Cesa-Bianchi et al. [14, 15, 17], Durvasula et al. [21], Haghtalab et al. [25, 27], Kannan et al. [30].

Online learning in metric spaces. Our problem is related to online learning in metric spaces [31], where the action space is endowed with a metric and the losses are induced by a sequence of Lipschitz functions defined onto it. Tight regret bounds are known, parameterized by some notion of dimension of the metric space, in both the full and the bandit models. The simple structure of our action space ($[0, 1]$ with the Euclidean distance) allows us to obtain tight bounds by either using a uniform grid (Theorems 6 and 7) or sample-based grids (Theorems 2 and 4), without resorting to the more elaborate techniques that characterize this line of research, e.g., zooming (which is typically used in the bandit feedback model to account for the lack of feedback). Also related to our model is the study of piecewise and regular Lipschitz functions [6, 20, 36]. In particular, Lemma 1 and Theorem 3 in Balcan et al. [6] imply our Theorem 6 in the special case of independent processes.[*]

2 THE LEARNING MODEL

We introduce formally the repeated bidding problem in first-price auctions. At each time step t, a new item arrives for sale, for which the learner holds some unknown valuation $V_t \in [0, 1]$. The learner bids some $B_t \in [0, 1]$ and, at the same time, a set of competitors bid for the same object. We denote their highest competing bid by $M_t \in [0, 1]$. The learner gets the item at cost B_t if it wins the auction (i.e., if $B_t \geq M_t$), and does not get it otherwise. Then, the learner observes some feedback Z_t and gains utility $\text{Util}_t(B_t)$, where, for all $b \in [0, 1]$, $\text{Util}_t(b) = (V_t - b)\mathbb{I}\{b \geq M_t\}$ (see the Protocol in Section 1). Crucially, at time t the learner does not

know its valuation V_t for the item before bidding, implying that its bid B_t only depends on its past observations $Z_1, \ldots, Z_{t-1}$ (and, possibly, some internal randomization). The goal of the learner is to design a learning algorithm $\mathcal{A}$ that maximizes its utility. More precisely, we measure the performance of an algorithm $\mathcal{A}$ by its *regret* $R_T(\mathcal{A})$ against the worst environment $\mathcal{S}$ in a certain class Ξ: $R_T(\mathcal{A}) = \sup_{\mathcal{S} \in \Xi} R_T(\mathcal{A}, \mathcal{S})$, where

$$R_T(\mathcal{A}, \mathcal{S}) = \sup_{b \in [0,1]} \mathbb{E}\left[\sum_{t=1}^{T} \text{Util}_t(b) - \sum_{t=1}^{T} \text{Util}_t(B_t) \right].$$

The expectation in the previous display is taken with respect to the randomness of the algorithm $\mathcal{A}$ which selects B_t, and (possibly) the randomness of the environment $\mathcal{S}$ generating the (V_t, M_t) pairs.

The environments. In this paper we consider both stochastic i.i.d. and adversarial environments.

- Stochastic i.i.d.: The pairs $(V_1, M_1), (V_2, M_2), \ldots$ are a stochastic i.i.d. process.
- Adversarial: The sequence $(V_1, M_1), (V_2, M_2), \ldots$ is generated by an oblivious adversary.

Following previous works in online learning (see Section 1.3), we also study versions of the above environments that are constrained to generate the sequence of (V_t, M_t) values using distributions that are "not too concentrated". To this end, we introduce the notion of smooth distributions.

Definition 1 (Haghtalab et al. [26]). *Let $\mathcal{X}$ be a domain that supports a uniform distribution ν. A measure μ on $\mathcal{X}$ is said to be σ-smooth if for all measurable subsets $A \subseteq \mathcal{X}$, we have $\mu(A) \leq \frac{\nu(A)}{\sigma}$.*

We thus also consider the following two types of environments.

- The σ-smooth stochastic i.i.d. environment, which is a stochastic i.i.d. environment where the common distribution of all pairs $(V_1, M_1), (V_2, M_2), \ldots$ is σ-smooth.
- The σ-smooth adversarial setting, where the pairs $(V_1, M_1), \ldots$ form a stochastic process such that, for each t, the distribution of the pair (V_t, M_t) is σ-smooth.

The feedback. After describing the environments that we study, we now specify the types of feedback the learner receives at the end of each round, from the richest to the least informative.

- Full feedback. The learner observes its valuation and the highest competing bid: $Z_t = (V_t, M_t)$.
- Transparent feedback. The learner always observes M_t, but V_t is only revealed if it gets the item: Z_t is equal to $(\star, M_t)$ if $B_t < M_t$ and to (V_t, M_t) otherwise.
- Semi-transparent feedback[†]. The learner observes V_t if it gets the item and M_t otherwise: Z_t is equal to $(\star, M_t)$ if $B_t < M_t$ and to $(V_t, \star)$ otherwise.
- The bandit feedback[‡]. The learner observes V_t if it gets the item and the symbol $\star$ otherwise: Z_t is $\star$ if $B_t < M_t$ and to V_t otherwise.

---

[*]Combining the second part of their Lemma 1 with their Theorem 3 to lift independence gives void guarantees in the general case (note that there is a typo in the statement of their Lemma 1: as it can be seen in the proof, the correct result is $k = P \cdot O(M \cdot \kappa \cdot w + \sqrt{M \log(P/\zeta)})$) and, without assuming independence, $P = T$ in our setting).

[†]This feedback is similar to the winner-only feedback in Han et al. [29].

[‡]We call this the bandit feedback because it is equivalent to receiving $\text{Util}_t(B_t)$ (with the extra information $\star$ to distinguish between losing the item and winning it with $V_t = B_t$, which does not affect regret guarantees).

3 THE STOCHASTIC I.I.D. SETTING

In this section, we investigate the problem of repeated bidding in first-price auctions with unknown valuations, when the pairs of valuations and highest competing bids are drawn i.i.d. from a fixed but unknown distribution. We start by proving in Section 3.1 that it is impossible to achieve sublinear regret under the bandit feedback model without any assumption on the distribution of the environment. Then, in Section 3.2, we give matching upper and lower bounds of order $T^{2/3}$ in the semi-transparent feedback model. Notably, the lower bound holds for smooth distributions, while the upper bound works for any (possibly non-smooth) distributions. Finally, in Section 3.3 we prove that both the full and transparent feedback yield the same minimax regret regime of order $\sqrt{T}$, regardless of the regularity of the distribution.

3.1 I.I.D. – Bandit Feedback

In the bandit feedback model, at each time step, the learner observes the valuation V_t (and nothing else) when the auction is won, and observes nothing when the auction is lost. The crucial difference with the other (richer) types of feedback is the amount of information received about M_t, which, in the bandit case, is just the relative position with respect to B_t (i.e., whether $M_t \leq B_t$ or $B_t < M_t$). This allows to hide in the interval $[0, 1]$ an optimal bid $b^\star$ which the learner cannot uncover over a finite time horizon. Following this idea, a difficult environment should randomize between two scenarios: a good scenario with large value $V_t = 1$ and M_t slightly smaller than $b^\star$ and a bad one with poor value $V_t = 0$ and M_t slightly larger than $b^\star$. Then, to avoid suffering linear regret, the learner has to find this tiny interval around $b^\star$ (the "needle in a haystack").

Theorem 1. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. environment with bandit feedback. Then, any learning algorithm $\mathcal{A}$ satisfies $R_T(\mathcal{A}) \geq \frac{1}{13}T$.*

Proof. We construct a randomized i.i.d. environment $\mathcal{S}$, such that any deterministic algorithm $\mathcal{A}$ suffers linear regret against it, and then apply Yao's minimax principle to conclude the proof. The randomized environment is simple: before starting the sequence, a uniform seed $b^\star$ is drawn uniformly at random in $(1/3, 1/2 - \varepsilon)$, where ε is a small parameter we set later. Then, the i.i.d. sequence $(V_1, M_1), (V_2, M_2), \ldots$ is drawn as follows: at each time step t with probability $1/2$ we have $(V_t, M_t) = (1, b^\star)$, otherwise $(0, b^\star + \varepsilon)$. The best bid in hindsight, $b^\star$, yields an overall expected utility of $\frac{T}{2}(1 - b^\star)$, which is at least $T/4$, as $b^\star$ belongs to $(1/3, 1/2)$.

We now upper bound the utility achievable by any deterministic algorithm $\mathcal{A}$ against $\mathcal{S}$. Fix any such algorithm, and consider its bids against any environment that selects the valuations V_t to be either 0 or 1 (as the one we just constructed). At each time step, the feedback that $\mathcal{A}$ receives is 0, 1 or $\star$ (when the item is allocated to one of the competitors), so that the history of the bids posted by $\mathcal{A}$ is naturally described by a ternary decision tree of height T, where each level corresponds to a time step and any node to a bid. Crucially, the leaves of this tree are finite (at most 3^T), which means that the algorithm $\mathcal{A}$ only posts bids in a finite subset N of $[0, 1]$. Now, let $\varepsilon = 3^{-2T}/12$; we have that, with probability at least $1 - 6N\varepsilon/(1-6\varepsilon) \geq 1 - e^{-T}$, the set $[b^\star, b^\star + \varepsilon]$ does not intersect N.

1: **input:** Time horizon T_0
2: $X_0 \leftarrow 0$ and $M^{(0)} \leftarrow 0$
3: **for** each round $t = 1, 2, \ldots, T_0$ **do**
4:      Post bid $B_t = 0$ and observe the highest competing bid M_t
5: Sort the observed highest competing bids in increasing order: $M^{(1)} \leq M^{(2)} \leq \cdots \leq M^{(T_0)}$
6: **if** $M^{(T_0)} = 0$ **then return** candidate bid X_0
7: **for** $i = 1, 2, \ldots$ **do**
8:      $j_{i-1}^\star \leftarrow \max\{j \in \{0, \ldots, T_0\} \mid X_{i-1} = M^{(j)}\}$
9:      $j_i \leftarrow \min\{j_{i-1}^\star + \lceil\sqrt{T_0}\rceil, T_0\}$, $X_i \leftarrow M^{(j_i)}$
10:      **if** $j_i = T_0$ **then** let $K \leftarrow i$ and **break**;
11: **return** Candidate bids $X_0, X_1, X_2, \ldots, X_K$

Note: the randomness is with respect to the uniform seed $b^\star$ drawn by $\mathcal{S}$, while the bound on the probability holds independently to the choice of the deterministic algorithm $\mathcal{A}$.

The total utility of $\mathcal{A}$ when $[b^\star, b^\star + \varepsilon]$ does not intersect N is easy to analyze: every time that $\mathcal{A}$ posts bids smaller than $b^\star$, then it never wins the item (zero utility). Instead, if it posts bids larger than $b^\star + \varepsilon$, then it always gets the item (whose average value is $1/2$), paying at least $b^\star + \varepsilon \geq 1/3$. Putting these two cases together, we have proved that at each time step the expected utility earned by the learner is at most $1/6 = 1/2 - 1/3$, when $[b^\star, b^\star + \varepsilon] \cap N = \varnothing$ (which happens with probability at least $1 - e^{-T}$). Finally, by combining the lower bound on the performance of $b^\star$ with the upper bound on the expected utility of the learner, we get $R_T(\mathcal{A}, \mathcal{S}) \geq (1 - e^{-T})(T/4 - T/6) \geq T/13$. □

3.2 I.I.D. – Semi-Transparent Feedback

In this section, we prove two results settling the minimax regret for the semi-transparent feedback where the environment is i.i.d. (and, possibly, smooth). First, we construct a learning algorithm, COLLECTING BANDIT, achieving $T^{2/3}$ regret against any i.i.d. environment. Then, we complement it with a lower bound of the same order (up to log terms) obtained even in a smooth i.i.d. environment.

3.2.1 A $T^{2/3}$ Upper Bound for the i.i.d. Environment. Our learning algorithm COLLECTING BANDIT is composed of two phases. First, for $T_0 = \Theta(T^{2/3})$ rounds, it collects samples from the highest competing bid random variables $M_1, M_2, \ldots, M_{T_0}$ by posting dummy bids $B_1 = B_2 = \cdots = B_{T_0} = 0$. Among these values (plus the value $X_0 = 0$), the algorithm selects $\Theta(\sqrt{T_0})$ candidate bids according to their ordering, in such a way that the empirical frequencies of bids $M_1, M_2, \ldots, M_{T_0}$ landing strictly in between two consecutive selected values are at most $\Theta(1/\sqrt{T_0})$ (see the pseudocode of COLLECT BIDS for details). Second, for the remaining time steps, it runs any bandit algorithm, using as candidate bids the ones collected in the first phase (see COLLECTING BANDIT for details). Note that, in this second phase, the (less informative) bandit feedback would be enough to run the algorithm: the additional information provided by the semi-transparent feedback is only exploited in the initial "collecting bids" phase.

As a first step, we state a simple concentration result pertaining the i.i.d. process $M, M_1, M_2, \ldots, M_{T_0}$, for $T_0 \in \mathbb{N}$. If $\mathcal{I}$ is the family

of all the subintervals of $[0, 1]$ and $\delta \in (0, 1)$, we define

$$\mathcal{E}_\delta^{T_0} = \bigcap_{I \in \mathcal{I}} \left\{ \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{I}\{M_t \in I\} - \mathbb{P}[M \in I] \right| < 8\sqrt{\frac{\ln(1/\delta)}{T_0}} \right\} .$$

The family $\mathcal{I}$ of all the subintervals of $[0, 1]$ has VC dimension 2 (see, e.g., Mitzenmacher and Upfal [34, Chapter 14.2]). Therefore, $\mathcal{E}_\delta^{T_0}$ is realized with probability at least $1 - \delta$, via standard sample complexity bound for ε-samples (see, e.g., Mitzenmacher and Upfal [34, Theorem 14.15]). This is summarized in the following lemma.

Lemma 1. *For every $T_0 \in \mathbb{N}$ and $\delta \in (0, 1)$, we have $\mathbb{P}[\mathcal{E}_\delta^{T_0}] \geq 1 - \delta$.*

For the sake of readability, we introduce the following notation:

Notation 1. *Let $X = \{x_0, \ldots, x_K\}$ be any finite grid with $0 = x_0 < x_1 < \cdots < x_K \leq 1$, we denote by $k_X \colon [0, 1] \to \{0, 1, \ldots, K\}$ the function that maps each $b \in [0, 1]$ to the unique k such that $b \in [x_k, x_{k+1})$, with the convention that $x_{K+1} = 2$.*

We now prove a lemma that allows us to control the expected cumulative utility of any bid in $[0, 1]$ with that of the best bid in a discretization (without relying on any smoothness assumption).

Lemma 2. *Consider any finite grid $X = \{x_0, \ldots, x_K\}$, with $0 = x_0 < x_1 < \cdots < x_K \leq 1$, and assume that the process $M, M_1, M_2, \ldots$ of the highest competing bids form an i.i.d. sequence. For all $b \in [0, 1]$ and $T_0, T_1 \in \mathbb{N}$ with $T_0 < T_1$, $\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(b)\right]$ is at most*

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t\left(x_{k_X(b)}\right)\right] + (T_1 - T_0)\mathbb{P}\left[x_{k_X(b)} < M < x_{k_X(b)+1}\right] .$$

Proof. Fix any $b \in [0, 1]$, $T_0, T_1 \in \mathbb{N}$ with $T_0 < T_1$, and a time step $t \in \{T_0 + 1, \ldots, T_1\}$. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Util}_t(b)\right] &= \mathbb{E}\left[(V_t - b)\mathbb{I}\{b \geq M_t\}\right] \\
&\leq \mathbb{E}\left[(V_t - x_{k_X(b)})\left(\mathbb{I}\{x_{k_X(b)} \geq M_t\} + \mathbb{I}\{b \geq M_t > x_{k_X(b)}\}\right)\right] \\
&\leq \mathbb{E}\left[\mathrm{Util}_t(x_{k_X(b)})\right] + \mathbb{P}\left[x_{k_X(b)} < M_t \leq b\right] \\
&\leq \mathbb{E}\left[\mathrm{Util}_t(x_{k_X(b)})\right] + \mathbb{P}\left[x_{k_X(b)} < M_t < x_{k_X(b)+1}\right] .
\end{aligned}$$

Summing over all times t and recalling that M_t and M share the same distribution, yields the conclusion. □

As a corollary of Lemmas 1 and 2 we obtain a similar discretization error guarantee when the grid of points X is random.

Lemma 3. *Fix any $T_0 \in \mathbb{N}$ and $\delta \in (0, 1)$. Let $X = \{X_0, \ldots, X_K\}$ be a random set containing a random number K of points satisfying $0 = X_0 < X_1 < \cdots < X_K \leq 1$. Assume that the random variables $K, X_0, X_1, \ldots, X_{K+1}$ are $\mathcal{H}_{T_0}$-measurable, where $\mathcal{H}_{T_0}$ is the history up to and including time T_0. Assume that the process $(V_1, M_1), (V_2, M_2), \ldots$ of the valuations/highest competing bids form an i.i.d. sequence. Then, for all $b \in [0, 1]$ and $T_1 \in \mathbb{N}$ with $T_1 > T_0$, we have:*

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(b)\right] \leq \mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t\left(X_{k_X(b)}\right)\right]$$

$$+ (T_1 - T_0)\left(\frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{P}\left[X_{k_X(b)} < M_t < X_{k_X(b)+1}\right] + 8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right).$$

Collecting Bandit (CoBa)

1: **input:** Time horizon T, bandit algorithm $\widetilde{\mathcal{A}}$ for gains in $[-1, 1]$
2: $T_0 \leftarrow \lceil T^{2/3} \rceil$
3: Run Collect Bids with horizon T_0 and obtain $X_0, X_1, \ldots, X_K$
4: Initialize $\widetilde{\mathcal{A}}$ on $K + 1$ actions (one for each candidate bid X_i) and $T - T_0$ as time horizon
5: **for** each round $t = T_0 + 1, T_0 + 2, \ldots, T$ **do**
6:      Receive from $\widetilde{\mathcal{A}}$ the bid $B_t = X_{I_t}$ for some $I_t \in \{0, 1, \ldots, K\}$
7:      Post bid B_t and observe feedback Z_t
8:      Reconstruct $\mathrm{Util}_t(B_t)$ from Z_t and feed it to $\widetilde{\mathcal{A}}$

We are now ready to present the main theorem of this section.

Theorem 2. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. environment with semi-transparent feedback. Then there exists a learning algorithm $\mathcal{A}$ such that*

$$R_T(\mathcal{A}) \leq 16\left(13 + \sqrt{\ln T}\right)T^{2/3} .$$

Proof. We prove that Collecting Bandit yields the desired bound when its learning routine $\widetilde{\mathcal{A}}$ is (a rescaled version of) MOSS [3]: since MOSS is designed to run with gains in $[0, 1]$ while the utilities we observe are in $[-1, 1]$, we first apply the reward transformation $x \mapsto \frac{x+1}{2}$ to the observed utilities. This costs a multiplicative factor of 2 on the regret guarantees of MOSS. Leveraging the fact that the empirical frequency between two consecutive X_k and X_{k+1} generated by Collect Bids is at most $2/\sqrt{T_0}$ by design and applying Lemma 3 with $T_1 = T$ to the random variables $X_0, X_1, \ldots, X_K$, we get, for all $b \in [0, 1]$, that $\mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t(b)\right]$ is upper bounded by

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t\left(X_{k_X(b)}\right)\right] + (T - T_0)\left(\frac{2}{\sqrt{T_0}} + 8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) = (\star) .$$

Now, applying the tower rule to the expectation conditioning to the history $\mathcal{H}_{T_0}$ up to time T_0, we can use the fact that the regret of the rescaled version of MOSS is upper bounded by $98\sqrt{(K + 1)(T - T_0)}$ and the number of points $K + 1$ collected by Collect Bids is at most $\sqrt{T_0} + 1$ to obtain

$$(\star) \leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} \mathrm{Util}_t(B_t)\right] + 98\sqrt{(\sqrt{T_0} + 1)(T - T_0)}$$

$$+ (T - T_0)\left(\frac{2}{\sqrt{T_0}} + 8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right) .$$

Finally, tuning $\delta = 1/T_0$, upper bounding the cumulative regret over the first T_0 rounds with T_0, and recalling that $T_0 = \lceil T^{2/3} \rceil$, yields the conclusion. □

3.2.2 A $T^{2/3}$ Lower Bound for the Smooth i.i.d. Environment. We prove that the $\widetilde{O}(T^{2/3})$ bound achieved by Collecting Bandit is indeed optimal, up to logarithmic terms. Our lower bound consists in carefully embedding into our model a hard multiarmed bandit instance with $K = \Theta(T^{1/3})$ arms, which entails a lower bound of order $\Omega(\sqrt{KT}) = \Omega(T^{2/3})$. This proof agenda involves various challenges: we want to embed a discrete construction of K independent
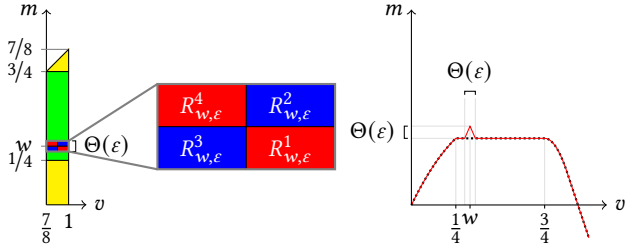
Figure 2: Left: The support of the base density f lies inside the yellow and green regions. The perturbation $g_{w,\varepsilon}$ of f occurs inside the green region, where the four rectangles $R^1_{w,\varepsilon}, \ldots, R^4_{w,\varepsilon}$ (in red and blue) lie. Right: The corresponding qualitative plots of $b \mapsto \mathbb{E}[\mathrm{Util}_t(b)]$ (black, dotted) and $p \mapsto \mathbb{E}^{w,\varepsilon}[\mathrm{Util}_t(b)]$ (red, solid).

actions into our continuous framework, where the utilities of different bids are correlated, while enforcing smoothness. Furthermore, the semi-transparent feedback is richer than the bandit one. We report here a proof sketch and refer the interested reader to the full version of the paper [16] for the missing details.

THEOREM 3. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. σ-smooth environment with semi-transparent feedback, for $\sigma \in (0, 1/66]$. Then, any learning algorithm $\mathcal{A}$ satisfies, for $T \geq 8$, $R_T(\mathcal{A}) \geq \frac{3}{10^4} T^{2/3}$.*

PROOF SKETCH. Define, for all $v, m \in [0, 1]$, the density

$$f(v,m) = \mathbb{I}_{\left[\frac{7}{8},1\right]}(v)\left(\frac{1}{(v-m)^2}\mathbb{I}_{\left[\frac{1}{4}, v-\frac{1}{8}\right]}(m) + \frac{4}{v-1/4}\mathbb{I}_{\left[0,\frac{1}{4}\right)}(m)\right) \text{\S}.$$

Let $\mathbb{P}^0$ be a probability measure such that $(V, M), (V_1, M_1), \ldots$ is a $\mathbb{P}$-i.i.d. sequence where each pair (V, M) has common probability density function f. Denoting by $\mathbb{E}^0$ the expectation with respect to $\mathbb{P}^0$, we have, for any bid $b \in [0, 1]$ and any time step t

$$\mathbb{E}^0\left[\mathrm{Util}_t(b)\right] = b\left(\frac{1}{2} + (1-4b)\ln\frac{6}{5}\right)\mathbb{I}_{\left[0,\frac{1}{4}\right)}(b) + \frac{1}{8}\mathbb{I}_{\left[\frac{1}{4},\frac{3}{4}\right]}(b)$$
$$- \left(4b^2 - 6b + \frac{17}{8}\right)\mathbb{I}_{\left[\frac{3}{4},\frac{7}{8}\right]}(b) + \left(\frac{15}{16} - b\right)\mathbb{I}_{\left[\frac{7}{8},1\right]}(b).$$

This function grows with b on $[0, 1/4)$, has a plateau of maximizers $[1/4, 3/4]$, then decreases on $(3/4, 1]$ (see Figure 2, right). We introduce the perturbation space Ξ:

$$\Xi = \left\{(w, \varepsilon) \in [0,1]^2 : w - \varepsilon \geq \tfrac{1}{4} \text{ and } w + \varepsilon \leq \tfrac{3}{4}\right\}$$

and define, for all $(w, \varepsilon) \in \Xi$, the four rectangles

$$R^1_{w,\varepsilon} = [15/16, 1] \times [w - \varepsilon, w), \quad R^2_{w,\varepsilon} = [15/16, 1] \times [w, w + \varepsilon),$$

$$R^3_{w,\varepsilon} = [7/8, 15/16) \times [w - \varepsilon, w), \quad R^4_{w,\varepsilon} = [7/8, 15/16) \times [w, w + \varepsilon).$$

For all $(w, \varepsilon) \in \Xi$, we introduce the probability density function $f_{w,\varepsilon}$ as follows $f_{w,\varepsilon} = f + g_{w,\varepsilon}$, where the perturbation $g_{w,\varepsilon}$ is defined as follows

$$g_{w,\varepsilon}(v,m) = \frac{16}{9}\left(\mathbb{I}_{R^1_{w,\varepsilon} \cup R^4_{w,\varepsilon}}(v,m) - \mathbb{I}_{R^2_{w,\varepsilon} \cup R^3_{w,\varepsilon}}(v,m)\right).$$

§Note, we use the notation $\mathbb{I}_A(x)$ to denote the indicator function that has value 1 when $x \in A$, and 0 otherwise.

We refer to the left plot in Figure 2 for a visualization of the support of the $f_{w,\varepsilon}$. For all $(w, \varepsilon) \in \Xi$, let $\mathbb{P}^{w,\varepsilon}$ be a probability measure such that $(V, M), (V_1, M_1), (V_2, M_2), \ldots$ is a $\mathbb{P}^{w,\varepsilon}$-i.i.d. sequence where each pair (V, M) has common probability density function $f_{w,\varepsilon}$. Denoting by $\mathbb{E}^{w,\varepsilon}$ the expectation with respect to $\mathbb{P}^{w,\varepsilon}$, we have, for any bid $b \in [0, 1]$ and any t

$$\mathbb{E}^{w,\varepsilon}\left[\mathrm{Util}_t(b)\right] = \mathbb{E}^0\left[\mathrm{Util}_t(b)\right] + \frac{\varepsilon}{144}\Lambda_{w,\varepsilon}(b)$$

where $\Lambda_{u,r}$ is the tent map centered at u with radius r defined as $\Lambda_{u,r}(x) = \max\{1 - |x - u|/r, 0\}$. In words, in a perturbed scenario $\mathbb{P}^{w,\varepsilon}$ the expected utility is maximized at the peak of a spike centered at w with length and height $\Theta(\varepsilon)$ perturbing the plateau area $[1/4, 3/4]$ of maximum height (see Figure 2, right). Define, for all times $t \in \mathbb{N}$, the feedback function $\psi_t \colon [0, 1] \to \left([0, 1] \times \{\star\}\right) \cup \left(\{\star\} \times [0, 1]\right)$, as follows:

$$b \mapsto \begin{cases} (V_t, \star) & \text{if } b \geq M_t \\ (\star, M_t) & \text{if } b < M_t \end{cases}$$

and note that, in our semi-transparent feedback model, the feedback Z_t received after bidding B_t at time t is $\psi_t(B_t)$. Crucially, for each $(w, \varepsilon) \in \Xi$ and each $b \in [0, 1] \setminus [w - \varepsilon, w + \varepsilon]$, the distribution of $\psi_t(b)$ under $\mathbb{P}^{w,\varepsilon}$ coincides with the distribution of $\psi_t(b)$ under $\mathbb{P}^0$. In push-forward notation, it holds that

$$\mathbb{P}^{w,\varepsilon}_{\psi_t(b)} = \mathbb{P}^0_{\psi_t(b)}. \tag{1}$$

Now, let $K \in \mathbb{N}$, $\varepsilon = 1/(4K)$, $w_k = 1/4 + (2k-1)\varepsilon$ and $\mathbb{P}^k = \mathbb{P}^{w_k,\varepsilon}$ (for each $k \in [K]$). At a high level, we built a problem with two crucial properties: (i) we know in advance the region where the optimal bid belongs to (i.e., the interval $[1/4, 3/4]$), but (ii) when the underlying scenario is determined by the probability measure $\mathbb{P}^k$, the learner has to detect inside this potentially optimal region where a spike of height (and length) $\Theta(\varepsilon)$ occurs (to avoid suffering suffer $\Omega(\varepsilon T)$ regret). 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 interval $[w_k - \varepsilon, w_k + \varepsilon]$ if the underlying probability is $\mathbb{P}^k$, suffering instantaneous regret of order ε whenever the underlying probability is $\mathbb{P}^j$, with $j \neq k$. Given that we partitioned the potentially optimal region $[1/4, 3/4]$ into $\Theta(1/\varepsilon)$ disjoint intervals where these perturbations can occur, the feedback structure implies that each of these intervals deserves its dedicated exploration.

To better highlight this underlying structure, in the full version, we show that our problem is not easier than a simplified K-armed stochastic bandit problem, where the instances we consider are determined by the probability measures $\mathbb{P}^1, \ldots, \mathbb{P}^K$. In this bandit problem, when the underlying probability measure is induced by some $\mathbb{P}^k$, the corresponding arm k has an expected reward $\Theta(\varepsilon)$ larger than the others. Then, via an information-theoretic argument, we can show that any learner would need to spend at least order of $1/\varepsilon^2$ rounds to explore each of the K arms (paying $\Omega(\varepsilon)$ each time) or else, it would pay a regret $\Omega(\varepsilon T)$. Hence, the regret of any learner, in the worst case, is lower bounded by $\Omega\left(\frac{K}{\varepsilon^2}\varepsilon + \varepsilon T\right) = \Omega\left(K^2 + \frac{T}{K}\right)$ (recalling our choice of $\varepsilon = 1/(4K)$). Picking $K = \Theta(T^{1/3})$ yields a lower bound of order $T^{2/3}$. □

3.3 I.I.D. – Transparent/Full Feedback

This section completes the study of the stochastic i.i.d. environment by determining the minimax regret when the learner has access to full or transparent feedback.

3.3.1 A $\sqrt{T}$ Upper Bound for the i.i.d. Environment. While with semi-transparent feedback, the learning algorithm has to rely on dummy bids $B_1 = \cdots = B_{T_0} = 0$ to gather information about the distribution of the highest competing bids, with the transparent one, this information is collected for free at each bidding round. To use this extra information, we present a wrapper W.T.FPA (for a sequence of base learning algorithms for the transparent feedback model) whose purpose is restarting the learning process with a geometric step to update the set of candidate bids. We assume that each of the wrapped base algorithms $\widetilde{\mathcal{A}}_\tau$ can take as input any finite subset $\mathcal{X} \subset [0, 1]$ and returns bids in $\mathcal{X}$. Furthermore, for all T', we let $\mathcal{R}_{T'}(\widetilde{\mathcal{A}}_\tau, \mathcal{X})$ be an upper bound on the regret over T' rounds of $\widetilde{\mathcal{A}}_\tau$ with input $\mathcal{X}$ against the best fixed $x \in \mathcal{X}$. Formally, we require that for any two times $T_0 < T_1$ such that $T' = T_1 - T_0$, the quantity $\mathcal{R}_{T'}(\widetilde{\mathcal{A}}_\tau, \mathcal{X})$ is an upper upper bound on $\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(x) - \sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(B_t)\right]$, where $B_t \in \mathcal{X}$ is the sequence of prices played by $\widetilde{\mathcal{A}}_\tau$ (with input $\mathcal{X}$) when started at round $t = T_0 + 1$ and ran up to time T_1. Without loss of generality, we assume that $T' \mapsto \mathcal{R}_{T'}(\widetilde{\mathcal{A}}_\tau, \mathcal{X})$ is non-decreasing.

W.T.FPA (Wrapper for Transparent First-Price Auctions)

1: **input:** Base algorithms $\widetilde{\mathcal{A}}_1, \widetilde{\mathcal{A}}_2, \ldots$
2: **initialization:** $s \leftarrow 0$
3: **for** each epoch $\tau = 1, 2, \ldots$ **do**
4:     $\mathcal{X}_\tau \leftarrow \{0\} \cup \{M_1, \ldots, M_s\}$ (with $\mathcal{X}_1 = \{0\}$)
5:     Start $\widetilde{\mathcal{A}}_\tau$ with input $\mathcal{X}_\tau$ and run it for $t = s+1, \ldots, s + 2^{\tau-1}$
6:     Update $s \leftarrow s + 2^{\tau-1}$

Proposition 1. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. environment with transparent feedback. Then the regret of W.T.FPA run with base algorithms $\widetilde{\mathcal{A}}_1, \widetilde{\mathcal{A}}_2, \ldots$ satisfies*

$$R_T(\text{W.T.FPA}) \leq \sum_{\tau=2}^{\lceil \log_2(T+1)\rceil} \mathcal{R}_{2^{\tau-1}}(\widetilde{\mathcal{A}}_\tau, \mathcal{X}_\tau) + 3 + 16(\sqrt{2} + 2)\sqrt{T \ln T} .$$

PROOF. Fix an arbitrary epoch $\tau \in \{2, \ldots, \lceil \log_2(T+1)\rceil\}$; we want to bound the regret suffered there by W.T.FPA using Lemma 3. Using the notation of the lemma, let $\mathcal{X} = \mathcal{X}_\tau$, $K + 1 = |\mathcal{X}|$, $T_0 = \sum_{\tau'=1}^{\tau-1} 2^{\tau'-1} = 2^{\tau-1} - 1$ (the time passed from the beginning of epoch 1 up to and including the end of epoch $\tau-1$), $T_1 = \min\{T_0 + 2^{\tau-1}, T\}$ (the end of epoch τ), and let $X_0 < X_1 < \cdots < X_K$ be the distinct elements of $\mathcal{X}$ in increasing order, where we note that $X_0 = 0$, $X_K \leq 1$, and we set $X_{K+1} = 2$. Let $\mathcal{H}_{T_0}$ be the history, including time T_0.

Applying Lemma 3 (together with the fact that the empirical frequency between any two consecutive values X_k and X_{k+1} is 0 by design), and exploiting the monotonicity of $T' \mapsto \mathcal{R}_{T'}(\widetilde{\mathcal{A}}_\tau, \mathcal{X}_\tau)$ for

the last epoch (if $T_0 + 2^{\tau-1} > T$), we obtain, for all $b \in [0, 1]$ and $\delta \in (0, 1)$,

$$\mathbb{E}\left[\sum_{t=T_0+1}^{T_1} \mathrm{Util}_t(b)\right] \leq \sum_{t=T_0+1}^{T_1} \mathbb{E}\left[\mathrm{Util}_t\left(X_{k_{\mathcal{X}}(b)}\right)\right] + 2^{\tau-1}\left(8\sqrt{\frac{\ln(1/\delta)}{T_0}} + \delta\right)$$

$$\leq \sum_{t=T_0+1}^{T_1} \mathbb{E}\left[\mathrm{Util}_t(B_t)\right] + \mathcal{R}_{2^{\tau-1}}(\widetilde{\mathcal{A}}_\tau, \mathcal{X}_\tau) + 2^{\tau-1}\left(8\sqrt{\frac{\ln(1/\delta)}{2^{\tau-1}-1}} + \delta\right).$$

Summing over epochs $\tau \in \{2, \ldots, \lceil \log_2(T+1)\rceil\}$, upper bounding by 1 the regret incurred in the first epoch, and tuning $\delta = 1/T$ yields the conclusion. □

Now we are only left to design appropriate base algorithms $\widetilde{\mathcal{A}}_1, \widetilde{\mathcal{A}}_2, \ldots$ for the transparent feedback to wrap W.T.FPA around.

The Exp3.FPA algorithm. To this end, we introduce the Exp3.FPA algorithm (designed to run with transparent feedback), which borrows ideas from online learning with feedback graphs [2]. Similar algorithms for related settings have been previously proposed by Weed et al. [39] and Feng et al. [23]. For the familiar reader, note that our setting can be seen as an instance of online learning with strongly observable feedback graphs. In contrast to a black-box application of feedback-graph results, we shave off a logarithmic term (in the time horizon) by using a dedicated analysis. For any $x \in [0, 1]$, we denote by δ_x the Dirac distribution centered at x.

Exp3.FPA

1: **input:** Finite $\mathcal{X} \subset [0, 1]$, and exploration rate $\gamma \in (0, 1)$
2: Let $\bar{x} \leftarrow \max\{\mathcal{X}\}$ and, for all $x \in \mathcal{X}$, let $w_1(x) \leftarrow 1$
3: **for** each round $t = 1, 2, \ldots$ **do**
4:     Post bid $B_t \sim p_t \leftarrow (1-\gamma)\frac{w_t}{\|w_t\|_1} + \gamma\delta_{\bar{x}}$
5:     For all $x \in \mathcal{X}$, define the reward estimate:
$$\widehat{g}_t(x) \leftarrow (V_t - x)\mathbb{I}\{x \geq M_t\}\frac{\mathbb{I}\{M_t \leq B_t\}}{\sum_{y \geq M_t} p_t(y)}$$
6:     For all $x \in \mathcal{X}$, update the weight:
$$w_{t+1}(x) \leftarrow w_t(x)\exp\left(\gamma\widehat{g}_t(x)\right)$$

Note that the transparent feedback is sufficient to compute the reward estimates in Line 5. We defer the proof of the following proposition to the full version.

Proposition 2. *Let $\mathcal{X} \subset [0, 1]$ be a finite set, $T \in \mathbb{N}$ a time horizon, and tune the exploration rate as $\gamma = \sqrt{\ln(|\mathcal{X}|)/(e-1)T}$. Then, the regret of Exp3.FPA against the best fixed bid in $\mathcal{X}$ is*

$$\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{Util}_t(x) - \sum_{t=1}^{T} \mathrm{Util}_t(B_t)\right] \leq 2\sqrt{(e-1)\ln(|\mathcal{X}|)T}$$

Putting together Propositions 1 and 2 yields the desired rate.

THEOREM 4. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. environment with transparent feedback. Then there exists a learning algorithm $\mathcal{A}$ such that*

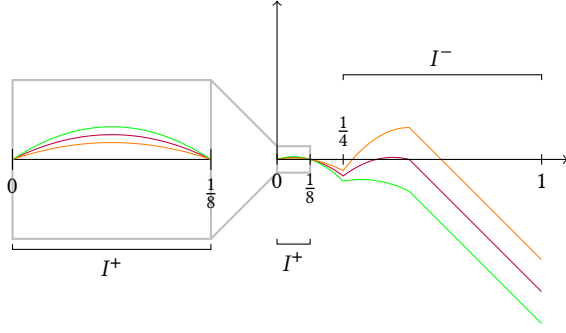$$R_T(\mathcal{A}) \leq 3 + 2(\sqrt{2} + 2)(\sqrt{2(e-1)} + 8)\sqrt{T \ln T} .$$

Figure 3: The expected utility function for three different distributions: $\mathbb{P}^0$ **in purple,** $\mathbb{P}^+$ **in orange, and** $\mathbb{P}^+$ **in green.**

PROOF. The statement of the theorem holds for W.T.FPA run with the base algorithm of each epoch τ being $Exp3.FPA$ tuned with $\gamma = \gamma(\tau) = \sqrt{\ln(|\mathcal{X}_\tau|)/\big((e-1)2^{\tau-1}\big)}$. Substituting the guarantees of Proposition 2 into those of Proposition 1 and recalling that $|\mathcal{X}_\tau| \leq 2^{\tau-1}$ for each epoch $\tau = 2, 3, \ldots$, yields the desired bound. □

3.3.2 A $\sqrt{T}$ Lower Bound for the i.i.d. Environment. We complement the positive result of Theorem 4 with a matching lower bound of order $\sqrt{T}$. The idea underlying our hard instance is to embed the well-known lower bound for prediction with (two) experts into our framework: we construct two smooth distributions that are "similar" but have two different optimal bids whose performance is separated so that no learner can identify the correct distribution without suffering less than $\sqrt{T}$ regret.

THEOREM 5. *Consider the problem of repeated bidding in first-price auctions in a stochastic i.i.d. σ-smooth environment with full feedback, for $\sigma \in (0, 1/9)$. Then, any learning algorithm $\mathcal{A}$ satisfies* $R_T(\mathcal{A}) \geq \frac{1}{2048}\sqrt{T}$.

PROOF. We prove the theorem by Yao's principle: we show that there exists a distribution over stochastic σ-smooth environments such that any deterministic learning algorithm $\mathcal{A}$ suffers $\Omega(\sqrt{T})$ regret against it, in expectation. We do that in two steps. First, for every $\varepsilon \in (0, 1/2)$ we construct a pair of $1/9$-smooth distributions that are hard to discriminate for the learner. Then, we prove that, for the right choice of ε, any learner suffers the desired regret against a uniform mixture of them. For visualization, we refer to Figure 3.

As a tool for our construction, we introduce a baseline probability measure $\mathbb{P}^0$, such that the sequence $(V, M), (V_1, M_1), (V_2, M_2), \ldots$ is $\mathbb{P}^0$-i.i.d., and (V, M) has distribution $\mathbb{P}^0_{(V,M)}$ whose density is

$$f^0(v, m) = 8\,(\mathbb{I}_{Q_+}(v, m) + \mathbb{I}_{Q_-}(v, m)),$$

where $Q_+ = (0, 1/4) \times (0, 1/4)$ and $Q_- = (3/4, 1) \times (1/4, 1/2)$. A convenient way to visualize this distribution is to draw a uniform random variable U_t in the square Q_+ and then toss an unbiased coin. If the coin yields heads, then (V_t, M_t) is equal to U_t, otherwise (V_t, M_t) coincides with U_t translated by $(3/4, 1/4)$. With some simple computation, it is possible to explicitly compute the expected utility of posting any bid $b \in [0, 1]$ when (V_t, M_t) is drawn following the

distribution $\mathbb{P}^0$ (and expectation $\mathbb{E}^0$):

$$\mathbb{E}^0[\text{Util}_t(b)] = \begin{cases} \frac{b}{4}(1 - 8b) & \text{if } b \in [0, 1/4) \\ -\frac{1}{8}(16b^2 - 14b + 3) & \text{if } b \in [1/4, 1/2) \\ \frac{1}{2}(1 - 2b) & \text{if } b \in [1/2, 1] \end{cases}$$

The function $\mathbb{E}^0[\text{Util}_t(b)]$ has two global maxima in $[0, 1]$, of value $1/128$, attained in $1/16$ and $7/16$ (see purple line in Figure 3).

For any $\varepsilon \in (0, 1/2)$, we also define two additional (perturbed) probability measures $\mathbb{P}^{\pm\varepsilon}$, such that the sequence $(V, M), (V_1, M_1), \ldots$ is $\mathbb{P}^{\pm\varepsilon}$-i.i.d. and the distribution $\mathbb{P}^{\pm\varepsilon}_{(V,M)}$ of (V, M) has density:

$$f^{\pm\varepsilon}(v, m) = 8(1 \pm \varepsilon)\mathbb{I}_{Q_+}(v, m) + 8(1 \mp \varepsilon)\mathbb{I}_{Q_-}(v, m).$$

Note, $||f^{\pm\varepsilon}||_\infty < 9$, while $||f^0||_\infty = 8$, therefore all the distributions considered in this proof are $1/9$-smooth. To visualize these new perturbed distributions, recall the construction of $\mathbb{P}^0_{(V,M)}$ using the coin toss and the uniform random variable U: in this case, the coin is biased, and the probability of tails is $(1\pm\varepsilon)/2$. It is possible to explicitly compute the expected utility under these perturbed distributions for any bid $b \in [0, 1]$: $\mathbb{E}^{\pm\varepsilon}[\text{Util}_t(b)]$ is equal to

$$\begin{cases} \frac{b}{4}(1 - 8b) \pm \varepsilon\frac{b}{4}(1 - 8b) & \text{if } b \in [0, \frac{1}{4}) \\ -\frac{1}{8}(16b^2 - 14b + 3) \pm \frac{\varepsilon}{4}(8b^2 - 11b + 2) & \text{if } b \in [\frac{1}{4}, \frac{1}{2}) \\ \frac{1}{2}(1 - 2b \mp \frac{3}{4}\varepsilon) & \text{if } b \in [\frac{1}{2}, 1] \end{cases} \quad (2)$$

For visualization, we refer to Figure 3 (bottom). The crucial property of the distributions we constructed is that the instantaneous regret of not playing in the "correct" region is $\Omega(\varepsilon)$; formally, we have the following result (missing proofs are deferred to the long version).

Claim 1. *There exists two disjoint intervals I_+ and I_- in $[0, 1]$ such that, for any $\varepsilon \in (0, 1/2)$ and any time t, the following hold:*

$$\max_{x \in [0,1]} \mathbb{E}^{\pm\varepsilon}[\text{Util}_t(x)] \geq \mathbb{E}^{\pm\varepsilon}[\text{Util}_t(b)] + \frac{1}{128}\varepsilon, \text{ for all } b \notin I_\pm$$

Since the two distributions are "ε-close[¶]", any learner needs at least $1/\varepsilon^2$ rounds to discriminate which ones of the two distributions it is actually facing, paying each error with an instantaneous regret of $\Omega(\varepsilon)$ (Claim 1). All in all, any learner suffers a regret that is $\Omega(\varepsilon \cdot \frac{1}{\varepsilon^2} + \varepsilon T)$, which is of the desired $\Omega(\sqrt{T})$ order for the right choice of $\varepsilon \approx T^{-1/2}$.

As the last step of the proof, we formalize the above argument. Fix $\varepsilon = 1/(4\sqrt{T})$ and rename $\mathbb{P}^{+\varepsilon} = \mathbb{P}^1$ and $\mathbb{P}^{-\varepsilon} = \mathbb{P}^2$. Similarly, denote with I_1 and I_2 the two intervals I_+ and I_- as in the statement of Claim 1. For each $j \in \{0, 1, 2\}$, consider the run of $\mathcal{A}$ against the stochastic environment which draws $(V_1, M_1), (V_2, M_2), \ldots$ i.i.d. from $\mathbb{P}^j$. Let N_1 be the random variable that counts the number of times that algorithm $\mathcal{A}$ posts a bid in I_1. Similarly, N_2 counts the number of times that it posts a bid in I_2. For $i = 1, 2$, we have the following crucial relation between the expected value of N_i under $\mathbb{P}^i$. Note, the results hold because the two distributions are so similar that the deterministic algorithm $\mathcal{A}$ bids in the wrong region a constant fraction of the time steps.

Claim 2. *The following inequality holds:* $\frac{1}{2}\sum_{i=1,2}\mathbb{E}^i[N_i] \leq \frac{3}{4}T$.

[¶]In the full version, we formally prove that their total variation is at most $\Theta(\varepsilon)$.

We finally have all the ingredients to conclude the proof. Consider an environment that selects uniformly at random either $\mathbb{P}^1$ or $\mathbb{P}^2$ and then draws the (V_t, M_t) i.i.d. following it. We prove that the algorithm $\mathcal{A}$ suffers linear regret against this randomized environment and, by a simple averaging argument, against at least one of them. Specifically, if $b_i^\star$ is the optimal bid in the scenario determined by $\mathbb{P}^i$, for $i \in \{1, 2\}$, we have

$$R_T(\mathcal{A}) \geq \frac{1}{2} \sum_{i=1,2} \mathbb{E}^i \left[\sum_{t=1}^T \text{Util}_t(b_i^\star) - \sum_{t=1}^T \text{Util}_t(B_t) \right]$$

$$\overset{(*)}{\geq} \frac{1}{1024\sqrt{T}} \sum_{i=1,2} \mathbb{E}^i \left[T - N_i \right] \overset{(\circ)}{\geq} \frac{1}{512\sqrt{T}} \left(T - \frac{3}{4}T \right) = \frac{\sqrt{T}}{2048}$$

where $(*)$ follows by Claim 1 and choice of ε, and $(\circ)$ by Claim 2. $\qquad\square$

4 THE ADVERSARIAL SETTING

In this section we complete the perspective on repeated bidding in first-price auction by investigating the adversarial environment. In particular, we consider two models: the standard one, where the sequence $(V_1, M_1), (V_2, M_2), \ldots$ is chosen upfront in a deterministic oblivious way, and the smooth environment, where the sequence $(V_1, M_1), (V_2, M_2), \ldots$ is some σ-smooth stochastic process. In Section 4.1 we construct an algorithm achieving $T^{2/3}$ regret in the bandit feedback model under the smoothness assumption; this result, together with the lower bound of the same order for the semi-transparent feedback (Theorem 3) settles the problem for these two feedback regimes. Then, in Section 4.2 we provide another upper bound, namely an algorithm achieving $\sqrt{T}$ regret in the transparent feedback model under the smoothness assumption; this result, together with the lower bound of the same order for the semi-transparent feedback (Theorem 5) settles the problem for these two feedback regimes. Finally, in Section 4.3 we provide a lower bound proving that the non-smooth adversarial environment is too hard to learn, even when the learner has access to full feedback.

4.1 Smooth – Bandit Feedback

The smoothness assumption regularizes the objective function: if (V_t, M_t) is smooth, then the expected utility is Lipschitz.

Lemma 4 (Lipschitzness). *Let (V_t, M_t) be a σ-smooth random variable in $[0, 1]^2$. Then the induced expected utility function $\mathbb{E}\left[\text{Util}_t(\cdot)\right]$ is $2/\sigma$-Lipschitz in $[0, 1]$:*

$$\left| \mathbb{E}\left[\text{Util}_t(y) - \text{Util}_t(x)\right] \right| \leq \frac{2}{\sigma}|y - x|, \quad \forall x, y \in [0, 1]. \quad (3)$$

PROOF. Let $x > y$ be any two bids in $[0, 1]$, we have:

$$\left| \mathbb{E}[\text{Util}_t(x) - \text{Util}_t(y)] \right|$$
$$= \left| \mathbb{E}\left[(V_t - x)\mathbb{I}\{M_t \leq x\} - (V_t - y)\mathbb{I}\{M_t \leq y\} \right] \right|$$
$$= \left| \mathbb{E}\left[(V_t - x)\mathbb{I}\{y < M_t \leq x\} + (x - y)\mathbb{I}\{M_t \leq y\}) \right] \right|$$
$$\leq \mathbb{P}\left[M_t \in [x, y]\right] + (x - y) \leq \frac{2}{\sigma}(x - y). \quad\square$$

Interestingly, we only need the marginal distribution of M_t to be σ-smooth for the previous lemma to hold. This Lipschitzness property has the immediate corollary that any fine enough discretization of $[0, 1]$ contains a bid whose utility is close the the optimal one.

Lemma 5 (Discretization Lemma). *Let $\mathcal{X}$ be any finite grid of bids in $[0, 1]$, and let $\delta(\mathcal{X})$ be the largest distance of a point in $[0, 1]$ to $\mathcal{X}$ (i.e., $\delta(\mathcal{X}) = \max_{p \in [0,1]} \min_{x \in \mathcal{X}} |p - x|$), then if each pair of random variables $(V_1, M_1), \ldots, (V_T, M_T)$ is σ-smooth, we have the following:*

$$\sup_{b \in [0,1]} \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(b) \right] - \max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(x) \right] \leq 3\frac{\delta(\mathcal{X})}{\sigma}T .$$

PROOF. Fix any such sequence and let $b^\star$ a fixed bid such that

$$\sup_{b \in [0,1]} \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(b) \right] \leq \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(b^\star) \right] + \frac{\delta(\mathcal{X})}{\sigma}T. \quad (4)$$

If $b^\star$ is in $\mathcal{X}$ there is nothing to prove, otherwise these exists $x^\star \in \mathcal{X}$ such that $|b^\star - x^\star| \leq \delta(\mathcal{X})$ (by definition of $\delta(\mathcal{X})$). It holds that

$$\sum_{t=1}^T \mathbb{E}\left[\text{Util}_t(b^\star) - \text{Util}_t(x^\star) \right] \overset{(L)}{\leq} \sum_{t=1}^T \frac{2}{\sigma}|b^\star - x^\star| \leq 2\frac{\delta(\mathcal{X})}{\sigma}T.$$

where (L) follows by Lipschitzness and Lemma 4. The right-hand side with Equation (4) concludes the proof of the lemma. $\qquad\square$

We can combine the above discretization lemma with any (optimal) bandits algorithm to get the desired bound on the regret. For details, we refer to the pseudocode of DISCRETIZED BANDIT.

DISCRETIZED BANDIT

1: **input:** Time horizon T, bandit algorithm $\widetilde{\mathcal{A}}$ for gains in $[-1, 1]$, grid of K bids $\mathcal{X}$
2: Initialize $\widetilde{\mathcal{A}}$ on K actions, one for each $x \in \mathcal{X}$, time horizon T
3: **for** each round $t = 1, 2, \ldots, T$ **do**
4:     Receive from $\widetilde{\mathcal{A}}$ the bid $B_t \in \mathcal{X}$
5:     Post bid B_t and observe feedback Z_t
6:     Reconstruct $\text{Util}_t(B_t)$ from Z_t and feed it to $\widetilde{\mathcal{A}}$

THEOREM 6. *Consider the problem of repeated bidding in first-price auctions in an adversarial σ-smooth environment with bandit feedback. Then there exists a learning algorithm $\mathcal{A}$ such that*

$$R_T(\mathcal{A}) \leq \frac{29}{\sigma} T^{2/3}.$$

PROOF. We prove that algorithm DISCRETIZED BANDIT with the right choice of learning algorithm $\widetilde{\mathcal{A}}$ and grid of bids $\mathcal{X}$ achieves the desired bound on the regret. As learning algorithm $\widetilde{\mathcal{A}}$ we use (a rescaled version of) the Poly INF algorithm [4]: since Poly INF is designed to run with gains in $[0, 1]$ while the utilities we observe are in $[-1, 1]$, we first apply the reward transformation $x \mapsto \frac{x+1}{2}$ to the observed utilities. This transformation costs a multiplicative factor of 2 in the regret guarantees of Poly INF.

The analysis builds on the discretization result in Lemma 5, by choosing as $\mathcal{X}$ the uniform grid of $\lceil T^{2/3} \rceil + 1$ equally spaced bids on $[0, 1]$ (note, $\delta(\mathcal{X})$ becomes $T^{-1/3}$). Fix any σ-smooth environment $\mathcal{S}$, by Lemma 5, the following chain of inequalities holds:

$$\max_{b \in [0,1]} \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(b) \right] \leq \max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(x) \right] + \frac{6}{\sigma}T^{2/3}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^T \text{Util}_t(B_t) \right] + \frac{6}{\sigma}T^{2/3} + 23T^{2/3} \leq \frac{29}{\sigma}T^{2/3}.$$

The second inequality follows from the guarantees of (the rescaled version of) Poly INF [4, Theorem 11]. □

4.2 Smooth – Transparent Feedback

For transparent feedback, we combine two tools: the adversarial discretization result (Lemma 5) and the algorithm Exp3.FPA for learning with transparent feedback on a finite grid. Note, using any other $\sqrt{KT}$ black box learning algorithm (like in the previous section for bandits) would yield a suboptimal regret bound of $T^{2/3}$.

THEOREM 7. *Consider the problem of repeated bidding in first-price auctions in an adversarial σ-smooth environment with transparent feedback. Then there exists a learning algorithm $\mathcal{A}$ such that*

$$R_T(\mathcal{A}) \le 6 \left(\frac{1}{\sigma} + \sqrt{\ln T} \right) \sqrt{T} .$$

PROOF. Consider algorithm Exp3.FPA on the uniform grid $\mathcal{X}$ of $\lceil \sqrt{T} \rceil + 1$ bids, with $\delta(\mathcal{X}) \le \sqrt{T}$. Fix any σ-smooth environment $\mathcal{S}$, Lemma 5 implies the following:

$$\max_{b \in [0,1]} \mathbb{E} \left[\sum_{t=1}^{T} \text{Util}_t(b) \right] \le \max_{x \in \mathcal{X}} \mathbb{E} \left[\sum_{t=1}^{T} \text{Util}_t(x) \right] + \frac{6}{\sigma} \sqrt{T}$$

$$\le \mathbb{E} \left[\sum_{t=1}^{T} \text{Util}_t(B_t) \right] + 6 \left(\frac{1}{\sigma} + \sqrt{\ln T} \right) \sqrt{T},$$

where the second inequality follows from Proposition 2. □

4.3 The (Non-Smooth) Adversarial Model

The positive results provided in the previous sections hold under either one of two conditions: the environment is stochastic and the learner has at least the semi-transparent feedback (Theorem 1 says that bandit feedback is not enough) or the environment uses smooth distributions. These settings allow the learner to compute a discrete class of representative bids efficiently. In this section, we formally argue that learning is impossible if any of these assumptions is dropped. Specifically, the standard adversarial environment that generates the sequence without any smoothness constraint is too strong. In particular, we construct a randomized sequence $(V_1, M_1), (V_2, M_2), \dots$ that induces any learner to suffer at least linear regret. This construction shares some similarities with the lower bound construction in Theorem 1, the main difference being that the best bid $b^\star$ is randomized and hidden in such a way that even a learner having access to full feedback cannot pin-point it.

THEOREM 8. *Consider the problem of repeated bidding in first-price auctions in an adversarial environment with full feedback. Then, any learning algorithm $\mathcal{A}$ satisfies $R_T(\mathcal{A}) \ge T/24$.*

PROOF. We prove the result via Yao's principle, showing that there exists a randomized environment $\mathcal{S}$ such that any deterministic learning algorithm suffers $T/24$ regret against it. The random sequence posted by $\mathcal{S}$ is based on two randomized auxiliary sequences $L_1, L_2, \dots$ and $U_1, U_2, \dots$ defined as follows. They are initiated to $L_0 = 1/2$, $U_0 = 2/3$. They then evolve recursively as follows:

$$\begin{cases} L_t = L_{t-1} + \frac{2}{3}\Delta_{t-1} \text{ and } U_t = U_{t-1}, \text{ with probability } \frac{1}{2}, \\ U_t = U_{t-1} - \frac{2}{3}\Delta_{t-1} \text{ and } L_t = L_{t-1}, \text{ with probability } \frac{1}{2}, \end{cases}$$

where $\Delta_{t-1} = U_{t-1} - L_{t-1}$. For each realized sequence of the (L_t, U_t) pairs, the actual sequence of the (M_t, V_t) selected by $\mathcal{S}$ is constructed as follows. At each time step t, the environment selects $(M_t, V_t) = (L_t, 1)$ or $(U_t, 0)$, uniformly at random; so that the distribution is characterized by two levels of independent randomness: the auxiliary sequence of shrinking intervals and the choice between $(L_t, 1)$ and $(U_t, 0)$.

We move our attention to the expected performance of the best fixed bid in hindsight. For each realization of the random auxiliary sequence, there exists a bid $B^\star$ such that (*i*) it wins all the auctions (V_t, M_t) of the form $(L_t, 1)$ (which we may call "good auctions" because they bring positive utility when won) and (*ii*) it loses all the auctions (V_t, M_t) of the form $(U_t, 0)$ (called "bad auctions" because they bring negative utility). Thus its expected utility at each time step is at least $1/6$: with probability $1/2$ the environment selects a good auction, which induces a utility of $(1 - L_t) \ge 1/3$. All in all, the optimal bid achieves an expected utility of at least $T/6$.

Consider now the performance of any deterministic algorithm $\mathcal{A}$: for any fixed time $t > 1$ and possible realization of the past observations, the learner posts some deterministic bid B_t. If $B_t < L_{t-1}$, then it gets 0 utility, so we only consider the following cases:

- If $B_t \in [L_{t-1}, L_{t-1} + \frac{1}{3}\Delta_{t-1})$, then the bidder gets the item with probability $1/4$ ($L_t = L_{t-1}$, V_t is set to 1 and $M_t = L_t$) with an expected utility of $(1-L_t)/4 \le 1/8$.
- If $B_t \in [L_{t-1} + \frac{1}{3}\Delta_{t-1}, L_{t-1} + \frac{2}{3}\Delta_{t-1})$, the bidder gets the item with probability $1/2$ (when $L_t = L_{t-1}$ and $U_t = U_{t-1} - \frac{2}{3}\Delta_{t-1}$) for an expected utility of $\frac{1}{4}(1 - 2L_{t-1} - \frac{1}{3}\Delta_{t-1}) \le 0$
- If $B_t \in [L_{t-1} + \frac{2}{3}\Delta_{t-1}, U_{t-1})$, the bidder gets the item with probability $3/4$ (when $L_t = L_{t-1}$ and when $U_t = U_{t-1}$, $V_t = 1$ and $M_t = L_t$) for an expected utility of $\frac{1}{4}(1 - L_{t-1}) - \frac{1}{4}(L_{t-1} + \frac{1}{3}\Delta_{t-1}) + \frac{1}{4}(1 - L_{t-1} - \frac{2}{3}\Delta_{t-1}) \le \frac{1}{8}$
- If $B_t \ge U_{t-1}$, the bidder always gets the item with a negative expected utility.

All in all, the expected utility of any deterministic algorithm is at most $T/8$. If we compare this quantity with the lower bound on the expected utility of the best bid in hindsight, we get the desired result: $\mathbb{E}[R_T(\mathcal{A}, \mathcal{S})] \ge T/6 - T/8 = T/24$. □

A final observation: the main ingredient in the proof is the elaborate auxiliary sequence. To construct it, we only needed the non-smoothness of M_t, while we may have chosen the valuations V_t to be smooth, say uniformly in $[0, 1/4]$ for the bad auctions and in $[3/4, 1]$ for the good ones.

5 CONCLUSION

Motivated by the recent shift from second to first-price auctions in online advertising markets, this paper comprehensively analyzes the online learning problem of repeated bidding in first-price auctions under the realistic assumption that the bidder does not know its valuation before bidding. We characterize the minimax regret achievable for different levels of transparency in the auction format and different data generation models, considering both the stochastic i.i.d. and the standard adversarial model, while also considering smoothness. Although our regret rates are tight in their dependence on the time horizon T, a natural open problem is studying their minimax dependence on the smoothness parameter σ. This

paper belongs to the long line of research that studies economic problems from the online learning perspective; an intriguing open problem consists in offering a unified framework to characterize in a satisfying way all these games with partial feedback, similar to what has been done for partial monitoring and feedback graphs.

ACKNOWLEDGMENTS

REFERENCES

[1] Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Fast rate learning in stochastic first price bidding. In *ACML*, volume 157 of *Proceedings of Machine Learning Research*, pages 1754–1769. PMLR, 2021.

[2] 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. doi: 10.1137/140989455.

[3] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, 2009.

[4] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

[5] Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. Learning to bid in contextual first price auctions. In *WWW*, pages 3489–3497. ACM, 2023.

[6] Maria-Florina Balcan, Travis Dick, and Ellen Vitercik. Dispersion for data-driven algorithm design, online learning, and private optimization. In *FOCS*, pages 603–614. IEEE Computer Society, 2018. doi: 10.1109/FOCS.2018.00064.

[7] Santiago R. Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab S. Mirrokni, and Jon Schneider. Contextual bandits with cross-learning. *NeurIPS*, 2019.

[8] 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. doi: 10.1287/moor.2014.0663.

[9] Dirk Bergemann and Johannes Hörner. Should first-price auctions be transparent? *American Economic Journal: Microeconomics*, 10(3):177–218, 2018.

[10] Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. No-regret learning in bilateral trade via global budget balance. In *STOC*. ACM, 2024.

[11] Jason Bigler. Rolling out first price auctions to Google Ad Manager partners. https://www.blog.google/products/admanager/rolling-out-first-price-auctions-google-ad-manager-partners/, 2019. Accessed April 7, 2023.

[12] 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.

[13] Nataša Bolić, Tommaso Cesari, and Roberto Colomboni. An online learning theory of brokerage. *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems*, 2024.

[14] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *EC*, pages 289–309. ACM, 2021. doi: 10.1145/3465456.3467645.

[15] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Repeated bilateral trade against a smoothed adversary. In

[16] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. *CoRR*, abs/2307.09478, 2023.

[17] Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Bilateral trade: A regret minimization perspective. *Mathematics of Operations Research*, 49(1):171–203, 2024. doi: 10.1287/moor.2023.1351.

[18] Xiaotie Deng, Xinyan Hu, Tao Lin, and Weiqiang Zheng. Nash convergence of mean-based learning algorithms in first price auctions. In *WWW*. ACM, 2022.

[19] Nishanth Dikkala and Éva Tardos. Can credit increase revenue? In *WINE*, volume 8289 of *Lecture Notes in Computer Science*, pages 121–133. Springer, 2013.

[20] Paul Duetting, Guru Guruganesh, Jon Schneider, and Joshua Ruizhi Wang. Optimal no-regret learning for one-sided lipschitz functions. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 8836–8850. PMLR, 2023.

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

[22] Michal Feldman, Brendan Lucier, and Noam Nisan. Correlated and coarse equilibria of single-item auctions. In *WINE*, volume 10123 of *Lecture Notes in Computer Science*, pages 131–144. Springer, 2016. doi: 10.1007/978-3-662-54110-4_10.

[23] Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to bid without knowing your value. In *EC*, pages 505–522. ACM, 2018.

[24] Zhe Feng, Guru Guruganesh, Christopher Liaw, Aranyak Mehta, and Abhishek Sethi. Convergence analysis of no-regret bidding algorithms in repeated auctions. In *AAAI*, pages 5399–5406. AAAI Press, 2021. doi: 10.1609/aaai.v35i6.16680.

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

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

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

[28] Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568*, 2020.

[29] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795*, 2020.

[30] Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.

[31] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *J. ACM*, 66(4):30:1–30:77, 2019. doi: 10.1145/3299873.

[32] Yoav Kolumbus and Noam Nisan. Auctions between regret-minimizing agents. In *WWW*, pages 100–111. ACM, 2022. doi: 10.1145/3485447.3512055.

[33] 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.

[34] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Second Edition*. Cambridge University Press, 2017. doi: 10.1017/CBO9780511813603.

[35] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011.

[36] Dravyansh Sharma, Maria-Florina Balcan, and Travis Dick. Learning piecewise lipschitz functions in changing environments. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3567–3577. PMLR, 2020.

[37] Sarah Sluis. Big changes coming to auctions, as exchanges roll the dice on first-price. https://adexchanger.com/platforms/big-changes-coming-auctions-exchanges-roll-dice-first-price/, 2017. Accessed July 3, 2023.

[38] 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. doi: 10.1145/990308.990310.

[39] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 1562–1583. JMLR.org, 2016.

[40] Matt Wong. Moving AdSense to a first-price auction. https://blog.google/products/ads-commerce/our-move-to-a-first-price-auction/, 2021. Accessed July 6, 2023.

[41] Wei Zhang, Brendan Kitts, Yanjun Han, Zhengyuan Zhou, Tingyu Mao, Hao He, Shengjun Pan, Aaron Flores, San Gultekin, and Tsachy Weissman. MEOW: A space-efficient nonparametric bid shading algorithm. In *KDD*. ACM, 2021.

[42] Wei Zhang, Yanjun Han, Zhengyuan Zhou, Aaron Flores, and Tsachy Weissman. Leveraging the hints: Adaptive bidding in repeated first-price auctions. *NeurIPS*, 2022.