Learning to Bid Without Knowing your Value

ZHE FENG, Harvard University, USA

CHARA PODIMATA, Harvard University, USA

VASILIS SYRGKANIS, Microsoft Research, New England, USA

We address online learning in complex auction settings, such as sponsored search auctions, where the value of the bidder is unknown to her, evolving in an arbitrary manner and observed only if the bidder wins an allocation. We leverage the structure of the utility of the bidder and the partial feedback that bidders typically receive in auctions, in order to provide algorithms with regret rates against the best fixed bid in hindsight, that are *exponentially faster* in convergence in terms of dependence on the action space, than what would have been derived by applying a generic bandit algorithm and almost equivalent to what would have been achieved in the full information setting. Our results are enabled by analyzing a new online learning setting with outcome-based feedback, which generalizes learning with feedback graphs. We provide an online learning algorithm for this setting, of independent interest, with regret that grows only logarithmically with the number of actions and linearly only in the number of potential outcomes (the latter being very small in most auction settings). Last but not least, we show that our algorithm outperforms the bandit approach experimentally and that this performance is robust to dropping some of our theoretical assumptions or introducing noise in the feedback that the bidder receives.

CCS Concepts: • **Information systems** → **Sponsored search advertising**; • **Theory of computation** → **Online learning algorithms**; **Algorithmic game theory and mechanism design**;

1 INTRODUCTION

A standard assumption in the majority of the literature on auction theory and mechanism design is that participants that arrive in the market have a clear assessment of their valuation for the goods at sale. This assumption might seem acceptable in small markets with infrequent auction occurrences and amplitude of time for participants to do market research on the goods. However, it is an assumption that is severely violated in the context of the digital economy.

In settings like online advertisement auctions or eBay auctions, bidders participate very frequently in auctions that they have very little knowledge about the good at sale, e.g. the value produced by a user clicking on an ad. It is unreasonable, therefore, to believe that the participant has a clear picture of this value. However, the inability to pre-assess the value of the good before arriving to the market is alleviated by the fact that due to the large volume of auctions in the digital economy, participants can employ *learning-by-doing* approaches.

In this paper we address exactly the question of *how would you learn to bid approximately optimally in a repeated auction setting where you do not know your value for the good at sale and where that value could potentially be changing over time.* The setting of learning in auctions with an unknown value poses an interesting interplay between exploration and exploitation that is not

standard in the online learning literature: in order for the bidder to get feedback on her value she has to bid high enough to win the good with higher probability and hence, receive some information about that underlying value. However, the latter requires paying a higher price. Thus, there is an inherent trade-off between value-learning and cost. The main point of this paper is to address the problem of learning how to bid in such unknown valuation settings with partial *win-only feedback*, so as to minimize the regret with respect to the best fixed bid in hindsight.

On one extreme, one can treat the problem as a Multi-Armed Bandit (MAB) problem, where each possible bid that the bidder could submit (e.g. any multiple of a cent between 0 and some upper bound on her value) is treated as an arm. Then, standard MAB algorithms (see e.g. [14]) can achieve regret rates that scale *linearly* with the number of such discrete bids. The latter can be very slow and does not leverage the structure of utilities and the form of partial feedback that arises in online auction markets. Recently, the authors in [42] addressed learning with such type of partial feedback in the context of repeated single-item second-price auctions. However, their approach does not address more complex auctions and is tailored to the second-price auction.

Our Contributions. Our first main contribution is to introduce a novel online learning setting with partial feedback, which we denote *learning with outcome-based feedback* and which could be of independent interest. We show that our setting captures online learning in many repeated auction scenarios including all types of single-item auctions, value-per-click sponsored search auctions, value-per-impression sponsored search auctions and multi-item auctions.

Our setting generalizes the setting of learning with feedback graphs [4, 35], in a way that is crucial for applying it to the auction settings of interest. At a high level, the setting is defined as follows: The learner chooses an action $b \in B$ (e.g. a bid in an auction). The adversary chooses an *allocation function* x_t, that maps an action to a distribution over a set of potential outcomes O (e.g. the probability of getting a click) and a *reward function* r_t that maps an action-outcome pair to a reward (utility conditional on getting a click with a bid of b). Then, an outcome o_t is chosen based on distribution $x_t(b)$ and a reward $r_t(b, o_t)$ is observed. The learner also gets to observe the function x_t and the reward function $r_t(\cdot, o_t)$ for the realized outcome o_t (i.e. in our auction setting: she learns the probability of a click, the expected payment as a function of her bid and, *if she gets clicked*, her value).

Our second main contribution is an algorithm which we call WIN-EXP, which achieves regret $O\left(\sqrt{T|O|\log(|B|)}\right)$. The latter is inherently better than the generic multi-armed bandit regret of $O\left(\sqrt{T|B|}\right)$, since in most of our applications $|O|$ will be a small constant (e.g. $|O| = 2$ in sponsored search) and takes advantage of the particular feedback structure. Our algorithm is a variant of the EXP3 algorithm [8], with a carefully crafted unbiased estimate of the utility of each action, which has lower variance than the unbiased estimate used in the standard EXP3 algorithm. This result could also be of independent interest and applicable beyond learning in auction settings. Our approach is similar to the importance weighted sampling approach used in EXP3 so as to construct unbiased estimates of the utility of each possible action. Our main technical insight is how to incorporate the allocation function feedback that the bidder receives to construct unbiased estimates with small variance, leading to dependence only in the number of outcomes and not the number of actions. As we discuss in the related work, despite the several similarities, our setting has differences with existing partial feedback online learning settings, such as learning with experts [8], learning with feedback graphs [4, 35] and contextual bandits [2].

This setting engulfs learning in many auctions of interest where bidders learn their value for a good only when they win the good and where the good which is allocated to the bidder is determined by some randomized allocation function. For instance, when applied to the case of single-item first-price, second-price or all-pay auctions, our setting corresponds to the case where

the bidders observe their value for the item auctioned at each iteration only when they win the item. Moreover, after every iteration, they observe the critical bid they would have needed to submit to win (for instance, by observing the bids of others or the clearing price). The latter is typically the case in most government auctions or in auction settings similar to eBay.

Our flagship application is that of *value-per-click sponsored search auctions*. These are auctions were bidders repeatedly bid in an auction for a slot in a keyword impression on a search engine. The complexity of the sponsored search ecosystem and the large volume of repeated auctions has given rise to a plethora of automated bidding tools (see e.g. [43]) and has made sponsored search an interesting arena for automated learning agents. Our framework captures the fact that in this setting the bidders observe their value for a click only when they get clicked. Moreover, it assumes that the bidders also observe the average probability of click and the average cost per click for any bid they could have submitted. The latter is exactly the type of feedback that the automated bidding tools can receive via the use of *bid simulators* offered by both major search engines [24–26, 38]. In Figure 1 we portray example interfaces from these tools, where we see that the bidders can observe exactly these allocation and payment curves assumed by our outcome-based-feedback formulation. Not using this information seems unreasonable and a waste of available information. Our work shows how one can utilize this partial feedback given by the auction systems to provide improved learning guarantees over what would have been achieved if one took a fully bandit approach. In the experimental section, we also show that our approach outperforms that of the bandit one even if the allocation and payment curves provided by the system have some error that could stem from errors in the machine learning models used in the calculation of these curves by the search engines. Hence, even when these curves are not fully reliable, our approach can offer improvements in the learning rate.
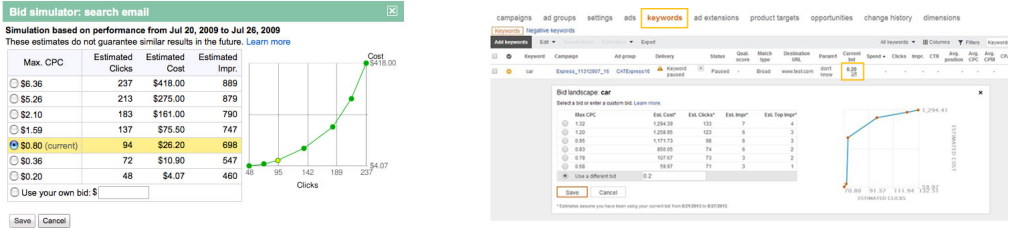

Fig. 1. Example interfaces of bid simulators of two major search engines, Google Adwords (left) and BingAds (right), that enables learning the allocation and the payment function. (sources [33, 41])

We also extend our results to cases where the space of actions is a continuum (e.g. all bids in an interval $[0, 1]$). We show that in many auction settings, under appropriate assumptions on the utility functions, a regret of $O\left(\sqrt{T \log(T)}\right)$ can be achieved by simply discretizing the action space to a sufficiently small uniform grid and running our WIN-EXP algorithm. This result encompasses the results of [42] for second price auctions, learning in first-price and all-pay auctions, as well as learning in sponsored search with smoothness assumptions on the utility function. We also show how smoothness of the utility can easily arise due to the inherent randomness that exists in the mechanism run in sponsored search.

Finally, we provide two further extensions: *switching regret* and *feedback-graphs over outcomes*. The former adapts our algorithm to achieve good regret against a sequence of bids rather than a fixed bid, which has implications on the faster convergence to approximate efficiency of the outcome (price of anarchy). Feedback graphs address the idea that in many cases the learner could be receiving information about other items other than the item he won (through correlations in the values for these items). This essentially corresponds to adding a feedback graph over

outcomes and when outcome o_t is chosen, then the learner learns the reward function $r_t(\cdot, o)$ for all neighboring outcomes o in the feedback graph. We provide improved results that mainly depend on the dependence number of the graph rather than the number of possible outcomes.

Related Work. Our work lies on the intersection of two main areas: No regret learning in Game Theory and Mechanism Design and Contextual Bandits.

No regret learning in Game Theory and Mechanism Design. No regret learning has received a lot of attention in the Game Theory and Mechanism Design literature [18]. Most of the existing literature, however, focuses on the problem from the side of the auctioneer, who tries to maximize revenue through repeated rounds without knowing a priori the valuations of the bidders [5, 6, 12, 13, 16, 20, 21, 23, 29, 32, 36, 37, 39]. These works are centered around different auction formats like the sponsored search ad auctions, the pricing of inventory and the single-item auctions. Our work is mostly related to Weed et al. [42], who adopt the point of view of the bidders in repeated second-price auctions and who also analyze the case where the true valuation of the item is revealed to the bidders only when they win the item. Their setting falls into the family of settings for which our novel and generic WIN-EXP algorithm produces good regret bounds and as a result, we are able to fully retrieve the regret that their algorithms yield, up to a tiny increase in the constants. Hence, we give an easier way to recover their results. Closely related to our work are the works of [22] and [9]. Dikkala and Tardos [22] analyzes a setting where bidders have to experiment in order to learn their valuations, and show that the seller can increase revenue by offering an initial credit to them, in order to give them incentives to experiment. Balseiro and Gur [9] introduce a family of dynamic bidding strategies in repeated second-price auctions, where advertisers adjust their bids throughout the campaign. They analyze both regret minimization and market stability. There are two key differences to our setting; first, Balseiro and Gur consider the case where the goal of the bidders is the expediture rate in a way that guarantees that the available campaign budget will be spent in an optimal *pacing* way and second, because of their target being the expenditure rate at every timestep t, they assume that the bidders get information about the value of the slot being auctioned and based on this information they decide how to adjust their bid. Moreover, several works analyze the properties of auctions when bidders adopt a no-regret learning strategy [11, 15, 40]. None of these works, however, addresses the question of learning more efficiently in the unknown valuation model and either invokes generic MAB algorithms or develops tailored full information algorithms when the bidder knows his value. Another line of research takes a Bayesian approach to learning in repeated auctions and makes large market assumptions, analyzing learning to bid with an unknown value under a Mean Field Equilibrium condition [1, 10, 28][1].

Learning with partial feedback. Our work is also related to the literature in *learning with partial feedback* [2, 14]. To establish this connection we observe that the *policies* and the *actions* in contextual bandit terminology translate into *discrete bids* and *groups of bids for which we learn the rewards* in our work. The difference between these two is the fact that for each *action* in contextual bandits we get a single reward, whereas for our setting we observe a *group* of rewards; one for each action in the group. Moreover, the fact that we allow for randomized outcomes adds extra complication, non existent in contextual bandits. In addition, our work is closely related to the literature in *online learning with feedback graphs* [3, 4, 19, 35]. In fact, we propose a new setting in online learning, namely, *learning with outcome-based feedback*, which is a generalization of learning with feedback graphs and is essential when applied to a variety of auctions which include sponsored search, single-item second-price, single-item first-price and single-item all-pay auctions. Moreover, the fact

---

[1]No-regret learning is complementary and orthogonal to the mean field approach, as it does not impose any stationarity assumption on the evolution of valuations of the bidder or the behavior of his opponents.

that the learner only learns the probability of each outcome and not the actual realization of the randomness, is similar in nature to a feedback graph setting, but where the bidder does not observe the whole graph. Rather, she observes a distribution over feedback graphs and for each bid she learns with what probability each feedback graph would arise. For concreteness, consider the case of sponsored search and suppose for now that the bidder gets even more information than what we assume and also observes the bids of her opponents. She still does not observe whether she would get a click if she falls on the slot below but only the probability with which she would get a click in the slot below. If she could observe whether she would still get a click in the slot below, then we could in principle construct a feedback graph that would say that for all bids were the bidder gets a slot her reward is revealed, and for every bid that she does not get a click, her reward is not revealed. However, this is not the structure that we have and essentially this corresponds to the case where the feedback graph is not revealed, as analyzed in [19] and for which no improvement over the full bandit feedback is possible. However, we show that this impossibility is amended by the fact that the learner observes the probability of a click and hence for each possible bid, she observes the probability with which each feedback graph would have happened. This is enough for a low variance unbiased estimate.

2 LEARNING IN AUCTIONS WITHOUT KNOWING YOUR VALUE

For simplicity of exposition, we start with a simple single-dimensional mechanism design setting, but our results extend to multi-dimensional (multi-item) mechanisms, as we will see in Section 4. Let n be the number of bidders. Each bidder has a value $v_i \in [0, 1]$ *per-unit of a good* and submits a bid $b_i \in B$, where B is a discrete set of bids (e.g. a uniform ϵ-grid of $[0, 1]$). Given the bid profile of all bidders, the auction allocates a unit of the good to the bidders. The allocation rule for bidder i is given by $X_i(b_i, b_{-i})$. Moreover, the mechanism defines a per-unit payment function $P_i(b_i, b_{-i}) \in [0, 1]$. The overall utility of the bidder is quasi-linear, i.e. $u_i(b_i, b_{-i}) = (v_i - P_i(b_i, b_{-i})) \cdot X_i(b_i, b_{-i})$.

Online Learning with Partial Feedback. The bidders participate in this mechanism repeatedly. At each iteration, each bidder has some value v_{it} and submits a bid b_{it}. The mechanism has some time-varying allocation function $X_{it}(\cdot, \cdot)$ and payment function $P_{it}(\cdot, \cdot)$. We assume that the bidder does *not* know her value v_{it}, nor the bids of her opponents $b_{-i,t}$, nor the allocation and payment functions, *before* submitting a bid.

At the end of each iteration, she gets an item with probability $X_{it}(b_{it}, b_{-i,t})$ and observes her value v_{it} for the item only when she gets one [2]. Moreover, we assume that she gets to observe her allocation and payment functions for that iteration, i.e. she gets to observe two functions $x_{it}(\cdot) = X_{it}(\cdot, b_{-i,t})$ and $p_{it}(\cdot) = P_{it}(\cdot, b_{-i,t})$. Finally, she receives utility $(v_{it} - p_{it}(b_{it})) \cdot \mathbb{1}\{\text{item is allocated to her}\}$ or in other words expected utility $u_{it}(b_{it}) = (v_{it} - p_{it}(b_{it})) \cdot x_{it}(b_{it})$. Given that we focus on learning from the perspective of a single bidder we will drop the index i from all notation and instead write, $x_t(\cdot), p_t(\cdot), u_t(\cdot), v_t$, etc. The goal of the bidder is to achieve small expected regret with respect to any fixed bid in hindsight: $R(T) = \sup_{b^* \in B} \mathbb{E}\left[\sum_{t=1}^{T} (u_t(b^*) - u_t(b_t))\right] \leq o(T)$.

3 ABSTRACTION: LEARNING WITH WIN-ONLY FEEDBACK

Let us abstract the learner's problem to a setting that could be of interest beyond auction settings.

Learning with Win-Only Feedback. Every day a learner picks an action b_t from a finite set B. The adversary chooses a reward function $r_t : B \to [-1, 1]$ and an allocation function $x_t : B \to [0, 1]$. The learner wins a reward $r_t(b)$ with probability $x_t(b)$. Let $u_t(b) = r_t(b)x_t(b)$ be the learner's

---

[2]E.g. in sponsored search, the good allocated is the probability of getting clicked, and you only observe your value if you get clicked.

expected utility from action b. After each iteration, if she won the reward then she learns the whole reward function $r_t(\cdot)$, while she *always* learns the allocation function $x_t(\cdot)$.

Can the learner achieve regret $O(\sqrt{T \log(|B|)})$ rather than bandit-feedback regret $O(\sqrt{T|B|})$?

In the case of the auction learning problem, the reward function $r_t(b)$ takes the parametric form $r_t(b) = v_t - p_t(b)$ and the learner needs to learn v_t and $p_t(\cdot)$ at the end of each iteration, when she wins the item. This is in line with the feedback structure we described in the previous section.

We consider the following adaptation of the EXP3 algorithm with unbiased estimates based on the information received. It is notationally useful throughout the section to denote with A_t the event of *winning a reward at time t*. Then, we can write: $\Pr[A_t | b_t = b] = x_t(b)$ and $\Pr[A_t] = \sum_{b \in B} \pi_t(b) x_t(b)$, where with $\pi_t(\cdot)$ we denote the multinomial distribution from which bid b is drawn. With this notation we define our WIN-EXP algorithm in Algorithm 1. We note here that our generic family of the WIN-EXP algorithms can be parametrized by the step-size η, the estimate of the utility $\tilde{u}_t$ that the learner gets at each round and the feedback structure that she receives.

Algorithm 1 WIN-EXP algorithm for learning with win-only feedback

---

Let $\pi_1(b) = \frac{1}{|B|}$ for all $b \in B$ (i.e. the uniform distribution over bids), $\eta = \sqrt{\frac{2\log(|B|)}{5T}}$
for each iteration t **do**
    Draw a bid b_t from the multinomial distribution based on $\pi_t(\cdot)$
    Observe $x_t(\cdot)$ and if reward is won also observe $r_t(\cdot)$
    Compute estimate of utility:
        If reward is won $\tilde{u}_t(b) = \frac{(r_t(b)-1)\Pr[A_t | b_t = b]}{\Pr[A_t]}$; otherwise, $\tilde{u}_t(b) = -\frac{\Pr[\neg A_t | b_t = b]}{\Pr[\neg A_t]}$.
    Update $\pi_t(\cdot)$ as in Exponential Weights Update: $\forall b \in B : \pi_{t+1}(b) \propto \pi_t(b) \cdot \exp\{\eta \cdot \tilde{u}_t(b)\}$

Bounding the Regret. We first bound the first and second moment of the unbiased estimates built at each iteration in the WIN-EXP algorithm.

LEMMA 3.1. *At each iteration t, for any action $b \in B$, the random variable $\tilde{u}_t(b)$ is an unbiased estimate of the true expected utility $u_t(b)$, i.e.: $\forall b \in B : \mathbb{E}[\tilde{u}_t(b)] = u_t(b) - 1$ and has expected second moment bounded by: $\forall b \in B : \mathbb{E}\left[(\tilde{u}_t(b))^2\right] \leq \frac{4\Pr[A_t | b_t = b]}{\Pr[A_t]} + \frac{\Pr[\neg A_t | b_t = b]}{\Pr[\neg A_t]}$.*

PROOF. Let A_t denote the event that the reward was won. We have:

$$\mathbb{E}[\tilde{u}_t(b)] = \mathbb{E}\left[\frac{(r_t(b)-1) \cdot \Pr[A_t | b_t = b]}{\Pr[A_t]}\mathbb{1}\{A_t\} - \frac{\Pr[\neg A_t | b_t = b]}{\Pr[\neg A_t]}\mathbb{1}\{\neg A_t\}\right]$$
$$= (r_t(b)-1)\Pr[A_t | b_t = b] - \Pr[\neg A_t | b_t = b]$$
$$= r_t(b)\Pr[A_t | b_t = b] - 1 = u_t(b) - 1$$

Similarly for the second moment:

$$\mathbb{E}\left[\tilde{u}_t(b)^2\right] = \mathbb{E}\left[\frac{(r_t(b)-1)^2 \cdot \Pr[A_t | b_t = b]^2}{\Pr[A_t]^2}\mathbb{1}\{A_t\} + \frac{\Pr[\neg A_t | b_t = b]^2}{\Pr[\neg A_t]^2}\mathbb{1}\{\neg A_t\}\right]$$
$$= \frac{(r_t(b)-1)^2 \cdot \Pr[A_t | b_t = b]^2}{\Pr[A_t]} + \frac{\Pr[\neg A_t | b_t = b]^2}{\Pr[\neg A_t]} \leq \frac{4\Pr[A_t | b_t = b]}{\Pr[A_t]} + \frac{\Pr[\neg A_t | b_t = b]}{\Pr[\neg A_t]}$$

where the last inequality holds since $r_t(\cdot) \in [-1, 1]$ and $x_t(\cdot) \in [0, 1]$. □

We are now ready to prove our main theorem:

THEOREM 3.2 (REGRET OF WIN-EXP). *The regret of the WIN-EXP algorithm with the aforementioned unbiased estimates and step size $\sqrt{\frac{2\log(|B|)}{5T}}$ is: $4\sqrt{T\log(|B|)}$.*

PROOF. Observe that regret with respect to utilities $u_t(\cdot)$ is equal to regret with respect to the translated utilities $u_t(\cdot) - 1$. We use the fact that the exponential weights update with an unbiased estimate $\tilde{u}_t(\cdot) \le 0$ of the true utilities, achieves expected regret of the form[3]:

$$R(T) \le \frac{\eta}{2} \sum_{t=1}^{T} \sum_{b \in B} \pi_t(b) \cdot \mathbb{E}\left[(\tilde{u}_t(b))^2\right] + \frac{1}{\eta} \log(|B|)$$

Invoking the bound on the second moment by Lemma 3.1, we get:

$$R(T) \le \frac{\eta}{2} \sum_{t=1}^{T} \sum_{b \in B} \pi_t(b) \cdot \left(\frac{4\Pr[A_t|b_t = b]}{\Pr[A_t]} + \frac{\Pr[\neg A_t|b_t = b]}{\Pr[\neg A_t]}\right) + \frac{1}{\eta} \log(|B|) \le \frac{5}{2}\eta T + \frac{1}{\eta} \log(|B|)$$

Picking $\eta = \sqrt{\frac{2\log(|B|)}{5T}}$, we get the theorem. □

4 BEYOND BINARY OUTCOMES: OUTCOME-BASED FEEDBACK

In the win-only feedback framework there were two possible outcomes that could happen: either you win the reward or not. We now consider a more general problem, where there are more than two outcomes and you learn your reward function for the outcome you won. Moreover, the outcome that you won is also a probabilistic function of your action. The proofs for the results presented in this section can be found in Appendix B in the full version[4].

Learning with Outcome-Based Feedback. Every day a learner picks an action b_t from a finite set B. There is a set of payoff-relevant outcomes O. The adversary chooses a reward function $r_t : B \times O \to [-1, 1]$, which maps an action and an outcome to a reward and he also chooses an allocation function $x_t : B \to \Delta(O)$, which maps an action to a distribution over the outcomes. Let $x_t(b, o)$ be the probability of outcome o under action b. An outcome $o_t \in O$ is chosen based on distribution $x_t(b_t)$. The learner wins reward $r_t(b_t, o_t)$ and observes the whole outcome-specific reward function $r_t(\cdot, o_t)$. She *always* learns the allocation function $x_t(\cdot)$ after the iteration. Let $u_t(b) = \sum_{o \in O} r_t(b, o) \cdot x_t(b, o)$ be the expected utility from action b.

We consider the following adaptation of the EXP3 algorithm with unbiased estimates based on the information received. It is notationally useful throughout the section to consider o_t as the random variable of the outcome chosen at time t. Then, we can write: $\Pr_t[o_t|b] = x_t(b, o_t)$ and $\Pr_t[o_t] = \sum_{b \in B} \pi_t(b)\Pr_t[o_t|b] = \sum_{b \in B} \pi_t(b) \cdot x_t(b, o_t)$. With this notation and based on the feedback structure, we define our WIN-EXP algorithm for learning with outcome-based feedback in Algorithm 2.

THEOREM 4.1 (REGRET OF WIN-EXP WITH OUTCOME-BASED FEEDBACK). *The regret of Algorithm 2 with $\tilde{u}_t(b) = \frac{(r_t(b,o_t)-1)\Pr_t[o_t|b]}{\Pr_t[o_t]}$ and step size $\sqrt{\frac{\log(|B|)}{2T|O|}}$ is: $2\sqrt{2T|O|\log(|B|)}$.*

---

[3] A detailed proof of this claim can be found in Appendix G.
[4] https://arxiv.org/pdf/1711.01333.pdf

Algorithm 2 WIN-EXP algorithm for learning with outcome-based feedback

---

Let $\pi_1(b) = \frac{1}{|B|}$ for all $b \in B$ (i.e. the uniform distribution over bids), $\eta = \sqrt{\frac{\log(|B|)}{2T|O|}}$

for each iteration t **do**

    Draw an action b_t from the multinomial distribution based on $\pi_t(\cdot)$

    Observe $x_t(\cdot)$, observe chosen outcome o_t and associated reward function $r_t(\cdot, o_t)$

    Compute estimate of utility:

$$\tilde{u}_t(b) = \frac{(r_t(b, o_t) - 1)\Pr_t[o_t|b]}{\Pr_t[o_t]} \tag{1}$$

    Update $\pi_t(\cdot)$ based on the Exponential Weights Update:

$$\forall b \in B : \pi_{t+1}(b) \propto \pi_t(b) \cdot \exp\{\eta \cdot \tilde{u}_t(b)\} \tag{2}$$

Applications to Learning in Auctions. We now present a series of applications of the main result of this section to several learning in auction settings, even beyond single-item or single-dimensional ones.

Example 4.2 (Second-price auction). Suppose that the mechanism ran at each iteration is just the second price auction. Then, we know that the allocation function $X_i(b_i, b_{-i})$ is simply of the form: $\mathbb{1}\{b_i \geq \max_{j \neq i} b_j\}$ and the payment function is simply the second highest bid. In this case, observing the allocation and payment functions at the end of the auction boils down to observing the highest other bid. In fact, in this case we have a trivial setting where the bidder gets an allocation of either 0 or 1 and if we let $B_t = \max_{j \neq i} b_{jt}$, then the unbiased estimate of the utility takes the simpler form (assuming the bidder always loses in case of ties) of: if $b_t > B_t : \tilde{u}_t(b) = \frac{(v_{it} - B_t - 1)\mathbb{1}\{b > B_t\}}{\sum_{b' > B_t} \pi_t(b')}$ and $\tilde{u}_t(b) = \frac{\mathbb{1}\{b \leq B_t\}}{\sum_{b' \leq B_t} \pi_t(b')}$ in any other case. Our main theorem gives regret $4\sqrt{T \log(|B|)}$. We note that this theorem recovers exactly the results of Weed et al. [42], by using as B a uniform $1/\Delta^o$ discretization of the bidding space, for an appropriately defined constant Δ^o (see Appendix B.1 for an exact comparison of the results).

Example 4.3 (Value-per-click auctions). This is a variant of the binary outcome case analyzed in Section 3, where $O = \{A, \neg A\}$, i.e. get clicked or not. Hence, $|O| = 2$, and $r_t(b, A) = v_t - p_t(b)$, while $r_t(b, \neg A) = 0$. Our main theorem gives regret $4\sqrt{T \log(|B|)}$.

Example 4.4 (Unit-demand multi-item auctions). Consider the case of K items at an auction where the bidder has value v_k for only one item k. Given a bid b, the mechanism defines a probability distribution over the items that the bidder will be allocated and also defines a payment function, which depends on the bid of the bidder and the item allocated. When a bidder gets allocated an item k she gets to observe her value v_{kt} for that item. Thus, the set of outcomes is equal to $O = \{1, \ldots, K + 1\}$, with outcome $K + 1$ associated with not getting any item. The rewards are also of the form: $r_t(b, k) = v_{kt} - p_t(b, k)$ for some payment function $p_t(b, k)$ dependent on the auction format. Our main theorem then gives regret $2\sqrt{2(K + 1)T \log(|B|)}$.

4.1 Batch Rewards Per-Iteration and Sponsored Search Auctions

We now consider the case of sponsored search auctions, where the learner participates in multiple auctions per-iteration. At each of these auctions she has a chance to win and get feedback on her value. To this end, we abstract the *learning with win-only feedback* setting to a setting where multiple rewards are awarded per-iteration. The allocation function remains the same throughout the iteration but the reward functions can change.

Outcome-Based Feedback with Batch Rewards. Every iteration t is associated with a set of *reward contests* I_t. The learner picks an action b_t, which is used at *all* reward contests. For each $\tau \in I_t$ the adversary picks an outcome specific reward function $r_\tau : B \times O \to [-1, 1]$. Moreover, the adversary chooses an allocation function $x_t : B \to \Delta(O)$, which is not τ-dependent. At each τ, an outcome o_τ is chosen based on distribution $x_t(b_t)$ and independently. The learner receives reward $r_\tau(b_t, o_\tau)$ from that contest. The overall realized utility from that iteration is the average reward: $\frac{1}{|I_t|} \sum_{\tau \in I_t} r_\tau(b_t, o_\tau)$, while the expected utility from any bid b is: $u_t(b) = \frac{1}{|I_t|} \sum_{\tau \in I_t} \sum_{o \in O} r_\tau(b, o) \cdot x_t(b, o)$. We assume that at the end of each iteration the learner receives as feedback the average reward function conditional on each realized outcome, i.e. if we let $I_{to} = \{\tau \in I_t : o_\tau = o\}$, then the learner learns the function: $Q_t(b, o) = \frac{1}{|I_{to}|} \sum_{\tau \in I_{to}} r_\tau(b, o)$ (with the convention that $Q_t(b, o) = 0$ if $|I_{to}| = 0$) as well as the realized frequencies $f_t(o) = \frac{|I_{to}|}{|I_t|}$ for all outcomes o.

With this at hand we can define the *batch-analogue* of our unbiased estimates of the previous section. To avoid any confusion we define: $\Pr_t[o|b] = x_t(b, o)$ and $\Pr_t[o] = \sum_{b \in B} \pi_t(b) \Pr_t[o|b]$, to denote that these probabilities only depend on t and not on τ. The estimate of the utility will be:

$$\tilde{u}_t(b) = \sum_{o \in O} \frac{\Pr_t[o|b]}{\Pr_t[o]} f_t(o) \left(Q_t(b, o) - 1\right) \tag{3}$$

We show the full algorithm with outcome-based batch-reward feedback in Algorithm 3.

Algorithm 3 WIN-EXP algorithm for learning with outcome-based batch-reward feedback

---

Let $\pi_1(b) = \frac{1}{|B|}$ for all $b \in B$ (i.e. the uniform distribution over bids), $\eta = \sqrt{\frac{\log(|B|)}{2T|O|}}$
for each iteration t **do**

Draw an action b_t from the multinomial distribution based on $\pi_t(\cdot)$
Observe $x_t(\cdot)$, chosen outcomes $o_\tau, \forall \tau \in I_t$, average reward function conditional on each realized outcome $Q_t(b, o)$ and the realized frequencies for each outcome $f_t(o) = \frac{|I_{to}|}{|I_t|}$.
Compute estimate of utility:

$$\tilde{u}_t(b) = \sum_{o \in O} \frac{\Pr_t[o|b]}{\Pr_t[o]} f_t(o) \left(Q_t(b, o) - 1\right) \tag{4}$$

Update $\pi_t(\cdot)$ based on the Exponential Weights Update:

$$\forall b \in B : \pi_{t+1}(b) \propto \pi_t(b) \cdot \exp\left\{\eta \cdot \tilde{u}_t(b)\right\} \tag{5}$$

COROLLARY 4.5. *The WIN-EXP algorithm with the latter unbiased utility estimates and step size* $\sqrt{\frac{\log(|B|)}{2T|O|}}$, *achieves regret in the outcome-based feedback with batch rewards setting at most:* $2\sqrt{2T|O|\log(|B|)}$.

It is also interesting to note that the same result holds if instead of using $f_t(o)$ in the expected utility (Equation (10)), we used its *mean value*, which is $x_t(o, b_t) = \Pr_t[o|b_t]$. This would not change any of the derivations above. The nice property of this alternative is that the learner does not need to learn the realized fraction of each outcome, but only the expected fraction of each outcome. This is already contained in the function $x_t(\cdot, \cdot)$, which we assumed was given to the learner at the end of each iteration. Thus, with these new estimates, the learner does not need to observe $f_t(o)$. In Appendix C we also discuss the case where different periods can have different number of rewards and how to extend our estimate to that case. The batch rewards setting finds an interesting application in the case of learning in sponsored search, as we describe below.

Example 4.6 (Sponsored Search). In the case of sponsored search auctions, the latter boils down to learning the average value $\hat{v} = \frac{1}{\#clicks} \sum_{clicks} v_{click}$ for the clicks that were generated, as well as the cost-per-click function $p_t(b)$, which is assumed to be constant throughout the period t. Given these quantities, the learner can compute: $Q(b, A) = \hat{v} - p_t(b)$ and $Q(b, \neg A) = 0$. An advertiser can keep track of the traffic generated by a search engine ad and hence, can keep track of the number of clicks from the search engine and the value generated by each of these clicks (conversion). Thus, she can estimate $\hat{v}$. Moreover, she can elicit the probability of click (aka click-through-rate or CTR) curves $x_t(\cdot)$ and the cost-per-click (CPC) curves $p_t(\cdot)$ over relatively small periods of time of about a few days. See for instance the Adwords bid simulator tools offered by Google [24–26, 38][5]. Thus, with these at hand we can apply our batch reward outcome based feedback algorithm and get regret that does not grow linearly with $|B|$, but only as $4\sqrt{T \log (|B|)}$. Our main assumption is that the expected CTR and CPC curves during this relatively small period of a few days remains approximately constant. The latter holds if the distribution of click-through-rates does not change within these days and if the bids of opponent bidders also do not significantly change. This is a reasonable assumption when feedback can be elicited relatively frequently, which is the case in practice.

5 CONTINUOUS ACTIONS WITH PIECEWISE-LIPSCHITZ REWARDS

Until now we only considered discrete action spaces. In this section, we extend our discussion to continuous ones; that is, we allow the action of each bidder to lie in a continuous action space $\mathcal{B}$ (e.g. a uniform interval in $[0, 1]$). To assist us in our analysis, we are going to use the discretization result in [31] [6]. For what follows in this section, let $R(T, \mathcal{B}) = \sup_{b^* \in \mathcal{B}} \mathbb{E} \left[\sum_{t=1}^{T} (u_t(b^*) - u_t(b_t)) \right]$ be the regret of the bidder, after T rounds with respect to an action space $\mathcal{B}$. Moreover, for any pairs of action spaces B and $\mathcal{B}$ we let: $DE(B, \mathcal{B}) = \sup_{b \in \mathcal{B}} \sum_{t=1}^{T} u_t(b) - \sup_{b' \in B} \sum_{t=1}^{T} u_t(b')$, denote the discretization error incurred by optimizing over B instead of $\mathcal{B}$. The proofs of this section can be found in Appendix E.

LEMMA 5.1. ([30, 31]) *Let $\mathcal{B}$ be a continuous action space and B a discretization of $\mathcal{B}$. Then:*

$$R(T, \mathcal{B}) \leq R(T, B) + DE(B, \mathcal{B})$$

Observe now that in the setting of Weed et al. [42] the discretization error was: $DE(B, \mathcal{B}) = 0$ if $\epsilon < \Delta^o$, as we discussed in Section 4 and that was *the key* that allowed us to recover this result, without adding an extra ϵT in the regret of the bidder. To achieve that, we construct the following general class of utility functions:

Definition 5.2 (Δ^o-Piecewise Lipschitz Average Utilities). A learning setting with action space $\mathcal{B} = [0, 1]^d$, is said to have Δ^o-Piecewise Lipschitz Cumulative Utilities if the average utility function $\frac{1}{T} \sum_{t=1}^{T} u_t(b)$ satisfies the following conditions: the bidding space $[0, 1]^d$ is divided into d-dimensional cubes with edge length at least Δ^o and within each cube the utility is L-Lipschitz with respect to the ℓ_∞ norm. Moreover, for any boundary point there exists a sequence of non-boundary points whose limit cumulative utility is at least as large as the cumulative utility of the boundary point.

LEMMA 5.3 (DISCRETIZATION ERROR FOR PIECEWISE LIPSCHITZ). *Let $\mathcal{B} = [0, 1]^d$ be a continuous action space and B a uniform ϵ-grid of $[0, 1]^d$, such that $\epsilon < \Delta^o$ (i.e. B consists of all the points*

---

[5]One could argue that the CTRs that the bidder gets in this case are not accurate enough. Nevertheless, even if they have random perturbations, we show in our experimental results that for reasonable noise assumptions, WIN-EXP is preferable compared to EXP3.

[6]Kleinberg [31] discuss the uniform discretization of continuum-armed bandits and Kleinberg et al. [30] extend the results for the case of Lipschitz-armed bandits.

whose coordinates are multiples of a given ϵ). Assume that the average utility function is Δ^o-Piecewise L-Lipschitz. Then, the discretization error of B is bounded as: $DE(B, \mathcal{B}) \leq \epsilon LT$.

If we know the Lipschitzness constant L mentioned above, the time horizon T and Δ^o, then our WIN-EXP algorithm for Outcome-Based Feedback with Batch Rewards yields regret as defined by the following theorem. In Appendix E, we also show how to deal with unknown parameters L, T and Δ^o by applying a standard doubling trick.

THEOREM 5.4. Let $\mathcal{B} = [0, 1]^d$ be the action space as defined in our model and let B be a uniform ϵ-grid of $\mathcal{B}$. The WIN-EXP algorithm with unbiased estimates given by $\tilde{u}_t(b) = \sum_{o \in O} \frac{\Pr_t[o|b]}{\Pr_t[o]} f_t(o) (Q_t(b, o) - 1)$ (6) on B with $\eta = \sqrt{\frac{\log(|B|)}{2T|O|}}, \epsilon = \min\left\{\frac{1}{LT}, \Delta^o\right\}$ achieves expected regret at most $2\sqrt{2T|O|d \log\left(\max\left\{\frac{1}{\Delta^o}, LT\right\}\right)} + 1$ in the outcome-based feedback with batch rewards and Δ^o-Piecewise $L-$Lipschitz average utilities [7].

Example 5.5 (First Price and All-Pay Auctions). Consider the case of learning in first price or all-pay auctions. In the former, the highest bidder wins and pays her bid, while in the latter the highest bidder wins and every bidder pays her bid whether she wins or loses. Let B_t be the highest other bid at time t. Then the average hindsight utility of the bidder in each auction is [8]:

$$\frac{1}{T} \sum_{t=1}^T u_t(b) = \frac{1}{T} \sum_{t=1}^T v_t \cdot \mathbb{1}\{b > B_t\} - b \cdot \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{b > B_t\} \qquad \text{(first price)}$$

$$\frac{1}{T} \sum_{t=1}^T u_t(b) = \frac{1}{T} \sum_{t=1}^T v_t \cdot \mathbb{1}\{b > B_t\} - b \qquad \text{(all-pay)}$$

Let Δ^o be the *smallest* difference between the highest other bid at any two iterations t and t' [9]. Then observe that the average utilities in this setting are Δ^o-Piecewise 1-Lipschitz: Between any two highest other bids, the average allocation, $\frac{1}{T} \sum_{t=1}^T v_t \cdot \mathbb{1}\{b > B_t\}$, of the bidder remains constant and the only thing that changes is his payment which grows linearly. Hence, the derivative at any bid between any two such highest other bids is upper bounded by 1. Hence, by applying Theorem 5.4, our WIN-EXP algorithm with a uniform discretization on a ϵ-grid, for $\epsilon = \min\left\{\Delta^o, \frac{1}{T}\right\}$, achieves regret $4\sqrt{T \log\left(\max\left\{\frac{1}{\Delta^o}, T\right\}\right)} + 1$, where we used that $|O| = 2$ and $d = 1$ for any of these auctions.

5.1 Sponsored Search with Lipschitz Utilities

In this subsection, we extend our analysis of learning in the sponsored search auction model (Example 4.6) to the continuous bid space case, i.e., each bidder can submit a bid $b \in [0, 1]$. As a reminder, the utility function is: $u_t(b) = x_t(b)(\hat{v}_t - p_t(b))$, where $b \in [0, 1]$, $\hat{v}_t \in [0, 1]$ is the average value for the clicks at iteration t, $x_t(\cdot)$ is the CTR curve and $p_t(\cdot)$ is the CPC curve. These curves are implicitly formed by running some form of a Generalized Second Price auction (GSP) at each iteration to determine the allocation and payment rules. As we show in this subsection, the form of the GSP ran in reality gives rise to Lipschitz utilities, under some minimal assumptions. Therefore, we can apply the results in Section 5 to get regret bounds even with respect to the continuous bid space $\mathcal{B} = [0, 1]$ [10]. We begin by providing a brief description of the type of Generalized Second Price auction ran in practice.

---

[7]Interestingly, the above regret bound can help to retrieve two familiar expressions for the regret. First, when $L = 0$ (i.e. when the function is constant within each cube), in is the case for the second price auction analyzed in [42], $R(T) = 2\sqrt{2dT|O| \log\left(\frac{1}{\Delta^o}\right)} + 1$. Hence, we recover the bounds from the prior sections up to a tiny increase. Second, when $\Delta^o \to \infty$, then we have functions that are L-Lipschitz in the whole space $\mathcal{B}$ and the regret bound that we retrieve is: $R(T) = 2\sqrt{2dT|O| \log(LT)} + 1$, which is of the type achieved in continuous lipschitz bandit settings.

[8]For simplicity, we assume the bidder loses in case of ties, though we can handle arbitrary random tie-breaking rules.

[9]This is an analogue of the Δ^o used by [42] in second price auctions.

[10]The aforementioned Lipschitzness is also reinforced by real world data sets from Microsoft's sponsored search auction system.

Definition 5.6 (Weighted-GSP). Each bidder i is assigned a *quality score* $s_i \in [0, 1]$. Bidders are ranked according to their score-weighted bid $s_i \cdot b_i$, typically called the *rank-score*. Every bidder whose rank-score does not pass a reserve r is discarded. Bidders are allocated slots in decreasing order of *rank-score*. Each bidder is charged per-click the lowest bid she could have submitted and maintained the same slot. Hence, if a bidder i is allocated a slot k and ρ_{k+1} is the rank-score of the bidder in slot $k + 1$, then she is charged ρ_{k+1}/s_i per-click. We denote with $U_i(\mathbf{b}, \mathbf{s}, r)$, the utility of bidder i under a bid profile $\mathbf{b}$ and score profile $\mathbf{s}$.

The quality scores are typically highly random, dependent on the features of the ad and the user that is currently viewing the page. Hence, a reasonable modeling assumption is that the scores s_i at each auction are drawn i.i.d. from some distribution with CDF F_i. We now show that if the CDF F_i is Lipschitz (i.e. admits a bounded density), then the utilities of the bidders are also Lipschitz.

THEOREM 5.7 (LIPSCHITZNESS OF THE UTILITY OF WEIGHTED GSP). *Suppose that the score s_i of each bidder i in a weighted GSP is drawn independently from a distribution with an $L-$Lipschitz CDF F_i. Then, the expected utility $u_i(b_i, \mathbf{b}_{-i}, r) = \mathbb{E}_{\mathbf{s}}[U_i(b_i, \mathbf{b}_{-i}, \mathbf{s}, r)]$ is $\frac{2nL}{r}-$Lipschitz wrt b_i.*

Thus, we see that when the quality scores in sponsored search are drawn from L-Lipschitz CDFs $F_i, \forall i \in n$ and the reserve is lower bounded by $\delta > 0$, then the utilities are $\frac{2nL}{\delta}$-Lipschitz and we can achieve good regret bounds by using the WIN-EXP algorithm with batch rewards, with action space B being a uniform ϵ-grid, $\epsilon = \frac{\delta}{2nLT}$ and unbiased estimates given by Equation (6) or Equation (3). In the case of sponsored search the second unbiased estimate takes the following simple form:

$$\tilde{u}_t(b) = \frac{x_t(b) \cdot x_t(b_t)}{\sum_{b' \in B} \pi_t(b') x_t(b')} (\hat{v}_t - p_t(b) - 1) - \frac{(1 - x_t(b)) \cdot (1 - x_t(b_t))}{\sum_{b' \in B} \pi_t(b')(1 - x_t(b'))} \tag{7}$$

where $\hat{v}_t$ is the average value from the clicks that happened during iteration t, $x_t(\cdot)$ is the CTR curve, b_t is the realized bid that the bidder submitted and $\pi_t(\cdot)$ is the distribution over discretized bids of the algorithm at that iteration. We can then apply Theorem 5.4 to get the following guarantee:

COROLLARY 5.8. *The WIN-EXP algorithm run on a uniform ϵ-grid with $\epsilon = \frac{\delta}{2nLT}$, step size $\sqrt{\frac{\log(1/\epsilon)}{4T}}$ and unbiased estimates given by Equation (6) or Equation (3), when applied to the sponsored search auction setting with quality scores drawn independently from distributions with L-Lipschitz CDFs, achieves regret at most: $4\sqrt{T \log\left(\frac{2nLT}{\delta}\right)} + 1$.*

6 FURTHER EXTENSIONS

In this section, we discuss an extension to switching regret and the implications on Price of Anarchy and one to the feedback graphs setting.

6.1 Switching Regret and Implications for Price of Anarchy

We show below that our results can be extended to capture the case where, instead of having just one optimal bid b^*, there is a sequence of $C \geq 1$ *switches* in the optimal bids. Using the results presented in [27] and adapting them for our setting we get the following corollary (with proof in Appendix F).

COROLLARY 6.1. *Let $C \geq 0$ be the number of times that the optimal bid $b^* \in \mathcal{B}$ switches in a horizon of T rounds. Then, using Algorithm 2 in [27] with $\mathcal{A} \equiv$ WIN-EXP and any $\alpha \in (0, 1)$ we can achieve expected switching regret at most: $O\left(\sqrt{(C+1)^2 \left(2 + \frac{1}{\alpha}\right) 2d|O|T \log\left(\max\left\{LT, \frac{1}{\Delta^o}\right\}\right)}\right)$*

This result has implications on the price of anarchy (PoA) of auctions. In the case of sponsored search where bidders' valuations are changing over time adversarially but non-adaptively, our

result shows that if the valuation does not change more than C times, we can compete with any bid that is a function of the value of the bidder at each iteration, with regret rate given by the latter theorem. Therefore, by standard PoA arguments [34], this would imply convergence to an approximately efficient outcome at a faster rate than bandit regret rates.

6.2 Feedback Graphs over Outcomes

We now extend Section 5, by assuming that there is a directed feedback graph $G = (O, E)$ over the outcomes. When outcome o_t is chosen, the bidder observes not only the outcome specific reward function $r_t(\cdot, o_t)$, for that outcome, but also for any outcome o in the out-neighborhood of o_t in the feedback graph, which we denote with $N^{out}(o_t)$. Correspondingly, we denote with $N^{in}(o)$ the incoming neighborhood of o in G. Both neighborhoods include self-loops. Let $G_\epsilon = (O_\epsilon, E_\epsilon)$ be the sub-graph of G that contains only outcomes for which $\Pr_t[o_t] \geq \epsilon$ and subsequently, let N_ϵ^{in} and N_ϵ^{out} be the in and out neighborhoods of this sub-graph.

Based on this feedback graph we construct a WIN-EXP algorithm with step-size $\eta = \sqrt{\frac{\log(|B|)}{8T\alpha \ln\left(\frac{16|O|^2 T}{\alpha}\right)}}$, utility estimate $\tilde{u}_t(b) = \mathbb{1}\{o_t \in O_\epsilon\} \sum_{o \in N_\epsilon^{out}(o_t)} \frac{(r_t(b,o)-1)\Pr_t[o|b]}{\sum_{o' \in N_\epsilon^{in}(o)} \Pr_t[o']}$ and feedback structure as described in the previous paragraph. With these changes we can show that the regret grows as a function of the *independence number of the feedback graph*, denoted with α, rather than the *number of outcomes*. The full Algorithm can be found in Appendix A.

THEOREM 6.2 (REGRET OF WIN-EXP-G). *The regret of the WIN-EXP-G algorithm with step size* $\eta = \sqrt{\frac{\log(|B|)}{8T\alpha \ln\left(\frac{16|O|^2 T}{\alpha}\right)}}$ *is bounded by:* $R(T) \leq 2\sqrt{8\alpha T \log(|B|) \ln\left(\frac{16|O|^2 T}{\alpha}\right)} + 1$.

In the case of learning in auctions, the feedback graph over outcomes can encode the possibility that winning an item can help you uncover your value for other items. For instance, in a combinatorial auction for m items, the reader should think of each node in the feedback graph as a bundle of items. Then the graph encodes the fact that winning bundle o can teach you the value for all bundles $o' \in N^{out}(o)$. If the feedback graph has small dependence number then a much better regret is achieved than the dependence on $\sqrt{2^m}$, that would have been derived by our outcome-based feedback results of prior sections, if we treated each bundle of items separately as an outcome.

7 EXPERIMENTAL RESULTS

In this section, we present our results from our comparative analysis between EXP3 and WIN-EXP on a simulated sponsored search system that we built and which is a close proxy of the actual sponsored search algorithms deployed in the industry. We implemented the weighted GSP auction as described in definition 5.6. The auctioneer draws i.i.d rank scores that are bidder and timestep specific; as is the case throughout our paper, here we have assumed a stochastic auctioneer with respect to the rank scores. After bidding, the bidder will always be able to observe the allocation function. Now, if the bidder gets allocated to a slot and she gets clicked, then, she is able observe the *value* and the payment curve. Values are assumed to lie in $[0, 1]$ and they are obliviously adversarial. Finally, the bidders choose bids from some ϵ-discretized grid of $[0, 1]$ (in all experiments, apart from the ones comparing the regrets for different discretizations, we use $\epsilon = 0.01$) and update the probabilities of choosing each discrete bid according to EXP3 or WIN-EXP. Regret is measured with respect to the best fixed discretized bid in hindsight.

We distinguish three cases of the bidding behavior of the rest of the bidders (apart from our learner): i) all of them are *stochastic* adversaries drawing bids at random from some distribution, ii) there is a subset of them that are bidding *adaptively*, by using an EXP3 online learning algorithm

and iii) there is a subset of them that are bidding *adaptively* but using a WINEXP online learning algorithm (self play). Validating our theoretical claims, in all three cases, WIN-EXP outperforms EXP3 in terms of regret. We generate the event of whether a bidder gets clicked or not as follows: we draw a timestep specific threshold value in $[0, 1]$ and the learner gets a click in case the CTR of the slot she got allocated (if any) is greater than this threshold value. Note here that the choice of a timestep specific threshold imposes *monotonicity*, i.e. if the learner did not get a click when allocated to a slot with CTR $x_t(b)$, she should not be able to get a click from slots with lower CTRs. We ran simulations with 3 different distributions of generating CTRs, so as to understand what is the effect of different levels of click-through-rates on the variance of our regret: i) $x_t(b) \sim U[0.1, 1]$, ii) $x_t(b) \sim U[0.3, 1]$ and iii) $x_t(b) \sim U[0.5, 1]$. Finally, we address robustness of our results to errors in CTR estimation. For this, we add random noise to the CTRs of each slot and we report to the learners the allocation and payment functions that correspond to the erroneous CTRs. The noise was generated according to a normal distribution $\mathcal{N}(0, \frac{1}{m})$, where m could be viewed as the number of training samples on which a machine learning algorithm was ran in order to output the CTR estimate ($m = 100, 1000, 10000$).

For each of the following simulations, there are $N = 20$ bidders, $k = 3$ slots and we ran the experiment for each round for a total of 30 times. For the simulations that correspond to adaptive adversaries we used $a = 4$ adversaries. Our results for the *cumulative regret* are presented below. We measured ex-post regret with respect to the realized thresholds that determine whether a bidder gets clicked or not. Note that the solid plots correspond to the emprical mean of the regret, whereas the opaque bands correspond to the 10-th and 90-th percentile.

Different discretizations. In Figure 2 we present the comparative analysis of the estimated average regret of WIN-EXP vs EXP3 for different discretizations, ϵ, of the bidding space when the learner faces adversaries that are stochastic, adaptive using EXP3 and adaptive using WINEXP. As it was expected from the theoretical analysis, the regret of WIN-EXP, as the disretized space ($|B|$) increases exponentially, remains almost unchanged compared to the regret of EXP3. In summary, *finer discretization of the bid space helps our WIN-EXP algorithm's performance, but hurts the performance of EXP3.*
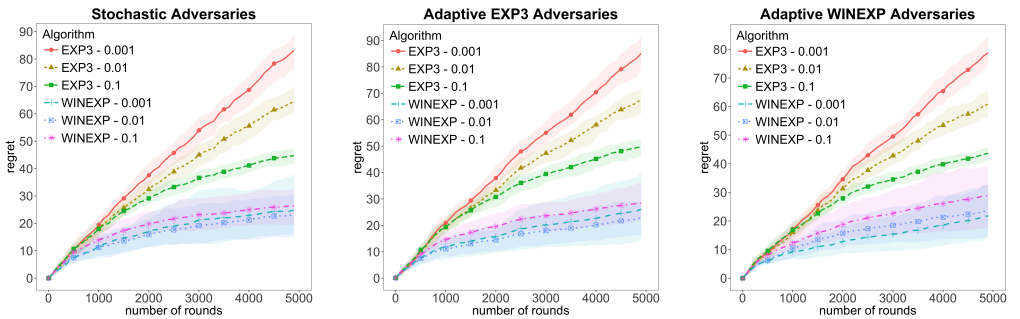

Fig. 2. Regret of WIN-EXP vs EXP3 for different discretizations ϵ (CTR $\sim U[0.5, 1]$).

Different CTR Distributions. In Figures 3, 4 and 5 we present the results of the regret performance of WIN-EXP compared to EXP3, when the learner discretizes the bidding space with $\epsilon = 0.01$ and when she faces stochastic, adaptive adversaries using EXP3 and adaptive adversaries using WINEXP, respectively. For all three cases, the estimated average regret of WIN-EXP is less than the estimated average regret that EXP3 yields.
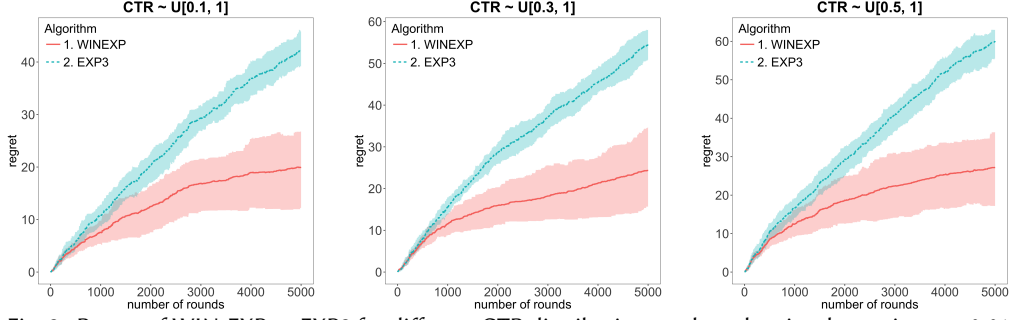
Fig. 3. Regret of WIN-EXP vs EXP3 for different CTR distributions and stochastic adversaries, $\epsilon = 0.01$.
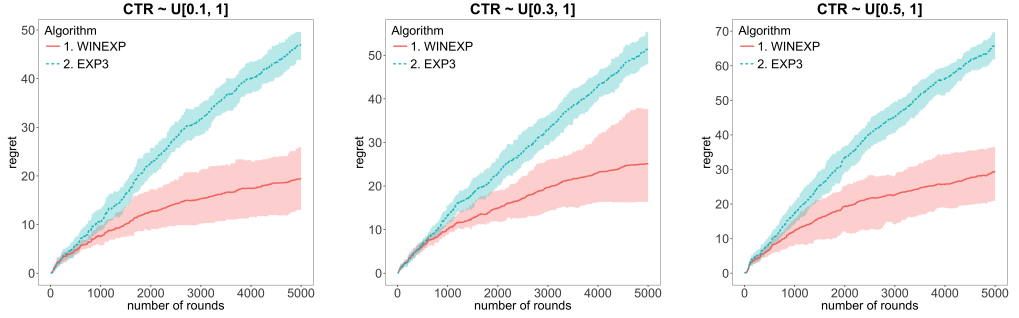

Fig. 4. Regret of WIN-EXP vs EXP3 for different CTR distributions and adaptive EXP3 adversaries, $\epsilon = 0.01$.
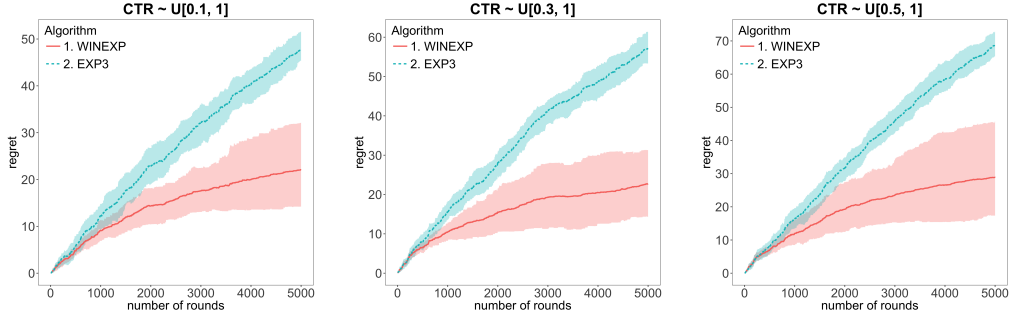

Fig. 5. Regret of WIN-EXP vs EXP3 for different CTR distributions and adaptive WINEXP adversaries, $\epsilon = 0.01$.

Robustness to Noisy CTR Estimates. In Figures 6, 7, 8 we empirically tested the robustness of our algorithm to random perturbations of the allocation function that the auctioneer presents to the learner, for perturbations of the form $\mathcal{N}\left(0, \frac{1}{m}\right)$, where m could be viewed as the number of training examples used from the auctioneer in order to derive an approximation of the allocation curve. When the number of training samples is relatively small ($m = 100$) the empirical mean of WINEXP outperforms EXP3 in terms of regret, i.e., it is more robust to such perturbations. As the number of training samples increases, WINEXP clearly outperforms EXP3. The latter validates one of our claims throughout the paper; namely, that even though the learner might not see the exact allocation curve, but a randomly perturbed proxy, WIN-EXP still performs better than the EXP3.
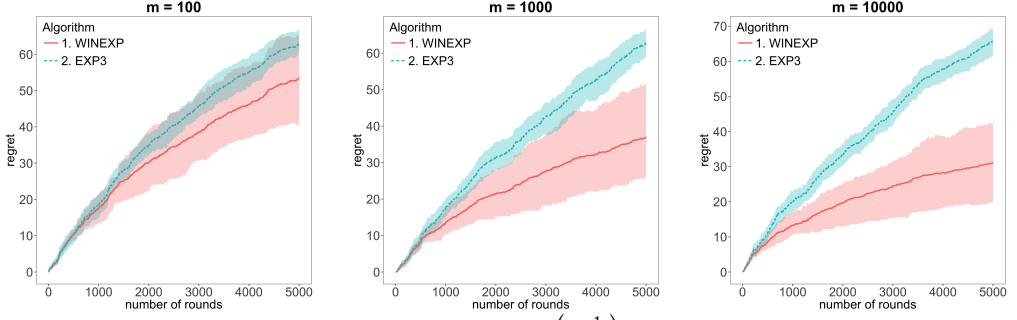
Fig. 6. Regret of WIN-EXP vs EXP3 with noise $\sim \mathcal{N}\left(0, \frac{1}{m}\right)$ for stochastic adversaries, $\epsilon = 0.01$.
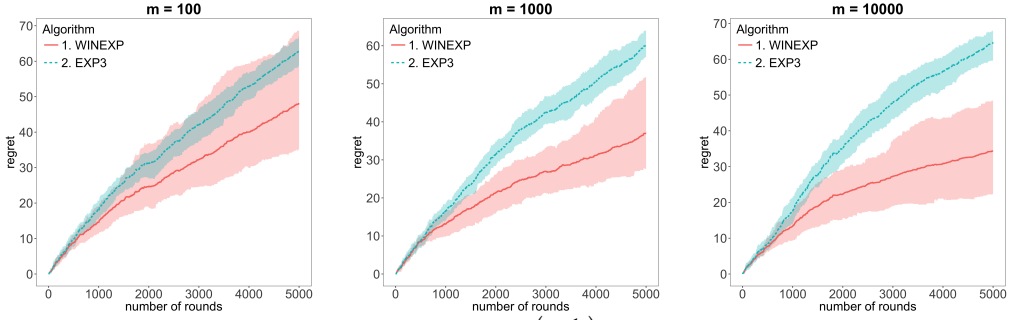

Fig. 7. Regret of WIN-EXP vs EXP3 with noise $\sim \mathcal{N}\left(0, \frac{1}{m}\right)$ for adaptive EXP3 adversaries, $\epsilon = 0.01$.
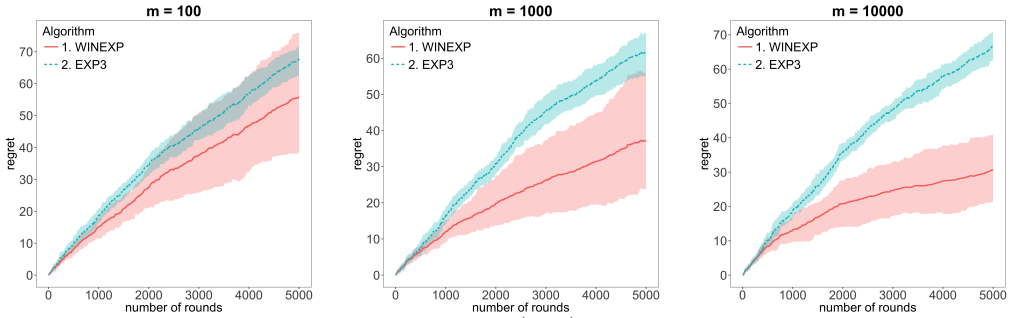

Fig. 8. Regret of WIN-EXP vs EXP3 with noise $\sim \mathcal{N}\left(0, \frac{1}{m}\right)$ for adaptive WINEXP adversaries, $\epsilon = 0.01$.

8 CONCLUSION

We addressed learning in repeated auction scenarios were bidders do not know their valuation for the items at sale. We formulated an online learning framework with partial feedback which captures the information available to bidders in typical auction settings like sponsored search and provided an algorithm which achieves almost full information regret rates. Hence, we portrayed that not knowing your valuation is a benign form of incomplete information for learning in auctions. Our experimental evaluation also showed that the improved learning rates are robust to violations of our assumptions and are valid even when the information assumed is corrupted. We believe that exploring further avenues of relaxing the informational assumptions (e.g., what if the value is only later revealed to a bidder or is contingent upon the competitiveness of the auction) or being

more robust to erroneous information given by the auction system is an interesting future research direction. We believe that our outcome-based learning framework can facilitate such future work.

REFERENCES

[1] Sachin Adlakha and Ramesh Johari. 2013. Mean field equilibrium in dynamic games with strategic complementarities. *Operations Research* 61, 4 (2013), 971–989.

[2] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. 2014. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Eric P. Xing and Tony Jebara (Eds.), Vol. 32. PMLR, Bejing, China, 1638–1646.

[3] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. 2015. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*. 23–35.

[4] Noga Alon, Nicolo Cesa-bianchi, Claudio Gentile, and Yishay Mansour. 2013. From Bandits to Experts: A Tale of Domination and Independence. In *Advances in Neural Information Processing Systems 26*, C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger (Eds.). 1610–1618.

[5] Kareem Amin, Rachel Cummings, Lili Dworkin, Michael Kearns, and Aaron Roth. 2015. Online Learning and Profit Maximization from Revealed Preferences.. In *AAAI*. 770–776.

[6] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. 2014. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*. 622–630.

[7] Sanjeev Arora, Elad Hazan, and Satyen Kale. 2012. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing* 8, 1 (2012), 121–164.

[8] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.

[9] Santiago Balseiro and Yonatan Gur. 2017. Learning in Repeated Auctions with Budgets: Regret Minimization and Equilibrium. (2017).

[10] Santiago R Balseiro, Omar Besbes, and Gabriel Y Weintraub. 2015. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science* 61, 4 (2015), 864–884.

[11] Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. 2008. Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM, 373–382.

[12] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. 2004. Online learning in online auctions. *Theoretical Computer Science* 324, 2-3 (2004), 137–146.

[13] Avrim Blum, Yishay Mansour, and Jamie Morgenstern. 2015. Learning Valuation Distributions from Partial Observation.. In *AAAI*. 798–804.

[14] Sébastien Bubeck, Nicolo Cesa-Bianchi, and others. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[15] Ioannis Caragiannis, Christos Kaklamanis, Panagiotis Kanellopoulos, Maria Kyropoulou, Brendan Lucier, Renato Paes Leme, and Éva Tardos. 2015. Bounding the inefficiency of outcomes in generalized second price auctions. *Journal of Economic Theory* 156 (2015), 343–388.

[16] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. 2015. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory* 61, 1 (2015), 549–564.

[17] Nicolò Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge University Press.

[18] Shuchi Chawla, Jason D. Hartline, and Denis Nekipelov. 2014. Mechanism design for data science. In *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*. 711–712.

[19] Alon Cohen, Tamir Hazan, and Tomer Koren. 2016. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*. 811–819.

[20] Richard Cole and Tim Roughgarden. 2014. The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 243–252.

[21] Peerapong Dhangwatnotai, Tim Roughgarden, and Qiqi Yan. 2015. Revenue maximization with a single sample. *Games and Economic Behavior* 91 (2015), 318–333.

[22] Nishanth Dikkala and Éva Tardos. 2013. Can Credit Increase Revenue?. In *Web and Internet Economics - 9th International Conference, WINE 2013, Cambridge, MA, USA, December 11-14, 2013, Proceedings*. 121–133.

[23] Michal Feldman, Tomer Koren, Roi Livni, Yishay Mansour, and Aviv Zohar. 2016. Online Pricing with Strategic and Patient Buyers. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3864–3872.

[24] Google. 2018. AdWords Bid Simulator. https://support.google.com/adwords/answer/2470105?hl=en&ref_topic=3122864. (2018). [Online; accessed 15-February-2018].

[25] Google. 2018. Bid Landscapes. https://developers.google.com/adwords/api/docs/guides/bid-landscapes. (2018). [Online; accessed 15-February-2018].

[26] Google. 2018. Bid Lanscapes. https://developers.google.com/adwords/api/docs/reference/v201710/DataService.BidLandscape. (2018). [Online; accessed 15-February-2018].

[27] András Gyorgy, Tamás Linder, and Gábor Lugosi. 2012. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory* 58, 11 (2012), 6709–6725.

[28] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. 2011. Mean field equilibria of dynamic auctions with learning. *ACM SIGecom Exchanges* 10, 3 (2011), 10–14.

[29] Yash Kanoria and Hamid Nazerzadeh. 2014. Dynamic Reserve Prices for Repeated Auctions: Learning from Bids - Working Paper. In *Web and Internet Economics - 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings*. 232.

[30] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. 2008. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM, 681–690.

[31] Robert D Kleinberg. 2005. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*. 697–704.

[32] Tomer Koren, Roi Livni, and Yishay Mansour. 2017. Bandits with Movement Costs and Adaptive Pricing. In *Proceedings of the 2017 Conference on Learning Theory (Proceedings of Machine Learning Research)*, Satyen Kale and Ohad Shamir (Eds.), Vol. 65. PMLR, Amsterdam, Netherlands, 1242–1268.

[33] Search Engine Land. 2014. Bing Ads Launches âĂIJBid LandscapeâĂİ, A Keyword Level Bid Simulator Tool. https://searchengineland.com/bing-ads-launches-bid-landscape-keyword-level-bid-simulator-tool-187219. (2014). [Online; accessed 15-February-2018].

[34] Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. 2016. Learning and efficiency in games with dynamic population. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 120–129.

[35] Shie Mannor and Ohad Shamir. 2011. From Bandits to Experts: On the Value of Side-Observations.. In *NIPS*, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 684–692.

[36] Andres M Medina and Mehryar Mohri. 2014. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 262–270.

[37] Andrés Muñoz Medina and Sergei Vassilvitskii. 2017. Revenue Optimization with Approximate Bid Predictions. *CoRR* abs/1706.04732 (2017). arXiv:1706.04732 http://arxiv.org/abs/1706.04732

[38] Microsoft. 2018. BingAds, Bid Landscapes. https://advertise.bingads.microsoft.com/en-us/resources/training/bidding-and-traffic-estimation. (2018). [Online; accessed 15-February-2018].

[39] Michael Ostrovsky and Michael Schwarz. 2011. Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of the 12th ACM conference on Electronic commerce*. ACM, 59–60.

[40] Tim Roughgarden. 2009. Intrinsic robustness of the price of anarchy. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 513–522.

[41] Search Marketing Standard. 2014. Google AdWords Improves The Bid Simulator Tool Feature. http://www.searchmarketingstandard.com/google-adwords-improves-the-bid-simulator-tool-feature. (2014). [Online; accessed 15-February-2018].

[42] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. 2016. Online learning in repeated auctions. In *Conference on Learning Theory*. 1562–1583.

[43] Wordstream. 2018. Bid Management Tools. https://www.wordstream.com/bid-management-tools. (2018). [Online; accessed 15-February-2018].

APPENDIX

The Appendix is available in our full version on arXiv.