Selling to a No-Regret Buyer

MARK BRAVERMAN, Princeton University, USA
JIEMING MAO, Princeton University, USA
JON SCHNEIDER, Princeton University, USA
MATT WEINBERG, Princeton University, USA

We consider the problem of a single seller repeatedly selling a single item to a single buyer (specifically, the buyer has a value drawn fresh from known distribution $\mathcal{D}$ in every round). Prior work assumes that the buyer is fully rational and will perfectly reason about how their bids today affect the seller's decisions tomorrow. In this work we initiate a different direction: the buyer simply runs a no-regret learning algorithm over possible bids. We provide a fairly complete characterization of optimal auctions for the seller in this domain. Specifically:

- If the buyer bids according to EXP3 (or any "mean-based" learning algorithm), then the seller can extract expected revenue arbitrarily close to the expected welfare. This auction is independent of the buyer's valuation $\mathcal{D}$, but somewhat unnatural as it is sometimes in the buyer's interest to overbid.
- There exists a learning algorithm $\mathcal{A}$ such that if the buyer bids according to $\mathcal{A}$ then the optimal strategy for the seller is simply to post the Myerson reserve for $\mathcal{D}$ every round.
- If the buyer bids according to EXP3 (or any "mean-based" learning algorithm), but the seller is restricted to "natural" auction formats where overbidding is dominated (e.g. Generalized First-Price or Generalized Second-Price), then the optimal strategy for the seller is a pay-your-bid format with decreasing reserves over time. Moreover, the seller's optimal achievable revenue is characterized by a linear program, and can be unboundedly better than the best truthful auction yet simultaneously unboundedly worse than the expected welfare.

CCS Concepts: • **Theory of computation → Online learning algorithms**; **Algorithmic game theory**; **Convergence and learning in games**;

Additional Key Words and Phrases: mechanism design, auctions, multi-armed bandits, no-regret learning

1 INTRODUCTION

Consider a bidder trying to decide how much to bid in an auction (for example, a sponsored search auction). If the auction happens to be the truthful Vickrey-Clarke-Groves auction [7, 15, 31], then the bidder's decision is easy: simply bid your value. If instead, the bidder is participating in a Generalized First-Price (GFP) or Generalized Second-Price (GSP) auction, the optimal strategy is less clear. Bidders can certainly attempt to compute a Bayes-Nash equilibrium of the associated

This work is supported by NSF CAREER award (CCF-1149888), NSF CCF-1215990, NSF CCF-1525342, NSF CCF-1412958, a Packard Fellowship in Science and Engineering, the Simons Collaboration on Algorithms and Geometry and NSF CCF-1717899.

Authors' addresses: Mark Braverman, Princeton University, 35 Olden St, Princeton, NJ, 08540, USA, mbraverm@cs.princeton.edu; Jieming Mao, Princeton University, 35 Olden St, Princeton, NJ, 08540, USA, jiemingm@cs.princeton.edu; Jon Schneider, Princeton University, 35 Olden St, Princeton, NJ, 08540, USA, js44@cs.princeton.edu; Matt Weinberg, Princeton University, 35 Olden St, Princeton, NJ, 08540, USA, smweinberg@princeton.edu.

ACM EC'18, June 18–22, 2018, Ithaca, MA, USA. ACM ISBN 978-1-4503-5829-3/18/06…$15.00
https://doi.org/10.1145/3219166.3219233

game and play accordingly, but this is unrealistic due to the need for accurate priors and extensive computation.

Alternatively, the bidders may try to learn a best-response over time (possibly offloading the learning to commercial bid optimizers). We specifically consider bidders who *no-regret learn*, as empirical work of [27] shows that bidder behavior on Bing is largely consistent with no-regret learning (i.e. for most bidders, there exists a per-click value such that their behavior guarantees no-regret for this value). From the perspective of a revenue-maximizing auction designer, this motivates the following question: **If a seller knows that buyers are no-regret learning over time, how should they maximize revenue?**

This question is already quite interesting even when there is just a single item for sale to a single buyer. We consider a model where in every round t, the seller solicits a bid $b_t \in [0, 1]$ from the buyer, then allocates the item according to some allocation rule $x_t(\cdot)$ and charges the bidder according to some pricing rule $p_t(\cdot)$ (satisfying $p_t(b) \leq b \cdot x_t(b)$ for all t, b).[1] Note that the allocation and pricing rules (henceforth, auction) can differ from round to round, and that the auction need not be truthful. Each round, the bidder has a value v_t drawn independently from $\mathcal{D}$, and uses some no-regret learning algorithm to decide which bid to place in round t, based on the outcomes in rounds $1, \dots, t-1$ (we will make clear exactly what it means for a buyer with changing valuation to play no-regret in Section 2, but one can think of v_t as providing a "context" for the bidder during round t). The same mathematical model can also represent a population $\mathcal{D}$ of many indistinguishable buyers with fixed values who each separately no-regret learn - see Section 2.3 for further details.

One default strategy for the seller is to simply to set Myerson's revenue-optimal reserve price for $\mathcal{D}$, $r(\mathcal{D})$, in every round (that is, $x_t(b_t) = I(b_t \geq r(\mathcal{D}))$, $p_t(b_t) = r(\mathcal{D}) \cdot I(b_t \geq r(\mathcal{D}))$ for all t, where $I(\cdot)$ is the indicator function). It's not hard to see that *any* no-regret learning algorithm will eventually learn to submit a winning bid during all rounds where $v_t > r(\mathcal{D})$, and a losing bid whenever $v_t < r(\mathcal{D})$. Note that this observation appeals only to the fact that the buyer guarantees no-regret, and makes no reference to any specific algorithm the buyer might use. So if $\mathsf{Rev}(\mathcal{D})$ denotes the expected revenue of the optimal reserve price when a single buyer is drawn from $\mathcal{D}$, the default strategy guarantees the seller revenue $T \cdot \mathsf{Rev}(\mathcal{D}) - o(T)$ over T rounds. The question then becomes whether or not the seller can beat this benchmark, and if so by how much.

The answer to this question isn't a clear-cut yes or no, so let's start with the following instantiation: how much revenue can the seller extract if the buyer runs EXP3 [3]? In Theorem 3.1, we show that the seller can actually do *much* better than the default strategy: it's possible to extract revenue per round equal to (almost) the full expected welfare! That is, if $\mathsf{Val}(\mathcal{D}) = \mathbb{E}_{v \leftarrow \mathcal{D}}[v]$, there exists an auction that extracts revenue $T \cdot \mathsf{Val}(\mathcal{D}) - o(T)$ for all $\mathcal{D}$.[2] It turns out this result holds not only for EXP3, but for any learning algorithm with the following (roughly stated) property: if at time t, the mean reward of action a is significantly larger than the mean reward of action b, the learning algorithm will choose action b with negligible probability. We call a learning algorithm with this property a "mean-based" learning algorithm and note that many commonly used learning algorithms - EXP3, Multiplicative Weights Update [1], and Follow-the-Perturbed-Leader [16, 18, 19] - are 'mean-based' (see Section 2 for a formal definition).

We postpone all intuition until Section 3.1 with a worked-through example, but just note here that the auction format is quite unnatural: it "lures" the bidder into submitting high bids early on by giving away the item for free, and then charging very high prices (but still bounded in $[0, 1]$)

[1] Of course, the pricing rule can be implemented by charging $p_t(b)/x_t(b)$ whenever the item is awarded if ex-post individual rationality is desired.

[2] The order of quantifiers in this sentence is correct: it is actually the same auction format that works for all $\mathcal{D}$.

near the end. The transition from “free” to “high-price” is carefully coordinated across different bids to achieve the revenue guarantee.

This result motivates two further directions. First, do there exist other no-regret algorithms for which full surplus extraction is impossible for the seller? In Theorem 3.2, we show that the answer is yes. In fact, there is a simple no-regret algorithm $\mathcal{A}$, such that when the bidder uses algorithm $\mathcal{A}$ to bid, the default strategy (set the Myerson reserve every round) is optimal for the seller. We again postpone a formal statement and intuition to Section 3.2, but just note here that the algorithm is a natural adaptation of EXP3 (or in fact, any existing no-regret algorithm) to our setting.

Finally, it is reasonable to expect that bidders might use off-the-shelf no-regret learning algorithms like EXP3, so it is still important to understand what the seller can hope to achieve if the buyer is specifically using such a “mean-based” algorithm (formal definition in Section 2). Theorem 3.1 is perhaps unsatisfying in this regard because the proposed auction is so unnatural. It turns out that the key property separating natural untruthful auctions (e.g. GSP/GFP) from the unnatural auction above is whether overbidding is a dominated strategy. That is, in our unnatural auction, if the bidder truly hopes to guarantee low regret they must seriously consider overbidding (and this is how the auction lures them into bidding way above their value). In both GSP and GFP, overbidding is dominated, so the bidder can guarantee no regret while overbidding with probability 0 in every round.

The final question we ask is the following: if the buyer is using EXP3 (or any “mean-based” algorithm), never overbids (we call such a bidder *conservative*), how much revenue can the seller extract using an auction where overbidding is dominated in every round? It turns out that the auctioneer can still outperform the default strategy, but not extract full welfare. Instead, we identify a linear program (as a function of $\mathcal{D}$) that tightly characterizes the optimal revenue the seller can achieve in this setting when the buyer’s values are drawn from $\mathcal{D}$. Moreover, we show that the auction that achieves this guarantee is natural, and can be thought of as a pay-your-bid auction with decreasing reserves over time. Finally, we show that this “mean-based revenue” benchmark, $\mathsf{MBRev}(\mathcal{D})$ lies truly in between the Myerson revenue and the expected welfare: for all c, there exists a distribution $\mathcal{D}$ over values such that $c \cdot T \cdot \mathsf{Rev}(\mathcal{D}) < \mathsf{MBRev}(\mathcal{D}) < \frac{1}{c} \cdot T \cdot \mathsf{Val}(\mathcal{D})$. In other words, the seller’s mean-based revenue may be unboundedly better than the default strategy, yet simultaneously unboundedly far from the expected welfare. We provide formal statements and a detailed proof overview of these results in Section 3.3. To briefly recap, our main results are the following:

(1) If the buyer uses a “mean-based” learning algorithm like EXP3, the seller can extract revenue $(1-\varepsilon)T \cdot \mathsf{Val}(\mathcal{D}) - o(T)$ for any constant $\varepsilon > 0$ (Theorem 3.1).

(2) There exists a natural no-regret algorithm $\mathcal{A}$ such that when the buyer bids according to $\mathcal{A}$, the seller’s default strategy (charging the Myerson reserve every round) is optimal (Theorem 3.2).

(3) If the buyer uses a “mean-based” algorithm only over undominated strategies, the seller can extract revenue $\mathsf{MBRev}(\mathcal{D})$ using an auction where overbidding is dominated in every round. Moreover, we characterize $\mathsf{MBRev}(\mathcal{D})$ as the value of a linear program, and show it can be simultaneously unboundedly better than $T \cdot \mathsf{Rev}(\mathcal{D})$ and unboundedly worse than $T \cdot \mathsf{Val}(\mathcal{D})$ (Theorems 3.6, 3.4 and 3.8).

Our plan for the remaining sections is as follows. Below, we overview our connection to related work. Section 2 formally defines our model. Section 3 works through a concrete example, providing intuition for all three results. Section 4 discusses conclusions and open problems.

1.1 Related Work

There are two lines of work that are most related to ours. The first is that of *dynamic auctions*, such as [2, 21–23, 28]. Like our model, there are T rounds where the seller has a single item for sale to a single buyer, whose value is drawn from some distribution every round. However, the buyer is fully strategic and processes fully how their choices today affect the seller's decisions tomorrow (e.g. they engage with deals of the form "pay today to get the item tomorrow"). Additional closely related work is that of Devanur et al. studying the Fishmonger problem [12, 17]. Here, there is again a single buyer and seller, and T rounds of sale. Unlike our model, the buyer draws a value from $\mathcal{D}$ once during round 0 and that value is fixed through all T rounds (so the seller could try to learn the buyer's value over time). Also unlike our model, they study perfect Bayesian equilibria (where again the buyer is fully strategic, and reasons about how their actions today affect the seller's behavior tomorrow).

In contrast to these works, while buyers in our model do care about the future (e.g. they value learning), they don't reason about how their actions today might affect the seller's decisions tomorrow. Our model better captures settings where full information about the auction is not public (and fully strategic reasoning is simply impossible without the necessary information).

Other related work considers the *Price of Anarchy* of simple combinatorial auctions when bidders no-regret learn [9, 27, 29, 30]. One key difference between this line of work and ours is that these all study welfare maximization for combinatorial auctions with rich valuation functions. In contrast, our work studies revenue maximization while selling a single item. Additionally, in these works the seller commits to a publicly known auction format, and the only reason for learning is due to the strategic behavior of other buyers. In contrast, buyers in our model have to learn *even when they are the only buyer*, due to the strategic nature of the seller.

Recent work has also considered learning from the perspective of the seller. In these works, the buyer's (or buyers') valuations are drawn from an unknown distribution, and the seller's goal is to learn an approximately optimal auction with as few samples as possible [6, 8, 11, 13, 14, 24, 25]. These works consider numerous different models and achieve a wide range of guarantees, but all study the learning problem from the perspective of the *seller*, whereas the buyer is simply myopic and participates in only one round. In contrast, it is the buyer in our model who does the learning (and there is no information for the seller to learn: the buyer's values are drawn fresh in every round).

Finally, no-regret learning in online decision problems is an extremely well-studied problem. When feedback is revealed for every possible action, one well-known solution is the multiplicative weight update rule which has been rediscovered and applied in many fields (see survey [1] for more details). Another algorithmic scheme for the online decision problem is known as Follow the Perturbed Leader [16, 18, 19]. When only feedback for the selected action is revealed, the problem is referred to as the multi-armed bandit problem. Here, similar ideas to the MWU rule are used in developing the EXP3 algorithm [3] for adversarial bandit model, and also for the contextual bandit problem [20]. Our algorithm in Theorem 3.2 bears some similarities to the low swap regret algorithm introduced in [4]. See the survey [5] for more details about the multi-armed bandit problem. Our results hold in both models (i.e. whether the buyer receives feedback for every bid they could have made, or only the bid they actually make), so we will make use of both classes of algorithms.

In summary, while there is already extensive work related to repeated sales in auctions, and even no-regret learning with respect to auctions (from both the buyer and seller perspective), our work is the first to address how a seller might adapt their selling strategy when faced with a no-regret buyer.

2 MODEL AND PRELIMINARIES

We consider a setting with 1 buyer and 1 seller. There are T rounds, and in each round the seller has one item for sale. At the start of each round t, the buyer's value $v(t)$ (known only to the buyer) for the item is drawn independently from some distribution $\mathcal{D}$ (known to both the seller and the buyer). For simplicity, we assume $\mathcal{D}$ has a finite support[3] of size m, supported on values $0 \leq v_1 < v_2 < \cdots < v_m \leq 1$. For each $i \in [m]$, v_i has probability q_i of being drawn under $\mathcal{D}$.

The seller then presents K options for the buyer, which can be thought of as "possible bids" (we will interchangeably refer to these as *options*, *bids*, or *arms* throughout the paper, depending on context). Each arm i is labelled with a bid value $b_i \in [0, 1]$, with $b_1 < \ldots, < b_K$. Upon pulling this arm at round t, the buyer receives the item with some allocation probability $a_{i,t}$, and must pay a price $p_{i,t} \in [0, a_{i,t} \cdot b_i]$. These values $a_{i,t}$ and $p_{i,t}$ are chosen by the seller during time t, but remain unknown to the buyer until he plays an arm, upon which he learns the values for that arm. All of our positive results (i.e. strategies for the seller) are *non-adaptive* (in some places called *oblivious*), in the sense that that $a_{i,t}, p_{i,t}$ are set before the first round starts. All of our negative results (i.e. upper bounds on how much a seller can possibly attain) hold even against *fully adaptive* sellers, where $a_{i,t}$ and $p_{i,t}$ can be set *even after learning the distribution of arms the buyer intends to pull in round t*.

In order for the selling strategies to possibly represent natural auctions, we require the allocation/price rules to be monotone. That is, if $i > j$, then for all t, $a_{i,t} \geq a_{j,t}$ and $p_{i,t} \geq p_{j,t}$. In other words, bidding higher should result in a (weakly) higher probability of receiving the item and (weakly) higher expected payment. We'll also insist on the existence of an arm 0 with bid $b_0 = 0$ and $a_{0,t} = 0$ for all t; i.e., an arm which charges nothing but does not give the item. Playing this arm can be thought of as not participating in the auction.

2.1 Bandits and experts

Our goal is to understand the behavior of such mechanisms when the buyer plays according to some no-regret strategy for the multi-armed bandit problem. In the classic multi-armed bandit problem a learner (in our case, the buyer) chooses one of K arms per round, over T rounds. On round t, the learner receives a reward $r_{i,t} \in [0, 1]$ for pulling arm i (where the values $r_{i,t}$ are possibly chosen adversarially). The learner's goal is to maximize his total reward.

Let I_t denote the arm pulled by the principal at round t. The *regret* of an algorithm $\mathcal{A}$ for the learner is the random variable $\text{Reg}(\mathcal{A}) = \max_i \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{I_t,t}$. We say an algorithm $\mathcal{A}$ for the multi-armed bandit problem is *δ-no-regret* if $\mathbb{E}[\text{Reg}(\mathcal{A})] \leq \delta$ (where the expectation is taken over the randomness of $\mathcal{A}$). We say an algorithm $\mathcal{A}$ is *no-regret* if it is δ-no-regret for some $\delta = o(T)$.

In the multi-armed bandits setting, the learner only learns the value $r_{i,t}$ for the arm i which he pulls on round t. In our setting, the learner will learn $a_{i,t}$ and $p_{i,t}$ explicitly (from which they can compute $r_{i,t}$). Our results (both positive and negative) also hold when the learner learns the value $r_{i,t}$ for *all* arms i (we refer this full-information setting as the *experts setting*, in contrast to the partial-information *bandits setting*). Simple no-regret algorithms exist in both the experts setting and the bandits setting. Of special interest in this paper will be a class of learning algorithms for the bandits problem and experts problem which we term 'mean-based'.

Definition 2.1 (Mean-Based Learning Algorithm). Let $\sigma_{i,t} = \sum_{s=1}^{t} r_{i,s}$. An algorithm for the experts problem or multi-armed bandits problem is *γ-mean-based* if it is the case that whenever $\sigma_{i,t} < \sigma_{j,t} - \gamma T$, then the probability that the algorithm pulls arm i on round t is at most γ. We say an algorithm is *mean-based* if it is *γ-mean-based* for some $\gamma = o(1)$.

[3] If $\mathcal{D}$ instead has infinite support, all our results hold approximately after discretization to multiples of ε. If $\mathcal{D}$ is bounded in $[0, H]$, then all our results hold after normalizing $\mathcal{D}$ by dividing by H.

Intuitively, 'mean-based' algorithms will rarely pick an arm whose current mean is significantly worse than the current best mean. Many no-regret algorithms, including commonly used variants of EXP3 (for the bandits setting), the Multiplicative Weights algorithm (for the experts setting) and the Follow-the-Perturbed-Leader algorithm (experts setting), are mean-based (see full version of this paper).

Contextual bandits. In our setting, the buyer has the additional information of their current value for the item, and hence is actually facing a *contextual bandits* problem. In (our variant of) the contextual bandits problem, each round t the learner is additionally provided with a *context* c_t drawn from some distribution $\mathcal{D}$ supported on a finite set C (in our setting, $c_t = v(t)$, the buyer's valuation for the item at time t). The adversary now specifies rewards $r_{i,t}(c)$, the reward the learner receives if he pulls arm i on round t while having context c. If we are in the full-information (experts) setting, the learner learns the values of $r_{i,t}(c_t)$ for all arms i after round t, where as if we are in the partial-information (bandits) setting, the learner only learns the value of $r_{i,t}(c_t)$ for the arm i that he pulled.

In the contextual bandits setting, we now define the regret of an algorithm $\mathcal{A}$ in terms of regret against the best "context-specific" policy π; that is, $\mathrm{Reg}(\mathcal{A}) = \max_{\pi:C\to[K]} \sum_{t=1}^{T} r_{\pi(c_t),t}(c_t) - \sum_{t=1}^{T} r_{I_t,t}(c_t)$, where again I_t is the arm pulled by M on round t. As before, we say an algorithm is δ-low regret if $\mathbb{E}[\mathrm{Reg}(M)] \leq \delta$, and say an algorithm is no-regret if it is δ-no-regret for some $\delta = o(T)$.

If the size of the context set C is constant with respect to T, then there is a simple way to construct a no-regret algorithm M' for the contextual bandits problem from a no-regret algorithm M for the classic bandits problem: simply maintain a separate instance of M for every different context $v \in C$ (in the contextual bandits literature, this is sometimes referred to as the S-EXP3 algorithm [5]). We call the algorithm we obtain this way its *contextualization*, and denote it as $\mathrm{cont}(M)$.

If we start with a mean-based learning algorithm, then we can show that its contextualization satisfies an analogue of the mean-based property for the contextual-bandits problem (proof in the full version of this paper).

Definition 2.2 (Mean-Based Contextual Learning Algorithm). Let $\sigma_{i,t}(c) = \sum_{s=1}^{t} r_{i,s}(c)$. An algorithm for the contextual bandits problem is *γ-mean-based* if it is the case that whenever $\sigma_{i,t}(c) < \sigma_{j,t}(c) - \gamma T$, then the probability $p_{i,t}(c)$ that the algorithm pulls arm i on round t if it has context c satisfying $p_{i,t}(c) < \gamma$. We say an algorithm is *mean-based* if it is *γ-mean-based* for some $\gamma = o(1)$.

Theorem 2.3. *If an algorithm for the experts problem or multi-armed bandits problem is mean-based, then its contextualization is also a mean-based algorithm for the contextual bandits problem.*

Finally, we will refer to learning algorithms that never overbid as *conservative*. We will sometimes abuse notation and instead refer to a buyer employing a conservative algorithm as conservative.

2.2 Welfare and monopoly revenue

In order to evaluate the performance of our mechanisms for the seller, we will compare the revenue the seller obtains to two benchmarks from the single-round setting of a seller selling a single item to a buyer with value drawn from distribution $\mathcal{D}$.

The first benchmark we consider is the *welfare* of the buyer, the expected value the buyer assigns to the item. This quantity clearly upper bounds the expected revenue that the seller can hope to extract per round.

Definition 2.4. The *welfare*, $\mathrm{Val}(\mathcal{D})$ is equal to $\mathbb{E}_{v\sim\mathcal{D}}[v]$.

The second benchmark we consider is the *monopoly revenue*, the maximum possible revenue attainable by the seller in one round against a rational buyer. Seminal work of Myerson [26] shows that this revenue is attainable by setting a fixed price ("monopoly/Myerson reserve") for the item, and hence can be characterized as follows.

Definition 2.5. The *monopoly revenue* (alternatively, *Myerson revenue*) $\mathsf{Mye}(\mathcal{D})$ is equal to $\max_p p \cdot \Pr_{v\sim\mathcal{D}}[v \geq p]$.

2.3 A final note on the model

For concreteness, we chose to phrase our problem as one where a single bidder whose value is repeatedly drawn independently from $\mathcal{D}$ each round engages in no-regret learning with their value as context. Alternatively, we could imagine a population of m different buyers, each with a *fixed* value v_i. Each round, exactly one buyer arrives at the auction, and it is buyer i with probability q_i. The buyers are indistinguishable to the seller, and each buyer no-regret learns (without context, because their value is always v_i). This model is mathematically equivalent to ours, so all of our results hold in this model as well if the reader prefers this interpretation instead.

3 AN ILLUSTRATIVE EXAMPLE

In this section, we overview an illustrative example to show the difference between mean-based and non-mean-based learning algorithms, and between conservative and non-conservative learners. We will not prove all claims in this section (nor carry out all calculations) as it is only meant to illustrate and provide intuition. Throughout this section, the running example will be when $\mathcal{D}$ samples $1/4$ with probability $1/2$, $1/2$ with probability $1/4$, and 1 with probability $1/4$. Note that $\mathsf{Val}(\mathcal{D}) = 1/2$ and $\mathsf{Rev}(\mathcal{D}) = 1/4$.

3.1 Mean-Based Learning

Let's first consider what the seller can do with an auction when the buyer is running a mean-based (non-conservative) learning algorithm like EXP3. The seller will let the buyer bid 0 or 1. If the buyer bids 0, they pay nothing but do not receive the item (recall that an arm of this form is required). If the buyer bids 1 in round t, they receive the item and pay some price p_t as follows: for the first half of the game ($1 \leq t \leq T/2$), the seller sets $p_t = 0$. For the second half of the game ($T/2 < t \leq T$), the seller sets $p_t = 1$.

Let's examine the behaviour of the buyer, recalling that they run a mean-based learning algorithm, and therefore (almost) always pull the arm with highest cumulative utility. The buyer with value 1 will happily bid 1 all the way through, since he is always offered the item for less than or equal to his value for the item. The buyer with value $1/2$ will bid 1 for the first $T/2$ rounds, accumulating a surplus (i.e., negative regret) of $1/2$ per round. For the next $T/2$ rounds, this surplus slowly disappears at the rate of $1/2$ per round until it disappears at time T, so the bidder with value $1/2$ will bid 1 all the way through. Finally, the bidder with value $1/4$ will bid 1 for the first $T/2$ rounds, accumulating surplus at a rate of $1/4$ per round. After round $T/2$, this surplus decreases at a rate of $3/4$ per round, until at round $2T/3$ his cumulative utility from bidding 1 reaches 0 and he switches to bidding 0.

Now let's compute the revenue. From round $T/2$ through $2T/3$, the buyer always buys the item at a price of 1, so the seller obtains $T/6$ revenue. Finally, from round $2T/3$ through T, the buyer purchases the item with probability $1/2$ and pays 1. The total revenue is $0 + T/6 + T/6 = T/3$. Note that if the seller used the default strategy, they would extract revenue only $T/4$.

Where did our extra revenue come from? First, note that the welfare of the buyer in this example is quite high: the bidder gets the item the whole way through when $v \geq 1/2$, and two-thirds of the way through when $v = 1/4$. One reason why the welfare is so high is because we give the item away for free in the early rounds. But notice also that the utility of the buyer is quite low: the buyer actually has zero utility when $v \leq 1/2$, and utility $1/2$ when $v = 1$. The reason we're able to keep the utility low, despite giving the item away for free in the early rounds is because we overcharge the bidders in later rounds (and they choose to overpay, exactly because their learning is mean-based).

In fact, by offering additional options to the buyer, we show that *it is possible for the seller to extract up to the full welfare from the buyer* (e.g. a net revenue of $T/2 - o(T)$ for this example). As in the above example, our mechanism makes use of arms which are initially very good for the buyer (giving the item away for free, accumulating negative regret), followed by a period where they are very bad for the buyer (where they pay more than their value). The trick in the construction is making sure that the good/bad intervals line up so that: a) the buyer purchases the item in every round, no matter their value (this is necessary in order to possibly extract full welfare) and b) by round T, the buyer has zero (arbitrarily small) utility, no matter their value.

Getting the intervals to line up properly so that any mean-based learner will pick the desired arms still requires some work. But interestingly, our constructed mechanism is non-adaptive and prior-independent (i.e. the same mechanism extracts full welfare *for all* $\mathcal{D}$). Theorem 3.1 below formally states the guarantees. The construction itself and the proof appear in the full version of this paper.

Theorem 3.1. *If the buyer is running a mean-based algorithm, for any constant* $\varepsilon > 0$, *there exists a strategy for the seller which obtains revenue at least* $(1 - \varepsilon)\mathsf{Val}(\mathcal{D})T - o(T)$.

Two properties should jump out as key in enabling the result above. The first is that the buyer *only* has no regret towards fixed arms and *not* towards the policy they would have used with a lower value (this is what leads the buyer to continue bidding 1 with value $1/2$ even though they have already learned to bid 0 with value $1/4$). This suggests an avenue towards an improved learning algorithm: have the bidder attempt to have no regret not only towards each fixed arm, but also towards the policy of play produced when having different values. This turns out to be exactly the right idea, and is discussed in the following subsection below.

The second key property is that we were able to "lure" the bidders into playing an arm with a free item, then overcharge them later to make up for lost revenue. This requires that the bidder consider pulling an arm with maximum bid exceeding their value, which will never happen for a conservative bidder. It turns out it is still possible to do better than the default strategy against conservative bidders, but not as well as against non-conservative mean-based bidders. Section 3.3 explores conservative mean-based bidders for this example.

3.2 Better Learning

In our bad example above, the buyer with value $1/2$ for the item slowly spends the second half of the game losing utility. While his behaviour is still no-regret (he ends up with zero net utility, which indeed is at least as good as only bidding 0), he would have been much happier to follow the actions of the buyer with value $1/4$, who started bidding 0 at $2T/3$.

Using this idea, we show how to construct a no-regret algorithm for the buyer (Algorithm 1) such that the seller receives at most the Myerson revenue every round. We accomplish this by extending an arbitrary no-regret algorithm (e.g. EXP3) by introducing "virtual arms" for each value, so that each buyer with value v has low regret not just with respect to every fixed bid, but also no-regret with respect to the policy of play as if they had a different value v' for the item (for all $v' < v$). In some ways, our construction is very similar to the construction of low internal-regret (or swap-regret) algorithms from low external-regret algorithms. The main difference is that instead of having low regret with respect to swapping actions, we have low regret with respect to swapping *contexts* (i.e. values). Theorem 3.2 below states that the seller cannot outperform the default strategy against buyers who use such algorithms to learn.

Theorem 3.2. *There exists a no-regret algorithm (Algorithm 1) for the buyer against which every seller strategy extracts no more than* $\mathsf{Mye}(\mathcal{D})T + O(m\sqrt{\delta T})$ *revenue.*

Algorithm 1 No-regret algorithm for buyer where the seller achieves no more than $\mathsf{Mye}(\mathcal{D})T+o(T)$ revenue.

```
Let $M$ be a $\delta$-no-regret algorithm for the classic multi-armed bandit problem, with $\delta = o(T)$.
  Initialize $m$ copies of $M$, $M_1$ through $M_m$.
Instance $M_i$ of $M$ will learn over $K + i - 1$ arms.
The first $K$ arms of $M_i$ (“bid arms”) correspond to the $K$ possible menu options $b_1, b_2, \ldots, b_K$.
The last $i - 1$ arms of $M_i$ (“value arms”) correspond to the $i - 1$ possible values (contexts)
  $v_1, \ldots, v_{i-1}$.
for $t = 1$ to $T$ do
    if buyer has value $v_i$ then
        Use $M_i$ to pick one arm from the $K + i - 1$ arms.
        if the arm is a bid arm $b_j$ then
            Pick the menu option $j$ (i.e. bid $b_j$).
        else if the arm is a value arm $v_j$ then
            Sample an arm from $M_j$ (but don't update its state). If it is a bid arm, pick the corre-
              sponding menu option. If it is a value arm, recurse.
        end if
        Update the state of algorithm $M_i$ with the utility of this round.
    end if
end for
```

A more further discussion of the algorithm along with a proof of Theorem 3.2 appear in the full version of this paper. The key observation in the proof is that “not regretting playing as if my value were v'” sounds a lot like “not preferring to report value v' instead of v.” This suggests that the aggregate allocation probabilities and prices paid by any buyer using our algorithm should satisfy the same constraints as a truthful auction, proving that the resulting revenue cannot exceed the default strategy (and indeed the proof follows this approach).

Finally, observe that the following corollary immediately follows. Because the seller cannot hope to get more than $\mathsf{Mye}(\mathcal{D})T + o(T)$ per round when the buyer is using Algorithm 1, and the buyer cannot hope to do better than telling the truth against a truthful auction, it is in fact a Nash for the buyer to use Algorithm 1 and the seller to set price equal to the Myerson reserve every round.

Corollary 3.3. *It is an* $o(T)$*-Nash equilibrium for the seller to set the Myerson reserve* $p(\mathcal{D})$ *in every round (any bid* $\geq p(\mathcal{D})$ *reserve wins the item and pays* $p(\mathcal{D})$*), and the buyer to use Algorithm 1.*

3.3 Mean-Based Learning and Conservative Bidders

Recall in our example that to extract revenue $T/3$, bidders with values $1/4$ and $1/2$ had to consider bidding 1. If bidders are conservative, they will simply never do this.

Although the auction in Section 3.1 is no longer viable, consider the following auction instead: in addition to the zero arm, the bidder can bid $1/4$ or $1/2$. If they bid $1/2$ in any round, they will get the item with probability 1 and pay $1/2$. If they bid $1/4$ in round $t \leq T/3$, they get nothing. If they bid $1/4$ in round $t \in (T/3, T]$, they get the item and pay $1/4$. Let's again see what the bidder will choose to do, remembering that they will always pull the arm that has provided highest cumulative utility (due to being mean-based).

Clearly, the bidder with value $1/4$ will bid $1/4$ every round (since they are conservative, they won't even consider bidding $1/2$), making a total payment of $2T/3 \cdot 1/4 \cdot 1/2 = T/12$. The bidder with value $1/2$ will bid $1/2$ for the first $T/3$ rounds, and then immediately switch to bidding $1/4$, making a total payment of $T/3 \cdot 1/2 \cdot 1/4 + 2T/3 \cdot 1/4 \cdot 1/4 = T/12$.

The bidder with value 1 will actually bid $1/2$ for the entire T rounds. To see this, observe that their cumulative surplus through round t from bidding $1/2$ is $t \cdot 1/2 \cdot 1/4 = t/8$ (t rounds by utility $1/2$ per round by probability $1/4$ of having value 1). Their cumulative surplus through round t from bidding $1/4$ is instead $(t - T/3) \cdot 3/4 \cdot 1/4 = 3t/16 - T/16 \leq t/8$ (for $t \leq T$). Because they are mean-based, they will indeed bid $1/2$ for the entire duration due to its strictly higher utility. So their total payment will be $T \cdot 1/2 \cdot 1/4 = T/8$. The total revenue is then $7T/24 > T/4$, again surpassing the default strategy (but not reaching the $T/3$ achieved against non-conservative buyers).

Let's again see where our extra revenue comes from in comparison to a truthful auction. Notice that the bidder receives the item with probability 1 conditioned on having value $1/2$, and also conditioned on having value 1. Yet somehow the bidder pays an average of $1/3$ conditioned on having value $1/2$, but an average of $1/2$ conditioned on having value 1. *This could never happen in a truthful auction*, as the bidder would strictly prefer to pretend their value was $1/2$ rather than 1. But it is entirely possible when the buyer does mean-based learning, as evidenced by this example.

In the full version of this paper, we define $\mathsf{MBRev}(\mathcal{D})$ as the value of the LP in Figure 1. In Theorems 3.6 and 3.4, we show that $\mathsf{MBRev}(\mathcal{D})T$ tightly characterizes (up to $\pm o(T)$) the optimal revenue a seller can extract against a conservative buyer. The proofs can be found in the full version of this paper.

$$
\begin{aligned}
\textbf{maximize} \quad & \sum_{i=1}^{m} q_i(v_i x_i - u_i) \\
\textbf{subject to} \quad & u_i \geq (v_i - v_j) \cdot x_j, \ \ \forall\, i, j \in [m] : i > j \\
& u_i \geq 0, 1 \geq x_i \geq 0, \ \ \forall\, i \in [m]
\end{aligned}
$$

Fig. 1. The mean-based revenue LP.

Before stating our theorems, let us parse this LP. q_i is a constant representing the probability that the buyer has value v_i (also a constant). x_i is a variable representing the average probability that the bidder gets the item with value v_i, and u_i is a variable representing the average utility of the bidder when having value v_i. Therefore, this bidder's average value is $v_i x_i$, the average price they pay is $v_i x_i - u_i$, and the objective function is simply the average revenue. The second constraints are just normalization, ensuring that everything lies in $[0, 1]$. The first line of constraints are the interesting ones. These look a lot like IC constraints that a truthful auction must satisfy, but something's missing: the LHS is clearly the utility of the buyer with value v_i for "telling the truth," but the utility of the buyer for "reporting v_j instead" is $(v_i - v_j) \cdot x_j + u_j$ (so the u_j term is missing on the RHS).

Here is a brief proof outline for why no seller can extract more revenue than $\mathsf{MBRev}(\mathcal{D})$:

(1) Since the buyer has no regret conditioned on having value v_i, their utility is at least as high as playing arm j every round, for all $j \leq i$.

(2) Since the auction never charges arm j more than v_j (conditioned on awarding the item), the buyer's utility for playing arm j every round is at least $y_j \cdot (v_i - v_j)$, where y_j is the average probability that arm j awards the item.

(3) Since the auction is monotone, and the buyer never considers overbidding, if the buyer gets the item with probability x_j conditioned on having value v_j, we must have $y_j \geq x_j$.

These three facts together show that no seller can extract more than $\mathsf{MBRev}(\mathcal{D})$ against a no-regret buyer who doesn't overbid. Observe also that step 3 is *exactly* the step that doesn't hold for buyers who consider overbidding (and is exactly what's violated in our example in Section 3.1): if the buyer ever overbids, then they might receive the item with higher probability than had they just played their own arm every round.

Theorem 3.4. *Any strategy for the seller achieves revenue at most* $\mathsf{MBRev}(\mathcal{D})T + o(T)$ *against a conservative buyer.*

The full proof of Theorem 3.4 appears in the the full version of this paper - all of the key ideas have been overviewed above.

It turns out that the previous theorem is tight; there exists an auction (taking the form of a first-price auction with descending reserve) which achieves revenue $\mathsf{MBRev}(\mathcal{D})T$ against a conservative mean-based buyer. More specifically, this auction is defined by a threshold r_t that decreases over time. If at time t you bid $b_t \geq r_t$, then you receive the item and must pay b_t; otherwise, you receive nothing and pay nothing. Moreover, the threshold function r_t which achieves optimal revenue is determined from the optimal solution to the mean-based LP: the threshold r_t drops from v_i to v_{i+1} at round x_i (where the x_i belong to some optimal solution).

To show that this is a valid strategy for the seller, we need to show that the values x_i are monotone increasing. Luckily, this follows simply from the structure of the mean-based revenue LP.

Lemma 3.5. *Let* $x_1, x_2, \ldots, x_m, u_1, u_2, \ldots, u_m$ *be an optimal solution to the mean-based revenue LP. Then for all* $i < j$, $x_i < x_j$.

Proof. We proceed by contradiction. Suppose that the sequence of x_i are not monotone; then there exists an $1 \leq i \leq m-1$ such that $x_i > x_{i+1}$. Now consider another solution of the LP, where we increase x_{i+1} to x_i, keeping the value of all other variables the same. This new solution does not violate any constraints in the LP since for all $j > i+1$, $u_j \geq (v_j - v_i) \cdot x_i \geq (v_j - v_{i+1}) \cdot x_i$. However this change increases the value of the objective by $v_{i+1}q_{i+1}(x_i - x_{i+1}) > 0$, thus contradicting the fact that $x_1, \ldots, x_m, u_1, \ldots, u_m$ was an optimal solution of the mean-based revenue LP. □

We now show that this strategy indeed achieves $\mathsf{MBRev}(\mathcal{D})T$ against a conservative buyer.

Theorem 3.6. *For any constant* $\varepsilon > 0$, *there exists a strategy for the seller gets revenue at least* $(\mathsf{MBRev}(\mathcal{D}) - \varepsilon)T - o(T)$ *against a buyer running a mean-based algorithm who overbids with probability* 0. *The strategy sets a decreasing cutoff* r_t *and for all* t *awards the item with probability* 1 *to any bid* $b_t \geq r_t$ *for price* b_t, *and with probability* 0 *to any bid* $b_t < r_t$.

Proof. We will show that: i) the buyer with value v_i receives the item for at least $x_iT - o(T)$ turns (receiving $v_i x_i T - o(T)$ total utility from the items), and ii) this buyer's net utility is at most $(u_i + \varepsilon)T + o(T)$. This implies that this buyer pays the seller at least $x_i v_i T - (u_i + \varepsilon)T - o(T)$ over the course of the T rounds; taking expectation over all v_i completes the proof.

Assume the buyer is running a γ-mean-based learning algorithm. Consider the buyer when they have value v_i. Note that

$$\sigma_{j,t}(v_i) = (v_i - v_j + \varepsilon) \cdot \max(0, t - (1 - x_j)T).$$

We first claim that after round $(1-x_i)T+\gamma T/\varepsilon$, the buyer will buy the item (i.e., choose an option that results in him getting the item) each round with probability at least $1 - m\gamma$. To see this, first note that $\sigma_{i,t}(v_i) \geq \gamma T$ when $t \geq (1-x_i)T + \gamma T/\varepsilon$. Then, since the cumulative utility of any arm is 0 until it starts offering the item, it follows from the mean-based condition that the buyer will pick a specific arm that is not offering the item with probability at most γ, and therefore choose some good arm with probability at least $1 - m\gamma$. It follows that, in expectation, the buyer with value v_i receives the item for at least $(1-m\gamma)(x_iT - \gamma T/\varepsilon) = x_iT - o(T)$ turns.

We now proceed to upper bound the overall expected utility of the buyer. For each index $j \leq i$, let S_j be the set of t where $\sigma_{j,t}(v_i) > \sigma_{j',t}(v_i)$ for all other j'. Note that since each $\sigma_{j,t}(v_i)$ is a linear function in t (when positive), each S_j is either the empty set or an interval (y_jT, z_jT). Since all the v_i are distinct, note that these intervals partition the interval $((1-x_i)T, T)$ (with the exception of up to m endpoints of these intervals); in particular, $\sum_{j \geq i}(z_j - y_j) = x_i$.

Let $\varepsilon' = \min_j(v_{j+1} - v_j)$. Note that, if $t \in (y_jT + \gamma T/\varepsilon', z_jT - \gamma T/\varepsilon')$, then for all $j' \neq j$, $\sigma_{j,t}(v_i) > \sigma_{j',t}(v_i) + \gamma T$. This follows since $\sigma_{j,t}(v_i) - \sigma_{j',t}(v_i)$ is linear in t with slope $v_j - v_{j'}$, and $|v_j - v_{j'}| > \varepsilon'$. It follows that if t is in this interval, then the buyer will choose option j with probability at least $1 - m\gamma$ (by a similar argument as before).

Define $j(t) = \arg\max_j \sigma_{j,t}(v_i)$ to be the index of the arm with the current largest cumulative reward, and let $\sigma_{max,t}(v_i) = \sum_{s=1}^{t} r_{j(s),s}(v_i)$ be the cumulative utility of always playing the arm with the current highest cumulative reward for the first t rounds. The following lemma shows that $\sigma_{max,T}(v_i)$ is close to $\max_j \sigma_{j,T}(v_i)$. (In other words, playing the best arm every round and playing the best-at-the-end arm every round have similar payoffs if the historically best arm does not change often).

Lemma 3.7. *$|\sigma_{max,T}(v_i) - \max_j \sigma_{j,T}(v_i)| \leq m$.*

Proof. Let $W = |\{t | j(t) \neq j(t+1)\}|$ equal the number of times the best arm switches values; note that since each $\sigma_{j,t}(v_i)$ is linear, W is at most m. Let $t_1 < t_2 < \cdots < t_W$ be the values of t such that $j(t) \neq j(t+1)$. Additionally define $t_0 = 1$ and $t_{W+1} = T$. Then, dividing the cumulative reward $\sigma_{max,t}$ into intervals by these t_i, we get that

$$\begin{aligned}
\sigma_{max,t}(v_i) &= \sum_{s=1}^{t} r_{j(s),s}(v_i) \\
&= \sum_{i=1}^{W+1} (\sigma_{j(t_i),t_i}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i)) \\
&= \sigma_{j(T),T}(v_i) + \sum_{i=1}^{W+1} (\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i)) \\
&= \max_j \sigma_{j,t}(v_i) + \sum_{i=1}^{W+1} (\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i))
\end{aligned}$$

It therefore suffices to show that $|\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i)| \leq 1$ for all i. To see this, note that (by the definition of $j(t)$), $\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i) > 0$, and that $\sigma_{j(t_{i-1}),t_{i-1}+1}(v_i) - \sigma_{j(t_i),t_{i-1}+1}(v_i) < 0$. However,

$$(\sigma_{j(t_{i-1}),t_{i-1}+1}(v_i) - \sigma_{j(t_i),t_{i-1}+1}(v_i)) = (\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i)) + (r_{j(t_{i-1}),t_{i-1}+1}(v_i) - r_{j(t_i),t_{i-1}+1}(v_i))$$

Since $0 \leq r_{j,t}(u) \leq 1$, it follows that $|\sigma_{j(t_{i-1}),t_{i-1}}(v_i) - \sigma_{j(t_i),t_{i-1}}(v_i)| \leq 1$. This completes the proof. □

Let $\sigma_T(v_i) = \sum_{t=1}^{T} \mathbb{E}[r_{I_t,t}(v_i)]$ denote the expected cumulative utility of this buyer at time T. We claim that $\sigma_T \leq \max_j \sigma_{j,T}(v_i) + o(T)$. To see this, recall that, for $t \in (y_jT + \gamma T/\varepsilon', z_jT - \gamma T/\varepsilon')$, $\Pr[I_t \neq j] \leq m\gamma$, and therefore $\mathbb{E}[r_{I_t,t}] \leq r_{j,t} + m\gamma$. Furthermore, note that for $t \in S_j$, $j(t) = j$, so $r_{j,t} = r_{j(t),t}$ and $\mathbb{E}[r_{I_t,t}] \leq r_{j(t),t} + m\gamma$. It follows that

$$\begin{aligned}
\sigma_T(v_i) &= \sum_{t=1}^{T} \mathbb{E}[r_{I_t,t}(v_i)] \\
&\leq \sum_{t=(1-x_i)T}^{T} \mathbb{E}[r_{I_t,t}(v_i)] \\
&= \sum_{j=1}^{i} \sum_{t=y_jT}^{z_jT} \mathbb{E}[r_{I_t,t}(v_i)] \\
&\leq \sum_{j=1}^{i} \left(\frac{2\gamma T}{\varepsilon'} + \sum_{t=y_jT+\gamma T/\varepsilon'}^{z_jT-\gamma T/\varepsilon'} \mathbb{E}[r_{I_t,t}(v_i)] \right) \\
&\leq \sum_{j=1}^{i} \left(\frac{2\gamma T}{\varepsilon'} + \sum_{t=y_jT+\gamma T/\varepsilon'}^{z_jT-\gamma T/\varepsilon'} (r_{j(t),t}(v_i) + m\gamma) \right) \\
&\leq \frac{2m\gamma T}{\varepsilon'} + m\gamma T + \sum_{t=1}^{T} r_{j(t),t}(v_i) \\
&= \frac{2m\gamma T}{\varepsilon'} + m\gamma T + \sigma_{max,T}(v_i) \\
&\leq \frac{2m\gamma T}{\varepsilon'} + m\gamma T + m + \max_j \sigma_{j,T}(v_i) \\
&= \max_j \sigma_{j,T}(v_i) + o(T).
\end{aligned}$$

Finally, note that

$$\begin{aligned}
\max_j \sigma_{j,T}(v_i) &= \max_{j<i} (v_i - v_j + \varepsilon) x_j T \\
&\leq (\max_{j<i} (v_i - v_j) x_j + \varepsilon) T \\
&= (u_i + \varepsilon) T
\end{aligned}$$

It follows that $\sigma_T(v_i) \leq (u_i + \varepsilon)T + o(T)$, as desired.

□

Finally, we show that this quantity $\mathsf{MBRev}(\mathcal{D})$ is in fact significantly different from both $\mathsf{Val}(\mathcal{D})$ and $\mathsf{Rev}(\mathcal{D})$; in particular, it is a constant-factor approximation to neither. In particular, the multiplicative gap between $\mathsf{MBRev}(\mathcal{D})$ and $\mathsf{Rev}(\mathcal{D})$ can grow as large as $\log\log H$ for distributions $\mathcal{D}$ supported on $[1, H]$. In comparison, the gap between $\mathsf{Val}(\mathcal{D})$ and $\mathsf{Rev}(\mathcal{D})$ can grow as large as $\log H$ on this same interval, and in fact both gaps are maximized for the same distribution: the equal-revenue curve $\mathcal{D}_{ERC}$ truncated at H.

Theorem 3.8. *For distributions $\mathcal{D}$ supported on $[1, H]$, $\mathsf{MBRev}(\mathcal{D}) = O(\log\log H)$, and there exist $\mathcal{D}$ supported on $[1, H]$ such that $\mathsf{MBRev}(\mathcal{D}) = \Theta(\log\log H)$. For this same $\mathcal{D}$, $\mathsf{Val}(\mathcal{D}) = \Theta(\log H)$.*

The proof of Theorem 3.8 is included in the full version of this paper. The proof is divided into two parts (after extending the definition of $\mathsf{MBRev}(\mathcal{D})$ to hold for continuous distributions $\mathcal{D}$): 1. showing that $\mathsf{MBRev}(\mathcal{D}_{ERC}) \leq O(\log\log H)$, and 2. showing that $\mathsf{MBRev}(\mathcal{D}_{ERC}) \geq O(\log\log H)$.

To show the first part, it suffices to simply demonstrate a solution to the mean-based LP with value at least $O(\log\log H)$. It suffices to choose $x(v) = \frac{\log v}{\log H}$ (equivalently, the reserve for the associated second-price auction should exponentially decay over time).

To show the second part, we examine the dual of the LP. Effectively, this involves rewriting $\mathsf{MBRev}(\mathcal{D})$ in the form

$$\mathsf{MBRev}(\mathcal{D}) = \max_x \mathbb{E}_{v_i \sim \mathcal{D}} \left[v_i x_i - \max_j (v_i - v_j) x_j \right]$$

(in particular, note that for a fixed choice of x, $u_j = \max_j (v_i - v_j) x_j$), and finding an appropriate function $j(i)$ (which corresponds to an assignment to the dual).

3.4 A Final Note on the Example

While reading through our examples, the reader may think that the mean-based learner's behavior is clearly irrational: why would you continue paying above your value? Why would you continue paying more than necessary, when you can safely get the item for less?

But this is exactly the point: a more thoughtful learner can indeed do better (for instance, by using the algorithm of Section 3.2). It is also perhaps misleading to believe that the bidder should "obviously" stop overpaying: we only know this because we know the structure of the example. But in principle, how is the bidder supposed to know that the overcharged rounds are the new norm and not an anomaly? Given that most standard no-regret algorithms are mean-based, it's important to nail down the seller's options for exploiting this behavior.

4 CONCLUSION AND FUTURE DIRECTIONS

We consider a revenue-maximizing seller with a single item (each round) to sell to a single buyer. We show that when the buyer uses mean-based algorithms like EXP3, the seller can extract revenue equal to the expected welfare with an unnatural auction. We then provide a modified no-regret algorithm $\mathcal{A}$ such that the seller cannot extract revenue exceeding the monopoly revenue when the buyer bids according to $\mathcal{A}$. Finally, we consider a mean-based buyer who never overbids. We tightly characterize the seller's optimal revenue with a linear program, and show that a pay-your-bid auction with decreasing reserves over time achieves this guarantee. Moreover, we show that the mean-based revenue can be unboundedly better than the monopoly revenue while simultaneously worse than the expected welfare. In particular, for the equal revenue curve truncated at H, the monopoly revenue is 1, the expected welfare is $\ln(H)$, and the mean-based revenue is $\Theta(\ln(\ln(H)))$.

While our work has already shown the single-buyer problem is quite interesting, the most natural direction for future work is understanding revenue maximization with multiple learning buyers. Of our three main results, only Theorem 3.2 extends easily (that if every buyer uses our modified learning, the default strategy, which now runs Myerson's optimal auction every round, is optimal). Our work certainly provides good insight into the multi-bidder problem, but there are still clear barriers. For example, in order to obtain revenue equal to the expected welfare, the auction must necessarily also maximize welfare. In our single-bidder model, this means that we can give away the item for free for $\Omega(T)$ rounds, but with multiple bidders, such careless behaviour would immediately make it impossible to achieve the optimal welfare. Regarding the mean-based revenue, while there is a natural generalization of our LP to multiple bidders, it's no longer clear how to achieve this revenue against conservative bidders, as all the relevant variables now implicitly depend on the actions of the other bidders. These are just examples of concrete barriers, and there are likely interesting conceptual barriers for this extension as well.

Another interesting direction is understanding the consequences of our work from the perspective of the buyer. Aside from certain corner configurations (e.g. the seller extracting the buyer's full welfare), it's not obvious how the buyer's utility changes. For instance, is it possible that the buyer's utility actually *increases* as the seller switches from the default strategy to the optimal mean-based revenue? Does the buyer ever benefit from using an "exploitable" learning strategy, so that the seller can exploit it and make them both happier?

REFERENCES

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. 2012. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing* 8, 6 (2012), 121–164. DOI: http://dx.doi.org/10.4086/toc.2012.v008a006

[2] Itai Ashlagi, Constantinos Daskalakis, and Nima Haghpanah. 2016. Sequential Mechanisms with Ex-post Participation Guarantees. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*. 213–214. DOI: http://dx.doi.org/10.1145/2940716.2940775

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2003. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* 32, 1 (Jan. 2003), 48–77. DOI: http://dx.doi.org/10.1137/S0097539701398375

[4] Avrim Blum and Yishay Mansour. 2007. From External to Internal Regret. *Journal of Machine Learning Research* 8 (2007), 1307–1324. http://dl.acm.org/citation.cfm?id=1314543

[5] Sébastien Bubeck and Nicolò Cesa-Bianchi. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5, 1 (2012), 1–122. DOI: http://dx.doi.org/10.1561/2200000024

[6] Yang Cai and Constantinos Daskalakis. 2017. Learning Multi-item Auctions with (or without) Samples. In *FOCS*.

[7] Edward H. Clarke. 1971. Multipart Pricing of Public Goods. *Public Choice* 11, 1 (1971), 17–33.

[8] 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 (STOC '14)*. ACM, New York, NY, USA, 243–252. DOI: http://dx.doi.org/10.1145/2591796.2591867

[9] Constantinos Daskalakis and Vasilis Syrgkanis. 2016. Learning in Auctions: Regret is Hard, Envy is Easy. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. 219–228. DOI: http://dx.doi.org/10.1109/FOCS.2016.31

[10] Constantinos Daskalakis and S. Matthew Weinberg. 2012. Symmetries and Optimal Multi-Dimensional Mechanism Design. In *the 13th ACM Conference on Electronic Commerce (EC)*.

[11] Nikhil R. Devanur, Zhiyi Huang, and Christos-Alexandros Psomas. 2016. The Sample Complexity of Auctions with Side Information. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing (STOC '16)*. ACM, New York, NY, USA, 426–439. DOI: http://dx.doi.org/10.1145/2897518.2897553

[12] Nikhil R. Devanur, Yuval Peres, and Balasubramanian Sivan. 2015. Perfect Bayesian Equilibria in Repeated Sales. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '15)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 983–1002. http://dl.acm.org/citation.cfm?id=2722129.2722196

[13] Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E. Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. 2017. Oracle-Efficient Learning and Auction Design. In *FOCS*.

[14] Yannai A. Gonczarowski and Noam Nisan. 2017. Efficient Empirical Revenue Maximization in Single-parameter Auction Environments. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2017)*. ACM, New York, NY, USA, 856–868. DOI: http://dx.doi.org/10.1145/3055399.3055427

[15] Theodore Groves. 1973. Incentives in Teams. *Econometrica* 41, 4 (1973), 617–631.

[16] James Hannan. 1957. Approximation to bayes risk in repeated play. In *Contributions to the Theory of Games*. 3:97–139.

[17] Nicole Immorlica, Brendan Lucier, Emmanouil Pountourakis, and Samuel Taggart. 2017. Repeated Sales with Multiple Strategic Buyers. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 167–168.

[18] Adam Kalai and Santosh Vempala. 2002. Geometric Algorithms for Online Optimization. In *Journal of Computer and System Sciences*. 26–40.

[19] Adam Kalai and Santosh Vempala. 2005. Efficient Algorithms for Online Decision Problems. *J. Comput. Syst. Sci.* 71, 3 (Oct. 2005), 291–307. DOI: http://dx.doi.org/10.1016/j.jcss.2004.10.016

[20] John Langford and Tong Zhang. 2008. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.). Curran Associates, Inc., 817–824. http://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information.pdf

[21] Siqi Liu and Christos-Alexandros Psomas. 2017. On the Competition Complexity of Dynamic Mechanism Design. *CoRR* abs/1709.07955 (2017). arXiv:1709.07955 http://arxiv.org/abs/1709.07955

[22] Vahab S. Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. 2016. Dynamic Auctions with Bank Accounts. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. 387–393. http://www.ijcai.org/Abstract/16/062

[23] Vahab S. Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. 2016. Optimal dynamic mechanisms with ex-post IR via bank accounts. *CoRR* abs/1605.08840 (2016). arXiv:1605.08840 http://arxiv.org/abs/1605.08840

[24] Jamie Morgenstern and Tim Roughgarden. 2015. The Pseudo-dimension of Near-optimal Auctions. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 136–144. http://dl.acm.org/citation.cfm?id=2969239.2969255

[25] Jamie Morgenstern and Tim Roughgarden. 2016. Learning Simple Auctions. In *29th Annual Conference on Learning Theory (Proceedings of Machine Learning Research)*, Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (Eds.), Vol. 49. PMLR, Columbia University, New York, New York, USA, 1298–1318. http://proceedings.mlr.press/v49/morgenstern16.html

[26] Roger B. Myerson. 1981. Optimal Auction Design. *Mathematics of Operations Research* 6, 1 (1981), 58–73.

[27] Denis Nekipelov, Vasilis Syrgkanis, and Eva Tardos. 2015. Econometrics for Learning Agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC '15)*. ACM, New York, NY, USA, 1–18. DOI: http://dx.doi.org/10.1145/2764468.2764522

[28] Christos Papadimitriou, George Pierrakos, Christos-Alexandros Psomas, and Aviad Rubinstein. 2016. On the Complexity of Dynamic Mechanism Design. In *Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '16)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1458–1475. http://dl.acm.org/citation.cfm?id=2884435.2884535

[29] Tim Roughgarden. 2012. The Price of Anarchy in Games of Incomplete Information. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 862–879. DOI: http://dx.doi.org/10.1145/2229012.2229078

[30] Vasilis Syrgkanis and Eva Tardos. 2013. Composable and Efficient Mechanisms. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing (STOC '13)*. ACM, New York, NY, USA, 211–220. DOI: http://dx.doi.org/10.1145/2488608.2488635

[31] William Vickrey. 1961. Counterspeculations, Auctions, and Competitive Sealed Tenders. *Journal of Finance* 16, 1 (1961), 8–37.