Exploiting mean-based bidders in symmetric settings

Matteo Russo

Independent work in partial fulfillment of
the requirements for the certificate in the
Program in Applied and Computational Mathematics
PACM

Advisor: Professor Mark Braverman

Department of Computer Science
Princeton University
Princeton, New Jersey, USA

April 30, 2020

Abstract

With the advent of the Internet and of powerful search engines such as Google, private businesses, who want their ads to be the first to be displayed, would also like to comprehend how to bid effectively whilst automatizing the process. Notably, during the past two decades, the paradigm of Machine Learning has blindly and ubiquitously spread throughout a vast variety of application-specific domains: one of those is indeed automatic bidding in Sponsored-Search Auctions. In recent years, the problem of repeatedly selling an item to a single-buyer has been considered. Hereby, a strategic seller prices an item dynamically so to extract the whole utility from a buyer that plays most historically-rewarding actions. In this paper, we extend this framework by substituting the contrasting seller and buyer with two bidders who have to bid for a single item repeatedly while: we denote it by the term Welfare Redistribution Auction (WRA for short). At each time step, the mean-based bidder as well as the strategic one draw independently valuations f *and* v *respectively from two distributions* $\mathcal{F}$ *and* $\mathcal{D}$ *which have both support over the set of possible actions* $\mathcal{B} = \{\mathsf{b}_1, \ldots, \mathsf{b}_K\}$*. They, therefore, bid for the item. The winner gets the item and pays the adversary some amount. We formulate the conjecture of whether there exists an algorithm for which the strategic bidder may extract the whole surplus from the mean-based bidder up to a sublinear factor in number of rounds. We, thus, test, in a variety of WRA settings, whether Reinforcement Learning agents such as Deep Q-Learning agents perform well against mean-based ones. The empirical evidence demonstrates how even very blazoned algorithms are not potential candidates for the conjecture strategy to exist, thus, opening future lines of research.*

1 Introduction

1.1 Motivation and Goal

The employment of learning algorithms in a Repeated Auctions setting has been of the uttermost importance since the huge increase in popularity of search engines like Google or Microsoft's Bing. The reason why auctions are so fundamental for both search engines and single websites owners is that advertisement on these platforms is performed by auctions: fixed a relevance score, single owners bid against each other to aspire to the first, second, etc. place on the search engine result list. Over the years, bidding algorithms have become progressively more and more refined; still, this has not brought parity or equality in the capability of bidding and being the first to be sponsored. This is mainly due to two reasons: the first one being that the difference in liquidity amount between companies may be abyssal and the market power of the richer companies is surely a crucial factor in the advantage. Unlike one may think, this is not the only reason why this can occur. Too many times, in fact, bidders assume that other bidders are not strategic and are oblivious of the adversarial behaviour other bidders may have. In other words, one naïve given bidder may think of employing an optimization algorithm that, based on a private valuation of the item (in this case the place on the search results list) and the observed frequency distribution over other bidders' actions (in this case bids) at the current time step, outputs the most lucrative action for the next time step. However, a smarter bidder, who knows his or her opponent(s) play according to such kind of shortsighted algorithm, should make this kind of bidder believe a strategically crafted frequency distribution over bids for some time and, later, use this false belief state to extract as much reward from the opponent as possible: we will call this type of player 1-layer strategic bidder. The dynamics we have described may be so detrimental to non-strategic bidders and so beneficial to strategic ones, that the latter category may exploit the former in a very serious way, making themselves gain much more than they should due to strategic attitudes. Nevertheless, one could naturally be interested in comprehending if a 2-layer strategic bidder may be able to exploit a 1-layer one through similar strategic considerations and whether, inductively, a k-layer strategic bidder may be able to exploit a $(k-1)$-layer one. Hence, **the goal of this independent work is twofold: on one hand, it aims to demonstrate provable guarantees on revenue and regret for what regards the exploitation from a strategic bidder to a non-strategic one. On the other hand, we demonstrate with strong empirical evidence that even more modern and blazoned approaches like Deep Q-Learning could be easily exploited.**

1.2 Overview

During the second half of the 20th century, both from a theoretical standpoint and from a commercial one, auctions have been of great interest to the economic, mathematical and computer science world. In the most general sense, auctions can be seen as markets with multiple items sold by multiple sellers to multiple buyers (or bidders) with the following four assumptions [10]:

1. All of the bidders are risk-neutral.
2. Each bidder has a private valuation for the item independently drawn from some probability distribution.
3. The bidders possess symmetric information.
4. The payment is represented as a function of only the bids.

Canonically, however, the literature has focused mainly on single seller auctions as well as single seller (or auctioneer) and single-item ones, as they already show a certain degree of complexity. Ideal auctions or mechanisms may be seen as those designed to possess (informally) the following four properties.

1. Ex-ante individual rationality (**IR**): for all parties involved in the mechanism, their expected payoff should be non-negative, so that they all have an initial incentive to

participate in the trade.

2. Weak balanced budget (**WBB**): the trade should not be subsidized by the auctioneer.

3. Nash equilibrium incentive compatibility (**NEIC**): if all other bidders report their true value for this mechanism, then, the given bidder is best off by reporting her true value (this has to hold for all bidders).

4. Ex-post Pareto efficiency (**PE**): the auction should be concluded with the item assigned to the bidder who has valued it the most.

In this setting, the goal of the auctioneer is to assign a price (possibly the maximum one) to the item so that the four properties are satisfied. Nevertheless, Myerson-Satterthwaite's theorem [11] establishes that there exists no mechanism satisfying **IR**, **WBB**, **NEIC** and **PE**. Let us now consider a bidder that wants to adopt her optimal bidding strategy in an auction. If the auction, like the Vickrey-Clarke-Groves auction [6, 8, 13], satisfies **NEIC**, then her decision is simply to bid truthfully according to her value. Otherwise, as for instance in a Generalized First-Price (GFP) or Generalized Second-Price (GSP) auction, the bidder's optimal bidding strategy would not be as simple and each bidder would try to compute the bid associated to a Bayes-Nash Equilibrium, which is computationally too expensive and, thus, unrealistic, if we acknowledge the fact that most auctions are today carried out by artificial agents on the web. An alternative to the *a priori* static setting is that of a learning bidder and, specifically, a low-external regret bidder [2, 3] or low-policy regret bidder [1] trying to progressively approach a Coarse-Correlated Equilibrium or a Policy Equilibrium respectively. Even in these settings, however, the strategy space might be too huge to be explored in its entirety. And to make matters worse, Braverman et al. in [4] showed that a low-external regret bidder can be hugely exploited by a strategic seller or another strategic bidder who knows that the former is playing according to an LER algorithm like `MWU` or `EXP3`.

1.3 Summary of Approach, Implementation and Results

The paper can be roughly divided in three parts, which are delineated below.

Part I. We explain the repeated single-item auction where, for each time step, a mean-based buyer (defined in the next sessions) and a strategic seller perform a trade of an item and, in particular, where the mean-based buyer learns over time which bid is best performing and where the strategic seller prices item to extract the whole surplus from the buyer.

Part II. We, therefore, extend the framework to what we name a Welfare Redistribution Auction (WRA). Hereby, a mean-based bidder and a strategic one compete for the allocation of an item at each time step and split the remaining amount among themselves. Once again, the mean-based bidder learns over time which bid is best performing and where the strategic bidder bids to extract the whole surplus from the game.

Part III. In the last part, we run some experiments to see how different kinds of algorithms perform in the WRA setting and we come across important considerations such as the one about the fact that `DQN` is not at all more powerful than a simple `EXP3` and might be exploited as easily.

2 Problem Background and Related Work

Despite being a relatively recent problem to be addressed, in the past few years, algorithmic learning in auctions has become a focus of attention for both the Algorithmic Game Theory community as well as for the Theoretical Machine Learning and Stochastic Control community. As a matter of fact, the most useful related setting to introduce in order for us to truly comprehend the problem at hand is that of the so-called Multi-Armed Bandit Problem (MABP) [12] which is a very well-known and well-studied problem in Probability Theory. It could be informally described as follows: suppose a gambler (or principal) $\mathcal{M}$ goes into a Las Vegas Casino and sees n slot machines one next to the other to choose from at each of the finite number of rounds. Each machine will output, if he or she plays machine i at round t, a reward $\rho_{it}(\mathcal{M}) \sim \mathcal{D}_i$, where, for all i, distribution $\mathcal{D}_i$ is unknown to the gambler. His or her goal is, thus, to maximize the total reward, as in the sum of all rewards received at each round over the T rounds. This problem is a classic instance of the more general "Exploration-Exploitation Dilemma", where the principal has to decide whether to "exploit" a subset of slot machines that has previously rewarded well or "explore" new subsets of slot machines with the opportunity to gain larger rewards and risk of receiving smaller ones. Relaxing the assumption that arms draw their valuations stochastically and have to truthfully declare those, the Strategic MABP setting arises. Hereby, if the principal (or auctioneer) plays machine (or bidder) i at round t, a reward $\rho_{it}(\mathcal{M}) - x_{it}$ will be output, where x_{it} is strategically chosen by arm corresponding to action $b_j \in \mathcal{B}$ at each round $t \in [T]$ and no other arm can visualize this value (tacit scenario). As mentioned, Braverman et al. in [4] showed that if $\mathcal{M}$ is a low-external regret algorithm used for the stochastic version of MABP, then there exists an MABP instance such that in a $o(T)$-Nash Equilibrium, the principal is rewarded at most $o(T)$ (in the tacit scenario). They also showed that, if $\mu_i = \mathbb{E}[\rho_i]$ and μ' is the second largest of those expectations, then, there exists a truthful mechanism such that the principal is guaranteed an expected reward of $\mu' T - o(T)$. Here, not only does the principal face the "Exploration-Exploitation Dilemma" but also the arms need to trade-off between whether to keep high $x_i t$'s, hence, inducing the principal not to pull them often over time but keeping a substantial amount each time they do get pulled, or to declare high $\rho_{it}(\mathcal{M}) - x_{it}$'s, inducing the principal to pull them often over time but keeping a small amount each time they do get pulled. Braverman et al. in [5] prove that what the auctioneer may be able to extract from bidders is substantial in the case where agents participating to an auction bid according to a low-external regret learning algorithms such as `EXP3`. Furthermore, lines of work such as [7] or [9], the authors devise explicit algorithms to exploit `EXP3`-like players and are quite successful in extracting revenue from them. Arora et al. in [1] explain the three main adversarial scenarios in learning equilibria setting: the first is the oblivious setting where the opposing bidder to a given bidder does not react to the given bidder actions and where payoffs can be thought as predetermined. The second is the adaptive setting where the opposing bidder to a given bidder does react to the given bidder actions and seeks the maximization of her own payoff. The third kind of scenario is the m-memory bounded one where the opposing bidder to a given bidder does react to the given bidder actions but only being able to store the adversary's past m actions and, of course, still seeking to maximize her own payoff. We may observe how an oblivious adversary is equivalent to a 0-memory bounded one and an adaptive adversary is equivalent to a ∞-memory bounded one. As thoroughly explained in the paper, the notion of external regret captures neither the adaptive nor the m-memory bounded scenario insofar as it does not express how a given bidder would react had the opposing bidder chosen another bid. Thus, we need a new formulation of the concept of regret, which Arora et al. define as m-memory Policy Regret, for $m \in \mathbb{N}$. This notion of regret captures the possibility that the opposing bidder deviates herself when the given bidder deviates in the first place.

3 Approach

The following section is organized as follows:

- In subsection 3.1, we outline the game theoretical framework we are operating in, formally defining the notion of strategy space, utility as well as regret.
- In subsection 3.2, we describe the `EXP3` algorithm and summarize the provable guarantees on regret and utility this algorithm allows.
- In subsection 3.3, we introduce the auction model named Welfare Redistribution Auction (WRA for short), which generalizes the notion of selling to a no-regret buyer as per Braverman et al. in [5].
- In subsection 3.4, we formulate the extended problem formally and explain the reasons why this problem looks much harder than its predecessor.

3.1 Framework

Given the problem described below (which is between two players), we will outline the theoretical framework denoting one player as bidder $\mathsf{B}_\triangle$ and the other as bidder $\mathsf{B}_\square$.

Definition. *Let us use the following notation:*

- $\mathcal{S}^\mathsf{T}$ *is the set of strategies available to* $\mathsf{B}_\triangle$ *and* $\mathsf{B}_\square$*, given that the bids come from the same set* $\mathcal{B}$ *for both bidders.*
- $\Delta\mathcal{S}^\mathsf{T}$ *is the simplex of mixed strategies over* $\mathcal{S}^\mathsf{T}$*. This means that each bidder at each round may choose one of the* K *bids available according to some probability distribution over bids.*

We thus see how the number of pure strategies she has is $|\mathcal{S}^T| = K^T$, which requires a bit-representation of $T \log K$ which is much too intractable for all practical purposes, let alone the (uncountable) number of mixed strategies. Moreover, let us use the following definition of utility.

Definition. *Let* $\sigma_i(t) \in \mathbb{R}^n$ *represents a probability vector, which means* $\forall\ t \in [T],\ \|\sigma_i(t)\| = 1$ *and* $\forall\ j \in [n],\ \sigma_{ij}(t) \geq 0$*;* $\pi_i(t) \in \mathbb{R}^n$ *represents the vector encoding all pure payoffs resulting from the auction above for player* $i \in \{0, 1\}$ *at time* $t \in [T]$*. Then, let us define the utility from a mixed strategy at time step* t *as follows:*

$$u_i\left(\sigma_i, \pi_i, t\right) = \sigma_i(t)^\intercal \pi_i(t)$$

In order to represent the utility of a sequence of actions $\{b(1), \ldots, b(t)\} \in \Delta\mathcal{S}^t$ *until time step* $t \in [T]$ *for bidder* $i \in \{0, 1\}$*, we will write* $u_i\left(b(1), \ldots, b(t)\right)$*.*

Having defined the above notions, we are almost ready to define the concept of regret which, together with the concept of utility, will be central for our analysis. Let us, thus, consider an algorithm $\mathcal{A}$ that in t steps of a given game has to choose on a sequence of actions $a_1, \ldots, a_t$ with corresponding payoffs $\pi_{a_1}(1), \ldots, \pi_{a_t}(t)$.

Definition. *We can, hence, define the sum of these obtained payoffs as follows*

$$\mathcal{G}_{\mathcal{A},t} := \sum_{s \in [t]} \pi_{a_s}(s)$$

At the same time, we need to define what the payoff for a different course of action would have been. That means that fixed an arbitrary sequence of actions $\mathbf{\Psi} = (\psi_1, ..., \psi_t)$*, we could* *analogously define*

$$\mathcal{G}_{\psi,t} := \sum_{s \in [t]} \pi_{\psi_s}(s)$$

As we may guess, we would like to find an algorithm $\mathcal{A}$ such that the difference between the maximum total payoff for a special sequence of actions $\boldsymbol{\psi}^*$ and its expected sum of payoffs is minimal. This way, looking backwards, the algorithm has minimized the amount of payoff it has wasted by not playing another sequence of actions. This leads us to the following crucial definition.

Definition. *We define weak regret the following quantity*

$$\mathcal{R}_{\mathcal{A}} := \left(\max_{\psi \in \mathcal{S}^T} \sum_{t \in [T]} \pi_{\psi}(t) \right) - \mathbb{E}\left[\mathcal{G}_{\mathcal{A}}\right]$$

Whereby, $\mathcal{G}_{\mathcal{A}} := \mathcal{G}_{\mathcal{A},T}$.

Next, we will introduce the `EXP3` algorithm, a learning procedure for the MABP problem which guarantees sublinear regret.

3.2 EXP3 and mean-based algorithms

The `EXP3` algorithm, as per [3], has been designed to tackle the problem of selecting actions in the adversarial bandit setting with a guarantee on the sublinearity of its regret at the end of the time period $[T]$. Below, we will first introduce the adversarial bandit setting, then, describe the algorithm itself, and, finally, prove some useful bounds on the regret `EXP3` achieves.

Definition. *Let us define an adversarial bandit problem as a pair* (K, π)*, whereby* K *denotes the number of actions (i.e.* $|\mathcal{B}| = K$*), and* π *represents the sequence of payoffs received in the game as outlined in the previous section, that is* $\pi_{a_1}(1), \dots, \pi_{a_t}(t)$*.*

We are, thus, ready to introduce the EXP3 algorithm, along with its psudocode, we give a brief explanation of what it does and we delineate the most important regret guarantees.

```
EXP3 algorithm
$\eta \in [0,1]$, $\boldsymbol{w} \leftarrow \mathbf{1}$ with $\boldsymbol{w} \in \mathbb{R}^K$
FOR $s \in [T-1]$:
    SET: $\boldsymbol{\alpha}_y(s) \leftarrow (1-\eta)\frac{\boldsymbol{w}_y(s)}{||\boldsymbol{w}(s)||} + \frac{\eta}{K}$, $\forall\, y \in [K]$
    DRAW: $y(s) \sim \boldsymbol{\alpha}(s)$
    OBSERVE: reward $\boldsymbol{\pi}_y(s)$ and let $\widetilde{\boldsymbol{\pi}}_y(s) = \frac{\boldsymbol{\pi}_y(s)}{\boldsymbol{\alpha}_y(s)} \cdot \mathbb{1}_{\{y(s)=x\}}$ for $x \in [K]$
    UPDATE: $\boldsymbol{w}_y(s+1) \leftarrow \boldsymbol{w}_y(s) \cdot \exp\left(\eta \cdot \widetilde{\boldsymbol{\pi}}_y(s)\right)$
RETURN $\boldsymbol{\alpha}(T)$
```

In EXP3, the variable η is a parameter which tunes exploration in the sense that the closer it is to 1, the more EXP3 will pick an action uniformly at random (in case when $\eta = 1$, EXP3 is simply an equiprobable randomized algorithm over actions). The peculiarity of EXP3 is that, after having initialized a weight parameter over each action to 1, it progressively updates the probability of playing a specific action by observing the reward it gets from playing that very action in an online fashion. This algorithm performs not too much worse than the best of the K experts in hindsight, implying that the weak regret is always sublinear. Nevertheless, this does not necessarily imply that the overall utility of EXP3 will be high, especially in the presence of a strategic adversary who knows how to properly exploit EXP3, as shown in [5].

Theorem. *For any* $\mathsf{K} > 0$*,* $\eta \in (0, 1]$ *and any stopping time* $\mathsf{T} \in \mathbb{N}$*, an* ***EXP3*** *algorithm with* $\eta = \sqrt{\frac{\log \mathsf{K}}{8\mathsf{TK}}} \in o(1)$*,*

$$\mathcal{R}_{\textit{EXP3}} \leq 2\sqrt{\mathsf{TK} \log \mathsf{K}} \in o(\mathsf{TK})$$

Proof.

Let $\boldsymbol{W}(s) = \sum_{y \in [K]} \boldsymbol{w}_{\mathbf{y}}(s)$ and note that

$$\mathbb{E}_{\boldsymbol{\alpha}(s)} \left[\widetilde{\boldsymbol{\pi}}_x(s)\right] = \sum_{k \in [K]} \boldsymbol{\alpha}_k(s) \cdot \frac{\boldsymbol{\pi}_x(s)}{\boldsymbol{\alpha}_x(s)} \cdot \mathbb{1}_{\{k=x\}} = \boldsymbol{\pi}_x(s)$$
$$\mathbb{E}_{\boldsymbol{\alpha}(s)} \left[\widetilde{\boldsymbol{\pi}}_{y(s)}(s)\right] = \sum_{k \in [K]} \boldsymbol{\alpha}_{y(s)}(s) \cdot \frac{\boldsymbol{\pi}_{y(s)}(s)}{\boldsymbol{\alpha}_{y(s)}(s)} \leq K$$

Hence,

$$\begin{aligned} \frac{\boldsymbol{W}(s+1)}{\boldsymbol{W}(s)} &= \sum_{k \in [K]} \frac{\boldsymbol{w}_k(s+1) \cdot \exp\left(\eta \cdot \widetilde{\boldsymbol{\pi}}_k(s)\right)}{\boldsymbol{W}(s)} = \sum_{k \in [K]} \frac{\boldsymbol{\alpha}_k(s) - \frac{\eta}{K}}{1 - \eta} \cdot \exp\left(\eta \cdot \widetilde{\boldsymbol{\pi}}_k(s)\right) \\ &\leq \sum_{k \in [K]} \frac{\boldsymbol{\alpha}_k(s) - \frac{\eta}{K}}{1 - \eta} \cdot \Big(1 + \eta \cdot \widetilde{\boldsymbol{\pi}}_k(s) + (e-2) \cdot \eta^2 \cdot \widetilde{\boldsymbol{\pi}}_k^2(s)\Big) \\ &\leq 1 + \frac{1}{1-\eta} \sum_{k \in [K]} \boldsymbol{\alpha}_k(s) \cdot \Big(1 + \eta \cdot \widetilde{\boldsymbol{\pi}}_k(s) + (e-2) \cdot \eta^2 \cdot \widetilde{\boldsymbol{\pi}}_k^2(s)\Big) \end{aligned}$$

The second to last inequality follows from the fact that $e^x \leq 1 + x + (e-2) \cdot x^2$ for any $x \leq 1$. Thus, we could the log of both sides since the log-function is monotone:

$$\log \frac{\boldsymbol{W}_{s+1}}{\boldsymbol{W}_s} \leq \frac{1}{1-\eta} \sum_{k \in [K]} \boldsymbol{\alpha}_k(s) \cdot \Big(1 + \eta \cdot \widetilde{\boldsymbol{\pi}}_k(s) + (e-2) \cdot \eta^2 \cdot \widetilde{\boldsymbol{\pi}}_k^2(s)\Big)$$

Therefore, we have

$$\log \frac{\boldsymbol{W}_{T+1}}{\boldsymbol{W}_1} \leq \frac{1}{1-\eta} \sum_{s \in [T]} \sum_{k \in [K]} \boldsymbol{\alpha}_k(s) \cdot \Big(1 + \eta \cdot \widetilde{\boldsymbol{\pi}}_k(s) + (e-2) \cdot \eta^2 \cdot \widetilde{\boldsymbol{\pi}}_k^2(s)\Big)$$

Moreover,

$$\log \frac{\boldsymbol{W}_{T+1}}{\boldsymbol{W}_1} = \log \sum_{k\in[K]} \exp\left(\eta \cdot \sum_{s\in[T]} \widetilde{\boldsymbol{\pi}}_k(s)\right) - \log K \geq \eta \cdot \sum_{s\in[T]} \widetilde{\boldsymbol{\pi}}_k(s) - \log K$$

Finally, for any $k \in [K]$, by the above, we have that

$$\begin{aligned} &\mathbb{E}\left[(1-\eta)\cdot \sum_{s\in[T]} \widetilde{\boldsymbol{\pi}}_k(s) - \sum_{s\in[T]}\sum_{k\in[K]} \boldsymbol{\alpha}_k(s)\cdot \widetilde{\boldsymbol{\pi}}_k(s)\right] \\ &\leq (1-\eta)\frac{\log K}{\eta} + (e-2)\cdot\eta\cdot\mathbb{E}\left[\sum_{s\in[T]}\sum_{k\in[K]} \boldsymbol{\alpha}_k(s)\cdot \widetilde{\boldsymbol{\pi}}_k^2(s)\right] \end{aligned}$$

This implies that for any $k \in [K]$,

$$\sum_{s\in[T]} \boldsymbol{\pi}_k(s) - \mathbb{E}\left[\sum_{s\in[T]} \boldsymbol{\pi}_{y(s)}(s)\right] \leq \eta T + \frac{\log K}{\eta} + (e-2)\eta TK$$

Henceforth,

$$\mathcal{R}_{\texttt{EXP3}} := \left(\max_{k\in[K]} \sum_{s\in[T]} \boldsymbol{\pi}_k(s)\right) - \mathbb{E}\left[\sum_{s\in[T]} \boldsymbol{\pi}_{y(s)}(s)\right] \leq \eta T + (e-1)\eta TK \leq \eta T(1+2K)$$

Choosing $\eta = \sqrt{\frac{\log K}{8TK}} \in o(1)$, we have the desired result, that is

$$\mathcal{R}_{\texttt{EXP3}} \leq 2\sqrt{TK \log K} \in o(TK)$$

This concludes the proof. □

The power of this theorem lies in the fact that we are able to guarantee sublinear regret for such a simple online learning algorithm. At a first glance, this might seem to be the same as saying that`EXP3` is also guaranteed high utility. In fact, this is not the case not only for the `EXP3` algorithm but for any kind of *mean-based algorithm* [5] which we hereby define.

Definition. *Following the notation of above for $\pi_x(s)$, let $\mathcal{G}_{x,t} := \sum_{s \in [t]} \pi_x(s)$. An algorithm is γ-mean-based if whenever $\mathcal{G}_{x,t} < \mathcal{G}_{y,t} - \gamma T$, the probability that an algorithm $\mathcal{A}$ pulls arm x on round t is at most γ. An algorithm is mean-based if it is γ-mean-based for some $\gamma \in o(1)$.*

Hereby, we summarize how to construct a setting where a strategic adversary could easily exploit a mean-based algorithm following [5]. Indeed, Braverman et al. in section D of the appendix of [5] prove that several algorithms are mean-based algorithms. Some examples are `MWU`, `FTL` and `EXP3`: all of them have similar regret guarantees and all of them as we will see shortly are easily exploitable. We have show-cased the characteristics of `EXP3` as an instance of a broader class of algorithms. Hence, let us consider the following setting: in each of the T total rounds, we have a single mean-based buyer interested in getting an item from a strategic seller who has the power of setting the price between 0 and 1 at each time step and offering a finite number of options n to the the buyer, each associated to a bid b_i, for $i \in [n]$. The buyer has valuation drawn from an unknown-to-the-seller distribution $\mathcal{F}$ and, by the fact of being mean-based, it may be able to learn over time and will intuitively pay the price that yields the best historical performance, i.e. yields the lowest regret. The following theorem, based on appendix B of [5], establishes that mean-based algorithms perform very poorly in the setting outlined above. In particular, the authors of [5] construct a strategic algorithm for the seller renamed in this work as `FOOL-BUYER` (outlined below). Hereby, we decided to report this simple strategic algorithm needed to achieve an almost full extraction of the bidder's value as this gives deep insights into the approaches of the following sections. The rationale behind `FOOL-BUYER`, as the name suggests, is to make the buyer believe that he or she could the item for free at the beginning, irrespective of his or her item valuation. Then later, the seller will charge the buyer the most, and, since the latter is a mean-based player, it will be more convenient for him or her to keep on buying the item until his or her own utility does not become 0. This guarantees of courses sublinear regret but, at the same time, it guarantees that the buyer will get vanishingly small utility by the end of the game, which will flow into the seller's coffers.

Idea. Let us assume that $\mathbf{supp}(\mathcal{F}) \not\subseteq [1-\varepsilon, 1]$ as otherwise the strategic seller could at each time step price the item at $1-\varepsilon$ and guarantee the revenue in the theorem statement. Let us define the following quantities, recalling that $\mathbf{supp}(\mathcal{F}) = \mathcal{B} = \{b_1, \dots, b_K\}$, whose elements are listed in ascending order without loss of generality. Thus,

- $\rho = \min\left(b_K, 1-\frac{\varepsilon}{2}\right)$ (in case of prior-independence, we need $\rho = 1-\frac{\varepsilon}{2}$).
- $\delta = \frac{1-\rho}{1-b_1}$ (in case of prior-independence, we need $\delta = \frac{\varepsilon}{2}$).

The seller in general will give the buyer $n = \frac{\log\frac{\varepsilon}{2}}{\log(1-\delta)}$. Let us also denote by $p_i(s)$ the price of the item i at time step s and by $q_i(s)$ the indicator of whether item i has been allocated at time step s. We are ready to outline the strategy for the seller.

FOOL-BUYER **algorithm** [5]

```
FOR $s \in \left[(1-(1-\delta)^{i-1})T\right]$:
    ALLOCATE: $p_i(s) = 0$, $q_i(s) = 0$
FOR $s \in \left[\left(1-(1-\delta)^{i-1}\right)T+1, \left(1-\rho\,(1-\delta)^{i-1}\right)T\right]$:
    ALLOCATE: $p_i(s) = 0$, $q_i(s) = 1$
FOR $s \in \left[\left(1-\rho\,(1-\delta)^{i-1}\right)T+1, T\right]$:
    ALLOCATE: $p_i(s) = 1$, $q_i(s) = 1$
```

For the first time period, the seller charges 0 and the buyer does not get the item, then, the seller still charges 0 but the buyer gets the item and for the final time period the seller charges 1 and the buyer gets the item.

Theorem. *For a mean-based buyer, for any constant $\varepsilon > 0$, the* ***FOOL-BUYER*** *strategy guarantees utility at least $(1-\varepsilon)\mu T - o(T)$ to the seller, where $\mu := \mathbb{E}_{f\sim\mathcal{F}}[f]$.*

Proof.

A full proof of this theorem may be found in theorem B.1 in [5].

□

The proof makes the dichotomy between regret minimization and utility maximization evident. In particular, the first does not imply the second and, moreover, a regret-minimizing mean-based player may end up with a an $o(T)$ utility. Next, we will outline a very simple auction model we have thought of in order to show that a mean-based algorithm such as EXP3 can be provably exploited by a simple iterative technique like the one of above even when the strategic player (seller) draws a private valuation from a private distribution $\mathcal{D}$.

3.3 Auction Model: Welfare Redistribution Auction

We would like to study the simplest discretized auction setting where we can exploit a regret-based learner. This setting needs to have the following properties:

1. Any agent has a finite number of actions to choose from.
2. At each time step, each agent has valuation v that is drawn independently from the same distribution $\mathcal{D}$.
3. The payoff bi-matrix is symmetric and time independent.
4. The payment scheme is monotone.

In order to start our analysis, let us consider the following simple game which has the following auction model specification.

Action Space. The action space is $\mathcal{B} = \{\ell, m, h\}$ (meaning low, medium and high), for some values of ℓ, m, h such that $\ell < m < h$.

Bidders. There are two competing bidders who have valuation $v \sim \mathcal{D}$ and $f \sim \mathcal{F}$, such that $\mathbf{supp}(\mathcal{D}) = \mathbf{supp}(\mathcal{F}) = \mathcal{B}$. We will assume for the rest of the analysis that $\mathcal{F}$ represents parametrically the true value distribution of the `EXP3` player and that $\mathcal{D}$ is the true value distribution of the player trying to exploit the `EXP3` algorithm over time.

Procedure. At each time step $t \in [T]$, we make the two bidders independently draw valuations $(v, f) \sim \mathcal{D} \times \mathcal{F}$ and let them bid based on the history but unaware of the other's valuation or distribution.

Payment Scheme. Let us denote by $p_{m\ell}, p_{h\ell}, p_{hm}$ the amount of money a player needs to pay to the opponent when respectively he or she bids medium and the other bids low, he or she bids high and the other bids low and he or she bids high and the other bids medium. Of course, since the payment scheme needs to be monotone, $p_{m\ell} < p_{h\ell} < p_{hm}$. We also need to impose the condition that $p_{m\ell} + p_{h\ell} \neq p_{hm} + \frac{1}{2}$, in order for the players not to be indifferent between bidding low, medium or high.

Payoff Bi-Matrix. In compact form, we can represent the payoff bi-matrix for the two competing bidders as follows:

		Bidder $\mathcal{F}$		
		ℓ	m	h
	ℓ	$\left(\frac{v}{2}, \frac{f}{2}\right)$	$(p_{m\ell}, f - p_{m\ell})$	$(p_{h\ell}, f - p_{h\ell})$
Bidder $\mathcal{D}$	m	$(v - p_{m\ell}, p_{m\ell})$	$\left(\frac{v}{2}, \frac{f}{2}\right)$	$(p_{hm}, f - p_{hm})$
	h	$(v - p_{h\ell}, p_{h\ell})$	$(v - p_{hm}, p_{hm})$	$\left(\frac{v}{2}, \frac{f}{2}\right)$

As we may observe, this payoff bi-matrix possesses some notion of symmetry. Indeed, whenever one overbids the other, the amount paid to the opponent is the same (clearly, there cannot be any symmetry around the payoff itself given that generally $\mathcal{D}$ and $\mathcal{F}$ are different). Whenever the two bidders tie, they each receive expected payoff equal to half their valuation: this could be seen as each player getting the item with probability $\frac{1}{2}$ and paying nothing to the opponent.

3.4 Conjecture: Mean-based bidders in symmetric settings

In order to understand whether we are able to replicate and extend the `FOOL-BUYER` algorithm in this new setting (i.e. WRA), we need to establish how much, theoretically, the strategic bidder could extract from the mean-based one. Thus, we could formulate this theoretical best as follows:

$$\begin{aligned}
\text{BEST} &= \mathbb{E}_{(v,f)\sim\mathcal{D}\times\mathcal{F}}\left[\sum_{s\in[T]}\left(\frac{v+f}{2}+\max(v-f,0)\right)\right] \\
&= \sum_{s\in[T]}\mathbb{E}_{(v,f)\sim\mathcal{D}\times\mathcal{F}}\left[\frac{v+f}{2}+\max(v-f,0)\right] \\
&= \sum_{s\in[T]}\mathbb{E}_{(v,f)\sim\mathcal{D}\times\mathcal{F}}\left[\max(v,f)\right] \\
&= T\cdot\mathbb{E}_{(v,f)\sim\mathcal{D}\times\mathcal{F}}\left[\max(v,f)\right] \\
&:= \mathfrak{s}T
\end{aligned}$$

The expression above for BEST is such because the term $\max(v-f,0)$ is the added utility to the world from bidder with distribution $\mathcal{D}$ joining, and we are trying to express the idea that all the benefit from bidder with distribution $\mathcal{D}$ joining accrues only to him or herself. In fact, we could think of the strategic bidder as a seller who loses money if he or she does not sell the item for the price he or she should.

The problem at hand looks indeed much harder than its predecessor for multiple reasons which can be outlined below:

- On one hand, the strategic bidder is now constrained by its valuation of the item and, thus, as mentioned earlier, we could think of him or her as a seller that cannot afford not to sell the item for the amount he or she should.

- Moreover, the payment scheme should now reflect a trade-off between efficiency and truthfulness. This means that if the payment scheme is such that bidding truthfully in the single shot game is the dominant strategy or dominant in expectation in the repeated game, then, there is no strategy for which we can exploit the mean-based bidder.

- The found strategy might be prior-dependent which would considerably limit the scope of the study. Indeed, if exploitation requires knowledge of the mean-based bidder's private distribution, then, we would not truly be exploiting the mean-based assumption but, to some extent, some unrealistic knowledge about the prior.

We, thus, formulate the following conjecture.

Conjecture. *For a mean-based bidder, for any constant $\varepsilon > 0$, for some values of bids $\mathcal{B} = \{b_1, \dots, b_K\}$ and some payment scheme $\{p_{ij}\}_{(i,j)\in\mathcal{B}^2}$, there exists a strategy which guarantees utility at least $(1-\varepsilon)\mathfrak{s}T - o(T)$ to the strategic bidder, where $\mathfrak{s} := \mathbb{E}_{(\nu,f)\sim\mathcal{D}\times\mathcal{F}}[\max(\nu, f)]$.*

As a matter of fact, it is not at all trivial to conclude that there exists a payment scheme for which the strategic bidder extracts the full surplus from the mean-based one.

4 Implementation and Evaluation

The following section is organized as follows:

- In subsection 3.1, we outline the experimental setting by explicitly setting the parameters of ℓ, m, h as well as $p_{m\ell}, p_{h\ell}, p_{hm}$ such that $p_{m\ell} < p_{h\ell} < p_{hm}$ and $p_{m\ell} + p_{h\ell} \neq p_{hm} + \frac{1}{2}$. We will also provide a short rationale of why exactly we chose this kind of setting by showing that in this kind of game, with the specified payment scheme, truthful bidding is not necessarily dominant on underbidding or overbidding.

- In subsection 3.2 and 3.3, we run two baselines settings: on one hand, the game of above for two EXP3 agents competing against each other. On the other hand, an EXP3 agent against a DQN (Deep Q-Learning) one. We started exploring various Reinforcement Learning algorithms insofar as we thought of them as potential candidates for "strategic bidders" since the function they optimize takes into account future rewards much more heavily than mean-based ones and is, therefore, much less localized. However, we experimentally find that both types of agents arrive at similar convergence regret as well as similar utility throughout the game and, henceforth, DQN does not buy anything to an agent than a simpler EXP3.

4.1 Experiment design

We could summarize the experimental design as follows for $T = 40,000$:

Action Space. The action space is $\mathcal{B} = \{\ell, m, h\} = \{0, 9, 10\}$.

Bidders. There are two competing bidders who have valuation $v \sim \mathcal{D} = [0, 1, 0]$ and $f \sim \mathcal{F} = [0, 1, 0]$.

Payment Scheme. We have $\{p_{m\ell}, p_{h\ell}, p_{hm}\} = \left\{\frac{\ell+m}{4}, \frac{\ell+h}{4}, \frac{m+h}{4}\right\} = \left\{\frac{9}{4}, \frac{5}{2}, \frac{19}{4}\right\}$. Indeed, the payment scheme is monotone, $\frac{9}{4} < \frac{5}{2} < \frac{19}{4}$ and the non-indifference conditioned is also ensured given that $\frac{9}{4} + \frac{5}{2} \neq \frac{19}{4} + \frac{1}{2}$.

Payoff Bi-Matrix. Then, the payoff bi-matrix for the two bidders is

		Bidder $\mathcal{F}$		
		0	9	10
	0	$\left(\frac{\nu}{2}, \frac{f}{2}\right)$	$\left(\frac{9}{4}, f - \frac{9}{4}\right)$	$\left(\frac{5}{2}, f - \frac{5}{2}\right)$
Bidder $\mathcal{D}$	9	$\left(\nu - \frac{9}{4}, \frac{9}{4}\right)$	$\left(\frac{\nu}{2}, \frac{f}{2}\right)$	$\left(\frac{19}{4}, f - \frac{19}{4}\right)$
	10	$\left(\nu - \frac{5}{2}, \frac{5}{2}\right)$	$\left(\nu - \frac{19}{4}, \frac{19}{4}\right)$	$\left(\frac{\nu}{2}, \frac{f}{2}\right)$

Hereby, we provide a short rationale of why we chose this payment scheme. In short, if truthful bidding were to be a dominant strategy in the single time step game, then, despite not being not necessarily dominant in the repeated game as the Repeated Prisoners' Dilemma shows, it will be the case that `EXP3` will converge to the truthful strategy which would be hard to exploit. Let us, therefore, turn our attention to the game at hand and denote by b, c the bids of bidder $\mathcal{D}$ and bidder $\mathcal{F}$ respectively:

(**Overbidding**) Assume $b > \nu$:

- If $\nu > c$, then bidder $\mathcal{D}$ wins anyway and the payoff would even be higher had he or she bid truthfully: we have payoffs $\nu - \frac{b+c}{4} < \nu - \frac{\nu+c}{4}$, the first of which is untruthful and the second is truthful.

- If $b < c$, then bidder $\mathcal{D}$ loses the item but gets a higher payoff than if he had bid truthfully: we have payoffs $\frac{b+c}{4} > \frac{\nu+c}{4}$, the first of which is untruthful and the second is truthful.

- If $\nu < c < b$, then bidder $\mathcal{D}$ wins the item and the payoff is $\nu - \frac{b+c}{4} < \nu - \frac{\nu}{2} = \frac{\nu}{2} < \frac{\nu+c}{2}$: the latter term is the payoff of truthful bidding.

(**Underbidding**) Assume $b < v$:

- If $v < c$, then bidder $\mathcal{D}$ would lose anyway but the payoff would be higher had he bid truthfully: we have payoffs $\frac{b+c}{4} < \frac{v+c}{4}$, the first of which is untruthful and the second is truthful.

- If $c < b$, then bidder $\mathcal{D}$ wins the item anyway but gets a higher payoff than if he had bid truthfully: we have payoffs $v - \frac{b+c}{4} > v - \frac{v+c}{4}$, the first of which is untruthful and the second is truthful.

- If $b < c < v$, then bidder $\mathcal{D}$ loses the item and the payoff is $\frac{b+c}{4} = \frac{b+c}{2} - \frac{b+c}{4} < v - \frac{b+c}{4}$: the latter term is the payoff of truthful bidding.

In other words, there is one overbidding and one underbidding scenario where bidder $\mathcal{D}$ is incentivized to lie about his or her valuation.

4.2 EXP3 vs. EXP3

4.2.1 Utility

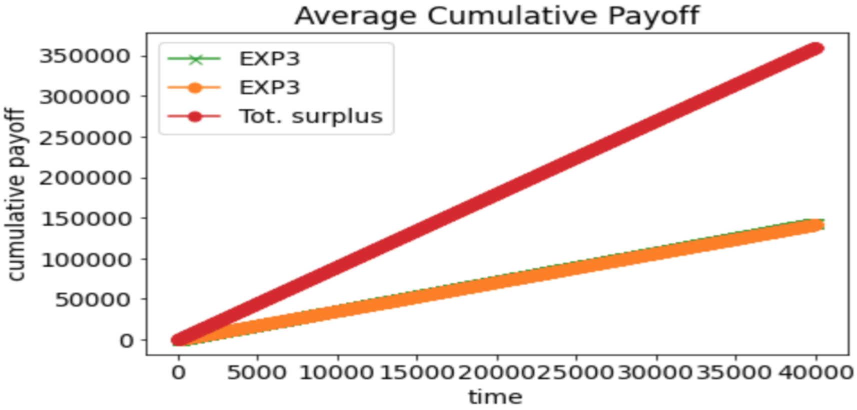

Figure 1: This graph depicts the two EXP3 agents' cumulative payoffs (utilities) by the end of the game in green and orange as well as the theoretical best (total surplus) which is the cumulative sum of $\max(v, f)$ as outlined in previous sections in red. As expected, the two agents perform comparably in the sense that their utility is substantially the same over time.

4.2.2 Regret and Average Payoff

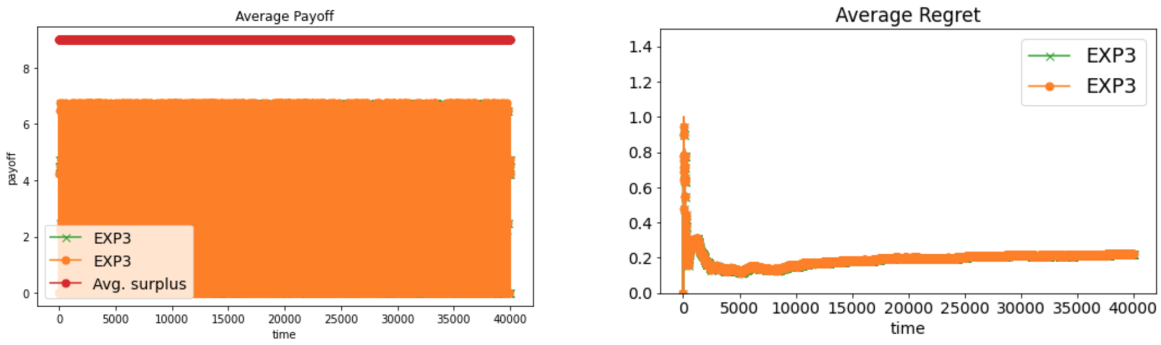

Figure 2: These graphs depict respectively the two `EXP3` average payoffs at each time step in green and orange as well as how the regret evolves over time. Not only does the regret converge to the same value for both but it does so in the same way, which is not unexpected since the two agents are the same.

4.2.3 Actions

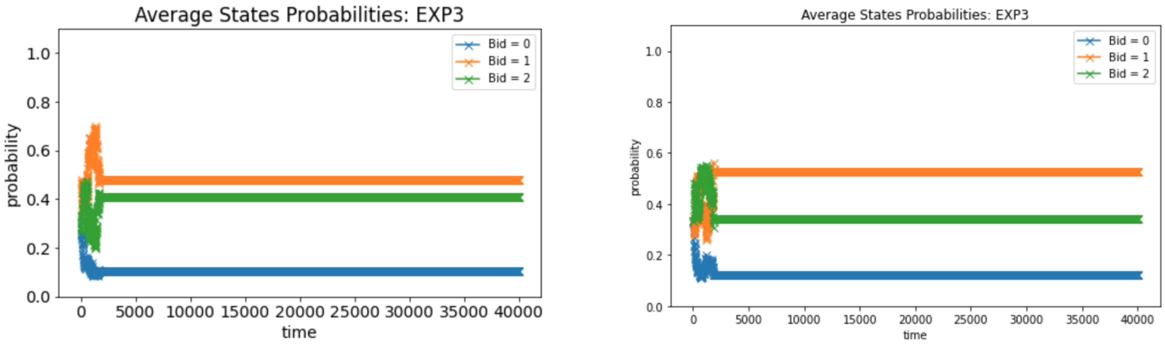

Figure 3: These graphs depict respectively how the two `EXP3` agents have learnt to bid over time: given that they both have distribution of value equal to $[0, 1, 0]$, despite the first numerical instabilities, they converge essentially to the same behavior. Let us not that this is close to truthful bidding but not exactly so.

4.3 EXP3 vs. DQN

4.3.1 Utility

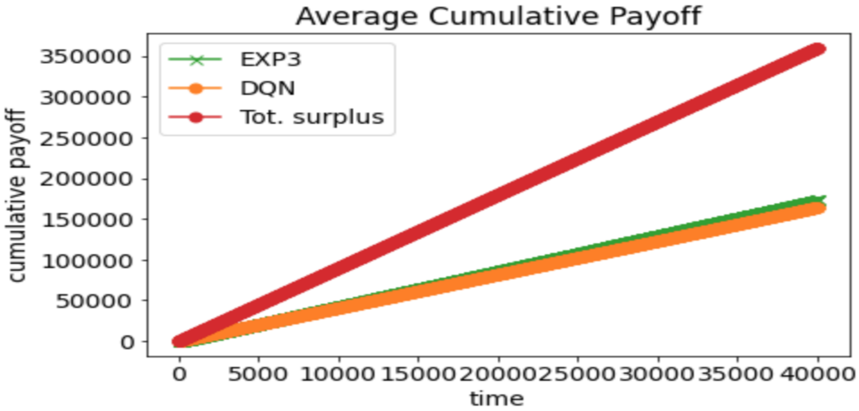

Figure 4: This graph depicts the EXP3 and DQN agents' cumulative payoffs (utilities) by the end of the game in green and orange as well as the theoretical best (total surplus) which is the cumulative sum of $\max(v, f)$ as outlined in previous sections in red. One would expect episodic Deep Q-Learning to have better performance than a simple EXP3 but indeed their utility is substantially the same over time. This result is confirmed throughout a variety of parameter specifications.

4.3.2 Regret and Average Payoff

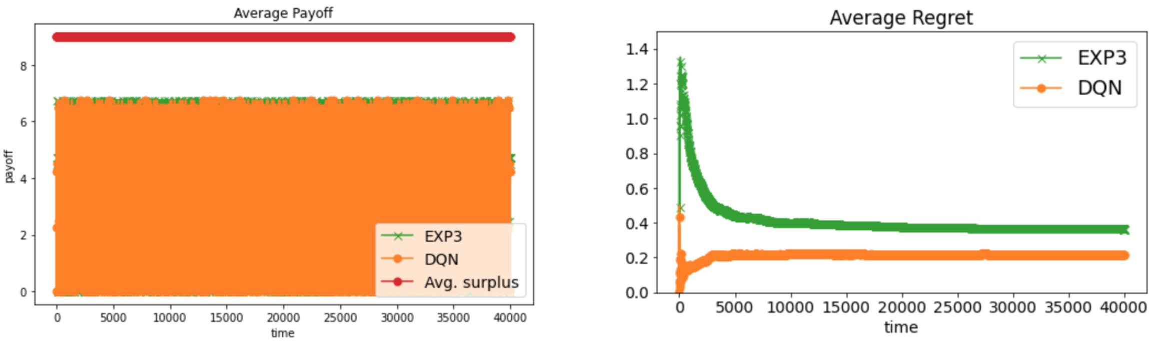

Figure 5: These graphs depict respectively the EXP3 and DQN agents' average payoffs at each time step in green and orange as well as how the regret evolves over time. The payoffs are very similar one to the other but, as we will see in the next couple of pictures, this is not due to the fact hat the two algorithm converge to the same strategy. It is also interesting to see that there is a sharp drop in regret immediately for the DQN agent which mildly increases over time but still remains under the converging regret of the EXP3 agent.

4.3.3 Actions

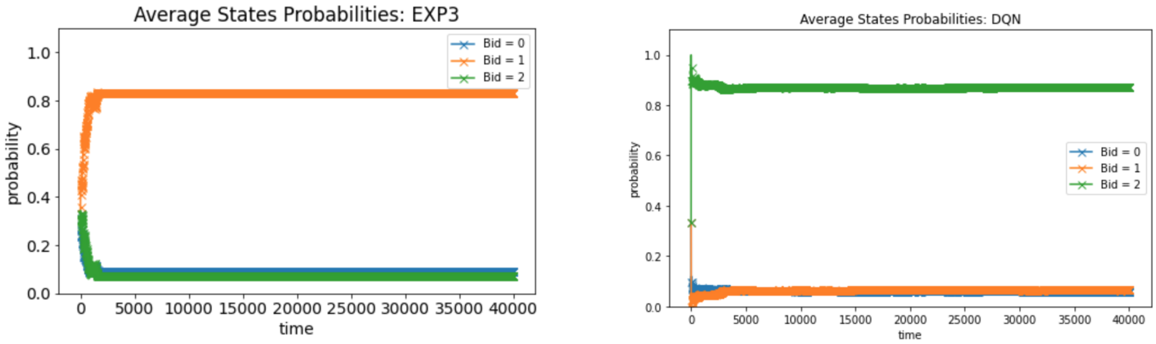

Figure 6: These graphs depict respectively how the EXP3 and DQN agents have learnt to bid over time: they both have distribution of value equal to $[0, 1, 0]$, but the DQN agent converges much faster to a purely untruthful strategy, whereas the EXP3 converges to almost pure truthful bidding as a best response to the DQN agent.

5 Conclusion

5.1 Summary

In this research paper, we have considered the exploitation of mean-based bidders, thus, extending the framework of selling to a no-regret buyer which has been developed in [5] by Braverman et al.. We delved into the formulation of the conjecture and explained the reasons why this problem looks much harder than its predecessor. We, then, ran experiments and established that we could extend the class of mean-based algorithms to Reinforcement Learning ones.

5.2 Future Work

The future work can be framed by considering two possible scenarios. On one hand, if the conjecture holds, then we would have two natural subscenarios depending whether or not `DQN` belongs to the mean-based algorithms class or not. In the former case, then, the conclusion is simple: any `DQN`-based bidding algorithm would be easily exploitable by a strategic bidder. In the latter case, then, it might be that the conjecture is easily extendable to a wider class of algorithms and not just mean-based ones. Possibly, the most interesting conclusion would appear if the conjecture happened not to hold. In this very case, the simple addition of the value constraint for the strategic bidder makes any mean-based algorithm unexploitable for any sort of payment scheme. Although this would be a strong statement, it also would entail that strategic bidding is far less powerful than strategic selling and, thus, the problem is provably much more complex than its predecessor which regarded an all-powerful seller against a mean-based buyer as per [5].

6 Acknowledgments

I would like to thank Professor Mark Braverman for his incredibly helpful support and feedback during the development of our research process. His wonderful help and great insights gave me the opportunity to improve the quality of my Independent Work. Finally, I would like to thank Pier Giuseppe Sessa and the ETHZ LAS Lab lead by Professor Andreas Krause for the code that they have made available on GitHub for public use.

7 Ethics

This report represents my work in accordance to University regulations. I pledge my honor that I have not violated the Honor Code during the composition of my Independent Work.

References

[1] Raman Arora, Michael Dinitz, Teodor V. Marinov, and Mehryar Mohri. Policy regret in repeated games. *CoRR*, abs/1811.04127, 2018.

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

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

[4] Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Multi-armed bandit problems with strategic arms. *CoRR*, abs/1706.09060, 2017.

[5] Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Selling to a no-regret buyer. *CoRR*, abs/1711.09176, 2017.

[6] Edward Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, 1971.

[7] Yuan Deng, Jon Schneider, and Balusubramanian Sivan. Strategizing against no-regret learners. 2019.

[8] Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–31, 1973.

[9] Hoda Heidari, Mohammad Mahdian, Umar Syed, Sergei Vassilvitskii, and Sadra Yazdanbod. Pricing a low-regret seller. 48:2559–2567, 20–22 Jun 2016.

[10] Randolph McAfee and John McMillan. Auctions and bidding. *Journal of Economic Literature*, 25(2):699–738, 1987.

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

[12] Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. pages 437–448, 2005.

[13] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37, 1961.