Repeated Dollar Auctions: A Multi-Armed Bandit Approach

Marcin Waniek
University of Warsaw
vua@mimuw.edu.pl

Long Tran-Thanh
University of Southampton
l.tran-thanh@soton.ac.uk

Tomasz Michalak
University of Oxford &
University of Warsaw
tomasz.michalak@cs.ox.ac.uk

ABSTRACT

We investigate the repeated version of Shubik's [34] dollar auctions, in which the type of the opponent and their level of rationality is not known in advance. We formulate the problem as an adversarial multi-armed bandit, and we show that a modified version of the ELP algorithm [25], tailored to our setting, can achieve $\tilde{O}(\sqrt{|\mathcal{S}_0|T})$ performance loss (compared to the best fixed strategy), where $|\mathcal{S}_0|$ is the cardinality of the set of available strategies and T is the number of (sequential) auctions. We also show that under some further conditions, these bound can be improved to $\tilde{O}(|\mathcal{S}_0|^{1/4}\sqrt{T})$ and $\tilde{O}(\sqrt{T})$, respectively. Finally, we consider the case of spiteful players. We prove that when a non-spiteful player bids against a malicious one, the game converges in performance to a Nash equilibrium if both players apply our strategy to place their bids.

CCS Concepts

•Theory of computation → Online learning algorithms; Computational pricing and auctions;

Keywords

dollar auction; online learning; multi-armed bandits

1. INTRODUCTION

Conflicts are an integral part of human culture [8]. Whether it is a competition of rival technological companies [38], a battle of lobbyists trying to acquire a government contract [15] or the arms race of modern empires [31], conflict analysis has always been of interest to game theorists and economists alike [18]. The analysis of conflict is also of natural interest analysed in AI and, especially, in Multi-Agent Systems (MAS) literature [27, 37, 11]. Typically conflicts in MAS used to be considered as failures or synchronisation problems [40, 36]. More recently, however, the need for more advanced studies has been recognized [10], including analyses of conflict generation, escalation, or detection in MAS.

A very simple dollar auction introduced by Shubik [34] provides a powerful framework to formalise and study conflict situations. In this auction two participants compete for a dollar by making bids just as in the regular English auction. However, both the winner and the loser have to pay their final bids. In other words, this is an all-pay auction which means that, once a player decides to participate and make her first bid, any invested resources are irretrievably lost. Consequently, the only way to gain any profit (or minimize the loss) is to win the dollar. However, since both players are in the same situation, the dollar auction often ends up with an irrational conflict escalation [23].

Interestingly, however, in a celebrated work, O'Neill [31] proved that the first player has a winning strategy in the dollar auction assuming pure strategies. In particular, if she makes a bid that is a specific combination of the auction's stake, the budget of both players and the minimal bid increment, her opponent will never win the auction and should therefore pass. For instance, if the price increment is \$0.05 and both players have budgets of \$2.50 each, then the first player that has a chance to bid should offer precisely \$0.60 and the other player should drop out. As a result, escalation should never occur. The result by O'Neill suggests that conflict escalation should never occur and if it does so in real-life experiments, it has to be related to human bounded-rationality or other factors not accounted for by the auction model.

However, O'Neil's result was re-examined by Leininger [23] who showed the advantage of the first bidder vanishes considering mixed strategies equilibria and the auction escalates. Demange [13] proved the same for settings where players are uncertain about each other's strength. In addition, Waniek *et al.* [39] showed that when used against a spiteful or malicious player, the optimal strategy proposed by O'Neill can lead to severe loss.

One of the key deficiencies of all of the above results is that they concern one-shot auction case, where the goal is to investigate what would be an optimal bidding strategy in the single auction. However, many real-world applications consist of a sequence of auctions repeatedly played against the same opponent [7, 20, 17]. In fact, many examples of R&D competition, arms races, lobbying, or interactions in MAS can be seen not as a single instance of an auction, but rather as a series of confrontations. We will show later in this paper that the nature of repeated encounters can be leveraged in order to achieve good performance.

Typically in such repeated settings, there is little (if any) prior knowledge about the opponent's incentives, nor about her rationality level. Second, while it is typical to assume some (perfect or bounded) rationality model of the players, it might be the case that the chosen bids do not follow such

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

assumptions (e.g., due to lack of computational power, or lack of perfect execution). As such, efficient bidding strategies need to consider the uncertainty in both the incentive and rationality of the opponent.

In this paper, we reconsider Shubik's dollar auction but now in the repeated setting. We ask whether it is possible to design efficient bidding strategies, given the high level of uncertainty and no prior knowledge of the opponent. Put differently, we investigate whether a participant's knowledge about the opponent (gained in the previous encounters) can give the participant an edge over her opponent so that the escalation does not occur. To this end, we view the problem from the perspective of online learning theory, where the auctions are played against an adversary.

More specifically, we formulate the problem using an adversarial multi-armed bandit model [5, 25], where at each round we choose from a (finite) set of possible strategies $\mathcal{S}_0$ to play against the opponent. At the end of each round, we observe the outcome of the auction and update our estimation on the strategy's fitness (*i.e.*, how much expected profit the strategy can achieve). Our model also assumes that there is some level of interdependence between the different strategies. That is, by observing an outcome of a specific strategy, we can get some *side information* about the fitness of other strategies. For example, if a particular strategy wins the auction with a (final) bid b, then any other strategies that shares the same behaviour as the current one up to b (*i.e.*, the latter is a *prefix* of the former), would also win the auction. As such, we can also update our fitness estimate of the interdependent strategies as well.

Given this, we prove that by applying ELP (*i.e.*, "Exponentially weighted algorithm with Linear Programming")—a state-of-the-art online learning algorithm proposed by [25]—we can achieve $\tilde{O}(\sqrt{|\mathcal{S}_0|T})$ performance loss [1] against the best fixed bidding strategy (*i.e.*, playing the same strategy each round) in hindsight. In particular, since the regret is sub-linear due to the square root in $\tilde{O}(\sqrt{|\mathcal{S}_0|T})$, it implies that the average regret per time step is converging to 0 as T tends to infinity. Thus, the behaviour of ELP converges to that of the best fixed strategy in hindsight. While this result is a direct consequence of the performance analysis of ELP from [25], our main contributions are to improve this result to more concrete cases where we have some additional conditions. Specifically:

- By allowing some additional side-information structure (about the other strategies) that we obtain when observing the outcome of the chosen strategy, we prove that the performance loss can be bounded with $\tilde{O}(|\mathcal{S}_0|^{1/4}\sqrt{T})$, which is an improvement of $|\mathcal{S}_0|^{1/4}$ (note that this improvement is significant, especially when $|\mathcal{S}_0|$ is large);

- When $\mathcal{S}_0$ is the class of non-spiteful optimal bidding strategies with different budget limits, and the opponent is deterministic (*i.e.*, she chooses the next bids in a deterministic manner), the performance loss of the ELP-based bidding strategy is at most $\tilde{O}(\sqrt{T})$.

Finally, we show that when the auction is played between a non-spiteful and a malicious player, if both use ELP, they will converge to a Nash equilibrium.

[1]The logarithmic terms are hidden in the $\tilde{O}$ notation.

The remainder of the paper is organised as follows. We start by introducing the basic definitions and notation, and then outline our multi-armed bandit based bidding model. We then describe the ELP algorithm, and analyse its performance in different settings. The case of non-spiteful against malicious player is discussed afterwards, followed by the conclusions. Appendix provides notation table for the reader's convenience.

2. PRELIMINARIES

In this section, we formally describe the two building blocks of our problem, namely the dollar auction and the multi-armed bandit model.

2.1 The Dollar Auction

The dollar auction [34] is an all-pay auction between two players $N = \{0, 1\}$. We will often denote them by i and j. The winner of an auction is given the stake $s \in \mathbb{N}$. The players can make bids that are multiples of a minimal bid increment δ. Without loss of generality, we assume that $\delta = 1$.

While Shubik [34] considered the dollar auction without any budget limit, following O'Neill [31], most subsequent studies considered the setting in which players have limited budgets $b_0, b_1 \in \mathbb{N}$. In this paper, we assume that both players have equal budgets, *i.e.*, $b_0 = b_1 = b$.

The players move in turns and the starting player is determined randomly at the beginning of the auction. Moreover, the starting player i can either make the first bid, pass and leave the auction, or let j move first. Once the first bid has been made, any player making the move can either make a bid higher than the opponent, or pass. In other words, offering the turn to the opponent is only permitted before the first bid is made. We denote by X_b the set $\{(x, y) \in \{0, \ldots, b-1\} \times \{0, \ldots, b-1\} : y > x\} \cup \{(0, 0)\}$ where x is the last bid of the player currently choosing her bid and y is the last bid of her opponent. Whenever it is a player's turn to make a bid, the dollar auction would be in one of the states in X_b.

We assume that all the strategies in our setting are deterministic. Let us now consider $\mathcal{S}$ the set of all possible strategies in the dollar auction. We have $\mathcal{S} \subset X_b \to \{0, \ldots, b\}$, where for any strategy $f \in \mathcal{S}$, the value $f(x, y)$ represents the bid to be made by the player whose last bid was x and the last bid of her opponent was y. For valid strategies we have $f(x, y) \geq y$, where $f(0, 0) = 0$ represents the decision to let the opponent move first if the player starts (as it is never rational for any player to pass if her opponent has not made any bids yet), and $f(x, y) = y$ represents the decision to pass in all other cases. Set of strategies S contains all valid strategies, *i.e.*, strategies that meet conditions described above.

An auction ends when one of the players either passes or makes the bid that her opponent cannot top (in a setting with equal budgets this would mean bidding the entire budget). The stake is then given to the higher bidder. However, since this is an all-pay auction, both players have to pay their final bids. The profit of player i from an auction that ends with bids y_i and y_j is therefore:

$$p_i = \begin{cases} s - y_i & \text{if } y_i > y_j, \\ -y_i & \text{if } y_i < y_j. \end{cases}$$

2.2 The Multi-Armed Bandit Model

We now describe the *adversarial bandit* model [6], a variation of the *multi-armed bandit* problem [33]. In particular, it consists of T rounds, and a player can choose one of K actions (traditionally described as different slot machines, or bandits, hence the name of the multi-armed bandit model) to play at each round. However, before the player chooses her action to play, an adversary assigns an arbitrary reward vector $r(t) \in [0,1]^K$, where $r(t) = (r_1(t), \ldots, r_K(t))$, without revealing it to the player. After the player chooses her action i_t to play, the value $r_{i_t}(t) \in [0,1]$, which represents reward for performing action i_t in round t, is revealed to the player as her (collected) reward. The goal of the player then is to maximise her total collected reward over T rounds.

Let A denote the algorithm of playing actions chosen by the player, which selects action i_{t+1} at each in round $t+1$ to perform (the choice of action may be based on the observed results $r_{i_1}(1), \ldots, r_{i_t}(t)$ of the previously chosen actions $i_1, \ldots, i_t$). The total collected reward, or the return, of algorithm A after T rounds is as follows:

$$R_A(T) = \sum_{t=1}^{T} r_{i_t}(t).$$

Since the value of each $r(t)$ is not revealed to the player, it is impossible to maximise the term above without having any additional information about $r(t)$. Instead, the performance of an algorithm is measured with a *regret*, which is defined as the difference between the return of the best fixed action over time and the return of the algorithm, namely:

$$U_A(T) = \max_i \sum_{t=1}^{T} r_i(t) - \sum_{t=1}^{T} r_{i_t}(t).$$

Within our bandit formulation, the repeated dollar auction problem provides side information. Within the bandit literature, the first work that fits our setting is the work of Mannor and Shamir [25], which introduced the multi-armed bandit model with side information. In particular, within their model, action j is a *neighbour* of action i chosen by the algorithm in round t if and only if for some fixed parameter d we are able to create unbiased estimator $\hat{r}_j(t)$ of the reward $r_j(t)$, *i.e.*, $\mathbb{E}[\hat{r}_j(t) | \text{action } i \text{ chosen in round } t] = r_j(t)$ and $\mathbb{P}(|\hat{r}_j(t)| \leq d) = 1$. Intuitively, we can use the knowledge gained from the execution of action i to estimate the reward of actions in its neighbourhood.

This extra information can be represented as a sequence of graphs $G_1, \ldots, G_T$ with K nodes each. Let $N_i(t)$ denote the neighbourhood of node i in graph G_t, *i.e.*, an edge from i to j exists in graph G_t if and only if $j \in N_i(t)$. Mannor and Shamir proposed an algorithm, called ELP, to efficiently tackle this bandit model. Here we will rely on the model of Mannor and Shamir and the ELP algorithm, and fit them to our setting of the repeated dollar auction.

3. META-BIDDING ALGORITHM

We now turn to the description of our bidding algorithm. In particular, this algorithm is a meta-strategy that consists of multiple dollar auction strategies played at each round, in order to minimize the regret from the series of auctions. To do so, we first describe the multi-armed bandit formulation of our problem. We then define the neighbourhood of a strategy within the strategy space. Finally we discuss how to tailor the ELP algorithm to our setting.

3.1 Model Description

In our setting, a player takes part in a series of T dollar auctions against the same opponent, each auction corresponds to a single round in the multi-armed bandit model. We assume that the player has at her disposal a set of strategies $\mathcal{S}_0 \subseteq \mathcal{S}$. Before each round t the player has to choose a strategy $f_t \in \mathcal{S}$ to be used in this particular round. Choice of strategy corresponds to the choice of an action (a playing arm) in the multi-armed bandit model. On the other hand, her opponent also chooses her strategy (equivalent of setting payoffs for each action). We do not put any explicit restrictions on the strategy of the opponent.

The auction is played using chosen strategies and the player achieves profit $p_{f_t}(t) \in \{-b+1, \ldots, s\}$. The profit is then mapped to reward $r_{f_t}(t)$, corresponding to the reward in the multi-armed bandit model. We assume that always $r_{f_t}(t) \in [0,1]$ (we normalise the rewards for the sake of simplicity). In case of a non-spiteful player (*i.e.*, player who just wishes to maximize her profit) reward depends only on her profit. We discuss the case of a malicious player, whose reward depends on the profit of her opponent in Section 6.

Given this, a meta-strategy is an algorithm that chooses a strategy to be used in each auction, based on the knowledge about previous auctions. Note that it corresponds to the algorithm of choosing an arm for each round in the multi-armed bandit model. The goal of the meta-strategy A is to minimize the regret after a series of auctions, regret being the difference between sum of rewards after using single best strategy in hindsight and after using a sequence of strategies chosen by the meta-strategy:

$$U_A(T) = \max_{g \in \mathcal{S}_0} \sum_{t=1}^{T} r_g(t) - \sum_{t=1}^{T} r_{f_t}(t).$$

3.2 Neighbourhood of a Strategy

We now define the relation of being a *prefix* strategy. In particular, strategy $g \in \mathcal{S}$ is a prefix of strategy $f \in \mathcal{S}$ if and only if $\forall_{(x,y) \in X_b} g(x,y) = f(x,y) \vee g(x,y) = y$. Intuitively, in every moment of the auction g either makes the same bid as f or passes. We can notice that f is also the prefix of itself. We can finally define *neighbourhood of the strategy* f as $N_f = \{g \in \mathcal{S} : g \text{ is a prefix of } f\}$. In fact, the following lemma justifies that playing a strategy provides side information about its prefixes.

LEMMA 1. *Assuming that player used strategy f in the dollar auction, we know the result of an auction if player used strategy g, that is a prefix of f.*

PROOF. After the auction using strategy f in round t we know a sequence of states $(x_1, y_1), \ldots, (x_n, y_n)$ when player was asked for a bid. If $\forall_i g(x_i, y_i) = f(x_i, y_i)$ then reward $r_g(t)$ for using g in auction would have been exactly the actual reward for using f, *i.e.*, $r_g(t) = r_f(t)$. Otherwise, reward for using g would have been $r_g(t) = -x_j$, where j is the lowest index such that $g(x_j, y_j) = y_j$, as the player would pass earlier than when using strategy f. □

3.3 The ELP Algorithm

We now describe how the ELP algorithm [25] can be adapted as a meta-bidding method in the dollar auction. The adaptation is based on the usage of previously defined set of prefixes

Algorithm 1 The ELP algorithm

Input: $\mathcal{S}_0 \subseteq \mathcal{S}$, β, $(\gamma(t))_{t=1}^T$, $(\cup_{f\in\mathcal{S}_0}\{q_f(t)\})_{t=1}^T$
  $\forall_{f\in\mathcal{S}_0} w_f(1) \leftarrow \frac{1}{|\mathcal{S}_0|}$
  for $t = 1, \ldots, T$ **do**
    for $g \in \mathcal{S}_0$ **do**
      $P_g(t) \leftarrow (1-\gamma(t))\frac{w_g(t)}{\sum_{h\in\mathcal{S}_0} w_h(t)} + \gamma(t)q_g(t)$
    $f_t \sim P(t)$
    Play f_t getting reward $r_{f_t}(t)$
    for $g \in N_{f_t}$ **do**
      Compute $\hat{r}_g(t)$
    for $g \in \mathcal{S}_0$ **do**
      if $f_t \in N_g$ **then**
        $\tilde{r}_g(t) \leftarrow \frac{\hat{r}_g(t)}{\sum_{h\in N_{g(t)}} P_h(t)}$
      else
        $\tilde{r}_g(t) \leftarrow 0$
    for $g \in \mathcal{S}_0$ **do**
      $w_g(t+1) \leftarrow w_g(t)\exp(\beta\tilde{r}_g(t))$

as the neighbourhood of the strategy, in order to determine the potential profit of other strategies.

Algorithm 1 presents the pseudocode of ELP. The algorithm uses strategies from the provided set $\mathcal{S}_0 \subseteq \mathcal{S}$ to play a series of T dollar auctions, selecting a single strategy for each auction. Typically for multi-armed bandits algorithms, the ELP algorithm chooses between exploration (checking new strategies) and exploitation (using so far most promising strategies to generate profit). This trade-off is expressed by parameters $\gamma(t)$ and $q_f(t)$ of the algorithm. We will discuss the exact values of these parameters, that provide low regret bounds, later in the paper.

In round t algorithm selects strategy $f_t \in \mathcal{S}_0$ using a probability distribution $P(t)$ incorporating knowledge gained in previous rounds. Knowing reward $r_{f_t}(t)$ received for using selected strategy, algorithm computes reward $\hat{r}_g(t)$ that would have been received for playing strategy g from the neighbourhood of f_t in round t. Method of computing this reward is given in the proof of Lemma 1. Basically, for auction where player was making decision in states $(x_1, y_1), \ldots, (x_n, y_n)$ we have $\hat{r}_g(t) = r_f(t)$ if $\forall_i g(x_i, y_i) = f(x_i, y_i)$ and $\hat{r}_g(t) = -x_j$ (where j is the lowest index such that $g(x_j, y_j) = y_j$) otherwise.

This reward is used to construct probability distribution in the next round. The higher the reward, the more probably the strategy is going to be used in the subsequent rounds.

3.4 Regret Analysis

We now give an upper bound for the regret of the ELP algorithm:

THEOREM 2. *Assume that the ELP algorithm is run using $\beta \in (0, \frac{1}{2|\mathcal{S}_0|})$, with $(q_f(t))_{f\in\mathcal{S}_0}$ parameters set such that $\forall_f q_f(t) \geq 0$, $\sum_f q_f(t) = 1$ and the value of*

$$\min_{f\in\mathcal{S}_0} \sum_{g\in N_f(t)} q_g(t)$$

is maximal (such values can be computed with linear programming). Moreover assume that $\gamma(t)$ parameters are set to

$$\gamma(t) = \frac{\beta}{\min_{f\in\mathcal{S}_0}\sum_{g\in N_f(t)} q_g(t)}.$$

In addition, let $\varepsilon = \min_{t,g}\gamma(t)q_g(t) > 0$. Then the upper bound on the regret of ELP is

$$U_A(T) \leq 9\beta T\alpha(G)\log\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon} + \frac{\log(|\mathcal{S}_0|)}{\beta}.$$

where $\alpha(G)$ is the independence number of the underlying neighbourhood graph G of the strategies in $\mathcal{S}_0$.

PROOF. Following the proof of Theorem 3 in [25][2], we can show that

$$U_A(T) \leq \beta T\alpha G + 2\beta\sum_{t=1}^T \frac{P_f(t)}{\sum_{s\in N_f(t)} P_s(t)} + \frac{\log|\mathcal{S}_0|}{\beta}$$

From Lemma 13 of [2], we can further bound the second term of the RHS as follows:

$$\begin{aligned}\beta\sum_{t=1}^T \frac{P_f(t)}{\sum_{s\in N_f(t)} P_s(t)} &\leq 2\alpha(G)\log\Big(1 + \frac{\frac{|\mathcal{S}_0|^2}{\alpha(G)\varepsilon} + |\mathcal{S}_0| + 1}{\alpha(G)}\Big) \\ &+ 2\alpha(G)\end{aligned}$$

By using elementary algebra, the RHS can be further bounded as:

$$\beta\sum_{t=1}^T \frac{P_f(t)}{\sum_{s\in N_f(t)} P_s(t)} \leq 2\alpha(G)\log\Big(e^2\frac{\frac{|\mathcal{S}_0|^2}{\alpha(G)\varepsilon} + |\mathcal{S}_0| + \alpha(G) + 1}{\alpha(G)}\Big)$$

which can be further bounded with

$$\begin{aligned}\beta\sum_{t=1}^T \frac{P_f(t)}{\sum_{s\in N_f(t)} P_s(t)} &\leq 2\alpha(G)\log\Big(9\frac{4\frac{|\mathcal{S}_0|^2}{\alpha(G)\varepsilon^2}}{\alpha(G)}\Big) \\ &= 2\alpha(G)\log\Big(\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}\Big)\end{aligned}$$

Here we exploit the facts that $1 \leq \alpha(G) \leq |\mathcal{S}_0|$ and $0 < \varepsilon < 1$. This implies that

$$U_A(T) \leq 9\beta T\alpha G\log\Big(\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}\Big) + \frac{\log|\mathcal{S}_0|}{\beta}$$

which concludes the proof. □

Choosing the value of β such that both components of the bound are equal, we get the following:

COROLLARY 3. *For $\beta = \sqrt{\frac{\log(|\mathcal{S}_0|)}{9T\alpha(G)\log\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}}}$ the bound defined in Theorem 2 equals to*

$$U_A(T) \leq 6\sqrt{\alpha(G)\log\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}\log(|\mathcal{S}_0|)T}.$$

Note that here we have $U_A(T) = \tilde{O}(\sqrt{\alpha(G)T})$. However, as $\mathcal{S}_0$ can be arbitrary, in the worst case scenario (when there is no side information at all by playing any particular strategy from $\mathcal{S}_0$), we have $U_A(T) = \tilde{O}(\sqrt{|\mathcal{S}_0|T})$, which can be quite large if the set of possible strategies $\mathcal{S}_0$ is large. A

[2]Note that the original proof of this theorem only considers oblivious opponents. However, by using the argument introduced in [32], we can apply the proof to adaptive adversaries as well.

possible way to improve the bound is to make the underlying neighbourhood graph denser. However, it is not trivial to provide such improvements. For example, assuming that even with a neighbourhood graph with a minimum degree of constant m (*i.e.*, m is independent from $|\mathcal{S}_0|$) would not help much. In fact, we state the following:

THEOREM 4. *There exists $\mathcal{S}_0 \in \mathcal{S}$ with minimum neighbourhood degree of m and a (randomised) opponent strategy, such that if $T > 374|\mathcal{S}_0|^3$, we have $U_A(T) \geq 0.6\sqrt{\frac{|\mathcal{S}_0|}{m+1}T}$ for any meta-strategy A.*

PROOF. From Theorem 4 of [25], we know that under the conditions given in the theorem, the following holds:

$$U_A(T) \geq 0.06\sqrt{\alpha(G)T}$$

Now, we just need to show that there exists $\mathcal{S}_0$ with minimum neighbourhood degree of m but $\alpha(G) \geq \frac{|Str_0|}{m+1}$. To do so, consider a set of groups of $m+1$ strategies such that within each group, there is one chosen strategy and the rest are a prefix of the chosen one. It is clear that the underlying neighbourhood graph G has at least $\frac{|\mathcal{S}_0|}{m+1}$ cliques, which implies that $\alpha(G) \geq \frac{|Str_0|}{m+1}$. □

That is, even in this case, we still have $U_A(T) = \Omega(\sqrt{|\mathcal{S}_0|T})$. Nevertheless, in the following sections, we investigate some special cases of the problem, in which this regret bound can be significantly improved.

4. BIDDING WITH RANDOM NEIGHBOURHOOD GRAPH

We first start with the scenario when some additional information are randomly added to the model. In particular, we assume that apart from the side information observed for prefixes of the chosen strategy, we have access to some additional information that can reveal the potential outcome of strategies other then the prefixes of the chosen one.

4.1 Random Neighbourhood

By random additional information, we mean that there is an oracle in the system, to whom we can send our requests to infer about the outcome of other (not-played) strategies, after observing the outcome of our chosen one. This assumption of having such an oracle is quite common in the online learning literature [42, 25]. Note that we do not consider query costs here in our model. In addition, we also assume that the result of a query is probabilistic, that is, it is revealed with some probability $v > 0$. As such, we can represent the neighbourhood graph of the dollar auction as a random graph $G(\mathcal{S}_0, v)$.

Random neighbourhood can model some additional expert knowledge acquired by the player. For example, during the negotiations (modelled by the dollar auction) some strategies can be considered more aggressive than others (and likely giving similar result), despite not being the prefix of each other (in the strict sense described in the previous sections). Another possible interpretations of the additional information include results of intelligence gathering in the arms race scenario (identifying the maximal number of troops or weapon units that our opponent can use during the conflict) or industrial espionage in the R&D competition scenario (abandoning research on technologies already tested by the competition).

4.2 Regret Analysis

Given the random neighbourhood model, we now turn to the regret analysis of the algorithm. In particular, we add the following conditions to the model:

$$|\mathcal{S}_0| \geq 8000 \quad (1)$$

$$1 \geq v \geq |\mathcal{S}_0|^{-\frac{1}{10}} \quad (2)$$

The first condition is reasonable as we typically consider large sets of strategies. The second condition guarantees that the neighbourhood graph is quite dense in general. We state the following:

THEOREM 5. *Suppose that conditions (1)-(2) hold. Under the same conditions of Theorem 2, with probability at least $1 - o(|\mathcal{S}_0|)$, we have the following regret bound for ELP [3]:*

$$U_A(T) \leq 3|\mathcal{S}_0|^{\frac{1}{4}}\sqrt{2\log\frac{36|\mathcal{S}_0|}{\varepsilon^2}\log(|\mathcal{S}_0|)T}.$$

PROOF. From [16] we have that with at least $1 - o(|\mathcal{S}_0|)$ probability, the following holds:

$$\alpha(G) \leq \frac{2}{v}\Big(\log|\mathcal{S}_0|v - \log\log|\mathcal{S}_0|v - \log 2 + 2\Big)$$

Since $|\mathcal{S}_0| > 8000$ and $v \geq |\mathcal{S}_0|^{-\frac{1}{10}}$, this can further bounded as:

$$\alpha(G) \leq \frac{2}{v}\Big(\log|\mathcal{S}_0|v + 2\Big) \leq 4\frac{\log|\mathcal{S}_0|}{v}$$

Note that we use $v \leq 1$ here. Now, since $v \geq |\mathcal{S}_0|^{-\frac{1}{10}}$, we have that

$$\alpha(G) \leq 4\frac{\log|\mathcal{S}_0|}{|\mathcal{S}_0|^{-\frac{1}{10}}} \leq \sqrt{|\mathcal{S}_0|}$$

The last inequality holds if $|\mathcal{S}_0| \geq 8000$. Applying this to Theorem 2 concludes the proof. □

Note that here we have $U_A(T) = \tilde{O}(|\mathcal{S}_0|^{\frac{1}{4}}\sqrt{T})$, which is a $|\mathcal{S}_0|^{\frac{1}{4}}$ improvement, compared to the regret bound of the general case. In the next section, we will show how this bound can be further improved by taking a different approach of modifying the neighbourhood model.

5. BIDDING WITH THRESHOLD

Beside adding further side information to the model, another way to simplify the underlying structure of the neighbourhood graph, and thus, to improve the induced regret bound, is to restrict the type of behaviour each player can follow. In this spirit, in this section we investigate a class of meta strategies that provides this simplification.

To do so, we first discuss the optimal non-spiteful strategy, proposed by O'Neill, that provides the basis of our class of behaviours. We then describe how to create a class of strategies, which we call optimal strategies with threshold. Finally, we show that by restricting to this class of strategies, we can further improve the previous regret bound with another factor of $|\mathcal{S}_0|^{\frac{1}{4}}$.

[3]Note that the $o(\cdot)$ notation here means that $o(|\mathcal{S}_0|) \to 0$ as $|\mathcal{S}_0|$ tends to infinity.

5.1 The Optimal Non-Spiteful Strategy

We first start with the discussion of the pure strategy in the dollar auction against a non-spiteful opponent that was proposed by O'Neill [31]. In particular, O'Neill proved that this strategy is optimal in terms of maximising the expected utility of the non-spiteful player, when played against a non-spiteful opponent.

Let the initial bid of the strategy be $x_0 = (b-1) mod (s-1)+1$, where b is the budget of both players, s is the stake and 1 is the minimal bid increment. In addition, let $x_i = x_0 + i(s-1)$ for $i > 0$ and let $x_{-1} = 0$. Given this, the optimal strategy of O'Neill $\hat{f}$ for $y \in \{x_{i-1}, \dots, x_i - 1\}$ is:

$$\hat{f}(x,y) = \begin{cases} x_i & \text{if } x = x_{i-1}, \\ y & \text{otherwise.} \end{cases}$$

5.2 Optimal Strategies with Threshold

Based on the optimal strategy of O'Neill, we now propose a special class of strategies, using of which allow us to achieve better regret bound for ELP. In particular, let $\hat{f} \in \mathcal{S}$ be the optimal strategy by O'Neill described in the previous section. We call a strategy $f_\theta \in \mathcal{S}$ a *strategy with threshold* θ if and only if $\forall_{y<\theta} f_\theta(x,y) = \hat{f}(x,y)$ and $\forall_{y\geq\theta} f_\theta(x,y) = y$

Note that a strategy with threshold follows the optimal strategy by O'Neill up to a moment, when opponent's bid exceed threshold and then passes. The underlying intuition of this class of strategies is that in many cases, the bidder is rational (*i.e.*, bidding optimally), but is also risk-aware. To model the latter, we assume that the bidder is only willing to bid up to a certain threshold value, although her bidding budget would allow her to go beyond that threshold. As such, we argue that this class of strategies with threshold represents a realistic behaviour in many real-world situations.

5.3 Regret Analysis

We now turn to the regret analysis of ELP for the case of bidding with threshold. In particular, we assume that we only have access to the above defined class of non-spiteful strategies with threshold.

THEOREM 6. *Assume that set $\mathcal{S}_0$ contains all strategies with threshold $\theta \in \{0, \dots, b\}$ and no other strategies. Under the abovementioned conditions and the settings of Theorem 2, the regret bound of the ELP algorithm is at most:*

$$U_A(T) \leq 9\beta T \log \frac{6|\mathcal{S}_0|}{\varepsilon} + \frac{\log(|\mathcal{S}_0|)}{\beta}.$$

PROOF. From Theorem 2 we have:

$$U_A(T) \leq 9\beta T \alpha G \log\left(\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}\right) + \frac{\log|\mathcal{S}_0|}{\beta}$$

where $\alpha(G)$ is the independence number of the information graph G.

Given two strategies with threshold f_θ and f_Θ, where $\theta < \Theta$, there is always an edge $f_\Theta \to f_\theta$ in the information graph for all rounds. If auction was lost while using f_Θ, then auction would have also been lost using f_θ and the result would be equal to $-x$, where x is a last bid of f_θ. If auction was won while using f_Θ, we know the last bid of opponent's strategy y. Since f_θ and f_Θ follow the same sequence of bids, then for $\theta \leq y$ auction would have been lost with the result of $-x$ (x being the last bid made by f_θ) and for $\theta > y$ auction would have been won with the result of $s - x$ (x being the lowest bid made by f_θ higher than y). Therefore, for the setting describe in the theorem, G is a clique and $\alpha(G) = 1$, which concludes the proof. □

Again, choosing the value of β such that both components of the bound are equal, we get the following:

COROLLARY 7. *For $\beta = \sqrt{\frac{\log(|\mathcal{S}_0|)}{9T\log\frac{6|\mathcal{S}_0|}{\varepsilon}}}$ the bound defined in Theorem 6 equals:*

$$U_A(T) \leq 6\sqrt{\log\frac{6|\mathcal{S}_0|}{\varepsilon}\log(|\mathcal{S}_0|)T}.$$

Note that $U_A(T) = \tilde{O}(\sqrt{T})$, which is an improvement of a factor of $|\mathcal{S}_0|^{\frac{1}{2}}$, compared to the regret bound of the general case.

6. AUCTION AGAINST MALICIOUS OPPONENT

So far we have shown that by applying ELP, we can still achieve low regret performance, despite the fact that there is no prior knowledge about the opponent's motives, neither about her rationality. We now introduce the concept of spitefulness and consider an auction against a malicious opponent. As argued by Waniek *et al.* [39], there are various real-world applications, in which the dollar auctions are played between *spiteful* players, whose objective function involves some linear combination of maximising one's own utility and minimising the opponent's utility. For instance, one may argue that certain moves of the USA or the Soviet Union during the Cold War were designed to harm the opponent even if it meant incurring some extra cost.

6.1 The Concept of Spitefulness

We interpret *spitefulness* as the desire of a player to hurt her opponent. We follow the definition of spitefulness introduced by Brandt *et al.* [9].

Any spiteful player i is characterised by a *spitefulness coefficient* $\sigma_i \in [0,1]$. The higher the coefficient, the more the player i is interested in minimizing the profit of her opponent j. The utility function of a spiteful player is then:

$$u_i = (1-\sigma_i)p_i - \sigma_i p_j.$$

where p_i and p_j are the profits of corresponding players.

A player with the spitefulness coefficient $\sigma = 0$ is called a *non-spiteful* player, and she is only interested in maximizing her own profit. On the other hand, a player with spitefulness coefficient $\sigma = 1$ is called a *malicious* player, and she is only interested in minimizing the profit of her opponent.

6.2 Mixed Nash Equilibrium

We now consider a case when a non-spiteful bidder plays the repeated dollar auctions against a malicious opponent, not knowing that the opponent is malicious. In addition, our opponent might not have full information about our non-spitefulness either and might not follow a rational behaviour. Given this, we assume that her strategy is taken from a set of $\mathcal{S}_1 \subseteq \mathcal{S}$.

If we knew the set $\mathcal{S}_1$ in advance (and the opponent also knows our set of strategies $\mathcal{S}_0$), a mixed Nash equilibrium can be easily calculated using standard arguments. However,

as this knowledge is typically not available in advance, it is impossible to calculate the Nash equilibrium. Nevertheless, we show that if both players apply the ELP algorithm to play their strategies, we can converge to a Nash equilibrium point in performance.

THEOREM 8. *Let $f_i(t) \in \mathcal{S}_i, i \in \{0,1\}$ denote the bidding strategy of player i at round t, and we denote by $f_j(t)$ the strategy chosen by i's opponent. Let*

$$\tilde{f}_i = \frac{1}{T}\sum_{t=1}^{T} f_i(t),$$

$$\tilde{f}_j = \frac{1}{T}\sum_{t=1}^{T} f_j(t)$$

mixed strategies denote the average of these chosen strategies. Given this, for each T, there exists a mixed Nash equilibrium profile (f_i^*, f_j^*) such that

$$\left| r_i(f_i^*, f_j^*) - r_i(\tilde{f}_i, \tilde{f}_j) \right| \leq \tilde{O}\Big(\sqrt{\frac{|\mathcal{S}_0|}{T}}\Big)$$

PROOF. For the sake of simplicity we now map the profit of a player i to reward $r_i \in [-1,1]$ (instead of $r_i \in [0,1]$ as before). Reward $r_i = 1$ of the non-spiteful player i corresponds to a profit equal to the budget $p_i = b$, and reward $r_i = -1$ of the non-spiteful player corresponds to a loss equal to the budget $p_i = -b$. Reward of the malicious player is expressed as utility described in the previous section, *i.e.*, reward $r_i = 1$ of the malicious player i corresponds to a loss of the non-spiteful player j equal to the budget $p_j = -b$, and reward $r_i = -1$ of the malicious player i corresponds to a profit of the non-spiteful player j equal to the budget $p_j = b$.

Given that assumption, note that when a non-spiteful player bids against a malicious one, our problem can be regarded as a repeated zero-sum game. In fact, within this game, each player i (with $i \in \{0,1\}$) chooses from set of bidding strategies $\mathcal{S}_i$, the received reward of player i is r_i, and the received reward of her opponent (due to the malicious behaviour) is $r_j = -r_i$.

It is well known that in a 2-player zero-sum game, the minimax value exists and does coincide with the (mixed) Nash equilibrium. Given this, we only need to show that the average behaviour of the players, when both apply ELP, converges in performance to the minimax solution. To do so, let u^* denote the minimax value, which can be defined as follows:

$$u^* = \max_{f_i}\min_{f_j} r_i(f_i, f_j) = \min_{f_j}\max_{f_i} r_i(f_i, f_j)$$

where f_i and f_j are mixed strategies of players i and j, respectively. For now we consider the case that player i (non-spiteful) uses ELP, and player j can play any meta-strategy $\{f_j(t)\}_{t=1}^{T}$. In addition, let $BR(f_i(t))$ denote the best response mixed strategy against player i's $f_i(t)$ bidding strategy. By denoting

$$\Delta_T = 6\sqrt{\frac{\alpha(G)\log\frac{6|\mathcal{S}_0|}{\alpha(G)\varepsilon}\log(|\mathcal{S}_0|)}{T}},$$

from Theorem 2 we have that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} r_i(t) &\geq \frac{1}{T}\max_{f\in\mathcal{S}_i}\sum_t r_i(f, f_j(t)) - \Delta_T \\ &= \frac{1}{T}\max_{\mathbf{f}}\sum_t r_i(\mathbf{f}, f_j(t)) - \Delta_T\end{aligned}$$

where $\mathbf{f}$ is a mixed strategy over $\mathcal{S}_i$. The first inequality is obtained from Theorem 2, while the second one is from the fact that the mixed strategy belongs to the convex hull of the pure strategies, and thus, cannot exceed the best pure strategy. This can be further bounded as follows:

$$\frac{1}{T}\max_{\mathbf{f}}\sum_t r_i(\mathbf{f}, f_j(t)) - \Delta_T \geq \frac{1}{T}\max_{\mathbf{f}}\sum_t \min_{\mathbf{g}} r_i(\mathbf{f}, \mathbf{g}) - \Delta_T$$

where $\mathbf{g}$ is a mixed strategy over $\mathcal{S}_j$. This can be further bounded as:

$$\begin{aligned}\frac{1}{T}\max_{\mathbf{f}}\sum_t \min_{\mathbf{g}} r_i(\mathbf{f}, \mathbf{g}) &\geq \max_{\mathbf{f}}\min_{\mathbf{g}} r_i(\mathbf{f}, \mathbf{g}) - \Delta_T \\ &= u^* - \Delta_T\end{aligned}$$

That is, we have

$$\frac{1}{T}\sum_{t=1}^{T} r_i(t) \geq u^* - \Delta_T$$

Similarly, we can prove (from the malicious player's perspective) that

$$\frac{1}{T}\sum_{t=1}^{T} r_i(t) \leq u^* + \Delta_T$$

which concludes the proof. □

Note that this convergence in performance is a weaker property, compared to the convergence in strategy (*i.e.*, the average behaviour of the players converge to the distribution of the mixed strategy in the equilibrium point), as the latter automatically implies the former, while the other way is not true in general.

7. RELATED WORK

In this section we describe the bodies of literature concerning important aspects of our setting: all-pay auctions, repeated auctions and multi-armed bandit models.

7.1 All-Pay Auctions and the Dollar Auction

All-pay auctions were studied as a model multiple settings where expenditure of the resource does not always guarantee profit [24]. Some of them include political campaigns, R&D competitions of oligopolistic companies [38], realizing crowdsourcing enterprises [14] and international situation analysis [34]. Aside from theoretical studies, there is also experimental research [12]. The dollar auction is a widely studied type of an all-pay auction[23, 22]. It is often used in a classroom experiments serving a study of conflict escalation among people [29, 19].

Recently, Waniek *et al.* [39] offered a preliminary study of a dollar auction in which a spiteful player plays against a rational opponent. The authors considered the setting in which a non-spiteful bidder unwittingly bids against a spiteful one. In various scenarios in this setting the conflict escalates. In particular, the spiteful bidder is able to force the

non-spiteful opponent to spend most of the budget. Still, it is the spiteful bidder who often wins the prize. Furthermore, a malicious player with a smaller budget is likely to plunge the opponent more than a malicious player with a bigger budget. Thus, a malicious player should not only hide his real preferences but also the real size of his budget.

7.2 Repeated Auctions

Modelling various processes as repeated auctions is a well established idea [41]. Among other applications, it was used for allocating tasks to robots [30] and assigning channels in radio network [17]. Many authors concentrate on the problem of possible collusion in repeated auctions [3, 35, 26], focusing their attention on the effect of cooperation between players on the income of an auctioneer [7]. Most of the published results consider second-price auctions [20, 7], however all-pay auction format was also studied [28].

7.3 Multi-Armed Bandits

The first multi-armed bandit models were designed to tackle problems with stochastic rewards [33, 1, 4]. The first adversarial setting was proposed by Auer *et al.* [5, 6]. However, in the early years, adversarial bandit models were believed to be only suitable against non-adaptive opponents. The argument that justified the usage of multi-armed bandits against adaptive opponents was first proposed by Poland [32], whose techniques was adopted by many more recent works. Mannor and Shamir[25] extended the adversarial bandit model to the case when additional side information is provided to improve the regret bounds of the model. Their work was later improved by [2] and [21]. In particular, the former improved the regret bound, while the latter focuses on the cases when the neighbourhood graph cannot be fully revealed in advance (*i.e.*, before the chosen arm is played).

8. CONCLUSIONS

In this paper, we investigated the problem of repeated dollar auctions in which players do not have prior knowledge about the opponent's spitefulness, neither her rationality level. We proposed an adversarial multi-armed bandit model to efficiently tackle this problem, in which the goal is to maximise the players' total utility over time. We showed that by using an adversarial bandit based meta-strategy, ELP, we can indeed provably achieve good performance. We then further improved this performance by considering additional settings of the model: (i) dollar auctions with random neighbourhood graphs; and (ii) playing optimal strategies with thresholds. We also proved that in a special case of non-spiteful player *versus* malicious player, if both use ELP to make their decisions, the game converges to a mixed Nash equilibrium in performance.

As a potential future work, we aim to extend our work to repeated auctions where the auctions share the same budget. That is, apart from making decision about which strategy to use per each auction, the players also have to choose the budget allocated to a particular auction, without exceeding an *a priori* given global limit. As our current techniques significantly rely on the fact that the current budgets are given, it seems to be highly not trivial how to extend our model to such settings. Furthermore, until now, the literature on the dollar auction focused on the case of two players. Hence, it would be very interesting to extend all the results to multiple players.

Acknowledgments

Tran-Thanh gratefully acknowledges funding from the UK Research Council for project "ORCHID", grant EP/I011587/1. Tomasz Michalak was supported by the European Research Council under Advanced Grant 291528 ("RACE").

APPENDIX

Table 1: Notation

Symbol	Meaning
i,j	Players participating in the auction
s	Stake of the auction
δ	Minimal bid increment
b	Budget of both players
X_b	States where player can make a bid
$\mathcal{S}$	Set of all valid strategies in the auction
$\mathcal{S}_0$	Set of available strategies
p_i	Profit of player i
T	Number of rounds (separate auctions)
A	Algorithm selecting strategies
$R_A(T)$	Total reward of algorithm A after T rounds
$U_A(T)$	Regret of algorithm A after T rounds
$r_f(t)$	Reward for using strategy f in round t
G	Neighbourhood graph
$\alpha(G)$	Independence number of graph G
N_f	Neighbourhood of strategy f
f_t	Strategy selected to use in round t
β,$\gamma(t)$,$q_f(t)$	Parameters of the ELP algorithm
$w_f(t)$	Weight of choosing strategy f in round t
$P_g(t)$	Probability of selecting strategy f in round t
$\hat{r}_f(t)$	Estimated reward for using str. f in round t
v	Prob. of revealing additional information
$G(\mathcal{S}_0, v)$	Random neighbourhood graph of strat. $\mathcal{S}_0$
θ	Threshold of a strategy
σ_i	Spitefulness coefficient of player i

REFERENCES

[1] R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.

[2] N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2013.

[3] M. Aoyagi. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory*, 112(1):79–105, 2003.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite–time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[7] S. Bikhchandani. Reputation in repeated second-price auctions. *Journal of Economic Theory*, 46(1):97–119, 1988.

[8] K. E. Boulding. Conflict and defense: A general theory. 1962.

[9] F. Brandt, T. Sandholm, and Y. Shoham. Spiteful bidding in sealed-bid auctions. In *Computing and Markets*, 2005.

[10] J. Campos, C. Martinho, and A. Paiva. Conflict inside out: A theoretical approach to conflict from an agent point of view. AAMAS '13, pages 761–768. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[11] C. Castelfranchi. Conflict ontology. In *Computational Conflicts*, pages 21–40. Springer, 2000.

[12] E. Dechenaux, D. Kovenock, and R. M. Sheremeta. A survey of experimental research on contests, all-pay auctions and tournaments. WZB Discussion Paper SP II 2012-109, Berlin, 2012.

[13] G. Demange. Rational escalation. *Annales d'ÃL'conomie et de Statistique*, (25/26):pp. 227–249, 1992.

[14] D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *EC'09*, pages 119–128. ACM, 2009.

[15] H. Fang. Lottery versus all-pay auction models of lobbying. *Public Choice*, 112(3-4):351–371, 2002.

[16] A. M. Frieze. On the independence number of random graphs. *Discrete Mathematics*, 81(2):171–175, 1990.

[17] Z. Han, R. Zheng, and H. V. Poor. Repeated auctions with bayesian nonparametric learning for spectrum access in cognitive radio networks. *Wireless Communications, IEEE Transactions on*, 10(3):890–900, 2011.

[18] A. J. Jones. *Game theory: Mathematical models of conflict*. Elsevier, 2000.

[19] J. H. Kagel and D. Levin. Auctions: a survey of experimental research, 1995–2008. 2008.

[20] J. L. Knetsch, F.-F. Tang, and R. H. Thaler. The endowment effect and repeated market trials: Is the vickrey auction demand revealing? *Experimental economics*, 4(3):257–269, 2001.

[21] T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pages 613–621, 2014.

[22] V. Krishna and J. Morgan. An analysis of the war of attrition and the all-pay auction. *journal of economic theory*, 72(2):343–362, 1997.

[23] W. Leininger. Escalation and cooperation in conflict situations the dollar auction revisited. *Journal of Conflict Resolution*, 33(2):231–254, 1989.

[24] Y. Lewenberg, O. Lev, Y. Bachrach, and J. S. Rosenschein. Agent failures in all-pay auctions. IJCAI/AAAI, 2013.

[25] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

[26] R. C. Marshall and L. M. Marx. Bidder collusion. *Journal of Economic Theory*, 133(1):374–402, 2007.

[27] H. J. Müller and R. Dieng. On conflicts in general and their use in ai in particular. In *Computational conflicts*, pages 1–20. Springer, 2000.

[28] J. Münster. Repeated contests with asymmetric information. *Journal of Public Economic Theory*, 11(1):89–118, 2009.

[29] J. K. Murnighan. A very extreme case of the dollar auction. *Journal of Management Education*, 26(1):56–69, 2002.

[30] M. Nanjanath and M. Gini. Repeated auctions for robust task execution by a robot team. *Robotics and Autonomous Systems*, 58(7):900–909, 2010.

[31] B. O'Neill. International escalation and the dollar auction. *Journal of Conflict Resolution*, 30(1):33–50, 1986.

[32] J. Poland. Fpl analysis for adaptive bandits. *3rd Symposium on Stochastic Algorithms, Foundations and Applications (SAGA'05)*, pages 58–69, 2005.

[33] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the AMS*, 55:527–535, 1952.

[34] M. Shubik. The dollar auction game: A paradox in noncooperative behavior and escalation. *Journal of Conflict Resolution*, pages 109–111, 1971.

[35] A. Skrzypacz and H. Hopenhayn. Tacit collusion in repeated auctions. *Journal of Economic Theory*, 114(1):153–169, 2004.

[36] C. Tessier, L. Chaudron, and H. J. MÃijller. Agents' conflicts: New issues. In *Conflicting Agents*, volume 1 of *Multiagent Systems, Artificial Societies, And Simulated Organizations*, pages 1–30. Springer US, 2002.

[37] W. W. Vasconcelos, M. J. Kollingbaum, and T. J. Norman. Normative conflict resolution in multi-agent systems. *AAMAS'09*, 19(2):124–152, 2009.

[38] X. Vives. Technological competition, uncertainty, and oligopoly. *Journal of Economic Theory*, 48(2):386–415, 1989.

[39] M. Waniek, A. Niescieruk, T. Michalak, and T. Rahwan. Spiteful bidding in the dollar auction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 667–673. AAAI Press, 2015.

[40] D. Weyns and T. Holvoet. Regional synchronization for simultaneous actions in situated multi-agent systems. pages 497–510, 2003.

[41] E. Wolfstetter. Auctions: an introduction. *Journal of economic surveys*, 10(4):367–420, 1996.

[42] J. Y. Yu and S. Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184. ACM, 2009.