

# Applying Opponent Modeling for Automatic Bidding in Online Repeated Auctions

Yudong Hu

University of Chinese Academy of Sciences  
Beijing, China  
huyudong201@mailsucas.ac.cn

Tiande Guo

University of Chinese Academy of Sciences  
Beijing, China  
tdguo@uacs.ac.cn

Congying Han\*

University of Chinese Academy of Sciences  
Beijing, China  
hancy@ucas.ac.cn

Hao Xiao

Institute of Electrical Engineering, Chinese Academy of Sciences  
Beijing, China  
xiaohao09@mail.iee.ac.cn

## ABSTRACT

Online auction scenarios, such as bidding searches on advertising platforms, often require bidders to participate repeatedly in auctions for identical or similar items. Most previous studies have only considered the process by which the seller learns the prior-dependent optimal mechanism in a repeated auction. However, in this paper, we define a multiagent reinforcement learning environment in which strategic bidders and the seller learn their strategies simultaneously and design an automatic bidding algorithm that updates the strategy of bidders through online interactions. We propose Bid Net to replace the linear shading function as a representation of the strategic bidders' strategy, which effectively improves the utility of strategy learned by bidders. We apply and revise the opponent modeling methods to design the PG (pseudo-gradient) algorithm, which allows bidders to learn optimal bidding strategies with predictions of the other agents' strategy transition. We prove that when a bidder uses the PG algorithm, it can learn the best response to static opponents. When all bidders adopt the PG algorithm, the system will converge to the equilibrium of the game induced by the auction. In experiments with diverse environmental settings and varying opponent strategies, the PG algorithm maximizes the utility of bidders. We hope that this article will inspire research on automatic bidding strategies for strategic bidders.

## KEYWORDS

Auction Theory; Strategic Bidder; Opponent Modeling; Multiagent Reinforcement Learning

### ACM Reference Format:

Yudong Hu, Congying Han, Tiande Guo, and Hao Xiao. 2024. Applying Opponent Modeling for Automatic Bidding in Online Repeated Auctions. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, Auckland, New Zealand, May 6 – 10, 2024, IFAAMAS, 9 pages.

\*Corresponding author.



This work is licensed under a Creative Commons Attribution International 4.0 License.

*Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

## 1 INTRODUCTION

The rapid development of electronic commerce has generated a significant demand for online auctions [35]. Employing appropriate mechanisms to facilitate the sale of items, such as advertisements on electronic platforms, can effectively improve revenue [1, 5, 10, 29]. Among the most prevalent mechanisms in online auctions are the VCG auction (Vickrey-Clarke-Groves) [32, 33] and the GSP auction (Generalized Second Price) [10, 12]. Both of these auctions are instances of the prior-independent auction mechanism, which can be executed efficiently, but does not maximize revenue.

With the advent of deep reinforcement learning, it has become feasible to learn optimal auctions and maximize seller revenue [14, 17]. Deep neural network techniques, such as Myerson Net [9], ALGNet [27], MenuNet [30], and Deep GSP [38], have been applied to derive optimal mechanisms based on bidding data, demonstrating efficiency even in multi-bidder, multi-item auction scenarios. These methods incorporate prior-dependence of bidders' value distribution to the mechanism through the utilization of bidding data sourced from online auctions. By creating mechanisms that adhere to the principle of Incentive Compatibility (IC), the seller can collect truthful bidding data from online auctions and employ them to estimate the value distribution among bidders [3].

Unlike the seller's perspective, recent studies have introduced strategic bidding approaches in online auctions [25]. These methods exploit the process in which the seller learns the value distribution of the bidders using bid samples [4, 8] and design strategic bidders to increase the utility with specific bidding strategies. The distribution reporting model [31] states that in incentive compatibility auctions, bidders can improve their utility by bidding according to a specific "fake distribution". When the seller employs the Myerson Net learning mechanism, adversarial learning strategies maximize the utility of strategic bidders [24], particularly when other participants in the system adopt truthful bidding strategies [26].

The primary aim of this paper is to formulate an approach that enables bidders to automatically learn bidding strategies. Our approach diverges from previous research on strategic bidding, such as ZIP (Zero Intelligence Plus) [2, 7], which focused on the learning process of bidders through repeated interactions within static auction mechanisms. In our work, we depict the repeated auction environment as a multiagent reinforcement learning system, where both the seller and individual bidders are treated as agents. Each

agent employs independent algorithms to acquire strategies. Although bidders engage in repeated auctions for the same item, it is important to note that the value distribution and strategies of agents can evolve over time. This requires bidders to predict changes in the mechanism induced by bidding strategies and learn how to bid in response to a dynamic mechanism.

Opponent modeling research offers robust approaches to strategy learning within the aforementioned system, where all agents independently update their strategies with strategy prediction [36]. By anticipating transitions in the opponent’s strategy, an agent can acquire strategies that yield superior rewards in repeated games. In particular, LOLA (Learning with Opponent-Learning Awareness) [11] is the first algorithm capable of achieving cooperation in the Prisoner’s Dilemma game through independent opponent modeling processes for each agent. Subsequent research, COLA (Consistency in Opponent-Learning Awareness) [34], states that the system will reach the target equilibrium when the predictions of the opponent’s strategy (lookahead rate) maintain consistency.

In this paper, we provide a complete definition of automatic bidding for bidders in online auctions and design algorithms with theoretical support. Our contributions are summarized as follows: 1) We model the behavior of the seller and bidders in repeated auctions as an induced game in which all agents learn strategies in repeated interactions. We adopt Myerson Net as the seller strategy and design an automatic bidding strategy for bidders. Our objective is that this MARL system converges to an equilibrium in which bidders have optimal utility. 2) We propose Bid Net to represent the strategies learned by the bidders. The network structure of the Bid Net has the advantage of satisfying the IR (individual rationality) and accurate gradient propagation with NeuralSort. We illustrate that the Bid Net is an efficient improvement of the simple linear strategy through experiment. 3) We design the PG (pseudo-gradient) algorithm based on the opponent modeling method. The PG algorithm is an automatic bidding method for updating the bidding strategy based on the prediction of changes in the parameters of the auction mechanism. We prove that when a bidder employs the PG algorithm, it can learn the optimal bidding strategy under the current prior-dependent mechanism. When all bidders adopt the PG algorithm, the system will converge to the Nash equilibrium of the induced game.

In our experiments, we compare the Bid Net with the previously used linear shading bidding strategy [24]. The results show that Bid Net can learn utility-maximizing bidding strategies, even in environments where the parameters of the seller’s mechanism are constantly updated through learning. In an experiment in which all bidders use the same algorithm learning strategy, the results show that only the PG algorithm can stably learn the target equilibrium strategy of the induced game and maximize the average utility of the bidders. To illustrate the effectiveness of our automatic bidding algorithm, we test it in different environmental settings and with various opponent strategies. The PG algorithm achieves higher utility when other bidders use a static or dynamic strategy and can be applied to arbitrary environmental parameter settings. These experiments illustrate the applicability of our automatic bidding algorithm to online repeated auctions in terms of effectiveness and generalizability.

## 2 NOTATION AND BACKGROUND

In this paper, we focus on single-item auctions as defined in [15]. The values of the  $n$  bidders are drawn from their value distributions  $v_i \sim F_i$ .  $F = F_1 \times \dots \times F_n$  is the joint distribution. For each bidder  $i$ , its bidding function is denoted as  $B_i$ , with the actual bid being  $b_i = B_i(v_i)$ . We use  $\pi_i$  to denote the strategy of bidder  $i$ , when it comes from the learnable parameters. An example of a bidding strategy is the linear shading strategy, expressed as  $B_i(v_i) = \alpha_i \cdot v_i$ . The truthful bidding strategy  $B_i(v_i) = v_i$  is the simplest strategy.

We use  $M$  to represent the seller’s mechanism, which comprises both the allocation rule  $\vec{a}$  and the payment rule  $\vec{p}$ . The seller receives joint bids  $\vec{b} = (b_1, \dots, b_n)$  and generates allocation  $\vec{a}(\vec{b}) = (a_1(\vec{b}), \dots, a_n(\vec{b}))$  and payment  $\vec{p}(\vec{b}) = (p_1(\vec{b}), \dots, p_n(\vec{b}))$  according to its mechanism  $(\vec{a}, \vec{p}) = M(\vec{b}, \theta)$ . Here,  $\theta$  signifies the strategy parameter for the seller. In the context of the traditional first-price auction, the item is allocated to the bidder with the highest bid, denoted as  $(a_i = 1 \text{ if } b_i = \max(\vec{b}), \text{ otherwise } 0)$ , and the payment equals the bid of the winning bidder, indicated as  $(p_i = b_i \text{ if } a_i = 1, \text{ otherwise } 0)$ .

### 2.1 Traditional Prior-dependent Auction

The main results of auction theory have been consolidated in [20]. In addition, [22] discussed the representation and reasoning processes associated with auctions. For prior-independent mechanisms, such as first-price and second-price auctions, the seller’s revenue is the same when the bidders respond optimally [23]. This has inspired research on prior-dependent auctions with IC constraint, where the truthful bidding strategy is the dominant strategy for bidders.

The well-known revenue-maximizing mechanism is the Myerson auction [23], which is the optimal mechanism that satisfies the IC constraint under the premise that the value of the bidders is public information. In the Myerson auction, each bidder’s value  $v_i$  is transformed into a virtual value  $w_i = g_i(v_i)$ , after which a second-price auction with a reserve price of 0 is conducted using the set of virtual values  $\vec{w}$ . The virtual value function  $g_i$  is determined by the distribution  $F_i$  and the density function  $f_i$  associated with  $v_i$ .

$$g_i(v_i) = v_i - \frac{1 - F_i(v_i)}{f_i(v_i)}.$$

In fact, the virtual value function  $g_i$  plays a key role in determining both the allocation and payment of the mechanism.

At this point, it is assumed that the joint value distribution of the other bidders is represented as  $F_{-i}$ . Consequently, the utility of the bidder  $i$  can be expressed as:

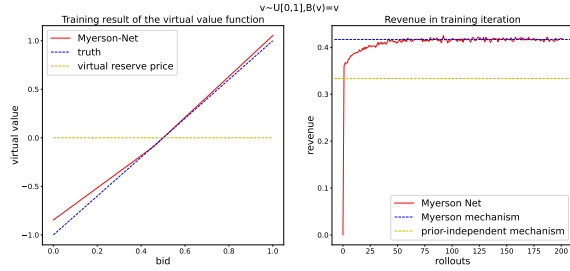
$$U_i(\vec{a}, \vec{p}, v_i) = \int_{F_{-i}} [a_i(b_i, B_{-i}(v_{-i}))v_i - p_i(b_i, B_{-i}(v_{-i}))] f_{-i}(v_{-i}) dv_{-i}.$$

Here,  $dv_{-i} = dv_1 dv_2 \dots dv_{i-1} dv_{i+1} \dots dv_n$  and  $b_i = B_i(v_i)$ . In the context of single-item auctions, it is necessary for the allocation to be 0 or 1. As a result, the above equation can be reformulated as:

$$U_i(\vec{a}, \vec{p}, v_i) = \int_{F_{-i}} \mathbb{I}(a_i = 1) [v_i - p_i(b_i, B_{-i}(v_{-i}))] f_{-i}(v_{-i}) dv_{-i},$$

Here,  $\mathbb{I}$  denotes the indicator function, and  $a_i = a_i(b_i, B_{-i}(v_{-i}))$ . The seller’s revenue can be defined as:

$$R(\vec{a}, \vec{p}, F) = \int_F \sum_{i=1}^n p_i(\vec{b}) dv_1 \dots dv_n.$$



**Figure 1: The learned virtual value function after 200 iterations of Myerson-Net in single-item, two-bidder auction. The line labeled truth indicates that it is derived from Myerson’s Lemma, and the line labeled Myerson-Net is from the output of the network. The revenue curve represents the revenue of the mechanism in the learning process.**

The Myerson mechanism has been proven to satisfy the IC constraint, indicating that the truthful bidding strategy  $B_i(v_i) = v_i$  is the optimal response to the Myerson mechanism [23].

We consider the standard setting as an example. There are two bidders and their value distribution is identical, represented as  $v_i \sim U[0, 1]$ . If the seller employs the Myerson mechanism, the virtual value function is  $g_i(v_i) = 2v_i - 1$ . In response, both bidders choose to adopt the truthful bidding strategy as their best response to the Myerson mechanism. The expected utility of each bidder is  $\frac{1}{12}$  and the expected revenue of the seller is  $\frac{5}{12}$ . However, if the seller adopts the first-price or second-price auction instead, the revenue generated is only  $\frac{1}{3}$ .

## 2.2 Online Repeated Auction and the Myerson Net

In this paper, online auction means the auctions held in online marketplaces, such as bidding searches on advertising platforms [6]. In this context, bidders frequently participate in multiple auctions for identical or similar items. This circumstance presents opportunities to implement prior-dependent mechanisms, even when access to the value distribution of bidders is limited. The seller can establish a mechanism that adheres to the IC constraints, requiring bidders to employ a truthful bidding strategy. By collecting bid samples from repeated auctions, the seller can estimate the value distribution of the bidders, denoted as  $\vec{F}$ , and subsequently adjust the mechanism parameters to converge towards the optimal Myerson auction.

Myerson Net [9] presents a technique that leverages a monotonically increasing parameterized neural network  $G$  to acquire virtual value functions  $\vec{g}$  and derive optimal auction mechanisms.

**Definition 1** (Myerson Net). *For joint bids  $\vec{b} = (b_1, b_2, \dots, b_n)$  from bidders  $i = 1 \dots n$ , the seller first transforms each bid into a virtual value  $w_i = G_i(b_i, \theta)$  ( $\theta$  is the network parameter). The item is then assigned to the bidder with the highest positive virtual value. Payment is the minimum bid required for the winner to win:*

$$p_i = \mathbb{I}(a_i = 1)b_i^*, \quad b_i^* = \underset{b_i}{\operatorname{argmin}}[G_i(b_i, \theta) = \max_{j \in 1 \dots n} G_j(b_j, \theta) > 0].$$

From the definition we can obtain the following property.

**THEOREM 2.1.** *Myerson Net satisfies the IC constraints in single-item auctions. For a Myerson Net  $M$ , we add a strictly increasing*

*function  $Q$  to each virtual value function  $G'_i = Q \circ G_i$  that satisfies  $G_i(b_i) = 0 \iff Q \circ G_i(b_i) = 0$  to obtain another mechanism  $M'$ . Then  $M'$  and  $M$  have the same allocation and payment rule.*

The proof of Theorem 2.1 is given in the supplementary material. We use  $\cong$  to represent the two different networks with the same allocation and payment mechanism. Since the Myerson Net satisfies the IC constraints, we can assume that the bidders will adopt the truthful bidding strategy. Using the real bids obtained in repeated auctions, the seller updates the network parameter  $\theta$  with the objective of maximizing revenue  $r$ , which results in the convergence of the virtual value function  $G$  to the optimal function  $\vec{g}$  corresponding to the true value distribution of the bidders. This suggests that the system converges to the revenue-maximizing Myerson mechanism when bidders bid truthfully in this environment.

Figure 1 provides an example of the application in the standard setting. The figure on the left shows that the seller has learned the optimal virtual value function in 200 iterations. The figure on the right shows the change in revenue in the learning iterations. We can see that the seller using Myerson Net obtains optimal revenue, which exceeds the prior-independent mechanism.

## 2.3 Strategic Bidder and the Induced Game

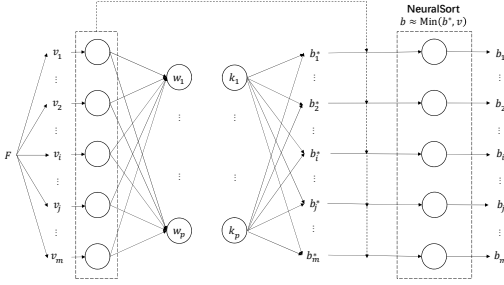
Myerson Net provides an approach as a revenue-maximizing mechanism for online repeated auctions. Since Myerson Net satisfies the incentive compatibility constraint, the optimal strategy for bidders in a single auction round is the truthful bidding strategy. However, some previous studies have found that bidders in repeated auctions have more efficient strategies, which requires them to bid according to some specific distribution instead of true value distribution. Although bidders initially lose some utility under the IC-constrained mechanism, untruthful bids can lead the seller to misestimate the value distribution and increase bidders’ long-term utility.

Tang’s research [31] defines the induced game of auction when the seller adjusts the mechanism according to the bids. As an example, the Myerson mechanism induces a game in which the players contain only bidders. In this game, bidders adopt specific bidding strategies rather than the truthful strategy, and their utility is derived from the Myerson mechanism that runs based on bids.

**Definition 2** (Induced game of Myerson mechanism  $M$ ). *The induced game is represented as  $(N, A, \vec{U})$ , where  $N = \{1, \dots, n\}$  is the set of bidders,  $A = \{B_1, \dots, B_n\}$  is the bidding function set, and  $\vec{U} = (U_1, \dots, U_n)$  is the utility function. Given the joint action  $(B_1, \dots, B_n)$ , the utility is derived by applying the Myerson mechanism  $M$  with the assumption  $B_i(v_i) = v_i$ .*

We present the partial payment matrix for the induced game of the single-item, two-bidder auction in the supplementary material and we have the following property.

**THEOREM 2.2.** *We assume that there are two bidders and their value distribution is  $U[0, 1]$  (standard setting), the induced game is as detailed in Definition 2. For strategic bidders with a limited strategy space  $B(v) = \alpha v$  (referred to as the linear shading strategy space), the Nash equilibrium of the induced game yields  $B_i(v_i) = \frac{5}{14}v_i$ . However, in cases where bidders have access to arbitrary monotone increasing strategies, a Nash equilibrium emerges with bidding strategies represented as  $B_i(v_i) = \frac{v_i+1}{4}$ .*



**Figure 2: The network for strategic bidder (Bid Net), which takes the value of the bidder as input. NeuralSort is a differentiable sorting operator that can output an approximate sorted sequence while preserving the gradient.**

For the two bidders in the standard setting, the expected utility in each round of the auction when they employ a truthful bidding strategy is  $\frac{1}{12}$ . When their bidding strategy is  $B_i(v_i) = \frac{v_i+1}{4}$ , the expected utility is  $\frac{1}{6}$ . This suggests that strategic bidders can increase utility through specific bidding strategies in the induced game. The proof of Theorem 2.2 is given in the supplementary material.

### 3 MULTIAGENT REINFORCEMENT LEARNING BASED AUTOMATIC BIDDING METHOD

To analyze the equilibrium and learning process of the induced game in Section 2.3, we introduce the environment where the seller and strategic bidders learn strategies in a repeated auction for maximum reward. We assume that the value distribution of the bidders is unavailable to the seller. Bidders and the seller will update their strategies based on the results of each auction round and the corresponding rewards. The above single-item auction is repeated until the strategies of all agents converge.

#### 3.1 Bid Net

From Theorem 2.2 we can see that bidders learn different equilibria when their strategy space is limited. Since an optimal mechanism can be learned through strategy networks, we consider setting up a similar structure for the strategy learning of bidders.

If we consider static strategies, then for the game induced by the Myerson auction, the Nash equilibrium in the standard setting is:

$$B_i(v_i) = \frac{v_i + 1}{4}, \quad g_i(B_i(v_i)) = \frac{1}{2}v_i, \quad i \in \{1, 2\}.$$

However, its drawback is that  $B(v) \leq v$  does not always hold. This equilibrium strategy may lead to negative utilities for bidders with  $v < \frac{1}{3}$ . When bidders do not have access to the dynamic mechanism update method, adopting a bidding strategy that does not satisfy IR (individual rationality) may reduce the bidders' expected utility. Therefore, we require that the bidding strategy obtained by the network satisfies the following requirements:

- 1)  $B(v)$  is increasing function about  $v$ ,
- 2)  $v \geq B(v) \geq 0$ .

We propose a parametric network (Bid Net) to represent the bidder's strategy. The network first inputs the received values  $v_i$  from the distribution  $F_i$  into a multilayer perceptron. Monotonic outputs  $b_i^* = B_i^*(v_i)$  are combined with regularization constraints

$B_i(v_i) = \text{Min}[B_i^*(v_i), v_i]$ . This ensures that the output of the network meets both of these requirements. Since Bid Net training requires the use of a gradient propagation algorithm and the operators associated with sorting (*Min*, *Max*, *Sort*) cannot propagate the gradient, we use NeuralSort [13, 18] as an approximation in our network. The structure of the Bid Net is shown in Figure 2.

#### 3.2 Modeling Repeated Auction as an MARL System

We assume that bidders and the seller interact in an infinitely repeated auction with identical items and update their strategies after each auction round. At the moment  $t$ , the seller first announces the current parameter  $\theta$  of the mechanism. The strategy of the bidders  $\vec{\pi}$  and the seller  $\theta$  determines the utility  $u_i^t = a_i^t v_i^t - p_i^t$  and the revenue  $r^t$  of this auction round. For strategic bidders and the seller, the observations they receive are the joint bidding  $\vec{b}$ , mechanism parameter  $\theta$ , and their rewards ( $u_i^t$  or  $r^t$ ). When the value distribution is constant, the joint bidding distribution  $\vec{b}$  is directly determined by the joint strategy  $\vec{\pi}$ . Therefore, we use strategies  $\vec{\pi}$  instead of bids  $\vec{b}$  to represent the action of the bidders in repeated auctions.

Based on the observations and the reward maximization objectives, the agents will update their strategy for the next auction round. We usually assume that the strategic bidder adjusts the previous strategy based on observation:

$$\pi_i^{t+1} = \pi_i^t + \Delta \pi_i^t(\pi_{-i}^t, \theta^t, u_i^t).$$

This forms an MARL system when both the seller and bidders update their strategies through learning. The convergence of the strategy in repeated auctions is consistent with the equilibrium of the induced game, where the bidding strategy and mechanism reach a stable point. Our goal is to provide a learning algorithm for bidders that maximizes their utility as the system converges. When the bidder  $i$  is a naive learner [11], it will maximize its utility assuming that the other agent strategies remain unchanged:

$$\pi_i^{t+1} = \underset{\pi_i}{\text{argmax}} U_i(\pi_i, \pi_{-i}^t, \theta^t).$$

We can design gradient-based methods with learning rate  $\gamma$ :

$$\Delta \pi_i^t = \gamma \cdot \nabla_{\pi_i} U_i(\pi_i^t, \pi_{-i}^t, \theta^t).$$

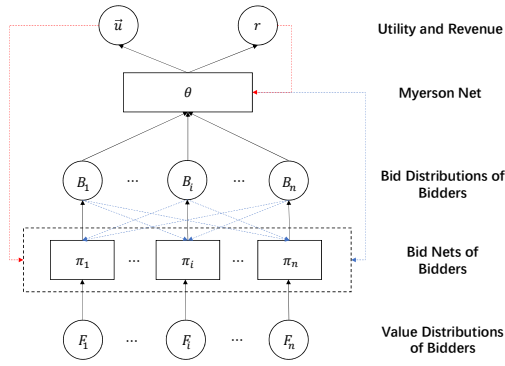
While the naive learner can only engage in inefficient learning process, the opponent modeling approach is able to improve the reward of agents by predicting the change in the opponent's strategy and selecting the corresponding best response:

$$\pi_i^{t+1} = \underset{\pi_i}{\text{argmax}} U_i(\pi_i, \pi_{-i}^t + \Delta \hat{\pi}_{-i}, \theta^t + \Delta \hat{\theta}).$$

A simple assumption in [11] is that the other agents are naive learners with lookahead rate  $\gamma'$ , which means that:

$$\Delta \hat{\pi}_j^t = \gamma' \cdot \nabla_{\pi_j^t} U_j(\pi_j^t, \pi_{-j}^t, \theta^t), \quad \forall j \neq i \text{ and } \Delta \hat{\theta}^t = \gamma' \cdot \nabla_{\theta^t} R(\theta^t, \vec{\pi}^t).$$

In the following sections, we will specifically discuss the design of the opponent modeling algorithm applicable to repeated auctions and propose our strategy learning method in this MARL system.



**Figure 3: Gradient propagation direction of the repeated auction induced MARL system. The red line represents a direct gradient, which comes from the revenue and utility. The blue line represents the indirect gradient, which comes from the impact of the bidder’s strategy on other players.**

### 3.3 Opponent Modeling Based Automatic Bidding Method

In this section, we introduce the automatic bidding method for the MARL system in Section 3.2. Since the seller and other bidders also update their strategies, agents must consider both the current utility and the impact on subsequent states when evaluating the strategy  $\pi_i$ . We adopt representation similar to RL (reinforcement learning) and use  $Q_i$  to denote the utility expectation of the strategy for the bidder  $i$ , then we have the following:

$$Q_i(\pi_i^t) = U_i(\pi_i^t, \pi_{-i}^t, \theta^t) + \lambda \max_{\pi_i} Q_i(\pi_i, \pi_{-i}^{t+1}, \theta^{t+1}),$$

where  $\lambda$  is the discount factor.

For the agent  $i$  who needs to choose a strategy at moment  $t$ , the information available is historical observations before moment  $t$ . Therefore, it is difficult to obtain unbiased estimates of  $Q_i$ . A direct approximation is  $\hat{Q}_i = U_i(\pi_i^t, \pi_{-i}^t, \theta^t)$ , which means that the bidder assumes that the subsequent strategy of other agents ( $\pi_{-i}^{t+1}, \theta^{t+1}$ ) is not relevant to its strategy  $\pi_i^t$ . Then its choice of action at the next moment will be based on the prediction of the opponent’s action:

$$\pi_i^t = \operatorname{argmax}_{\pi_i} U_i(\pi_i, \hat{\pi}_{-i}^t, \hat{\theta}^t),$$

$$\hat{\pi}_{-i}^t = \hat{\pi}_{-i}^t(\pi_i^{t-1}, \pi_{-i}^{t-1}, \theta^{t-1}), \quad \hat{\theta}^t = \hat{\theta}^t(\pi_i^{t-1}, \pi_{-i}^{t-1}, \theta^{t-1}).$$

For the naive learner, it will assume that the strategies of other agents remain unchanged:

$$\hat{\pi}_{-i}^t = \pi_{-i}^{t-1}, \quad \hat{\theta}^t = \theta^{t-1}.$$

Then its gradient update direction for the strategy will be:

$$\nabla_{\pi_i^{t-1}} U_i(\pi_i^{t-1}, \pi_{-i}^{t-1}, \theta^{t-1}).$$

More accurate estimations for  $U_i(\pi_i^t, \pi_{-i}^t, \theta^t)$  usually require first-order approximations:

$$\begin{aligned} & U_i(\pi_i, \pi_{-i} + \Delta \hat{\pi}_{-i}, \theta + \Delta \hat{\theta}) \\ & \approx U_i(\pi_i, \pi_{-i} + \Delta \hat{\pi}_{-i}, \theta) + (\Delta \hat{\theta})^\top \nabla_{\theta} U_i(\pi_i, \pi_{-i} + \Delta \hat{\pi}_{-i}, \theta) \\ & \approx U_i(\pi_i, \pi_{-i}, \theta) + (\Delta \hat{\pi}_{-i})^\top \nabla_{\pi_{-i}} U_i(\pi_i, \pi_{-i}, \theta) \\ & + (\Delta \hat{\theta})^\top \nabla_{\theta} [U_i(\pi_i, \pi_{-i}, \theta) + (\Delta \hat{\pi}_{-i})^\top \nabla_{\pi_{-i}} U_i(\pi_i, \pi_{-i}, \theta)]. \end{aligned}$$

We can obtain different predictions for  $(\Delta \hat{\pi}_{-i}, \Delta \hat{\theta})$  with opponent modeling methods. The LOLA [11] algorithm assumes that the opponent is a naive learner, which means  $\Delta \hat{\theta} = \eta \cdot \nabla_{\theta} R(\bar{\pi}, \theta)$ . Assuming that the opponent is a LOLA learner leads to high-order LOLA (HOLA). Similar opponent modeling algorithms include SOS (stable opponent shaping) [16], COLA (consistent learning with opponent learning awareness) [34], and others, which effectively learn equilibrium with higher rewards in different game environments.

COLA [34] points out that the effectiveness of opponent modeling algorithms depends on the consistency of the lookahead rate of all agents. This means that agents need to choose strategy-update approaches with consistency to ensure the convergence of the system. However, we have discussed earlier the asymmetry of the repeated auction environment: bidders have private information (value function) that allows them to adopt specific strategies to improve utility. Training strategies directly using opponent modeling algorithms does not maximize the utility of bidders.

Figure 3 shows the direction of propagation of the network parameter gradient for bidders and the seller in the MARL environment of repeated auctions. We illustrate the necessity of computing the indirect gradient represented by the blue dashed line in the figure in repeated auctions by proving the following property.

**THEOREM 3.1.** *If all bidders use only the gradient which comes directly from the utility to update the strategy under the Myerson Net of the seller, their gradient-based strategy updating will converge to the truthful bidding strategy  $B_i(v_i) = v_i$ . The stable state of the system is where all bidders bidding truthfully.*

The proof of Theorem 3.1 is given in the supplementary material. This Theorem indicates that the strategy network of bidder must be trained with the indirect gradient, which requires predicting the impact of bidding strategies on mechanism changes. In order to obtain a more accurate prediction of the  $Q_i$  function for bidders, we estimate  $\max_{\pi_i} Q_i(\pi_i, \pi_{-i}^{t+1}, \theta^{t+1})$  by modeling  $\Delta \hat{\pi}_{-i}^{t+1}, \Delta \hat{\theta}^{t+1}$  as functions of  $\pi_i^t$ :

$$\Delta \hat{\pi}_{-i}^{t+1} = \Delta \hat{\pi}_{-i}^{t+1}(\pi_i^t), \quad \Delta \hat{\theta}^{t+1} = \Delta \hat{\theta}^{t+1}(\pi_i^t).$$

Then we have:

$$\begin{aligned} \hat{Q}_i(\pi_i^t, \pi_{-i}^{t+1}, \dots) & = U_i(\pi_i^t, \hat{\pi}_{-i}^t, \hat{\theta}^t) + \lambda U_i(\pi_i^{t+1}, \hat{\pi}_{-i}^t + \Delta \hat{\pi}_{-i}(\pi_i^t), \hat{\theta}^t \\ & + \Delta \hat{\theta}(\pi_i^t)) + \lambda^2 \dots \end{aligned}$$

The convergence of the system in the repeated auction is equivalent to the convergence of the strategies of each agent. We can simplify the above equation by assuming that  $\pi_i^t = \pi_i^{t+1} = \dots$  and then it will become:

$$\hat{Q}_i(\pi_i^t) = U_i(\pi_i^t, \hat{\pi}_{-i}^t, \hat{\theta}^t) + \sum_{k=1}^{\infty} \lambda^k U_i(\pi_i^t, \hat{\pi}_{-i}^t + \Delta^k \hat{\pi}_{-i}(\pi_i^t), \hat{\theta}^t + \Delta^k \hat{\theta}(\pi_i^t)),$$

where  $\Delta^k \hat{\pi}_{-i}(\pi_i^t)$  means the prediction of  $\pi_{-i}$  after  $k$  updates.

We use  $\hat{\pi}_{-i}(\pi_i^t), \hat{\theta}(\pi_i^t)$  to represent the strategy of other agents when the system converges, which means that:

$$\hat{\pi}_{-i}(\pi_i^t) = \hat{\pi}_{-i}^t + \Delta^\infty \hat{\pi}_{-i}(\pi_i^t), \quad \hat{\theta}(\pi_i^t) = \hat{\theta}^t + \Delta^\infty \hat{\theta}(\pi_i^t).$$

We can always require that the system converge in finite time by periodic reducing the step size of the strategy updating. Assuming

**Algorithm 1:** Inner loop process of pseudo-gradient algorithm

---

**Input:** Initial Parameters  $\pi_i^{t-1}, \Delta\pi_i, \pi_{-i}^{t-1}, \theta^{t-1}$ ; termination step  $T$ ; lookahead rate  $\eta'$ .

- 1: Generate Myerson Net with parameter  $\theta^*$  which satisfies  $M(\theta^*) \cong M(\theta^{t-1})$ , and set  $\pi_{-i}^* = \pi_{-i}^{t-1}$
- 2: **for**  $k := 1, \dots, T$  **do**
- 3:   Generate sample  $S_k = (b_1, \dots, b_n)$  with  $(\pi_i^{t-1} + \Delta\pi_i, \pi_{-i}^*)$
- 4:   Compute gradient  $\nabla_{\theta} R(\theta^*, S_k)$
- 5:    $\theta^* = \theta^* + \eta' \nabla_{\theta} R(\theta^*, S_k)$ ,  $\pi_{-i}^* = \pi_{-i}^*$
- 6: **end for**

**Output:**  $\hat{\theta}^t = \theta^*$ ,  $\hat{\pi}_{-i}^t = \pi_{-i}^*$

---

that the system converges after  $T$  updates, we have:

$$\hat{Q}(\pi_i^t) = \sum_{k=0}^{T-1} U_i^{k+t} + (\lambda^T + \lambda^{T+1} + \dots) U_i(\pi_i^t, \hat{\pi}_{-i}(\pi_i^t), \hat{\theta}(\pi_i^t)).$$

When  $\lambda \rightarrow 1$ , the second item will be sufficiently larger than the first. Then we have:

$$\hat{Q}(\pi_i^t) \approx (\lambda^T + \lambda^{T+1} + \dots) U_i(\pi_i^t, \hat{\pi}_{-i}(\pi_i^t), \hat{\theta}(\pi_i^t)).$$

Thus, the strategy selection function of the agent is as follows:

$$\pi_i^t = \operatorname{argmax}_{\pi_i} U_i(\pi_i, \hat{\pi}_{-i}(\pi_i^t), \hat{\theta}(\pi_i^t)).$$

When a single bidder's strategy is fixed in a repeated auction, the strategy updates of other bidders and the seller are synchronized and will affect each other. The convergence result is determined by the algorithm used by these agents. To avoid discussing the possibility of different stable points in the system, we assume that the strategy updates of other agents are independent and the bidders' strategies change slowly  $\Delta \hat{\pi}_{-i}(\pi_i^t) \approx 0$ . We found this assumption to be valid in our experiments, for predicting the strategies of other agents under the condition that their private information is unknown will lead to a large bias.

We refer to this process (calculating  $\hat{\pi}_{-i}(\pi_i^t)$ ,  $\hat{\theta}(\pi_i^t)$ ) using  $\pi_i^t$  as the inner loop part of the algorithm. When the seller adopts Myerson Net as the mechanism, its strategy updating process according to historical bids is similar to a naive learner. We can simulate this process in the inner loop through the Myerson Net by constraining the bidder strategies to  $\pi_i^t, \pi_{-i}^{t-1}$ . The complete procedure for the inner loop is given in Algorithm 1.

Based on the strategy predictions of the other agents obtained from the inner loop, we can derive the bidding strategy that maximizes the expected utility. To avoid the instability caused by rapid strategy changes, we restrict the step size by  $|\Delta\pi_i| \leq d$ . Then our goal is to solve the optimization problem:

$$\max_{|\Delta\pi_i| \leq d} U_i(\pi_i + \Delta\pi_i, \hat{\pi}_{-i}(\pi_i + \Delta\pi_i), \hat{\theta}(\pi_i + \Delta\pi_i))$$

Since the parameters in the optimization objective contain the output of the inner loop, it is difficult to calculate the strategy gradient directly. We give the approximate calculation based on the PG (pseudo-gradient) by defining the pseudo-gradient obtained

**Algorithm 2:** Pseudo-gradient algorithm

---

**Input:** Initial Parameters  $\pi_i^0, \pi_{-i}^0, \theta^0$ , termination step  $T'$ , learning rate  $\eta$ , number of inner loop  $K'$  and constant  $s$  and  $l$ .

- 1: **for**  $t := 1, \dots, T'$  **do**
- 2:   Strategic bidder process: read the output of  $t - 1$  from environment and set  $\hat{\pi}_{-i} = \pi_{-i}^{t-1}, \hat{\theta} = \theta^{t-1}$
- 3:   **for**  $j := 1, \dots, K'$  **do**
- 4:     Randomly generate  $\Delta\pi_i^j$ , which satisfies  $\langle \Delta\pi_i^j, \Delta\pi_i^{j_0} \rangle < l$  ( $\forall j_0 \in \{j-1, \dots, j-s\}$ )
- 5:     Using the inner loop process to obtain  $\hat{\theta}(\Delta\pi_i^j)$
- 6:   **end for**
- 7:    $grad(\pi_i^{t-1}, \Delta\pi_i^j) = \frac{U_i(\pi_i^{t-1} + \Delta\pi_i^j, \hat{\pi}_{-i}, \hat{\theta}) - U_i(\pi_i^{t-1}, \pi_{-i}, \theta)}{|\Delta\pi_i^j|}$
- 8:    $grad^*(\pi_i^{t-1}) = \operatorname{argmax}_{\Delta\pi_i^j} [grad(\pi_i^{t-1}, \Delta\pi_i^j) \mid grad(\cdot, \cdot) > 0]$
- 9:    $\pi_i^t = \pi_i^{t-1} + \eta \cdot grad^*(\pi_i^{t-1})$
- 10:   Other bidders and the seller update their strategies according to their algorithms
- 11: **end for**

**Output:**  $(\pi_i^{T'}, \pi_{-i}^{T'}, \theta^{T'})$

---

from  $\Delta\pi_i$  as:

$$grad(\pi_i, \Delta\pi_i) = \frac{U_i(\pi_i + \Delta\pi_i, \pi_{-i} + \Delta\hat{\pi}_{-i}(\Delta\pi_i), \theta + \Delta\hat{\theta}(\Delta\pi_i)) - U_i(\pi_i, \pi_{-i}, \theta)}{|\Delta\pi_i|}$$

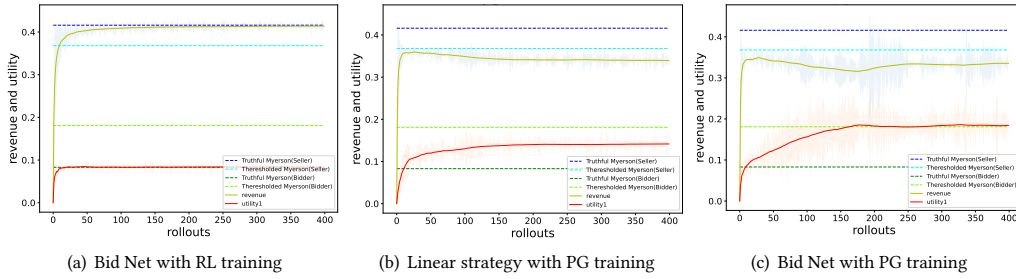
From the definition we can see that, given the strategy update  $\Delta\pi_i$ , it is possible to calculate the pseudo-gradients with the inner loop. For the algorithm to choose an update step that is close to the direction of the true gradient, we generate a set of different directions of  $\Delta\pi_i$  and select the positive gradient with the largest absolute value from all pseudo-gradients as the update direction of the Bid Net. The complete procedure for the PG algorithm is given in Algorithm 2. We want the algorithm to converge to the equilibrium of the repeated auction-induced game. For a single-item Myerson auction that induces a game among bidders, we give proof of the convergence in the supplementary material.

**THEOREM 3.2.** *Assuming that other bidders use static strategies  $\pi_{-i}$ ,  $\hat{\theta}(\Delta\pi_i)$  obtained from the inner loop belonging to the optimal mechanisms and  $K' \rightarrow \infty$  in algorithm 2, the strategic bidder using the PG algorithm will converge to the strategy with optimal utility. When all bidders adopt the PG algorithm, the equilibrium of the Myerson auction-induced game is the only stable point of the system.*

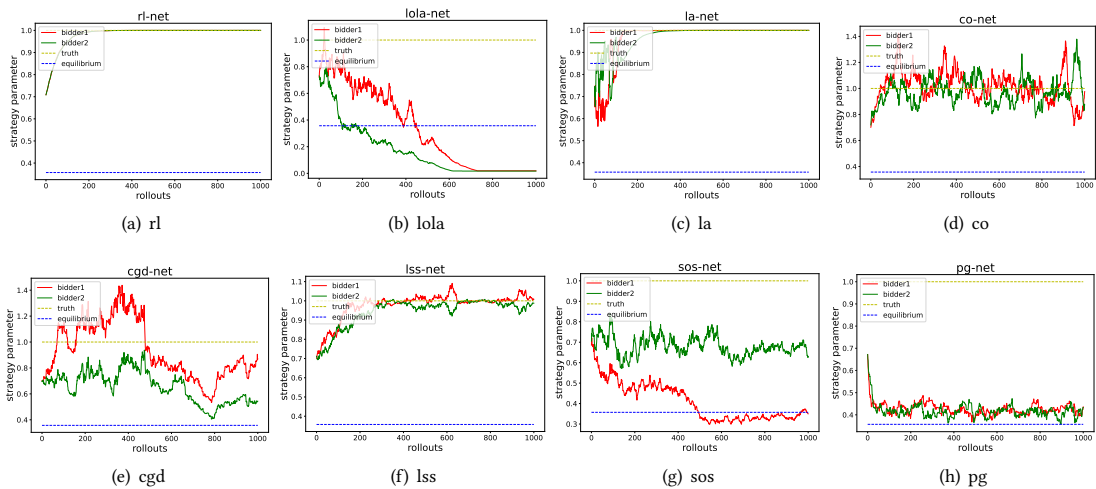
In our experiments, we find that the result of the algorithm satisfies this theorem even if the inner loop does not converge and  $K'$  is finite. Our algorithm PG converges to the equilibrium of all the games induced by the auction, even in multiple-bidders settings.

## 4 EXPERIMENTS AND DISCUSSION

Our experiment consists of three parts. 1) Experiment in the standard setting for Bid Net employed by individual bidders. We compare the learned Bid Net strategies with linear shading strategies



**Figure 4: The utility of a strategic bidder in scenarios where another bidder consistently employs the truthful bidding strategy, while the seller’s strategy is derived from the Myerson Net. The solid red line represents the utility of the strategic bidder, while the solid yellow line represents the revenue of the seller. The dashed line labeled "Truthful Myerson" represents the theoretical utility and revenue when the strategic bidder adheres to the truthful bidding strategy. The dashed line labeled "Theresholded Myerson" illustrates the theoretical utility and revenue when the strategic bidder employs the optimal bidding strategy.**



**Figure 5: The strategy parameters  $\alpha_i$  of strategic bidders during the learning process. Both bidders are designated strategic bidders and employ the same algorithm. The seller strategy is derived from Myerson Net. The line labeled "truth" represents the truthful strategy  $\alpha_i = 1$ , and the line labeled "equilibrium" represents the equilibrium strategy of the induced game  $\alpha_i = \frac{5}{14}$ .**

[24]. In this configuration, another bidder consistently employs the truthful bidding strategy, while the seller’s strategy is derived from the Myerson Net. 2) Experiment with the PG algorithm. As part of an ablation experiment, we utilize linear shading networks to represent bidder strategies. We compare the PG algorithm with other opponent modeling algorithms and evaluate performance using metrics such as the average utility of the bidders and the deviation of strategy from equilibrium in the induced game. 3) Experiments in diverse environments with varied settings and opponent models. We expand our experiment to include various environments that feature distinct settings and opponent models. Our aim is to demonstrate the effectiveness of the combination of PG algorithm and Bid Net in optimizing bidders’ utility in online auctions.

#### 4.1 Experiment for Bid Net

We compare linear shading strategies and Bid Net, which is trained using PG and RL algorithms. Our evaluation takes place in the standard setting, with another bidder consistently employing the truthful bidding strategy. The results are given in Figure 4.

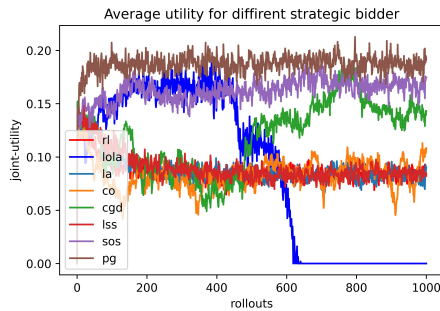
From the figure, we can see that the network trained using RL converges to the truthful bidding strategy. The linear shading strategy trained through PG improves the utility of the strategic bidder, but does not approach the optimal strategy. The strategy produced by Bid Net trained with the PG algorithm closely approximates the optimal strategy. This experiment underscores the efficiency of Bid Net in accurately representing bidding strategies, while linear shading strategies do not contain the optimal strategy. The gap between the seller’s revenue and the theoretical value may arise from the fact that we set a very small reserve price.

#### 4.2 Experiment for PG Algorithm

We compare the utility of agents with different strategy learning algorithms. The environment is the standard setting, and all bidders employ the linear shading strategy  $B_i(v_i) = \alpha_i v_i$ . In this environment, both bidders will adopt the same algorithm to learn the strategy. At the end of each auction round  $t$ , bidders and the seller will observe the strategies  $(\alpha_i^t, \alpha_{-i}^t, \theta^t)$  and the corresponding rewards  $(u_i^t, u_{-i}^t, r^t)$ . Subsequently, the seller updates the mechanism in accordance with the Myerson Net, while the bidders adjust

Setting	RL	LOLA	SOS	PG (linear shading)	PG (Bid Net)
① $B_2(v_2) = v_2$ $v_1 \sim U[0, 1], v_2 \sim U[0, 1]$	0.09	0.15	0.12	0.18	0.20
② $B_2(v_2) = 0.25v_2 + 0.25$ $v_1 \sim U[0, 1], v_2 \sim U[0, 1]$	0.09	0.07	0.10	0.12	0.16
③ $B_2(v_2) = v_2$ $v_1 \sim U[0, 2], v_2 \sim U[0, 1]$	0.22	0.37	0.33	0.33	0.46
④ $B_2(v_2) = v_2$ $v_1 \sim (\frac{n}{N} + 1) \cdot U[0, 1], v_2 \sim U[0, 1]$	0.16	0.26	0.09	0.27	0.33
⑤ $B_2(v_2) = v_2, B_3(v_3) = v_3$ $v_1 \sim U[0, 1], v_2 \sim U[0, 1], v_3 \sim U[0, 1]$	0.06	0.09	0.08	0.04	0.11
⑥ $B_2(v_2) \sim$ LOLA updating $v_1 \sim U[0, 1], v_2 \sim U[0, 1]$	0.12	0.16	0.16	0.20	0.31
⑦ $B_2(v_2) \sim$ SOS updating $v_1 \sim U[0, 1], v_2 \sim U[0, 1]$	0.12	0.13	0.15	0.18	0.32

**Table 1: The utility of strategic bidder across different opponent strategies and environments. Row labels signify the environment and strategy of the other bidders. The column labels indicate the algorithms employed by the strategic bidder.  $n$  and  $N$  denote the number of current rollouts and the total number of rollouts.**



**Figure 6: The average utility for the strategic bidders.**

their strategies based on their respective learning algorithms. An equilibrium of the above system is  $\alpha_1 = \alpha_2 = \frac{5}{14}$ .

We compare the PG algorithm with opponent modeling and equilibrium solving algorithms (RL, LOLA [11], LA [37], CO [21], CGD [28], LSS [19], SOS [16]) in the aforementioned environment. Figure 5 illustrates the evolution of agent strategy parameters during training. The results demonstrate that the majority of algorithms converge towards the truthful bidding strategy, aligning with the conclusion in Theorem 3.1. Algorithms that incorporate opponent predictions improve utility but display a degree of instability. Notably, the PG algorithm’s training results closely approximate the Nash equilibrium strategy. The slight deviation from equilibrium could be attributed to the implementation of a sufficiently small reserve price. The average utility of various agent systems is shown in Figure 6, where the PG algorithm attains the highest average utility in the automated bidding task, consistent with Theorem 3.2.

### 4.3 Experiments with Different Opponents and Environment Settings

We conducted a series of experiments within diverse environments and with varying opponent strategies. Experiments in which other bidders use static strategies include: ① **Standard setting with a truthful bidding opponent.** In this scenario, another bidder consistently employs the truthful bidding strategy. ② **Standard setting with a Nash equilibrium opponent.** In this setup, another bidder adheres to the Nash equilibrium bidding strategy. ③ **Asymmetric value distribution.** Here, the private value distribution

of bidders exhibits asymmetry. ④ **Dynamic value distribution.** In this experiment, the value distribution function of the strategic bidder evolves over time. ⑤ **Single-item three-bidders auction.** This experiment extends to a scenario with more than two bidders.

Experiments involving dynamic strategies employed by other bidders encompass: ⑥ **LOLA algorithm opponent.** Here, another bidder updates its strategy using the LOLA algorithm. ⑦ **SOS algorithm opponent.** In this case, another bidder adjusts its strategy using the SOS algorithm. In the two sets of experiments described above, the opponent bidder uses linear shading strategies.

We evaluate the effectiveness of the algorithms to improve the utility of the strategic bidder after 400 iterations. The results are presented in Table 1. The algorithms used for comparison (RL, LOLA, SOS) are experimented with both Bid Net and linear shading strategies, and the results in the table are taken from the better of the two. We see that the Bid Net trained with PG always maximizes the utility of the strategic bidder. More details of the setting and figures of the experiment are given in the supplementary material.

## 5 CONCLUSION

In this study, we have developed an automatic bidding approach for bidders in repeated auctions. We introduce the Bid Net as a representation of bidding strategies, and we have proposed the PG algorithm to train this network. We have shown the effectiveness of PG in learning optimal responses when faced with static opponents, as well as its convergence to induced equilibriums when all agents adopt PG simultaneously. Through a series of experiments, we have highlighted the superiority of Bid Net over the linear shading function and showcased the efficacy of the PG algorithm by comparing it with other opponent modeling algorithms. PG has proven to significantly enhance the utility of strategic bidders in varying environments and with diverse strategies employed by other agents. We hope that this work will contribute to more research on strategic bidders in auctions and automatic bidding.

## ACKNOWLEDGMENTS

This paper is supported by National Key R&D Program of China (2021YFA1000403), and National Natural Science Foundation of China (Nos.11991022), and the DNL Cooperation Fund, CAS (DNL202023).



## REFERENCES

- [1] Reza Refaei Afshar, Jason Rhuggenaath, Yingqian Zhang, and Uzay Kaymak. 2021. A Reward Shaping Approach for Reserve Price Optimization using Deep Reinforcement Learning. In *2021 International Joint Conference on Neural Networks*. 1–8.
- [2] Anthony Bagnall and Iain Toft. 2006. Autonomous adaptive agents for single seller sealed bid auctions. *Autonomous Agents and Multi-Agent Systems* 12 (2006), 259–292.
- [3] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. 2018. A General Theory of Sample Complexity for Multi-Item Profit Maximization. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 173–174.
- [4] Maria-Florina F Balcan, Tuomas Sandholm, and Ellen Vitercik. 2016. Sample Complexity of Automated Mechanism Design. In *Advances in Neural Information Processing Systems*, Vol. 29.
- [5] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. 2018. Reinforcement Mechanism Design for E-Commerce. In *Proceedings of the 2018 World Wide Web Conference*. 1339–1348.
- [6] Carlos Carrion, Zenan Wang, Harikesh Nair, Xianghong Luo, Yulin Lei, Peiqin Gu, Xiliang Lin, Wenlong Chen, Junsheng Jin, Fanan Zhu, et al. 2023. Blending advertising with organic content in e-commerce via virtual bids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15476–15484.
- [7] Dave Cli. 1997. Minimal-intelligence agents for bargaining behaviors in market-based environments. *Hewlett-Packard Labs Technical Reports* (1997).
- [8] Richard Cole and Tim Roughgarden. 2014. The Sample Complexity of Revenue Maximization. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*. 243–252.
- [9] Paul Duetting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. 2019. Optimal Auctions through Deep Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 1706–1715.
- [10] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. *American Economic Review* 97, 1 (2007), 242–259.
- [11] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 122–130.
- [12] Renato D. Gomes and Kane S. Sweeney. 2009. Bayes-Nash Equilibria of the Generalized Second Price Auction. In *Proceedings of the 10th ACM Conference on Electronic Commerce*. 107–108.
- [13] Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. 2019. Stochastic Optimization of Sorting Networks via Continuous Relaxations. In *International Conference on Learning Representations*. 1–23.
- [14] Yash Kanoria and Hamid Nazerzadeh. 2014. Dynamic Reserve Prices for Repeated Auctions: Learning from Bids. In *Web and Internet Economics*. 232–232.
- [15] Vijay Krishna. 2009. *Auction theory*. Academic press.
- [16] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. 2019. Stable Opponent Shaping in Differentiable Games. In *International Conference on Learning Representations*. 1–20.
- [17] Ningyuan Li, Yunxuan Ma, Yang Zhao, Zhijian Duan, Yurong Chen, Zhilin Zhang, Jian Xu, Bo Zheng, and Xiaotie Deng. 2023. Learning-Based Ad Auction Design with Externalities: The Framework and A Matching-Based Approach. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1291–1302.
- [18] Xiangyu Liu, Chuan Yu, Zhilin Zhang, Zhenzhe Zheng, Yu Rong, Hongtao Lv, Da Huo, Yiqing Wang, Dagui Chen, Jian Xu, Fan Wu, Guihai Chen, and Xiaoqiang Zhu. 2021. Neural Auction: End-to-End Learning of Auction Mechanisms for E-Commerce Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and amp; Data Mining*. 3354–3364.
- [19] Eric Mazumdar, Michael I. Jordan, and S. Shankar Sastry. 2019. On Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-Sum Games. <https://arxiv.org/abs/1901.00838>
- [20] R Preston McAfee and John McMillan. 1987. Auctions and bidding. *Journal of economic literature* 25, 2 (1987), 699–738.
- [21] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2017. The Numerics of GANs. In *Advances in Neural Information Processing Systems*, Vol. 30. 1–11.
- [22] Munyque Mittelmann, Sylvain Bouveret, and Laurent Perrusset. 2022. Representing and reasoning about auctions. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 20.
- [23] Roger B. Myerson. 1981. Optimal Auction Design. *Mathematics of Operations Research* 6, 1 (1981), 58–73.
- [24] Thomas Nedelec, Jules Baudet, Vianney Perchet, and Noureddine El Karoui. 2021. Adversarial Learning in Revenue-Maximizing Auctions. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 955–963.
- [25] Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, Vianney Perchet, et al. 2022. Learning in repeated auctions. *Foundations and Trends® in Machine Learning* 15, 3 (2022), 176–334.
- [26] Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet. 2019. Learning to bid in revenue-maximizing auctions. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 4781–4789.
- [27] Jad Rahme, Samy Jelassi, and S. Matthew Weinberg. 2021. Auction Learning as a Two-Player Game. In *International Conference on Learning Representations*. 1–16.
- [28] Florian Schaefer and Anima Anandkumar. 2019. Competitive Gradient Descent. In *Advances in Neural Information Processing Systems*, Vol. 32. 1–11.
- [29] Weiran Shen, Binghui Peng, Hanpeng Liu, Michael Zhang, Ruohan Qian, Yan Hong, Zhi Guo, Zongyao Ding, Pengjun Lu, and Pingzhong Tang. 2020. Reinforcement Mechanism Design: With Applications to Dynamic Pricing in Sponsored Search Auctions. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 02, 2236–2243.
- [30] Weiran Shen, Pingzhong Tang, and Song Zuo. 2019. Automated Mechanism Design via Neural Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 215–223.
- [31] Pingzhong Tang and Yulong Zeng. 2018. The Price of Prior Dependence in Auctions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 485–502.
- [32] Hal R. Varian. 2009. Online Ad Auctions. *American Economic Review* 99, 2 (2009), 430–34.
- [33] Hal R. Varian and Christopher Harris. 2014. The VCG Auction in Theory and Practice. *American Economic Review* 104, 5 (2014), 442–45.
- [34] Timon Willi, Alistair Hp Letcher, Johannes Treutlein, and Jakob Foerster. 2022. COLA: Consistent Learning with Opponent-Learning Awareness. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. 23804–23831.
- [35] Song Yao and Carl F Mela. 2008. Online auction demand. *Marketing Science* 27, 5 (2008), 861–885.
- [36] Chongjie Zhang and Victor Lesser. 2010. Multi-agent learning with policy prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 927–934.
- [37] Chongjie Zhang and Victor Lesser. 2010. Multi-Agent Learning with Policy Prediction. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 927–934.
- [38] Zhilin Zhang, Xiangyu Liu, Zhenzhe Zheng, Chenrui Zhang, Miao Xu, Junwei Pan, Chuan Yu, Fan Wu, Jian Xu, and Kun Gai. 2021. Optimizing Multiple Performance Metrics with Deep GSP Auctions for E-Commerce Advertising. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 993–1001.