Optimal Bidding in Repeated Wireless Spectrum Auctions with Budget Constraints

Mehrdad Khaledi and Alhussein A. Abouzeid

Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180-3590, USA
Email: khalem@rpi.edu, abouzeid@ecse.rpi.edu

***Abstract*—Small operators who take part in secondary wireless spectrum markets typically have strict budget limits. In this paper, we study the bidding problem of a budget constrained operator in repeated secondary spectrum auctions. In existing truthful auctions, truthful bidding is the optimal strategy of a bidder. However, budget limits impact bidding behaviors and make bidding decisions complicated, since bidders may behave differently to avoid running out of money. We formulate the problem as a dynamic auction game between operators, where knowledge of other operators is limited due to the distributed nature of wireless networks/markets. We first present a Markov Decision Process (MDP) formulation of the problem and characterize the optimal bidding strategy of an operator, provided that opponents' bids are i.i.d. Next, we generalize the formulation to a Markov game that, in conjunction with model-free reinforcement learning approaches, enables an operator to make inferences about its opponents based on local observations. Finally, we present a fully distributed learning-based bidding algorithm which relies only on local information. Our numerical results show that our proposed learning-based bidding results in a better utility than truthful bidding.**

Keywords*—*Wireless Spectrum Sharing, Game Theory, Markov Decision Process, Learning, Markov Games.

I. Introduction

Spectrum scarcity has become a major challenge due to the rapid growth in mobile wireless communications. Several measurement studies indicate that the problem lies in inefficient use of the available wireless spectrum rather than scarcity of the spectrum [1]. Secondary spectrum markets have emerged to improve the spectrum utilization, where a primary license owner (PO) can lease its idle spectrum band(s) to unlicensed secondary users (SU) for a short period of time. The common approach for leasing the spectrum is holding an auction among SUs.

Several auction mechanisms have been proposed in the literature for re-allocating the spectrum in secondary markets that mostly focus on a single round of auction (one-shot mechanisms) [2], [3]. However, in practice, secondary spectrum auctions repeat frequently. The difficulty arises as the SUs can learn some information about their opponents and the environment over time, which consequently complicate their bidding decisions.

In a repeated auction environment, the major problem of an SU is to find an optimal biding strategy that maximizes its long-term utility. The decision making process of SUs in a repeated spectrum auction is studied in [4]. In their model, SUs choose between participating in the auction by bidding their true valuations, or staying out of the auction to just monitor the results. Assuming independent and identically distributed (i.i.d.) SUs' valuations, a threshold is derived for SUs above which they should participate in the auction. In a more general context, [5] utilizes Bayesian auction games for resource allocation in wireless networks, which entails maintaining beliefs about private information of others. From a PO perspective, the spectrum pricing competition has been studied extensively [6]–[8]. For instance, [8] studies a pricing game between two spectrum providers and presents the optimal provider selection strategy for SUs. In such a model, SUs choose a spectrum provider solely based on the offered price, then bid their true valuations.

We study repeated spectrum auctions in presence of budget constrained SUs, since in real world scenarios, bidders have limits on the amount of money they can spend. According to an analysis of FCC's spectrum auctions [9], many local wireless operators have budget limits, and these limits affect their bidding behaviors. Each operator typically starts with an initial budget to invest in the spectrum market. The operator improves its utility after winning an auction and getting high quality channel access for its services. In case of losing an auction, the operator may resort to opportunistic generalized access mode which does not provide a quality of service guarantee, [10]. Therefore, an operator needs to bid wisely and plan its budget to get the most value from its participation in multiple rounds of auction. It should be noted that we use the terms operator and SU interchangeably in this paper.

Our goal is to characterize an optimal bidding strategy for a budget constrained SU in repeated secondary spectrum auctions. To the best of our knowledge, optimizing the bidding strategy of an SU in presence of budget limits has not been previously considered in the literature. The challenge presented by budget limits is that it makes the bidding decisions complex, since an SU needs to take into account both the competition in the market and its own budget constraints. *Truthful bidding is no longer the optimal bidding strategy when SUs are budget constrained, as SUs may behave differently to avoid running out of money.* Thus, in contrast with prior works [4], [11] that assume SUs always bid their true valuations, budget constrained SUs have a wide variety of strategies for bidding. Therefore, SUs face a budget planning problem and they need to find utility-maximizing bids without exceeding their

budgets.

The significance of our work is that we propose solutions for the bidding problem of a budget constrained SU, with and without i.i.d competing bids. We characterize the optimal bidding strategy of an SU, when opponents' bids are i.i.d. For the case when no information about other SUs is available, we present a learning-based bidding algorithm that relies only on local information, and is well-suited to wireless environments/markets. It is worth noting that budget optimization has been studied in the context of online keyword advertising [12]–[14]. An underlying assumption in the analysis of such studies is that a bidder faces large (theoretically infinite) number of i.i.d bidders. However, such an assumption does not typically hold in the context of wireless spectrum markets, since there are limited number of competing SUs.

It should also be noted that our approach is different from the literature of dynamic auction design, where the objective is to design efficient or revenue-optimal mechanisms in dynamic environments (e.g. [15]). Instead of designing a complex mechanism that focuses on the PO's side, we consider repetition of simple auction mechanisms, and we study the dynamics of such a system from SUs' point of view. In this setting, we analyze the bidding strategies of an SU. In fact, an SU is faced with a trade-off between the possibility of getting a surplus in the current auction and the possibility of getting a larger but uncertain surplus in future auctions, subject to its budget limit.

In this paper, we make the following contributions. We formulate the budget-constrained spectrum sharing problem as a repeated auction game in which SUs compete to get one of the available channels. We first present a Markov Decision Process (MDP) formulation of the problem and characterize the optimal bidding strategy of an SU, assuming that opponents' bids are i.i.d. Next, we generalize the formulation to a Markov game, where an SU can make inferences about its competitors based on its local observations, and i.i.d. bids assumption is not required. Finally, we present a fully distributed learning-based bidding algorithm which relies only on local information.

The rest of this paper is organized as follows. Section II presents the system model used in this paper. In Section III, we present a formulation of the optimal bidding problem of a budget constrained SU. We characterize the optimal bidding strategy of an SU, assuming that the SU faces i.i.d opponent bids in Section IV. In Section V, we present a fully distributed learning-based bidding algorithm for an SU, which does not require i.i.d. bids assumption. Numerical results are presented in Section VI. Finally, Section VII concludes the paper and outlines possible avenues for future work.

II. System Model

We consider a network consisting of a set of secondary users/operators (SUs) who are willing to buy channel access for their services from a primary spectrum owner (PO). SUs are budget constrained and compete with other SUs to get one of the k available channels. The auction is held by the PO and is repeated over time which is indexed by $t = 0, 1, 2, \cdots$.

An SU's valuation for a channel is the benefit for that specific SU of obtaining that channel. Similar to [2], [8], SU's valuation for a channel can be related to the achievable capacity of that channel. It is worth noting that the model presented in this paper is not limited to any specific valuation function. We assume that at each time step, each SU can observe its current valuation, and that valuations evolve according to a Markov probability model. Let v_i^t denote the valuation of SU i at time t, then $P(v_i^{t+1}|v_i^t, v_i^{t-1}, \cdots, v_i^0) = P(v_i^{t+1}|v_i^t)$. Each SU knows its own valuation probability transition model which can be learnt over time. [15] presents a model in which SUs learn their valuations over time.

In this paper, we utilize the well-known Vickrey-Clarke-Groves (VCG) auction [16] in each round. At time step t, the VCG mechanism takes the SUs' bids as input and determines the output for each SU as

$$o_i^t = \{(x_i^t, p_i^t)|x_i^t \in \{0,1\} \wedge \sum_i x_i^t \leq 1\}, \forall i, \qquad (1)$$

where the output consists of the allocation indicator, which determines whether a channel is allocated to SU i or not, and the payment that SU i needs to make.

According to the VCG mechanism, k identical channels are allocated to the SUs with k highest bids. The winning SUs need to pay the externality[1] that they cause on other SUs. Since channels are identical, the winners pay the $(k+1)$th highest bid. Therefore, we have $p_i^t = p^t = (k+1)$th highest bid if $x_i^t = 1$, and $p_i^t = 0$ otherwise. In such an auction, $(k+1)$th highest bid is a *threshold* bid for winning the auction and winners pay that threshold.

The auction mechanism in each step can be summarized as follows. First, SU i observes its valuation v_i^t. Second, SU i decides what to bid in the current round which is denoted by b_i^t. Third, The PO holds the auction based on the VCG mechanism. Finally, SU i observes its own bidding result o_i^t, defined in (1).

We focus on the bidding problem faced by an SU in the described repeated auction environment. At each time step, an SU's bid depends not only on its valuation, but also on its remaining budget and the behavior of its competitors. In conventional auction settings, where SUs are not budget constrained, it is in SUs' best interest to bid their true valuations. Thus, truthful bidding is the best strategy of an SU regardless of its opponents. However, in presence of budget limits, truthful bidding is no longer the best strategy. In fact, an SU needs to plan its budget and find its optimal bidding strategy accordingly. In addition, the SU needs to take into account the behavior of its opponents in its decision making process. However, due to the distributed nature of network, knowledge about other SUs is limited, and each SU may learn some information about its opponents by repeatedly participating in the auction.

An instance of the problem setting is depicted in Figure 1 where three secondary operators compete for channel access in a repeated auction. Each SU typically starts with an initial budget to invest in the spectrum market. An SU improves its utility after winning an auction and getting channel access for its services. At each time step, an SU can explore and learn

[1] In other words, an SU pays the difference between the social welfare of the others with and without its participation [16].

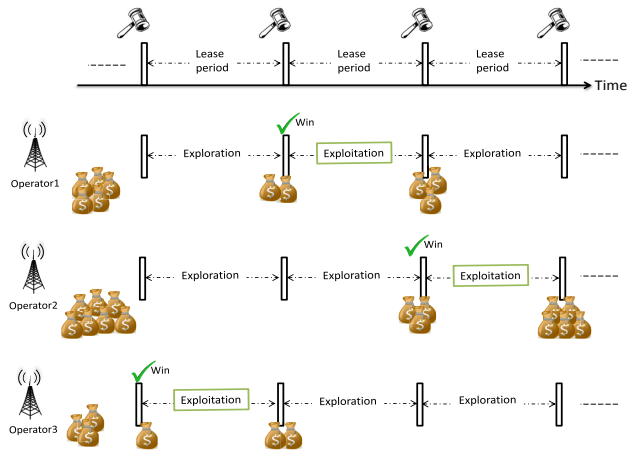

Fig. 1. Three budget constrained operators compete for a single channel in a repeated auction. If an operator wins, it can exploit the channel and earn an income. By participating in several rounds of auction, operators get the chance to explore and learn about their opponent's bids.

more information about its opponents or exploit the current information and bid to win a channel.

III. Optimal Bidding Problem Formulation

In this section, we formulate the optimal bidding problem of a budget constrained SU in the repeated auction environment described in Section II. Let m_i^t be the remaining budget of SU i at time t. SU i observes its valuation at time t and places a bid b_i^t, which results in an immediate utility of $(v_i^t - p_t)\mathbb{1}_{p_t < b_i^t \leq m_i^t}$. The SU's problem is to find a bidding strategy that maximizes its long-term discounted utility. The SU's objective at time t can be written as:

$$\mathbb{E}\Big[\sum_{t'=t}^{\infty} \delta^{t'-t}\,(v_i^{t'} - p_{t'})\mathbb{1}_{p_{t'} < b_i^{t'} \leq m_i^{t'}}\Big] \tag{2}$$

where the expectation is taken over the winning threshold, $p_{t'}$, and $0 \leq \delta < 1$ is the discount factor that controls how important future rewards are in current decisions (with larger values of δ giving more weight to future situations, as opposed to immediate rewards).

The bidding problem of SU i can be modeled as a Markov decision process (MDP) that is described by a tuple $< S_i, B_i, q_i, r_i >$. S_i corresponds to a finite set of states of SU i, where state of an SU is specified by its valuation and its remaining budget. Formally, the state of SU i at time t is defined as $s_i^t = (m_i^t, v_i^t)$. B_i denotes a finite set of actions for SU i, where an action corresponds to placing a bid, b_i^t. State transition probability for SU i is represented by q_i. Therefore, $q_i(s_i^{t+1}|s_i^t, o_i^t)$ is the probability that the state of SU i changes from s_i^t to s_i^{t+1} when the auction output is o_i^t. State of an SU transits as follows. SU i's valuation is drawn i.i.d over time, and its budget evolves according to the following equation

$$m_i^{t+1} = \begin{cases} m_i^t + a_i - p_t, & x_i^t = 1 \\ m_i^t, & \text{Otherwise} \end{cases} \tag{3}$$

where a_i denotes a fixed income that the SU earns from getting channel access permission, and p_t is the threshold amount that the SU pays for winning the auction at time t.

The immediate reward of SU i in an auction round is denoted by r_i, which is the difference between SU's valuation and its payment if the SU wins a channel.

$$r_i^t(s_i^t, b_i^t, o_i^t) = \begin{cases} v_i^t - p_t, & x_i^t = 1 \\ 0, & \text{Otherwise} \end{cases} \tag{4}$$

With the MDP formulation, the SU i's objective is to find a stationary strategy π_i that maps its current state (valuation and remaining budget) into a bid to maximize its long-term discounted utility given by

$$\max_{\pi_i \in \Pi_i} \mathbb{E}\Big[\sum_{t'=t}^{\infty} \delta^{t'-t}\, r_i^t(s_i^t, b_i^t, o_i^t)\Big]. \tag{5}$$

IV. Optimal Bidding with i.i.d SUs

In this section, we characterize the optimal bidding strategy of an SU assuming that the SU faces i.i.d opponent bids. This assumption implies that the SU knows the probability distribution of the winning threshold. While the assumption of i.i.d bidders is common in the prior work [4], in Section V, we present a learning-based approach that does not require i.i.d opponent bids.

We define the value function of the described MDP as the maximum (over all bidding strategies) expected discounted utility of an SU. Let $U(m)$ be the value function starting from budget m, using the dynamic programming principle we can write

$$U(m) = \mathbb{E}_v\Big[\max_{b \leq m} \mathbb{E}_p\big[(v - p + \delta U(m - p + a))\mathbb{1}_{p<b} + \delta U(m)\mathbb{1}_{p \geq b}\big]\Big] \tag{6}$$

The SU wins if it bids strictly more than the winning threshold p. In this case, the SU gets an immediate reward of $v - p$ in addition to the discounted expected future utility of starting with budget $m - p + a$. If the SU loses the auction, it gets the discounted expected utility with the same initial budget. It is worth noting that since we consider the bidding problem of a typical SU, we can omit the SU index for simplicity of notation. Also, we can leave out the time index in the above recursive formula.

For every possible winning threshold p, the SU's optimal bid can be found by simulating a single-shot VCG auction in which the winning threshold is represented by a function f defined as:

$$f(p, m) = p + \delta(U(m) - U(m - p + a)). \tag{7}$$

The function f defines the costs associated with winning a round of auction. The first term p is the immediate cost that the winning SU needs to pay. The second term in (7) is the exploitation cost which is incurred when the SU wins the current round of auction and starts the next round with budget $m - p + a$, compared to the case of losing the current auction and starting the next round with the same budget. In fact, exploitation cost is the discounted utility difference between winning and not winning the current round of auction.

Now, the optimal bid can be defined as a function of the current state (consisting of budget and valuation) as follows:

$$b^*(m,v) = \arg\max_{b \le m} \mathbb{E}_p\big[(v - f(p,m))\mathbb{1}_{p<b\le m}\big]. \quad (8)$$

In the following theorem, we characterize the optimal bid of an SU.

Theorem 1: The optimal bidding strategy of a budget constrained SU in the described repeated VCG auction (Section II) is characterized as

$$b^*(m,v) = \min(m, f^{-1}(v,m))$$

where $f^{-1}(v,m)$ is the z such that $f(z,m) = v$.

Proof: Due to space limits, we put the proof in our technical report [17]. ■

The specified optimal bid in Theorem 1 depends on the value function (6) of the MDP. Therefore, in order to calculate the optimal bid, the SU needs to compute $U(m)$. Let U^t be the value function at time t, we can find U^t for $t = 1, 2, \cdots$ iteratively as follows:

$$U^{t+1}(m) = \delta U^t(m) + \mathbb{E}\big[v - (p + \delta(U^t(m) - U^t(m-p+a)))\big]^+,$$

with the initial value of $U^0(m) = 0$ for $\forall m$. It is worth noting that the above equation is another form of the value function defined in (6). If the SU loses the auction, the expectation term is zero in the above equation and the SU gets $\delta U^t(m)$. When the expectation term is positive and the SU wins the auction, $\delta U^t(m)$ terms cancel out and the SU gets $v - p + \delta U^t(m - p + a)$.

V. Learning-based Optimal Bidding Strategy

In this section, we find an optimal bidding strategy of an SU without the i.i.d bids requirement. For this purpose, we formulate the bidding problem as a Markov game (also called a stochastic game)[2] [18].

An n-user Markov game can be described by a tuple $< S,\ B_1,\ \cdots,\ B_n,\ r_1,\ \cdots,\ r_n,\ q >$ where S is the state space, B_i is the set of actions for user i, r_i is the reward function for user, and q determines the state transition probabilities. Given state $s \in S$, each user independently chooses an action $b_i \in B_i$, and receives a reward r_i. Then, the state transits to the next state based on transition probability function q which follows the Markov property.

It is worth noting that in a Markov game, states are defined globally and for the environment. That is, all users make their decisions based on a common environment state, and the system state evolves as a result of joint actions. In accordance with Section III, we consider a local state space S_i for each SU i. We define the global state space as $S = \Pi_i S_i$, and we let $S_{-i} = \Pi_{j\neq i} S_j$ be the joint state of all SUs other than i. The global state of the system at time slot t is defined as $s^t = (s_i^t, s_{-i}^t)$.

Also, in such a Markov game, each SU's reward depends on the global state and the joint action of all SUs. However, due to the distributed nature of wireless networks/markets, exact information about other SUs is not available. Therefore, an SU needs to learn about its opponents through observations made from participating in the auction.

It should be noted that, in contrast with [4] that assumes SUs can stay out of the auction and monitor the results, we assume that an SU can make observations only through participating in the auction. We also assume that SUs cannot observe each other's bids, and that no information is exchanged among SUs. Thus, we define the observation of an SU as its previous states, bids, and auction outcomes for that SU, in addition to the SU's current state. Formally, we define the observation of SU i at time t as $(s_i^{t''}, b_i^{t'}, o_i^{t'})$ for $t' = 0, \cdots, t-1$ and $t'' = 0, \cdots, t$.

We utilize *model-free reinforcement learning* approaches in which an SU learns its optimal bidding strategy without knowing the reward function or the state transition probabilities. Q-learning [19], [20] is a well-known example of model-free reinforcement learning algorithms. The main idea of Q-learning is to define a Q-function that represents the quality of a state-action pair. Then, for a given state, the optimal strategy would be to choose an action that gives the highest value for Q-function.

A. State Space Classification

In a Markov game, Q-functions are defined over the global state and joint actions of all SUs. However, as mentioned earlier SUs cannot observe states and actions of each other. Thus, SU i needs to approximate the state of others, S_{-i}. Since the winning threshold fully represents the state and behavior of other SUs, it suffices for an SU to keep an estimate of the winning threshold. Therefore, winning threshold can be used as the representative state of competing SUs. In order to reduce the time and space complexity of learning, we use a similar state classification as in [11] to classify the representative state space. Let $\mathcal{T}$ be the maximum value for the winning threshold. SU i uniformly decomposes the range $[0, \mathcal{T}]$ into N_i intervals as $[\mathcal{T}_0, \mathcal{T}_1), [\mathcal{T}_1, \mathcal{T}_2), \cdots, [\mathcal{T}_{N_i-1}, \mathcal{T}_{N_i}]$, where $\mathcal{T}_0 \le \mathcal{T}_1 \le \cdots \le \mathcal{T}_{N_i} = \mathcal{T}$.

Depending upon the outcome of the auction, SU i gets to know different information about its competitors. Let $\tilde{s}_{-i}^t$ be the approximated state of other SUs at time t, we have the following two cases:

1) If SU i wins the auction at time t, the winning threshold can be observed. Therefore, the representative state of other SUs is determined as

$$\tilde{s}_{-i}^t = n, \quad \text{if} \quad p_t \in [\mathcal{T}_{n-1}, \mathcal{T}_n)$$

2) When SU i loses the current round of auction, the only information available to the SU is that its bid was lower than the winning threshold. Thus, the representative state can be chosen as

$$\tilde{s}_{-i}^t = n, \quad \text{if} \quad b_i^t \in [\mathcal{T}_{n-1}, \mathcal{T}_n)$$

It is worth noting that the choice of N_i leaves a tradeoff between complexity and performance for SU i. Higher values of N_i results in more accurate approximation of S_{-i}, but at the cost of increased complexity.

[2] The theory of MDP focuses on a single-user stationary environment. Game theory, on the other hand, studies the interaction of multiple users. Markov games extend game theory to MDP-like environments. In other words, Markov games generalize MDP to environments with multiple interacting users.

B. Transition Probability Estimation

SU i also needs to estimate the transition probabilities for representative state of other SUs. For this purpose, SU i maintains an $N_i \times N_i$ matrix Y. Each element $y_{n,m}$ of the matrix indicates the number of transitions from $\tilde{s}_{-i}^t = n$ to $\tilde{s}_{-i}^{t+1} = m$. SU i can update the matrix Y through its observations and state approximation described in previous subsection. Then, we can approximate the transition probabilities as follows:

$$q_{-i}(\tilde{s}_{-i}^{t+1} = m | \tilde{s}_{-i}^t = n) = \frac{y_{n,m}}{\sum_m y_{n,m}}$$

C. The Learning Algorithm

In this section, we present a learning-based bidding algorithm for an SU which depends only on the local observations of the SU. The learning algorithm is similar to the well-known Q-learning [20] method, except that we include budget constraints of SUs, and we use state classification and transition probability approximation of other SUs, since the information about other SUs are limited in the network.

We define the Q-function of SU i at time t as follows. The quality of action b_i, when state of SU i is s_i and the representative state of others is $\tilde{s}_{-i}$, equals

$$Q_i^t(s_i, \tilde{s}_{-i}, b_i) = \begin{cases} (1-\alpha_i^t) Q_i^{t-1}(s_i, \tilde{s}_{-i}, b_i) + \alpha_i^t (r_i^t \\ \qquad\qquad + \delta V_i^t(s_i, \tilde{s}_{-i})), \\ \qquad\qquad \text{if } s_i^t = s_i, \tilde{s}_{-i}^t = \tilde{s}_{-i}, b_i^t = b_i \\ \\ Q_i^{t-1}(s_i, \tilde{s}_{-i}, b_i) \qquad \text{Otherwise} \end{cases} \tag{9}$$

where $0 \le \alpha_i^t < 1$ is the SU's learning rate, r_i^t is the immediate reward as defined in (4). The function $V_i^t(s_i, \tilde{s}_{-i})$ represents the value of the joint state $(s_i, \tilde{s}_{-i})$, which is the expected discounted utility starting from that state.

$$V_i^t(s_i^t, \tilde{s}_{-i}^t) = \sum_{s_i^{t+1}, \tilde{s}_{-i}^{t+1}} \Big[q_i(s_i^{t+1} | s_i^t, o_i^t) q_{-i}(\tilde{s}_{-i}^{t+1} | \tilde{s}_{-i}^t) \max_{b_i \le m_i^{t+1}} \big\{ Q_i^{t-1}(s_i^{t+1}, \tilde{s}_{-i}^{t+1}, b_i) \big\} \Big] \tag{10}$$

In other words, the quality of a state-action pair (9) is the immediate utility plus the discounted expected value of future states, and the value of a joint state (10) is the quality of the best action for that state. The results in [20] show that the estimated values for Q and V converge to their true values if learning rates satisfy certain conditions. Therefore, if an SU learns the Q values, it can specify its optimal strategy, which is choosing the bid (action) with the highest Q value subject to its budget constraints. Thus, SU i chooses its bid at time t according to the following strategy:

$$\pi_i^*(s_i^t, \tilde{s}_{-i}^{t-1}) = \arg\max_{b_i \le m_i^t} \Big\{ \sum_{\tilde{s}_{-i}^t} q_{-i}(\tilde{s}_{-i}^t | \tilde{s}_{-i}^{t-1}) Q_i^{t-1}(s_i^t, \tilde{s}_{-i}^t, b_i) \Big\} \tag{11}$$

The SU chooses a bid that maximizes its expected Q value, where the expectation is taken over the possible representative state of other SUs for the current time step. This is because SU i can infer about other SUs' state only after bidding and observing the auction results. Given the information from previous time step and with the aid of transition probability approximation (Section V-B), the SU can find the expected current state of other SUs.

Algorithm 1 Learning-based bidding for SU i

Initialize the Q_i values to zero for all possible states and bids
Initialize $n(s)$ values to zero for all possible joint states s
for Each time step t **do**
  Observe the current state s_i^t
  With probability $\epsilon(s_i^t, \tilde{s}_{-i}^{t-1}) = c/n(s_i^t, \tilde{s}_{-i}^{t-1})$ choose a random bid, and with probability of $1-\epsilon(s_i^t, \tilde{s}_{-i}^{t-1})$ use the greedy strategy in (11) to place a bid
  $n(s_i^t, \tilde{s}_{-i}^{t-1}) + +$
  Observe the auction outcome o_i^t and receive the immediate reward r_i^t
  Estimate the state of other SUs $\tilde{s}_{-i}^t$ and update the corresponding transition probabilities as described in Sections V-B and V-A
  Compute the value of state $s_i^t, \tilde{s}_{-i}^t$ using (10)
  Update the Q value $Q_i^t(s_i^t, \tilde{s}_{-i}^t, b_i^t)$ according to (9)
end for

The results in [21] indicate that the greedy strategy that always chooses an action which maximizes the Q values may not provide enough exploration for the user to guarantee optimal performance. A very common approach is to add some randomness to the policy. We use ϵ-greedy with decaying exploration in which, the SU chooses a random exploratory bid at the joint state s with probability $\epsilon(s) = c/n(s)$, where $0 < c < 1$ and $n(s)$ is the number of times the joint state s has been observed so far. The SU chooses the greedy Q-maximizing bid (i.e. (11)) with probability of $1 - \epsilon(s)$. In this approach the probability of exploration decays over time as the SU learns more.

The learning-based bidding algorithm for SU i is summarized in Algorithm 1. The time complexity of the algorithm is dominated by finding state values (10) which can be done in $O(|S_i| \times N_i \times |B_i|)$, where $|S_i|$ is the state space size for SU i, N_i is the number of classes for other SUs' states, and $|B_i|$ is the bid space for SU i. In terms of space complexity, the SU needs to keep a table of size $|S_i| \times N_i \times |B_i|$ for Q values.

VI. Numerical Results

In this section, we evaluate the performance of our proposed bidding algorithm[3]. We compare our learning-based bidding algorithm (Algorithm 1) versus truthful bidding which is known to be the optimal bidding strategy without budget limits. In truthful bidding, an SU bids its true valuation when budget allows, and bids zero if the remaining budget is lower than true valuation. Since bidding algorithms intend to maximize utility of an SU, our performance metric of interest is the accumulated utility that an SU obtains over time.

The parameters in our numerical evaluations are set as follows. The SU starts with initial budget of 500, its valuation

[3]Due to space constraints, we only provide one comparison diagram with truthful bidding here, and we put more results in our technical report [17].

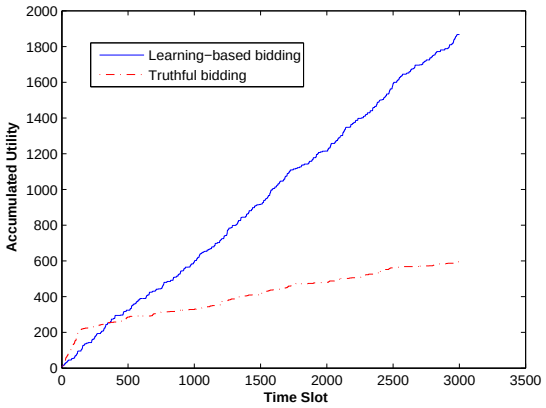

Fig. 2. The proposed learning-based bidding algorithm outperforms truthful bidding after the first 300 rounds.

at each time slot is drawn randomly from discrete uniform distribution with maximum of 10, and SU's budget evolves according to (3). The discount factor δ is set to 0.8, the fixed income of SU for getting channel access, a, is 2, the learning rate α is constant over time and equals 0.5. We set the number of classes (intervals) for representing other SUs to $N = 5$, and we choose 0.2 for the constant c in Algorithm 1. The auction is repeated for 3000 rounds.

Fig. 2 shows the accumulated utility of an SU using our learning-based bidding versus truthful bidding. As can be seen, our proposed algorithm outperforms truthful bidding after the first 300 rounds. This is due to the fact that truthful bidding does not take into account budget planning. Therefore, the SU bids aggressively at first, which significantly reduces its remaining budget to the extent that the SU does not have enough competitive ability for the remaining auction rounds. On the other hand, our learning-based bidding method considers the effect of bids on future and plans the budget wisely, which results in a better performance in the long run.

VII. Conclusion

In this paper, we studied the bidding problem of a budget constrained SU in repeated secondary spectrum auctions. We presented an MDP formulation of the problem and characterized the optimal bidding strategy of an SU, assuming that opponents' bids are i.i.d. Then, we generalized the formulation to a Markov game that allows an SU to make inferences about its opponents based on local observations. Using model-free reinforcement learning approaches, we proposed a fully distributed learning-based bidding algorithm which relies only on local information. Through numerical evaluations, we showed that our learning-based bidding method outperforms truthful bidding, in terms of utility.

Acknowledgment

This material is based upon work partially supported by the National Science Foundation under grant numbers 1422153 and 1456887.

References

[1] FCC Spectrum Policy Task Force, *Report of the spectrum efficiency working group*, Available: http://www.fcc.gov/sptf/reports.html, Nov. 2002.

[2] M. Khaledi and A. Abouzeid, "Auction-based spectrum sharing in cognitive radio networks with heterogeneous channels," in *Information Theory and Applications Workshop (ITA), 2013*, 2013, pp. 1–8.

[3] M. Hoefer and T. Kesselheim, "Secondary spectrum auctions for symmetric and submodular bidders," *ACM Trans. Econ. Comput.*, vol. 3, no. 2, pp. 9:1–9:25, Apr. 2015.

[4] Z. Han, R. Zheng, and H. Poor, "Repeated auctions with bayesian nonparametric learning for spectrum access in cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 3, pp. 890–900, March 2011.

[5] K. Akkarajitsakul, E. Hossain, and D. Niyato, "Distributed resource allocation in wireless networks under uncertainty and application of bayesian game," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 120–127, August 2011.

[6] L. Duan, J. Huang, and B. Shou, "Competition with dynamic spectrum leasing," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, April 2010, pp. 1–11.

[7] D. Niyato, E. Hossain, and Z. Han, "Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: A game-theoretic modeling approach," *Mobile Computing, IEEE Transactions on*, vol. 8, no. 8, pp. 1009–1022, 2009.

[8] A. Min, X. Zhang, J. Choi, and K. Shin, "Exploiting spectrum heterogeneity in dynamic spectrum market," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 12, pp. 2020–2032, 2012.

[9] P. Cramton, "The fcc spectrum auctions: An early assessment," *Journal of Economics and Management Strategy*, vol. 6, no. 3, pp. 431–495, 1997. [Online]. Available: http://dx.doi.org/10.1111/j.1430-9134.1997.00431.x

[10] FCC, *Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550-3650 MHz Band*, Available: https://apps.fcc.gov/edocs_public/attachmatch/FCC-15-47A1.pdf, Apr. 2015.

[11] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 4, pp. 1904–1919, May 2009.

[12] J. Feldman, S. Muthukrishnan, M. Pal, and C. Stein, "Budget optimization in search-based advertising auctions," in *Proceedings of the 8th ACM conference on Electronic commerce*. ACM, 2007, pp. 40–49.

[13] K. Amin, M. Kearns, P. Key, and A. Schwaighofer, "Budget optimization for sponsored search: Censored learning in mdps," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, 2012, pp. 54–63.

[14] R. Gummadi, P. B. Key, and A. Proutiere, "Optimal bidding strategies in dynamic auctions with budget constraints," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, Sept 2011, pp. 588–588.

[15] M. Khaledi and A. Abouzeid, "Dynamic spectrum sharing auction with time-evolving channel qualities," *Wireless Communications, IEEE Transactions on*, vol. 14, no. 11, pp. 5900–5912, Nov 2015.

[16] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. New York, NY, USA: Cambridge University Press, 2007.

[17] M. Khaledi and A. Abouzeid, "Optimal Bidding in Repeated Wireless Spectrum Auctions with Budget Constraints," Tech. Rep., 2016. [Online]. Available: https://www.dropbox.com/s/5b1qs5lvi25fs9o/techrep.pdf?dl=0

[18] M. L. Littman, "Value-function reinforcement learning in markov games," *Cognitive Systems Research*, vol. 2, no. 1, pp. 55 – 66, 2001.

[19] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge England, 1989.

[20] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[21] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policyreinforcement-learning algorithms," *Mach. Learn.*, vol. 38, no. 3, pp. 287–308, Mar. 2000.