

ENERGY EFFICIENT *BID* LEARNING PROCESS IN AN AUCTION BASED COGNITIVE RADIO NETWORK

Oloyede, Ayopo Abdulkarim¹ and Grace, David²

¹Department of Telecommunication Science, University of Ilorin, Ilorin, Nigeria

²Communications and Signal Processing Research Group, University of York Yo105DD, York UK

ABSTRACT

This paper proposes a learning based auction model for cognitive radio network using the concept of Bayesian and Q-learning. A learning process is introduced to aid energy efficiency in an auction based cognitive system. By using Q-learning to learn the bid price, this paper showed that for the learning users, the amount of energy consumed per file sent can be reduced when compared to the non-learning users. Furthermore, to overcome the deficiencies of traditional Q-learning we bias the exploration process with Bayesian learning. This helps the exploration process to converge faster, thereby further reducing the energy consumption by the learning users in the system and the system delay.

Keywords: Machine Learning; Dynamic Spectrum Access; Cognitive Radio; Q-learning; Spectrum Auction.

1. INTRODUCTION

Cognitive radio network is as wireless systems that are designed to interact and observe the radio environment. They learn based on the interaction with the environment to allow for opportunist access to the radio spectrum (Akyildiz, Won-Yeol, Vuran, and Mohanty, 2008). Based on the proposed concepts for the implementation of cognitive radio an auction model is a fair process that can be used in opportunistic access of the radio spectrum, Jia et al. (2009) agreed that. An auction process allows both the users and the service provider to gain some market knowledge about the value attached to the radio spectrum, given the dynamic nature of a wireless network. Before the use of an auction for a cognitive based wireless network, there is a need to study how the bidding/ auction process can be applied to a cognitive wireless system. This is because the use of an auction could introduce delay into the system especially when

there is an unannounced reserve price (Oloyede and Grace, 2013b). Furthermore, an auction process for DSA is slightly different from a conventional auction because energy and other resources are consumed to conduct the auction itself.

Oloyede and Grace (2013b) designed an auction model for dynamic spectrum access, showing that the players, who are the users generating the bids, are likely to be software agents rather than humans. Therefore, there is a need to allow the players to have complete information or have the ability to independently use learning algorithms to decode the incomplete information that might be provided. In Oloyede and Grace (2013b), the bid equations are based on the assumption that the bidders have a prior knowledge of certain information such as the number of potential bidders submitting a bid in an

auction round. Such assumptions cannot always be valid especially when modelling a real world scenario. This is because the auctioneers might not know the number of potential bidders that are submitting a bid until after the bids are submitted. Even when such an assumption is valid the number of bidders has to be fed back via a controlled and reliable medium. However, to design a simple and effective bidding process for a wireless network, the amount of information exchange between the bidders and the auctioneer must be reduced to the minimum. This will reduce the delay and overhead involved in unnecessary information exchange. Furthermore, the problem of energy wasted as a result of the rejection of a large number of bids needs to be addressed. This is because the losers during the auction process waste energy in participating, whereas they could be better off not participating. Hence, the learning process might help the bidders learn the appropriate bid value that is above the reserve price so that the user does not pay too much for the radio spectrum. This is because machine learning allows for historical information to be incorporated. Learning can also help in reducing the amount of energy wasted as a result of losing an auction process.

This paper examines if machine learning is suitable for DSA, the different types of Machine Learning model that can be used in an auction based DSA scheme. Machine Learning is based on the concept of artificial intelligence which involves solving problems based on experience (Zhe and Qiu, 2011). The experience might be transferred or be based on a form of reward and penalty which accumulates additively. This can be with or without a discount factor. The learning agent uses the information learnt at time t based on action $a_t \in A(S_t)$ to transit from one state (S) to another at time $t + 1$, receiving a reward (r_t). The main objective is to maximise such reward ($\pi: S \rightarrow A$). Where $A(S_t)$ and π is the set of available action and policy at time t respectively. In this paper, only few of the users in the system adopt the learning model. This is because if all the users in the system adopt the same method of

learning, the learning process might have a difficulty in achieving convergence. This is because all the users are learning about the others and they might just keep increasing their bids until they run out of budget. Hence this paper examines if learning the best bidding price can help in improving the efficiency of the auction process and the system.

Recently, the concept of a learning based auction process for Dynamic Spectrum Access (DSA) has emerged in the literature. Zhe and Qiu (2011) proposed a Q-learning bidding algorithm for secondary users. Zhu, Rong and Poor (2011) modelled a repeated auction learning process for secondary users based on a distributed learning algorithm for a single channel network based on a second price auction with entry fees. The bidders can either choose to bid or stay out depending on the user's requirement. The results show that the schemes perform better in resource allocation compared to a bidding process without learning. Lunde et al. (2011) proposed a double auction reinforcement learning to improve performance in terms of packet loss, bidding efficiency and transmission rate of the secondary users. In this work, we model dynamic spectrum access for cognitive radio networks using a multi-winner first price sealed bid auction with a reserve price. An auction process is carried out at the beginning of each auction period. Each user that requires access to the radio spectrum submits a bid at the beginning of every period, after which a number of users equal to the number of available channels emerge as winners. The winners are allocated the channel for a fixed duration of time. This marks the end of a bidding round after which another bidding round begins. In some parts of this work, we assume that users can either bid myopically or learn how to bid using Machine learning. Myopic bidders pick a low value randomly within a range of values without considering the reserve price and continue to use such value. Learning bidders learn the appropriate bid value. Such value should be above the reserve price set by the Wireless Service Provider (WSP). It should also give the best possible utility value to the user. This

is to prevent the rejection of bids leading to energy wastage. Furthermore, we observe and compare the amount of energy consumed by the myopic and learning users. The learning process is carried out using Q-learning. We also examine a technique to speed up the exploration process by biasing the Q-learning using Bayesian learning.

The rest of this paper is organised as follows: The

modeling parameters are discussed in section I. Section II explains the learning model. The simulation scenario is given in section III Section IV, evaluates the scenario where learning can be used in an auction based DSA process. Based on the scenario where learning is applicable, the performance of different learning techniques is analysed in section V. Conclusions are drawn in the last section.

2. SYSTEM MODEL

An uplink infrastructure network is considered consisting of the users, the spectrum broker and the database. The elements in the network are as described in Figure 2. The system consists of N users and N_{TC} transmission channels in each cell. During an auction period (t), the number of users seeking access to the radio spectrum are time varying. They are represented as N_{USA} . After the auction process, a number of winners emerge. These winners are represented as N_{WU} . A channel for transmission can only be utilised by one user at a time and users cannot communicate directly to another user except via the central entity (base station). During any auction period, the database provides the information to the auctioneer regarding the available channels in each cell before the auction process. The number of available channel in each cell as provided by the database is represented as N_{AC} .

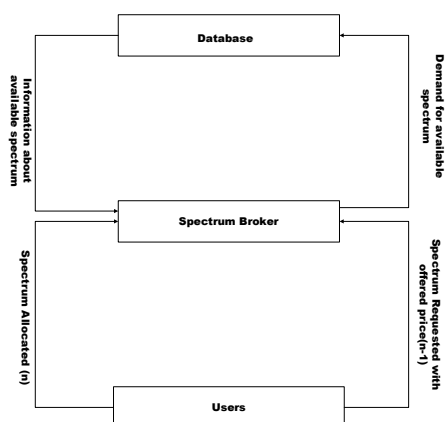


Figure 2. The database approach to spectrum auction

After the auction process, the winning users are allocated a channel, but not all winning user that is allocated a channel is able to transmit successfully because of inference or noise from adjacent users sharing same channel. N_{UT} is used to represents the number of users who are able to transmit after the auction and the allocation process and for the reason stated earlier $N_{UT} \leq N_{WU}$.

2.1 The Users Bid

The bid of a user is the price offered to gain access to the radio spectrum. In our previous work in (Oloyede and Grace, 2013b), the bid or the valuation of a user is determined using a random process. However, to simplify and understand the analysis done in the later part of this work, it is assumed that the offered bid that can be picked by each of the users is from a fixed range of discrete values and the values depend on the traffic arrivals (N_{USA}). Therefore, the range is referred to as the traffic load bin. The range for each traffic load bin is determined by first examining the average minimum and maximum number of users requesting the use of the channels. The number of users arriving depends on the traffic load in the system. Hence, the average minimum and maximum number of users arriving for each traffic

load bin ($N_{USA_{min}}^{T_L}$ and $N_{USA_{max}}^{T_L}$, where superscript T_L is the traffic load) are respectively determined as:

$$L_p = \begin{cases} \lfloor C_r N_{USA_{min}}^{T_L} \rfloor, \\ \lfloor C_r N_{USA_{max}}^{T_L} N_{TC} \rfloor, \dots \end{cases} \quad (1)$$

Where: L_p represents the traffic load bins

For the lower range, it is assumed that $N_{AC} = N_{TC}$. This is in order to determine the minimum possible reserve price. For the upper range (for each traffic loads), it is assumed that $N_{AC}=1$, in order to determine the maximum possible reserve price for a particular traffic load. The value of N_{USA} used in both cases is the average arrival for each traffic load. The range is then approximated to the nearest integer number as stated in table 1. Hence, it is assumed that each of the users intending to transmit on the radio spectrum picks an offered bid value from one of the bins depending on the belief of the user regarding the traffic load in the system. Five traffic load bins are assumed in this paper as shown below:

Table 1: Traffic Load Bin

Traffic Load	Value of offered Bid
1	12-40
2	35-45
3	40-50
4	45-55
5	48-60

The traffic load bins reflect the increase in competition occurring in the system as the traffic load increases. As shown in the table, the values overlap for each traffic load because the estimated value of N_{USA} is overlapping too. In some other scenarios (not considered in this

work), the bidding values in each of the bins can be associated with the quality of service that the user expects from the WSP.

2.2 Auction Process

The auction model adopted in this work is the simultaneous sealed bid auction where no bidder knows the bid of any other user and in the same auction period, no user can amend the offered price after putting in the bid. The simultaneous process is adopted in order not to introduce additional delay into the system. For an auction process to take place, the value of the N_{USA} is always greater than the N_{AC} . Hence this is taken into account when formulating the auction period later in this paper. During the auction process, the amount paid by any of the winning bidders depends on bids submitted by the users.

2.3 The Reserve Price

The reserve price (r) is the minimum price paid before the spectrum is allocated. It is introduced because the demand for the radio spectrum is both time and space dependent and therefore, when the demand for the radio spectrum is low, the reserve price helps to retain the minimum selling price of the service provider. The reserve price might also prevent the use of the spectrum if the value is set too high. Therefore, the reserve price is formulated by taking into account the current traffic load in the system, the frequency band in use, the total number of channels in the system and the number of channels in use. The reserve price is expressed as:

$$r \text{ (price unit)} = \frac{N_{USA} C_r N_{TC}}{N_{AC}} \quad \dots \quad (2)$$

2.4 The Energy Model

The energy model used for the mobile users in an up-link configuration in this paper can be represented as a 2 state Markov chain as shown in figure 1 below. In the

energy model a user can transit from 1 to 4 as shown and explained below:

1. A user who has file(s) to send moves into the OFF state and continue to be in this state until such user is among the winning bidders.
2. A user who is among the winning bidders moves from the OFF state to the ON state.
3. The user remains in the ON state until after transmission if transmission is successful or until when the user receives a failed signal either due to low offered bid compared to the reserve price or due to poor quality channel.
4. After transmission the user moves back to the OFF state before switching completely off if no file is to be sent again. However if the user has another file to send, the user remains and attempt again in the off state. The complete off mode (not in figure 1) is the mode a user is in when there is no file to be sent.

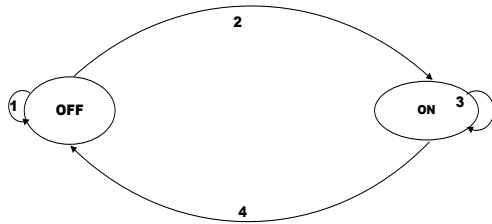


Figure 3: Energy and system model as a two-state Markov chain

In the ON state, it is assumed that a transmission is successful provided the Signal to Noise plus Interference Ratio (SNIR) and the reserve price are above the set thresholds, after which the user moves back to the OFF state. If the user moves from the ON state to the OFF state, and the thresholds are not met, then the energy consumed in processing the request of the user during the state transition is considered as energy wasted. A processing time which is the time taken to process the received bid is also assumed. All users that move from the ON state to the OFF state have the same processing time.

2.5 Users Utility Function

The utility function plays an important role in determining the achievable performance of a system. It describes the level of satisfaction or the preference of a user based on the QoS received (Youping, Shiwen, Neel, and Reed, 2009). It can be used in radio resource management to determine the level of satisfaction of the users. The utility function can be described using different ways, but the choice of the function is critical in achieving the desired performance. The utility function, $u_i(t)$, adopted in this paper shows the satisfaction derived by a user from the value of the bid placed. It is assumed that users want to win the bid with the least possible amount. The utility function measures how much a user deviates from the lowest winning bid. Without a loss of generality, we use an exponential utility function because of its rapidly increasing nature. In a bidding round, a bid below the reserve price gives a utility value of zero. The lower the value of the winning bid the higher the value of utility. The closer the bid of a winning bidder to the minimum winning bid the higher the utility of the user. If the user is not among the winning bidders then the utility is zero. Set K contains the winning bids and b_1 to b_k represents the bids of the winning users

$$K = \{b_1, b_2, b_3 \dots b_k \dots\} \quad (3)$$

$$\delta = \min K$$

$$(4) u_i(t) = \begin{cases} 2^{\frac{\delta}{b_i}} - 1 & \text{For winning bidders} \\ 0 & \text{otherwise} \end{cases} \dots (5)$$

Where set N_{WU} contains the bids of the winning bidders and b_i represents the bid of user i . A power utility function is used because of its rapidly increasing nature as explained in chapter 5. The closer the bids of the winning bidders to the minimum winning bid the higher the utility of such bidder. If a user is not among the winning bidders then the utility of such user is zero.

3. THE LEARNING MODEL

In wireless networks, Reinforcement Learning (RL) is the most widely learning model used (Hui and Brown, 2000). Hence, reinforcement learning with some modifications is examined.

3.1 Linear Reinforcement Learning

Reinforcement learning involves the interaction of learning agents with an environment. The agents are learning from the environment by using a process of trial and error (Wiering and van Otterlo, 2012). A learning user obtains a reward after exploring each of the available actions or the possible bidding values, as is the case in this paper. This is done to understand the consequences associated with each bidding value. After the exploration process, the learning users begin exploiting the best action learnt.

Linear Reinforcement Learning (LRL) (T.Jiang, 2011) is usually described using a Markov decision process and using a tuple such as $(\mathbf{S}_s, \mathbf{A}_s, P(s|s'), \mathbf{R}_s(s|s'), \gamma_s)$. Where $\mathbf{S}_s$ is set containing the finite states available, $\mathbf{A}_s$ is the set containing the finite number of actions available at each state $\mathbf{S}_s$, $P(s|s')$ is the probability that an action by user i leads to a another action s' , $\mathbf{R}_s(s|s')$ is the reward or punishment for a user and γ_s is the reward discount factor. Usually γ_s has a value between 0 and 1. $(s|s')$ represents the transition from an initial state s to another state s' . The learning process can be described as Markov because at time t , the current state is the only information required by a learning user to move from one state to the other since $|s' = [s_{t+1}|s_1 \dots s_t]$. This is true until a point when the learning process converges. The present state captures all the relevant historical information. Therefore the previous history is no longer necessary. The reward/punishment can sometimes be discounted, because the undiscounted reward might not fully represent the uncertainty in the learning process (Jiang, Grace, and Liu, 2011; Puterman, 2009). According to Wiering

and van Otterlo (2012), the optimal outcome ($V^{\pi^*}(s)$) can be learnt using the value function defined as $V^{\pi}(s)$. The learning agent updates its knowledge based on the feedback from the value function. There are different approaches that can be used to update a value function (Konidaris, Osentoski, and Thomas, 2011). However, this paper adopts a linear value function because of the simplicity associated with the updating rule (Konidaris et al., 2011). Linear approximation allows for a linear weight to be associated with the previous weight as shown below:

$$W(t) = \gamma_1 W_{t-1} + \gamma_2 W_t \quad \dots \quad (6)$$

Where $W(t)$ is the final weight, W_{t-1} is the weight at time $t - 1$, W_t is the weight at time t which is added to the previous weight to give the weight for the present time. γ_1 and γ_2 are the weighting factors associating with the learning process for each state.

The weighting factors may or may not be the same and the ways in which the weight W_{t-1} (the addition of $\gamma_2 W_t$) is updated in time t to give $W(t)$ is known as the updating rule. The summary of weighting factor used in as shown in table 2.

Table 2: Weight factor values

SCHEMES	Reward	Punishment
Discounted Reward $f(U(b_i))$	$f(U(b_i))$	0

The reward function $f(U(b_i))$ means that the weight of the reward is dependent of the value of the utility function based on the bid submitted as explained earlier. One of the main problems of reinforcement learning is its

slowness to converge (Wang, Won, Hsu, and Lee, 2012). This is because of the large number of trials involved for the learning process to be effective. It uses different methods to allow for faster convergence such as taking the average, using the greedy policy and Q learning. Some other methods can be found in MacKay (2003).

3.2 Q Reinforcement Learning (QL)

The policy adopted by the learning user usually affects the learning convergence speed. QL is a type of reinforcement learning which assigns values to pairs of the state actions. Every state has a number of possible actions and the reward earned depends on the action taken (scalar reward). The reward received by taking an action is based on how close the action taking is to the best known action. In the proposed auction process where users aim is to win with the least possible amount, the closer the bid of any of the winning bidder to the minimum winning bid the higher the reward obtained. This is because the acceptable values of the winning bids can be any value down to the reserve price. The Q learning algorithm is similar to the reinforcement learning earlier explained. However, the difference is in the update of the Q table. The Q learning for policy π is updated as shown below (Watkins and Dayan, 1992):

$$Q^\pi(s, a) = R_s + \gamma_s \max_a Q(s', a') \quad \dots \quad (7)$$

Where R_s is the reward obtained by observing the new state by taking action a , γ_s is the weighting factor and $\max_a Q(s', a')$ is the maximum possible reward for the new state that the learning user moves into by taking action a .

The QL approach used in this paper is similar to Sui and Leung (2009). The bidders only have information regarding their own bid history and not that of any other user in the system. A user can submit only a single bid in a bidding round. In this paper, the reserve price is not known to the bidders. This is because the reserve price is

set by the WSP whose aim is to maximise profit. However, in a real-world auction model, it is likely that if the reserve price is known, the bidders might only offer bids slightly above the reserve price hence, the reason this assumption is modified from the one made in Oloyede and Grace (2013a). An auction process is started by asking the bidders to submit their sealed bids to the WSP. Based on the value of the submitted bid, the bidders obtain a utility value using the utility equation in (5). The aim of the user is to win the bid with a high utility value, however, the lower the offered price of the bidder, the lower the probability of the bidder winning the bid and the higher the bid of the user, the higher the probability of the user paying too much for the spectrum compared to the price paid by others. Therefore, any of the two extremes can lead to a lower value of utility. Hence, the dilemma faced by a bidder is to decide the best bid that offers the best utility value.

Using the QRL, each of the possible bidding values has a reward received after a learning event. A learning event occurs after each iteration, meaning that after each bidder submits a bid value, a reward is obtained. The reward is added to the previous reward weight. After the conclusion of the learning exploration, the learning user exploits the bidding value that gives the highest weight. The reward is obtained as explained below:

First, all the possible bidding values for a traffic load bin (L_p) can be represented in an ascending order as shown in equation 8, where subscript p represents each of the traffic load bins (i.e. 1 to 5 in this case) and the superscript y represents each of the discrete ascending values which fall in the p^{th} traffic load bin:

$$L_p = [b_p^1, b_p^2, b_p^3, \dots, b_p^y] \quad \dots \quad (8)$$

A user picks one of the possible bid values and the winning ratio (W_R) as shown below. Where N_{SW} is the number of times a user selects and wins with a particular bid value divided by $N_{\tau_a}(t)$. This is the total number of iterations carried out up until time t for a bid value.

$$W_R(t) = \frac{N_{SW}(t)}{N_{\tau_a}(t)} \quad \dots \quad (9)$$

The winning ratio is then multiplied by the utility function obtained from equation 5 to obtain the weight as shown below:

$$R_W = W_R U \quad \dots \quad (10)$$

The reward associated with a bidding value is updated using to obtain the total weight as shown below:

$$W_p^y = \sum_1^t R_W \quad \dots \quad (11)$$

Where: W_p^y is the learning weight which is the summation of the reward the users obtained for learning up to time t using a specific offered bid value from traffic load bin p .

The weight for all the possible bids in the single traffic load bin p (L_p) value can be as shown below:

$$w_p = [w_p^1, w_p^2, w_p^3 \dots w_p^y] \quad \dots \quad (12)$$

A learning user then picks the bid value that has the maximum weight (π^*) after the exploration period as shown below:

$$\pi^* = \max(w_p) \quad \dots \quad (13)$$

In some of the scenarios considered, where the user picks from any of the traffic load bins the weight for all the traffic load bins is as shown below in form of a matrix.

$$W = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & \dots & y \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ p \end{matrix} & \begin{bmatrix} w_1^1 & w_1^2 & w_1^3 & \dots & w_1^y \\ w_2^1 & w_2^2 & w_2^3 & \dots & w_2^y \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_p^1 & w_p^2 & w_p^3 & \dots & w_p^y \end{bmatrix} \end{matrix} \quad \dots \quad (14)$$

The reason for using this method is mainly because a scenario where the users want to win with the least possible amount is modelled hence, the utility function is Available online at www.bayerojet.com

used in obtaining the weights during the learning process.

The summary of the Q learning process is as shown below:

- 1: Users pick a bid value randomly from the bin.
- 2: The utility function of the user based on the bid is calculated.
- 3: Other records are also calculated in equation (8).
- 4: The weight table is computed at time t .
- 5: After τ trials, the best bidding value is exploited $\pi^* = \max(W)$.

Another learning method that is considered in this paper is the Bayesian reinforcement learning as explained below:

3.3 Bayesian framework for Reinforcement Learning (BRL)

The problems with uncertainty and delay associated with random exploration in reinforcement and Q-learning might lead to a sub-optimal convergence point if the exploration period is not long enough. This brings about the need to manage the uncertainty and enhance the exploration process. One such approach is Bayesian learning. The Bayesian algorithm allows the learning agent to make a decision based on the most likely events that could happen, using prior experience. Bayesian learning can be implemented in an auction based DSA as proposed in this work because the users are bidding very often to request for the spectrum. Therefore, the users use the previous bidding rounds as the prior knowledge and update it as the bidding progresses. This allows for a faster and smooth movement from the exploration to exploitation behaviour. Bayes' theorem is applied as shown below:

$$P_r(A \setminus B) = \frac{P_r(B \setminus A) \times P_r(A)}{P_r(B)} \quad \dots \quad (15)$$

Where: $P_r(A)$ is the prior probability distribution of

hypothesis A , $P_r(B)$ is the probability of the training data B (likelihood) and $P_r(B \setminus A)$ is the probability of A given B (posterior probability).

Bayesian learning offers several advantages over the previous model. It provides an ideal format to reach a compromise between the exploration and the exploitation stages by providing information on states which the player might not have explored. It also allows for the incorporation of prior probabilities to determine the user's transition. However, it has a disadvantage in that it can only be applied when prior information is available.

In this paper, a learning user generates the initial prior knowledge based on an assumption regarding the common knowledge of the price for the spectrum and the utility of the user. The transition matrix is assumed to be sparse as only a certain number of the bid values in the bins have a non-zero probability. The results examine the concept of uniform and non-uniform prior probabilities known as the rectangular and triangular prior probabilities respectively. The rectangular prior probability assumes the same values of the utility function for all possible bidding values except the bidding values which fall below the reserve price. They are assigned a zero prior probability. While for the triangular prior probability the utility function is calculated using equation (16).

$$\text{Prior probability} = U_i b_i \quad \dots \quad (16)$$

The posterior probability is calculated using the Bayes equation after converting the likelihood into a probability. After then the same process as explained in the Q-learning model is used except that the total number of iterations to be carried out (τ) is reduced. This is further explained in the simulation description. There are other methods of calculating the prior and the likelihood. A similar method can be found in (Singer, 1999). In a real world scenario, the prior probability distribution can be based on facts such as users bidding above or below a certain percentage of the guide price as as-

sumed in (Zeng and Sycara, 1998). The posterior distribution represents the uncertainty in the system for the transition probability for each of the state action pair as explained earlier. The summary is stated below:

- 1: u and the prior probabilities are calculated.
- 2: The likelihood is generated from QL and converted into probability.
- 3: Bayes rule is applied.
- 4: Users pick the action with the highest utility and the winning percentage (π^*).

3.4 Learning Rate (LR) and Learning Efficiency (LE)

Learning rate is generally a time-dependent factor which decreases as the time of learning increases. Generally, it is inversely proportional to the time it takes to learn.

$$LR = \left(\frac{1}{\tau}\right) t_a \quad \dots \quad (17)$$

Where: τ is the number of iterations before a steady state is reached and t_a is the time taken for an iteration to be carried out. It is assumed that the time taking for each of the iterations is the same. However, learning rate does not tell an accurate story about the learning process of an individual user. This is because training samples contribute differently to the final optimal value (Cherkassky and Mulier, 2007) therefore, the learning efficiency is examined as shown below:

$$LE = \frac{L_{UC}}{L_{TC}} \quad \dots \quad (18)$$

Useful learning cost (L_{UC}) is defined as the number of trials carried out while trying the offered bid value that eventually becomes the optimal bidding price while the total learning cost (L_{TC}) is the total number of trials before convergence. This paper examines 3 potential learning models that can be used given an auction based DSA network. The scenario is now formulated and subsequently examined.

4. SIMULATION SCENARIO

Assume that the traffic model is developed and driven by a Poisson process. This means that the length of the file and the interarrival time are exponentially distributed. To evaluate the performance of the proposed learning algorithms, an auction scenario using a multi-winner first price sealed bid auction with a reserve price is modelled. This is done in an uplink scenario with one WSP and N in the system. It is assumed that the users are connection oriented while using the WINNER II B2 propagation model as detailed in (Kyösti et al., 2007) for the propagation loss. All the channels in the system are assumed to be identical and no channel has a better quality than the other. However, users may experience different fading and path loss depending on the location of the transmitting users. Hence, each of the users in the same location, experiencing the same fading and path loss perceives the quality of the channels to be the same. N_{USA} out of the N possible users in the system request the use of the channel by submitting a bid during every bidding periods ($b_1, b_2, b_3 \dots b_{N_{USA}}$). The bids are picked from the traffic load bin as explained earlier. The value of N_{USA} varies depending on the traffic load in the system. The initial bidding value is chosen randomly from the possible bidding values in the bins. Based on the bids submitted by the N_{USA} bidders, N_{WU} winners emerge at any bidding period t . The number of winners that emerge is dependent of the number of available channels in the system therefore $N_{WU} = N_{AC}$. Such winners are allocated the channel after the auction process as shown in the flow chart in figure 3. However not all the users allocated the channels are able to transmit because of the reserve price and the channel quality. The number of users that are able to transmit after the auction and allocation period is represented as N_{UT} . The users that are allocated the channels at time t use the channel for a fixed period of time depending on the transmitting characteristics of the users. This is based on the TSB

in(Burr, Papadogiannis, and Jiang, 2012) before releasing the channel. Throughout this paper, only 4 users are picked randomly among all the users in the system to be the learning users, while the others are not learning.

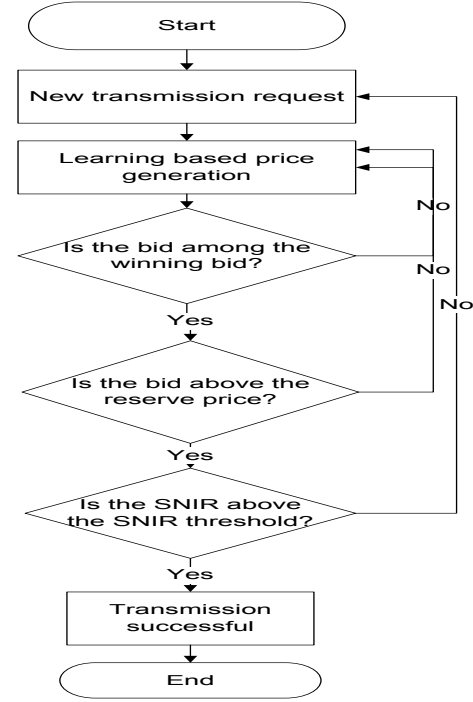


Figure 3: Learning Based Auction Process

The relationships that exist between the users in the system (N), users seeking the spectrum (N_{USA}) and the allocated users (N_{WU}) is as shown below and the total number of channels in the system (N_{TC}) and the number of available channels (N_{AC}).

$$N_{WU} \subseteq N_{USA} \subseteq N \quad \dots \quad (19)$$

$$N_{AC} \subseteq N_{TC} \quad \dots \quad (20)$$

During the allocation process, the winning users are

allocated the channels provided the offered bid is above the reserve price and the SNIR of each of the winning bidders is above the set threshold $SNIR_{threshold}$. The other parameters used are as specified in table 3.

Table 3: Parameters used

Parameters	Value
Cell radius	2km
Interference threshold	-40dBm
Users in a cell	200
Number of cell	19
Noise floor	-114 dB/MHz
$SINR_{max}$	21 dB
$SINR_{threshold}$	1.8 dB
Cr	0.7
Max number of channels per	4
Height of base station	15 m
Height of mobile station	1 m
Budget	100000 Price Units
Transmit power for users	0.9 W/bit
Energy consumed by device	0.5 Watt sec
Power used in bidding	0.25% of the transmit

4.1 Evaluation of Learning in an Auction Process

This section examines the scenarios where learning can be adopted and the implications of the learning process converging at a sub-optimal point. This is done in terms of the energy and delay characteristics of the system. The weight associated with each offered bid examined has a wide range of values. Hence the weights are normalized as shown below:

$$W_n = \frac{w_p^y}{N_{\tau_a}} \quad \dots \quad (21)$$

Where W_n is the normalised final weight, W_p^y is the weight for the y^{th} value in traffic load bin P and N_{τ_a} is the total number of events or iteration for a particular offered value for the entire exploration period. Before discussing the results, one important concept

relating to the understanding of the results is explained.

4.2 Learning Convergence

The learning process should lead to a convergence point as defined below:

Definition 1: The learning process is said to have converged if a user works out the optimal bidding price. In this work, the optimal bidding price must obey rules 1 and 2. **Rule 1:** The converged bidding price is the least price that gives the highest reward weight. **Rule 2:** The convergence bidding price must optimise the utility value of the user, allowing the learning user to consume the least possible energy and the least possible delay in the system.

To prove the above rules, the different scenarios are examined and some assumptions are relaxed in order to understand the effects of such assumption on the performance of the system. The scenarios are modelled with the assumption that the auctioneer knows the optimal bidding price. This means that after the bids are submitted and analysed the auctioneer is aware of the price unit that gives the best utility performance. This assumption is reasonable because in a real world auction scenario, the auctioneer has the knowledge of the optimal bidding price after analysing the submitted bids. This is because the auctioneer collects all the bids from all the sources and determines the winner. However, the users do not have this knowledge hence, the need for learning the optimal bidding price.

It is a difficult task to show the performance of each of the bidding values and there is no point showing the values below the reserve price since such values can never be accepted by the auctioneer in the model. Hence the weight of some of the bid values after a number of iterations is shown in table 3. Based on the results in table 4, the optimal value that allows the bidder to win at least 50% of the time is above 49 price units.

Table 4: The number of times a bid value is used successful by a learning bidder

Bid Value	50	100	200	300	400	500
47	0	0	0	0	0	0
48	0	0	0	0	0	0
49	4	12	19	49	71	96
50	27	57	113	184	255	343
51	39	78	155	240	323	401
52	41	92	189	286	356	463

The table shows that the reserve price isok above 48 price units at a traffic load of 4 Erlangs. No user can win with any price below 49 price units. Therefore in the remaining results in the sub section, only values above 49 price units are explored. In other results, a similar range is examined depending on the traffic load in the system.

Before going into the modeling scenario, the importance of introducing equation 6 in the learning process is examined along with the convergence when more than one value has the same learning weight is determined. It can be seen that because the users want to win with the least possible amount, the utility equation allows the learning process to converge at the lowest price that ensures the least amount of energy is consumed by the system. If the price utility function is not taken into consideration there might be more than one convergence point after the exploration period, then from rule 1, the convergence value is the least value. This must be achieved before the learning process converges. The proof of this is quite simple. This is because for a bidding price to

converge to the optimal bidding price, it must win the auction process most of the time. Hence, if a value wins the auction process, then any value above such value must also win assuming it was offered by a bidder in the same system. It can be seen from table 5 which shows the learning weight at a traffic load of 3 Erlangs that when the utility function is not taking into consideration that more than one bid value give the same weight value (i.e. 50-55 bid values). If this occurs from rule 1, the least offered price is the best bidding price however, to avoid this from happening, the utility function as defined in 5 is used so that the bid value that performs best can be easily identified.

Table 5: Weight of possible bidding values in traffic load bin

Bid	4	4	4	4	4	5	5	5	5	5	5
Weigh	2	3	3	3	4	9	9	9	9	9	9

From the above table, the converged bidding price is 50 price units based on rules 1 and 2. The reason for having the same weight after 50 price units is because any value above 49 price units the learning bidder wins all the time. It can also be seen from the above table that if the utility is not taken into consideration, it is difficult to have an optimal (single) value with the highest reward weight. This helps to show how the convergence in this work is defined and why equation 6 was introduced.

5 RESULTS AND DISCUSSIONS

In order to show the consequences of the learning process converging at a non-optimal and the optimal point on the delay and energy consumption of the system, the total energy consumption and average delay per file sent for such convergence points are examined. Figure 4 shows the total energy consumed by only the learning

users in the system if the learning process converge between 48 price units and 51 price units for a traffic load of between 1 and 4 Erlangs. It can be seen that below 4 Erlangs, all offered bids consume a similar amount of energy.

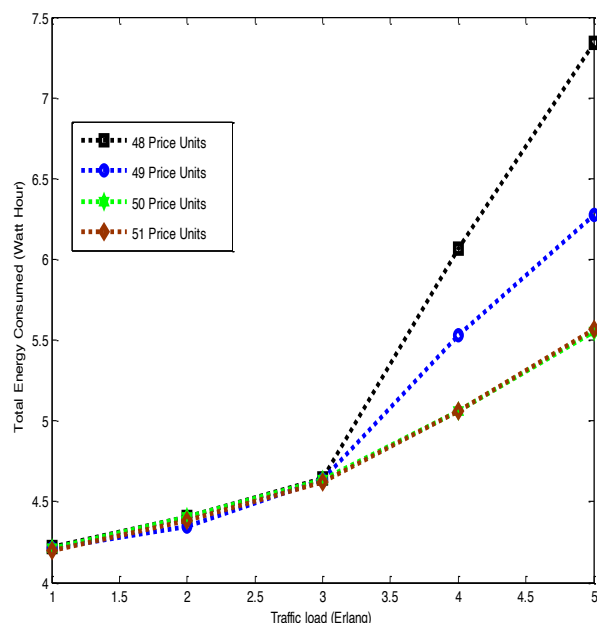


Figure 4: Energy consumed per file sent for different offered bid

This is because at lower traffic loads, such values (48 to 52 price units) are always higher than the bid offered by the non-learning users therefore, the exploiting users always win the auction process. However, at traffic load of 4 Erlangs, the offered bid of 48 or 49 price units gives higher energy consumed compared to offered bid of 50 and 51 price units. This is because with a traffic load of 4 Erlangs the offered bid of 48 and 49 price units are sometimes lower or the same as the bids of the non-learning bidders. Therefore, the non-learning bidders sometimes win the auction process at the expense of the learning users. The offered bid of 50 and 51 price units gives the lowest and the same amount of energy consumption per file sent at traffic load of 4 Erlangs because the exploiting users are able to win the bid using those values all the time. This again shows that at traffic load of 4 Erlangs the optimal price is above 49 price units.

Furthermore, the implications of the exploitation of such values on the delay experienced by the learning users in the system are examined in order to understand the effects. Figure 5 shows the average delay against Available online at www.bayerojet.com

traffic load for different offered bid values between 48 and 51 price units. It can be seen that below 4 Erlangs the energy consumption is the same for all offered bids and at 4 Erlangs the delay is almost the same for 50 and 51 price units. This is due to the reasons as explained earlier for figure 4.

Furthermore, 50 and 51 price units also gives the least amount of delay because the learning bidder wins all the time if they offer either 50 or 51 price units as their offered bid. From figures 4 and 5, it can be seen that the best offered bid that guarantees a bidder winning at 4 Erlangs of traffic load is either 50 price units or above. However in terms of the utility (using equation (5)), it can be seen from figure 6 that 50 price units offer a better utility to the user.

Another important parameter to show is the utility of the users when the learning process converges at the optimal and non-optimal point to understand how satisfied the users are during the auction process. Figure 6 shows the utility obtained by the learning users when exploiting 50 and 51 price units respectively. Only the two offered bid prices are examined because they give the least possible consumed energy and delay as shown in Figures 4 and 5 respectively. At an offered traffic level below 3 Erlangs, the utility obtained by exploiting 50 and 51 price units is very low because the exploiting users are winning the bids but they pay too much for the spectrum.

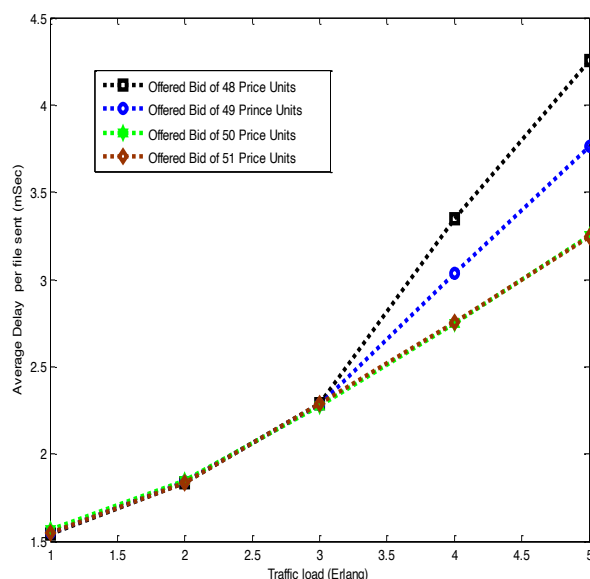


Figure 5: Delay per file sent for different offered bid

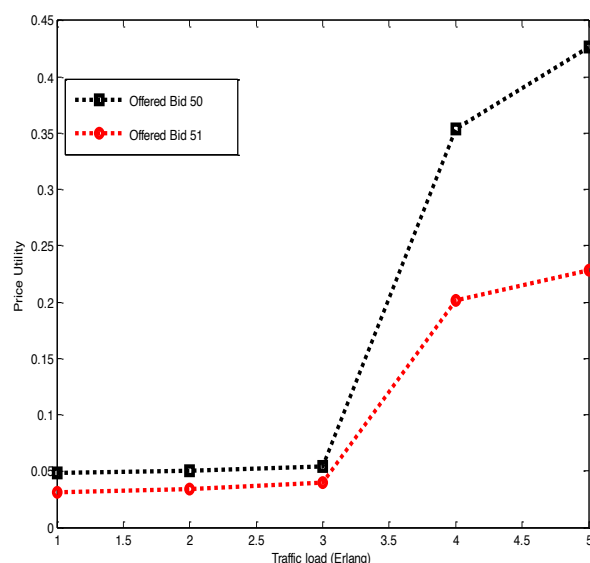


Figure 6: Utility obtained using 50 and 51 price units

This might lead to a situation known as ‘the winners curse’, which is defined as a situation where a bidding user exceeds the intrinsic value of the item purchased. At an offered traffic load of 4 Erlangs and above, the utility of the exploiting users improves because the bids offered by the non-learning users are relatively close to

what is offered by the exploiting users. As expected, the offered bid of 50 price units gives a better utility compared to 51 price units.

The above results have demonstrated that by using the proposed scheme the learning process can only be said to have converged if a price units of 50 gives the highest weight during the learning process. From the definition of the convergence in this scenario it can be seen that 50 price units is the convergence bidding price. This is because the convergence bidding price must use the least consumed energy, deliver the least delay and achieve the best utility as stated in rule 2.

5.1 Evaluation of Different Models of Learning

In order to examine how fast each of the learning models converges, the three learning models are examined and compared in this section. The price utility is taken into consideration and the traffic load bin as explained is used in modelling this section. Figures 7 (a-d) – 9 (a-d) show the weight of each of the offered bids after 100, 260, 400 500 iterations for figures (a), (b), (c) and (d) respectively when using linear reinforcement learning, Q-learning and Bayesian learning respectively. First, it can be seen from the all the above scenarios that at 500 events, the converged price gave the highest reward weight. This is because the price utility was taken into consideration by assuming that the users want to win with the least possible value. It can also be seen that it takes about 400, 250 and 100 iterations for the learning process to converge using reinforcement learning, Q-learning and Bayesian learning respectively. This shows that Bayesian learning converges fastest among the three learning methods. This is because Bayesian learning incorporates prior knowledge and this aids the learning process to converge faster than the other two methods examined.

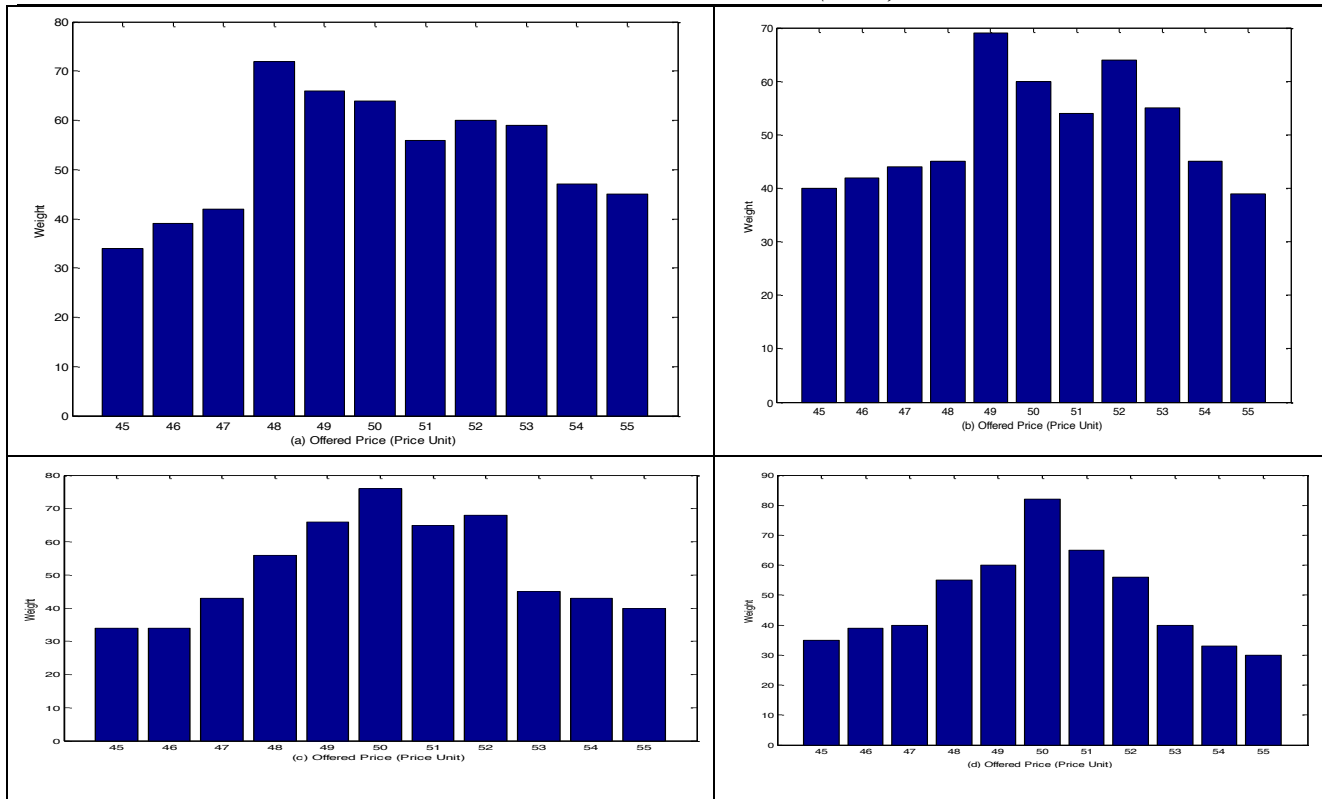


Figure 7: Linear Reinforcement learning for (a) 100 (b) 260 (c) 400 (d) 500 Events

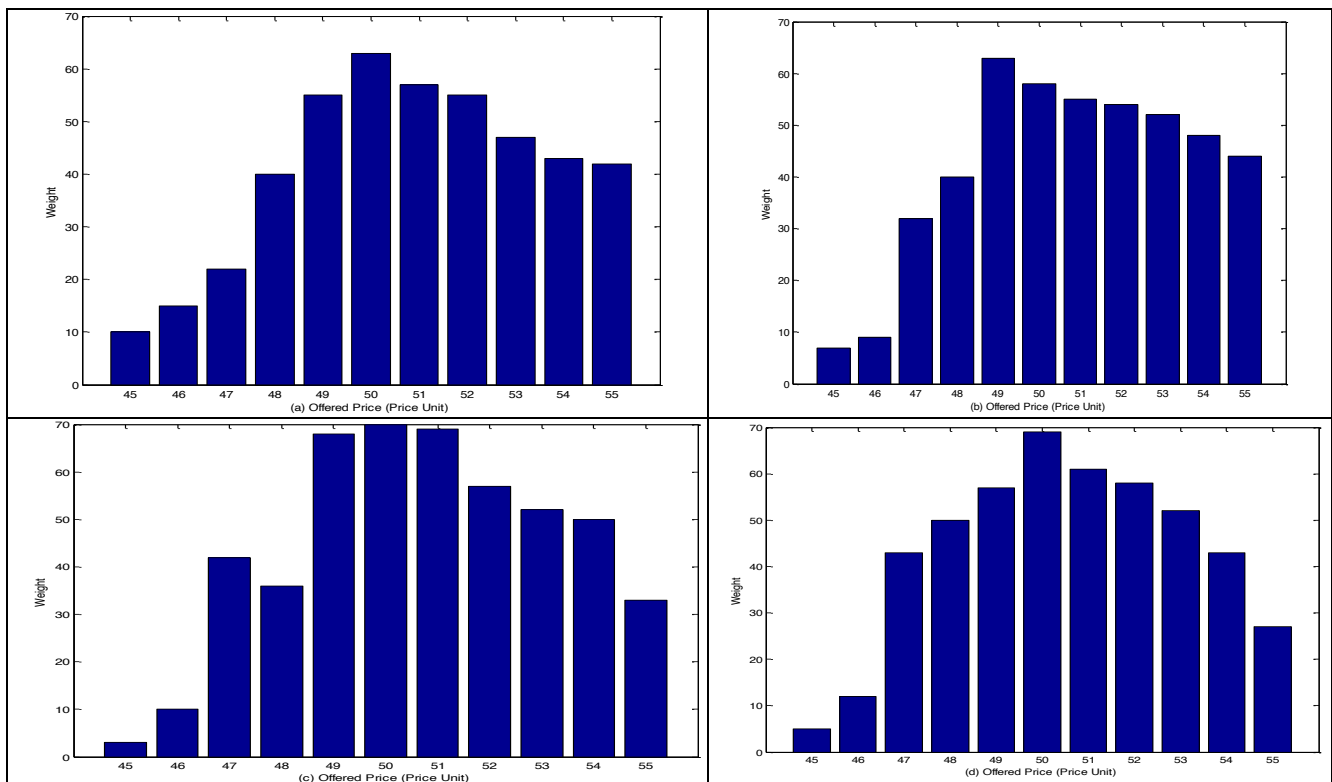
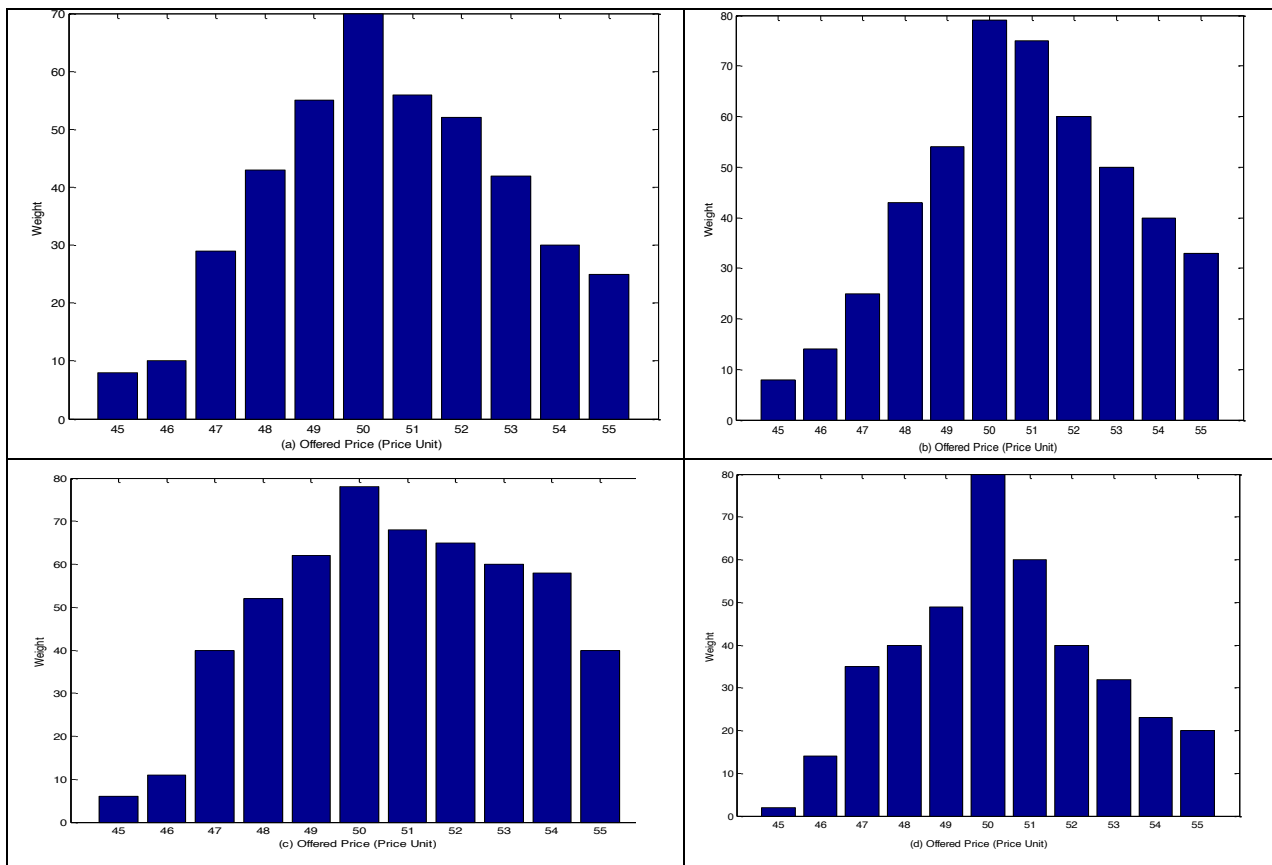


Figure 8: Q Learning for (a) 100 Events (b) 260 Events (c) 400 Events (d) 500 Events**Figure 9: Bayesian learning for (a) 100 (b) 260 (c) 400 (d) 500 Events**

After demonstrating that the Bayesian learning model converges fastest of the 3 methods, it is important to show the performance of the three learning models in terms of the energy consumed and the delay. This is to examine if the Bayesian learning method also helps with other performance metrics of the proposed system. Figure 10 shows the average energy consumed using different learning methods and that of the non-learning user. It is worth pointing out that the moving average is used in generating the results in this paper. This figure also shows that Bayesian learning which converges fastest compared to the other two learning methods gives the least consumed energy after convergence. It can be seen that for a learning user, as the number of events increases, the average energy consumed per file

sent decreases. This is as a result of the learning process moving closer to the optimal point during the exploration process.

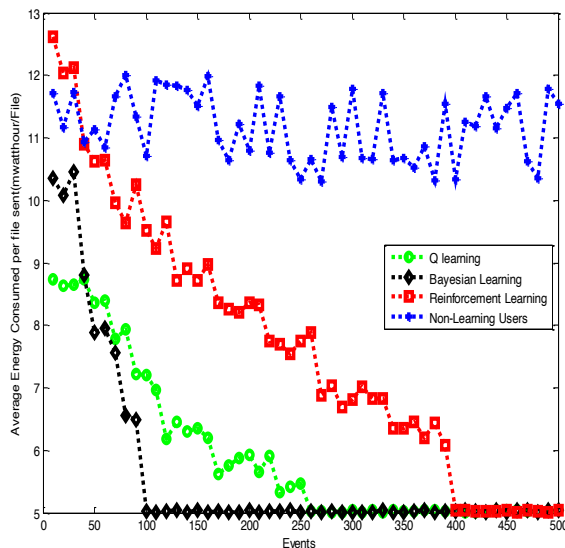


Figure 10: Average energy consumed for linear reinforcement learning Q-Learning, Bayesian learning and non-learning bidders

As the number of events increases, the user learns the optimal bid until a steady state is reached. For the non-learning bidder the average energy consumed at steady state is higher when compared with the learning users. This is because the user does not learn the optimal value, rather the user wants to maximise utility.

Another important parameter in any wireless network is the delay. Hence, the delay experienced during the exploration process is examined. Figure 11 shows the average delay per file sent in the system for the different learning methods and the non-learning users.

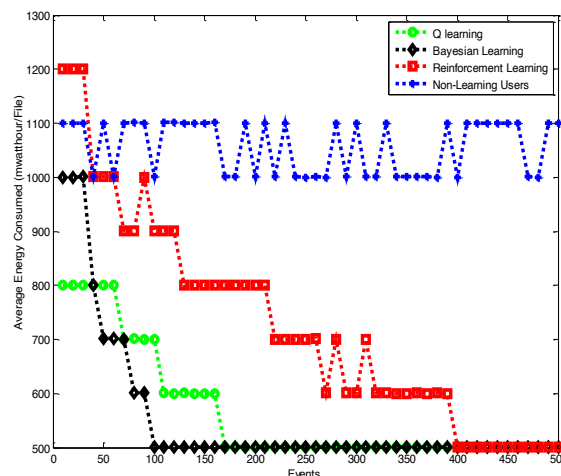


Figure 11: Average delay for reinforcement learning Q Learning, linear Bayesian learning and non-learning bidders

The average used is the moving average. It can be seen that the average delay reduces as the number of events decreases for the learning users. This is because as the learning progresses, the users learn to bid above the reserve price and reduces the probability of a file being rejected as a result of price. Just like in the energy graph, Bayesian learning reaches a steady state faster than Q-learning and linear reinforcement because of the prior probability that has been incorporated. This shows that Bayesian learning helps in reducing the delay in the system during the learning process.

5.2 Learning Rate and Learning Efficiency

So far the performance of the system using modeling techniques has been shown. The three learning models using the learning rate and learning efficiency equations as explained earlier are now examined. These metrics are also important because they demonstrate how effective and how fast the models converge. The results are shown in Table 6 below.

Table 6: Learning rate and efficiency of different learning schemes

Learning Model	Learning Rate	Learning Efficiency
Linear Reinforcement	0.0011	0.24323
Q Learning	0.0016	0.2645
Bayesian Learn-	0.005	0.27536

models it can be seen that Bayesian learning performs best, since it takes the shortest time to converge. However, all the 3 models give approximately the same value in terms of the learning efficiency. This is because all the three models eventually converge at the optimal point and the useful learning increases as the total learning increases. This shows that all the three learning models can be used but Bayesian learning converges faster followed by Q-learning and then reinforcement learning.

In terms of the learning rate as calculated for all the three

6. CONCLUSIONS

This paper examined the concept of bid learning in an auction based dynamic spectrum access network. It showed how users can employ bid learning to learn the bid of their competitors so as to be able to place better bids using reinforcement learning, Q-learning and

Bayesian learning. Different learning methods were explored and the paper showed that Bayesian learning performs best in terms of low energy consumption and low delay compared to the other learning models such as Q-learning and reinforcement.

REFERENCES

- Akyildiz, I. F., Won-Yeol, L., Vuran, M. C., and Mohanty, S. (2008). A survey on spectrum management in cognitive radio networks. *Communications Magazine, IEEE*, 46(4), 40-48. doi: 10.1109/MCOM.2008.4481339
- Burr, A., Papadogiannis, A., and Jiang, T. (2012, 1-1 April 2012). *MIMO Truncated Shannon Bound for system level capacity evaluation of wireless networks*. Paper presented at the Wireless Communications and Networking Conference Workshops (WCNCW), 2012 IEEE.
- Cherkassky, V., and Mulier, F. M. (2007). *Learning from Data*: John Wiley and Sons INC.
- Hui, T., and Brown, T. X. (2000). Adaptive call admission control under quality of service constraints: a reinforcement learning solution. *Selected Areas in Communications, IEEE Journal on*, 18(2), 209-221. doi: 10.1109/49.824799
- Jia, J., Zhang, Q., Zhang, Q., and Liu, M. (2009). *Revenue generation for truthful spectrum auction in dynamic spectrum access*. Paper presented at the Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing, New Orleans, LA, USA.
- Jiang, T., Grace, D., and Liu, Y. (2011). Two-stage reinforcement-learning-based cognitive radio with exploration control. *Communications, IET*, 5(5), 644-651. doi: 10.1049/iet-com.2009.0803
- Konidaris, G., Osentoski, S., and Thomas, P. S. (2011). *Value Function Approximation in Reinforcement Learning Using the Fourier Basis*. Paper presented at the AAAI.
- Kyösti, P., Meilä, J., Hentilä, L., Zhao, X., Jämsä, T., Schneider, C., . . . Rautiainen, T. (2007). IST-4-027756 WINNER II D1.1.2 V1.2 WINNER II Channel Models.

- <http://www.cept.org/files/1050/documents/winners2020-final%20report.pdf>
- Lunde, x, n, J., Koivunen, V., Kulkarni, S. R., and Poor, H. V. (2011, 3-6 May 2011). *Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks*. Paper presented at the New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*: Cambridge university press.
- Oloyede, A., and Grace, D. (2013a). Energy Efficient Dynamic Spectrum Pricing for Cellular Systems using the Concept of Green Payments. *Submitted for Review with Wireless Personal Communications*.
- Oloyede, A., and Grace, D. (2013b). *Energy Efficient Soft Real Time Spectrum Auction for Dynamic Spectrum Access*. Paper presented at the 20th International Conference on Telecommunications Casablanca.
- Puterman, M. L. (2009). *Markov decision processes: discrete stochastic dynamic programming* (Vol. 414): Wiley. com.
- Singer, N. F. Y. (1999). Efficient Bayesian parameter estimation in large discrete domains. *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, 11, 417.
- Sui, X., and Leung, H.-F. (2009). *A Q-Learning based adaptive bidding strategy in combinatorial auctions*. Paper presented at the Proceedings of the 11th International Conference on Electronic Commerce.
- T.Jiang. (2011). *Reinforcement Learning-based Spectrum Sharing for Cognitive Radio*. (PhD thesis), University of York. Retrieved from http://etheses.whiterose.ac.uk/1960/1/Full_Thesis_Exams_Office_Tao_Jiang.pdf
- Wang, Y., Won, K. S., Hsu, D., and Lee, W. S. (2012). Monte carlo bayesian reinforcement learning. *arXiv preprint arXiv:1206.6449*.
- Watkins, C. J., and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.
- Wiering, M., and van Otterlo, M. (2012). *Reinforcement Learning: State-of-the-art* (Vol. 12): Springer.
- Youping, Z., Shiwen, M., Neel, J. O., and Reed, J. H. (2009). Performance Evaluation of Cognitive Radios: Metrics, Utility Functions, and Methodology. *Proceedings of the IEEE*, 97(4), 642-659. doi: 10.1109/JPROC.2009.2013017
- Zeng, D., and Sycara, K. (1998). Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, 48(1), 125-141.
- Zhe, C., and Qiu, R. C. (2011, 17-20 March 2011). *Q-learning based bidding algorithm for spectrum auction in cognitive radio*. Paper presented at the Southeastcon, 2011 Proceedings of IEEE.
- Zhu, H., Rong, Z., and Poor, H. V. (2011). Repeated Auctions with Bayesian Nonparametric Learning for Spectrum Access in Cognitive Radio Networks. *Wireless Communications, IEEE Transactions on*, 10(3), 890-900. doi: 10.1109/TWC.2011.010411.100838.