Comparison of Q-Learning and Genetic Algorithm for Narrow-Band Cognitive Radio Networks

Apurva
Communication Systems
Fraunhofer FKIE
Wachtberg, Germany
apurva.apurva@rwth-aachen.de

Stefan Couturier
Communication Systems
Fraunhofer FKIE
Wachtberg, Germany
stefan.couturier@fkie.fraunhofer.de

Michael Reyer
Institute for Theoretical Information Technology
RWTH
Aachen, Germany
reyer@ti.rwth-aachen.de

***Abstract*—Narrow-band communication is widely used in military operations. Due to the limited bandwidth, efficient usage of the available spectrum with limited overhead is required. In this work, we present an efficient Dynamic Spectrum Management model for distributed multi-hop networks using narrow-band waveform. The system design is made robust using a collaborative feedback technique. Each node maintains a channel availability matrix and historical channel performance information based on this feedback. The channel allocation and radio parameters are optimized using Genetic Algorithm and Q-Learning algorithm. We observe that the robust system design, along with Genetic Algorithm and Q-Learning, efficiently mitigates the interference impact on a transmission link and consequently improves the overall transmission success rate and throughput for an end-to-end transmission.**

***Index Terms*—Cognitive Radio Network, Multi-Hop, Distributed, Narrowband, Q-Learning, Genetic Algorithms**

I. Introduction

Tactical networks used in military often suffer from a higher probability of disturbances due to node mobility, interference, congestion etc. In a multi-hop tactical network, for a single transmission there are multiple transmission links. This leads to an additional challenge of an extended interference area, making it more susceptible to communication failure. Therefore, optimization for end-to-end communication is required [1].

Cognitive Radio Networks (CRN) are a promising solution to this issue. They have revolutionized the spectrum usage by introducing flexibility in reconfiguration of the radio parameters, e.g. the frequency for transmission. Hence, as compared to the traditional fixed spectrum allocation policy, the concept of Dynamic Spectrum Management (DSM) - facilitated by CRN - incorporates intelligence in the network nodes to perceive the RF environment and make use of the spectrum in line with the network goals [2]. Here, we study the application of DSM, in particular to a multi-hop network using a narrow band waveform.

DSM entails overhead in terms of the information exchanged between nodes to facilitate the dynamic usage e.g. exchange of information regarding the present spectrum availability, the next frequency to use, possible interference etc. For narrow-band CRN, the bandwidth is small. Hence, this overhead must be minimized for any efficient communication to happen.

In this work, we first introduce the system design to facilitate DSM in a multi-hop network with little overhead. For an end-to-end transmission, each node is able to monitor the channels independently as well as receive feedback from other nodes about the channel availability and quality information (e.g. Signal to Noise and Interference Ratio (SNIR) that can be achieved compared to other frequencies). Thus, the short-term channel performance in the vicinity is identified and the transmission is recovered as soon as interference is detected on any of the links with much less overhead.

In addition to that, the DSM solution is capable of learning; i.e. the more information about the spectrum is available over time, the higher transmission success rate and throughput performance is achieved. Thus, the long-term channel behaviour is optimized. For this purpose, we decided to use two of the widely used algorithms in decision making, Genetic Algorithm (GA) and reinforcement learning.

GA benefits from parallel evaluation of different dimensions and handling of constraints and is therefore a suitable choice for our problem [3]. Alternatively, reinforcement learning employs a learning strategy suitable to a continuously varying environment by taking actions and observing the reward. Q-Learning is a type of reinforcement learning where the transition probabilities to the next state are unknown [4]. In this paper, we analyze the two different approaches - GA and Q-Learning. Both approaches will be compared regarding their performance in a CRN.

Section II provides an overview about the related work. In section III, the system and its protocols to facilitate DSM in a multi-hop network are described. The Q-Learning and GA approaches developed for this paper are given in section IV. Section V presents some simulation results on the system, and section VI summarizes our findings.

II. Related Work

In this paper several concepts known from literature are used, which mainly address the MAC layer and the decision making of the CRN nodes. This section introduces the concepts and the background of our work.

Our system adapts the principles of the Logical Link Control (LLC) and the Media Access Control (MAC) layer of the NATO narrow-band waveform (NBWF) described in [5]. For applying DSM to the NBWF we need to consider the MAC strategies to transmit control information reliably [6]. This can either be done in a Common Control Channel (CCC) for all nodes, as proposed in [7], or by using multiple control channels, as described in [8]. We need to consider the constraint of using a single radio. As described in [9], control information exchange is often separated from user data traffic in a timely manner. Similar to [9], our system is organized in a way to separate control information exchange from user data exchange by providing specific time slots to each node for transmitting control information.

GA has widely been used in CRN and various solutions are available in literature. However, we observe that there is comparatively less focus on tactical networks which as stated previously can be multi-hop and distributed. E.g. in [10], the authors present a joint sensing and channel allocation problem. However, it entails the need of a cluster head which is not suitable for a distributed set-up in a tactical network. In [11], [12] and [13] the decision-making is limited to a pair of nodes, without the possibility of extending it to multi-hop scenarios. However, in [14], the authors formulate a matrix based GA implementation for multiple user case. Although, this is suitable for a centralised network structure and is not formulated for a multi-hop network, it is a good study for multiple user network. The authors in [13] use GA and optimize network throughput by solving a single objective. GA can also be formulated to optimize multiple objectives simultaneously. In [15], the selection of transmission parameters such as frequency, power, modulation etc. based on multiple objectives is demonstrated using GA. However, the authors focus only on the adaptation problem and do not demonstrate the communication protocol or the system design and the work is not extended to a multi-hop scenario. The authors in [16], solve a multiple objective problem using GA in a multi-hop scenario. However, the solution is on a network level and does not consider a distributed network. In this work, we demonstrate the how each node in a multi-hop network can individually make decisions based on its RF environment and yet co-ordinate for successful end-to-end transmission.

In [17], the various use cases of reinforcement learning including Q-Learning in DSM are studied. In [18], the channel selection is done based on the Q-values but no other transmission parameter is considered in this approach.The authors of [19] propose three different variations of Q-Learning algorithms for improving the success of transmissions. The results show that exploitation of actual channel condition improve the network performance. In [20], the authors demonstrate the stateless Q-Learning where the Q-value is associated with different channels. Both approaches, the exploitation of the channel conditions and the stateless Q-Learning concept, have been used for our research, but the algorithms behind them were adapted to our needs, as further described in Section III. In [21], the choice of transmission parameters such as frequency and modulation is done based on Q-Learning. However, it does not consider decision making and implementation for multi-hop network.

III. SYSTEM DESCRIPTION

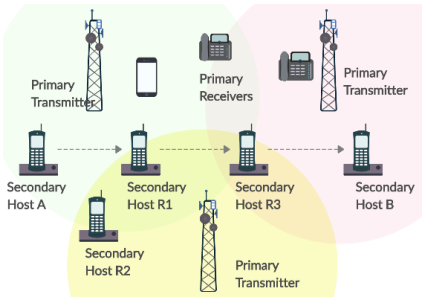

Fig. 1. Network Diagram

The network model consists of a network of primary users and a network of secondary users. In Fig. 1, the network diagram with three primary transmitters and five secondary hosts is shown. The primary network can be any network operating in VHF (Very High Frequency) radio band, i.e. its transmitters occupy some of the frequency bands which could be used by the secondary network. We focus our study on the design of DSM capable secondary network which is able to sustain communication in the face of interference from the primary transmission. In the subsequent sections, the complete network design for a secondary multi-hop CRN for opportunistic spectrum access and management is presented. Any DSM capable system should be capable of four tasks [2]:

- Sense the spectrum to determine spectrum opportunities.
- Negotiate transmission parameters using control signalling.
- Decide the optimal transmission parameters including the transmission channel.
- Vacate the channel on the appearance of the primary and continue its service with minimal disruption.

To accomplish these tasks, the communication framework including the frame structure, channel negotiation, and data transmission for a half-duplex cognitive node is developed. As described above, the design is based on the link layer of the NATO NBWF described in [5] with the following details:

- Band of operation is the VHF radio band.
- Channel bandwidth is 25 kHz.
- Opportunistic spectrum access by secondary users.
- Single radio (half-duplex communication) on a single channel at a time.
- Time-slot based transmission with a frame length of 202.5ms constituted of nine time slots.

A. Assumptions

The main aim of this work is to demonstrate the efficiency of multi-hop transmission in narrow-band for an end-to-end transmission. Therefore, we make certain assumptions for simplicity of design and analysis.

- The CCC has a fixed frequency that is not interfered.
- The routing path is fixed.
- The free-space path loss model is used.
- The system is synchronized, and there is no multipath time delay.
- No co-operation between the primary and secondary network.

B. Frame Structure

The frame structure defines the resources in time and frequency, based on the link layer design of the NBWF [5]. To segregate the tasks of DSM, different logical channels are defined as shown in Fig. 2.

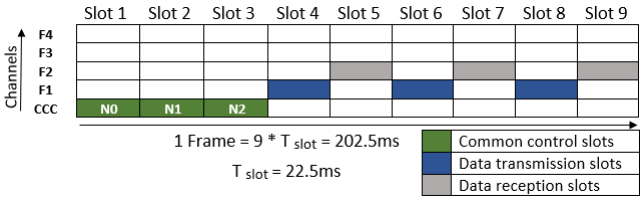

Fig. 2. Exemplary Frame Structure for an Even Hop Node

The CCC is used for the exchange of control messages for communication initiation, termination or other control signaling purposes. As all nodes use the same control channel frequency, each node has a fixed control slot for transmission to avoid interference. The control slots are the first three slots of the Control Sub-Frame (CSF), where a CSF block defines the number of CSF frames required for all the nodes in a network to be assigned a control slot. The CSF blocks are interleaved with a Sensing Sub-Frame (SSF), where a SSF is a frame where the first three slots are used for sensing. In each slot, a node senses one channel to check if it is occupied or not. A block of SSFs refers to the total number of SSFs required to sense all channels in the system. An exemplary frame sequence for an even hop node in a network with six nodes (N0 to N5) and 16 channels (C0 to C15) is shown in Fig. 3.

Since the communication is half-duplex, different time-slots are required for transmission and reception. The transmission and reception channels shown in this figure are for an even hop node. For the next hop nodes (odd hop), the position of the transmission and reception channel will be reversed which will be explained further in sub-section III-C.

The frequency for the transmission channel is not fixed, but is negotiated with the next hop node(s), and for the reception channel, it is negotiated with the previous hop node(s). In the example in Fig.2, the transmission frequency is F1, while the reception frequency is F2.

C. Data Flow

In the previous sub-sections, the tasks with a DSM perspective were identified and segregated by defining the activities and resources. In the subsequent sub-sections, the communication flow based on the activities and resources mentioned above will be presented. We first look at the control message exchange for the communication set-up, which is exemplarily shown in Fig. 4.

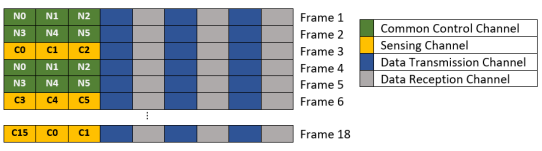

Fig. 3. Frame Sequence for an Even Hop Node

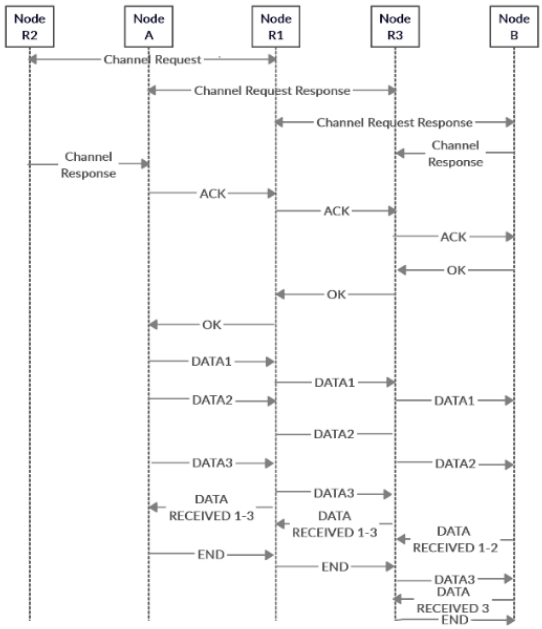

Fig. 4. Data Flow

The communication flow consists of exchange of control messages between the hops to negotiate the channel for transmission.

Transmission setup: The source node broadcasts a channel request message on CCC to the next hop to initiate communication. The subsequent hops send a combined channel request response in response to the request from the previous hop and to initiate a request to the next hop. These are broadcast messages with the channel status information i.e. if the channels are interfered (I), free (F) or occupied (O) as observed by the node. All the neighbouring nodes update their availability matrix based on this information as shown in Fig. 5 which is the availability matrix for Node A. In this figure channel 2 is a suitable choice for Node A as it is free for itself and its next-hop Node R1 (see Fig. 1)

ACK messages are exchanged between the hops to finalize the channels to be used and OK message is used to confirm the end-to-end channel negotiation between hops. For data transmission in half duplex communication, separate transmission and reception slots are defined. However, it may be noticed

	Channel 1	Channel 2	Channel 3	Channel 4
Node R1	F	F	I	I
Node R2	O	I	I	F
Node A	F	F	I	I

Fig. 5. Availability Matrix

that one node is transmitting (e.g. an even hop) and the next hop (i.e. an odd numbered hop in the transmission path) is receiving and vice versa.

Transmission start: Therefore, in order to allow for time critical services, the even hops can transmit simultaneously and the odd hops can transmit simultaneously, as long as they do not interfere with each other's communication (i.e., use different frequencies for transmission). E.g., for a transmission between Node A and B (see Fig. 6), the even hops A-R1 and R3-B can transmit simultaneously at the time instant 't'. The odd hop R1-R3 will transmit at the next time instant. In a complex environment, the even and odd hops can be determined based on the Time-to-Live (TTL) value which is decremented at each hop. The transmitted packets are acknowledged collectively to reduce control overhead (e.g., as shown in Fig. 4, packets 'Data1' to 'Data3' are acknowledged collectively by 'Data Received 1-3' by 'Node R1') .

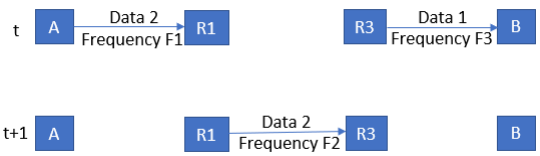

Fig. 6. Half-duplex transmission scheme

Transmission end: When all data has been sent and its reception has been acknowledged, a specific message indicating the end of the transmission is sent.

Problem handling: During an ongoing communication, when the AckTimeout timer expires (i.e. acknowledgement for the packets sent is not received), Channel Switch message is broadcast to change the channel. Channel Switch message can also be initiated if the channels conditions degrade.

IV. DYNAMIC SPECTRUM MANAGEMENT ALGORITHMS

The design of the dynamic spectrum management algorithms is influenced by the network architecture and the main goal to be achieved. Firstly, it can be observed that the network architecture is distributed and that it supports multi-hop communications. Therefore, each node should have individual decision-making capability. Secondly, the main goal is to optimize and improve the secondary network performance based on the cognitive decision making capability of the secondary node. In a multi-hop secondary network with an extended interference area for a single transmission link, the chance of transmission failure is high, as there is no fixed available channel for transmission. Therefore, the packet success rate for the transmissions characterizes the performance of the network.

The packet success rate can be improved by reducing the interference and subsequently improving the SNIR, decreasing the Packet Error Rate (PER), and improving the throughput. This is achieved by optimizing frequency, transmission power, and modulation scheme. However, some of these objectives conflict with each other. E.g., higher throughput would require a higher modulation scheme. However, a higher modulation scheme may lead to an increased Bit Error Rate (BER) and consequently an increased PER. Similarly, increased transmission power may lead to increased interference with other nodes of the secondary network, thus countering any improvement in the system.

Consequently, dynamic spectrum access is combined with the problem of network optimization. In our problem we use the performance metrics such as PER, SNIR and throughput and formulate this as a multi-objective optimization problem:

$$\max_{x \in X} f(x) \tag{1}$$

where,

$$f(x) = \left(1 - \text{PER(x)}_{avg}\right) * \text{SNIR(x)}_{avg} * \text{Throughput(x)}_{avg} \tag{2}$$

X is the feasible set of decision vectors including transmission power, frequency, and modulation scheme.

In our system, there are 16 frequency channels and four different modulation schemes. In addition to that, we define a set of eight different transmission powers, which can be set by the DSM algorithm. Fig. 7 shows a binary encoding of these transmission parameters, which will be used for the GA.

The system performance is looked at every time instance *m* (where, $m \in \mathbb{N}$). For each time instance *m* a node is in a state x_m. Each performance metric i.e. PER, SNIR, and throughput, is measured by the receiver during an ongoing transmission and sent as a feedback to the transmitter in the cumulative acknowledgement for the received packets i.e. in the DATA RECEIVED message (see Fig. 4). These metric values are assessed (Metric_m) at the time instances and a moving average value of each performance metric for the state x_m is calculated and indicated by $\text{Metric(x}_m)_{avg}$. The transmitter uses this average value instead of the instantaneous values in order to make the function $f(x)$ (as defined in Equation(2)) immune to small disturbances. The average value is updated as follows: For $x_m = x$,

$$\text{Metric(x)}_{avg} \leftarrow 0.8 * \text{Metric}_m\text{(x)} + 0.2 * \text{Metric(x)}_{avg} \ , \tag{3}$$

if it is at least the second occurrence of x
or,

$$\text{Metric(x)}_{avg} = \text{Metric}_m \ . \tag{4}$$

Two algorithms, namely the Genetic Algorithm and Q-Learning, which have been adapted to the constraints of a secondary network, are used to solve this multi-objective

optimization problem. Figuratively speaking, these algorithms learn the long-term channel behavior. Thus, in combination with the immediate or short-term channel availability information (channel sensing information or via control information from neighbours), an efficient DSM algorithm is obtained. Each node in the transmission path can take decision based on these algorithms. The transmitting nodes can decide the transmission parameters at the beginning of a transmission and trigger a state change whenever a better state is available. However, there is a cost associated with the change in state in terms of signalling overhead. Hence, in order to minimize the overhead, a change in state is triggered only when a channel is interfered.

A. Genetic Algorithm (GA)

GA is a metaheuristic used in optimization and search problems that follows Darwin's principle of natural selection and evolution. GA is well suited to multi-objective optimization problems due to the capability of GA to evaluate the objective function in different dimensions in parallel and to handle constraints [3]. In context to the multi-objective optimization problem for DSM, a population of chromosomes is generated, where each chromosome encodes a candidate solution (see Fig. 7) [15]. For example, the binary code for a chromosome in Fig. 7 encodes a frequency channel number 11, transmission power step 6 and modulation code number 2 i.e. 16-QAM, as BPSK, QPSK, 16-QAM and 64-QAM are encoded by 0, 1, 2 and 3. Each bit in the binary code represents the gene of the chromosome. Any change in the binary code will lead to a new combination of frequency channel number, transmission power and modulation scheme. E.g. Fig. 7 can be understood as follows:

$x = (101111010), X \in \{0,1\}^9$, where the bits encode the decision variables.

Fig. 7. Binary encoding of transmission parameters

For an end-to-end transmission, the following steps take place at each hop between a transmitter-receiver pair:

- Transmitter sends data to the receiver.
- Receiver on receiving the data calculates the performance metric (i.e. PER, SNIR and throughput).
- Receiver sends these metric values with the acknowledgement of the received data as a feedback.
- Transmitter updates these metric values for the tuple of transmission parameters (frequency, transmission power, and modulation scheme) for the ongoing transmission.

For our system, we determine the fitness based on observed network values of the performance metric, instead of theoretical values. The fitness of each chromosome is evaluated using an objective function that characterizes the solution. Equation (2) gives the function that has been used in this work. The population undergoes selection, crossover and mutation. The parent chromosomes are selected based on fitness proportional selection. The selected parents perform crossover to produce offspring. The new generation is created by removal of less fit individuals and introduction of offspring in the population. The chromosomes are driven towards an optimal solution by the process of repeated operations of selection, crossover, and mutation or any other operation defined to make the process more robust. One iteration of the algorithm is performed every time any update in the performance metric for the transmission parameter is observed.

As discussed previously, an important aspect in DSM for the secondary network is the availability of channels. A basic GA works on the principle of convergence with the population moving towards an optimal solution. However, the algorithm may converge towards an unavailable channel specially if the population is too small, i.e. the last performance of the chromosome might be good, but the channel could have recently been occupied, for which the performance metrics can only be obtained when transmission is already initiated on this channel. E.g., the GA population converged to a solution of frequency channel value 3, power 1.3 mW and modulation 16-QAM. However, if spectrum sensing declares channel 3 to be interfered at present for the node, the solution cannot be used. Therefore, additional operations are defined to prevent this.

In this work, we propose a continuously adaptive GA. In order to make population diverse and valid, two additional operations have been introduced. First, uniqueness of individuals is maintained in the population after every generation to avoid the domination of the population by a certain solution and subsequent invalidation due to unavailability of channels. Second, removal of unfit individuals or individuals consisting of unavailable channels is carried out in every generation. New random individuals that are fit (channels are free) are introduced in the population to keep the population size constant. If enough fit individuals are not available, random individuals are added. However, for this use case, the the secondary network is not limited by number of available channels as our primary focus is to observe the algorithm behaviour in learning and selection of optimal transmission parameters and seamless continuity of services for an end-to-end transmission on observing any interference from the primary network.

The algorithm progresses by using sub-optimal results and gradually learning and moving towards optimal results, as more long-term information about the channels is observed. When the acknowledgement for the transmitted packets are not received in time, a channel switch message is broadcast by the transmitter with the decision variable solution in the form of the fittest chromosome.

B. Q-Learning

Q-Learning is a model free reinforcement-learning algorithm that is based on the goal of finding an optimal policy that determines actions to different states. A policy is a function

that takes states as inputs and suggests actions. An optimal policy in Q-Learning maximizes the expected cumulative discounted reward of being in a particular state and taking an action (quality of an action), including the reward for the current state and all the successive states. The algorithm learns the instant rewards by interacting with the environment and then estimating the Q-Value for each state-action pair by value iteration. The estimate gets better over time as more states are explored and the information of the Q-Value of these state-action pairs is propagated to the other pairs. The action can be taken based on informed decision according to the optimal policy (also known as exploit) or can be randomly taken to explore new states and learn. This is known as the ϵ-Greedy policy, where the decision to explore or exploit can be taken based on the probability ϵ [4].

Unlike GA, where transmitter only made the decision of choosing the transmission parameters, each transmitter in this case can make two decisions. Firstly, when to change the channel and secondly, which transmission parameters to choose. Each state in Q-Learning is a tuple of frequency, modulation and power. Just as mentioned previously in GA, The transmitter interacts with the environment, receives feedback from the receiver in the form of performance metric. The action to take is based on the Q-Value. In this work, two variations of the modified Q-Learning are introduced. We first explore the quality of the action of going from one state to another, referred to as the Double State QL Switch. Second, we explore the possibility where the quality of the action is defined by the quality of transmission in a particular state, referred to as the Single State QL Switch adapted from the stateless Q-Learning mentioned in [20].

With respect to our optimization problem, the state, action, reward and Q-Value are as follows:

State $x \in X$

Action Change x by policy $\pi(x) \in X$

Reward

$$R(x) = f(x) * \eta \tag{5}$$

where

$$\eta = \frac{\text{Packets acknowledged in transmission interval}}{\text{Total packets sent in the transmission interval}}$$

and $f(x)$ from (2).

Q Value

- Single State

$$\begin{aligned} \hat{Q}(x_n) \leftarrow\ & (1-\alpha)\,\hat{Q}(x_n) + \\ & \alpha\left(R(x_n) + \gamma \max_x \hat{Q}(x)\right) \end{aligned} \tag{6}$$

- Double State

$$\begin{aligned} \hat{Q}(x_{n-1}, x_n) \leftarrow\ & (1-\alpha)\,\hat{Q}(x_{n-1}, x_n) + \\ & \alpha\left(R(x_n) + \gamma \max_x \hat{Q}(x_n, x)\right) \end{aligned} \tag{7}$$

where $\alpha \in (0,1]$ is the learning rate,
$\gamma \in [0,1]$ is the discount factor, and
n is the time instance of transmission in one state.

Optimal Policy at each time instance is given by

- Single State QL Switch

$$\pi^*(x) = \text{argmax}_x Q(x) \tag{8}$$

- Double State QL Switch

$$\pi^*(x) = \text{argmax}_x Q(x_{n-1}, x) \tag{9}$$

Following would be the steps between a transmitter receiver pair at each hop:

- The transmitter receives the performance metric values in feedback from the receiver.
- The transmitter calculates and stores, the instantaneous reward based on (5).
- Q-Value is calculated based on (6) or (7).
- When required, depending on the optimal policy given in (8) or (9), the transmission parameters are chosen by the transmitter as the next state.

Although the Q-Value takes into account the long-term reward for a secondary network, a quick recovery in case of appearance of the primary user is necessary. By weighing the function with η, the reward function starts degrading when no acknowledgement is received for the transmitted packets and a switch can be made earlier in time. Hence, the transmitter decides on when to vacate the channel as well as the transmission parameters in this case.

The states that include unavailable channels (which are either occupied by neighbors or have been sensed to be occupied by the primary user recently) or the states that have not been visited, the knowledge base in the node does not have any new statistics, as no new transmission has yet been made on these channels. Hence, the rewards do not reflect the correct picture. The only information about these channels or states is reflected in the availability matrix, which is updated based either on sensing or on information from the neighbors. It is evident that these states will not be chosen as the next states. Therefore, the Q-value update, and thus, its optimal policy selection (during the exploit phase of the ϵ-Greedy policy) is restricted to valid states at that time instance. So, as the gradient of the Q-value function falls below a certain threshold, channel switch is initiated.

V. SIMULATION RESULTS

The simulation model is implemented in OMNeT++ using the INET framework. The model includes a primary network consisting of 16 pairs of transmitter and receiver. Transmission is based on a two state Markov model. Each pair further communicates on a fixed frequency and, therefore, simulates the channel behavior as seen by the secondary nodes. Hence, there are 16 i. i. d. non-overlapping channels of bandwidth 25 kHz from 50.05 MHz to 50.8 MHz. The number of primary transmitter-receiver pairs can be increased for simulating a higher number of channels.

As depicted in Fig. 1, the secondary network consists of four nodes, which try to exchange information over multiple

hops. The secondary user source node also transmits based on a two-state Markov model. For analysis purposes, we have simulated only one ongoing transmission in the secondary network at a time.

The primary user activity has been modeled to simulate a slowly varying channel condition compared to the secondary user activity. In addition, enough channels are available (not interfered by primary) to the secondary network, considering the channel requirement of three transmission links in this model. This is because the focus of the work is to select the best possible channel opportunity for performance maximization and to provide seamless spectrum mobility in case of interference on the channel.

A. Parametrization of Genetic Algorithm (GA)

Regarding GA, an important parameter is the population size. We assessed the transmission success under varying population sizes for multi-hop transmissions. The simulation has been run for a simulation time of 6000s and averaged over 40 simulations. The average of the packet success rate gives the number of successfully received packets compared to the number of sent packets. The efficiency also regards the amount of control messages and thus shows the ratio of successful user traffic to all user traffic including control traffic.

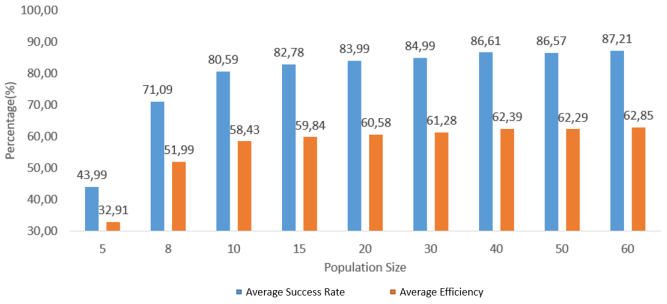

Fig. 8. Multi-hop transmission success under varying population sizes

The results are shown in Fig. 8. A small population size leads to poor performance. The performance does not vary significantly after a population size of 30. Consequently, we used this value for the performance simulations in the following section. Generally, the optimal population sizes may vary with the size of the chromosomes and the required number of hops in relation to the number of available (non-interfered) channels.

B. Performance comparison

For comparing the performance of the algorithms, we created a scenario with multi-hop communication and varying interference on all channels dynamically by inducing mobility in primary users. This scenario was then run several times for five different cases - no DSM, brute force search, GA, and Q-Learning with Single and Double State approach. In the no DSM case, we randomly selected a channel at the beginning of each transmission and remained there until the end of the transmission. Brute force search is a greedy algorithm, where the next best state (highest $f(x)$) out of the available states (states for which the channels are neither interfered nor occupied) based on the measurements stored in the database, is used for the next transmission. Brute force search is disadvantageous in higher search space scenarios due to its high computational complexity. GA limits the search space to its population size and would be advantageous even in high search space scenarios that include multiple transmission parameters. Also, brute force does not take into consideration the long term channel behaviour which is taken into account in Q-Learning. The results of the performance comparison are shown in Fig. 9.

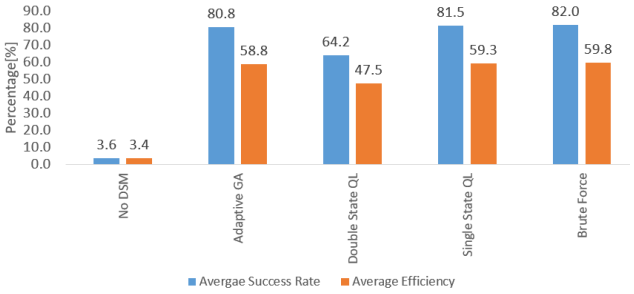

Fig. 9. Performance comparison

The results for the no DSM case show almost a complete loss of communication. The DSM solutions manage to identify spectrum holes and to use them for their transmissions. We may conclude that the GA and Single State Q-Learning perform better than the Double State Q-Learning algorithm. Both the GA and the Single State Q-Learning are able to achieve performance similar to brute force search. This can also be seen when looking at the throughput of the nodes, as e.g. depicted for relay node R3 in Fig. 10. The figure shows the average throughput at node R3 for the different DSM algorithms, taken from one iteration of the performance simulation. It can be observed that curves for brute force, GA, and Single State Q-Learning are similar, while the curve for Double State Q-Learning shows a significantly lower average throughput.

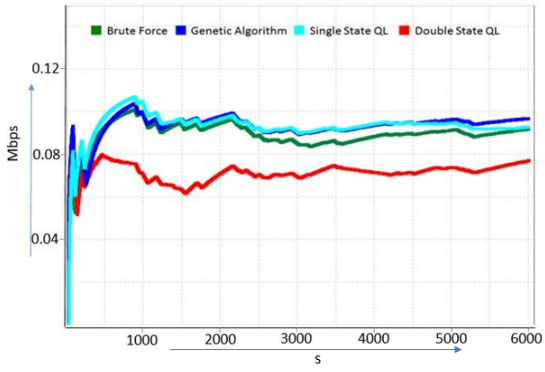

Fig. 10. Throughput at relay node R3

C. Discussion

A possible reason for the worse performance of the Double State Q-Learning may be that the Q-Values, which are the basis for the decision about the next state, are dependent on both the current state and the next state. Hence, the state space for decision-making is squared. Consequently, it will take much more time until all possible options in Q-Learning have been explored and can be used for decision-making. In addition to that, the ϵ-Greedy policy parameter is yet to be analyzed.

An important aspect to notice is that our network does not take into account energy efficiency. Hence, the fittest chromosomes tend towards using higher transmission power. However, as the system is extended to allow for multiple simultaneous transmissions, an increased transmission power also leads to increased system interference and thus, the fittest chromosomes tend towards using smaller transmission power.

We also considered perfect channel conditions where the system is synchronised, there is no multi-path delay with the free space pathloss model. It is believed that system can be extended to other pathloss models for urban environment. However, CCC needs to be designed to avoid loss of communication when it is interfered and system performance is to be measured including these constraints.

VI. SUMMARY AND OUTLOOK

In this paper, we have introduced the system design for a DSM secondary network in VHF band in a half-duplex scenario. Two algorithms – GA and Q-Learning – were explored to learn the radio environment and to make informed decision on radio parameters. For GA, it has been observed that the population size and the diversity of population are very important to prevent the population from getting invalidated due to channel unavailability. On the other hand, a large population will increase the computational complexity while searching for a valid candidate. On average, the GA and the Single State Q-Learning outperform the Double State Q-Learning. They even achieve a performance quite similar to brute force. Yet, GA is more efficient, as it does not need to take into account the objective function of all transmission parameter sets, but only of a selected subset. Q-Learning is beneficial compared to brute force, as it provides a long-term learning capability and does not only consider the most recent objective value. However, both GA and Q-Learning show significant improvement over a multi-hop secondary network without any DSM capability in terms of both success rate and efficiency.

In the future, the performance of the algorithms will have to be measured under multiple simultaneous transmissions in the network. Furthermore, the parameters to introduce dynamic switching for GA need to be explored and tuned further for Q-Learning to improve the performance.

REFERENCES

[1] S. Couturier, T. Bräysy, B. Buchin, J. Krygier, V. Le Nir, N. Smit, T. Tuukkanen, and E. Verheul, "End-to-end optimization for tactical cognitive radio networks," in *2018 International Conference on Military Communications and Information Systems (ICMCIS)*, pp. 1–8, 2018.

[2] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, pp. 40–48, Apr. 2008.

[3] D. Coley, *An Introduction to Genetic Algorithms for Scientist and Engineers*. June 2014.

[4] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.

[5] S. Haavik and B. Libaek, "Link layer design for a military narrowband radio network," 2010.

[6] M. Sami, N. K. Noordin, M. Khabazian, F. Hashim, and S. Subramaniam, "A survey and taxonomy on medium access control strategies for cooperative communication in wireless networks: Research issues and challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2493–2521, 2016.

[7] D. Raychaudhuri and Xiangpeng Jing, "A spectrum etiquette protocol for efficient coordination of radio devices in unlicensed bands," in *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, PIMRC*, vol. 1, pp. 172–176, 2003.

[8] Liangping Ma, Xiaofeng Han, and Chien-Chung Shen, "Dynamic open spectrum sharing mac protocol for wireless ad hoc networks," in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN*, pp. 203–213, 2005.

[9] J. So and N. H. Vaidya, "Multi-channel mac for ad hoc networks: Handling multi-channel hidden terminals using a single transceiver," in *Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '04, (New York, NY, USA), pp. 222–233, Association for Computing Machinery, 2004.

[10] A. P. Shrestha, J. Won, S. Yoo, M. Seo, and H. Cho, "Genetic algorithm based sensing and channel allocation in cognitive ad-hoc networks," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 109–111, 2016.

[11] S. Chantaraskul and K. Moessner, "Implementation of a genetic algorithm-based decision making framework for opportunistic radio," *IET Communications*, vol. 4, no. 5, pp. 495–506, 2010.

[12] J. Elhachmi and Z. Guennoun, "Cognitive radio spectrum allocation using genetic algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, p. 133, May 2016.

[13] N. M. Hidayati Robbi, I. W. Mustika, and Widyawan, "A modified genetic algorithm for resource allocation in cognitive radio networks," in *2018 4th International Conference on Science and Technology (ICST)*, pp. 1–5, 2018.

[14] Z. Zhao, Z. Peng, S. Zheng, and J. Shang, "Cognitive radio spectrum allocation using evolutionary algorithms," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4421–4425, 2009.

[15] C. J. R. Thomas W. Rondeau, Bin Le and C. W. Bostian, "Cognitive radios with genetic algorithms: Intelligent control of software defined radios," Nov 2004.

[16] R. Han, Y. Gao, C. Wu, and D. Lu, "An effective multi-objective optimization algorithm for spectrum allocations in the cognitive-radio-based internet of things," *IEEE Access*, vol. 6, pp. 12858–12867, 2018.

[17] Y. Wang, Z. Ye, P. Wan, and J. Zhao, "A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks," *Artificial Intelligence Review*, vol. 51, pp. 493–506, Mar. 2019.

[18] A. Das, S. C. Ghosh, N. Das, and A. D. Barman, "Q-learning based co-operative spectrum mobility in cognitive radio networks," in *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, pp. 502–505, 2017.

[19] L. R. Faganello, R. Kunst, C. B. Both, L. Z. Granville, and J. Rochol, "Improving reinforcement learning algorithms for dynamic spectrum allocation in cognitive sensor networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 35–40, 2013.

[20] N. Morozs, T. Clarke, D. Grace, and Q. Zhao, "Distributed q-learning based dynamic spectrum management in cognitive cellular systems: Choosing the right learning rate," in *2014 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6, 2014.

[21] X.-L. Huang, X.-W. Tang, and F. hu, "Dynamic spectrum access for multimedia transmission over multi-user, multi-channel cognitive radio networks," *IEEE Transactions on Multimedia*, p. 1, July 2019.