A Machine Learning Approach for Resource Allocation in Wireless Industrial Environments

Idayat O. Sanusi and Karim M. Nasr
Faculty of Engineering and Science, University of Greenwich,
Kent, ME4 4TB, United Kingdom
{i.o.sanusi, k.m.nasr}@gre.ac.uk

***Abstract*— In this paper, we present a machine learning technique for channel selection in a Device to Device (D2D)-enabled cellular network targeting a wireless industrial environment. The presented Base Station Assisted (BSA) reinforcement learning technique uses a distributed local Q-table for the agents (users), to prevent global information gathering within the cellular network. A stateless Q-learning approach is adopted to reduce the complexity of learning and the dimension of the Q-table. After the training of the D2D agents, the Q-tables of the D2D users are uploaded to the base station for resource allocation to be implemented centrally. Simulations results show that the presented technique provides a Radio Resource Management (RRM) solution with a good Quality of Service (QoS) performance compared to other conventional approaches.**

Keywords*— *5G and beyond networks; Radio Resource Management; Distributed Algorithms; Device-to-Device Communication; Reinforcement Learning.

I. INTRODUCTION

Device-to-Device communication (D2D) is considered as a promising solution for ultra-reliable low-latency use cases because of associated advantages in terms of reduced latency, improved reliability and throughput. The integration of D2D into future industrial wireless networks and smart manufacturing facilitates the creation of massive machine-type connections [1]. Machine-Type Communication (MTC) is expected to support large numbers of smart devices, predominantly with small data volume requirements, which aggregates into a massive amount of data from parallel transmissions of a large number of devices. Achieving ultra-high reliability and ultra-low latency pose challenges in terms of bandwidth requirements. Yilmaz et al. [2] studied the transmission bandwidth needed to enable Ultra-Reliable Low-Latency Communication (URLLC) for factory automation and found that the system bandwidth depends on the number of connected pieces of user equipment and the behaviour of their traffic. The scarcity of radio resources and the limitations on the available system bandwidth makes spectrum sharing a necessity for D2D implementation of MTC for factory automation [3]. RRM schemes need to be efficiently designed for interference management and coordination while guaranteeing tight URLLC demands.

Channel reuse among active devices in the same cells will generate interference which degrades system performance. Interference management is crucial to ensure efficient utilisation of available spectrum resources which is also particularly challenging for D2D deployments in underlay cellular networks.

The two major approaches to Radio Resource Management (RRM) are based on centralised and distributed methods. The centralised scheme requires global information gathering by base stations which often results in a high signalling overhead and increased complexity which tend to increase with the number of users, thus making it impractical. The distributed approach does not need a central entity; resource allocation is implemented by users therefore reducing the amount of information gathering and processing by base stations. However, a distributed RRM algorithm may also increase signalling overheads due to the high amount of information interchange among devices [4].

Reinforcement Learning (RL) has recently gained a lot of attention because of its suitability for the decision-making process where there is unknown or partial channel information. RL has been widely studied for intelligent power and spectrum allocations for D2D communication in cellular networks. Asheralieva and Miyanga [5] formulated the resource allocation problem as a non-cooperative game model among D2D users and a mixed strategy Nash Equilibrium was obtained. However, the Quality of Experience (QoS) metrics of cellular users sharing channel resources with D2D links were not taken into account in the reward model. A multi-agent actor-critic structure is proposed in [6] which involves interactions between users and centralised sharing of all historical information. This leads to an increase in signalling overheads and information exchange. In [5]-[8], the reward function captured the QoS metric of the cellular users in a centralised Q-learning approach, which leads to increased signalling overheads as well.

In this paper, we present a semi-distributed reinforcement learning scheme for D2D resource allocation in a cellular network in an industrial setting. After the decentralised training of the agents, their Q-tables uploaded are forwarded to the base station for a centralised resource allocation. The reward function is modeled in such a way that there is no information exchange related to other agents' action or reward. To address the problem of the 'curse of dimensionality' associated with Q-learning, a stateless Q-learning is adopted to reduce the dimension of the Q-table, nonetheless capturing the QoS demands of the D2D users.

The paper is organised as follows: The problem formulation and system model are presented in Section II. In Section III, a stateless reinforcement learning algorithm for base station-assisted resource allocation is presented. Section IV presents some results and discussions of the simulation scenarios considered. Conclusions and directions for future work are summarised in Section V.

 ISBN: 978-1-61208-956-0

II. SYSTEM MODEL

We consider D2D and cellular users coexisting within a cellular network for uplink spectrum-sharing. There are N Cellular User Equipment (CUEs) represented by a set C and M D2D User Equipment (DUEs) denoted by a set D randomly deployed within the coverage of the base station (BS) in a single cell system. The cellular users have strict performance requirements in the form of minimum Signal to Interference plus Noise Ratio (SINR) values to guarantee their throughput. The D2D links also have minimum SINR thresholds to guarantee their throughput demands, in addition, to the reliability and delay constraints. We assume that each CUE has been pre-allocated a resource block. The transmit power of the CUEs and DUEs are denoted by P_{c_i} and P_{d_j}, respectively. We denote $g_{c,B}$, $g_{d_T,B}$, g_{d_T,d_R} and g_{c,d_R} as the channel gains from the CUE c_i to the BS, the interference link from the DUE transmitter d_T to the BS, the D2D link from the DUE transmitter d_T to the receiver d_R and the interference link form the CUE transmitter to the DUE receiver d_R, respectively.

The instantaneous received signal-to-interference-plus-noise-ratio (SINR) at the BS from ith CUE and jth DUE over ith sub-channel at time slot t is given as:

$$\Gamma_{c_i}(t) = \frac{P_{c_i} g_{c,B}(t)}{\sigma^2 + \sum_{d_j \in D} \lambda_j^i(t) P_{d_j} g_{d_T,B}(t)} \quad (1)$$

$$\Gamma_{d_j}(t) = \frac{P_{d_j} g_{d_T,d_R}(t)}{\sigma^2 + \sum_{c_i \in C} \lambda_j^i(t) P_{c_i} g_{c,d_R}(t)} \quad (2)$$

$\lambda_j^i \in \{0,1\}$ denotes the binary resource reuse indicator, $\lambda_j^i = 1$ implying jth DUE selects ith CUE sub-channel at time slot t and $\lambda_j^i(t) = 0$ otherwise.

The data rates of the ith CUE and jth DUE is at time slot t are given by:

$$\mathrm{T}_{c_i}(t) = W_i \log_2\left(1 + \Gamma_{c_i}(t)\right) \quad (3)$$

$$\mathrm{T}_{d_j}(t) = W_i \log_2\left(1 + \Gamma_{d_j}(t)\right) \quad (4)$$

where W_i is the bandwidth of each resource block. The variance of the additive white Gaussian noise (AWGN) is denoted by σ^2. The resource allocation problem for D2D communication in cellular network is NP hard and cannot be solved directly and often requires global information gathering which increases complexity. The channel gains for links q to r can be expressed as follows:

$$g_{q,r} = G_r \gamma_{q,r} \chi_{q,r} L_{q,r}^{-\alpha_r} \quad (5)$$

where G_r is the pathloss constant, $\gamma_{q,r}$ is the small-scale fading gain due to multipath propagation and assumed to have an exponential distribution with unit mean. The large-scale fading comprises pathloss with exponent α_r and shadowing which has a slow fading gain $\chi_{q,r}$ with a log-normal distribution. $L_{q,r}$ is the distance from terminal q to terminal r [9]. The channel gain from D2D link d_j of transmitter d_T to the receiver d_R is g_{d_T,d_R}, the channel gain of the interference link from d_T to the base station is $g_{d_T,B}$ and from CUE c_i to DUE d_j receiver is h_{c,d_R} and $h_{c,B}$, is the channel gain from CUE c_i to the base station BS. The channel gain g_{d_T,d_R} and g_{c,d_R} can be estimated at the DUE receiver, d_R and made available at its transmitter, d_T instantaneously [10]. Similarly, $g_{c,B}$ and $g_{d_T,B}$ can be obtained at BS through local information since uplink transmission is considered. The reliability of the DUE $d_j \in D$, $\xi_{d_j}(t)$, is defined as the probability of packet delay exceeding a predefined delay bound, $l_{d_j,\max}$, for channel i at slot t is less than a threshold [11]. Only the transmission delay is considered in this work. The objective of the system is to maximise the total throughput, T_R, of paired CUE and DUEs while satisfying the QoS demands.

$$\max_{\lambda_j^i} T_R = W_i\left(\lambda_j^i\left(\sum_{c_i \in C} \log_2\left(1 + \Gamma_{c_i}\right) + \sum_{d_j \in D} \log_2(1 + \Gamma_{d_j})\right)\right) \quad (6)$$

subject to:

$$\lambda_j^i \Gamma_{c_i} - \Gamma_{c_i,\min} \geq 0 \qquad \forall c_i \in C \quad (6a)$$

$$\Pr\left(l_{d_j} > l_{d_j,\max}\right) < 1 - \xi_{d_j}^* \quad \forall d_j \in D \quad (6b)$$

$$\sum_{c_i \in C} \lambda_j^i \leq 1 \qquad \forall d_j \in D \quad (6c)$$

$$\sum_{d_j \in D} \lambda_j^i \leq 1 \qquad \forall c_i \in C \quad (6d)$$

The minimum CUE SINR, $\Gamma_{c_i,\min}$, to guarantee the throughput requirement of the CUEs is defined in constraint (6a). Constraint (6b) takes into account reliability and delay, where l_{d_j} is the packet delay constraint for packet transmission of DUE d_j. The expression captures the fact that the end-to-end delay should be less than $l_{d_j,\max}$ with a probability of at least $1 - \xi_{d_j}^*$. Constraints (6c) and (6d) are channel association criteria. The reliability of the DUE links in (6c) is evaluated using an empirical estimation of number of packets transmitted similar to [11], from d_T to d_R whose delay is within the budget $l_{d_j,\max}$ over the total number of packets sent to d_R at time slot t i.e.,

$$\xi_{d_j}(t) = 1 - \Pr\left(l_{d_j} > l_{d_j,\max}\right) \approx 1 - \frac{L_{d_j}(t)}{B_{d_j}(t)} \cong \frac{L'_{d_j}(t)}{B_{d_j}(t)} \quad (7)$$

where $L_{d_j}(t)$ is the number of packets for which $l_{d_j} > l_{d_j,\max}$ and $L'_{d_j}(t)$ is the number of packets transmitted with $l_{d_j} \leq l_{d_j,\max}$ (or number of packet delivered within the delay bound). $B_{d_j}(t)$ is total packet transmitted by DUE d_j at time slot t. Reliability can also be measured in terms of the outage probability, which is the probability that the measured SINR is lower than a minimum is less than a predefined threshold. The closed expression of the outage probability of jth DUE conditioned on the selected ith channel at time slot t is given below [12].

$$p_R(t) = \Pr\left(\Gamma_{d_j} \leq \Gamma_{d_j,\min}\right) = 1 - \frac{\mathrm{P}_{d_j} g_{d_T,d_R} \exp\left(-\frac{\Gamma_{d_j,\min}\sigma^2}{\mathrm{P}_{d_j} g_{d_T,d_R}}\right)}{\mathrm{P}_{d_j} g_{d_T,d_R} + \Gamma_{d_j,\min} \mathrm{P}_{c_i} g_{c,,d_R}} \leq p_{R_0} \quad (8)$$

where $p_R(t)$ is the measured outage probability of DUE d_j at time slot t and p_{R_0} is the maximum tolerable outage

probability of d_j. The reliability of the DUE in terms of outage probability is expressed as:

$$\xi_{d_j}(t) = 1 - p_R(t) \tag{9}$$

Transmission delay is given as the ratio of packet size transmitted within delay bound to transmission rate [13]. From (7), (8) and (9) the transmission of jth DUE on the ith RB is formulated as:

$$l_{d_j}(t) = \frac{L'_{d_j}(t)}{W_i \log_2(1 + \Gamma_{d_j})} \tag{10}$$

The resource allocation problem for D2D communication in a cellular network is complex and a direct solution is not feasible. We present next a base station-assisted resource allocation scheme which adopts a semi-distributive RRM approach.

III. Stateless Reinforcement Learning for Base Station-Assisted Resource Allocation

The goal of the agents is to maximise the throughput in a D2D-enabled cellular network. At each time slot t, a DUE, observes a state s^t and takes an action a^t from the action space, (i.e., select an RB k_i), according to the policy π. Q-learning enables an agent to determine the optimal strategy that maximises its long term expected cumulative reward [14]. The Q-value is updated as follows:

$$Q^{t+1} = \begin{cases} Q^t(s^t, a^t) + \sigma\left[r^t + \eta \max_a Q^t(s^{t+1}, a^{t+1}) - Q^t(s^t, a^t)\right] \\ \qquad \text{if } s = s^t,\ a = a^t \\ Q^t(s^t, a^t)\ , \text{ otherwise} \end{cases} \tag{11}$$

where $\sigma \in [0,1]$ is the learning rate. With $\sigma = 0$, the Q-values are never updated, hence no learning has taken place; setting σ to a high value such as means that learning can occur quickly and $0 \leq \eta \leq 1$ is the discount factor used to balance immediate and future reward [14].

The state-action dimension is reduced by adopting a stateless learning approach. For the considered scenario, any action $a_i \in A$ taken by an agent will result in the end of an episode i.e., states 0 and 1 are terminal states, where $S^i_{d_j}(t) = 1$ is the goal state of the DUEs. Therefore, the learning environment can be modelled entirely using a stateless Q-learning i.e., action-reward only since the state transition is not required. An agent can choose its action based solely on its Q-value and the updated Q-value of the chosen action is based on the current Q-value and the immediate reward from selecting that action. The update function in (11) is re-formulated as follows:

$$Q^{t+1}(a^t) = \begin{cases} Q^t(a^t) + \sigma[r(a^t) - Q^t(a^t)], \text{if } a = a^t \\ Q^t(a^t), \qquad \text{otherwise} \end{cases} \tag{12}$$

where $r(a^t)$ is the immediate reward of selecting a. In contrast to the standard Q-value update function in (11), it can be seen in (12) that not only the state-action formation (s, a) is not necessary, but also the information of the next state s^{t+1} is not required because the actions lead to a terminal state. Therefore, the Q-table is defined in terms of the actions only and updated using the immediate reward. This results in $1 \times |N|$ dimension Q-table for jth DUE.This method reduces the learning complexity and the Q-table dimension.

The traditional cellular users in the network need to be protected from the interference caused by the DUEs for their minimum SINR to be satisfied. This may be achieved by integrating the SINR of the CUE, Γ_{c_i} in the state space or reward function modelling. This way, the DUEs can obtain the information from BS at time slot t as in [15]-[17]; hence, the DUEs get a reward if the CUE SINR $\Gamma_{c_i} \geq \Gamma_{c_i,\min}$, on the and a penalty otherwise. Rather than the BS exchange the measured CUE SINR, Γ_{c_i}, with the DUEs for every action a^t taken at each time slot, we adopt a scheme in which the BS keeps a look-up table of the ith CUE based on the actions on the DUEs.

BSA Reinforcement Learning Algorithm

1: Initialise the action-value function for the DUEs
$\left[Q_{d_j}(a) = 0 | Q_{d_j}(a) \equiv Q^i_{d_j}(a^t)\ , i = 1,2, \dots, N\right] \forall\ d_j \in D$
2: Initialise the action-value function for the BS for the actions of the jth DUE on the ith RB
$\left[Q_{c_i}(a) = 0 | Q_{c_i}(a) \equiv Q^j_{c_i}(a^t),\ j = 1,2, \dots, M\right] \quad \forall\ c_i \in C$
3: for $d_j \in D$ $1 \leq j \leq M$ do
4: **while** not converge **do**
5: generate a random number $x \in \{0,1\}$
6: **if** $x < \varepsilon$ **then**
7: Select action a_i^t randomly
8: **else**
9: Select action a_i^t=$\underset{a \in A}{\text{argmax}}\, Q_{d_j}(a^t)$
10: **end**
11: Evaluate ξ_{d_j}, Γ_{d_j} and l_{d_j} of $d_j \in D$ for the action a^t
12: Measure the SINR, ξ_{c_i}, of CUE $c_i \in C$ for the action a^t taken by $d_j \in D$
13: Observe immediate reward of $d_j \in D$ and $c_i \in C$,
14: Update action-value for action of $d_j \in D$ on the ith RB $Q^i_{d_j}(a) = Q^i_{d_j}(a) + \sigma\left[r_{d_j}(a^t) + Q^i_{d_j}(a)\right]$
15: Update action-value for $c_i \in C$ for action a^t of jth DUE $Q^j_{c_i}(a) = Q^j_{c_i}(a) + \sigma\left[r_{c_i}(a^t) + Q^j_{c_i}(a)\right]$
16: **end while**
17: **end for**
18: Load $Q_{d_j}(a)$ to the BS $\forall\ d_j \in D$
19: **for** $d_j \in D$ $1 \leq j \leq M$ **do**
20: Obtain $Q(a) = \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\}$ $i = 1,2, \dots, N$
21: $\bar{Q}(a) \subseteq Q(a) | \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\} \in \mathbb{R}^+$, where $\mathbb{R}^+$ positive real number
22: $Q_{\text{TOT}} = Q^i_{d_j}(a) + Q^j_{c_i}(a) \qquad \forall\ q \in \bar{Q}(a)$
23: **end for**
24: Set up a list for unmatched DUE $D_u = \{d_j : \forall\ d_j \in D_u\}$
25: **while** $D_u \neq \emptyset$ **do**
26: Rank D_u in increasing order of $|0\ \bar{Q}(a)|$
27: Start DUE $d_j \in D_u$: $\bar{Q}(a) \neq \emptyset$ with the least $|\ \bar{Q}(a)|$
28: $c_i^* = \max_{r_i \in R} Q_{\text{TOT}}$
29: $D_u = D_u - d_j$
30: $\bar{Q}(a) = \bar{Q}(a) \backslash c_i^* \qquad \forall\ d_{j'} \in D_u | j' \neq j$
31: **end while**

There are a number of methods to select an action based on the current evaluation of the Q-value at every time slot t using a policy denoted by $p_{a_j}^t$. These methods are used to balance exploration and exploitation [18]. Epsilon greedy (ε-greedy) is one of the methods of choosing an optimal Q-value.

The reward function is modelled such that it relies only on local observations and can be implemented in a distributive manner. The rewards of the jth DUE and ith CUE for taking an action a_i^t is expressed in terms of the achievable throughput using the Shannon capacity formula. Therefore, the reward is directly related to the objective function of the optimisation problem. The following is a summary of the Base Station Assisted (BSA) Reinforcement Learning Algorithm:

The jth DUE only gets a reward when all state variables are 1 (i.e., the minimum QoS demands are met) while ith CUE gets a reward if its minimum SINR is satisfied at each time slot for the action taken by jth DUE. From the reward function defined above, learning can be implemented independently in a decentralised manner such that each agent maintains a local Q-table. There is no information exchange relating to other agents' actions or rewards and no cooperation is needed between the agents, which results in reduced signalling overheads and reduced complexity compared with a centralised Q-learning approach.

IV. Performance Evaluation

The performance of the presented BSA scheme is verified by considering a single-cell network in an industrial scenario. The simulation setup and channel models are summarised in Tables I and II. The network dynamics are captured by generating the channel fading effects randomly. The throughput is the main metric used to evaluate the performances of the algorithms. The performances of BSA are compared with centralised optimisation and the game theoretic Deferred Acceptance (DA) schemes [9].

TABLE I. MAIN SIMULATION PARAMETERS [9]

Parameter	Value
Carrier frequency, f_c	2GHz
System bandwidth	10MHz
Number of resource blocks (RB), K	50
RB bandwidth	180 kHz
Maximum CUE transmit power, $P_{c_i,\max}$	23dBm
Maximum DUE transmit power, $P_{d_j,\max}$	13dBm
D2D distance, L_{d_T,d_R}	$10\text{m} \leq L_{d_T,d_R} \leq 20\text{m}$
CUE SINR Threshold, $\Gamma_{c_i,\min}$	7 dB
DUE SINR Threshold, $\Gamma_{d_j,\min}$	3 dB
Noise power density	-174 dBm/Hz
Number of CUEs, N	50
Number of DUEs, M	50
Reliability for DUE, p_{R_0}	10^{-5}
Exploration rate, ε	0.7
Learning rate, σ	0.9
DUE Maximum Delay, $l_{d_j,\max}$	50ms
DUE Message Size, B_{d_j}	15kB

TABLE II. CHANNEL MODEL FOR LINKS [9]

Parameter	In-factory DUE link	UE-UE link	BS-UE link
Pathloss model	$\mathbf{36.8\log_{10}(d[m} + \mathbf{35.8}$	$\mathbf{40\log_{10}(d[m])} + \mathbf{28}$	$\mathbf{37.6\log_{10}(d[m} + \mathbf{15.3}$
Shadowing	4dB	6dB	8dB
Fast fading	Rayleigh Fading	Rayleigh Fading	Rayleigh Fading

The throughput performance of matched DUEs as a function of the number of DUEs in the system M, is shown in Fig. 1. It can be concluded that the sum throughput of the DUEs increases with the number of cellular users M for all the considered algorithms. As expected, the number of admitted DUEs increases with the introduction of new DUEs to the system but remains unchanged if a valid cellular resource-sharing partner cannot be found because the minimum QoS requirements are not satisfied.

The centralised optimisation and BSA approaches are comparable, while the DA method shows the least performance. The BS-A algorithm outperforms the DA algorithms by up to 9.69% increase in the DUE throughput performance. However, it is semi-distributive as the final resource allocation is implemented by the BS whereas the DA approach is decentralised (the channel selection is user-centric with no BS intervention to achieve autonomy). Players can make their resource allocations choices to maximise their individual and ultimately achieve system stability.

The performance of the sum throughput of the matched UEs (that is valid pairings between CUEs and DUEs) with respect to the number of cellular users M is presented in Fig. 2. The sum throughput increases with M. The BS-A approach shows better performance at $M \leq 35$ with up to 12.05% increase in sum throughput compared to the centralised approach while the centralised approach performed better at $M > 35$ with up to 9.39% increase in throughput. The DA algorithm again shows the least performance with up to 11.29% decrease compared with the BS-A technique. The effect of the outage probability of the p_{R_0}, and delay threshold of the DUEs $l_{d_j,\max}$ on the sum rate of the matched UEs for all algorithms is shown in Fig. 3 and Fig. 4. The sum throughput of the matched UEs increases with p_{R_0} and $l_{d_j,\max}$. This is because higher p_{R_0} causes the interference from the CUEs to be more tolerable by the DUEs, therefore making potential CUE-DUE pairing possible. Similarly, higher $l_{d_j,\max}$ increases the sum throughput at fixed outage probability and payload since the delay requirement is less stringent. More DUEs are able to satisfy the delay constraint and the number of admitted DUEs are increased.

V. Conclusion and Future Work

We presented a semi-distributed Base Station Assisted (BSA) scheme for Radio Resource Management (RRM) of a network with D2D and cellular users, targeting wireless industrial scenarios. The reinforcement learning based

 ISBN: 978-1-61208-956-0

approach relies on distributed training of the D2D agents. Subsequently, the look-up tables for the D2D agents are loaded to the base station for centralised channel allocation. Simulation results show that the throughput of the presented approach is comparable to traditional centralised optimisation and demonstrates an improved performance relative to the deferred acceptance (DA) scheme. The future work will focus on evaluating the trade-off between performance, complexity and signaling overheads for the BSA scheme relative to other techniques.

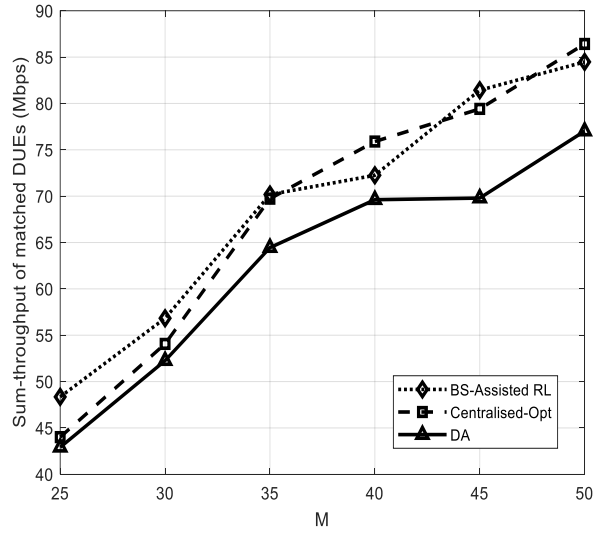

Fig. 1. Sum-rate of matched DUEs with varying number of DUEs, M in the System, for $N = 50$

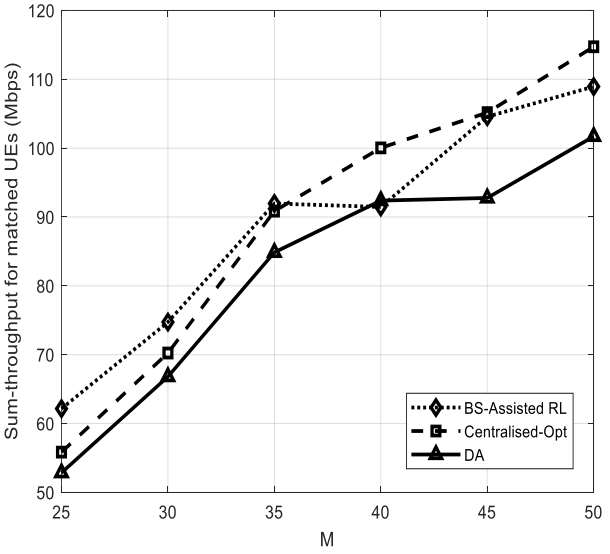

Fig. 2. Sum Throughput of matched UEs as a function of the number of DUEs M, in the system, for $N = 50$

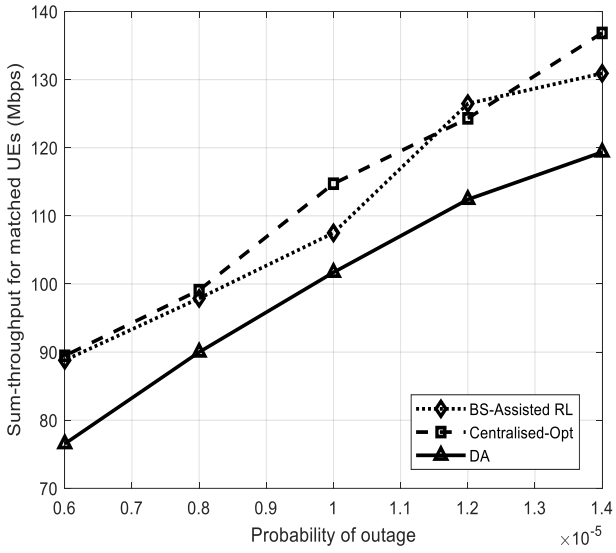

Fig. 3. Effect of the DUE outage ratio p_{R_0}, on the sum throughput for $N = M = 50$, $l_{d_j,\max} = 50$ms

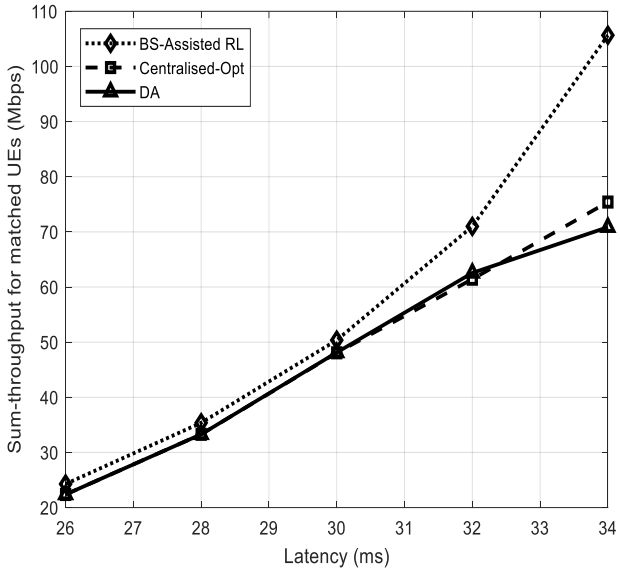

Fig. 4. Effect of the delay bound, $l_{d_j,\max}$ on the sum throughput of matched CUE-DUE pair for $N = M = 50$, $p_{R_0} = 10^{-5}$

REFERENCES

[1] J. Wan et al., “Toward dynamic resources management for IoT-based manufacturing,” IEEE Communications Magazine, vol. 56, no. 2, pp. 52-59, Feb. 2018.

[2] O.N. Yilmaz et al., “Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case,” in Proc. of 2015 IEEE International Conference on Communication Workshop (ICCW), pp. 1190-1195, Sep. 2015.

[3] N. Brahmi, O. N. Yilmaz, K. W. Helmersson, S. A. Ashraf and J. Torsner, “Deployment strategies for ultra-reliable and low-latency communication in factory automation,” in Proc. of 2015 IEEE Globecom Workshops (GC Wkshps), pp. 1-6, Feb. 2016.

 ISBN: 978-1-61208-956-0

[4] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," IEEE Transactions on Vehicular Technology, vol. 69, no. 2, pp. 1828-1840, Dec. 2019.

[5] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in cellular networks," IEEE transactions on communications, vol. 64, no. 9, pp. 3996-4012, Jul. 2016.

[6] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," IEEE Transactions on Vehicular Technology, pp. 1828-1840, Feb, 2020.

[7] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, "Q-learning based power control algorithm for D2D communication," in Proc. of IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1-6, Sep. 2016.

[8] K. Zia et al., "A distributed multi-agent RL-based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets". IEEE Access, no.7, pp. 6733-6745, Jan. 2019.

[9] I.O. Sanusi, K.M. Nasr and K. Moessner, "Radio resource management approaches for reliable Device-to-Device (D2D) communication in wireless industrial applications," IEEE Transactions of Cognitive Communication and Networking, vol. 7, no. 3, pp.905-916, Oct. 2021.

[10] L. Liang, H. Ye and G.Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2282-2292, Aug. 2019.

[11] A.T. Kasgari and W. Saad, "Model-free ultra-reliable low delay communication (URLLC): A deep reinforcement learning framework," in Proc. IEEE International Conference on Communications (ICC), pp. 1-6, May 2019.

[12] H. Wang and X. Chu, "Distance-constrained resource-sharing criteria for device-to-device communications underlaying cellular networks," Electronics letters, vol. 48, no. 9, pp. 528-530, Apr. 2012.

[13] H. Yang, X. Xie and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-delay IoV communication networks," IEEE Transactions on Vehicular Technology, vol. 68, no. 5, pp. 4157-4169, Jan. 2019.

[14] F.E. Souhir, A. Belghith and F. Zarai, "A reinforcement learning-based radio resource management algorithm for D2D-based V2V communication," in Proc. 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 1367-1372, Jun. 2019.

[15] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," IEEE Transactions on Vehicular Technology, pp. 1828-1840, Feb. 2020.

[16] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, "Q-learning based power control algorithm for D2D communication," in Proc. of IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Sep. 2016.

[17] Y. Wei, Y. Qu, M. Zhao, L. Zhang and F.R. Yu, "Resource allocation and power control policy for Device-to-Device communication using multi-Agent reinforcement learning," Computers, Materials & Continua, vol. 63, no. 3, pp.1515-1532, May 2020.

[18] J. Kim, J. Park, J. Noh and S. Cho, "Autonomous power allocation based on distributed deep learning for device-to-device communication underlaying cellular network," IEEE Access, Vol. 8, 107853-107864, Jun. 2020.

 ISBN: 978-1-61208-956-0