A Machine Learning Approach for Resource Allocation in Wireless Industrial Environments

Idayat O. Sanusi and Karim M. Nasr

{i.o.sanusi, k.m.nasr }@gre.ac.uk

Faculty of Engineering and Science, University of Greenwich, United Kingdom

Outline

- **Introduction and motivation**
- **Aim and objectives**
- **Methodology**
- **Results and discussion**
- **Conclusions and directions for future work**

Introduction and Motivation

- **Device-to-Device communication (D2D)** is considered a key enabling technology for **Ultra-Reliable Low-Latency Communication (URLLC).**

- Achieving ultra-high reliability and ultra-low latency pose challenges in terms of bandwidth requirements

- The **scarcity of radio resources** and the limitations on the available system bandwidth makes **spectrum sharing** a necessity for D2D implementation of machine-type communication (MTC) targeting factory automation

- **Radio Resource Management (RRM)** schemes need to be efficiently designed for **interference management** and coordination while guaranteeing tight URLLC (QoS/QoE) demands

The Different RRM Approaches

- **Centralised approach:** requires global information gathering by base stations often results in a **high signalling overhead** and increased complexity, thus making it impractical for ultra-dense networks.

- **Distributed approach:** terminal-centric and **supports self-organisation**; therefore reducing the amount of information gathering and processing by base stations, but may also **increase signalling overheads due to the high amount of information interchange among devices**.

- **Hybrid approach:** **combines centralised and distributed approaches** in allocating resources among devices with **a trade-off between performance, signalling overhead and complexity**.

Aim and Objectives

- Aim: To maximise the overall system throughput while satisfying the QoS requirements of the cellular users (CUEs), c_i and D2D users (DUEs), d_j.

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

subject to:

$$\lambda_j^i\Gamma_{c_i} - \Gamma_{c_i,min} \geq 0 \qquad \forall c_i \in C \qquad \text{(CUE SINR)}$$

$$Pr\left(l_{d_j} > l_{d_j,max}\right) < 1 - \xi_{d_j}^{*} \qquad \forall d_j \in D \qquad \text{(DUE reliability and latency)}$$

$$\sum_{c_i \in C}\lambda_j^i \leq 1 \qquad \forall d_j \in D \qquad \text{(Channel association)}$$

$$\sum_{d_j \in D}\lambda_j^i \leq 1 \qquad \forall c_i \in C$$

- Objectives: To determine the achieved throughput

Methodology: Stateless Reinforcement Learning

- In Q-learning, 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. The Q-value is updated as follows:

$$Q^{t+1} = \begin{cases} Q^t(s^t,a^t) + \sigma\left[r^t + \eta \max_{a'} Q^t(s^{t+1},a^{t+1}) - Q^t(s^t,a^t)\right] & \text{if } s = s^t, \quad a = a^t \\ Q^t(s^t,a^t), & \text{otherwise} \end{cases}$$

- For our work, the 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.

- 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.

Methodology: Stateless Reinforcement Learning

- The learning environment can be modelled entirely using a **stateless Q-learning** i.e., **action-reward** only since the state transition is not required.. The update function is re-formulated as follows:

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

$r(a^t)$ is the immediate reward of selecting a

- The performance requirements of the CUEs are considered by adopting a scheme in which the base station keeps a look-up table of the ith CUE based on the actions on the DUEs, rather than the BS exchange the measured CUE SINR with the DUEs for every action a^t taken at each time slot as done in other works. Therefore reducing the signalling overheads.

Methodology

Base Station Assisted (BSA) Reinforcement Learning

- The jth DUE only gets a reward when 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.
- For the BSA method, after the training phase, each DUE loads its Q-value table, to the BS for centralised matching.
- The BS will allocate cellular RB to D2D links in such a way that spectrum sharing is optimised and network throughput is maximised.
- There is no need for information exchange between the UEs to find a preferred candidate.

Algorithm Details (1/2)

Distributed Training of DUEs

1: Initialise the action-value function for the DUEs

$$\left[Q_{d_j}(a) = 0 \middle| Q_{d_j}(a) \equiv Q^i_{d_j}(a^t), i = 1,2,\dots,N\right] \forall d_j \in D$$

2: Initialise the action-value function for the BS for the actions of the jth DUE on the ith RB

$$\left[Q_{c_i}(a) = 0 \middle| Q_{c_i}(a) \equiv Q^j_{c_i}(a^t),\ j = 1,2,\dots,M\right] \quad \forall c_i \in C$$

3: for $d_j \in D$ $1 \le j \le M$ do

4: **while** not converge **do**

5: generate a random number $x \in \{0,1\}$

6: **if** $x < \varepsilon$ **then**

7: Select action a^t_i randomly

8: **else**

9: Select action $a^t_i = \underset{a \in A}{\operatorname{argmax}} Q_{d_j}(a^t)$

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**

Algorithm Details (2/2)

Centralised Channel Allocation

```
Load $Q_{d_j}(a)$ to the BS        $\forall\, d_j \in D$
for $d_j \in D$  $1 \le j \le M$  do
    Obtain $Q(a) = \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\}$  $i = 1,2, \dots, N$
    $\bar{Q}(a) \subseteq Q(a) | \left\{Q^i_{d_j}(a), Q^j_{c_i}(a)\right\} \in \mathbb{R}^+$, where $\mathbb{R}^+$
     positive real number
     $Q_{\mathrm{TOT}} = Q^i_{d_j}(a) + Q^j_{c_i}(a)$        $\forall\, q \in \bar{Q}(a)$
end for
Set up a list for unmatched DUE $D_u = \left\{d_j : \forall\, d_j \in D_u\right\}$
while $D_u \neq \emptyset$ do
     Rank $D_u$ in increasing order of $|0\ \bar{Q}(a)|$
     Start DUE $d_j \in D_u$: $\bar{Q}(a) \neq \emptyset$ with the least $|\ \bar{Q}(a)|$
      $c_i^* = \max_{r_i \in R} Q_{\mathrm{TOT}}$
      $D_u = D_u - d_j$
      $\bar{Q}(a) = \bar{Q}(a) \backslash c_i^*$       $\forall\, d_{j'} \in D_u \,|\, j' \neq j$
end while
```

Results (1/2)

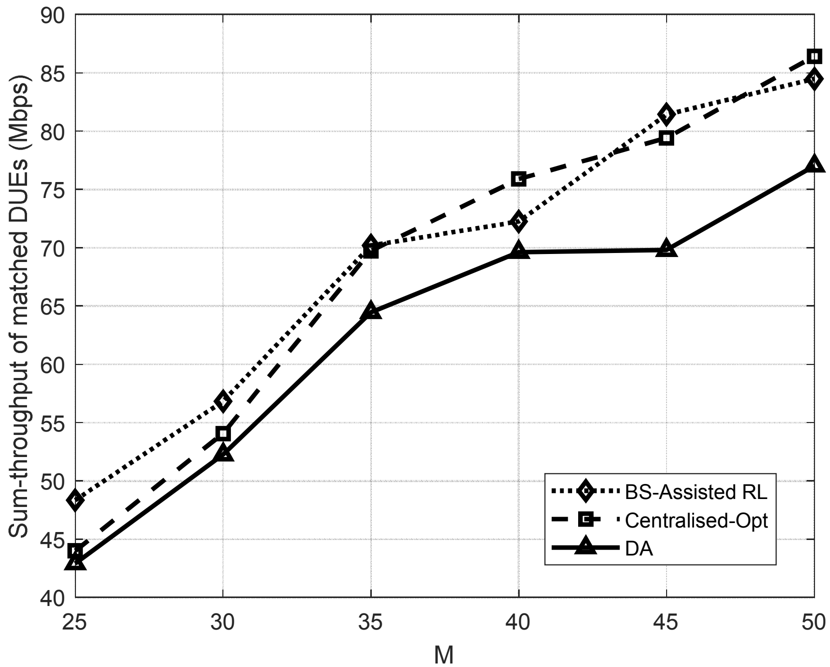

Fig. 1. Sum-rate of matched DUEs with varying number of DUEs, M in the System, for $N = 50$

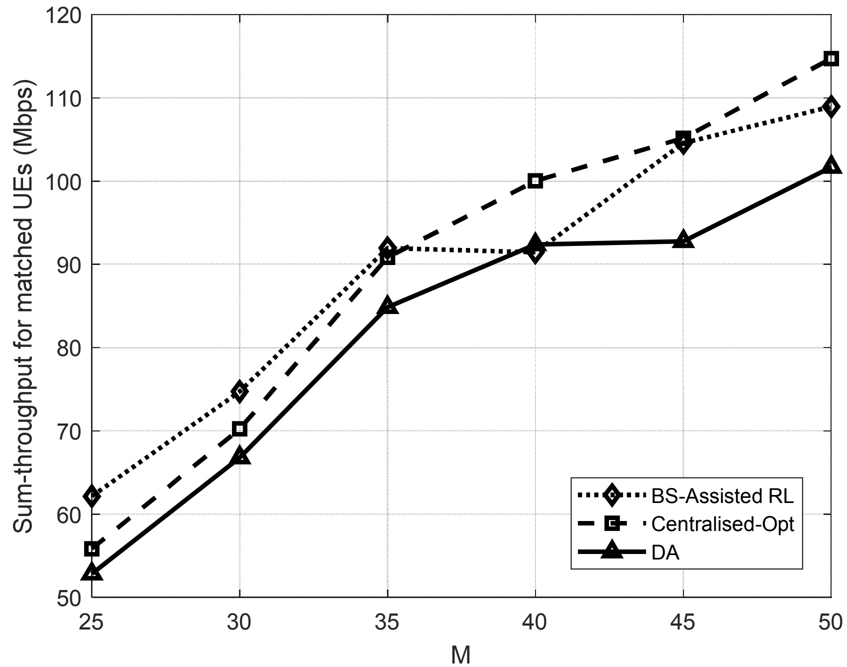

Fig. 2. Sum Throughput of matched UEs as a function of the number of DUEs M, in the system, for $N = 50$

Results (2/2)

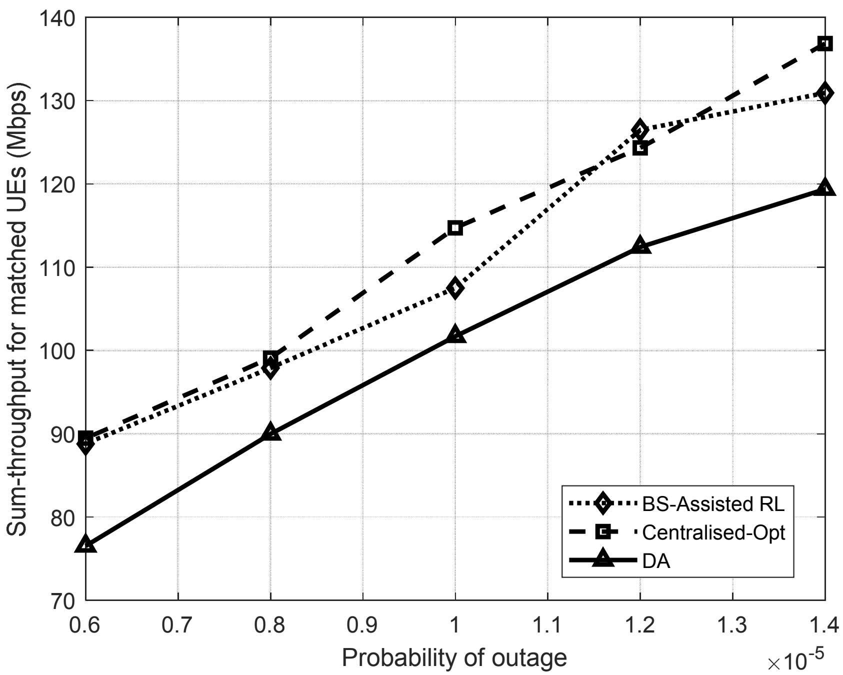

Fig. 3. Effect of the DUE outage ratio p_{R_0}, on the sum throughput for $N = M = 50$, $l_{d_j,\max} = 50$ms

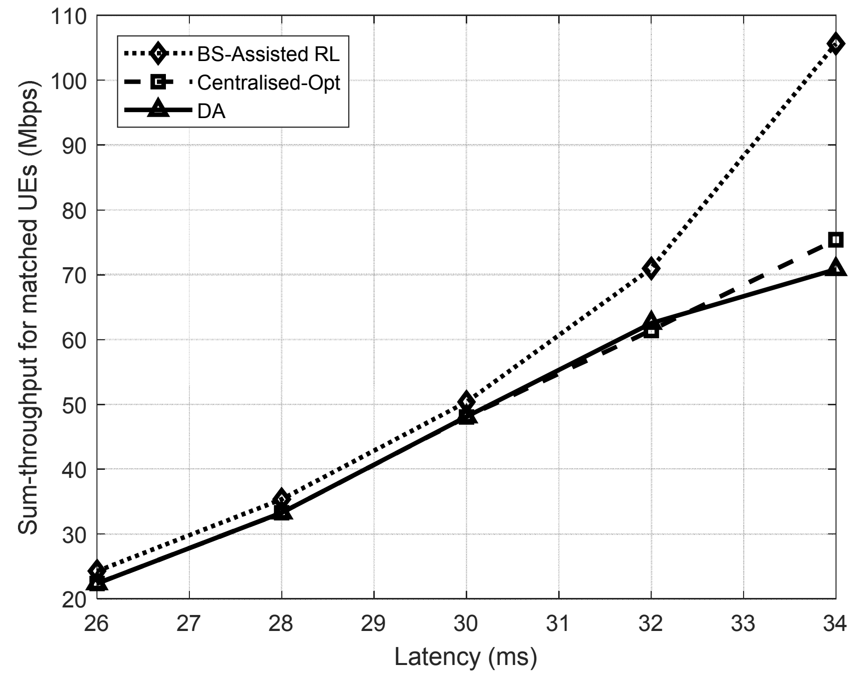

Fig. 4. Effect of the delay bound, $l_{d_j,\max}$ on the sum throughput of matched CUE-DUE pair for $N = M = 50$, $p_{R_0} = 10^{-5}$

Conclusions and Directions for Future Work

- 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 was presented.

- The reinforcement learning based approach presented 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 BSA 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 signalling overheads for the BSA algorithm relative to other techniques.

Thank you for your attention!