2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

# Reinforcement Learning for Joint Resource and Power Allocation in D2D Communications

Ifrah Saeed, Andrew C. Cullen, Zainab R. Zaidi, Sarah Erfani and Tansu Alpcan

*Abstract*— **Device-to-Device (D2D) communication is critical in many public safety scenarios where access to network infrastructure is not guaranteed. In such out-of-coverage situations, random resource allocation is a straightforward way for direct communication by each D2D pair. However, such a resource allocation plan often results in severe interference and low throughput, which can degrade crucial communications performance. In contrast, the problem of centralised joint resource block selection and power control optimisation to maximise the total data rate of all D2D pairs, known as the sum-rate, is non-convex and NP-hard but can effectively reduce interference. To address the inability to solve this non-convex centralised problem setting due to limited information in resource-constrained out-of-coverage scenarios, we propose two distributed reinforcement learning schemes. Our methods allow D2D pairs to autonomously make decisions on their joint resource block and power allocation to minimise their mutual interference and maximise their sum-rate, while maintaining the quality of service constraints. To evaluate these methods in out-of-coverage scenarios, we conduct extensive performance evaluations using the ns-3 network simulator designed for public safety LTE-D2D. We demonstrate that our algorithms, in the absence of any centralised coordination and neighbouring information, autonomously reach time-averaged sum-rates that are within $98.2\%$ and $98.6\%$ of the rates achieved by the centralised optimisation solution.**

## I. INTRODUCTION

It is disconcertingly true that the times when communication is most crucial—such as natural disasters or other emergencies—are also when network infrastructure is most likely to fail due to systemic issues and congestion. These failures can severely impact the ability of first responders to coordinate effectively. In such out-of-coverage scenarios, where traditional network infrastructure and coverage are unavailable, direct communication among responders is critical. LTE-based Device-to-Device (D2D) networks, leveraging existing LTE infrastructure, have emerged as a technological solution for maintaining communication in these situations [1], [2]. These networks facilitate direct communication between devices, prioritise public safety traffic during emergencies, and address the limitations of Land Mobile Radio (LMR) systems, which do not support broadband applications [3].

However, in these out-of-coverage D2D networks, resource allocation is typically performed randomly, resulting in significant interference between devices. Therefore, these D2D networks must be organised in a way that effectively minimises interference and maximises overall throughput, all within the constraints of limited resources including resource blocks and power. This is typically formalised as a joint resource block and power allocation optimisation problem that aims to maximise the combined data rates of D2D pairs, known as sum-rates, while satisfying quality of service (QoS) constraints [4]. Existing optimisation-based solutions to this non-convex, NP-hard problem rely on assumptions of access to detailed network information, such as instantaneous global network data [5] or extensive instantaneous signalling from D2D pairs [6]. However, these assumptions are impractical and render these methods unsuitable for resource-constrained, out-of-coverage scenarios.

In contrast to optimisation, Reinforcement Learning (RL) has demonstrated the ability to effectively scale to complex environments without requiring prior knowledge of environment dynamics [7]–[9]. In D2D networks, single-agent RL methods enable devices to adapt to the dynamic environment through experience, while Multiagent RL (MARL) allows them to understand both the dynamics of the environment and effective coordination with other pairs [4], [10]–[14]. These RL frameworks have successfully maximised the sum-rate in underlay networks by minimising interference between cellular and D2D users [10]–[12] and in overlay networks by reducing mutual interference between D2D users [4], [13], [14]. However, it must be noted that the training environments for these RL methods are typically either static, semi-static or episodically dynamic, which do not fully capture the dynamic nature of wireless environments. Moreover, while these implementations do not rely on instantaneous information as optimisation-based solutions, they make analogous assumptions regarding access to the channel state or signalling information, making them similarly unsuitable for out-of-coverage scenarios.

To address this limitation, recent works have developed methods designed for out-of-coverage D2D scenarios by employing distributed resource allocation schemes to improve system coverage probability [15]; priority-based preemptive scheduling schemes [16]; and investigating the channel parameters that affect the channel reliability and the amount of delay introduced during communication to ensure a guaranteed bit rate of public safety network applications [17]. However, while these methods address the unique needs of out-of-coverage scenarios, they do not solve the sum-rate maximisation problem and have not explored RL algorithms.

To fill these gaps, our work proposes two new RL-

I. Saeed and T. Alpcan are, and Z.R. Zaidi was with the Department of Electrical and Electronic Engineering, The University of Melbourne, Australia. ifrah.saeed@student.unimelb.edu.au; tansu.alpcan@unimelb.edu.au, zzaidi@ieee.org

A.C. Cullen and S. Erfani are with the School of Computing and Information Systems, The University of Melbourne, Australia {andrew.cullen, sarah.erfani}@unimelb.edu.au

based algorithms, named **Independent-D2D (ID2D)** and **Distributed-D2D (DD2D)**, which are designed to overcome the limitations of existing optimisation and RL methods and are specifically tailored for sum-rate maximisation in resource-constrained, out-of-coverage scenarios. Of these, ID2D uses locally available information at all times; while DD2D employs a centralised training and decentralised execution MARL framework [18]–[20] that requires only local information on deployment. To ensure the realism of both approaches, our simulations are built upon ns3-ai [21] and the ns-3 LTE-D2D network simulation model modified for public safety scenarios [22], following the Proximity Services (ProSe) standard for LTE networks that is developed by the 3rd Generation Partnership Project (3GPP) [23].

Our main contributions are as follows:

- We design two new distributed RL algorithms for sum-rate maximisation that enable D2D pairs to autonomously make decisions on resource block and power allocation using only local historical information without the aid of network infrastructure.
- We develop our experiments using ns-3 network simulator designed for public safety scenarios, ensuring the applicability of our methods in practical situations.
- We demonstrate the validity and utility of our techniques through extensive simulations, showing that our *distributed* methods that use limited historical information, can achieve performance comparable to the unrealistic *centralised* methods that use instantaneous, global data.

## II. Model and Background

In this section, we present the system and channel models, an overview of RL, and necessary mathematical notation.

### A. System Model

We consider $N$ out-of-coverage D2D transmitter-receiver (Tx-Rx) pairs that are deployed in a specific geographic location to perform public safety tasks. Since the D2D pairs are not connected to the network, each pair is equipped with a predetermined resource pool of $M$ resource blocks to facilitate sidelink D2D communication. $N$ D2D pairs share these $M$ resource blocks and for the case $N > M$, interference between D2D pairs is likely to occur when multiple devices choose the same resource block or transmits at high power.

We denote D2D pair $i$ as the $i$-th D2D pair where $i \in \mathcal{I} = \{1, 2, ..., N\}$ is the set of D2D pairs and resource block $j \in \mathcal{J} = \{1, 2, ..., M\}$ represents the set of available $M$ resource blocks. We assume that all Tx in D2D pairs have data to send to their respective Rx, are synchronised, and use the same set of sidelink parameters and resource pool configuration. Here power control and resource block allocation are respectively performed at the physical and mac layers.

### B. Channel Model

We consider a standard wireless channel model for D2D communications, in which the channel gain between Tx-Rx pair $i$ using resource block $j$ at time $t$ is a function of pathloss $\mathcal{X}_{i,i}$, log-normal shadowing $\beta_{i,i}$, and fast-fading $g_{(i,i,j)t}$ that is given by [4], $h_{(i,i,j)t} = \mathcal{X}_{i,i}\beta_{i,i}|g_{(i,i,j)t}|^2$.

### C. RL Framework

RL is used to learn optimal decisions in a dynamic environment by trial and error over time. A single-agent RL is modelled as a Markov Decision Process (MDP) that consists of a set of states $X$, an action set $U$, a state transition probability function $\mathcal{T}$, a discount factor $\gamma \in [0, 1]$, and a reward function $R$. At each timestep $t$, with $x \in X$ and $u \in U$, the agent finds a policy $\pi(u|x)$ that maximises the expectation of a long-term cumulative discounted reward $r$ as $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(x_t, u_t)\right]$ [24].

Multiagent RL (MARL) framework is usually used in the presence of multiple agents in the system. It can be modelled as Markov games [25] and is defined by a state set $X$, action sets $U_1, ..., U_N$ for each of $N$ agents, a state transition function $\mathcal{T} : X \times U_1 \times ... \times U_N \xrightarrow{P(X)} X$, a discount factor $\gamma$, and a reward function $R_i : X \times U_1 \times ... \times U_N \to \mathbb{R}$ for each agent. $P(X)$ defines the probability distribution over possible next states, given the current state and actions for each agent. For the partial-observable case, each agent $i$ receives an observation $o_i$, containing partial information of the global state $x \in X$. Since the state transitions are the result of the joint action of all the agents, the joint action at any time $t$ is represented as $u_t = \{u_{1t}, ..., u_{Nt}\}$. The objective of each agent $i$, $J_i$, is to learn a policy $\pi_i$ that maximises its expected discounted returns taking into account the presence of other agents in the environment, and is given by

$$J_i(\pi_i) = \mathbb{E}_{u_1 \sim \pi_1, ..., u_N \sim \pi_N, x \sim \mathcal{T}} \left[ \sum_{t=0}^{\infty} \gamma^t r_{it}(x_t, u_t) \right].$$

## III. Problem Definition

To minimise interference and maximise the sum-rate among D2D pairs in out-of-coverage networks, we consider the problem of jointly optimising resource block $j$ selection and transmit power $p_i$ for each pair $i$ while satisfying the QoS requirements at transmission time $t$. If $B$ is the resource block bandwidth and $r_{(i,j)t}$ is the data rate of pair $i$ using $j$, the optimisation problem takes the form [4]

$$\max_{\alpha_t, \mathcal{P}_t} \sum_{i=1}^{N} \sum_{j=1}^{M} B \times r_{(i,j)t} \tag{1}$$

$$\text{where} \quad r_{(i,j)t} = \log_2\left(1 + \xi_{(i,j)t}\right),$$

$$\xi_{(i,j)t} = \frac{\alpha_{(i,j)t} h_{(i,i,j)t} p_{it}}{\sigma^2 + \sum\limits_{k \in \mathcal{I}, k \neq i} \alpha_{(k,j)t} h_{(k,i,j)t} p_{kt}}$$

such that $\forall t \in T$,

$$P_{min} \leq p_{it} \leq P_{max}, \forall i \in \mathcal{I}, \tag{1a}$$

$$\alpha_{(i,j)t} \in \{0, 1\}, \forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \tag{1b}$$

$$\sum_{j=1}^{M} \alpha_{(i,j)t} \leq 1, \forall i \in \mathcal{I}, \tag{1c}$$

$$\xi_{(i,j)t} \geq \xi_{thresh}, \forall i \in \mathcal{I}, \forall j \in \mathcal{J}. \tag{1d}$$

Here, $P_{min}$ and $P_{max}$ are respectively the minimum and maximum transmit power levels, $\xi_{thresh}$ is the minimum signal to interference and noise ratio (SINR) requirement, and $\sigma^2$ is additive white Gaussian noise (AWGN) power. For a given D2D pair $i$ that uses the resource block $j$ at transmission time $t$, $\xi_{(i,j)t}$ is its SINR, $p_{it}$ is its transmit power, and $h(i,i,j)$ is the channel gain. We use $\alpha_{(i,j)t} = 1$ to indicate that $i$ uses $j$ to transmit at time $t$ and $\alpha_{(i,j)t} = 0$ otherwise. $\alpha_t$ and $\mathcal{P}_t$ represent the sets containing pairs' resource block allocation and power selection at $t$. Moreover, each D2D pair chooses at most one resource block to transmit given by constraint (1c) and QoS constraint (1d) specifies that each pair SINR should be greater than the minimum threshold value.

Eq. (1) is a non-convex and mixed integer nonlinear programming (MINLP) problem, which is challenging to solve [4]. Constructing such a solution requires *global* and *instantaneous* channel state information at each transmission, both of which are difficult to realise in practical situations. Additionally, the problem must be re-solved with every change in channel state, making it computationally intensive. Therefore, while both optimisation and RL require significant computational time to solve this problem, RL's cost is primarily incurred during training. On deployment after the training, the learned policies are able to reliably adapt to previously unseen, dynamic environments with minimal cost. This allows RL methods to reach higher scalability at reduced signalling overheads during in-coverage scenarios, and motivates our new algorithms for out-of-coverage scenarios, that only rely upon local information to maximise sum-rates.

## IV. RL based Resource Block Selection and Power Control Algorithms

In this section, we provide details of our RL-based methods, Independent-D2D (ID2D) and Distributed-D2D (DD2D), with which D2D pairs can autonomously optimise resource block selection and transmit power by only using local historical information to maximise sum-rate in out-of-coverage scenarios. The key elements of our methods are:

- *Agent:* Each D2D pair $i$ is an agent that makes the decision of the resource block and power level at the transmission time $t$. We use D2D pair and agent interchangeably in this work.
- *State Space:* The state space should provide enough information for agents to make decisions that maximise the overall reward in RL. In out-of-coverage scenarios, only partially observable local information is available to the agent. Therefore, for our algorithms, an agent $i$'s observation at time $t$ incorporates its choice of resource block $j_i$, power level $p_i$, and SINR $\xi_i$ received by the corresponding Rx at the previous time $(t-1)$, taking the form

$$o_{it} = \left\{ j_{i(t-1)}, p_{i(t-1)}, \xi_{i(t-1)} \right\}, \tag{2}$$

unlike methods [14] that needs multiple past observations. While historical information of a time-varying channel is inherently outdated, it provides sufficient information on the interference caused by other agents to allow for inferring both environment dynamics and the appropriate response at the current state.

- *Action Space:* The transmit power is discretised into $L$ power levels between $P_{min}$ and $P_{max}$. If each power level is given by $l \in [1, ..., L]$ then transmit power $p \in \mathcal{P}$ is calculated as $p = P_{min} + (P_{max} - P_{min})\, l/L$. The action taken by the agent $i$ at the transmission time $t$ constitutes the chosen resource block $j_i$ and power level $p_i$ is then

$$u_{it} = \{ j_i, p_i \mid \forall j_i \in \mathcal{J}, p_i \in \mathcal{P} \}. \tag{3}$$

- *Reward:* Each agent $i$ receives a positive reward equal to its throughput if all the constraints in Eq. (1) are satisfied. Our RL algorithms are constructed in a way that the first three constraints in the optimisation problem cannot be violated. However, to ensure that the constraint (1d) is satisfied, a negative reward of $-0.2$ is given if this constraint is violated, following prior works [10]. This negative reward for sub-threshold behaviour discourages agents from choosing sub-optimal actions. Thus, the reward function is defined as

$$r_{it} = \begin{cases} \sum_{j=1}^{M} B \times r_{(i,j)t} & \text{if } \xi_{it} \geq \xi_{thresh} \\ -0.2 & \text{if } \xi_{it} < \xi_{thresh} \end{cases}, \tag{4}$$

in which each agent $i$ tends to maximise its own reward for the distributed implementation of the algorithm.

- *Time-averaged Sum-rate:* As the sum-rate can vary due to the dynamic environment, the time-averaged sum rate

$$F_t = \frac{\sum_{t'=1}^{t} f_{t'}}{t}, \text{with } f_t = \sum_{i=1}^{N} \sum_{j=1}^{M} B \times r_{(i,j)t} \tag{5}$$

is preferable, calculated by averaging the sum-rate $f_t$ at a given time $t$ up to that $t$ [10].

### A. Independent-D2D (ID2D)

ID2D is a fully-distributed actor-critic based RL implementation in which each agent learns independently without explicitly considering the presence or actions of other agents, who are treated as part of the environment. As there is no coordination or communication between agents, $o_i$ is defined for each agent $i$ without access to the global state $x$. Each agent has an independent actor and a critic network, where the actor considers the agent's current observation and learns policy, and the critic scores the actor based on its choice of action. Each critic separately updates its parameter $\psi_i$ via a value function approximation, and each actor updates its parameter $\theta_i$ based on the critique from its critic.

Specifically, all critics independently learn their own state-action value functions to estimate expected discounted returns $Q_i^{\psi_i}(o_{it}, u_{it}) = \mathbb{E}[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{it'}(o_{it'}, u_{it'})]$. We use a one-layer Multi-Layer Perceptron (MLP) embedding function $g_i$ on critic input before passing it to the critic network. Each $Q$-function is learned through off-policy temporal-difference learning by minimising the regression loss as

$$\mathcal{L}_{Q_i}(\psi_i) = \mathbb{E}_{(o_i, u_i, r_i, o_i') \sim D} \left[ \left( Q_i^{\psi_i}(g_i(o_i, u_i)) - y_i \right)^2 \right],$$

$$\text{where } y_i = r_i(o_i, u_i) + \gamma \mathbb{E}_{u_i' \sim \pi_{\bar{\theta}_i}(o_i')} \left[ Q_i^{\bar{\psi}_i}(g_i(o_i', u_i')) \right],$$

in which $D$ represents the experience replay buffer and the parameter of corresponding target policy $\pi$ is $\bar{\theta}_i$. The target $Q$-function of each agent $Q_i^{\bar{\psi}_i}(g_i(o_i', u_i'))$ is the exponential moving average of past $Q$-functions with $\bar{\psi}_i$ is the parameter of agent $i$'s target critic. The actor network of each agent $i$ learns the policy $\pi_{\theta_i}$ that is updated by ascent through

$$\nabla_{\theta_i} J(\pi_{\theta_i}) = \mathbb{E}_{o_i \sim D, u_i \sim \pi_{\theta_i}} \big[\nabla_{\theta_i} \log(\pi_{\theta_i}(u_i|o_i)) \; Q_i^{\psi_i}(g_i(o_i, u_i))\big] \tag{6}$$

### B. Distributed-D2D (DD2D)

DD2D is a centralised training and decentralised execution based actor-critic MARL algorithm in which each agent $i$ applies its learned policy based on only its own observation similar to ID2D. However, in contrast to ID2D, DD2D allows agents access the observations $o = (o_1, ..., o_N)$ and actions $u = (u_1, ..., u_N)$ of all agents 1 to $N$ during centralised critic training, thereby enhancing their capacity to obtain a more comprehensive understanding of the global system state. This also helps with overcoming the non-stationarity in a multiagent environment. Therefore, the critic for agent $i$ takes the form

$$Q_i^{\psi}(o, u) = f_i\left(g_i(o_i, u_i), v_{-i}(g_{-i}(o_{-i}, u_{-i}))\right),$$

in which $-i$ represents all agents except $i$, $g_i$ is a one-layer MLP embedding function, $f_i$ is a two-layer MLP, $v_{-i}$ is a function that takes other agents embeddings encoded with an embedding function $g_{-i}$ and passes them through a leaky Rectified Linear Unit after linearly transforming them by a shared matrix [19]. In DD2D, the critic networks of all agents are trained together through temporal difference learning to minimise the joint regression loss, with the critic loss being

$$\mathcal{L}_Q(\psi) = \sum_{i=1}^{N} \mathbb{E}_{(o,u,r,o') \sim D}\left[\left(Q_i^{\psi}(o, u) - y_i\right)^2\right], \quad \text{where } y_i = r_i(o_i, u_i) + \gamma \mathbb{E}_{u' \sim \pi_{\bar{\theta}_i}(o')}\left[Q_i^{\bar{\psi}}(o', u')\right]. \tag{7}$$

Each agent's policy $\pi_{\theta_i}$ is updated through gradient ascent

$$\nabla_{\theta_i} J(\pi_\theta) = \mathbb{E}_{o \sim D, u \sim \pi_\theta}\big[\nabla_{\theta_i} \log\left(\pi_{\theta_i}(u_i|o_i)\right)\left(A_i(o, u)\right)\big],$$

where $A_i(o, u) = Q_i^{\psi}(o, u) - b(o, u_{-i})$ is the advantage function using a multiagent baseline and is shown to help overcome the multiagent credit assignment problem [18], [19]. The baseline marginalises out the actions of the given agent from $Q_i^{\psi}(o, u)$ and is calculated for our discrete action space as $b(o, u_{-i}) = \sum_{u_i'} \pi(u_i'|o_i) Q_i(o, (u_i', u_{-i}))$. Here, we exclude the agent $i$'s action from the input to the critic and get the $Q$-value for all actions for $i$. Moreover, actions for policy gradients are sampled from all agents' current policies to avoid over generalisation [19].

*Remark 1:* Our two proposed methods, ID2D and DD2D, are suitable for use in public safety scenarios when the network infrastructure is not guaranteed. Both methods do not require instantaneous global or neighbouring information during deployment and rely only on local information from the last transmission data defined in Eq. (2). DD2D is trained in a centralised way that helps get a rich training experience. However, each agent makes independent decisions after training, relying only on its local observations and the policy it learned during training.

## V. Experimental Evaluation

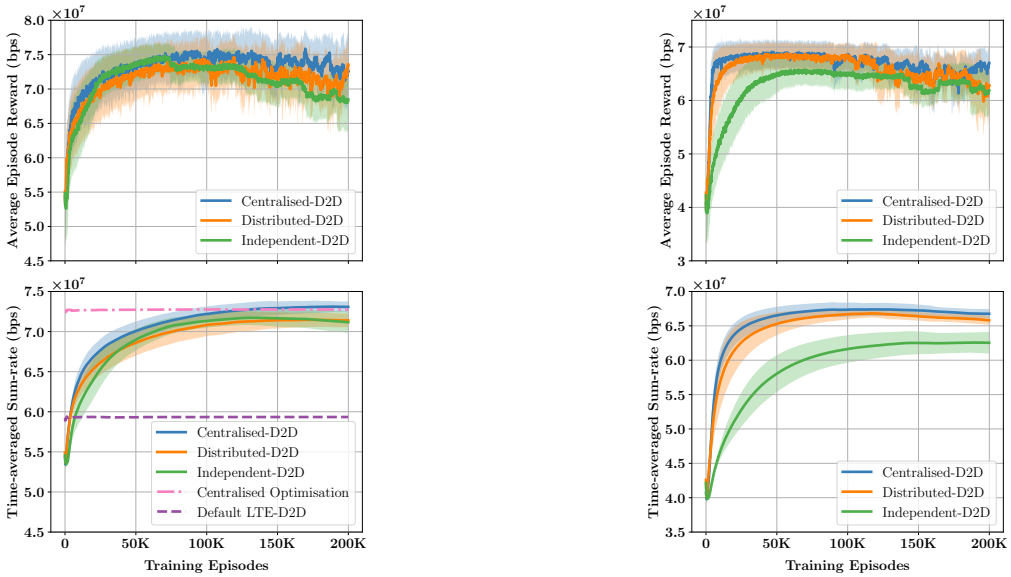
In this section, we conduct experiments to evaluate the performance of our proposed schemes in out-of-coverage scenarios using ns-3 simulator modified for LTE-D2D simulations [22] and ns3-ai [21]. The simulations are conducted with $N = 20$ D2D pairs, each randomly placed within a $500\text{m} \times 500\text{m}$ area [4]. For each D2D pair, the receiver is positioned randomly around the transmitter at a distance between 10m and 50m [4], [10]. The resource pool has $M = 15$ resource blocks, each with a bandwidth of 180kHz. Power levels are discretised into 10 levels, ranging from $P_{min} = 10\text{dBm}$ to $P_{max} = 31\text{dBm}$, which is a standard for public safety applications [23]. We use ns-3 LTE-D2D defined Hybrid 3GPP Outdoor to Outdoor Propagation Loss Model [23], log-normal distribution with a standard deviation of 7dB for shadowing, and Nakagami-m model with $m = 1$ for Rayleigh Fading. We set SINR threshold ($\xi_{thresh}$) as 7.8dB and AWGN ($\sigma^2$) as $-174\text{dBm/Hz}$. To determine the sum-rate, we assume that all control packets preceding data packets reach the respective receivers without loss.

For the RL algorithms, all neural networks used for policies and critics have two hidden layers, each with a dimension of 128, and employ a leaky Rectified Linear Unit as the activation function. After every 10 steps, we perform updates for the policies and critics and for each update, we sample mini-batches of size 100 from the replay buffer. We then perform gradient descent on the $Q$-function loss objective and the policy objective using Adam as the optimiser for both with a learning rate of 0.001. After these, we perform the soft update on the parameters of target critic(s) and policies with our learned critics and policies parameters using the update rate $\tau$ set to 0.001. We use a discount factor $\gamma$ of 0.99.

### A. Comparative Study

To analyse our ID2D and DD2D methods, we compare them with a Centralised Optimisation method, the default LTE-D2D method using the maximum transmit power, and a Centralised-D2D (CD2D) method. Note that the training environment is dynamic because of the introduction of Rayleigh fading. For the Centralised Optimisation method, we use Genetic Algorithm from the MATLAB Global Optimisation Toolbox [26] that solves the optimisation problem in Eq. (1) separately for every transmission time or episode for the current channel state information. The LTE-D2D method is the default random selection of resource blocks by ns-3 LTE-D2D out-of-coverage [22] for every episode using $P_{max}$. We also introduce CD2D, where each D2D pair has access to the state spaces of all other D2D pairs during both training and execution, providing a basis for comparing its performance to other RL methods where each agent accesses only its own state space.

We use the reward defined in Eq. (4) and the time-averaged sum-rate defined in Eq. (5) as performance metrics for our




(a) Standard D2D scenario

(b) D2D scenario with significantly increased interference

Fig. 1: Reward (Eq. (4)) and Time-averaged Sum-rate (Eq. (5)) in 200K training episodes by our methods and baseline methods. Average (solid line) and standard deviation (shaded region) from six independent simulations, with smoothing applied over 100 episodes for RL.

algorithms. For RL algorithms, both the metrics are averaged over six independent simulations each, following established RL practices [19], [20] and are smoothed over a window size of 100. A single simulation is performed for each of the centralised optimisation and LTE-D2D methods with smoothing over a 100-sized window size.

Fig. 1(a) illustrates the learning behaviour of our RL algorithms and their comparative performance with the Centralised Optimisation method and default LTE-D2D base case. The top figure illustrates the achieved rewards by RL algorithms in their learning process so they cannot be compared to non-learning methods. The bottom graph captures the time-averaged sum-rate achieved by our proposed algorithms and all other methods. While no guarantee of convergence exists for such RL methods, experimental results demonstrate that they attain their maximum achieved metric before their performance is degraded as they continue to explore and train. This is standard with RL algorithms [18], [20] and the model giving the best metric during training is used in deployment rather than the model learned by the end of training. Compared to Centralised Optimisation, ID2D and DD2D achieve 98.6% and 98.2% of the time-averaged sum-rate respectively. However, when considered relative to the default LTE-D2D method, ID2D and DD2D exhibit significant improvements of 20.7% and 20.3% respectively. This improvement is due to the LTE-D2D method employing random resource allocation in out-of-coverage scenarios, which can result in multiple devices choosing the same resource blocks, thereby causing low sum-rates.

Counterintuitively, ID2D, CD2D and DD2D all achieve similar maximum reward and sum-rate values, even though the former only had access to local information, in contrast to the latter two which were trained with information from all D2D pairs. This is likely due to the SINR data in the state space that implicitly contains enough interference

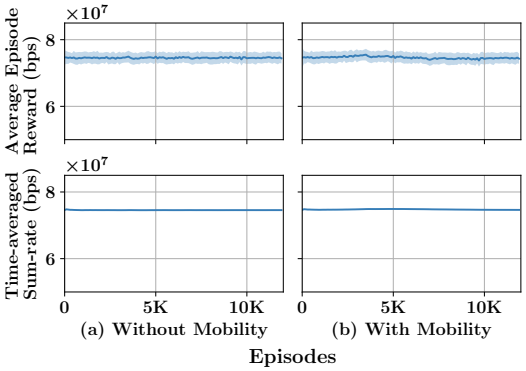



Fig. 2: Reward and Time-averaged Sum-rate achieved by ID2D model trained until 75K episodes and smoothed over 100 episodes.

information for the ID2D to learn appropriate actions. However, the need to learn this information in ID2D requires additional training time at the start compared to others. Moreover, as a single-agent RL algorithm, ID2D does not account for the non-stationarity introduced by other D2D pairs, which decreases its stability. This is evidenced by its abrupt divergence after attaining its maximum reward. In future work, we aim to fine-tune hyperparameters for our RL methods to achieve enhanced performance and stability.

Fig. 2 illustrates the performance of ID2D trained until 75K episodes when it reaches its maximum reward value. It is evident from the graphs that the model has learned to provide high reward and sum-rate in deployment. We re-evaluate the learned model performance when D2D pairs are moving at the speed of 5km/h [4]. Fig. 2(b) confirms that the trained model is robust enough to perform equally well in the case of mobility. In future work, we plan to test the model's performance across different speeds and mobility models.

To verify our intuition behind the ID2D's high performance, we significantly increase the interference between the D2D pairs by constraining the scenario area to 200m×200m,

well below the expectations for a realistic out-of-coverage scenario. Through this, we observe a significant increase in DD2D performance compared to ID2D in Fig. 1(b), confirming that the information from other users becomes important to learn the environment dynamics in the presence of high interference in a highly congested area. However, local information inherently including interference data is enough in the standard D2D scenario that resulted in high performance for ID2D in Fig. 1(a).

## VI. Conclusion

In this paper, we explore the potential of using RL for resource allocation in resource-constrained, out-of-coverage public safety D2D communications. We propose two distributed RL-based methods for autonomous resource selection and power control to solve the sum-rate maximisation problem while maintaining the QoS requirements. To be applicable to real-world scenarios where centralised network infrastructure is not guaranteed to exist, our algorithms use limited individual information from the previous transmission time without accessing the channel state information or signalling data from other D2D pairs on deployment. Simulation results demonstrate that our methods, ID2D and DD2D respectively attain $98.6\%$ and $98.2\%$ time-averaged sum-rates compared to the centralised optimisation framework, which relies on access to instantaneous channel and global information. We also show that DD2D outperforms ID2D in highly congested areas with significant interference, as centralised training improves the understanding of the environment and inter-agent dynamics in these conditions. The ability of our methods to learn and adapt in the dynamic wireless environment and achieve this level of performance demonstrates their utility for resource-constrained, out-of-coverage scenarios. In future, we intend to integrate our proposed schemes with 5G New Radio architecture, which provides enhanced QoS operations with backward LTE compatibility and a new feedback channel [27].

## VII. Acknowledgments

This research was supported in part by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative, and the Australian Research Council Linkage Project under the grant LP190101287.

## References

[1] M. Matracia, N. Saeed, M. A. Kishk, and M.-S. Alouini, "Post-Disaster Communications: Enabling Technologies, Architectures, and Open Challenges," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1177–1205, 2022.

[2] S.-Y. Lien, C.-C. Chien, F.-M. Tseng, and T.-C. Ho, "3GPP device-to-device communications for beyond 4G cellular networks," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 29–35, 2016.

[3] "Public Safety Mobile Broadband Strategic Review," Australian Government, Final Report, 2022.

[4] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep Reinforcement Learning for Joint Channel Selection and Power Control in D2D Networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1363–1378, 2021.

[5] S. N. Swain, S. Mishra, and C. S. R. Murthy, "A novel Spectrum Reuse Scheme for Interference Mitigation in a Dense Overlay D2D Network," in *Proceedings of the IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*, 2015, pp. 1201–1205.

[6] J. Lyu, Y. H. Chew, and W.-C. Wong, "A Stackelberg Game Model for Overlay D2D Transmission With Heterogeneous Rate Requirements," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8461–8475, 2016.

[7] I. Saeed, T. Alpcan, S. M. Erfani, and M. B. Yilmaz, "Distributed Nonlinear Model Predictive Control and Reinforcement Learning," in *Proceedings of the Australian & New Zealand Control Conference*, 2019, pp. 1–3.

[8] V. Lima, M. Eisen, and A. Ribeiro, "Learning Constrained Resource Allocation Policies in Wireless Control Systems," in *Proceedings of the IEEE Conference on Decision and Control*, 2020, pp. 2615–2621.

[9] D. Wei, P. Yi, and J. Lei, "Multi-Agent Deep Reinforcement Learning for Large-Scale Platoon Coordination with Partial Information at Hubs," in *Proceedings of the IEEE Conference on Decision and Control*, 2023, pp. 6242–6248.

[10] Z. Sun and M. R. Nakhai, "Channel Selection and Power Control for D2D Communication via Online Reinforcement Learning," in *Proceedings of the IEEE International Conference on Communications*, 2021, pp. 1–6.

[11] P. Gong, C. Wang, J. Sheu, and D. Yang, "Distributed DRL-based Resource Allocation for Multicast D2D Communications," in *Proceedings of the IEEE Global Communications Conference*, 2021, pp. 01–06.

[12] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3143–3156, 2021.

[13] T.-W. Ban, "A Deep Learning-Based Transmission Scheme Using Reduced Feedback for D2D Networks," *IEEE Access*, vol. 10, pp. 102 316–102 324, 2022.

[14] Z. Lu, C. Zhong, and M. C. Gursoy, "Dynamic Channel Access and Power Control in Wireless Interference Networks via Multi-Agent Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 1588–1601, 2022.

[15] J. Wang, R. A. Rouil, and F. J. Cintron, "Distributed Resource Allocation Schemes for Out-of-Coverage D2D Communications," in *Proceedings of the IEEE Global Communications Conference*, 2019, pp. 1–7.

[16] S. Feng and H.-A. Choi, "Compatible QPP Scheduling for LTE D2D in Out-of-Coverage Public Safety Scenarios," in *Proceedings of the IEEE International Symposium on Local and Metropolitan Area Networks*, 2021, pp. 1–7.

[17] S. Feng, H.-A. Choi, D. Griffith, and R. Rouil, "On Selecting Channel Parameters for Public Safety Network Applications in LTE D2D Communications," in *Proceedings of the IEEE 17th Annual Consumer Communications and Networking Conference*, 2020, pp. 1–6.

[18] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual Multi-Agent Policy Gradients," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 2974–2982.

[19] S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-agent Reinforcement Learning," in *Proceedings of the International Conference on Machine Learning*, vol. 97, 2019, pp. 2961–2970.

[20] I. Saeed, A. C. Cullen, S. Erfani, and T. Alpcan, "Domain-Aware Multiagent Reinforcement Learning in Navigation," in *Proceedings of the International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[21] H. Yin, P. Liu, K. Liu, L. Cao, L. Zhang, Y. Gao, and X. Hei, "Ns3-Ai: Fostering Artificial Intelligence Algorithms for Networking Research," in *Proceedings of the Workshop on ns-3*, 2020, p. 57–64.

[22] R. Rouil, F. J. Cintrón, A. Ben Mosbah, and S. Gamboa, "Implementation and Validation of an LTE D2D Model for ns-3," in *Proceedings of the Workshop on ns-3*, 2017, pp. 55–62.

[23] "Study on LTE Device to Device Proximity Services (Release 12)," 3GPP, Tech. Rep. TR 36.843 V12.0.1, 2014-03.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[25] M. L. Littman, "Markov Games as a Framework for Multi-agent Reinforcement Learning," in *Proceedings of the International Conference of Machine Learning*, 1994, pp. 157–163.

[26] T. M. Inc., "Global Optimization Toolbox version: 4.8 (R2022b)," Natick, Massachusetts, United States, 2023.

[27] F. Cintron, D. Griffith, C. Liu, R. Rouil, Y. Sun, J. Wang, P. Liu, C. Shen, A. B. Mosbah, and S. G. Quintiliani, "Study of 5G New Radio (NR) Support for Direct Mode Communications," *NIST, Tech. Rep.*, 2021.