IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2021, accepted July 16, 2021, date of publication August 9, 2021, date of current version August 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102230

Antenna Beamwidth Optimization in Directional Device-to-Device Communication Using Multi-Agent Deep Reinforcement Learning

NILOOFAR BAHADORI, MAHMOUD NABIL, AND ABDOLLAH HOMAIFAR, (Member, IEEE)
Department of Electrical and Computer Engineering, North Carolina A&T State University, Greensboro, NC 27411, USA

Corresponding author: Abdollah Homaifar (homaifar@ncat.edu)

The authors would like to acknowledge the support from the Air Force Research Laboratory (AFRL) and the Office of undersecretary of Defense (OSD) for sponsoring this research under agreement number FA8750-15-2-0116.

ABSTRACT Exploiting the millimeter wave (mmWave) band is an attractive solution to accommodate the bandwidth-intensive applications in device-to-device (D2D) communications. The directional nature of communications at mmWave frequencies and mobility of devices require beam alignment at both transmitter and receiver ends. The beam alignment signaling overhead leads to a loss in the network's throughput. There exists a trade-off between antenna beamwidth and the achievable throughput. Although a narrower antenna beam increases the directivity gain, it leads to a higher signaling overhead and less stable D2D links which reduce the network's throughput. Therefore, optimizing the antenna beamwidth is crucial to maintain the D2D users' quality-of-experience (QoE). In this paper, we propose a novel distributed antenna beamwidth optimization algorithm based on multi-agent deep reinforcement learning. We model D2D links as agents that interact with the communication environment concurrently and learn to refine their antenna beamwidth policies. Agents aim to maximize the network sum-throughput and maintain reliable communication links while taking into account the application-specific quality-of-service (QoS) requirements and the cost associated with beam alignment. Online deployment of the proposed algorithm is distributed and does not require any coordination among users. The performance of the proposed antenna beamwidth optimization algorithm is compared with other widely used baseline algorithms. Numerical results show that our proposed algorithm improves the network performance significantly and outperforms existing approaches.

INDEX TERMS Device-to-device, mmWave, antenna beamwidth optimization, multi-agent, deep reinforcement learning.

I. INTRODUCTION

Device-to-device (D2D) communication allows user equipments (UEs) to communicate over direct links rather than traversing the cellular infrastructure. D2D communication is envisioned to improve the network's performance by offloading the cellular network and providing ubiquitous coverage for commercial, public safety and critical communication applications [1]. However, implementation of D2D communication is limited mainly due to spectrum scarcity in the sub-6 GHz band. Exploiting the abundant unlicensed spectrum in the millimeter-wave (mmWave) band for D2D communications is seen as an attractive solution to addressing the spectrum scarcity bottleneck [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

Radio propagation at mmWave band encounters several obstacles such as severe path-loss and sensitivity to blockages [3]. The small wavelength of mmWave signals, however, facilitates the implementation of large directional and high-gain antenna arrays on D2D devices, which helps to compensate for additional path-loss [3]. This, in turn, introduces a new challenge to D2D communications. Achieving the maximum directivity gain in a highly directional mmWave band system requires the transmitters and receivers to be aligned. Beam alignment incurs significant signaling overhead to the system, which reduces the network's throughput significantly. There exists a trade-off between antenna beamwidth and achievable throughput [4]. Selecting a narrower antenna beam, although leads to higher antenna gain, incurs longer beam alignment overhead and reduces the link stability time. Therefore, one needs to optimize the

antenna beamwidth prior to data transmission to maximize the network's sum-throughput and maintain users' quality-of-experience (QoE). Modeling the antenna beamwidth optimization problem and finding a systematic algorithm to reach the optimal solution is even more challenging in the D2D environment as the mobility of the devices and the diverse quality of service (QoS) requirements, make network topology highly dynamic. In this paper, we focus on the antenna beamwidth optimization problem in a mmWave D2D network where D2D UEs optimize their antenna beamwidth based on their context information to maximize the network sum-throughput and to maintain reliable D2D links.

Despite the recent advances in antenna beamwidth tuning technologies [5] and the significant impact of antenna beamwidth optimization on network performance [6], it is a fairly unexplored research area. There exist few studies in the literature that discuss antenna beamwidth optimization [6]–[10]. Nevertheless, most of the existing works suffer from several limitations that make them inappropriate for D2D communications. For example, works in [6]–[11] are centralized and computationally expensive, [6]–[8], [10] increase the communication overhead significantly and [9], [10] are not robust against changes in the network topology, thus, cannot be applied to the D2D communication framework.

Compared with those methods, deep reinforcement learning (DRL) based algorithms present the most promising mechanisms to tackle complex optimization problems in communication networks. DRL enables an agent to make complex on-line decisions in dynamic and uncertain environments, given only sequences of observations and rewards without increasing the overall system overhead. However, existing DRL-based techniques such as [12], [13] implement an independent Q-learning (IQL) approach [14] through which each agent learns a policy based on its actions and observations and treats other agents as a part of the environment. Nevertheless, multi-agent environments are non-stationary since agents are learning and updating their policies concurrently [15]. Therefore, implementing IQL is not suitable for addressing multi-agent domains as it causes an agent's locally optimal action to become a globally non-optimal joint action [16]. In addition, non-stationarity introduced by IQL also inhibits exploiting *experience replay*, which is crucial in speeding up and stabilizing the DRL training process. *Therefore, addressing the antenna beamwidth optimization problem using a distributed and low-overhead DRL-based algorithm while considering the non-stationarity of the multi-agent environment is a paramount problem that is lacking in the literature.*

In this paper, our goal is to maximize the network sum-throughput through optimizing D2D UEs' antenna beamwidth. The interaction among D2D UEs along with the mobility of UEs and various D2D applications' QoS requirements, make antenna beamwidth optimization a challenging problem. Since ignoring user interactions and non-stationarity of the environment leads to non-optimal solutions, we model the beamwidth optimization problem as a multi-agent problem and exploit the recent progress of multi-agent DRL to develop a distributed antenna beamwidth optimization algorithm. The proposed algorithm enables D2D UEs to maximize the network's sum-throughput while maintaining reliable communication links to support various D2D commercial and public safety applications. The proposed algorithm considers mmWave propagation characteristics, directional communication, D2D users' mobility, payload size, QoS requirements, and achievable throughput versus antenna beamwidth trade-off, simultaneously. The main contributions are summarized as follows:

- We proposed a multi-agent DRL-based antenna beamwidth optimization algorithm to maximize the network sum-throughput. In addition, the D2D UEs joint antenna beamwidth policy maintains the reliability of the D2D links by assuring that D2D links transmit their payload successfully in the required time budget. The proposed algorithm has two phases: training and decentralized deployment. The training phase is performed offline, under different network topologies using a shared reward function. The training algorithm enables distributed agents to optimize their antenna beamwidth during the online implementation without requiring any inter-agent communications.
- We modeled the antenna beamwidth optimization problem as a cooperative multi-agent DRL; since implementing IQL does not guarantee convergence to an efficient joint solution. Fingerprint-based learning [17] is implemented to enable agents to track their fellow agents' policies and reach the optimal joint action solution. Using fingerprint learning facilitates experience replay which expedites and stabilizes the training phase of the multi-agent DRL-based antenna beamwidth optimization algorithm.
- The performance of the proposed algorithm is compared with the existing methods such as IQL [12], [13], random antenna beamwidth selection [18], and constant antenna beamwidth selection [19]. The simulation results show that our proposed algorithm outperforms other existing methods and improves the network sum-throughput and D2D links' reliability significantly.

The remainder of this paper is organized as follows. Section II reviews the relevant related work. The system model and assumptions along with network sum-throughput maximization problem formulation are described in Section III. A novel distributed antenna beamwidth optimization algorithm based on multi-agent DRL for solving the network sum-throughput maximization problem is proposed in Section IV. Simulation results are presented in Section V and finally, conclusions and the future work directions are discussed in Section VI.

II. RELATED WORK

Directional transmissions are used in the mmWave band to compensate for the high path-loss [3]. Therefore, beam alignment must be implemented at the transmitter and receiver

TABLE 1. Comparison with relevant schemes.

	Proposed alg.	[6]	[7]	[8]	[9]	[11]	[10]	[12]	[13]
Distributed deployment	✓	x	x	x	x	x	x	✓	✓
Low overhead	✓	x	x	x	✓	x	✓	✓	✓
Robust against network changes	✓	✓	✓	✓	x	x	✓	✓	✓
Facilitate experience replay	✓	x	x	x	x	x	x	x	x
Joint optimal solution	✓	x	x	x	x	x	x	x	x

ends in order to establish high-throughput physical links. Beam alignment between transceivers requires sending and receiving multiple pilot signals [4], which reduces the D2D links' throughput as D2D transceivers cannot transmit data during the beam alignment phase. Although reducing antenna beamwidth increases the directivity gain, it requires longer beam alignment overhead and is more prone to misalignment. Therefore, it is necessary to optimize antenna beamwidth according to D2D UEs context information. Despite its importance, antenna beamwidth optimization has yet to be explored properly. Existing antenna beamwidth optimization techniques can be categorized into the following groups: particle swarm optimization [6]–[8], dynamic programming [9], non-linear programming [10], deep learning [11], and DRL-based methods [12], [13].

The particle swarm optimization (PSO) algorithm is used in [6], [7] for improving system throughput of a vehicular communication network and a relaying small-cell network. In addition, the PSO algorithm is used in [8] for interference management in the D2D network by proposing a device association and beamwidth selection. Beam management is performed in [9] with the goal of maximizing network throughput using backward dynamic programming. A framework is proposed in [10] to simultaneously control the transmission power and the beam-level beamwidths of indoor mmWave transceivers to maximize the energy efficiency of the network using non-linear programming. In [11], a deep learning-based beam management and interference coordination (BM-IC) is proposed to maximize the sum-rate of a dense mmWave network. These techniques cannot be applied to mmWave band D2D networks since they suffer from several limitations. First, existing techniques such as [6]–[11] are centralized and computationally expensive as they require an online central controller to optimize the antenna beamwidth, thus, cannot be applied to the D2D communication framework. Moreover, most of the existing techniques such as [6]–[8], [10] require coordination and information exchange among network entities, which increases the communication overhead significantly and makes these approaches not scalable. Furthermore, existing techniques such as [9], [10] are not robust against changes in the network topology, where the dynamicity of the network entities can negatively impact the system performance.

Recently, reinforcement learning (RL) is shown to be a useful tool to tackle several complex optimization problems in communication networks [20] such as dynamic spectrum access [21] and resource allocation [22]. However, the learning process in RL is time-consuming. DRL takes advantage of multi-layer neural networks to expedite the learning process, thereby improving the learning speed and the performance of RL algorithms. A DRL-based approach is proposed in [12] that simultaneously optimizes beamwidth and transmit power of transceivers in the network. A self-tuning sectorization algorithm is proposed in [13] that optimizes base station MIMO broadcast beams for each cell. The authors in [23] addressed the problem of optimizing relay selection and antenna power allocation using a centralized hierarchical DRL algorithm. However, these works implement an IQL approach through which each agent learns a policy based on its actions and observations and treats other agents as a part of the environment. Using IQL causes an agent's locally optimal actions to become a globally non-optimal joint action [16]. Non-stationarity introduce by IQL also inhibit exploiting *experience replay*, which is crucial in speeding up the DRL training process.

Therefore, addressing the antenna beamwidth optimization problem using a decentralized and low-overhead DRL algorithm that considers user interactions and environment non-stationarity is lacking in the literature. To address these gaps, we propose a novel distributed antenna beamwidth optimization algorithm based on multi-agent DRL. Unlike [6]–[11], the proposed algorithm is decentralized, low-overhead, and robust to changes in the network topology. In addition, our proposed algorithm implements fingerprint learning to consider the interaction among users and the non-stationarity of the environment. Therefore, the proposed algorithm reaches an efficient joint solution unlike [12], [13]. Moreover, experience replay is facilitated through fingerprint learning to expedite the learning process significantly. Our goal is to maximize the D2D network's sum-throughput and maintain reliable communication links while taking into account the application-specific QoS and the cost associated with beam alignment. Table 1 compares the proposed antenna beamwidth optimization algorithm with other relevant schemes.

III. SYSTEM MODEL AND PROBLEM FORMULATION

This section describes the system model for mmWave D2D network and introduces the main elements that impact the antenna beamwidth policies. In addition, we formulate the network sum-throughput optimization and the D2D link reliability problem.

TABLE 2. Summary of notations.

Symbol	Description	Symbol	Description
$\mathcal{L}, \mathcal{L}'$	Set of D2D links and reliable D2D links.	C, α	Path-loss intercept and exponent.
θ, φ	Antenna angle, antenna beamwidth.	G, g	Antenna Main-lobe and side-lobe.
T_l^A, T_P	Beam alignment time, pilot transmission time.	V_l, μ_l	Relative speed and angle of DT m, DR n.
$\Delta\mu_l, \alpha$	Misalignment angle and misalignment threshold.	T_l^S, ψ_l	Link stability time, sector-level beamwidth.
λ_l	Beam management parameter.	η	Impact of beam alignment overhead.
$\Delta\tau$	Time slot duration.	$\Delta\tau'$	Effective time slot duration.
G_l^r, G_l^t	D2D receiver and transmitter antenna gain.	$t_l,$	Data throughput.
β	Blockage parameter.	$q_l(k)$	The remaining payload at beginning of time slot k.
$\delta_l(k)$	Transmitted payload during time slot k.	T_l, B_l	Time budget and payload.
Φ	Joint antenna beamwidth policy.	Γ_l	D2D link reliability.
$N_{\max}$	Number of time slots in a episode.	I_l	Received interference on link l.
$T_l^S(k), T_l(k)$	The remaining link stability time and time budget.	$r_s(k)$	Shared reward function.
e, ϵ	Episode number and exploration rate.	Γ	Network's link reliability1.

A. NETWORK TOPOLOGY

We consider a network of mobile UEs that communicate through D2D links established at the mmWave frequency band operating under time division duplexing (TDD). A co-channel deployment with bandwidth W, uniform transmit power, and half-duplex mode are assumed. Let $\mathcal{L} = \{1, \ldots, L\}$ denotes the set of D2D links in the network where each D2D link comprises a D2D transmitter and a D2D receiver. In this scenario, D2D links are already established using peer association mechanisms such as [24], [25]. Also, all D2D transmitters have a payload in their buffer B_l that must be transmitted in a limited time budget T_l according to their application's QoS requirement. D2D users move at variable speeds and directions.

B. D2D CHANNEL MODELING

To model the mmWave channel, the distance-dependent path-loss model for peer-to-peer communication proposed in [26] is adopted. Under this model the path-loss is defined as $PL(d_l) = Cd_l^{-\alpha}$, where C denotes the path-loss intercept, α is the path-loss exponent, and d_l represents D2D link length of a given D2D link $l \in \mathcal{L}$. Each communication link experiences i.i.d small-scale Nakagami fading with parameter N_h. Hence, the received signal power can be modeled as gamma random variable with parameter, $h_l \sim \Gamma(N_h, 1/N_h)$.

C. ANTENNA PATTERN

We assume that each D2D UE is equipped with a directional antenna and is enabled to rotate its antenna bore-sight toward the desired direction with a simple rotation around its location. Each D2D transceiver can pick a beamwidth from the set of its available beamwidths, Φ_l. Without loss of generality, we assume that D2D transceivers on a given link l adopt the same antenna beamwidth. This case can be extended to the case that D2D users implement different strategies.

The directional antenna pattern is modeled using the Gaussian antenna model as

$$G(\theta) = \begin{cases} G_m e^{-\rho\theta^2}, & |\theta| \le \varphi, \\ G_s, & \text{otherwise}, \end{cases} \tag{1}$$

where $\rho = \frac{2.028\ln(10)}{\varphi^2}$ and 2φ is the antenna half-power beamwidth. θ denotes the antenna angle relative to the antenna's bore-sight direction. $G_m = \frac{\pi 10^{2.028}}{42.64\varphi+\pi}$ and $G_s = 10^{-2.028}G_m$ are the maximum main-lobe gain and the side-lobe gain, respectively [27].

D. BEAM ALIGNMENT OVERHEAD

Achieving the maximum antenna gain in a highly directional mmWave band system requires the transceivers to be precisely aligned by finding the best transmit and receive antenna directions. Beam alignment between transceivers requires sending and receiving multiple pilot signals. In this work, the hierarchical beam alignment method is considered, where first the best wide-beam pair is found through an exhaustive search, and then the search is refined using a narrower beam level within the subspace of the best wide-beam pair [28]. Assuming the antenna wide-beam pairs are already aligned, the narrow-beam alignment time [4] can be written as

$$T_l^A = \left\lceil \frac{\psi_l}{\varphi_l} \right\rceil^2 T_P, \tag{2}$$

in which ψ_l and φ_l denote the wide- and narrow- level beamwidth of D2D transceivers on link l. T_P represents the pilot signal transmission time.

E. LINK STABILITY TIME

A D2D link is stable and appropriate for data transmission as long as its D2D transmitter and receiver antennas stay aligned. Misalignment in directional communication, due to the users' mobility, occurs when the received power cause drops less than a certain ratio, denoted by $\alpha \in [0, 1]$.

Consider D2D link l that its receiver and transmitter are located at point A and B, respectively, as shown in Figure 1. Assume that the transceivers antenna beams are aligned and the antenna main-lobe direction is fixed. Also, the receiver is moving with relative velocity V_l in the direction of the relative angle of μ_l (with respect to its antenna bore-sight direction). Since the bore-sight angle of D2D transceivers is fixed, the movement will cause beam misalignment. The pointing error of the D2D receiver toward its transmitter,

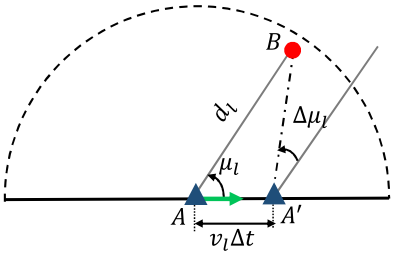

FIGURE 1. The blue triangle represents D2D receiver located at point *A* and red circle represents D2D transmitter located at point *B*. The green arrow shows the relative trajectory of D2D receiver.

Δt seconds later, denoted by $\Delta\mu_l$, can be obtained using the law of sines in triangle ABA' as

$$\frac{\sin(\Delta\mu_l)}{V_l\Delta t} = \frac{\sin(\mu_l)}{d_l}.$$

where d_l denotes the D2D links distance. Note that although receiver movement changes the distance d_l, the impact of distance difference is neglected and only the impact of movement on the angular difference is considered. Also, we assume that $V_l\Delta t \ll d_l$. For small $\Delta\mu_l$, we estimate $\sin(\Delta\mu_l) \simeq \Delta\mu_l$, therefore,

$$\Delta\mu_l \simeq \frac{V_l\Delta t\sin(\mu_l)}{d_l}. \tag{3}$$

Based on the definition, the link is stable if the relative antenna gain at the receiver is above a certain ratio, $\alpha \in [0, 1]$.

$$\alpha = \frac{G(\theta = \Delta\mu_l)}{G(\theta = 0)} = e^{-\rho\Delta\mu_l^2}. \tag{4}$$

Using (3) and (4) the link stability time, denoted by $T^S_{m,n}$, can be written as

$$T^S_l = \frac{d_l\varphi_l}{V_l\sin(\mu_l)}\sqrt{\frac{\ln(\frac{1}{\alpha})}{2.028\ln(10)}}. \tag{5}$$

It can be seen that higher antenna beamwidth and lower gain threshold increase the link stability time. Moreover, lower relative speed guarantees D2D links to be stable for longer.

F. PROBLEM FORMULATION

The beam misalignment of D2D transceivers caused by the relative movements of UEs, or availability of payload with different QoS requirements entail D2D UEs to perform beam management including beam alignment and antenna beamwidth optimization to maintain or improve their QoE. We consider the time-slotted communication framework with a slot duration of $\Delta\tau$, as shown in Figure 2. D2D UEs are allowed to perform beam management at the beginning of each time slot. Beam management is triggered upon antenna misalignment, availability of new payload in the D2D transmitter's queue, or change in the QoS requirements. Beam alignment leads to a loss in D2D links' throughput due to the time consumed to align transceivers' antenna beam, as explained in Section III-D. In other words, there exists a trade-off between antenna beamwidth and the D2D links' throughput. Selecting a narrower antenna beam leads to higher antenna gain based on (1), but it incurs higher beam alignment overhead as per (2). Consequently, the narrower antenna beam reduces the data transmission time and D2D links' throughput. Also, narrow antenna beams are less stable and more prone to misalignment according to (5). Therefore, to maintain the QoE (D2D link reliability) and increase the network's sum-throughput, D2D transceivers are required to optimize their antenna beamwidth according to the network conditions and context information.

The throughput on a given D2D link l with bandwidth W during time slot k can be defined as

$$t_l(k) = (1 - \lambda_l(k)\eta)W\log_2(1 + \mathrm{SINR}_l), \tag{6}$$

where $\lambda_l(k)$ is the beam alignment parameter, $\lambda_l(k) = 1$ indicates that beam alignment is performed at time slot k and $\lambda_l(k) = 0$, otherwise.[1] $\eta = \frac{T^A_l}{\Delta\tau'}$ captures the impact of beam alignment overhead, where $\Delta\tau'$ represent the effective time slot duration for payload transmission (Figure 2). Since D2D transceivers are not allowed to transmit data on unstable D2D links, the effective time slot duration for payload transmission is defined as the minimum of link stability time and maximum allowed time slot duration, i.e., $\Delta\tau' = \min(T^S_l, \Delta\tau)$.

The achieved signal-to-noise-plus-interference-ratio (SINR) on D2D link l can be written as

$$\mathrm{SINR}_l = \frac{PG^t_l(\theta_l)h_lG^r_l(\theta_l)PL(d_l)}{\sum\limits_{\substack{j\in\mathcal{L}\\ j\neq l}} PG^t_j(\theta_j)h_jG^r_l(\theta_l)PL(d_j) + \sigma^2},$$

where P represents the transmit power, $G^r_l(\theta^r_l)$, $G^t_l(\theta_l)$ and are the D2D receiver and transmitter antenna gain on link l, respectively. The leftmost term in the denominator represents the aggregated received interference at the receiver of link l from all other D2D transmitters, and σ^2 denotes the noise power. We assumed that duration of a time slot is shorter than channel coherence time. Therefore, the channel is considered non-varying during a time slot.

Let $q_l(k+1)$ be the remaining payload of D2D transmitter on link l at the beginning of time slot $k+1$ and is defined as

$$q_l(k+1) = q_l(k) - \delta_l(k), \tag{7}$$

where $\delta_l(k) = t_l(k) * \Delta\tau'$ denotes the amount of payload that is transmitted during time slot k. QoS constraint of the D2D link l is modeled as a limited time budget T_l through which the D2D payload B_l must be transmitted. The reliability of the D2D link l, as the measure of QoE of users, is defined as the ratio of the transmitted payload during time budget $T_l = N_l\Delta\tau$ with N_l time slots and can be written as

$$\Gamma_l = \frac{1}{B_l}\sum_{k=0}^{N_l}\delta_l(k) \tag{8}$$

[1] Antenna beam alignment optimization is not the focus of this work.

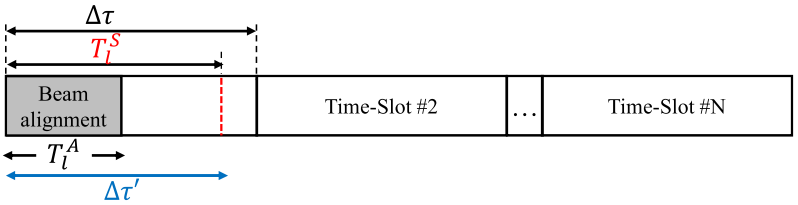

FIGURE 2. Time-slotted communication. Beam alignment can be performed in the beginning of each time-slot if necessary. T_l^S, T_l^A, $\Delta\tau'$ represent link stability time, beam alignment time, and effective time-slot duration, respectively.

The problem we are addressing in this work can be formulated as designing an antenna beamwidth selection policy such that it maximizes the network sum-throughput as

$$\underset{\Phi}{\text{Maximize}} \ \frac{1}{\Delta\tau}\sum_{l\in\mathcal{L}}\sum_{k=0}^{N_l}\delta_l(k) \tag{9a}$$

$$\text{subject to: } \varphi_l^2 \geq \psi_l^2\frac{T_P}{\Delta\tau'}, \quad \forall l\in\mathcal{L}, \tag{9b}$$

$$\varphi_l \leq \psi_l, \quad \forall l\in\mathcal{L}, \tag{9c}$$

where $\Phi = \{\varphi_1,\ldots,\varphi_l\}$ is the joint antenna beamwidth selection policy of D2D links. Constraint (9b) represents the lower band of feasible antenna beamwidth, and it holds since beam alignment time must be less than the effective time slot duration, i.e., $T_l^A \leq \Delta\tau'$. Constraint (9c) shows the antenna beamwidth upper bound.

The optimization problem (9) is difficult to solve analytically and is computationally hard due to the interaction among D2D links, especially in the D2D environment, which requires low-overhead distributed solutions. In this paper, we propose a solution based on multi-agent deep reinforcement learning to tackle this problem. The proposed framework considers the non-stationarity of the multi-agent environment and the interaction among users. Our goal is to enable D2D UEs to learn a joint antenna beamwidth optimization policy that maximizes the network sum-throughput in various network dynamics, only based on its local observation without online coordination or exchanging messages. Moreover, the reliability of the antenna beamwidth optimization policy is required to be assessed to assure that under such policy all D2D links' payloads are successfully transmitted, $\Gamma_l \geq 1, \forall l \in \mathcal{L}$.

IV. PROPOSED SOLUTION USING MULTI-AGENT DEEP REINFORCEMENT LEARNING

In this section, we describe the proposed solution to solve the network sum-throughput maximization problem through optimizing D2D UEs' antenna beamwidth in a mmWave D2D dynamic environment using cooperative multi-agent DRL. First, we explain the multi-agent DRL framework, where multiple agents interact in a common environment, take an action and try to learn a policy to maximize their shared reward. Then, we explain the details of the proposed antenna beamwidth optimization algorithm. The proposed algorithm is based on the multi-agent DRL framework and is used to solve the optimization problem (9). We define the states, actions, and rewards in the mmWave D2D multi-agent environment.

A. BACKGROUND ON MULTI-AGENT DEEP REINFORCEMENT LEARNING AND Q-LEARNING

A cooperative multi-agent DRL framework is a setting where agents concurrently interact with a shared environment and learn to coordinate together to achieve a common objective [29]. Agents interact with the environment according to partially observable (PO) Markov decision processes (MDP) (POMDP). In POMDP the system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state. An POMDP is defined as tuple $(\mathcal{L},\mathcal{S},\mathcal{U},\mathcal{T},\mathcal{R},\mathcal{Z},\mathcal{O})$, in which $\mathcal{L}$ is the set of agents, $\mathcal{S}$ denotes the state space, $\mathcal{U} = \times_l\mathcal{U}_l$ is the joint action space, $\mathcal{Z} = \times_l\mathcal{Z}_l$ is joint observation space. Each agent executes an action $u_l \in \mathcal{U}_l$ based on its policy π_l, forming a joint action $\mathbf{u}$ that make the current state $s \in \mathcal{S}$ transit to $\hat{s}$ with the probability of $\mathcal{T}(s,\mathbf{u},\hat{s})$. Agents in partially-observable environment receive observations of the latent state, denoted as z_l with the joint observation probability of $\mathcal{O}(\mathbf{z},\hat{s},\mathbf{u})$, where $\mathbf{z} = (z_1,\ldots,z_L)$. Consequently, at each time-step k, agents receive a shared reward $r(k) = \mathcal{R}(s(k),\mathbf{u}(k))$. Agents aim to maximize the expected discounted return $R(k) = \sum_{t=0}^{H}\gamma^t r(t+k+1)$ with horizon H, where $\gamma \in [0,1)$ is the discount factor, by finding the optimal policy π_l^*. Q-Learning [30] is used to find the best policy by estimating the Q-values of policies $Q_l^{\pi}(z_l,u_l) = \mathbb{E}_{\pi}[R(k)|z(k)=z_l,u(k)=u_l]$. In multi-agent DRL, agents interact with the environment without explicit knowledge of POMDP model. Due to partial observability and local non-stationarity of POMDP, learning the underlying POMDP model is complicated. Therefore, agents instead of learning functions $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{O}$, directly learn Q-values or policies.

Q-learning iteratively estimates the optimal Q-value function using backups. The optimal policy π^* maximizes the Q-value function, $Q_l^{\pi^*}(z_l,u_l) = \max_{\pi} Q_l(z_l,u_l)$. Deep Q-learning uses a deep neural network, known as deep Q network (DQN) parameterized by θ_l, $Q(z_l,u_l;\theta_l)$, to estimate Q-values. DQN relies on experience replay to accelerate and stabilize the training process. During training, actions are chosen according to ϵ-greedy policy that selects the currently estimated best action with probability $1-\epsilon$, and takes a random exploratory action with probability ϵ. Each agents' experience including current observation, action, reward and next observation as tuple $\langle z_l(k),u_l(k),r_l(k),z_l(k+1)\rangle$ is

stored in its *replay memory*. Replay memory is a first-in-first-out queue containing the set of latest experience tuples. The parameters θ_l are iteratively updated using stochastic gradient descent (SGD) by sampling batches of b experiences from the replay memory to minimize the squared temporal-difference (TD) error:

$$\mathcal{L}_l(\theta_l) = \sum_{i=1}^{b} \left[y_i^{DQN} - Q_l^{\pi}(z_l^b(k), u_l^b; \theta_l) \right]^2,$$

with a target $y_i^{DQN} = r_l^i(k) + \gamma \max_{u'} Q_l^{\pi}(z_l^b(k+1), u'; \theta_l^-)$ where θ_l^- are the parameters of a target network periodically copied from θ_l and kept constant for a number of iterations. The replay memory stabilizes learning, prevents the network from overfitting to recent experiences, and improves sample efficiency.

The widely used approach to solve multi-agent DRL is Independent Q-learning (IQL) [14], where each agent learns its DQN parameters only based on its observations and actions while treating other agents as a part of the environment. However, since all agents are learning and affecting the environment simultaneously, using IQL makes the environment non-stationary from the perspective of any individual agent. Non-stationarity and local observability of multi-agent environments cause locally optimal action to become globally non-optimal joint action [16]. In addition, the non-stationary nature of the environment makes the experience replay memory samples obsolete and negatively impacts the training performance [17].

The non-stationarity can be resolved if agents' observation state is augmented with an estimate of other agents' policies. One possible solution is augmenting each agent's observation space with its fellow agents' DQN parameters. However, this method is intractable in practice since a large number of DQN parameters complicates the learning process. Also, sharing and updating DQN parameters among agents increases the signaling overhead significantly that consequently reduces the D2D links' throughput. To overcome this problem, low-dimensional estimates (i.e., fingerprints) of other agents' policies can be added to agents' experience [17]. It is shown that augmenting the agents' experience tuple with fingerprints, including the training iteration number e and exploration rate ϵ disambiguates the age of training samples and stabilizes the replay memory significantly. Recently, this method has been used to address the non-stationarity of the environment in multi-agent wireless networks, such as spectrum sharing [31] and dynamic power allocation [32]. In this work, we use fingerprint-based learning to address the non-stationarity of the mmWave D2D environment.

B. PROPOSED FRAMEWORK

We model the D2D framework in Section III as POMDP and propose an beamwidth optimization algorithm based on the multi-agent DRL framework to enable D2D UEs to solve the optimization problem in (9a)-(9c). In this framework, the set of D2D links $\mathcal{L}$ are modeled as agents that are assigned with a common objective of maximizing the network sum-throughput. D2D links interact with the communication environment to gain experience which enables them to learn the optimal joint antenna beamwidth policy. The proposed framework has two phases, centralized training and distributed deployment. Each D2D link has a DQN that must be trained. During the training phase, each D2D link takes an action (selecting an antenna beamwidth) based on its observation and receives a shared reward from the environment which directs it toward learning the optimal policy through training its DQN. During the deployment phase, D2D links select an antenna beamwidth based on their observation using their trained DQN, without online coordination or message exchange. The schematic of the proposed framework is shown in Figure 3.

1) TRAINING PHASE

Since the optimal antenna beamwidth selection policy is unknown to the D2D links at the beginning of the training process, we consider the training process to be an episodic DRL where learning is a continuing task over a time horizon of $T = N_{\max}\Delta\tau$. D2D links' DQN are trained through running multiple episodes. At the beginning of each training episode, the environment parameters are randomly initialized including D2D UEs velocity, the direction of movement and antenna main lobe direction and beamwidth. Also, D2D transmitters are loaded with a payload that lasts until the end of the episode. Algorithm 1 presents the proposed offline antenna beamwidth selection training algorithm.

At the beginning of each time slot, if payload exists for transmission, $q(k) > 0$, D2D transmitter and receiver on each established D2D link examines the antenna alignment. If beam alignment is required (i.e., the transceivers' antennas are misaligned), D2D UEs on each link must select an antenna beamwidth and perform beam alignment, $\lambda(k) = 1$. Otherwise, D2D UEs stick to their previous policy $\lambda(k) = 0$, and the D2D transmitter continues to transmit its payload (lines 5-13).

a: ACTIONS AND oBSERVATIONS

D2D transmitter and receiver on each established D2D link select an action, i.e., antenna beamwidth $u_l \in \Phi_l$ based on their local observations and context information using an ϵ-greedy policy (lines 6 -13). The action space of each user is the set of the antenna beamwidths that can be generated by the user's antenna array, however, to satisfy (9c) antenna beamwidth must be less than the wide-level antenna beamwidth, i.e., $u_l < \psi_l$. We define the observation state of a D2D link l as

$$z_l(k) = \left\{ I_l, q_l(k), T_l(k), T_l^S(k), \lambda_l(k), e, \epsilon \right\}, \qquad (10)$$

where I_l denotes the measured interference power at the D2D receiver on link l. The interference can be accurately estimated by the receiver of each D2D link at the beginning of each time slot. We assume it is also available instantaneously at the transmitter through a delay-free feedback

Algorithm 1: Training Phase of the D2D Antenna Beamwidth Optimization Algorithm Using Multi-Agent DRL

1 **Initialization**: $k = 0$, set the D2D UEs' antenna beamwidth randomly, $\Phi(0) = \{u_1(0), u_2(0), \ldots, u_L(0)\}$.
2 **repeat**
3 **foreach** *D2D links* $l \in \mathcal{L}$ **do**
4 Receive observation $z_l(k)$.
5 **if** $T_l^S(k) = 0$ and $q_l(k) > 0$ **then**
6 $p =$ random().
7 **if** $p < \epsilon$ **then**
8 Pick a random action $u_l(k) \in \Phi_l$,
9 **else**
10 $u_l(k) = \arg\max_{u'} Q_l(z_l(k), u'; \theta_l)$
11 **end**
12 **else**
13 $u_l(k) = u_l(k-1)$.
14 **end**
15 Packet $\delta_l(k)$ and $q_l(k)$ and forward it to the central node.
16 **end**
17 The central node computes the shared reward $r_s(k)$ based on (11).
18 **if** $q_l(k) \neq 0$ **then**
19 $r_l(k) = r_s(k)$,
20 **else**
21 $r_l(k) = C$.
22 **end**
23 Receive observation $z_l(k+1)$.
24 **foreach** *D2D link* $l \in \mathcal{L}$ **do**
25 Store tuple $\langle z_l(k), u_l(k), r_l(k), z_l(k+1)\rangle$ in the replay memory.
26 Sample a batch of b samples from replay memory.
27 Calculate y_i^{DQN}.
28 Perform SGD to minimize $\mathcal{L}_l(\theta_l)$.
29 **if** $(k \ mod \ N_u) = 0$ **then**
30 Copy θ_l into θ_l^-.
31 **end**
32 **end**
33 $k = k + 1$
34 **until**
35 $k < N_{\max}$

channel. $q_l(k)$, $T_l^S(k)$ and $T_l(k)$ are the remaining D2D payload, the remaining link stability time and the remaining time budget to transfer the payload, respectively. Values of training episode e, and exploration rate ϵ are added to the observation tuple to address the non-stationarity of the environment and facilitate experience replay.

b: REWARD FUNCTION

Since a selfish action selection that solely maximizes each D2D link's throughput cannot guarantee to obtain global optimization. D2D links are trained by a shared reward to turn the environment into a fully cooperative[2] one. At the end of each time slot, the D2D transmitters evaluate the amount of the transmitted payload $\delta_l(k)$ and forward it to a central node (line 15). Then, the central node calculates the network average data throughput and broadcasts it to all agents (line 17-21). The shared reward function can be written as

$$r_s(k) = \frac{1}{|\mathcal{L}|\Delta\tau} \sum_{l \in \mathcal{L}} \delta_l(k), \tag{11}$$

where $|.|$ denotes the set cardinality. The central node can be the base station or a UE that is picked by D2D UEs. Note that, as soon as the agent delivered all of its payload, its reward becomes constant, C. The constant value should be big enough to ensure that the algorithm encourages reliable D2D links, i.e., $\Gamma_l > 1$. The observation tuple and the reward function parameters are selected to ensure that the optimization problem in (9) is satisfied. First, reward function (11) is in line with objective function (9) to maximize the network sum throughput. Parameters $T_l(k)$ and $T_l^S(k)$ in the observation tuple (10) ensure that the selected beamwidth provides sufficient time for successful data transfer in the required time budget as (9b). Also, monitoring $q_l(k)$ in the observation tuple and the constant reward C are designed to encourage the D2D link reliability such that $\Gamma_l \geq 1, \forall l \in \mathcal{L}$.

At the end of each time slot, experiences are stored in replay memory and the D2D links' DQN are trained using samples from their experience replay memory. In order to stabilize the learning, the parameter set of the target DQN, θ^- are duplicated from the training DQN parameter set θ every N_u episodes and is kept fixed in between [17] (lines 25-31).

2) DEPLOYMENT PHASE

During the deployment phase, at each time step, each D2D link assesses its context information including payload availability, beam alignment and received interference. Using the acquired information, the observation tuple in (10) is formed. Note that in the deployment phase values of e and ϵ are set to the values of the last training episode. The observation tuple is fed into the trained DQN. The outputs of the trained DQN network are Q-value of all possible actions (antenna beamwidth), as shown in Figure 3. Then, D2D UEs on each link select an antenna beamwidth with the maximum Q-value. Finally, all D2D links transmit their payload using their selected antenna beamwidth.

The centralized training procedure, which requires high computational capacity is performed offline under various network topologies and different initial conditions, which allows the D2D links to perform well during the decentralized execution time even with strong non-stationary conditions.

[2]By cooperative we mean that agents are aimed at maximizing a shared objective

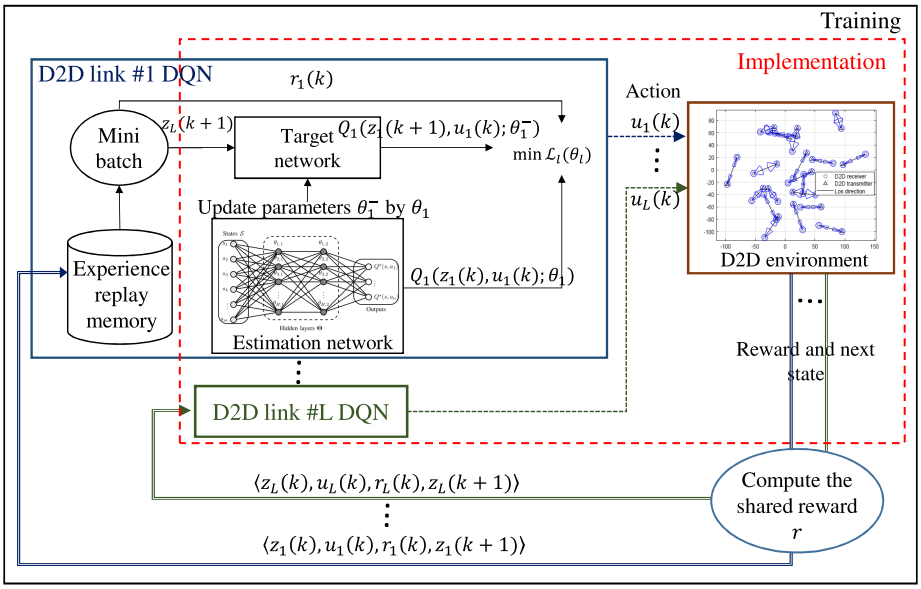

FIGURE 3. Illustration of the proposed beam management algorithm. Training is performed offline and requires a central node to compute the reward. The online implementation is distributed and dos not require any coordination.

V. NUMERICAL RESULTS

To demonstrate the effectiveness of the proposed algorithm, we compare its performance with two baseline models, i.e., *random antenna beamwidth selection* [18] and *constant antenna beamwidth* [19]. Also, to demonstrate the impact of non-stationarity of the multi-agent environment, the performance of the proposed antenna beamwidth optimization algorithm is compared with the classical IQL [12], [13], in which D2D links do not learn to cooperate and treat their fellow agents as a part of the environment.

We custom built our simulator consisting of the D2D interaction environment and the D2D links' DQN that is used to learn the antenna beamwidth selection policy. The D2D interaction environment is an area of the size 1 $km \times 1$ km, which is —given the transmit power of D2D users— large enough to avoid the boundary effect. In the simulation environment, D2D UEs are located uniformly in the simulation area. For each D2D transmitter, we assumed there exists a corresponding D2D receiver at distance d_l. Also, D2D UEs move according to the random walk model. D2D users' trajectories (speed and direction of movement) are drawn based on i.i.d. uniform random variables. D2D transceivers are equipped with a directional antenna for data transmission in the mmWave band. Also, we assume that all the D2D transmitters transmit at the same power. Simulation parameters shown in Table 3 are used, unless otherwise specified.

The DQN of each D2D link comprises three fully connected hidden layers, of 500, 250, and 120 neurons, respectively. The rectified linear unit (ReLU), $f(x) = \max(0, x)$, is used as the activation function and Adam optimizer is used to update network parameters with a learning rate of 0.001. The agents' DQN is trained offline using ϵ-greedy policy for a total of 3000 episodes. We want the D2D transceivers on each link to find the best policy fast, however, committing to a policy without sufficient exploration could also trap the D2D links in a locally optimal policy. To address the trade-off between exploration and exploitation, the exploration rate ϵ is linearly annealed from 1 to 0.02 over the first 2400 episodes and remains constant afterward. The hyper-parameters values of DQN in Table 4 are tuned through informal search. It is worth noting that in the training phase, the payload size of all D2D links is considered the same. However, the payload size and QoS requirements of D2D links vary during the deployment phase. Each episode of the training contains $N_{\max}$ time-slots with a duration of $\Delta\tau$. D2D transmitters are loaded with a payload with size B at the beginning of each episode which must be transmitted by the end of the episode.

TABLE 3. Simulation parameters:D2D environment.

Parameter	Value
Carrier frequency	28 GHz
Communication bandwidth (W)	100 MHz
Thermal noise density	-174 dBm/Hz
Transmission power	1 W (30 dBm)
Channel fading (h_l, N_h)	Gamma, 7
Free space path loss	-61.7 dB
Path loss exponent	2
Simulation area	1 km $\times$ 1 km
Antenna beamwidths (Φ_l)	[10:5:60]
D2D pairs distance (d_l)	$\sim U(30, 50)$
Pilot transmission time (T_p)	10 ms
Velocity of D2D UEs	$\sim U(2.8, 3.3)$ mph
Moving direction of D2D UEs	$\sim U(-\pi, \pi)$

TABLE 4. Simulation parameters:DQN.

Parameter	Value
Discount factor (γ)	0.95
Time slot duration ($\Delta\tau$)	0.1 s
Number of time slots in an episode ($N_{\max}$)	100
Training batch size (b)	1,500
Size of experience replay memory	100,000
Target network update frequency (N_u)	10
Training payload size (B)	25 MB

At the beginning of each training episode e, D2D UEs' velocity, the direction of movement and channel condition are set randomly. L D2D links are established between D2D transceivers in the network environment, and D2D UEs align their antenna towards their corresponding peer. At the beginning of each time slot k, D2D links' channel conditions are updated according to a Gamma random variable. Also, the location of D2D UEs is updated according to the random walk model. Based on the new location of users, beam alignment is performed if antennas are misaligned. At the beginning of each time slot, D2D UEs on each D2D link gathers their context information, including the remaining payload in the D2D transmitter's queue, the amount of interference, remaining link stability time and time budget to transmit the payload. Using the gathered information and fingerprints including current training episode number and exploration rate ϵ, D2D UEs form the observation tuple in (10). D2D UEs take an action according to the ϵ-greedy policy and receive a reward and a new observation. This information is stored in the replay memory which is used for training the DQN network. The size of the experience memory is limited. During the training, older samples will be replaced by new samples gradually.

To verify that the D2D UEs' joint policy maximizes the network sum-throughput while maintaining reliable D2D links, we evaluate the performance of the proposed antenna beamwidth optimization algorithm and the related baseline models in terms of D2D link reliability and throughput. Using (8), the network's link reliability, denoted by Γ_l, is defined as the ratio of D2D links that transmit their payload successfully in the limited time-budget (specified by the QoS requirement). The network's link reliability can be written as

$$\Gamma = \frac{|\mathcal{L}'|}{|\mathcal{L}|},$$

where $\mathcal{L}' = \{l \in \mathcal{L} \big| \Gamma_l \geq 1\}$.

Figure 4 compares the performance of the proposed algorithm in terms of network link reliability with the existing approaches. The results are averaged over 200 runs of Monte-Carlo simulations to thwart the effect of noisy results. It can be seen that increasing the payload size decreases the D2D links' reliability as fewer D2D links can successfully transmit their payload. However, our proposed antenna beamwidth optimization algorithm maintains the D2D link

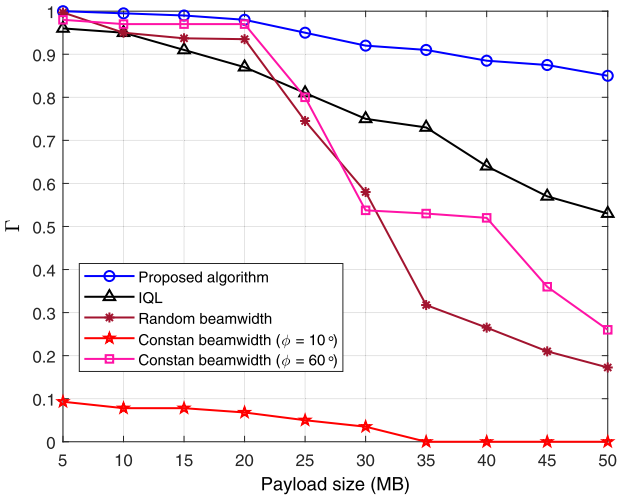

FIGURE 4. Performance comparison in terms of the D2D network link reliability vs. D2D payload size.

reliability at an acceptable level, i.e., more than 90% of D2D links transmit their payload successfully when the payload size is less than 35 MB. Also, the performance of random beamwidth selection and constant beamwidth selection deteriorates significantly as the payload size increases, since these approaches do not optimize antenna beamwidth according to the context information. It should also be noted that the IQL fails to guarantee D2D links reliability, which justifies the importance of considering the non-stationary of a multi-agent environment and enabling D2D links to track their fellow agents' policies to reach the best joint beam management policy.

Figure 5 compares the performance of the proposed antenna optimization training algorithm with the IQL algorithm in terms of D2D link throughput and reliability. The results are shown for four D2D links during an episode of the distributed deployment. Figures 5a and 5b compares the D2D transmitters' queue status during 100 time slots of a deployment scenario. It can be seen that the proposed algorithm enables all D2D UEs to maintain reliable links by transmitting their whole payload in the required time budget successfully. While using the IQL does not guarantee D2D links' reliability, since none of the D2D UEs could transmit their payload in the required time budget.

Figures 5c and 5d show the changes in the D2D links' throughput while transmitting their payload during the same deployment scenario. These figures illustrate that by implementing the proposed algorithm, D2D links learned to cooperate rather than acting selfishly as in IQL. It can be seen in Figure 5c that D2D links take turns to send their payloads according to their observation tuple. In this example, using the proposed algorithm, D2D link 2, at the beginning achieves a high throughput to transmit its payload while other D2D links keep their transmissions low to avoid causing interference and deteriorating D2D link 2's throughput. It can be seen that throughput of D2D link 2 goes to zero upon finishing transmitting its payload at time slot 45

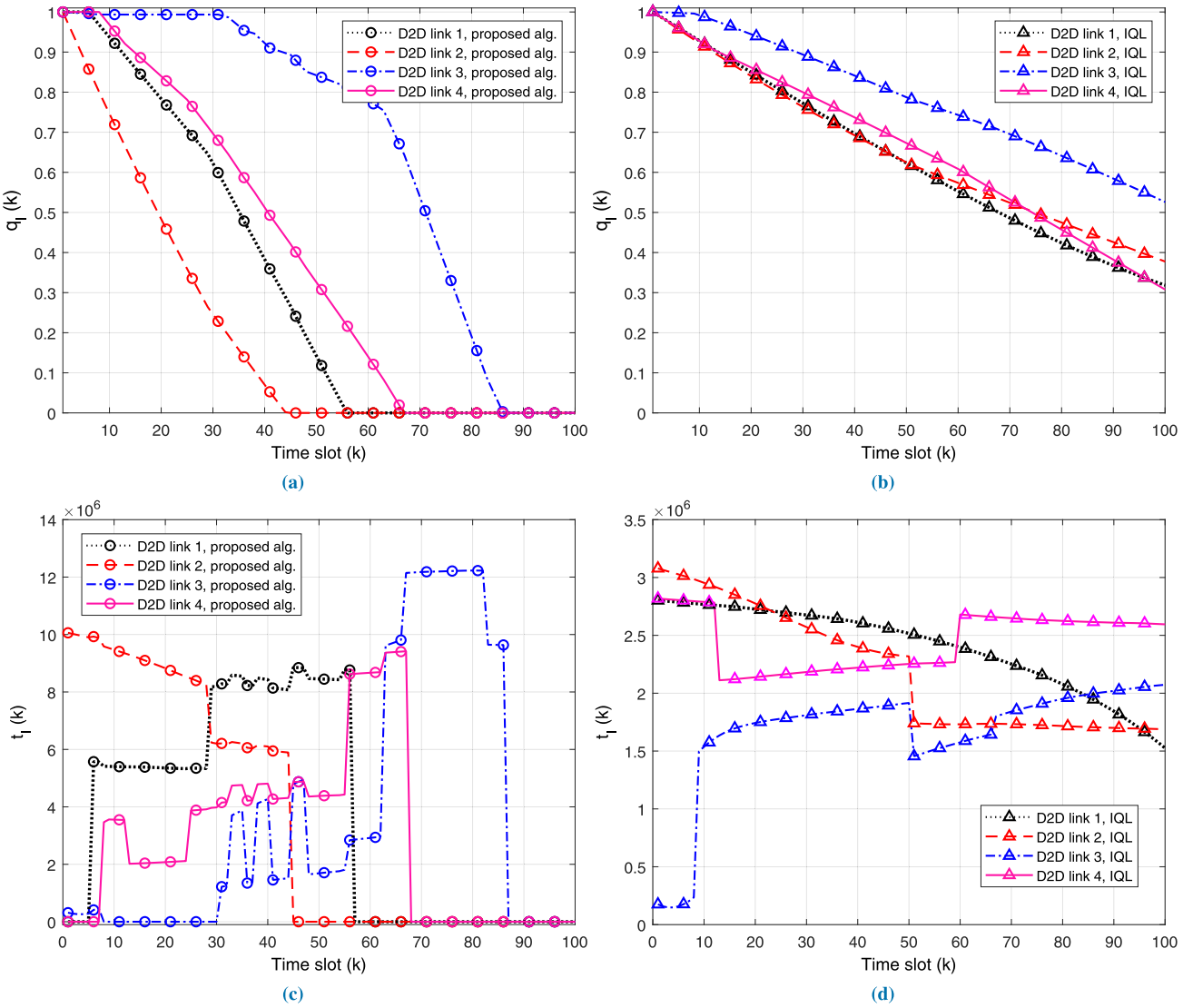

FIGURE 5. Reliability of the D2D links using the proposed multi-agent DRL-based algorithm is compared with the single-agent algorithm, where the non-stationarity of the environment is neglected. Four D2D links status are shown during 100 time slots within an implementation episode: (a) remaining D2D links' payload using the proposed algorithm, (b) remaining D2D links' payload using IQL, c) D2D links' throughput using the proposed algorithm, and (d) D2D links' throughput using the single-agent algorithm.

(also shown in Figure 5a). Afterward, in the same manner, D2D links 1, 4 and 3 take turns to transmit their payload, while others keep their activities at a minimum. Meanwhile, Figure 5d shows that using the IQL algorithm and ignoring the non-stationarity of the environment results in competition among the D2D links to increase their individual throughput during each time slot. The interference among D2D links leads to throughput reduction and failure in payload transmission. Compared to IQL, our proposed algorithm manages to keep the network sum-throughput very high. Also, the proposed algorithm provides very high throughput (about 10 Mbps) to each D2D link. In contrast, the throughput of D2D links using IQL is relatively low (about 3 Mbps). These results confirm that IQL is not a suitable approach for non-stationary multi-agent environments. Since using IQL D2D UEs merely take actions based on their own observations while other users are treated as a part of the environment.

Figure 6 compares the normalized reward received by D2D links using the proposed algorithm and IQL. The graph shows the reward of a given D2D link. The results are shown for 3000 training episodes. Note that, in IQL agents are not trained using a shared reward function. The growing trend of reward function using our proposed algorithm indicates its efficiency in enabling agents to cooperate and increase the reward function. While using IQL, the D2D link fails to refine its policy to improve its reward function. Also, the relatively stable and tight convergence of the proposed algorithm's reward function highlights its ability to find an effective joint policy. At the same time, IQL's massive fluctuation shows that the D2D link cannot converge to a good policy. Moreover, this

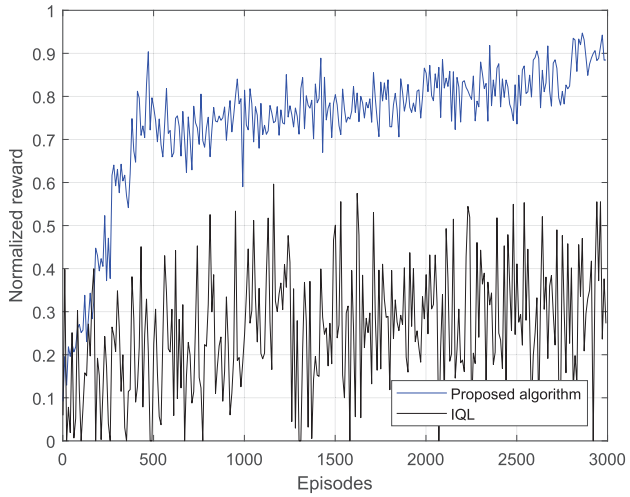

FIGURE 6. Normalized reward of D2D links during 3000 training episodes.

graph is a reliable measure to verify the sufficiency of the number of training episodes. The fast and tight convergence of the proposed algorithm indicates that 3000 episodes are sufficient to train the D2D links.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a novel multi-agent DRL-based algorithm to optimize D2D UEs' antenna beamwidth in a directional D2D network in the mmWave band. The proposed algorithm considers D2D UEs' mobility, payload size, QoS requirements, beam alignment cost and non-stationarity of the multi-agent environment. The proposed algorithm enables D2D links to learn an optimized antenna beamwidth selection policy to increase the network sum-throughput while maintaining the D2D link reliability. D2D links are trained offline using a shared reward function while the deployment of the proposed algorithm is distributed and does not require any online coordination. The training algorithm is based on the multi-agent DRL, and the non-stationarity of the environment is addressed by augmenting users' observation with a low dimensional fingerprint. Finally, the performance of the proposed antenna beamwidth optimization algorithm is evaluated through extensive simulations. Also, a performance comparison is performed with existing approaches, such as IQL and random beamwidth selection. Results show that our proposed algorithm improves network performance significantly and outperforms other approaches.

In the future, we plan to investigate adapting the proposed algorithm in indoor applications. In such environments, D2D users are more prone to perform beam alignment due to shorter beam stability time. The proposed beamwidth selection algorithm in this work will manage to compensate for the stability time by selecting the proper antenna beamwidth and render higher performance gains.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from Air Force Research Laboratory (AFRL) and the Office of the Secretary of Defense (OSD) for sponsoring this research under agreement number FA8750-15-2-0116. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL, OSD, or the U.S. Government.

REFERENCES

[1] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, Nov. 2014.

[2] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[3] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[4] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 1292–1297.

[5] Z. Wei, D. W. K. Ng, and J. Yuan, "NOMA for hybrid mmWave communication systems with beamwidth control," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 567–583, Jun. 2019.

[6] C. Perfecto, J. Del Ser, M. I. Ashraf, M. N. Bilbao, and M. Bennis, "Beamwidth optimization in millimeter wave small cell networks with relay nodes: A swarm intelligence approach," in *Proc. 22th Eur. Wireless Conf.*, 2016, pp. 1–6.

[7] C. Perfecto, J. Del Ser, and M. Bennis, "Millimeter-wave V2V communications: Distributed association and beam alignment," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2148–2162, Jun. 2017.

[8] Z. Zhang, C. Wang, H. Yu, M. Wang, and S. Sun, "Power optimization assisted interference management for D2D communications in mmWave networks," *IEEE Access*, vol. 6, pp. 50674–50682, 2018.

[9] S. Shahsavari, M. A. A. Khojastepour, and E. Erkip, "Beam training optimization in millimeter-wave systems under beamwidth, modulation and coding constraints," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–7.

[10] A. Saeed and O. Gurbuz, "Joint power and beamwidth optimization for full duplex millimeter wave indoor wireless systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[11] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, "Deep learning-based beam management and interference coordination in dense mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592–603, Jan. 2019.

[12] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, "Deep reinforcement learning for joint beamwidth and power optimization in mmWave systems," *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2201–2205, Oct. 2020.

[13] R. Shafin, H. Chen, Y.-H. Nam, S. Hur, J. Park, J. Zhang, J. H. Reed, and L. Liu, "Self-tuning sectorization: Deep reinforcement learning meets broadcast beam optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4038–4053, Jun. 2020.

[14] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 330–337.

[15] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems," *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 1–31, Feb. 2012.

[16] N. Fulda and D. Ventura, "Predicting and preventing coordination problems in cooperative q-learning systems," in *Proc. IJCAI*, 2007, pp. 780–785.

[17] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1146–1155.

[18] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[19] A. Thornburg, T. Bai, and R. W. Heath, Jr., "Performance analysis of outdoor mmWave ad hoc networks," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 4065–4079, Aug. 2016.
[20] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.
[21] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.
[22] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.
[23] H. Zhang, S. Chong, X. Zhang, and N. Lin, "A deep reinforcement learning based D2D relay selection and power level allocation in mmWave vehicular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 416–419, Mar. 2020.
[24] Y. Gao, Y. Xiao, M. Wu, M. Xiao, and J. Shao, "Dynamic social-aware peer selection for cooperative relay management with D2D communications," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3124–3139, May 2019.
[25] N. Namvar, N. Bahadori, and F. Afghah, "Context-aware D2D peer selection for load distribution in LTE networks," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2015, pp. 464–468.
[26] T. S. Rappaport, G. R. Maccartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[27] G. Yang and M. Xiao, "Performance analysis of millimeter-wave relaying: Impacts of beamwidth and self-interference," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 589–600, Feb. 2018.
[28] C. Liu, M. Li, S. V. Hanly, I. B. Collings, and P. Whiting, "Millimeter wave beam alignment: Large deviations analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1619–1631, Jul. 2017.
[29] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. AAAI/IAAI*, nos. 746–752, 1998, p. 2.
[30] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Dept. Psychol., Univ. Cambridge, Cambridge, U.K., 1989.
[31] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.
[32] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

NILOOFAR BAHADORI received the B.Sc. degree in electrical and electronics engineering from Isfahan University, in 2011, and the M.Sc. degree (Hons.) in electrical and radio frequency (RF) engineering from Semnan University, in 2013. She is currently pursuing the Ph.D. degree with North Carolina A&T State University, Greensboro, NC, USA. Her current research interests include device-to-device (D2D) and machine-to-machine (M2M) communication, mmWave band communication, the Internet of Things (IoT), the applications of machine learning in improving wireless networks, and game theory. She was a recipient of the 2019 IEEE Wireless Telecommunications Symposium (WTS) Best Paper Award.

MAHMOUD NABIL received the B.S. and M.S. degrees (Hons.) in computer engineering from Cairo University, Egypt, in 2012 and 2016, respectively, and the Ph.D. degree in electrical and computer engineering from Tennessee Tech University, Cookeville, TN, USA, in August 2019. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, North Carolina A&T University. He published many journals and conferences in different prestigious venues, such as the IEEE Internet of Things Journal, IEEE Transactions on Dependable and Secure Computing, IEEE Access, the International Conference on Communication (ICC), the International Conference on Pattern Recognition (ICPR), and the International Conference on Wireless Communication (WCNC). His research interests include security and privacy in smart grid, machine learning applications, vehicular ad hoc networks, and blockchain applications.

ABDOLLAH HOMAIFAR (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the State University of New York at Stony Brook, in 1979 and 1980, respectively, and the Ph.D. degree in electrical engineering from The University of Alabama, in 1987. He is currently the NASA Langley Distinguished Chair Professor and the Duke Energy Eminent Professor with the Department of Electrical and Computer Engineering, North Carolina A&T State University (NCA&TSU). He is also the Director of the Autonomous Control and Information Technology Institute, and the Testing, Evaluation, and Control of Heterogeneous Large-Scale Systems of Autonomous Vehicles (TECHLAV), NCA&TSU. Through his research, he has received funding in excess of 30 million from various U.S. funding agencies. He has written more than 350 technical publications, including book chapters and journal articles and conference papers. His current research interests include machine learning, unmanned aerial vehicles (UAVs), testing and evaluation of autonomous vehicles, optimization, and signal processing. He is a member of the IEEE Control Society, Sigma Xi, Tau Beta Pi, and Eta Kapa Nu. He also serves as an Associate Editor for the *Intelligent Automation and Soft Computing* journal. He serves as a Reviewer for the IEEE Transactions on Fuzzy Systems, Man, and Cybernetics, and *Neural Networks*.