RL-Based Waveform Adaptation With Partial Overlapping Tones in HetNets

Mrugen Deshmukh, Alphan Şahin, *Member, IEEE*, and İsmail Güvenç, *Fellow, IEEE*

***Abstract*— Partially-overlapping tones (POT) mitigate the co-channel interference in a wireless network by exploiting the space between adjacent subcarriers through intentional frequency offsets (FOs). In this letter, we use partially-overlapping tones (POT) in a two-tier heterogeneous network, where multiple small cells (SCs) interfere with a macro cell. We propose a multi-agent Q-learning-based approach to obtain transmit power and intentional FOs assigned to SCs for filtered multi-tones with various pulse shapes. We show that the proposed method reduces the total interference and in turn, increases the throughput in the network. We then compare the performance of the proposed approach to the existing schemes and demonstrate its advantage with numerical results.**

***Index Terms*— HetNets, partial overlapping, Q-learning.**

I. Introduction

A MULTI-AGENT reinforcement learning (MARL) system comprises individual entities that share and interact within the same environment [1]. By this definition, a heterogeneous network (HetNet), which consists of irregular deployments of different classes of small cells (SCs), can be considered as a multi-agent system, where multiple base stations (BSs) and user equipments (UEs) share the available resources. The density of a HetNet and the complexity of resource allocation sometimes lead to co-channel interference (CCI) scenarios. To solve this problem, resource allocation and interference mitigation schemes that use MARL are introduced in the literature. In [2], a multi-agent Q-learning algorithm that finds the optimal femtocell transmit powers in a two-tier HetNet is investigated. In [3], a dynamic resource allocation algorithm using RL is proposed for unmanned air vehicles (UAVs), where all UAVs make their decisions independently to minimize their total interference. In [4], the resource allocation problem is tackled with deep reinforcement learning (RL) in a multi-agent learning scenario. In [5], the authors formulate three different strategies based on deep Q-learning, convex optimization, and traditional Q-learning to solve the strategy and resource allocation problem in mobile edge computing (MEC) networks.

To address the CCI problem in wireless networks, partially-overlapping tones (POT), which exploit the intentional frequency shifts and the pulse shape used in a multi-carrier scheme, are proposed in [6] and [7].

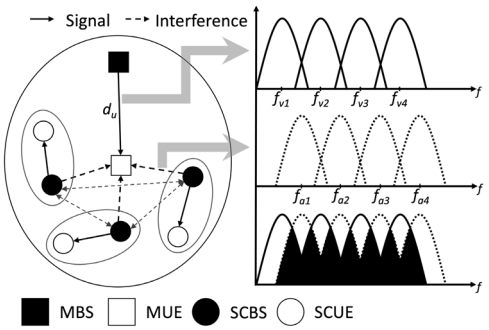


Fig. 1. Reducing CCI with POT in a HetNet scenario.

The transmitted signal in one of the links in partially-overlapping tones (POT) is shifted by an intentional frequency offset (FO) equal to a fraction of the frequency spacing between two tones (see Fig. 1). It was shown in [6] that the intentional FO can reduce the interference between the links and thus improve the throughput and error-rate performance if the orthogonality of the pulses is compromised. POT converts a hard problem (i.e., CCI), to an easier problem (i.e., self-interference) that can be solved using an equalizer. Nevertheless, utilizing POT in a large, multi-tier network is challenging as it requires sophisticated coordination.

In this letter, we formulate CCI as a function of the transmit power levels, intentional FOs, and the pulse shape used in each small cell, and introduce a new *multi-agent Q-learning framework* that aims to reduce the overall CCI in the network. To the best of our knowledge, MARL-based HetNet CCI mitigation considering simultaneously the waveform parameters and the transmit power for partial overlapping does not exist in the literature. Combining both aspects reduces the CCI significantly, which is corroborated by our link-level and system-level simulation results. By using orthogonal and other non-orthogonal schemes based on filtered multi-tone (FMT), we analyze the impact of waveform adaptation on interference. For assigning the intentional FO, we consider the effect of the pulse shape on the CCI within the HetNet. We develop two reward functions and observe their convergences for different configurations. We also compare the capacity and block-error rate (BLER) performance of the proposed algorithm with full-overlapping and a state-of-the-art algorithm in [2].

II. System Model

Consider a HetNet downlink scenario with U cells, where $U-1$ small cells (SCs) are deployed within a macro cell. The transmitted signals from the small cell base stations (SCBSs)

Manuscript received May 10, 2021; accepted June 4, 2021. Date of publication June 14, 2021; date of current version September 10, 2021. This research is supported by the National Science Foundation (NSF) CNS through the award number 1814727. The associate editor coordinating the review of this letter and approving it for publication was Z. Zhao. *(Corresponding author: Mrugen Deshmukh.)*

Mrugen Deshmukh and İsmail Güvenç are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: madeshmu@ncsu.edu; iguvenc@ncsu.edu).

Alphan Şahin is with the Electrical Engineering Department, University of South Carolina, Columbia, SC 29208 USA (e-mail: asahin@mailbox.sc.edu).

Digital Object Identifier 10.1109/LCOMM.2021.3088841

to their small cell user equipments (SCUEs), and from the macro base station (MBS) to its macro user equipment (MUE), interfere with each other. For simplicity, we assume that each MUE within the MBS is interfered by disjoint sets of SCBSs (other weak interference are neglected), and each SC consists of a single pair of SCBS and SCUE. Therefore, the rest of this letter considers the collision domain between $U-1$ SCs and a given MUE, and generalization to multiple MUEs/SCUEs is left as a future work.

We consider a large-scale fading model between any transmitter and the receiver in the uth cell for $u \in \{1, 2, \ldots, U\}$ as $P_u^{\mathrm{r}} = P_u^{\mathrm{t}} + K_u - 10\zeta_u \log_{10}\left(\frac{d_u}{d_0}\right) - \psi_u$ in dB, where $P_u^{\mathrm{r}} = G_u^2$ is the received power, G_u is the gain accounting for the channel effects and transmit power of the desired signal, P_u^{t} is the transmit power, K_u captures the attenuation and antenna characteristics in the link, ζ_u is the path loss exponent, d_0 is the reference distance for the antenna far-field, d_u is the distance between the transmitter and the receiver, and ψ_u represents log-normal shadowing [8].

We consider FMT for the transmitted signals from the SCBSs and the MBS. FMT can be thought of as simultaneous narrow-band single carrier transmissions on different frequencies, where the spacing between any two adjacent transmissions in the frequency domain is identical. Since FMT can be easily constructed as an orthogonal or a non-orthogonal waveform by using different pulse shapes, it can be effectively utilized with the concept of POT [6]. The transmitted FMT symbols can be described as $s_u(t) = \sum_{l=-\infty}^{\infty} \sum_{n=0}^{N-1} X_{ln}^u g_{ln}(t)$, where X_{ln}^u is the information symbol to be transmitted from the uth SC, l is the time index, n is the subcarrier index, N is the total number of subcarriers, and $g_{ln}(t)$ is the synthesis function [9] that maps X to the time-frequency domain in a lattice structure as $g_{ln}(t) = p(t - l\tau_0)e^{j2\pi n\nu_0 t}$, where $p(t)$ is the prototype filter being used, τ_0 is the time spacing between two consecutive symbols and ν_0 is the spacing between any two subcarriers. The received signal at the receiver of the uth cell can be calculated as

$$y_u(t) = \sum_{i=1}^{U} \int_{\tau_{i,u}} h_{i,u}(\tau_{i,u}, t) s_i(t - \tau_{i,u}) \mathrm{d}t + w(t), \quad (1)$$

where $h_{i,u}(\tau_{i,u}, t)$ is the multi-path channel between the transmitter at the ith cell and the receiver at the uth cell and $w(t)$ is the additive white Gaussian noise (AWGN).

After using a matched filter, the demodulated signal is obtained as $\tilde{X}_{mk}^u = \langle y_u(t), \gamma_{mk}(t) \rangle$, where $\gamma_{mk}(t) = \gamma(t - m\tau_0)e^{j2\pi k\nu_0 t}$ is the dual of the synthesis function at the receiver [9]. Accounting for the desired part, the CCI and noise in $\tilde{X}_{mk}^u$ as described in [6], the signal-to-interference-plus-noise ratio (SINR) of the MUE can be calculated as

$$\Gamma_{\mathrm{MUE}} = \frac{|G_u A_{mkmk}^u|^2}{I_{mk}^{(1)} + I_{mk}^{(2)} + \sigma^2}, \quad (2)$$

where

$$I_{mk}^{(1)} = \left| G_u \sum_{l=-K+1}^{K-1} \sum_{n=0}^{N-1} A_{lnmk}^u \right|^2,$$

and

$$I_{mk}^{(2)} = \left| \sum_{i \neq u} G_i \sum_{l=-K+1}^{K-1} \sum_{n=0}^{N-1} A_{lnmk}^i \right|^2,$$

in which G_i is the gain accounting for the channel effects and transmit power of the ith aggressor, X_{mk}^i is the symbol of the ith aggressor, σ^2 is the noise variance and A_{mkmk}^u and A_{nlmk}^i represent the coefficients obtained through the corresponding ambiguity functions of the desired signal and the ith aggressor, respectively, which can be calculated as:

$$A_{nlmk}^i = \int_{\tau} \int_{\nu} \int_{t} g_{ln}(t - \tau) e^{j2\pi\Delta f_i (t-\tau)} \gamma_{mk}(t) e^{j2\pi\nu t} \mathrm{d}t \mathrm{d}\tau \mathrm{d}\nu, \quad (3)$$

where Δf_i is the intentional FO given to the ith aggressor. The SINR of an SC can be calculated with a similar formulation.

From (3), it is clear that the intentional FO allocated to a user can influence the interference and in turn the SINR. Fig. 1 describes the partial overlapping concept. There are multiple SCs (aggressors) with one MUE (victim) nearby those aggressors: interference to other, far-away MUEs are assumed weak and hence neglected. Assume all users have the same transmitted powers. In the context of a HetNet, we consider the macro-cell to be a victim with $\{f_{v1}, f_{v2}, \ldots, f_{vN}\}$ as the center frequencies of its N subcarriers. A SC re-using the same frequencies is considered to be an aggressor and also has N sub-carriers with center frequencies $\{f_{a1}, f_{a2}, \ldots, f_{aN}\}$. When we use partial overlapping, one of the users' signal (aggressor's signal in this case) is shifted by a fraction of the carrier spacing between two subcarriers to reduce the interference between the two users. Thus, an intentional FO, i.e., $\Delta f_i = \beta_i(f_{a2} - f_{a1})$, is given to the ith aggressor, where β_i is a fractional value between 0 and 1. When multiple aggressors enter the network, each one of them can be assigned to an intentional FO to reduce the sum interference.

The pulse shape used in (3) determines the interference characteristics in (2). In this study, we consider a root-raised-cosine (RRC) filter with a roll-off factor α, which leads to an orthogonal waveform if the minimum normalized subcarrier spacing is $1 + \alpha$ and a Gaussian filter with a time-frequency dispersion parameter ρ, which causes a non-orthogonal waveform. For further details on filters, we refer the reader to [9].

III. Proposed Q-Learning Algorithm

The primary objective of this work is to maximize the capacity of the MUE by finding the optimal frequency offsets and transmit power levels for the interfering SCs, which can be expressed as

$$\max_{\substack{p_1', \ldots, p_U' \\ \beta_1', \ldots, \beta_U'}} \log_2(1 + \Gamma_{\mathrm{MUE}}), \text{ s.t. } p_{\forall i} \in (p_{\min}, p_{\max}), \beta_{\forall i} \in (0, 1),$$

where p_i is the transmit power of the ith interfering SC that can vary between $p_{\min}$ and $p_{\max}$ and β_i is the fraction of the carrier spacing between 0 and 1, i.e., the FO for the ith SCBS.

In a traditional single-agent Q-learning algorithm, the agent goes through numerous state-value iterations to observe long term rewards at different states for different actions and

optimizes for the action it should take in every state it reaches. In our system model, however, we have $U-1$ SCBSs that will take their actions individually. Therefore, in this paper, we formulate a multi-agent Q-learning algorithm to solve this problem of power and FO allocation to each SCBS. In a multi-agent system, each agent goes through a similar iterative process but accounts for a global (or collective) reward function that influences the actions of all the agents interacting within the environment [1]. In our case, the collective reward is formulated to optimize the capacity of the MUE and minimize the total interference in the network. We use the ϵ-greedy method for training our algorithm [10].

States: We use a set of three variables in our state, namely: 1) D_1 - a variable indicating the distance of the SC from the MBS, 2) D_2 - a variable indicating the distance of the SC from the MUE, and 3) Ω - a binary value indicating whether the SINR threshold of the MUE is satisfied or not. Mathematically, the state of each SC is described by the set $\mathcal{S}=\{D_1, D_2, \Omega\}$. Our state formulation is similar to that in [2].

Actions: A SCBS can: 1) change its transmit power level (p), and 2) change its FO (β). The action that each SCBS takes can be represented by the set $\mathcal{A}=\{p, \beta\}$.

In each time step of the simulation, each agent (or SCBS) will update its Q-value while iterating through the action set in every state for exploring future rewards. For a multi-agent Q-learning algorithm, the Q-value update is mathematically represented as [1]:

$$Q_i^{t+1}(x_i,a_i) = Q_i^t(x_i,a_i) + \alpha\Big[R_i^{t+1} + \gamma\Big(\max_{a\in\mathcal{A}} Q_i^t(x_i^{t+1},a_i) - Q_i^t(x_i,a_i)\Big)\Big], \quad (4)$$

where x_i represents the state of the ith SCBS ($x_i \in \mathcal{S}$), a_i represents the action of the ith SC, α is the learning rate, R^t is the reward calculated by the ith SCBS at time t, and γ is the discount factor that decides the trade-off between exploration of future rewards and exploitation of immediate rewards.

Reward: We consider two different rewards to evaluate our algorithm. The first reward function prioritizes the improvement in the capacity of the MUE and is given by

$$R_1^t = \lambda_{11}(C_{\text{MUE}} - C_0) + \lambda_{12}(C_i^t - C_i^{t-1}), \quad (5)$$

where λ_{11} and λ_{12} are scalar values that are used to tune the reward function, C_{MUE} is the capacity of the MUE, C_0 is the pre-defined threshold for the MUE, and C_i^t is the capacity for the ith SC at time instant t. The second reward function considers fairness of the resources being allocated to the SCBSs using Jain's fairness index [11] and is given by

$$R_2^t = \lambda_{21}(C_{\text{MUE}} - C_0) + \lambda_{22}(J_p + J_\beta), \quad (6)$$

where λ_{21} and λ_{22} are scalars to tune the reward function, J_p and J_β are Jain's fairness index values for the allocated transmit power and FOs respectively. The capacity for the uth user (including MUE and SCs) is calculated as

$$C_u = \log_2\left(1 + \frac{P_u^{\text{r}}\sigma_u^2}{\sigma_{\text{noise}}^2 + \sum_i P_i^{\text{r}}\sigma_i^2(\Delta f_i)}\right), \quad (7)$$

where P_u^{r} and P_i^{r} are the received powers for the desired user and the interfering user respectively, σ_{noise} is the variance in the noise in the channel, σ_u is the gain for the desired user obtained after calculating singular value decomposition (SVD) of all the correlated users in the interfering channel, and σ_i is the gain of the ith interferer (which is a function of the FO assigned to it). The SVD is used to decorrelate the channels for the desired as well as the interferer [6], [12].

A. System-Level Aspects

The steps to implement the proposed algorithm in a practical HetNet deployment are given as follows: 1) Each SCBS counts its state variables D1 and D2 based on its location with respect to the MBS and the MUE. Depending on whether the SINR of the MUE is above or below the threshold, the third variable Ω of the state is then decided for every SCBS. 2) Based on the state of an SCBS, it looks up the corresponding action values from the pre-trained Q-tables and chooses the trained transmit power level and FO to use. 3) If a new SCBS joins the network, the MBS has to update all the SCBSs since the Q tables are trained separately for different number of SCBSs in the network. 4) If there are any mobile SCBSs, the latency between the MBS and the SCBS should be short enough to track the state changes that may occur with the change in the location of the SCBS. In this study, we assume that SCBSs locations are fixed.

1) Latency: After the training is complete and the Q tables are generated, each SCBS would require feedback from the MBS to update its state Ω. This may cause additional latency in the network, depending on how frequently the SCBS is programmed to update its state. Since we consider stationary SCBSs and SCUEs, we do not consider latency constraints. As long as an SCBS can identify its state correctly, it does not have to concern with coherence time in a multi-path channel while assigning the transmit power level and FO to the SCUE.

2) Overhead: Any overhead caused due to the proposed algorithm lies in the communication between the MBS and all the SCBSs and storing their respective state values and Q tables. We consider all the base stations and equipment to have sufficient computing power, for this overhead to be negligible.

3) Convergence: We do not add any constraints for SCUEs to ease the burden on the RL algorithm to converge without adding too many restrictions since the action space here is fairly limited. While calculating the reward function, we do consider the improvement of SINR for SCUEs as well, so that the MARL algorithm does not always choose $p_{\min}$.

4) Fairness: To consider fairness in terms of the resources allocated to the SCs, we use Jain's fairness index in (6) that allows the algorithm to allocate resources fairly to the SCs instead of prioritizing for MUE performance. We discuss the implications of this and compare with the other reward function in the next section.

5) Complexity & Optimality: We consider a multi-agent Q-learning framework that has to find the appropriate action values from the set of possible actions $\mathcal{A}$ for all the possible states in the set $\mathcal{S}$ for U number of small cells in the network in a distributed manner. The complexity of our algorithm at each cell can be given by $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$, where $|.|$ denotes the cardinality of the set.

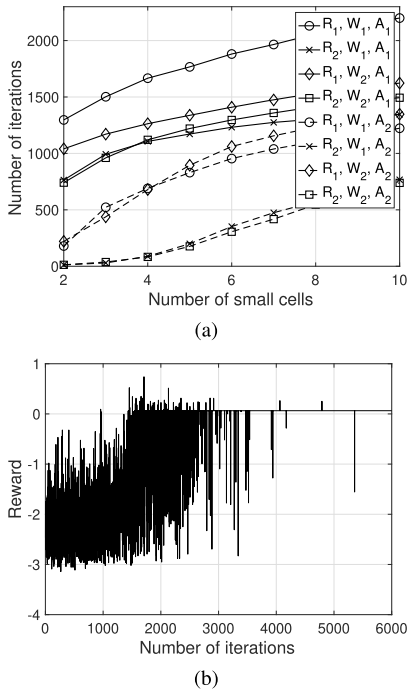


Fig. 2. (a) Convergence of the algorithm in different reward configurations, (b) Tracing the reward function value in one simulation.

Q-learning has been proven to converge to the optimal solution in [13]. The algorithm may at times converge prematurely with the ϵ-greedy approach yet it is still used in practice because it offers significant time reduction in convergence.

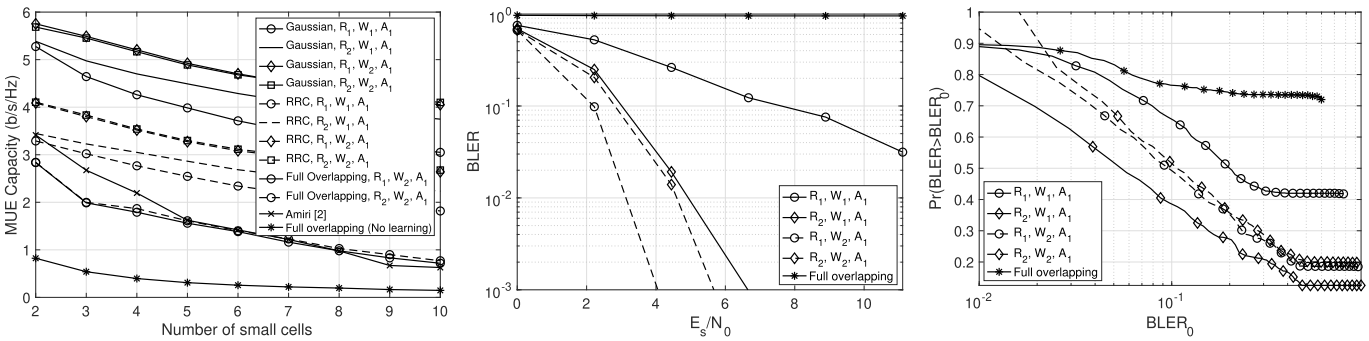
IV. Numerical Results

We assume that each MUE transmits with a physical resource block consisting of $K = 12$ symbols and $N = 12$ subcarriers. The subcarrier spacing and the carrier frequency are set to 16.67 kHz and 2 GHz, respectively. To simulate path loss between the MBS and the MUE, and between the MUE and the SCUE, we use the urban dual strip path loss models from [14]. The transmit power levels for the SCs vary from $p_{\text{min}} = 5$ dBm to $p_{\text{max}} = 15$ dBm in 10 steps. The fractional FO values are swept from 0 to 1 with increments of 0.1, i.e., $\beta \in \{0, 0.1, 0.2, \ldots, 0.9\}$. For the ϵ-greedy method, we start training with $\epsilon = 1$ and decay it with time as $\epsilon = \epsilon \times e^{-10^{-6}t}$. For the deployment, we consider a dense urban scenario with a dual strip apartment block, similar to the scenario being considered in [2]. In our simulations, the MUE is located approximately 140 m away from the MBS. The MUE is in a corridor inside a building. The dimensions of the corridor are assumed to be 10 m $\times$ 50 m. Adjacent to the corridor on both sides are total ten rooms of dimensions 10 m $\times$ 10 m. Each room has one SCBS randomly placed within it and interfering at the MUE. In all our simulations, the signal-to-noise ratio (SNR) between the MUE and the MBS is maintained to be 20 dB, whereas the SNR for the SC changes according to its allocated transmit power.

As a measure of the time it takes for our proposed algorithm to converge, we provide the number of iterations in Fig. 2a. Our simulation runs until the reward function starts converging to a constant value, for a given number of SCBSs in the network. We consider three different configurations of the scalar weights $W_1 = \{1, 1\}$, $W_2 = \{10, 1\}$ and $W_3 = \{1, 10\}$ for $\{\lambda_{11}, \lambda_{12}\}$ in R_1 in (5) and for $\{\lambda_{21}, \lambda_{22}\}$ in R_2 in (6) to observe how they affect convergence. We find that our simulation does not converge to a single value when using W_3, so we exclude those plots. Fig. 2a shows how changing the scalar weights in the reward function can affect the convergence of the algorithm. To show the effect of changing the candidate solution space size on convergence, we consider two different sized action spaces ($\mathcal{A}_i : \{p_i, \beta_i\}$) for training our algorithm. The action space $\mathcal{A}_1$ is the one described in the paragraph above and $\mathcal{A}_2 : |p_i| = 2,\ \beta \in \{0, 0.5\}$. In Fig. 2b, we show the value of the reward function against the number of iterations as our algorithm progresses. In this particular case, we have five SCs in the network fully overlapping with the MUE, the reward function used is R_2 and Gaussian filter is used. At every iteration, one SCBS chooses an action and the reward changes correspondingly. Eventually, as all SCBSs explore all the possible action values, they all converge to take a certain action that maximizes the reward function. Hence, these actions that the SCBSs converge to are essentially determined by the formulation of the reward function.

In Fig. 3a, we compare the capacity for the MUE obtained from the proposed algorithm for a given number of interfering SCs under various simulation settings. One of the settings is based on full overlapping, where all the SCs choose the maximum transmit power level available to them ($p = 15$ dBm) and do not use POT ($\beta = 0$). The second plot for comparison is a state-of-the-art algorithm proposed in [2]. As expected, the results in Fig. 3a show that the full overlapping exhibits the worst performance. The algorithm we compare to by Amiri *et. al.* in [2] uses a co-operative multi-agent Q-learning approach to assign optimal transmit power levels to all interfering users in a two-tier HetNet. On the other hand, our algorithm optimizes each SC for optimal transmit power as well as FOs. The additional benefits of using POT in addition to the power control are clear from the figure since our proposed algorithm provides a higher capacity for the MUE. Thus, our approach finds a trade-off for both actions of the SCs so that it improves the MUE capacity compared to any other simulation setting. For the proposed algorithm, we use two different configurations of the scalar weights for both of the reward functions. When we give equal scalar weights, i.e., $\{\lambda_{11}, \lambda_{12}\} = \{\lambda_{21}, \lambda_{22}\} = W_1$, we achieve higher capacity when R_2 is used. Here we give equal weight to improving the MUE capacity and to improve individual SCUE capacity if using R_1 or the fairness index if using R_2. When we use $\{\lambda_{11}, \lambda_{12}\} = \{\lambda_{21}, \lambda_{22}\} = W_2$ instead, the algorithm prioritizes improving the capacity of the MUE, in turn achieving a higher performance.

In Fig. 3a, we compare the proposed algorithm by using the two different prototype filters, i.e., Gaussian (non-orthogonal) and RRC (orthogonal). FMT with a Gaussian filter outperforms the one with RRC in all simulation settings. Even in the presence of self-interference due to the non-orthogonality of the pulse shapes, by allowing the transmission to fit more subcarriers for a given bandwidth, Gaussian filters lead to a lower CCI. These results also corroborate with the findings in our previous results in [15].



(a) MUE capacity vs. the number of small cells. (b) BLER vs. E_s/N_0 at the MUE (Fading channel). (c) CDF of the BLER at the SCUE (Fading channel).

Fig. 3. Performance results for the proposed algorithm in different settings.

We compare the proposed method in terms of the BLER of the MUE in Fig. 3b. Using $\{\lambda_{11}, \lambda_{12}\} = \{\lambda_{21}, \lambda_{22}\} = W_2$ during training prioritizes the capacity improvement of the MUE, the result of which is evident again here. When using R_1 with $\{\lambda_{11}, \lambda_{12}\} = W_1$, the instantaneous change in the SCUE capacity can be negative which sometimes causes negative reward values, eventually leading to non-ideal values for power and FO allocation. We use the Extended Pedestrian A model (EPA) for characterizing multi-path channel between the MBS and MUE. For the channel between the SCs and MUE, we use the ITU channel model for an indoor office [16]. All simulations have five SCs interfering with the MUE and consider quadrature phase shift keying (QPSK) modulated FMT symbols. At the receiver, we use a maximum likelihood sequence estimation (MLSE) equalizer. We also use a rate $1/2$ low-density parity check (LDPC) code, where the parity check matrix is generated according to the DVB-S.2 standard.

To observe the effect of the two reward functions on the performance of the SCUEs, we compare average cumulative distribution function (CDF) of the BLER at the SCUE in Fig. 3c. Using R_2 with equal scalar weights provides better performance than other configurations shown in the plot. Since using $\{\lambda_{11}, \lambda_{12}\} = \{\lambda_{21}, \lambda_{22}\} = W_2$ prioritizes power and FO allocation to improve MUE performance, the SCUE performance suffers as SCUEs are assigned the same FOs in some cases and they fully overlap with each other.

V. Conclusion

We propose a multi-agent Q-learning approach to mitigate interference in a HetNet that accounts for the PHY waveform being used, POTs and transmit power control of SCs. Our simulation results show superiority in throughput performance compared to a state-of-the-art algorithm. Since the action space of our algorithm is larger, this superiority comes at a cost of higher computational complexity. The proposed algorithm can be extended to a larger scenario that accommodates multiple tiers of layers in the HetNet and incorporates a deep-Q learning-based algorithm, which will potentially reduce the training complexity while allowing us to redefine the states to be based on exact locations of the agents.

References

[1] L. Busoniu, R. Babuska, and B. D. Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[2] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement learning for self organization and power control of two-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3933–3947, Aug. 2019.

[3] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[4] 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.

[5] Y.-C. Wu, T. Q. Dinh, Y. Fu, C. Lin, and T. Q. S. Quek, "A hybrid DQN and optimization approach for strategy and resource allocation in MEC networks," *IEEE Trans. Wireless Commun.*, early access, Feb. 7, 2021, doi: 10.1109/TWC.2021.3057882.

[6] A. Şahin, E. Bala, I. Güvenç, R. Yang, and H. Arslan, "Partially overlapping tones for uncoordinated networks," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3363–3375, Sep. 2014.

[7] Y. Ding, Y. Huang, G. Zeng, and L. Xiao, "Using partially overlapping channels to improve throughput in wireless mesh networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 11, pp. 1720–1733, Nov. 2012.

[8] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[9] A. Şahin, I. Güvenç, and H. Arslan, "A survey on multicarrier communications: Prototype filters, lattice structures, and implementation aspects," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1312–1338, 3rd Quart., 2014.

[10] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," 1984, *arXiv:cs/9809099*. [Online]. Available: https://arxiv.org/abs/cs/9809099

[12] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication* (Wiley Series in Telecommunications). Cambridge, U.K.: Cambridge Univ. Press, 2005.

[13] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[14] 3GPP, "Further advancements for EUTRA physical layer aspects (release 9)," Tech. Rep. TR 36.814 (V9.2.0), Mar. 2017.

[15] M. Deshmukh, M. M. U. Chowdhury, S. J. Maeng, A. Sahin, and I. Guvenc, "RL-based interference mitigation in uncoordinated networks with partially overlapping tones," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[16] *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*, document Rec. ITU-R M.1225, 1997.