Spectrum Allocation for Network Slices with Inter-Numerology Interference using Deep Reinforcement Learning

Marco Zambianco, Giacomo Verticale
Dipartimento di Elettronica, Informatica e Bioingegneria
Politecnico di Milano – Italy
{marco.zambianco, giacomo.verticale}@polimi.it

***Abstract*—Network slicing and mixed-numerology schemes are essential technologies to efficiently accommodate different services in 5G radio access networks (RAN). To fully take advantage of these techniques, the design of spectrum slicing policies needs to account for the limited availability of the radio resources as well as the inter-numerology interference generated by slices employing different numerologies. In this context, we formulate a binary non-convex problem that maximizes the aggregate capacity of multiple network slices. The resulting spectrum allocation minimizes the inter-numerology interference under the frequent channel fluctuations characterizing the various users. To address the computational complexity of the designed objective function, we leverage deep reinforcement learning (DRL) to design a model-free solution computation. In detail, the trained centralized DRL agent exploits the channel fading statistic in order to provide a spectrum allocation that minimizes the inter-numerology interference. Results reveal that the proposed DRL scheme achieves performance that is comparable to the optimal one. It also outperforms a baseline scheme that statically allocate the radio resources.**

I. Introduction

RAN slicing is a network feature that makes it possible to deploy multiple independent virtual networks on top of the same physical network infrastructure. Each virtual entity is commonly denoted as "network slice" and has access to a set of common radio resources that are shared with other slices. The advantage of this technique, which has become an essential tool in 5G networks, is the flexibility to tailor the network according to the service level agreement (SLA) requirements of each application [1].

To fully benefit from this technology, two main challenges need to be addressed: i) differently from other physical resources like hardware computational capabilities, radio spectrum is a limited resource whose quality is inherently related to the signal propagation environment, ii) inter-slice isolation is affected by the external interference that is mutually generated between coexisting slices on the same spectrum. Such issues are further exacerbated by the introduction of mixed-numerology schemes that are employed to enhance the transmission performance over the radio interface. Differently from conventional orthogonal frequency division multiple access (OFDMA) schemes, such technology makes it possible to multiplex time-frequency resources, denoted as resource blocks (RB), having heterogeneous subcarrier spacing on the same physical layer. On one hand, this access scheme provides the flexibility to accommodate different radio requirements, on the other hand, the loss of orthogonality between RBs of different numerologies generates interference that hinders the transmission performance [2].

For these reasons, the design of effective RAN slicing policies should jointly account for the dynamic quality of the radio resources together with the inter-numerology interference (INI). However, in our opinion, most work has addressed these issues separately and the analysis of their mutual impact has received little attention. Following these observations, we propose a centralized agent-based allocation of mixed-numerology spectrum slices by leveraging deep reinforcement learning (DRL), which has recently found many applications in the field of wireless communications [3] [4]. In detail, the main contributions of this work are:

- We design a binary non-convex optimization problem for the allocation of radio resources to multiple network slices that are multiplexed on a mixed-numerology physical layer. The objective function maximizes the aggregated capacity of each slice by accounting for the channel state information (CSI) of the users as well as for the INI power that is mutually generated between the coexisting network slices.
- We propose a DRL formulation of the optimal problem in order to overcome its computational complexity. More specifically, we design a DRL agent that learns how the relationship between the small-scale fading fluctuations and the INI power, generated between different slices, affects the aggregate capacity of their users. Based on such information, the trained agent can simultaneously allocate the spectrum to each network slice without incurring in a significant time overhead.
- We compare the DRL based resource allocation with the optimal solution and with a heuristic scheme that performs a static spectrum allocation. Results show the effectiveness of our approach in approximating the optimal solution as well as a consistent performance gain over the heuristic scheme.

The remainder of the paper is organized as follows. We present

978-1-7281-4490-0/20/$31.00 © 2020 IEEE

the related work in Section II. We describe the system model in Section III. We discuss the optimal problem formulation as well as the DRL based resource allocation in Section IV. We show the simulation results in Section V. Finally, the conclusion is drawn in Section VI.

II. Related work

A general overview of the basic concepts of RAN slicing and mixed-numerology access schemes as well as their main challenges can be found in [5] [6]. Recent work in these fields has investigated resource allocation schemes that address the related issues in a unified scenario.

The authors of [7] propose several online joint scheduling algorithms to allocate low latency and multi-broadband users at different time granularity on a shared physical layer. Similarly, the work in [8] addresses a user allocation problem leveraging the flexible frame structure provided by mixed-numerology schemes. In detail, the authors design a self-adaptive flexible transmission time interval scheduling strategy for low latency and multi-broadband services. However, although [7] and [8] consider a mixed-numerology access scheme, the INI effect is not included in the proposed algorithms.

The authors of [9] and of [10] propose a INI-based scheduling scheme to reduce the size of the guard interval between different numerologies and to mitigate the interference power, respectively. However, the designed solutions are based on heuristic approaches, which are not supported by an optimal problem formulation that analytically accounts for the INI. The authors of [11] provide an allocation scheme, formulated as a max-min Knapsack problem, that allocates radio resources of different numerologies to fulfill the latency requirements of each user. The proposed solution also includes the INI effect, but the analysis is limited to the interference generated by the users of higher numerologies. Moreover, the performed optimization accounts only for the macroscopic fading.

Differently from the aforementioned work, we provide an optimal formulation for the INI minimization. Then, we propose a DRL agent-based spectrum allocation that minimizes the INI power between different network slices while considering the small-scale fading fluctuations that affect the channel quality.

III. System Model

We consider a RAN where the network owner (NO) leases the available radio spectrum to M mobile virtual network operators (MVNO) providing different services. Each MVNO m manages a logical standalone radio interface that schedules the available radio resources among its users, U_m, in order to fulfill the SLA requirements. Similarly, the NO manages the amount of radio resources required by each MVNO in order to ensure the feasibility of the service provisioning. In this context, the RAN slicing architecture can be modeled as a two-layer radio resource scheduler. The top-layer is composed by the MVNOs schedulers, whereas the bottom-layer corresponds to the NO, which acts as a “slice scheduler” by managing the radio resource allocation between MVNOs. A scheme of the considered network slicing architecture is depicted in Fig. 1.

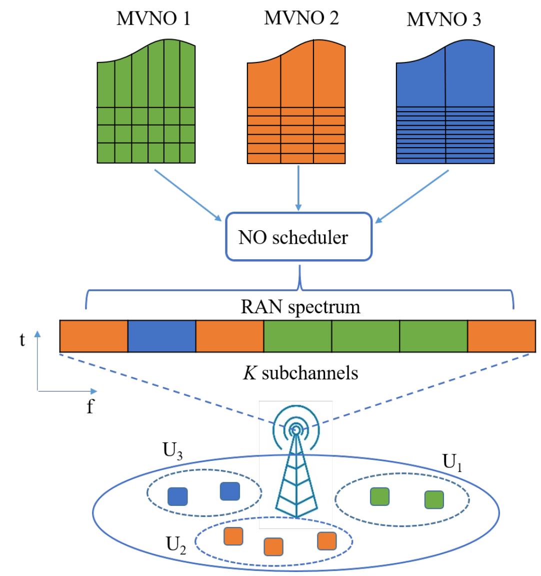

Fig. 1. RAN slicing architecture. The NO multiplexes the various MVNOs on the available RAN spectrum.

We assume that a mixed-numerology OFDMA scheme is employed as physical layer and we split the frequency-selective channel in K independent flat-fading subchannels, of bandwidth W, each one composed by a fixed number of contiguous RBs having the same numerology. To model different radio behaviors, each MVNO m selects a different numerology type, μ_m, which is tailored to the radio performance expected by its users. The numerology type defines the subcarrier spacing between symbols within each RB composing the various subchannels. Formally, we have that RBs of each MVNO m are characterized by a subcarrier spacing of $\Delta f_m = 15 \cdot 2^{\mu_m}$ kHz with $\mu_m \in \{0, 1, 2, 3, 4\}$ as dictated by the 3GPP specification [12].

We assume that the NO has full knowledge of the channel state information (CSI), so that each user u estimates the subchannel status and reports the related gain to the base station (BS) as feedback. We model the power gain in each subchannel as composed by large-scale fading, α_m^u, that is due to path loss and shadowing, and small scale fading, $h_m^u(k)$, that is assumed to be exponentially distributed with unit mean. Formally, we have that the subchannel gain of subchannel k reported by user u belonging to slice m can be written as

$$g_m^u(k) = \alpha_m^u h_m^u(k). \tag{1}$$

The simultaneous multiplexing of subchannels composed by RBs of different numerologies produces INI. As a matter of fact, the symbol orthogonality is ensured only within each subchannel as the employed numerology is homogeneous. To quantitatively model the INI dynamic, we rely on the analytical formulation proposed by authors in [13], which we adapted to fit the considered system scenario. In general, given two subchannels, k and k', of numerologies with subcarrier spacing

Δf_m and $\Delta f_{m'}$, the INI power affecting user $u \in U_m$ on subchannel k due to subchannel k' can be computed as, if $\Delta f_m < \Delta f_{m'}$,

$$I_m^u(k,k') \approx \frac{P_T(k')}{N_{k'}} \sum_{z=1}^{N_k} \sum_{v=1}^{N_{k'}} \frac{g_m^u(k')}{N_{k'} N_k} \left[\left| \frac{\sin[\frac{\pi}{N_k} w(z,v) \xi N_{k'}^{(T)}]}{\sin(\frac{\pi}{N_k} w(z,v))} \right|^2 + \xi \left| \frac{\sin[\frac{\pi}{N_k} w(z,v) N_{k'}^{(T)}]}{\sin[\frac{\pi}{N_k} w(z,v)]} \right|^2 \right], \tag{2}$$

otherwise, if $\Delta f_m > \Delta f_{m'}$, as

$$I_m^u(k,k') \approx \frac{P_T(k')}{N_{k'}} \sum_{z=1}^{N_k} \sum_{v=1}^{N_{k'}} \frac{g_m^u(k')}{N_{k'} N_k} \left| \frac{\sin[\frac{\pi}{N_{k'}} w(z,v) N_k]}{\sin[\frac{\pi}{N_{k'}} w(z,v)]} \right|^2 \tag{3}$$

where $N_k = W/\Delta f_m$ denotes the number of subcarriers in subchannel k, $P_T(k)$ is the power allocated on subchannel k, $N_k^{(T)} = N_k + N_k^{CP}$ is total number of subcarriers in subchannel k with N_k^{CP} corresponding to the number of subcarriers used as cyclic-prefix, $\xi = \lfloor N_k / N_{k'}^T \rfloor$ is the number of overlapping OFMD symbols within the same transmission frame, and $w(z,v)$ denotes the spectral distance between subcarriers of different numerolgies and it is computed as the total number of subcarriers separating subcarrier z from subcarrier v.

From (2) and (3) we note that INI power depends on the spectral distance between the subcarriers of different numerologies and the power allocated to each subcarriers. More precisely, the higher is the spectral gap between the victim and the aggressor subcarrier and the higher is the power of the aggressor subcarrier, the more interference is generated on the victim subcarrier.

We measure the subchannel quality perceived by each user u belonging to MVNO m and associated to each subchannel k as the signal-to-interference-plus-noise ratio (SINR)

$$\gamma_m^u(k) = \frac{P_T(k) g_m^u(k)}{\sigma_w^2 + \sum\limits_{m' \neq m} \sum\limits_{k' \neq k} x_{m',k'} I_m^u(k,k')}, \tag{4}$$

where $x_{m,k}$ is the binary subchannel allocation indicator assuming value $x_{m,k} = 1$ if subchannel k is allocated to MVNO m and $x_{m,k} = 0$ otherwise, and σ_w^2 is the white Gaussian noise power over each subchannel. From (4), we observe that the INI power generated in subchannel k depends on the subchannel gain related to the subchannels of different numerologies. Therefore, it is possible to exploit the independent fading effect on the various subchannel to reduce the INI power. This observation covers an important role in the design of the DRL algorithm that is going to be presented in Section IV.

IV. Problem formulation

In this section, we formalize the optimal spectrum allocation and we present the DRL framework proposed to approximate the optimal solution.

A. Optimal resource allocation

According to the system model previously described, the NO periodically computes the number of subchannels required by each MVNO based on the service demand. We denote as S_m the spectrum assignment policy that expresses the number of subchannels assigned to MVNO m such that $\sum_{m \in M} S_m = K$. The various subchannels are then allocated over the shared spectrum following two requirements:

1) The negative effect of INI is minimized.
2) Multi-user diversity is exploited to improve the aggregated capacity of each MVNO. In other words, the best subchannel, providing the highest aggregated capacity, should be allocated to each MVNO.

Following these observations, we design an optimization problem to compute the optimal subchannel allocation that maximizes the system capacity under the requirement 1) and 2). We formalize the resource allocation problem as

$$\max_x \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{1}{U_m} \sum_{u=1}^{U_m} x_{m,k} \cdot W \log_2 \left(1 + \gamma_m^u(k)\right) \tag{5}$$

subject to

$$\sum_{k=1}^{K} x_{m,k} = S_m \quad \forall m \in M \tag{6}$$

$$\sum_{m=1}^{M} x_{m,k} \leq 1 \quad \forall k \in K \tag{7}$$

$$x_{m,k} \in \{0,1\} \quad \forall m \in M, \forall k \in K. \tag{8}$$

Intuitively, the effect of such optimization is the allocation of each subchannel to the MVNO that can provide that highest throughput to its users while simultaneously minimizing the INI power between different slices. Problem constraints define a feasible resource allocation according to the proposed system model. In detail, (6) ensures that a suitable number of subchannels are allocated to each MVNO based on the spectrum assignment policy S_m. Equation (7) guarantees that every subchannel is allocated to a single MVNO only. Finally, (8) expresses the integer nature of the problem by means of the binary optimization variable $x_{m,k}$.

The resulting allocation increases the scheduling flexibility of each MVNO because their users have access to spectrum slices providing favorable radio conditions. In addition, slice isolation is enforced as the INI power is minimized. However, the computation of the solution of the proposed problem is challenging due to the non-convexity of (5) whose optimization variable is integer. Furthermore, the complexity is further exacerbated by the highly non-linear equation defining the INI as shown in (2) and (3). These issues make the discussed problem formulation unpractical for real systems due to the stringent time requirements required to adapt the slice allocation to the wireless channel dynamic.

For this reason, we propose an alternative approach to approximate the solution of (5) in a lower amount of time. In this regard, we design an allocation policy of the radio resources by leveraging DRL. The main advantage of this scheme is that it allows to compute an allocation policy under a model-free environment formulation. Consequently, we can overcome the challenges related to the design of a heuristic algorithm capable of effectively modeling the complex relationship between the fading fluctuations and the INI dynamic. In the next section we describe the proposed DRL scheme.

B. DRL-based resource allocation

DRL provides an iterative method to compute an optimal policy for solving a Markov Decision Process (MPD), where the transition probabilities from each state towards other states are unknown [14]. Formally, an MDP can be formulated by a 5-tuple composed by $\{S, A, p(s'|s,a), R(s,a), \beta\}$, where S and A denotes, respectively, the state space and action space, $p(s'|s,a)$ denotes the transition probability from state s at time t toward state s' at time $t+1$ and depends on the current state $s \in S$ and the action $a \in A$, $R(s,a)$ is the immediate reward that is obtained by performing action a under state s, that is discounted over time by a factor $\beta \in [0,1)$. This parameter models the diminishing returns of the current rewards in next time slot t' and expresses the uncertainty of the agent about the impact of its actions on future rewards. The goal of the learning is to find the optimal policy that allows to maximize the expected discounted reward, G_t, from any initial state s, i.e.

$$G_t = \sum_{i=0}^{\infty} \beta^i R_{t+1+i}, 0 \leq \beta < 1. \tag{9}$$

The optimal policy provides the probability of selecting each action from any state that maximizes the expected future rewards that are obtained by following such policy.

C. State space

We model the environment that is observable by the agent as the subchannel gains reported by each user and the INI power measured on every subchannel without accounting for the channel fading fluctuations. In other words, we compute the INI power that would be received by every user given the transmission power allocated on every subchannel. We obtain such quantity by setting $g_m^u(k) = 1$ in (2) and (3). Note that both observations do not require any significant signalling overhead as they can be easily retrieved by the NO at the BS. Formally, we have that in each time slot t the agent observes the environment that is characterized by the state space

$$S = \{G[k], I[k]\}_{k \in K} \tag{10}$$

where

$$G[k] = \{g_m^u(k)\}_{m \in M, u \in U_m}, \tag{11}$$

$$I[k] = \Big\{ \sum_{m' \neq m} \sum_{k' \neq k} x_{m',k'} I_m(k,k') \Big\}_{m \in M}. \tag{12}$$

We remark that the INI power computed as (12) is the same for all the users belonging to the same MVNO, hence we dropped the index u from the notation. The underlying idea behind this design choice is to let the agent learn how the INI power is affected by the fading fluctuations so that it can exploit the independent fading between subchannels to better mitigate the INI.

D. Action space

The action space is composed by all the feasible subchannel allocations. More specifically, starting from one feasible subchannel allocation that satisfies the spectrum assignment policy S_m, the total number of the available actions can be enumerated by computing the related unique permutations, which are equal to $N = |A| = \frac{K!}{S_1!...S_M!}$. Note that such action space design directly embeds constraints (6)-(7) within the action definition, hence it automatically denies the agent from selecting unfeasible allocation policies. The action set can be formalized as follows

$$A = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\} \tag{13}$$

where

$$\mathbf{X_i} = \begin{bmatrix} x_{1,1} & \ldots & x_{1,K} \\ \vdots & \ddots & \ldots \\ x_{M,1} & \ldots & x_{M,K} \end{bmatrix}_{i \in N}. \tag{14}$$

Every element in (13) models a specific subchannel allocation leveraging the subchannel indicator function as shown in (14). In details, the matrix rows represent the subchannel allocation associated to the M MVNOs, whereas the columns indicate whether the k-th subchannel is allocated to MVNO m.

E. Reward design

The reward function characterizes the learning performance of the agent while it interacts with environment. In the considered scenario, our objective is to design an allocation policy that maximizes the aggregated capacity. Hence, we directly employ (5) to model the reward obtained by the agent at each time step. More precisely, the agent gets a reward R_{t+1} that is evaluated using (5) according to the subchannel allocation chosen and the environment state modeled as (10) at time t.

To improve the learning stability, we discount each reward using a high discount rate β. The motivation is to incentive a greedy behavior of the agent in order to compute a subchannel allocation policy that maximizes the system throughput in the immediate time slots. As a matter of fact, the agent does not gain any substantial benefit in the knowledge of the expected reward in time slots that are far in the future given the current state. Instead, its goal is to approximate the behavior of (5) that maximizes the aggregate capacity at every new CSI update.

F. Algorithm overview

We adopt deep Q-learning (DQN) with experience replay [15] to compute the spectrum allocation policy approximating the optimal solution. In order to design a flexible agent that can adapt to different radio scenarios, we perform the training procedure on multiple episodes characterized by a different user distribution that is kept constant for all the episode duration. Each episode is composed by a fixed number of time steps corresponding to a new CSI reporting. The agent explores the environment using an ϵ-greedy policy. In details, with probability ϵ, the agent randomly selects a subchannel allocation regardless of its current state, otherwise, with probability 1-ϵ, it computes the slice allocation that maximizes the expected long-term reward based on both the INI power and CSI observations. The optimal policy π^* is computed from the state-action value function $Q_\pi(s,a)$, also denoted as Q-function, which is defined as the average discounted reward obtainable starting from state s, taking action a and following the policy π. Formally, we can write the optimal policy as

$$Q_{\pi^*}(s,a) = \max_{a \in A} Q_\pi(s,a) \tag{15}$$

where

$$Q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a] \tag{16}$$

with G_t defined in (9). Every time step, the Q-function (16) is updated as

$$\begin{aligned} Q(s_t,a_t) \leftarrow{} & (1-\alpha)Q_\pi(s_t,a_t) + \\ & \alpha[R(s_t,a_t) + \beta \max_{a_{t+1} \in A} Q_\pi(s_{t+1},a_{t+1})], \end{aligned} \tag{17}$$

where $\alpha \in [0,1]$ is the learning rate.

DQN employs a deep neural network (DNN) of weights $\{\boldsymbol{\theta}\}$ to approximate (16) and overcoming the memory complexity required to individually store the Q-function values for each state-action pair. A scheme of the DNN architecture is shown in Fig. 2.

The DNN weights $\{\boldsymbol{\theta}\}$ are updated in order to minimize the loss, $L(\boldsymbol{\theta})$, between the Q-function values computed on subsequent time steps. Formally, $L(\boldsymbol{\theta})$ is defined as

$$\begin{aligned} L(\boldsymbol{\theta}) = \sum_D [& R(s_t,a_t) + \beta \max_{a_{t+1} \in A} Q_\pi(s_{t+1},a_{t+1};\boldsymbol{\theta}') \\ & - Q_\pi(s_t,a_t;\boldsymbol{\theta})]^2, \end{aligned} \tag{18}$$

where $\{\boldsymbol{\theta}'\}$ corresponds to the weights of a second DNN that is used to stabilize the Q-function computation convergence and it is updated as $\{\boldsymbol{\theta}' = \boldsymbol{\theta}\}$ every few time steps. Parameter D is the size of the mini-batch that is randomly sampled from the experience-replay buffer. The latter collects and stores the most recent N tuples $(s_t, a_t, s_{t+1}, a_{t+1})$ generated by the agent during the training phase. This procedure improves the learning performance by breaking the correlation between subsequent weights updates [15]. In Algorithm 1, we summarize the training phase.

Although this procedure is performed offline and it can be computationally expensive, we highlight that the online deployment of the trained agent takes a considerable lower amount of time. The subchannel allocation is computed by the agent according to the optimal policy (15) that is easily derived by feeding the trained DNN with the current system state. Moreover, assuming that network parameters like RAN spectrum and number of active users are fixed, the agent requires to be re-trained only when a new spectrum assignment S_m is computed. This event is likely to rarely occur as it is required to accommodate unexpected traffic load variations between MVNOs. However, we acknowledge that the proposed DRL scheme scales poorly when the number of subchannels to allocate increases as shown in (13). We are going to address this issue in a future work by designing a suitable action reduction algorithm to enhance the agent scalibility in more complex scenarios.

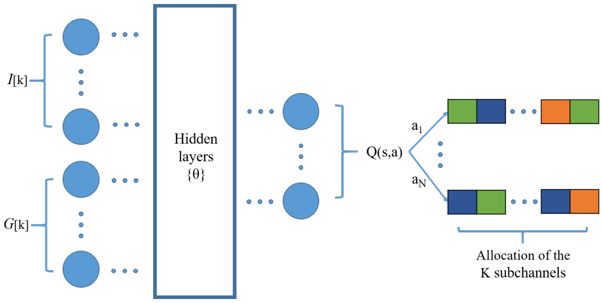

Fig. 2. Fully-connected DNN structure used to compute $Q(s,a)$. According to the input state, the action corresponding to the highest Q-function value is selected. Each action provides a subchannel allocation for all MVNOs based on the assignment policy S_m.

Algorithm 1: Training procedure

```
Result: Subchannel allocation according to the
        spectrum assignment policy $S_m$
Wireless environment and DRL agent initialization;
for each episode do
    for each MVNO $m \in M$ do
        Randomly place $U_m$ active users over the BS
         coverage area;
    end
    for each time step do
        for each subchannel $k \in K$ do
            Simulate INI power in every subchannel
             using (12);
            Generate CSI reporting of every user
             $u \in U_m$;
        end
        Observe the environment state $s_t$;
        Select the action $a_t$ according to the $\epsilon$-greedy
         policy;
        Store the tuple $(s_t, a_t, s_{t+1}, a_{t+1})$ in the
         experience-replay buffer;
        Uniformly sample a mini-batch of size $D$ from
         the experience-replay buffer;
        Update $\boldsymbol{\theta}$ by minimizing $L(\boldsymbol{\theta})$ in (18) using the
         sampled mini-batch;
    end
end
```

V. Performance Evaluation

We discuss the results obtained by the proposed DRL scheme. The whole simulation framework has been developed in MATLAB. As network scenario, we considered a single BS shared by up to 3 MVNOs having numerologies of subcarrier spacing ranging from 15 kHz to 60 kHz. Every subchannel has

TABLE I
SIMULATION PARAMETERS

Network parameters	**Value**
Subchannel transmission power	18 dBm
BS coverage radius	200 m
Carrier frequency	2.5 GHz
Available numerologies	$\{15, 30, 60\}$ kHz
Subchannel bandwidth	740 kHz
Active users per MVNO	4
Fading statistic	Rayleigh
Doppler shift	35 Hz
Fading update	1 ms
Path loss model [17]	$36.7 \log_{10} d + 33.05$
Shadowing standard deviation	4 dB
Noise power	-115 dBm
DRL parameters	**Value**
Learning rate	0.001
Discount factor	0.01
Experience-replay buffer size	1000
Mini-batch size	64
Episode duration	100 ms
Number of episodes	500

equal transmission power and we assume that each MVNO serves the same number of active users.

As DRL parameters, we employed a fully-connected DNN with 5 hidden layers of 300 neurons each, that are activated according to the Rectifier Linear Unit (ReLU), $f(x) = max(0, x)$, function. The loss function in (18) is minimized using RMSProp optimizer [16] and the related Q-function value is updated using a learning rate $\alpha = 0.001$ in (17). This parameter configuration was chosen experimentally by assessing the result quality with respect to a variable number of layers and neurons. The agent randomly explores the environment with probability ϵ, that is exponentially decremented at every new episode starting from $\epsilon = 1$ up to $\epsilon = 0.01$. We trained the agent for 500 episodes that are composed by 100 time steps of granularity 1 ms. Moreover, to improve the learning performance, we normalized the CSI reporting with respect to the path-loss of each user before feeding it to the DNN. We report the simulation parameters in Table I.

We analyze the performance of the DRL agent by comparing it with two baseline schemes, denoted as

- *Optimal allocation:* the optimal subchannel allocation is computed by means of an exhaustive search of all feasible solutions in (5) at every CSI update.
- *Static allocation:* the same subchannel allocation is chosen among the ones available regardless of the CSI update.

The first scheme allows to investigate the gap of the DRL solution from the optimal solution, whereas the second scheme allows to highlight the performance gain over a heuristic approach that does not jointly consider the channel fluctuations together with the INI power. We consider different simulations scenarios in order to provide better insights of the agent performance. In this regard, we assess the results for a variable number of subchannels. For each scenario, we consider 2 and 3 MVNOs employing numerologies with subcarrier spacing 15 kHz, 30 kHz and 15 kHz, 30 kHz, 60 kHz, respectively.

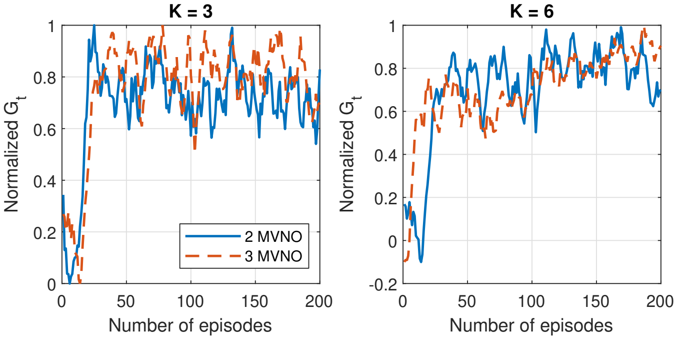

Fig. 3. Average reward obtained by the DRL agent when 2 MVNOs of numerologies $\{15, 30\}$ kHz and 3 MVNOs of numerologies $\{15, 30, 60\}$ kHz are considered. The number of subchannels is $k = 3$ and $k = 6$.

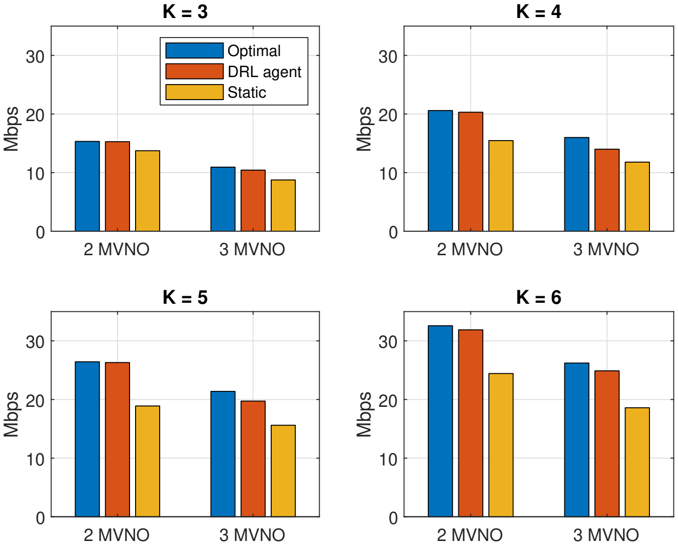

Fig. 4. Aggregate user throughput obtained with the optimal allocation, the DRL agent allocation and the static allocation. The active numerologies are $\{15, 30\}$ kHz and $\{15, 30, 60\}$ kHz when 2 MVNOs and 3 MVNOs are considered, respectively. The number of subchannels ranges from $k = 3$ to $k = 6$.

In Fig. 3, we show the average episode reward obtained by the agent during the first 200 episodes of training phase (we do not show all the 500 episodes for the sake of plot clarity). We observe that the agent requires more episodes to converge when the number of MVNOs and subchannels increases. Specifically, the reward curve takes more episodes to settle over a stationary trend when $K = 6$ due to the fact the a higher number of actions are available and a longer environment exploration is required to fully capture the underlying system dynamic.

In Fig. 4, we show the aggregated throughput achieved by the users belonging to the different MVNOs. We compare the DRL-based allocation with the optimal allocation and the static allocation previously discussed. We averaged the results across 50 independent episodes where, for the static allocation, we randomly chose a new subchannel allocation in each

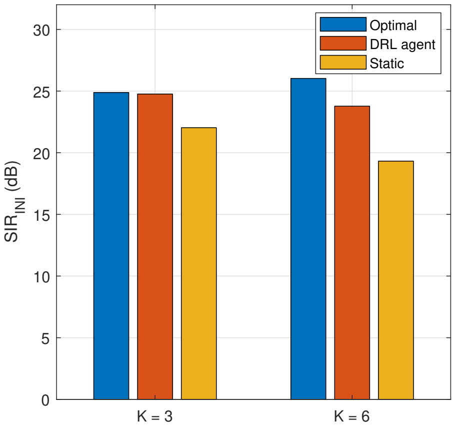

Fig. 5. Average SIR due to INI obtained with the optimal allocation, the DRL agent allocation and the static allocation when the number of subchannels is $k = 3$ and $k = 6$. The number of active MVNOs is $M = 3$.

episode. We observe that the DRL agent reaches performance that is very close to the optimal solution in most of the considered scenarios, where the minimum gap is achieved when 2 MVNOs are active. Conversely, we note a more visible performance loss in the 3 MVNOs case, that is due to the higher number of available subchannel allocations. Moreover, it is interesting to observe that increasing the number of subchannels has a limited impact on the agent performance. The motivation is that the INI dynamic is more dependent on the number of different active numerologies rather than the number of subchannels. Therefore, the agent can easily adapt the learning process to additional subchannels that are shared by the same number of MVNOs. The static allocation achieves a lower throughput in all scenarios thus highlighting the benefit of the considered approach to boost the system performance. In detail, we note that the highest gain is achieved when 3 MVNOs are active. This behavior is due to the fact that when more numerologies are multiplexed on the same spectrum the INI power generated between different slices increases since more non-orthogonal subchannels are contiguously allocated. Consequently, the advantage of the proposed subchannel allocation is more evident as more scheduling opportunities, that minimize INI, are available.

Finally, in Fig. 5, we plot the average signal-to-interference ratio (SIR) due to INI that is achieved by the various users. In general, we observe that the DRL agent provides higher average SIR values compared to the static allocation. Moreover, the DRL agent achieves performance close to the optimal allocation when the number of subchannels is $k = 3$, whereas a SIR degradation can be noted for $k = 6$ due to the more challenging allocation scenario.

VI. Conclusion

We proposed a spectrum allocation policy that maximizes the aggregated capacity of multiple MVNOs by taking into account inter-numerology interference as well as the small-scale fading effects of the various users. We designed a centralized DRL-based spectrum allocation scheme that provides an effective approximation of the optimal solution in an efficient amount of time. The DRL agent learns a suitable spectrum allocation by correlating the received reward with the INI power affecting each subchannel and the CSI reported by the users. Results showed that the proposed DRL scheme achieves performance very similar to the optimal allocation in most scenarios and that it outperforms a baseline scheme that statically allocates the radio resources to each MVNO.

References

[1] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5G: A tutorial overview of standards, trials, challenges, deployment, and practice,” *IEEE journal on selected areas in communications*, vol. 35, no. 6, pp. 1201–1221, 2017.

[2] A. B. Kihero, M. S. J. Solaija, A. Yazar, and H. Arslan, “Inter-numerology interference analysis for 5G and beyond,” in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–6.

[3] R. Li, Z. Zhao, Q. Sun, I. Chih-Lin, C. Yang, X. Chen, M. Zhao, and H. Zhang, “Deep reinforcement learning for resource management in network slicing,” *IEEE Access*, vol. 6, pp. 74 429–74 441, 2018.

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

[5] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, “5G RAN slicing for verticals: Enablers and challenges,” *IEEE Communications Magazine*, vol. 57, no. 1, pp. 28–34, 2019.

[6] A. A. Zaidi, R. Baldemair, V. Molés-Cases, N. He, K. Werner, and A. Cedergren, “OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN,” *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 78–83, 2018.

[7] A. Anand, G. De Veciana, and S. Shakkottai, “Joint scheduling of URLLC and eMBB traffic in 5G wireless networks,” *IEEE/ACM Transactions on Networking*, 2020.

[8] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, “Machine learning based flexible transmission time interval scheduling for eMBB and URLLC coexistence scenario,” *IEEE Access*, vol. 7, pp. 65 811–65 820, 2019.

[9] A. F. Demir and H. Arslan, “Inter-numerology interference management with adaptive guards: A cross-layer approach,” *IEEE Access*, vol. 8, pp. 30 378–30 386, 2020.

[10] A. Yazar and H. Arslan, “Reliability enhancement in multi-numerology-based 5G new radio using ini-aware scheduling,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 110, 2019.

[11] L. Marijanovic, S. Schwarz, and M. Rupp, “Multi-user resource allocation for low latency communications based on mixed numerology,” in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE, 2019, pp. 1–7.

[12] 3GPP, “NR; Physical channels and modulation,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, 2019, version 15.5.0.

[13] X. Zhang, L. Zhang, P. Xiao, D. Ma, J. Wei, and Y. Xin, “Mixed numerologies interference analysis and inter-numerology interference cancellation for windowed OFDM systems,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7047–7061, 2018.

[14] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[16] S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv preprint arXiv:1609.04747*, 2016.

[17] ITU-R, “Guidelines for evaluation of radio interface technologies for IMT-2020,” International Telecommunication Union (ITU), Tech. Rep., 2017.