Uplink Power Control Framework Based on Reinforcement Learning for 5G Networks

Francisco Hugo Costa Neto, Daniel Costa Araújo, Mateus Pontes Mota, Tarcisio F. Maciel, André L. F. de Almeida

***Abstract*—In this work, we propose an uplink power control (PC) framework compliant with the technical specifications of the fifth generation (5G) wireless networks. We apply the fundamentals of reinforcement learning (RL) to develop a power control algorithm able to learn a strategy that enhances the total data rate on the uplink channel and mitigates the neighbor cell interference. The base station (BS) uses a set of commands to specify by how much the user equipment (UE) transmit power should change. After implementing such commands, each UE reports a set of information to its serving BS, and this, in turn, predicts the next commands to achieve a suitable UE transmit power level. The BS converts the UE reports into rewards according to a predefined cost function, which impacts the longterm behavior of the UE transmit power. The simulation results indicate a near-optimum performance of the proposed framework in terms of total transmit power, total data rate, and network energy efficiency. It provides a self-exploratory power control strategy that overcomes soft dropping power control (SDPC) with similar signaling levels.**

***Index Terms*—uplink power control, reinforcement learning**

I. Introduction

Uplink power control (PC) constitutes an essential design problem of wireless communication networks. This important radio resource management technique provides mechanisms to increase the system capacity, coverage, and quality of service (QoS) while limiting interference to neighbor cells [1]. The fourth generation (4G) long-term evolution (LTE) supports several solutions based on well-founded technical literature to the uplink PC problem with distinct objectives regarding different deployment scenarios and services [2], [3], [4], [5].

Despite the advances promoted by LTE, there are challenging performance requirements imposed by new wireless communication networks use cases, such as enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine type communication (mMTC) [6]. In this context, the development of intelligent uplink PC strategies is essential to guarantee the QoS specifications of these use cases, specially for eMBB (higher data rates) [7] and mMTC (longer battery life for low-power devices) [8] at similar cost and energy consumption levels compared with 4G LTE networks [9].

Francisco Hugo Costa Neto, Mateus Pontes Mota, Tarcisio F. Maciel, and André L. F. de Almeida are with the Wireless Telecommunications Research Group, Department of Teleinformatics Engineering, Center of Technology, Federal University of Ceará, 60020-181, Fortaleza, Brasil (e-mail: hugo@gtel.ufc.br, mateus@gtel.ufc.br, maciel@gtel.ufc.br, andre@gtel.ufc.br). Daniel Costa Araújo is with the Department of Electrical Engineering, Campus Gama, University of Brasília, Brasil (e-mail: daniel.araujo@unb.br). (*Corresponding author: Francisco Hugo Costa Neto.*)

The fifth generation (5G) radio access technology, also called new radio (NR), has as one innovative aspect the support for a large number of antenna elements, which motivates the development of advanced multi-antenna techniques, such as beamspace massive multiple-input multiple-output (MIMO) [10]. Since multiple antenna technology has become a key component of the 5G networks, it has been adopted the beam-centric design of channels and signals [6]. Consequently, the uplink PC problems in 5G NR faces more challenging design concerns compared with 4G LTE and other classical solutions.

Then, the development of a new uplink PC paradigm to support 5G NR demands mentioned previously requires more sophisticated tools. In this context, machine learning (ML) emerges as a powerful source of mechanisms to uncover unknown properties of wireless networks, identify correlations, and suggest novel ways to optimize network deployment [11]. ML provides algorithms to forecast system behavior and find potential solutions by interacting with the environment. Enhancing the intelligence of wireless networks is essential to provide it with self-organization features, like self-configuration, self-optimization, and self-healing [12]. Moreover, 5G NR devices endowed with intelligence must be able to control the transmission power while relying on energy efficiency learning [13]. In this context, the design of an ML-based uplink PC solution compliant with the 5G NR radio access technology is a relevant research topic and is the focus of this work.

ML algorithms belonging to the reinforcement learning (RL) category are of particular interest to radio resource management. They learn from interactions with a dynamic environment on how to achieve a desired behavior. RL is a relevant tool to empower networks with autonomous self-adaptive algorithms provided with adaptability and capable of taking advantage of experience when making decisions [14]. These algorithms demonstrated pattern recognition ability and can be applied successfully to solve several radio resource management problems. For more details on RL applications, the reader may refer to [14], [15], [16] and the references therein.

In this work, we seek an answer to the question of whether RL-based uplink PC can help to mitigate inter-cell interference in 5G NR networks. Although this question is highly relevant from a research and system development perspective, it has not been exhaustively and suitably addressed by previous works, as discussed further in the next section. To this end, we propose a distributed multi-agent RL-based uplink PC compliant with 5G NR specifications [17], [18], [19], [20], [21], [22], [23].

Specifically, we propose a beam-based uplink PC framework to provide a self-exploratory solution where the base station (BS), considered as the decision-maker entity, learns the most suitable solution to the power control problem on the uplink. The BS uses a set of commands to specify by how much the user equipment (UE) transmit power should change. Based on the UE's reports generated as a response to those commands, the BS learns through experience what should be the best transmit power to be used by its associated UE. The BS converts the UE reports into rewards according to a predetermined cost function, which determines the long-term behavior of the uplink PC.

The remainder of this work is organized as follows. Section II discusses related works and our contributions. Section III describes the system model assumptions and Section IV reviews the fundamental concepts of RL. Then, Section V describes the 5G NR specifications to deploy uplink PC and the proposed RL framework. Simulation results are discussed in Section VI. Finally, the main conclusions are drawn in Section VII.

II. Literature Review and Contributions

The application of RL-based algorithms into uplink PC problems constitute a set of self-organized solutions capable of finding autonomously suitable transmit power levels. Consequently, they can reduce inter-cell interference and increase system data rate properly. They are suitable techniques and have been relevant in numerous studies in the field. In [24], the authors proposed a PC framework to manage the interference in a cognitive radio network. They modeled the wireless network as a multi-agent system, where the agents interact directly with the environment and learn a strategy to manage the power levels. In their model, the BSs represent the decision-maker entities which manage the radio resources allocated to their associated UEs. However, they focused on the downlink operation and presented results without considerations about third generation partnership project third generation partnership project (3GPP) specifications. We propose an uplink PC strategy which follows this multi-agent RL system modeling. Moreover, we incorporate practical 5G NR specifications into the proposed framework, such as required signaling, available power control commands, and hardware constraints.

In [25], the authors considered a more realistic cognitive radio network, modeled as a wireless regional area network (WRAN) compliant with the 802.22 [26] standards. Therein, they considered the downlink and uplink operation and situations of complete and incomplete information about the environment. The results indicated that a multi-agent RL system could automatically learn a policy to successfully manage the interference, without introducing signaling overhead in the system.

In [27], a decentralized uplink PC based on the multi-agent RL combined with a fractional power control (FPC) mechanism was proposed for LTE-based multi-tier networks. In their study, each UE decides the transmit power based on the channel conditions, namely, the uplink path loss. In this framework, each UE learns independently the transmit power without the need to wait for control signaling from an associated BS. It is show that the solution reduces the signaling in uplink transmission. However, due to the limited computational resources of the UE, it also reduces the processing capability of the decision-maker entity. In our approach, due to the beam-centric design of 5G NR, several PC mechanisms must be handled in parallel and in real-time. Consequently, the BS is considered as the entity endowed with intelligence. Moreover, we evaluate the impact of different levels of signaling on the decision-making ability.

The authors of [28] also investigated a learning-based power control based on an FPC mechanism in LTE systems. They presented a data-driven framework to model the interference patterns in orthogonal frequency division multiple access (OFDMA)-based networks. Based on the measurement of these interference patterns, the proposed learning algorithms define an optimal setting of the cell-specific power control parameters. Therein, the authors assumed that all path loss variables must be interpreted in a time-scale sense so that it averages the effect of fast-fading. In other words, the uplink PC mechanism proposed in [28] can compensate for path loss and large-scale variations such as shadowing, but does not adequately handle fast fading. In scenarios where these effects are prominent, this simplification may render an inappropriate representation of the channel conditions, restricting the success of that uplink PC solution.

The uplink PC frameworks proposed at [27] and [28] are based on the FPC mechanism. They employed the open-loop power control (OLPC) paradigm, i.e., they defined the transmit power according to large-scale channel conditions, namely, path loss measurements. The conventional FPC solution identifies UEs based only on the path loss. This is not entirely proper in more complex scenarios since interference conditions are not considered while allocating the transmit power. Consequently, this strategy usually results in high interference situations [29]. To overcome these issues, we perform an additional transmit power adjustment according to the closed-loop power control (CLPC) paradigm. In this case, we also consider the impact of transmit power commands taken earlier in the system on the choice of the uplink PC strategy.

To overcome these issues, in our work we perform an additional transmit power adjustment according to the impact of the commands taken earlier in the system. This is achieved by means of a multi-agent RL-based power control combined with the CLPC paradigm. In [30], the authors proposed an uplink PC framework in a cognitive network. In this paper, the decision-maker at each secondary UE performs a CLPC mechanism that perceives as a useful policy the actions that improve the signal to interference plus noise ratio (SINR) above a given threshold. This study considers that the UEs work in a non-cooperative manner, i.e., a secondary UE does not have any knowledge about the primary UE PC strategy. In our study, we investigate the impact of the cooperation among the UEs of the network on the CLPC mechanism. The proposed uplink PC considers a distributed strategy where the interaction among the decision-maker entities determines the knowledge acquisition process. In [31], the authors proposed a strategy which learns a policy that guides transmitters to adjust their power levels according to the CLPC paradigm under practical constraints, such as delayed information exchange

and incomplete channel state information (CSI). However, the authors do not consider the beam-centric design aspects defined in 5G NR. In our work, we develop a flexible CLPC strategy that takes into account the coordination among multiple beams and the limitation from 3GPP standards.

In [30] and [31], the authors used a deep reinforcement learning (DRL) method, called deep Q-network (DQN) [32]. Other studies also tackled transmit power strategies based on DRL algorithms. In [33], it is proposed a joint subcarrier and power allocation in a multi-cell orthogonal frequency division multiplexing (OFDM) system. In this paper, each BS implements a policy of resource allocation. They are the decision-maker entities and can exchange knowledge to jointly define a strategy of transmit power update. Zhang et al also developed a multi-agent DRL-based PC framework in [34]. Therein, each agent adaptively controls its transmit power based on the observed local states to minimize the interference in a multi-user video transmission system. In [35], a DRL-based framework is designed to solve the joint beamwidth and power allocation problem in a mmWave communication system. Despite the sophisticated mathematical tools, the authors did not incorporate practical aspects of implementation of 5G NR network into their solutions.

Therefore, DRL-based techniques have achieved remarkable attention in the last years, as it can be seen in [36], [37], [38] and references therein. Differently from conventional Q-learning, which uses a lookup table to store the knowledge, the DRL-based algorithms employ a deep neural network to represent this information. The lookup table is shown to be more computationally efficient than the neural network approach. On the other hand, the DRL-based approach has reduced memory requirements compared to the lookup table [24]. DRL-based techniques can provide suitable learning strategies in complex and broad-scale networks, where RL may not be able to discover an optimal strategy in a reasonable time. However, in this model the neural network is periodically trained based on distinct experiences obtained during the interactions with the environment, which can be computationally demanding [36].

However, the system model assumptions of our work result in a uplink PC problem where the advantages of DRL do not become good enough to outweigh the critical disadvantage of its use. Hence, our RL-based solution presents lower signaling and lower computational complexity than DRL-basedtechniques. Moreover, RL-based techniques have more flexibility since they do not require an offline training stage, being able to provide real-time learning. Therefore, we turn our attention to a multi-agent RL-based solution combined with the CLPC paradigm that is compliant with the 5G NR technical specifications.

The main difference of the NR uplink PC framework compared to its predecessor LTE is on the use of multiple control loops associated with beams. Specifically, each control loop may be associated with a specific pair of transmit and receive beams. For instance, one electronic device may have a beam associated with two MIMO layers (or even more) so that the device can manage multiple control loops at the same time. Without proper coordination among the loops, the multiple processes produce sub-optimal power solutions [6].

To the best of our knowledge, despite the relevance of the beam-centric power control to the management of interference in NR networks, we have not found in the literature works investigating this problem using RL-based algorithms. To fill this gap, we formulate a beam-based uplink PC framework combined with multi-agent RL to take into account the coordination among multiple beams. The proposed framework can provide a solution that enables the system to learn what should be the most appropriate solution to the power allocation on the uplink based on the signaling defined by the 5G NR. We summarize the main contributions of our work as follows:

1) development of an uplink PC framework to the management of interference in 5G NR networks compliant with the technical specifications from 3GPP Release 15;
2) formulation of a beam-based transmit PC based on the principles of multi-agent RL;
3) development of a signaling scheme to allow cooperation among the entities endowed with intelligence in a NR multi-cell system;
4) comparison of the proposed uplink PC frameworks with two classical algorithms, namely the optimal solution power control (OSPC) and the soft dropping power control (SDPC), in terms of total transmit power, total data rate, and network energy efficiency.

Notation: bold lowercase and uppercase letters represent column vectors and matrices, respectively. $(\cdot)^T$ and $(\cdot)^H$ stand for transpose and Hermitian of a matrix, respectively. Calligraphic upper-case letters denote sets, and $|\cdot|$ denotes set cardinality. $\mathbb{E}\{\cdot\}$ denotes expectation operator.

III. System Model

We consider a multi-cell system with C cells. Each cell has one BS equipped with M antennas, and serves L UEs equipped with N antennas each. We assume the uplink transmission, and all cells share the same frequency band. The UEs inside a cell are synchronized with their respective BS and periodically measure their associated B beam pairs.

The discrete received signal model at the lth UE of the cth cell is represented as

$$y_{l,c} = \underbrace{\overline{\mathbf{w}}_{l,c}^H \mathbf{H}_{l,c} \bar{\mathbf{f}}_{l,c} P_{l,c} x_{l,c}}_{\text{useful signal}} + \underbrace{\sum_{\substack{l'=1 \\ l' \neq l}}^{L} \overline{\mathbf{w}}_{l,c}^H \mathbf{H}_{l,c} \bar{\mathbf{f}}_{l',c} P_{l',c} x_{l',c}}_{\text{intra-cell interference}} + \underbrace{\sum_{l'=1}^{L} \sum_{\substack{c'=1 \\ c' \neq c}}^{C} \overline{\mathbf{w}}_{l,c}^H \mathbf{H}_{l,c} \bar{\mathbf{f}}_{l',c'} P_{l',c'} x_{l',c'}}_{\text{inter-cell interference}} + \underbrace{\overline{\mathbf{w}}_{l,c}^H \mathbf{z}}_{\text{filtered noise}}, \quad (1)$$

where $\overline{\mathbf{w}}_{l,c} \in \mathbb{C}^{N \times 1}$ is the receive beamforming vector, $\mathbf{H}_{l,c} \in \mathbb{C}^{N \times M}$ is the channel matrix, $\bar{\mathbf{f}}_{l,c} \in \mathbb{C}^{M \times 1}$ is the transmit beamforming vector, $P_{l,c}$ is transmit power, $x_{l,c}$ is the transmitted symbol, and $\mathbf{z}$ is the additive white Gaussian noise vector with zero mean and variance ϱ^2.

We assume a narrow band block-fading channel, which is constant within a time-frequency resource block. The channel follows a geometric model with a limited number K of scatterers [39]. Each scatter contributes with a single path between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3074892, IEEE Transactions on Vehicular Technology

BS and UE. Therefore, the channel matrix $\mathbf{H}_{l,c} \in \mathbb{C}^{N\times M}$ between the BS and the lth UE of the cth cell can be written as

$$\mathbf{H}_{l,c} = \sqrt{\rho_{l,c}} \sum_{k=1}^{K} \beta_k \mathbf{v}_{\text{UE}}(\phi^{UE}_{k,l,c}, \theta^{UE}_{k,l,c}) \mathbf{v}^H_{\text{BS}}(\phi^{BS}_{k,l,c}, \theta^{BS}_{k,l,c}), \quad (2)$$

where $\rho_{l,c}$ denotes the path loss between the BS and the lth UE of the cth cell and β_k is the complex gain of the kth path; $\phi^{UE}_{k,l,c}, \phi^{BS}_{k,l,c} \in [0, 2\pi]$ are the angles of departure (AoD) at the BS and UE, respectively; $\theta^{UE}_{k,l,c}, \theta^{BS}_{k,l,c} \in [0, \pi]$ are the angle of arrival (AoA) at the BS and UE, respectively.

We assume uniform rectangular arrays (URAs) at the BS and UEs. There are M_v vertical antenna elements and M_h horizontal antennas elements, such that $M = M_v M_h$. The array response at the BS is expressed as

$$\mathbf{v}_{\text{BS}}(\phi^{BS}_{k,l,c}, \theta^{BS}_{k,l,c}) = \frac{1}{\sqrt{M}}[1, \ldots, e^{\jmath\left((M_v-1)\frac{2\pi\Lambda}{\lambda}\cos\theta^{BS}_{k,l,c} + (M_h-1)\frac{2\pi\Lambda}{\lambda}\sin\phi^{BS}_{k,l,c}\sin\theta^{BS}_{k,l,c}\right)}] \quad (3)$$

where Λ is the antenna element spacing, and λ is the wavelength. The array response at UE can be written similarly.

The transmit and receive beamforming follow the so-called hybrid structure and are defined as $\overline{\mathbf{w}}_{l,c} = \tilde{\mathbf{W}}\mathbf{u}$ and $\bar{\mathbf{f}}_{l,c} = \tilde{\mathbf{F}}\mathbf{v}$, respectively. The beamforming vectors are modeled according to discrete Fourier transform (DFT) matrices $\mathbf{W} \in \mathbb{C}^{N\times N}$ and $\mathbf{F} \in \mathbb{C}^{M\times M}$ at the receiver and transmitter, respectively. Let us define $\tilde{\mathbf{W}} \in \mathbb{C}^{N\times B}$ and $\tilde{\mathbf{F}} \in \mathbb{C}^{M\times B}$ as the truncated receiver and transmit beam codebooks containing the B selected beam pairs. We assume a beam management framework to determine the best set of B transmit-receive beam pairs, which determines the structure of $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{F}}$. The vectors $\mathbf{u} \in \mathbb{C}^{B\times 1}$ and $\mathbf{v} \in \mathbb{C}^{B\times 1}$ correspond to the dominant left and right singular vectors of the equivalent channel, defined as $\hat{\mathbf{H}}_{l,c} = \tilde{\mathbf{W}}^H \mathbf{H}_{l,c} \tilde{\mathbf{F}} \in \mathbb{C}^{B\times B}$.

We employ the beam sweeping scheme proposed in [40] to determine the set of the most suitable transmit-receive beam pairs. The suitability of a beam pair is determined according to the connectivity provided by the transmitter and receiver beam directions (each one identified by a beamforming vector). As it can be seen in Fig. 1, the beam sweeping operation covers a spatial area with a set of beams according to pre-specified intervals and directions. It is carried out an exhaustive search in a set of directions (each one identified by a beamforming vector) that covers the whole angular space. The BS sequentially transmits synchronization signal (SS) blocks, that compose a SS burst set, and each SS block can be mapped to a certain angular direction.

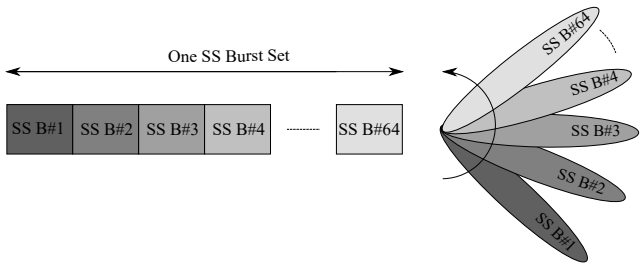


Fig. 1. Model of multiple time-multiplexed SS blocks within an SS burst set.

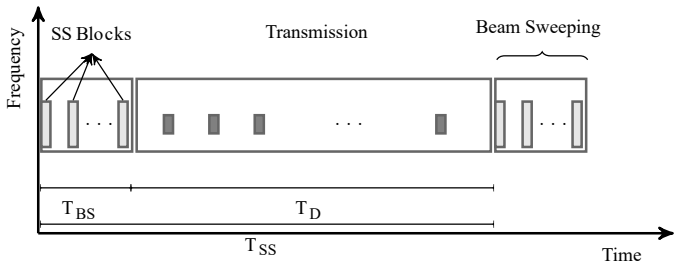


Fig. 2. Signaling model.

We consider a signaling period, of duration T_{SS}, divided into two time windows, as shown in Fig. 2. The first one contains a set of SS blocks with a duration T_{BS}. It is used to perform a beam sweeping procedure, that consists in the search of the best set of B transmit-receive beam pairs. The second window time is dedicated to data transmission using the selected beam pairs. This period has a duration T_D and each UE reports periodically the measured channel quality indicator (CQI) to the BS. The BS measures all possible combinations of transmit and receive beams from the codebooks $\mathbf{F}$ and $\mathbf{W}$, respectively, during the transmission of the SS blocks [40]. The selected beam pairs for kth signaling period between the BS and the lth UE of the cth cell are determined as the B highest values of υ_{b_r,b_t}, defined as

$$\upsilon_{b_r,b_t} = \frac{\|\mathbf{w}^H_{b_r} \mathbf{H}_{l,c,k} \mathbf{f}_{b_t}\|}{\varrho^2}, \quad (4)$$

where $\mathbf{w}_{b_r} \in \mathbb{C}^{N\times 1}$ is the b_rth column of the receive beam codebook $\mathbf{W}$ and $\mathbf{f}_{b_t}$ is the b_tth column of the transmit beam codebook $\mathbf{F}$. We assume that the angles of departure and arrival are constant over the beam sweeping period T_{BS}. It varies from 5 to 160 ms [6]. Therefore, the beam pairs remain constant within the time period T_D until the subsequent SS block arrives, when the beam pairs are re-evaluated.

Each individual uplink transmission is carried out from a specific antenna port, the identity of which is known by the device [6]. Each antenna port has its own specific reference signal, which is used by the device to estimate the CSI.

We define the SINR of the dth antenna port associated to the lth UE of the cth cell as

$$\Gamma_{d,l,c} = \frac{P_{d,l,c} |\overline{\mathbf{w}}^H_{l,c} \mathbf{H}_{l,c} \bar{\mathbf{f}}_{l,c}|^2}{\sum_{l'=1}^{L} \sum_{\substack{c'=1 \\ c'\neq c}}^{C} P_{l',c'} |\overline{\mathbf{w}}^H_{l,c} \mathbf{H}_{l,c} \bar{\mathbf{f}}_{l',c'}|^2 + \varrho^2}, \quad (5)$$

where the intra-cell interference mentioned in Eq. (1) is not considered since we assume a single active UE per time-frequency resource in a cell.

IV. Reinforcement Learning Background

RL is a ML paradigm that aims to discover the best behavior of a decision-maker entity, hereafter referred to as an agent, in an environment to optimize a function that measures the impact of the agent's decisions. We model the relationship between the agent and the environment using the concepts of state, action, and reward [41].

The state s_t in a discrete time step t is a value (or a set of values) that models the information that the agent has about the environment. The action a_t is an adjustment parameter used by the agent to interact with the environment. The reward r_t is a scalar function which indicates the immediate payoff from taking an action a_t in a state s_t [42].

Therefore, the interaction between the agent and the environment can be modeled as the transition from state s_t to s_{t+1}, restricted to the set $\mathcal{S}$ of all possible states. The transition is a consequence of an action a_t chosen in a set $\mathcal{A}$ of available actions and an associated reward r_{t+1}. Figure 3 shows a graphical representation of the interaction among these elements.

The goal of RL is the determination of the best policy, i.e, the most appropriate selection of actions according to the state of the environment. More specifically, the policy maps the perceived states of the environment to the action to be taken by the agent in those states. The agent finds its most desirable policy by taking into consideration the value of a state-action value function $Q(s_t, a_t)$. This function, also called Q-function, determines the overall expected discounted reward when starting in a state s_t and selecting an action a_t.

In this work, we adopt the Q-learning algorithm, which is an off-policy temporal difference algorithm initially proposed in [43]. This algorithm works by updating an estimate of the state-action value function based on the iterations of the agent with the environment. The state-action values are updated according to

$$\begin{aligned} &Q(s_t, a_t) = \\ &(1-\alpha)Q(s_t, a_t) \ + \alpha \left[r_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1}) \right], \end{aligned} \tag{6}$$

where $\alpha \in [0, 1]$ is the learning rate and $\gamma \in [0, 1]$ is the discount factor, which trades off the instantaneous and future rewards.

The learning process occurs through the balance between exploration, i.e., the sample of unseen parts of the state-action space, and exploitation of the accumulated knowledge [42]. We consider an adaptive ϵ-greedy algorithm strategy. Every time an agent takes an action a_t, it has a probability ϵ_t to be random (exploration) and a probability $1-\epsilon_t$ to select an action a_t based on previous experience (exploitation). The value of ϵ_t is gradually reduced over time from an initial value $\epsilon_{\max}$ until it reaches a minimum value $\epsilon_{\min}$.

The agent has to store the state-action values to be able to learn from the interactions with the environment. There are different mechanisms to represent these values, such as lookup tables or neural networks. For a more in-depth discussion on this topic, we refer the interested reader to [25]. In our work, we build a table $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to store the $Q(s_t, a_t)$ values. This mechanism has high requirements of computational memory in discrete state and action sets, but does not require any complex computational operation of training to store the acquired knowledge.

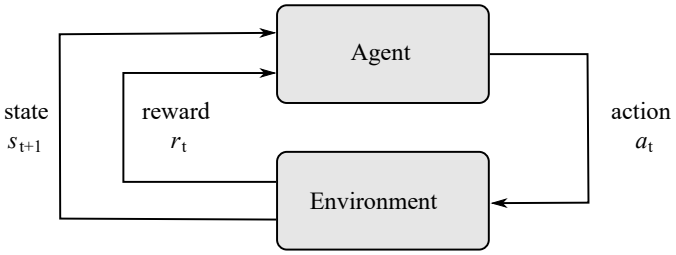


Fig. 3. Reinforcement learning interaction between elements.

V. Proposed Uplink Power Control Framework

In this section, we present the proposed beam-based uplink PC framework using multi-agent RL which is compliant to the 5G NR specifications [17], [18], [19], [20], [21], [22], [23]. The proposed approach consists of a joint power optimization of multiple antenna ports per UE using an RL-based technique since we consider multiple-antenna transmission/reception.

A. Uplink Power Control in 5G NR

NR uplink PC is the set of procedures that manage the transmit power of uplink physical channels, namely the physical uplink shared channel (PUSCH), the physical uplink control channel (PUCCH), and the physical random-access channel (PRACH), to guarantee suitable communication. We seek to determine the minimum transmit signal power necessary for appropriate decoding of the information conveyed through the physical channel. Furthermore, the uplink PC procedures must also limit the interference to other uplink transmissions. The transmit power control expressions of the uplink physical channels are very similar to each other. Their expressions are thoroughly detailed in [21]. The PUSCH is used for the transmission of uplink shared channel (ULSCH) data and control information. Thus, compared to PRACH and PUCCH, it presents a relation between power control and link adaptation that allows more flexibility. Consequently, the power control of the PUSCH has a greater scope of mechanisms and encompasses what can be done in PUCCH and PRACH.

Therefore, the uplink PC framework developed in our work considers the PUSCH expression Eq. (7) as the baseline power control. It can be concisely written, in dBm scale, as

$$\begin{aligned} P^{\text{PUSCH}} = \min\{&P_{\text{CMAX}}, P_0(\tau) + \varphi(\tau)\Upsilon(q) + \\ &10\log_{10}(2^{\mu}\Delta_{RB}) + \Delta_{TF} + \delta(\nu)\}, \end{aligned} \tag{7}$$

where

- P^{PUSCH} is the PUSCH transmit power;
- P_{CMAX} is the maximum allowed transmit power per carrier;
- $P_0(\tau)$ is the target received power;
- $\varphi(\tau)$ is the fractional path loss compensation factor;
- τ determines the transmission type, a network configurable parameter;
- $\Upsilon(q)$ is the estimation of the uplink path loss;
- q is the reference signal (RS) index;
- μ is the subcarrier spacing parameter;
- Δ_{RB} is the bandwidth of the resource assignment;
- Δ_{TF} models the required received power according to number of resource bits per resource element which also depends on the modulation scheme and channel coding rates;
- $\delta(\nu)$ is the power adjustment due to the closed loop power control;

- ν determines the closed loop process index.

Equation (7) indicates that the PUSCH power control is composed of OLPC and CLPC mechanisms. In the OLPC, the UE transmit power is adjusted according to estimates of the uplink path loss based on downlink measurements, as it can be seen by the expression $P_0(\tau)+\varphi(\tau)\cdot\Upsilon(q)$. The transmit power is adjusted to achieve the target $P_0(\tau)$, which is a configurable network parameter regulated to provide a target data rate, given the noise and interference levels at the receiver. The UE estimates the downlink path loss $\Upsilon(\tau)$ using the RS index τ for the active downlink. A UE does not simultaneously maintain more than four path loss estimates per transmission [21].

The parameter $\varphi(\tau)$ determines the compensation level of the path loss $\Upsilon(q)$. The full path loss compensation, which is done when we assume $\varphi(\tau) = 1$, ensures that the received SINR matches the requirement for the modulation and coding scheme (MCS) selected by the network, assuming that the UE transmit power does not reach its maximum value. On the other hand, a fractional path loss compensation, which arises when $0 < \varphi(\tau) < 1$, requires a lower transmit power, implying less interference to the other cells. However, the received power, and consequently, the SINR, decreases as the path loss increases. The data rate is also reduced to compensate this effect by switching to a lower MCS. The benefit of fractional path loss compensation is the reduced interference to neighbor cells. However, this benefit comes at the price of more significant variations in the service quality, with reduced data rate availability for UEs closer to the cell border. In this work, we assume a full path loss compensation to ensure service quality stability, requiring a CLPC mechanism to manage the uplink interference.

The transmit power should be proportional to the bandwidth assigned for the transmission, as indicated by the term $10\cdot\log_{10}(2^{\mu}\cdot\Delta_{RB})$ in Eq. (7). The transmit power must be proportional to the size of the resource block, where the subcarrier width is defined as $\Delta f = 2^{\mu}\cdot 15$ kHz. We assume a fixed bandwidth for the PUSCH transmission. Therefore, this term can be omitted from the power control expression.

The term Δ_{TF} models the impact of the number of bits per resource element and channel coding rates on the transmit power. This model is expressed as

$$\Delta_{TF} = 10\log(2^{1.25\eta}-1) + 10\log(\varrho), \tag{8}$$

where η is the number of information bits per resource element; ϱ models the impact of data transmission on PUSCH, and $\varrho = 1$ when the PUSCH includes ULSCH data. In our work, we assume that the PUSCH received power is matched to a certain MCS given by the selected value of $P_0(\tau)$. In this case, according to [1], we set Δ_{TF} to zero.

In the CLPC, the UE transmit power is adjusted according to power control commands provided by the network. This regulation is determined from prior network measurements of the received uplink power [6]. The term $\delta(\nu)$ defines the power control command and in the closed loop solution. Such commands are carried out in the transmit power control (TPC) field within uplink scheduling grants (downlink control information (DCI) formats $0-0$, $0-1$, and $2-2$ [21]). Each power control commands consists of 2 bits corresponding to four different steps: −1 dB, 0 dB, +1 dB, and +3 dB. These steps are associated with TPC command field values 0, 1, 2 and 3, respectively. Each command specifies the value in dB that a UE should add to its current transmit power.

The main extension of NR compared to LTE is the possibility for a beam-based power control. The parameters τ, q, and ν in Eq. (7) are associated with beam pairs for uplink PC. For instance, the uplink path loss estimate $\Upsilon(q)$ should reflect the path loss of the uplink beam pair q to be used for the PUSCH transmission. The network uses a set of downlink reference signals (CSI-RS, SS block) to estimate the path loss for a specific value of q, and each UE is limited to monitor up to $q_{\max} = 4$ parallel path loss estimation processes [21]. The network also provides a mapping from the possible sounding reference signals resource indicator (SRI) values provided in the scheduling grant to the different values of q. After a beam management process to determine the beam pairs, and consequently, the corresponding antenna ports, the path loss estimate is then used in the power control expression. Consequently, the parameters q and ν are directly associated to the antenna port index d.

Therefore, based on the assumptions previously described, the transmit power of the dth antenna port of the lth UE in the cth cell is represented as a simplification of Eq. (7) and can be written as

$$P^{\text{PUSCH}}_{d,l,c} = \min\{P_{\text{CMAX}}, P_0(\tau) + \varphi_{l,c}(\tau)\Upsilon_{l,c}(d) + \delta_{l,c}(d)\}. \tag{9}$$

For the reader interested in more details, please refer to [17], [18], [19], [20], [21], [22], [23] and references therein.

B. RL-Based Uplink PC Design

In our framework, each cell regards its BS as an agent, and the remaining of the system (other BSs, UEs, and ULSCH) represents the environment. The behavior of the Q-learning algorithm depends not only on the BS actions, but also on the actions taken by neighboring BSs, since all cells are assumed to operate at the same frequency and therefore suffer from inter-cell interference. In other words, there is states, actions, and rewards of different cells that are coupled, and influence one another. This is the main reason for considering a multi-agent RL approach in this work.

1) State Mapping: The state of the BS of the cth cell is a tuple $s_{t,c} = \{P_{1,c}, \cdots, P_{D,c}\}$ of powers associated with each antenna port. The index l is omitted for notation simplicity, since only one UE per resource block is considered. These powers are modeled as discrete values in dBm and are limited to minimum and maximum values, i.e., $s_{t,c} \in S_c = \{\{P^{\min}_{1,c}, \cdots, P^{\min}_{D,c}\}, \cdots, \{P^{\max}_{1,c}, \cdots, P^{\max}_{D,c}\}\}$. The power limits for the dth antenna port are defined according to

$$P^{\min}_{d,c} = P^{\text{SINR}}_{d,c} - \varphi_c(\tau)\Upsilon_c(d), \tag{10}$$

$$P^{\max}_{d,c} = P_{\text{CMAX}} - \varphi_c(\tau)\Upsilon_c(d), \tag{11}$$

where the power $P^{\text{SINR}}_{d,c}$ assures the minimum SINR, $\Gamma^{\min}_{d,c}$.

2) Action Mapping: The action of the cth agent is modeled as TPC commands $a_c = \{\delta_1, \cdots, \delta_D\}$ sent to the UE to update the transmit power of the antenna ports. Each TPC command δ_d is limited to the set $\{-1, 0, +1, +3\}$ defined in [21, Table 7.1.1-1]. In our model, each BS sends the commands to the associated UE and the transmit power of all antenna ports $\{P_{1,c}, \cdots, , P_{D,c}\}$ are updated simultaneously.

3) Reward Mapping: We consider a reward function based on a performance indicator of the network to determine the effects of the variation of the power level of the antenna ports. Note that the reward will be influenced by the SINR and by the level of cooperation between neighbor cells. The reward function is a convex sum of the cell's data rates, and can be written as

$$r_{t,c} = \rho_c \sum_{l=1}^{L} \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,l,c}) \; + \sum_{\substack{c'=1 \\ c' \neq c}}^{C} \rho_{c'} \left(\sum_{l=1}^{L} \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,l,l'}) \right), \tag{12}$$

where Ω is the resource block's bandwidth and $\rho_c \in [0, 1]$ is a weight factor that determines how much the data rate of the cth cell impacts the reward. These values are also shared between cells and they sum up one, i.e., $\sum_{c=0}^{C} \rho_c = 1$.

This expression represents a weighted average of the system's capacity taking into account all cells. In turn, the weights of this average represent the degree of importance that the agent considers for a given cell. The first term of Eq. (12) represents the performance indicator parameter related to the relationship between UEs and BS of a given cell. In other words, it describes the direct impact of the entity endowed with intelligence (the BS is the decision maker) on the associated UEs in the cell where it is the main entity. The second term of Eq. (12) measures the impact of the agent decision on the data rate of the other cells of the network. Therefore, a given action could be measured as beneficial if the data rate of the other cell increases. Thus, strategies that result in a reduction of the multi-user interference become interesting to the agents, even if the data rate of its cell remains the same or presents a reduction.

After an initial parameter configuration to define the power limits described by Eqs. (10) and (11), the process of taking actions and calculating rewards is carried out. The mapping between the states, actions and rewards is initially performed in an exploratory fashion. This means the BSs and UEs exchange commands and rewards in order to learn the relationship between the set of actions $\mathcal{A}$ and of states $\mathcal{S}$ to the observed rewards described by Eq. (6). Once this mapping is completed, the BS exploits it to choose the appropriate TPC command and send it to each UE. If any change in the scenario occurs, the mapping is updated, and another solution is provided. Such update requires minimal signaling.

We assume an adaptive ϵ-greedy algorithm where the value of ϵ is gradually reduced over discrete time steps t according to

$$\epsilon_t = \frac{\epsilon_{\max}}{\epsilon_{\max} + \xi t}, \tag{13}$$

where $\epsilon_{\max}$ is the initial exploration rate, and ξ is a fixed parameter that guarantees a given value to ϵ in a defined time step t. In the beginning of the each experiment, the agent explores intensely the state-action space and updates its matrix $[\mathbf{Q}]_{s_t,a_t} = Q(s_t, a_t)$. In the end, the probability of exploitation is higher than a minimum predetermined threshold, e.g., 90%.

4) RL-Based Signaling Scheme: Figure 4 represents the signaling scheme of the proposed multi-agent RL-based uplink PC framework when operating in a multi-cell scenario. Each cell has one BS that serves one UE per resource. At the step (1), each cell c defines the dimensions and the initial assessments of the matrix $Q_c \in \mathbb{R}^{|\mathcal{S}_c| \times |\mathcal{A}_c|}$, which stores the state-action values $Q(s_t^c, a_t^c)$ resultant of the interactions of the cth BS with the environment. In our model, we initialize the system considering matrix Q_c with all its entries equal to zero. That is, the system starts its learning process completely dummy, without any indication of how to proceed. We deliberately chose this condition to show that the algorithm is capable of learning and its results are very close to the optimal solution, as it can be seen in the performance evaluation in Section VI. In addition, this behavior shows that the algorithm is capable of learning in real time, even starting from a condition without any knowledge. This shows its great applicability in real systems, where online learning is a particularly important feature.

As a initial step, we need to define the spaces of states and actions, i.e., the sets $\mathcal{S}_c$ and $\mathcal{A}_c$, respectively. The cardinality of the set of actions $\mathcal{A}_c$ is a function of the number of TPC commands and antenna ports, since each action a_t^c is defined as a set of TPC commands sent to the antenna ports. Thus, it can be written as

$$|\mathcal{A}_c| = \iota^D \tag{14}$$

where ι is the number of TPC commands.

The cardinality of the set of states $\mathcal{S}_c$ is defined according to the number of power intervals and the number of antenna ports. The number of power intervals is a function of the step size among transmit power levels and the power limits $P_{d,c}^{\min}$ and $P_{d,c}^{\max}$. Thus, the cardinality is given by

$$|\mathcal{S}_c| = \left(\frac{P_{d,c}^{\max} - P_{d,c}^{\min}}{\chi} \right)^D \tag{15}$$

where χ is the size of the power step.

At the step (2), each BS sends a set of TPC commands to the associated UE. We assume uplink scheduling grant according to the DCI format $0-1$, where there are 2 bits reserved to adjust PUSCH transmission power [6]. We consider the format $0-1$ since it supports multi-antenna fields, like number of antenna ports, SRI, and sounding reference signals (SRS) request. These values are defined according to the ϵ-greedy algorithm. Hence, each agent determines an action modeled by the tuple $a_c = \{\delta_1, \cdots, \delta_D\}$ which defines the update of the UEs' transmit power levels. At the step (3), each BS observes the new power of its associated UE according to the SRS transmission. Then, at step (4), the BSs share their uplink measurements based on the SRS transmissions of the associated UE. Based on these measurements, at step (5), each BS calculates the reward associated with the action taken at step (2). Then, each BS updates the mapping between the spaces of actions and rewards at step (6). Finally, at step (7), based on the updated mapping, each BS determines the next TPC commands, i.e., the next update of the power of the associated UE.

In the following, Algorithm 1 summarizes the main steps of the proposed RL-based framework. These instructions are carried out independently by each agent. The main loop (instructions between lines 3 and 10) determines the sequential

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3074892, IEEE Transactions on Vehicular Technology

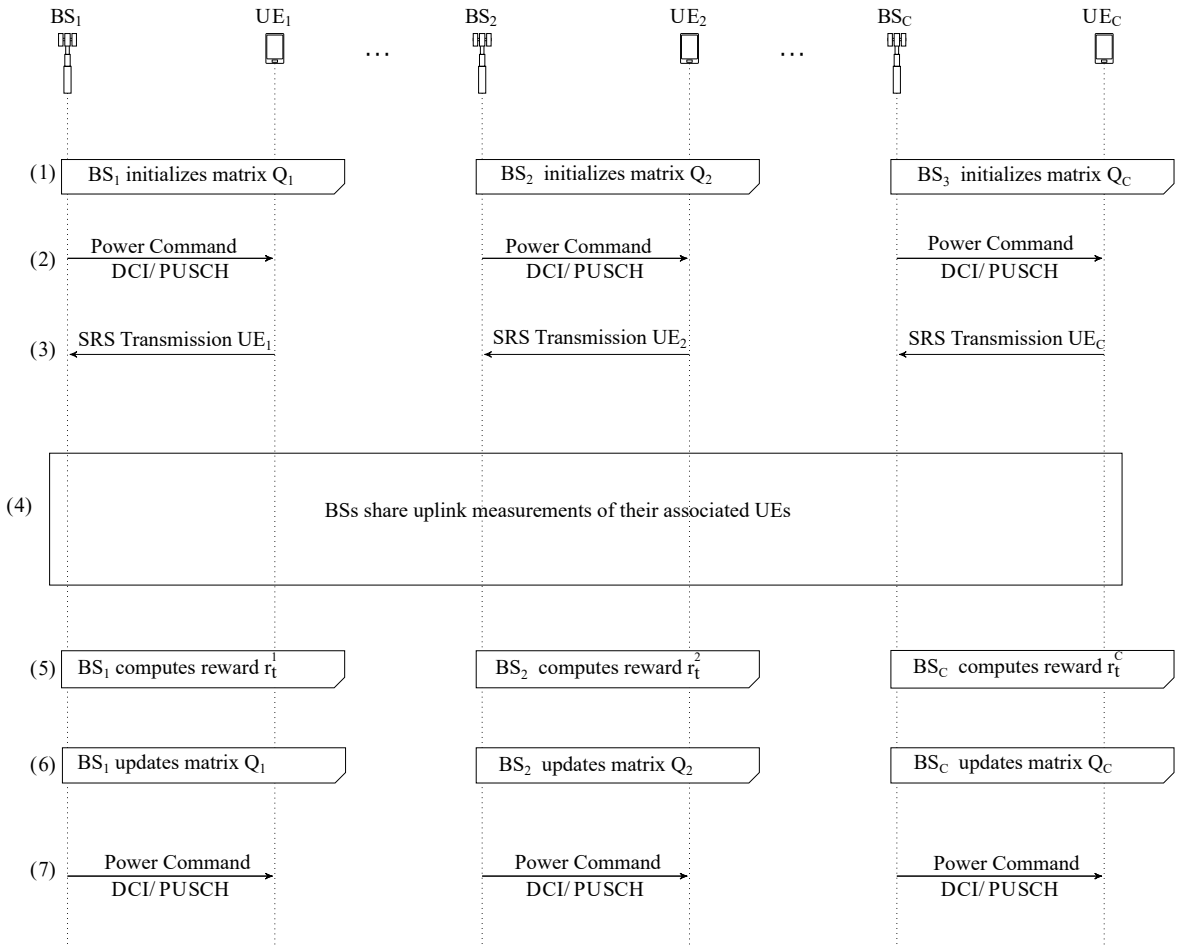


Fig. 4. Representation of the proposed uplink PC framework considering reception/transmission of measurement/information from/to another BS.

selection of actions according to the ϵ-greedy policy. The instruction with the greatest computational complexity is defined at line 9, as is requires the search for the maximum value in the matrix Q_c. We define the stop condition when t is equal to the number of iterations T. This approach is also considered in the comparison algorithms. Thus, the computational complexity of the proposed framework based on the pseudo-code Algorithm 1 is $O(TC|\mathcal{S}_c||\mathcal{A}_c|)$. For more details see Appendix A.

Algorithm 1 Proposed RL-based power control.

1: initialize $t = 0$;
2: sort initial power levels $s_{t,c} = \{P_{1,c}, \cdots, P_{D,c}\} \in \mathcal{S}_s$;
3: **while** stop condition not reached **do**
4: sort an uniform random number $e \in [0, 1]$;
5: determine ϵ_t according to Eq. (13);
6: **if** $(e < \epsilon_t)$ or $(t = 0)$ **then**
7: select randomly an action $a_{t,c} = \{\delta_1, \cdots, \delta_D\} \in \mathcal{A}_c$;
8: **else**
9: select the action $a_{t,c} = \{\delta_1, \cdots, \delta_D\} \in \mathcal{A}_c$ with the maximum $Q_c(s_{t,c}, a_{t,c})$;
10: **end if**
11: compute the transmit power update defined by the action $a_{t,c}$;
12: verify if the antenna port power limits $P_{d,c}^{\min}$ and $P_{d,c}^{\max}$ are respected;
13: execute the transmit power updates;
14: calculate the associated reward $r_{t,c}(s_{t,c}, a_{t,c}, s_{t+1,c})$ according to Eq. (12);
15: update the matrix Q_c with the $Q_c(s_{t,c}, a_{t,c})$ value according to Eq. (6);
16: $t = t + 1$;
17: **end while**

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed uplink PC framework. We consider a simulation scenario with three cells, where each cell has one BS and one associated UE, as can be seen in Fig. 5. We also consider a scenario with two antenna ports. Each UE moves along a linear track towards the cell border at a constant speed and a random direction. The cells share the same frequency band composed of 12 subcarriers, where orthogonal uplink transmission is assumed. The subcarrier spacing is 120 kHz, since we assumed a numerology based on $\mu = 3$, and the frequency carrier is 28 GHz. In this numerology, one time slot has a duration of 0.125ms. In this work, a time slot is also referred to as an iteration. The path loss follows the urban macro (UMa)-non-line of sight (NLOS) model [44, Table 7.4.1-1]. The shadowing is modeled as log-normal distributed with a standard deviation of 4 dB. The noise

TABLE I
GENERAL SIMULATION PARAMETERS.

Parameter	Value
Inter site distance	200 m
Min. dist. BS-UE (2D)	25 m
Angle sector	60°
BS height	15 m
UE height	1.5 m
UE track	linear
UE speed	5 km/h
BS antenna model	omnidirectional
BS antennas	8×8
UE antenna model	omnidirectional
UE antennas	2×2
Max. transmit power per carrier	24 dBm
Carrier frequency	28 GHz
Bandwidth	1.44 MHz
Number of subcarriers	12
Subcarrier spacing	120 kHz
Number of subframes	10
Number of symbols	14
Azimuth angle range	$[-60°, 60°]$
Elevation angle range	$[60°, 120°]$
Number of paths	10
Simulation rounds	100

power is modeled as $10\log_{10}(290 \cdot 10^{-23} \cdot \Omega)$ dBm. The main simulation parameters are listed in Table I.

We assume a UE power class 3, i.e., the maximum uplink transmit power per carrier is defined as $P_{\text{CMAX}} = 24$ dBm [45]. Therefore, each antenna port can assume one value from a total of 25 discrete power levels, separated in steps of 1 dBm. The state of the agent of the BS at the cth cell can be written as $s_{t,c} \in S_c = \{\{P^{\min}_{1,c}, P^{\min}_{2,c}\}, \cdots, \{P^{\max}_{1,c}, P^{\max}_{2,c}\}\}$. The maximum and minimum power limits are described by Eqs. (10) and (11). Notice that the sum of the powers of the antenna ports cannot exceed the maximum power P_{CMAX}

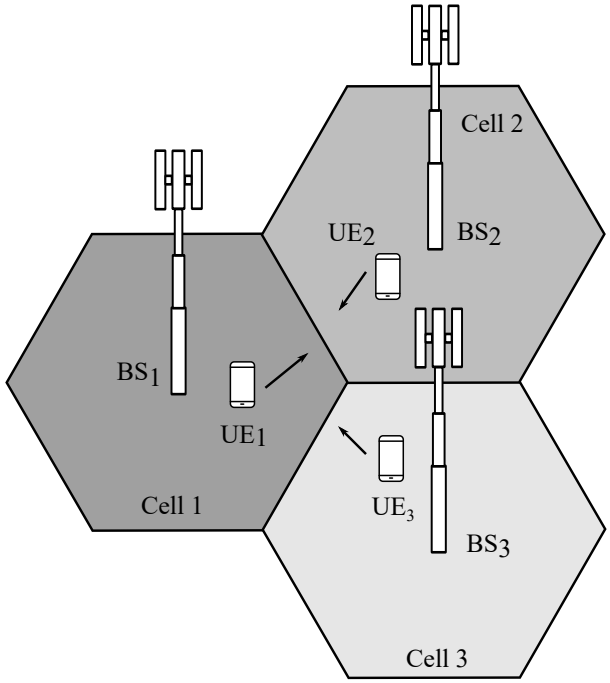


Fig. 5. Simulation scenario.

for each UE. The Q-learning algorithm determines the best policy for adjusting the transmit power of each antenna port, defined by $\{\delta_1, \delta_2\}$. Recall that we consider TPC in the set $\{-1, 0, +1, +3\}$ [21, Table 7.1.1-1].

Therefore, each cell implies an environment that has a total of $25^2 = 625$ states and each agent has $4^2 = 16$ actions. We assume an adaptive ϵ-greedy algorithm where the value of ϵ is gradually reduced over discrete time steps t according to Eq. (13). The initial exploration rate $\epsilon_{\max}$ is defined as 0.95, and ξ is a fixed parameter that guarantees $\epsilon_t = 0.50$ when $t = 2,000$. In the beginning of the each experiment, the agent explores intensely the state-action space and updates its Q-table. In the end, the probability of exploitation is higher than 90%. We summarize the Q-learning parameters in Table II.

We evaluate the proposed uplink PC solution in terms of (i) total uplink transmit power, (ii) total data rate, and (ii) network energy efficiency. The data rate of the cth cell, based on the Shannon's capacity formula, can be expressed as

$$\text{DR}_c = \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,c}), \tag{16}$$

where $\Gamma_{d,c}$ corresponds to $\Gamma_{d,1,c}$ since each cell has only one UE.

The network energy efficiency is the ratio between the total data rate and the total uplink transmit power. It can be written as

$$\text{EE} = \frac{\sum_{c=1}^{C} \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,c})}{\sum_{c=1}^{C} \sum_{d=1}^{D} P_{d,c}}, \tag{17}$$

where $P_{d,c}$ is the $P^{\text{PUSCH}}_{d,c}$ associated with the UE from the cth cell and the dth antenna port updated according to Eq. (9).

A. Evaluation of the Proposed RL-Based Uplink PC Design

Initially, we evaluate the behavior of the proposed RL-based uplink PC design. Our analysis focus on the impact of the reward function in the determination of the power management policy. According to Eq.(12), the reward function is defined as the sum of the data rate of each cell DR_c weighted by the design parameter ρ_c. In order to reduce the degree of freedom in our problem, we re-write the reward function as a convex sum of data rates, where the first term contains the data rate of the cell whose agent directly manages the transmit power and the second term accounts the data rate of the other cells in the system. This adaptation aims to simplify the determination of the best set of design parameters and to reduce the sensitivity of the learning system. Thus, the reward

TABLE II
MACHINE LEARNING PARAMETERS.

Parameter	Value
Number of iterations	21,000
Discount factor (γ)	0.10
Learning rate (α)	0.20
Initial exploration rate ($\epsilon_{\max}$)	0.95
Number of states	625
Number of actions	16

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3074892, IEEE Transactions on Vehicular Technology

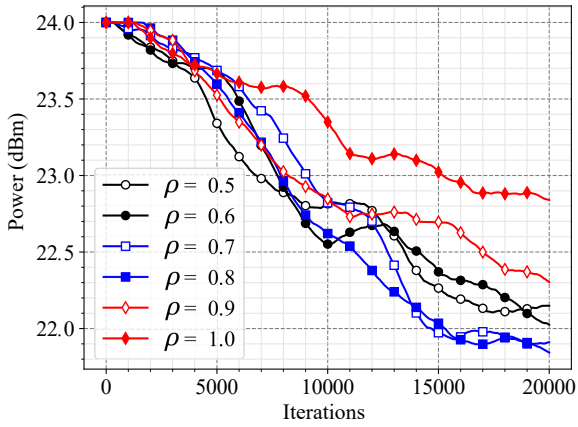
functions of the agents in our simulation model can be written as

$$r_{t,c} = \rho \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,c}) \; + (1-\rho) \sum_{\substack{c'=1 \\ c' \neq c}}^{C} \sum_{d=1}^{D} \Omega \log_2(1 + \Gamma_{d,c'}), \tag{18}$$

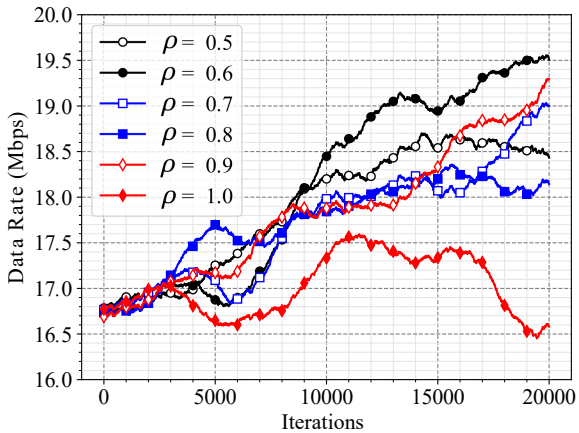
where ρ is the design parameter that regulates how the agent behavior is impacted by the gains or losses of the data rates of the other cells. Moreover, it also determines the level of cooperation among the agents in the learning process.

Figure 6a examines the total transmit power (in dBm) as a function of the number of iterations. We apply a simple moving averaging (SMA) with a window of 500 iterations to smooth the curves. All BSs operate their agents simultaneously according to the reward functions described by Eq. (18), a simplification from Eq. (12). The proposed uplink PC framework decreases the total transmission power over the iterations in comparison with the initial level from 1 dBm to 2 dBm, varying according to the value of the design parameter ρ. Therefore, the learning process resulting from the interaction with the system promotes different energy efficient resource management policies.

The behavior of the total transmit power varies according to the value of the design parameter ρ. On the one hand, when $\rho = 1.0$, the reward expression considers only the data rate of the cell where the BS is located. According to Q-learning principles, the dynamic of iteration between agent and environment is thought to maximize the long term reward. Consequently, the algorithm attempts to maximize the data rate of each cell. Based on Eq. (5), the transmit power is the only parameter controlled by the BS able to modify the SINR and, consequently the data rate. Therefore, the algorithm increases the power of all UEs indistinctly. Each UE updates its power autonomously, without any explicit observation of how its behavior affects the remaining of the system. Consequently, this parameter configuration reaches the highest power levels, namely 23 dBm. On the other hand, when $\rho < 1.0$, each agent seeks to maximize the weighted sum of the data rates, i.e., the algorithm also considers as advantageous an improvement in the data rate of the other cells. In that case, the uplink power stabilizes at lower levels. In comparison with $\rho = 1.0$, there is a reduction of 1 dBm in the total transmit power transmission when $\rho = 0.8$. This is the most efficient parameter setting considering only the reduction of transmit power. The design parameter $\rho \leq 0.6$ results in the most cooperative behavior of the system, which efficiently reduces the interference, despite the small increment of the transmit power.



(a) Total transmit power



(b) Total data rate

Fig. 6. Evaluation of the proposed RL-based uplink PC design considering different values of the design parameter ρ.

Figure 6b shows the total data rate (in Mbps) as a function of the number of iterations. The behavior observed when $\rho = 1.0$ promotes a high interference scenario, which reduces the SINR levels and jeopardizes the total data rate. The cooperation among agents, even if reduced, is capable of significantly improving the system conditions. The cooperation level parameter $\rho = 0.9$ promotes an enhancement of 15% of the total data rate in comparison with $\rho = 1$. The best performance of the system in terms of total rate is seen when $\rho = 0.6$, which increases in 20% the total data rate in comparison with the worst case.

Figure 7 indicates the average network energy efficiency in the last 2,000 iterations achieved by the proposed RL-based uplink PC considering different values of ρ, i.e., different levels of cooperation among agents. We observe higher levels of the network energy efficiency when $\rho < 1$. Considering this parameter setting, the proposed uplink PC framework reduces the transmit power and increases the total data rate. However, it requires an exchange of information between cells. That is, the enhancement of the network efficiency in at least 20% comes at the cost of a signaling exchange among BSs. The network energy efficiency gains with $\rho = 0.60$ are 40% higher than when $\rho = 1$. Therefore, the design parameter $\rho = 0.60$ achieves the highest network energy efficiency in the considered scenario.

Note that the determination of the most suitable design parameter value ρ to obtain the highest levels of network energy efficiency depends on the simulation scenario, being predominant the UE's location in the cell. On one hand, UEs located at the cell's edge experience high inter-cellular interference. In this case, lower rates of ρ represent the best choice, as this configuration seeks to balance reward's objectives more cooperatively, leading to lower levels of interference. On the other hand, UEs close to the BS are less affected. In this case, higher sigma rates can be chosen without the negative effects described above, since the increase in power does not significantly increase the levels of inter-cellular interference.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3074892, IEEE Transactions on Vehicular Technology

We have performed an exhaustive search for the parameter ρ to determine its behavior in a given scenario and to validate the proposed technique. However, from a practical point of view, the development of an adaptive strategy for this parameter would be worth to investigate. This is one of our prospects for future works.

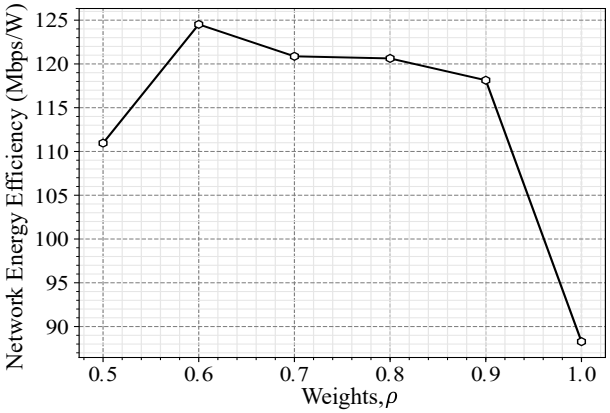


Fig. 7. Network energy efficiency achieved by the proposed RL-based uplink PC considering different values of the design parameter of cooperation among agents ρ.

B. Comparison with Classical Algorithms

In the following, we compare the performance of the proposed uplink PC framework with two classical solutions found in the literature.

1) Optimal Solution: the authors of [46] extended the results obtained in [47] to the power minimization problem with SINR constraints and fixed transmit-receive filters. This problem can be formally written as

$$\begin{aligned} &\max \min_{d,c} \ \frac{\Gamma_{d,c}}{\Gamma_{d,c}^{\text{Target}}} \\ &\text{s.t} \quad \sum_{c=1}^{C}\sum_{d=1}^{D} P_{d,c} \leq P_{\text{CMAX}} \end{aligned} \tag{19}$$

where $\Gamma_{d,c}^{\text{Target}}$ is the SINR target at the dth antenna port of the UE at the cth cell. In our simulations, the highest feasible SINR target is defined as $\Gamma_{d,c}^{\text{Target}} = \Gamma_{d,c}^{\max} = 6$ dB

The authors of the aforementioned articles prove that the optimal power allocation at the cth cell that maximizes the SINRs is provided by the dominant eigenvector of the matrix Λ_c, which can be written as

$$\Lambda_c = \begin{bmatrix} \mathfrak{D}_c \Psi_c^T & \mathfrak{D}_c \boldsymbol{\sigma} \\ \dfrac{1}{P_{\text{CMAX}}} \mathbf{1}^T \mathfrak{D}_c \Psi_c^T & \dfrac{1}{P_{\text{CMAX}}} \mathbf{1}^T \mathfrak{D}_c \boldsymbol{\sigma} \end{bmatrix} \tag{20}$$

where $\mathfrak{D}_c = \text{diag}\left\{ \dfrac{\Gamma_{1,c}^{\text{Target}}}{\|\overline{\mathbf{w}}_c^H \mathbf{H}_c \bar{\mathbf{f}}_c\|^2}, \dfrac{\Gamma_{2,c}^{\text{Target}}}{\|\overline{\mathbf{w}}_c^H \mathbf{H}_c \bar{\mathbf{f}}_c\|^2} \right\}$, $\boldsymbol{\sigma}$ is a vector of noise powers on all antenna ports, $\mathbf{1}$ is a vector of ones with appropriate dimension. The coupling matrix Ψ_c is defined as

$$[\Psi_c]_{b_r,b_t} = \begin{cases} \|\mathbf{w}_{b_r}^H \mathbf{H}_{l,c,k} \mathbf{f}_{b_t}\|^2, & \text{if } b_t \neq b_r. \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Algorithm 2 Optimal solution power control.

1: initialize $t = 0$;
2: define SINR targets of all antenna ports in all cells $\Gamma_{d,c}^{\text{target}}$;
3: **while** stop condition not reached **do**
4: define extended coupling matrix Λ_c;
5: perform the eigendecomposition of the extended coupling matrix Λ_c;
6: verify if the antenna port power limits $P_{d,c}^{\min}$ and $P_{d,c}^{\max}$ are respected;
7: execute the transmit power updates;
8: $t = t + 1$;
9: **end while**

Algorithm 2 summarizes the main steps of the OSPC algorithm. These instructions are carried out independently by each agent. The main loop (instructions between lines 3 and 9) determines the sequential selection of actions according to the eigen system discussed previously. The computational complexity of this strategy based on the pseudo-code Algorithm 2 is $O(TCD^2MN) + O(TCD^3)$. For more details see Appendix B.

2) Soft Dropping Power Control: initially proposed in [48], it is an iterative power allocation algorithm which promotes a self-regulation of the target SINR according to the transmit power and channel conditions. The transmit power $P_{d,c}^{\text{PUSCH}}(t+1)$ associated with the UE from the cth cell and the dth antenna port at the $(t+1)$th iteration is updated according

$$P_{d,c}^{\text{PUSCH}}(t+1) = P_{d,c}^{\text{PUSCH}}(t) + \zeta[\Gamma_{d,c}^{\text{Target}}(t) - \Gamma_{d,c}(t)], \tag{22}$$

where $\Gamma_{d,c}^{\text{Target}}$ is a power-dependent target SINR updated according to

$$\Gamma_{d,c}^{\text{Target}} = \min\left\{ \max\left\{ \left(\frac{\Gamma_{d,c}^{\max} - \Gamma_{d,c}^{\min}}{P_{d,c}^{\min} - P_{d,c}^{\max}} \right) P_{d,c}^{\text{PUSCH}}(t) + \Gamma_{d,c}^{\min}, \Gamma_{d,c}^{\min} \right\}, \Gamma_{d,c}^{\max} \right\}, \tag{23}$$

where $\Gamma_{d,c}^{\min}$ and $\Gamma_{d,c}^{\max}$ are lower and upper limits, respectively, such that $\Gamma_{d,c}^{\text{Target}} \in [\Gamma_{d,c}^{\min}, \Gamma_{d,c}^{\max}]$. The feasible SINR limits are defined through simulations as $\Gamma_{d,c}^{\text{Min}} = 0$ dB and $\Gamma_{d,ic}^{\text{Max}} = 6$ dB. Furthermore, aiming at the convergence of the algorithm, we define $\zeta = (1-\beta)^{-1}$ [49], [50], where β is defined as

$$\beta = \frac{\log_{10}(\hat{\Gamma}_{d,c}^{\text{Min}} / \hat{\Gamma}_{d,c}^{\text{Max}})}{\log_{10}(\hat{P}_{d,c}^{\text{Max}} / \hat{P}_{d,c}^{\text{Min}})} \tag{24}$$

where $\hat{\Gamma}_{d,c}^{\text{Min}}$ and $\hat{\Gamma}_{d,c}^{\text{Max}}$ are the minimum and the maximum SINR in linear scale, respectively; $\hat{P}_{d,c}^{\text{Min}}$ and $\hat{P}_{d,c}^{\text{Max}}$ denote the minimum and the maximum transmit power in linear scale, respectively. We randomly sort the initial power at each Monte Carlo simulation, i.e., $P_{d,c}^{\text{PUSCH}}(0) \in [P_{d,c}^{\min}, P_{d,c}^{\max}]$.

Algorithm 3 summarizes the main steps of the SDPC algorithm. These instructions are carried out independently by each agent. The computational complexity of this strategy based on the pseudo-code Algorithm 2 is $O(TCD)$. For more details see Appendix C.

Figure 8a shows the behavior total transmit power (in dBm) as a function of the number of iterations of the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3074892, IEEE Transactions on Vehicular Technology

Algorithm 3 Soft dropping power control.

1: initialize $t = 0$;
2: define the initial transmit power for each antenna port in each cell;
3: **while** stop condition not reached **do**
4: calculate the target SINR $\Gamma_{c,d}^{\text{target}}$ according to Eq. (23);
5: update the transmit power according to Eq.(22);
6: verify if the antenna port power limits $P_{d,c}^{\min}$ and $P_{d,c}^{\max}$ are respected;
7: execute the transmit power updates;
8: $t = t + 1$;
9: **end while**

RL-based uplink PC framework, hereafter referred to as RL-based power control (RLPC), compared with the two classical solutions. In the initial iterations, the RLPC presents the highest total transmit power. It exceeds the levels obtained by SDPC by 0.5 dBm. However, this disadvantage is reversed as the agents interact with the environment and the knowledge acquired is used in the decision making. At the end of the simulation, the RLPC outperforms the SDPC, with a significant reduction of the total transmit power (1.5 dBm), and its transmit power levels approach that observed with OSPC.

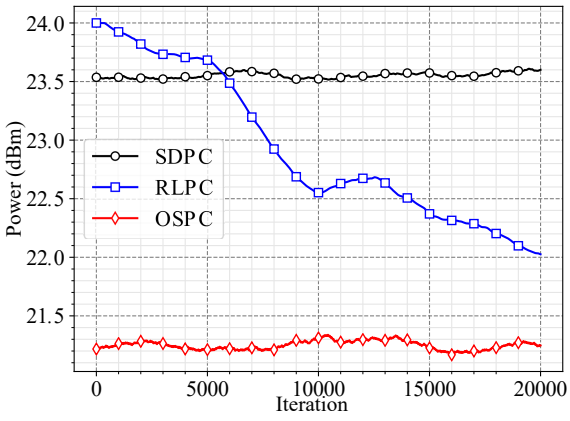
Figure 8b shows the total data rate (in Mbps) as a function of the number of iterations. The proposed RLPC has a slower convergence, but finds a power solution able to reduce the interference and increase the SINR. Consequently, we observe a continuous increment of the total data rate, that enhances 20% compared with SDPC and approaches the optimal solution.

Figure 8c depicts the behavior of the network energy efficiency as a function of the number of iterations. The SDPC has a static behavior, i.e., it is not able to learn new strategies from the interaction with the environment. The RLPC provides a self-exploratory energy-efficient solution which enhances its network energy efficient approximately 95%, achieving 75% of the performance of the optimal solution.

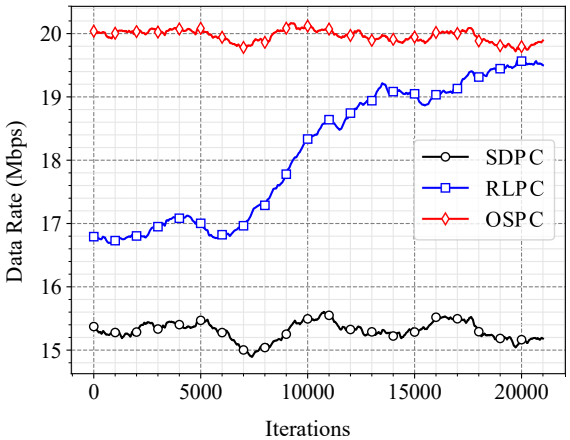
The RLPC has computational complexity higher than that of SDPC, since it requires the determination of the largest value of the Q table, while the SDPC only requires a comparison of two scalar values. The computational disadvantage of RLPC is compensated by the significant reduction in the transmit power and the substantial increase in the data rate in comparison with SDPC. OLPC obtains the best results in all analyzed parameter settings. This occurs at the expense of a high computational cost resulting from the self-decomposition operation involving a high-dimensional matrix. In addition, OLPC requires an intense signaling, since the channel matrix, precoders, decoders and SINR targets must be informed at each iteration. The RLPC algorithm requires a much lower signaling level, requiring only the parameter ρ (on the beginning of the process) and the data rate. We summarize in Table III the main comparison aspects of the algorithms under analysis.

VII. Conclusions

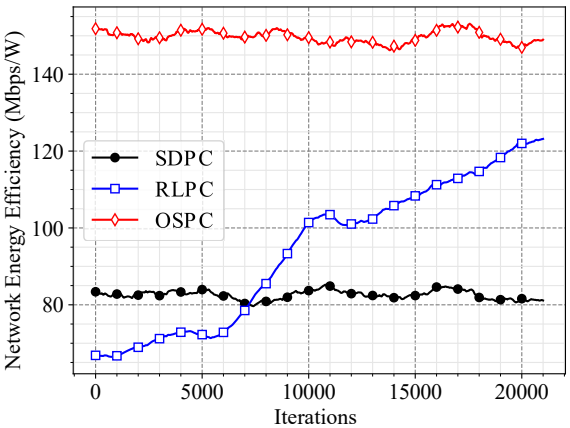
The proposed uplink PC framework based on multi-agent RL for a 5G NR network provides a self-exploratory solution. It enabled the system to learn the power control to fulfill the enhanced throughput on the uplink channel under neighbor cell interference mitigation. Simulation results show that the proposed uplink PC framework provides near-optimum performance in terms of total transmit power, total data



(a) Total transmit power



(b) Total data rate



(c) Network energy efficiency

Fig. 8. Comparison of the proposed RL-based uplink PC design with classical algorithms.

TABLE III
Comparison Uplink Power Control Algorithms.

Algorithm	Signaling	Complexity
SDPC	Low	Low
RLPC	Low	High
OSPC	High	High

rate, and network energy efficiency. The proposed signaling scheme provides a power control strategy that allows the cooperation among BSs to mitigate inter-cell interference. As a consequence, the proposed framework overcomes SDPC with similar signaling levels.

Appendix A
Computational Complexity of the RLPC Algorithm

Line 2 in Algorithm 1 defines the initialization of power levels of all D antenna ports in each cell. These powers are modeled as discrete values and are limited to minimum and maximum values ($P_{d,c}^{\min}$ and $P_{d,c}^{\max}$) determined by Eqs. (10) and (11), respectively. The number of possible power values is a function of the step size χ among transmit power levels and the power limits. Thus, initialization of power levels requires the sort of D values in a space of possible states (powers) with cardinality given by Eq. (15). A loop of C iterations is required to initialize the power levels on all cells. Therefore, the computational complexity related to this line is $O\left(\sum_{c=1}^{C}|\mathcal{S}_c|\right) = O\left(\sum_{c=1}^{C}\left(\frac{P_{d,c}^{\max}-P_{d,c}^{\min}}{\chi}\right)^D\right)$. For simplification purposes, we consider the space of states in all cells with the same size $|\mathcal{S}_c| = \left(\frac{P_{d,c}^{\max}-P_{d,c}^{\min}}{\chi}\right)^D$. Then, the computational complexity becomes $O(C|\mathcal{S}_c|)$.

The main loop involves the commands between lines 4 and 16. This loop is executed T times until the stop condition is determined. Line 4 determines the sort of a random number. This operation has complexity $CO(1) = O(C)$, since it is repeated by all agents. In the following, line 5 requires the calculation of the learning rate ϵ_t according to Eq. (12). This operation involves the calculation of scalar values, which has complexity $CO(1) = O(C)$. In one hand, if the condition ($e < \epsilon_t$) or ($t = 0$) is true, it is sorted a random set of actions from the space of action, as defined in line 7, which has computational complexity $O\left(\sum_{c=1}^{C}|\mathcal{A}_c|\right) = O(C\iota^D)$ (ι is the number of TPC commands). It is assumed that all agents have the same space of actions. On the other hand, if the condition ($e < \epsilon_t$) or ($t = 0$) is false, we select an action based on the evaluation of the maximum element of row $Q(s_{t,c},:)$. Therefore, the operation defined in line 9 has computational complexity $CO(|\mathcal{A}_c| = O(C|\mathcal{A}_c)) = O(C\iota^D)$.

Line 11 defines the computation of transmit power update of the antenna ports in all cells. Thus, the computational complexity of this operation is $CDO(1) = O(CD)$. The verification of the antenna port power limits in line 12 has computational complexity $CO(D) = O(CD)$ and the execution of transmit power updates in line 13 has complexity $CO(D) = O(CD)$. The calculation of the reward associated to the performed actions has computational complexity $CO(D) = O(CD)$. Finally, the update of the matrix Q defined in line 15 has computational complexity $CO(|\mathcal{S}_c||\mathcal{A}_c|) = O(C|\mathcal{S}_c||\mathcal{A}_c|)$.

The computational complexity of the proposed RL-based uplink PC framework based on the analyzed pseudo-code is $O(C|\mathcal{S}_c|) + T(O(C) + O(|\mathcal{A}_c|) + O(CD) + O(|\mathcal{S}_c||\mathcal{A}_c|)) \rightarrow O(TC|\mathcal{S}_c||\mathcal{A}_c|)$.

Appendix B
Computational Complexity of the OSPC Algorithm

Line 2 in Algorithm 2 defines the SINR targets of all D antenna ports in all C cells. This command has computational complexity $CD \cdot O(1) = O(CD)$. The next command at line 4 specifies the extended coupling matrix of each cell, which requires the operations defined at Eq. (20) and Eq.(21). The determination of the coupling matrix Ψ_c at Eq.(21) requires the manipulation of beamforming vectors and channel matrices, namely $\overline{\mathbf{w}}_{b_r} \in \mathbb{C}^{N\times 1}$, $\bar{\mathbf{f}}_{b_t} \in \mathbb{C}^{M\times 1}$, and $\mathbf{H} \in \mathbb{C}^{N\times M}$, respectively. The operation is repeated D^2 times since it is created a squared matrix with dimension $D\times D$. Thus, the computational complexity is $D^2 \cdot (O(N) + O(NM) + O(M)) \rightarrow O(D^2N) + O(D^2MN) + O(D^2M) \rightarrow O(D^2MN)$, where M and N are the number of receive and transmit antennas, respectively. Besides the coupling matrix Ψ_c, the definition of the extended coupling matrix Λ_c also requires the specification of the auxiliary matrix $\mathfrak{D}_c$. This matrix requires the computational of the norm of the vector $\overline{\mathbf{w}}_{b_r}^H \mathbf{H}\bar{\mathbf{f}}_{b_t}$ to determine the D elements of its main diagonal. Therefore, it has computational complexity $D\cdot(O(N)+O(MN)+O(M)) \rightarrow O(DN)+O(DMN)+O(DM) \rightarrow O(DMN)$.

The extended coupling matrix Λ_c is composed by different blocks of matrices. The determination of $\mathfrak{D}_c\Psi_c^T$ has computational complexity $O(D^2)$, since it requires the multiplication of matrices of dimension $D \times D$. The determination of $\frac{1}{P_{\text{PCMAX}}}\mathbf{1}^T\mathfrak{D}_c\Psi_c^T$ has computational complexity $O(D^2)$ since it involves the multiplication of arrays with dimensions $D \times 1$ and $D \times D$. The determination of $\mathfrak{D}_c\sigma$ also has computational complexity $O(D^2)$ since it involves again the multiplication of arrays with dimensions $D \times 1$ and $D \times D$. The determination of $\frac{1}{P_{\text{PCMAX}}}\mathbf{1}^T\mathfrak{D}_c\sigma$ has computational complexity $O(D^2)$ since it involves the multiplication of arrays with dimensions $D \times 1$ and $D \times D$. The composition of the extended coupling matrix has computational complexity $O(D^2)+O(D^2)+O(D^2)+O(D^2) \rightarrow O(D^2)$. Thus, the definition of the extended coupling matrix has computational complexity $T \cdot C \cdot O(D^2MN) + O(DMN) + O(D^2) \rightarrow O(TCD^2MN) + O(TCDMN) + O(CTD^2) \rightarrow O(TCD^2MN)$, where T is the total number of iterations.

In addition, line 5 requires the eigendecomposition of the extended coupling matrix Λ_c, so the computational complexity is $T \cdot C \cdot O((D+1)^3) = O(TCD^3)$. The verification of the power limits at line 6 has computational complexity $T \cdot C \cdot D \cdot O(1) = O(TCD)$. The implementation of the power commands has computational complexity $T \cdot C \cdot D \cdot O(1) = O(TCD)$. Thus, the total number of operations in big O notation is $O(CD) + O(TCD^2MN) + O(TCD^3) + O(CTD) \rightarrow O(TCD^2MN) + O(TCD^3)$.

Appendix C
Computational Complexity of the SDPC Algorithm

Line 2 in Algorithm 3 defines the initial transmit power of all D antenna ports in all C cells. As observed previously, this command has computational complexity $CD \cdot O(1) = O(CD)$. The main loop involves the commands between lines 4 and

8. This loop is executed T times until the stop condition is determined. Line 4 calculates the target SINR $\Gamma_{c,d}^{\text{target}}$ according to Eq. (23). This operation has computational complexity $C \cdot D \cdot O(1) = O(CD)$. Line 5 determines the update of the transmit power in all antenna ports of each cell according to Eq. (22). This command has computational complexity $C \cdot D \cdot O(1) = O(CD)$. Line 7 defines the execution of the transmit power updates, which has computational complexity $C \cdot D \cdot O(1) = O(CD)$. Therefore, the computational complexity of the soft dropping power control based on the previous pseudo-code is $O(CD) + T\left(O(CD) + O(CD) + O(CD)\right) \rightarrow O(TCD)$.

ACKNOWLEDGMENTS

This work was supported by the Ericsson Research, Sweden, and Ericsson Innovation Center, Brazil, under UFC.47 and UFC.48 Technical Cooperation Contracts Ericsson/UFC. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Francisco Hugo Costa Neto was supported by CAPES under the grant 88882.183549/2018-01. T. F. Maciel would like to acknowledge CNPq for its financial support under the grants 426385/2016-0 and 308621/2018-2. This work is also partially supported by CNPq (Proc. 306616/2016-5), CAPES - Finance Code 001, and CAPES/PRINT Proc. 88887.311965/2018-00.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-Advanced Pro and The Road to 5G*, 3rd ed. Academic Press, 2016, vol. 1.

[2] A. Simonsson and A. Furuskar, "Uplink Power Control in LTE - Overview and Performance, Subtitle: Principles and Benefits of Utilizing rather than Compensating for SINR Variations," in *IEEE Vehicular Technology Conference*, Sep. 2008, pp. 1–5.

[3] H. Zhang, N. Prasad, S. Rangarajan, S. Mekhail, S. Said, and R. Arnott, "Standards-Compliant LTE and LTE-A Uplink Power Control," in *IEEE International Conference on Communications*, Jun. 2012, pp. 5275–5279.

[4] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent Advances in Radio Resource Management for Heterogeneous LTE/LTE-A Networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2142–2180, 2014.

[5] G. Ku and J. M. Walsh, "Resource Allocation and Link Adaptation in LTE and LTE Advanced: A Tutorial," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1605–1633, 2015.

[6] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Academic Press, Aug. 2018, vol. 1.

[7] S. K. Sharma and X. Wang, "Toward Massive Machine Type Communications in Ultra-Dense Cellular IoT Networks: Current Issues and Machine Learning-Assisted Solutions," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 426–471, May 2020.

[8] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive Access for 5G and Beyond," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2020.

[9] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, M. Fallgren, and J. F. Monserrat, "The METIS 5G System Concept: Meeting the 5G Requirements," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 132–139, Dec. 2016.

[10] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective Multiple Antenna Technologies for Beyond 5G," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1637–1660, 2020.

[11] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks," *IEEE Access*, vol. 6, pp. 32 328–32 338, May 2018.

[12] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[13] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, and L. Hanzo, "Machine Learning Paradigms for Next-Generation Wireless Networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, Jul. 2017.

[14] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2392–2431, Jul. 2017.

[15] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019.

[16] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," *IEEE Access*, vol. 7, pp. 137 184–137 206, 2019.

[17] 3GPP, "General Description," 3rd Generation Partnership Project (3GPP), TS 38.201, Jan. 2018, v15.6.0.

[18] ——, "Services Provided by the Physical Layer," 3rd Generation Partnership Project (3GPP), TS 38.202, Oct. 2020.

[19] ——, "Physical Channels and Modulation," 3rd Generation Partnership Project (3GPP), TS 38.211, Oct. 2020, V15.8.0.

[20] ——, "Multiplexing and Channel Coding," 3rd Generation Partnership Project (3GPP), TS 38.212, Nov. 2020, V15.8.0.

[21] ——, "Physical Layer Procedures for Control," 3rd Generation Partnership Project (3GPP), TS 38.213, Nov. 2020, V15.8.0.

[22] ——, "Physical Layer Procedures for Data," 3rd Generation Partnership Project (3GPP), TS 38.214, Nov. 2020, V15.8.0.

[23] ——, "Physical Layer Measurements," 3rd Generation Partnership Project (3GPP), TS 38.215, Oct. 2020, V15.6.0.

[24] A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for Interference Control in OFDMA-Based Femtocell Networks," in *IEEE Vehicular Technology Conference*, May 2010, pp. 1–5.

[25] ——, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823–1834, May 2010.

[26] C. Cordeiro, K. Challapali, and D. Birru, "IEEE 802.22: An Introduction to the First Wireless Standard based on Cognitive Radios," *Journal of Communications*, vol. 01, no. 01, Apr. 2006.

[27] S. Dzulkifly, L. Giupponi, F. Said, and M. Dohler, "Decentralized Q-Learning for Uplink Power Control," in *IEEE International Workshop on Computer Aided Modelling and Design of Communication Links and Networks*, Sep. 2015, pp. 54–58.

[28] S. Deb and P. Monogioudis, "Learning-Based Uplink Interference Management in 4G LTE Cellular Systems," *IEEE/ACM Transactions on Networking*, vol. 23, no. 2, pp. 398–411, Apr. 2015.

[29] Z. Kaleem, A. Ahmad, and M. H. Rehmani, "Neighbors' Interference Situation-Aware Power Control Scheme for Dense 5G Mobile Communication System," *Telecommunication Systems*, vol. 67, no. 3, pp. 443–450, Mar. 2018.

[30] X. Li and J. Fang and W. Cheng and H. Duan and Z. Chen and H. Li, "Intelligent Power Control for Spectrum Sharing in Cognitive Radios: A Deep Reinforcement Learning Approach," *IEEE Access*, vol. 6, pp. 25 463–25 473, Oct. 2018.

[31] Y. S. Nasir and D. Guo, "Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[32] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-Level Control Through Deep Reinforcement Learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[33] Y. Hu, M. Chen, Z. Yang, M. Chen, and G. Jia, "Optimization of Resource Allocation in Multi-Cell OFDM Systems: A Distributed Reinforcement Learning Approach," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–6.

[34] T. Zhang and S. Mao, "Smart Power Control for Quality-Driven Multi-User Video Transmissions: A Deep Reinforcement Learning Approach ," *IEEE Access*, vol. 8, pp. 611–622, 2020.

[35] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, "Deep Reinforcement Learning for Joint Beamwidth and Power Optimization in mmWave Systems," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2201–2205, 2020.

[36] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[37] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Jiang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[38] K. I. Ahmed, H. Tabassum, and E. Hossain, "Deep Learning for Radio Resource Allocation in Multi-Cell Networks," *IEEE Network*, vol. 33, no. 6, pp. 188–195, Nov. 2019.

[39] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[40] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 173–196, 2019.

[41] C. J. C. H. Watkins and P. Dayan, "Technical Note: Q-Learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.

[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2018.

[43] C. J. C. H. Watkins, "Learning from Delayed Rewards," phdthesis, Kings's College, May 1989.

[44] 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz ," 3rd Generation Partnership Project (3GPP), TS 38.901, Jan. 2017, V.15.0.0.

[45] ——, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone," 3rd Generation Partnership Project (3GPP), TS 38.101-1, Oct. 2018.

[46] A. M. Khachan, A. J. Tenenbaum, and R. S. Adve, "Linear Processing for the Downlink in Multiuser MIMO Systems with Multiple Data Streams," in *IEEE International Conference on Communications*, Jun. 2006, pp. 1–6.

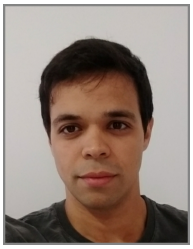
[47] M. Schubert and H. Boche, "Solution of the Multiuser Downlink Beamforming Problem with Individual SINR Constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[48] R. Yates, S. Gupta, C. Rose, and S. Sohn, "Soft Dropping Power Control," in *IEEE Vehicular Technology Conference*, May 1997, pp. 1–5.

[49] Tarcisio Ferreira Maciel, "Suboptimal Resource Allocation for Multi-User MIMO-OFDMA Systems," Phd Thesis, Technische Universitat Darmstadt, Sep. 2008.

[50] Yuri Victor Lima de Melo, "Power Control and Energy Efficiency Strategies for D2D Communications Underlying Cellular Networks," Master Thesis, Federal University of Ceara, Jul. 2015.

Francisco Hugo Costa Neto received the B.Sc. degree in Electrical Engineering in 2014 and the M.Sc. and Ph.D. degree in Telecommunications Engineering in 2016 and 2020, respectively, from Federal University of Ceará, Brazil. Currently, he is Postdoctoral Researcher with the Wireless Telecommunications Research Group (GTEL), Departament of Teleinformatics Engineering, Federal University of Ceará. His research interests include wireless communications, signal processing, machine learning, and optimization.



Daniel C. Araújo received the Diploma degree in Telecommunication Engineering from the University of Fortaleza (UNIFOR), Ceará, Brazil, in 2010, the M.Sc. degree in Telecommunication Engineering, and the PhD in Telecommunication Engineering from the University of Ceará (UFC), Ceará, Brazil, in 2012 and 2016, respectively. He was a member of the Wireless Telecommunication Group in the University of Ceará, from 2012 to 2019, when he worked as visiting researcher at Ericsson Research. Since 2019, he has been a Professor at the University of Brasília, Federal District, Brazil, where he serves as a Professor in the graduate course of electronic engineering and the Pos-Graduate Program of Electrical Engineering (PPGEE). His general research interests include statistical learning, communications, array signal processing.

Mateus P. Mota received the B.Sc and M.Sc degrees in teleinformatics engineering from the Federal University of Ceará (UFC), Brazil, in 2017 and 2020, respectively. In 2020, he joined Nokia Bell Labs France as an early stage researcher within the framework of the European Marie-Curie ITN Project Windmill. He is currently pursuing the Ph.D degree in Electrical Engineering at National Institute of Applied Sciences of Lyon (INSA Lyon), with a focus on Deep Reinforcement Learning for Wireless Communications.

Tarcisio F. Maciel received the B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Ceará (UFC), in 2002 and 2004, respectively, and the Dr.Ing. degree in electrical engineering from the Technische Universitat Darmstadt (TUD), Germany, in 2008. Since 2001,he has actively participated in several projects in a technical and scientific cooperation between the Wireless Telecom Research Group (GTEL), UFC, and Ericsson Research. From 2005 to 2008, he was a Research Assistant with the Communications Engineering Laboratory, TUD. Since 2008, he has been a member of the Postgraduation Program in Teleinformatics Enginnering, UFC. In 2009, he was a Professor of computer engineering with UFC-Sobral. Since 2010, he has been a Professor with the Center of Technology, UFC. His research interests include radio resource management, numerical optimization, and multiuser/multiantenna communications.

André L. F. de Almeida (Senior Member, IEEE) received a double Ph.D. degree in Sciences and Teleinformatics Engineering from the University of Nice, Sophia Antipolis, France, and the Federal University of Ceará, Fortaleza, Brazil, in 2007. He currently is an Associate Professor with the Teleinformatics Engineering Department of the Federal University of Ceará. He served as an Associate Editor for several journals, such as the IEEE Transactions on Signal Processing (2012-2014 and 2014-2016), the IEEE Signal Processing Letters (2016-2018 and 2018-2020). He currently serves as a Senior Area Editor for the IEEE Signal Processing Letters and as an Associate Editor for the IEEE Transactions on Vehicular Technology. Prof. Almeida is an elected member of the Sensor Array and Multichannel (SAM) Technical Committee of the IEEE Signal Processing Society (SPS) (2015-2018 and 2018-2021) and an elected member of the EURASIP Signal Processing for Multi-Sensor Systems Technical Area Committee (SPMuS TAC) (2016-2018 and 2019-2022). He also served as associate member of the Big Data Special Interest Group (SIG) of the IEEE SPS (2015-2018). He was a General Co-Chair of the 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'2017), Technical Co-Chair of the IEEE GlobalSIP'2018 and IEEE GlobalSIP'2019 Symposia on Tensor Methods for Signal Processing and Machine Learning, Technical Co-Chair of the 11th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'2020), and the General Co-Chair of the IEEE CAMSAP'2021. He is a level-1D productivity research fellow of the CNPq (the Brazilian National Council for Scientific and Technological Development). He was awarded multiple times visiting professor positions at the University of Nice Sophia-Antipolis, France (between 2012 and 2019). In 2018, he was elected an Affiliate Member of the Brazilian Academy of Sciences. Prof. Almeida is a Senior Member of the IEEE. His research interests lie in the area of signal processing for communications and sensor array and multichannel processing, including topics such as channel estimation and equalization, multi-antenna systems, blind and semi-blind signal processing, and direction of arrival estimation. An important part of his research has been dedicated to multilinear algebra and tensor decompositions with applications to communication systems and signal processing.