A Distributed Average Cost Reinforcement Learning approach for Power Control in Wireless 5G Networks

Antonio Ornatelli
Department of Computer, Control, and Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
ornatelli@diag.uniroma1.it

Alessandro Giuseppi
Department of Computer, Control and Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
giuseppi@diag.uniroma1.it

Andrea Tortorelli
Department of Computer, Control, and Management Engineering Antonio Ruberti
Sapienza University of Rome
Rome, Italy
tortorelli@diag.uniroma1.it

***Abstract*—This paper deals with the transmission power control problem in wireless networks. Such a problem represents a well known and relevant issue as it allows to efficiently manage the network's required energy and the interference experienced by end-users. With the widespread diffusion of smart devices, the relevance of this aspect further increased and has been identified as such also in 5G standards. The problem has been formalized as a Multi-Agent Reinforcement Learning approach (MARL) to guarantee scalability and robustness. These two aspects also drove the development of an original Distributed Average-Cost Temporal-Difference (TD) Learning algorithm. To adopt such an algorithm, a Markov Game formulation of the power control problem has also been derived. The effectiveness of the proposed distributed framework in reducing the total network's transmission power has been proved by means of simulations in a specific case study.**

***Index Terms*—Distributed Reinforcement Learning, Power Control, Average Cost TD Learning, Dynamic Consensus, Networked Multi-Agent System**

I. Introduction

Due to their modelling capabilities, in the last decade, Multi-Agent Systems (MASs) have been successfully applied to tackle a variety of complex control problems [1]–[4]. A MAS consists of a decision-making problem where a set of intelligent agents interacts with the environment, i.e., the system to be controlled. Such interaction consists of the agents performing actions impacting the environment and can be based on static rules (e.g., if-then-else conditions) or on learned policies. Being able to address complex and time-varying communication topologies as well as complex heterogeneous dynamics, the framework provided by MASs has been adopted in several application domains including communication networks [5]–[8], smart grids [9], robotics [10] and social networks [11].

To model and solve decision-making problems, the availability of some a priori knowledge regarding the system's dynamics plays a crucial role in aiding the achievement of the desired performances. However, in real-world applications, such a priori knowledge may not be available or not fully representative of the system to be controlled. The misalignment between the real and modelled system dynamics leads to approximations impairing the capabilities of the adopted mathematical model framework to capture relevant aspects and, in turn, bounding the effectiveness of the deployed control logic. In this respect, model-free approaches allow to overcome these issues and thus can be also deployed in time-varying and stochastic environments whereas certain aspects of the system's dynamics are unknown and/or complex to identify.

With this in mind, Reinforcement Learning (RL) [12] represents a powerful control methodology to solve complex decision-making problems as it allows to significantly reduce the system's required knowledge. More specifically, RL allows to learn (sub) optimal policies (i.e., functions mapping system's states and the control actions to be taken) by directly interacting with the system itself. The basic scheme of such interaction is the following: at each discrete time step t, the RL agent receives an observation o_t of the system to be controlled; based on such observation, the agent selects an action a_t which impacts the system causing a state transition from s_t to s_{t+1}; finally, the agent receives a reward r_{t+1} providing a measure of how good it was to choose action a_t when the system's state is s_t. This reward is then used by the agent to refine its behaviour and eventually learn the optimal policy. The described interaction scheme implies a closed-loop adaptive control scheme which renders RL approaches very suitable to solve complex decision-making problems.

When multiple RL agents (or MASs) are considered, the control problem is referred to as Multi-Agent Reinforcement Learning (MARL) [13]–[15]. MARLs can be classified into two main classes: i) cooperative-MARL, in which all the agents share a common goal, or ii) competitive-MARL, in which each agent selfishly pursues its own goals.

This project was partially supported by the *Sapienza University of Rome* under the project "Avvio alla ricerca 2021".

The aim of this paper is to present a cooperative-MARL control framework for transmission power control in heterogeneous wireless networks, as defined by the 5G and other emerging networks framework [16], [17]. More specifically, the objective consists in learning the best strategy to be adopted for managing, in an efficient way, the transmission power from data sources (e.g., transmitters/access points such as 5G NR, 4G, WiFi routers or sensors) to data destinations (e.g., receivers such as smartphones, sensors, actuators or microprocessors). This problem is particularly relevant since a coordinated and efficient management of the transmission power in a wireless network allows to mitigate interference among communication channels. Indeed, a lower interference leads to an overall higher transmission bit rate [18], [19] and, in turn, to a better usage of network resources.
In this respect, the framework provided by cooperative-MARLs is particularly suited for the considered problem. As a matter of fact, being able to enforce coordination between the agents (i.e., the transmitters) regarding the allocated transmission power allows to maximize the overall network capacity. This can be achieved since transmitters are able to retrieve information on the network status including the interference level which in typical communication networks is provided by the receivers.

To tackle scalability issues, the described power control problem will be addressed by developing a Distributed Average Cost Temporal-Difference (TD) Reinforcement Learning algorithm. The proposed distributed algorithm allows to define a scalable and reliable MARL control scheme minimizing communication exchanges between the agents and reducing the required computational complexity.

The remainder of this paper is organized as follows: Section II presents an overview of the *Multi-Agent Reinforcement Learning* framework; Section III introduces the state-of-the-art of *Average Temporal-Difference learning* (see Section III-A) and *Dynamic average consensus* (see Section III-B); Section IV describes the *proposed distributed control algorithm*; Section V describe the mathematical formulation of the *Wireless Power Control problem* (see Section V-A) and the *proposed Markov Game formulation* of the wireless power control problem (see Section V-B); Section VI presents the considered case study and simulation results of the proposed approach; finally, in Section VII, the obtained results are summarized and future developments discussed.

II. Background on Multi-Agent Reinforcement Learning

Single-agent RL problems can be easily formalized by exploiting the mathematical framework provided by Markov Decision Processes (MDPs). Indeed, MDPs allow to model complex decision-making problems characterized by the Markov (or *memoryless*) property which consists in requiring that, at any given time $t+1$, the environment's state s_{t+1} only depends on the previous state s_t and the action a_t taken at time t.

A (finite) MDP can be described by a tuple:

$$MDP = (S, A, \delta, r) \tag{1}$$

where

- S, referred to as *state space*, is a finite set including all possible environment's state;
- A, referred to as *action space*, is a finite set including all the possible actions which can be taken by the decision-maker (i.e., the RL agent);
- $\delta : P(S \times A \times S) \rightarrow [0, 1]$, referred to as the transition function, is the conditional probability $P(s'|s, a)$ of the successor state s', given the current state and action s, a;
- $r : S \times A \times S \rightarrow \mathbb{R}$ is the reward function which shall capture the performance of the controlled system in terms of its goals and objectives.

Given a MDP, the goal of RL algorithms is to find the *optimal policy* $\pi^* : S \times A \rightarrow A$ maximizing the cumulative expected reward over time. Hence, the optimal policy π^* can be defined as:

$$\pi^* = \arg\max_{\pi} E\left[\sum_{k=0}^{K} \gamma^k r_k\right], \qquad \forall s \in S \tag{2}$$

where $r_k = r(s_k, a_k, s_{k+1})$, $a_k = \pi(s_k)$, K is the control horizon and $\gamma \in [0, 1)$ is the discount factor weighting future rewards. In other words, the agent's objective consists in maximizing its long-term reward based on the current received feedback.

To take into account multiple collaborative agents, the discussed MDP framework needs to be generalized. Indeed, in this case, each agent would consider all the other agents as part of the environment. This, in turn, means that each agent adjusts its behaviour based on other agents' actions rather than on the environment state [13]–[15]. Following these considerations, it is possible to consider Markov Games (MGs) [20], also referred as Stochastic Games (SG) [21], [22], in place of MDPs. The framework provided by Markov Games allows each agent to perform actions considering (i) the environment's state and (ii) all the other agents' adaptive behaviour. Similarly to MDPs, MGs can be described by a tuple:

$$MG = (N, S, \{A^i\}_{i \in N}, \delta, \{r^i\}_{i \in N}) \tag{3}$$

where

- $N > 1$ is the set of agents;
- S is the finite set of environment's states;
- A^i is the finite set of actions of the i-th agent, and $A := A^1 \times A^2 \times \cdots \times A^N$ is referred to as the *joint action space*;
- δ is the transition function defined as the probability distribution over transitions $P(S \times A \times S) \rightarrow [0, 1]$ and can be expressed as the conditional probability $P(s'|s, a)$ of the successor state s', given the current state s and the current joint action;
- $r^i : S \times A^i \times S \rightarrow \mathbb{R}$ is the reward function of the i-th agent.

Given the described MG framework, one of the main issues of MARL solution algorithms is represented by the problem's complexity. Indeed, the problem dimension grows exponentially with respect to N (i.e., the number of agents) since each agent adds its own variables to the joint state-action spaces. Furthermore, RL algorithms are affected by the so-called *curse of dimensionality* since the dimension of state and action spaces grow exponentially with the number of possible actions and states. To address all these issues, in Section IV a distributed average cost TD reinforcement solution algorithm able to solve a cooperative MARL will be presented.

MARL problems can be classified based on the nature of agents' objectives:

- *Fully Cooperative Tasks* — In fully cooperative SGs, the agents share the same reward function and their learning goal consists in maximizing the common return. In this case, solution algorithms can be designed using coordination-free, coordination-based or indirect coordination methods;
- *Fully Competitive Tasks* — In fully competitive SGs, each agent maximizes its own goal assuming that the other agents' will try to minimize its obtained rewards (according to a minimax principle). In this case, solution algorithms can be tailored for each agent;
- *Mixed Tasks* — In mixed SGs, no constraints are imposed on the agents' reward functions. This class of MARLs is suited for heterogeneous scenarios in which there are both selfish and cooperative agents. More in detail, this problem formulation allows to take into account the fact that cooperative agents may encounter situations in which their immediate interests are in conflict with those of other agents. Concerning solution algorithms, game-theoretic considerations (e.g., the concept of equilibria) are the most adopted ones.
- *Explicit Coordination Mechanisms* — Explicit coordination mechanisms can be used for any type of MARL tasks (cooperative, competitive, or mixed): in this case, solution algorithms consider the fact that each agent's actions are coordinated (or negotiated) with all the other agents;

A second aspect based on which it is possible to classify MARL problems is represented by the considered information exchange logic. In this respect, three main settings can be identified:

- *Centralized MARLs* — In this case, a central controller collects and processes the environment's and local agents' data (i.e., joint actions, rewards, and observations). Such a centralized controller is in charge of providing each agent with the designed policy;
- *Decentralized Networked Agents MARLs* — This setting envisages a time-varying communication network connecting the agents. In other words, local information spreads across the network by local information exchanges between neighbouring agents;
- *Fully Decentralized MARLs* — In this scenario there is no explicit information exchange between the agents. More in detail, each agent makes decisions based on its local observations, without any coordination and/or aggregation of data. It should be noted that although local observations differ from agent to agent, they may contain some global information.

III. Background

A. Average Temporal-Difference learning

The present work is aimed at developing a Cooperative Distributed Average Cost Temporal-Difference algorithm allowing to solve the power control problem in wireless networks. To achieve this result, the starting point has been [23] where the authors presented a single agent formulation of the Average Cost Temporal-Difference learning problem. The proposed formulation considers an irreducible and aperiodic Markov chain $\{s_t | t = 0, 1, \dots\}$ on a finite state space $S = \{1, \dots, n\}$, with transition probability matrix P. The average cost per stage associated with state s is defined as $g(s)$, where s_t is the state at time t and the average cost is $\mu^* = E[g(s_t)]$. The considered objective function is a basic differential cost function $J^* = \sum_{t=0}^{\infty} P^t(g - \mu^* e)$ with e being a vector of all ones and P^t being transition probability matrix of the Markov chain. The learning problem consists in estimating $\hat{J}(s_t, r_t)$, i.e., the approximation of the differential cost function which is defined as:

$$\hat{J}(s, r) = \sum_{k=1}^{K} r(k)\phi_k(s) \tag{4}$$

where r is a vector of tunable parameters and each ϕ_k is a basis function defined in the state space S. To a given transition between two states s_t and s_{t+1} it is associated a temporal difference d_t defined as

$$d_t = g(s_t) - \mu_t + \hat{J}(s_{t+1}, r_t) - \hat{J}(s_t, r_t) \tag{5}$$

where μ_t and r_t are the estimation at time t of the average cost μ^* and parameters vector r, respectively. Said estimates are updated according to

$$\mu_{t+1} = (1 - \eta_t)\mu_t + \eta_t g(s_t) \tag{6}$$

$$r_{t+1} = r_t + \gamma_t d_t z_t \tag{7}$$

where z_t is referred to as *eligibility vector* and is defined as

$$z_{t+1} = \lambda z_t + \phi(s_{t+1}) \tag{8}$$

where γ_t and η_t are scalar step size and λ is a parameter in $[0, 1)$. With respect to equation 8, the eligibility vector z_1 at time $t = 1$ is simply equal to $\phi(s_1)$, i.e., $z_0 = 0$. Furthermore, the step sizes γ_t are positive, deterministic, and satisfy the following relations:

$$\sum_{t=0}^{\infty} \gamma_t = \infty \ , \ \sum_{t=0}^{\infty} \gamma_t^2 < \infty. \tag{9}$$

The reader is referred to [23] for a full discussion on the problem motivation, convergence proof and error estimates.

B. Dynamic Average Consensus

When a set of autonomous agents needs to reach a global agreement starting from local measures, several difficulties arise. Indeed, collecting said local data over large-scale and time-varying networks is a complex problem [25], [26]. To deal with this issue, the easiest solution consists in adopting a centralized control logic. However, this approach has several drawbacks such as (i) lack of robustness, (ii) increase in communication exchanges and required computational power, (iii) lack of scalability and (iv) confidentiality and privacy issues (since local data is spread over the whole network).

To deal with these issues, distributed communication protocols have been successfully applied. More specifically, dynamic average consensus algorithms, such as the one described in [27], allow autonomous agents to keep track of the average value of time-varying signals which are locally measured by each agent and shared only between neighbouring agents. For each agent i, the estimate $\tilde{\mu}$ of a given signal can be computed at each time step by means of the discrete First-Order Dynamic Average Consensus algorithm as follows:

$$\tilde{\mu}^i_{t+1} = \tilde{\mu}^i_t + \sum_{j \neq i} a_{i,j}(\tilde{\mu}^j_t - \tilde{\mu}^i_t) + \Delta x^i_t \tag{10}$$

where x^i_t is the signal measured by the i-th agent, $\Delta x^i_t = x^i_t - x^i_{t-1}$, and $a_{i,j}$ is the (i,j)-th entry of the adjacency matrix of the network's communication graph specifying if there exists a communication link between agents i and j.

IV. Distributed Average Temporal-Difference Learning

The MARL instance considered in this work is the fully cooperative one (see Section II). Concerning the characterization in terms of the information exchange logic, by considering a centralized controller the problem reduces to a MDP whose action space is equal to the joint action space of the MG as defined in equation (3). Indeed, the control problem' goal can be achieved by learning the optimal joint-action values with single agent algorithms thus removing the exponential growth of the problem's complexity with respect to the agents' number. However, as already mentioned, the adoption of a centralized controller can be characterized by scalability and robustness issues. Furthermore, since with this approach all the agents must send information to the central controller, also traffic congestion issues may arise.

Following these considerations, in this work a fully cooperative decentralized networked MARL will be considered. When considering decentralized approaches, the amount of exchanged information should be kept under control to avoid traffic congestion. This can be achieved by adopting appropriate communication protocols [24]–[27]. Furthermore, by adopting a dynamic average consensus protocol (see equation (10)), the complexity of the learning algorithm can be reduced as well allowing to deal with huge numbers of agents. More in detail, the proposed learning update rule is based on the Average TD learning [23] reported in Section III-A. This choice allows to update the agents' estimations by means of the average cost experienced by all the agents in the network; by doing so each agent updates its estimates taking into account the experienced cost of other agents.

A. Proposed Networked Agent Reinforcement Learning

The proposed distributed approach to tackle the power control in wireless networks relies on a multi-agent extension of the average cost TD learning algorithm already describe in Section III-A. To develop such an extension, the first step consists in coordinating the agents' actions. With respect to the classical average cost TD discussed in [23], in the proposed approach the average is computed considering all the costs experienced in the network by all the agents, and not the costs of a single agent.

With this solution, the agents share information about the local cost experienced with the neighbouring agents, and each agent can update its local estimate using the global average cost. However, the computation of the average cost between all the networked agents can be a cost expensive task, particularly with respect to the communication resources. To deal with this aspect, the dynamic average consensus algorithm presented in [27] has been embedded in the mentioned TD learning algorithm. This, in turn, enables a fully distributed computation of the average cost.

To solve the learning problem with this distributed approach, the update of the temporal difference in equation (5) is modified as follows for each agent i:

$$d^i_t = g^i(s_t) - \tilde{\mu}^i_t + \hat{J}^i(s_{t+1}, r_t) - \hat{J}^i(s_t, r_t) \tag{11}$$

where $\tilde{\mu}^i_t$ is the global average cost agreed by user i.

Furthermore, each agent updates its estimate of the average variable as follows:

$$\tilde{\mu}^i_{t+1} = \tilde{\mu}^i_t + \sum_{j \neq i} a_{i,j}(\tilde{\mu}^j_t - \tilde{\mu}^i_t) + \Delta\mu^i_t \tag{12}$$

where $\Delta\mu^i_t = \mu^i_t - \mu^i_{t-1}$, μ^i_t is computed according to equation (6), and $a_{i,j}$ is the (i,j)-th entry of the adjacency matrix of the network's communication graph. Equation (12) relies on the following parameters' updates:

$$\mu^{1,i}_{t+1} = (1 - \eta^i_t)\mu^{1,i}_t + \eta^i_t g(s_t)^i \tag{13}$$

$$\mu^{2,i}_{t+1} = \mu^{1,i}_t \tag{14}$$

$$r^i_{t+1} = r^i_t + \gamma^i_t d^i_t z^i_t \tag{15}$$

$$z^i_{t+1} = \lambda z^i_t + \phi^i(s_{t+1}) \tag{16}$$

$$\tilde{\mu}^i_{t+1} = \tilde{\mu}^i_t + \sum_{j \neq i} a_{i,j}(\tilde{\mu}^j_t - \tilde{\mu}^i_t) + \mu^{1,i}_t - \mu^{2,i}_t \tag{17}$$

V. Distributed Wireless Power Control

A. Mathematical Model

In wide-band wireless systems, when Code Division Multiple Access (CDMA) techniques are deployed, the downlink interference management is performed by each transmitter. This is achieved by modulating their transmission power taking into account the interference experienced by the receivers. Indeed, the higher signal interference is, the higher the transmission power should be to guarantee the quality of the signal at the receiver. However, higher transmission power implies higher interference for the receivers who are receiving different a signal. The interference level in the received signal can be measured by means of the Signal-to-Interference-plus-Noise Ratio *(SINR)* [19] which, for each transmitter i and each receiver j, is defined as:

$$SINR_{i,j} = \frac{P_{i,j} f(x_i, x_j)}{\sum_{k \in \mathbb{S}_{TX}, k \neq i} P_k f(x_k, x_j) + W_{i,j} N_0^j} \tag{18}$$

where $P_{i,j}$ is the i-th transmitting power towards the j-th receiver, $f(x_k, x_j)$ is the path loss function, it providing provides a measure of the power density reduction for transmitting a signal from the transmitter in position x_i and the receiver in position x_j, N_0^j is the thermal noise spectral density at the j-th receiver, $W_{i,j}$ is the bandwidth of the communication channel between transmitter i and receiver j, and $\mathbb{S}_{TX}$ is the set of nearby transmitters. Summarizing, equation (18) represents the *SINR* value experienced by the j-th receiver located at spatial point x_j for communications coming from the i-th transmitter, located at spatial point x_i, transmitting with power P_i.

SINR is a useful Key Performance Index (KPI) as it provides a measure of the reduction of the information transmitted in a communication path with a given power. Indeed, by combining equation (18) and the Shannon capacity theorem, it is possible to derive a measure of the maximum achievable data rate. That is, the channel capacity of a given communication path is defined as:

$$C_{i,j} = W_{i,j} \log(1 + SINR_{i,j}) \tag{19}$$

where $C_{i,j}$ is the channel capacity (or maximum rate of data). In other words, the higher the *SINR* and the channel bandwidth are, the higher is the allowed data rate.

B. Control Problem Formulation

The interference management problem presented in Section V-A can be formulated as a control problem aimed at maximizing the *SINR* experienced by each receiver. Such a problem can be formalized as an optimization problem as follows:

$$P_k = \underset{P \in \mathbb{R}^{N \times M}}{\arg\max}\, f(SINR_{i,j}, i \in \mathbb{S}_{TX}, j \in \mathbb{S}_{RX}) \tag{20}$$

subject to (18) and

$$P_{min}^i \leq \sum_{j \in \mathbb{S}_{RX}} P_{i,j}(k) \leq P_{max}^i \quad \forall i \in \mathbb{S}_{TX} \tag{21}$$

where $P_k \in \mathbb{R}^{N \times M}$ is the control matrix containing the transmission power of each transmitter towards each receiver, N and M are the number of transmitters and receivers in the network, respectively. $\mathbb{S}_{TX}$ and $\mathbb{S}_{RX}$ are the sets of transmitters and receivers, respectively, and P_{min}^i and P_{max}^i are the minimum and maximum total transmission power of the i-th transmitter, respectively.

Given the control problem formulation described by equations (18), (20), (21), it is straightforward to define the power control as a *MG*.

By recalling equation (3), this can be achieved by defining the elements of the tuple (N, S, A, δ, r). In the considered scenario, these elements can be defined as follows:

- N represents the number of the transmitters in the network;
- S is the state space observed by the agents and is given by:

  $$S = S^1 \times S^2 \times \cdots \times S^N \tag{22}$$

  where the state of a generic transmitter is defined as the measured *SINR* level for each receiver. Since the *SINR* is a continuous value, to reduce the algorithm complexity, its values are computed based on L discrete levels. By doing so, the state of each local agent at time k, $s_i(k) \in \mathbb{R}^M$, where M is the number of receivers connected to the i-th transmitter, can be described as a vector

  $$s_i(k) = [sinr_i^1, sinr_i^2, \ldots, sinr_i^M]^T \tag{23}$$

  where the generic scalar entry $sinr_i^j$ of the state represents the *SINR* level experienced by the generic receiver j. Note that the state space, as it has been defined, guarantees high flexibility since it is possible to trade off the model's description capabilities and the computational costs simply by increasing or decreasing the number of discrete levels L, respectively. Indeed, with these modelling choices, the number of possible states for the individual agent is equal to L^M which is a relatively small number. As a final remark, note that it is possible to generalize the proposed formulation by considering different discrete *SINR* levels for each transmitter-receiver pair based on the receivers' Quality of Service (QoS) level defined by the service provider contract, or other business or technical requirements. Therefore, the S^i state space is defined as

  $$S^i = \left\{ sinr(l)_i^j \,\middle|\, j \in [1, M], l \in [1, L] \right\} \tag{24}$$

- A is the joint action space defined as:

  $$A := A^1 \times A^2 \times \cdots \times A^N \tag{25}$$

  where A^i is the action space of the i-transmitter. From the above-described problem formulation, it follows that the control variable is the transmission power for each transmitter-receiver connection. In order to induce a smooth variation of the transmission power, and to reduce the dimension of the agent's action space, instead of directly considering the transmission power it is possible to consider as control actions at time k, $a_i(k) \in \mathbb{R}^M$

where M is the number of receivers connected to the i-th transmitter. With this modelling choice the action space of the i-th agent is described by

$$a_i(k) = [\lambda_i^1, \ldots, \lambda_i^M] \tag{26}$$

where the generic scalar entry λ_i^j represents the variation, with respect to the previous discrete time instant, of the transmission power from the i-th transmitter and the j-th receiver. Furthermore, such variations are limited to a small number of discrete levels Λ. With these modelling choices, the total number of possible actions for the individual agent is Λ^M, which is a relatively small number. Note that it is possible to generalize the proposed formulation by considering different discrete power variation levels for each transmitter, this choice can be driven by the maximum and minimum power variation the transmitter can implement. Therefore, the A^i action space can be rewritten as

$$A^i = \left\{ \lambda(l)_i^j \,\middle|\, j \in [1, M], l \in [1, \Lambda] \right\} \tag{27}$$

- δ is the transition function and its estimation is one of the objectives of the learning problem;
- r is the set of reward functions of each agent and is defined as a function of the state with a predefined value for each *SINR* level on the basis of receivers' needs. An example can be a linear function of the *SINR* since the higher the *SINR* level, the more the objective is satisfied according to the problem formulation above.

VI. Simulations and Results

In this section, the mathematical model presented in Section V and the proposed solution algorithm described in Section IV will be tested on a specific wireless power control scenario. The goal of the simulations is to prove that such a collaborative distributed algorithm is able to set the transmission powers to maximize the network's average bit rate as much as possible.

More in detail, 4 transmitters and 4 receivers have been considered. Each receiver is connected to a transmitter, and the resulting communication is affected by the transmission power of the other transmitters with a certain intensity level. Said intensity is proportional to the distance from the respective transmitters.

During the simulation, the receivers are assumed to be in three different positions. The receivers' positions variation changes the intensity of the interference of the transmitters for each receiver. Hence, it is expected that after a certain number of time steps the network average bit rate increases due to such position changes.

In Figure 1, the average of the bit rates experienced by the receiver is reported. The receivers' position variations are performed at steps 250000 and 350000 of the simulation. The figure shows that after both changes of receivers' positions the average bit rate level has a fast variation according to the current transmitters' power level and interference level at the receivers (see Section V-A). In fact, the proposed algorithm is able to increase the average bit rate level almost immediately, modulating the power level by performing actions as defined in Section V-B, in order to maximize the reward (i.e., the average experienced bit rate) computed in a distributed way according to the proposed algorithm in Section IV-A.

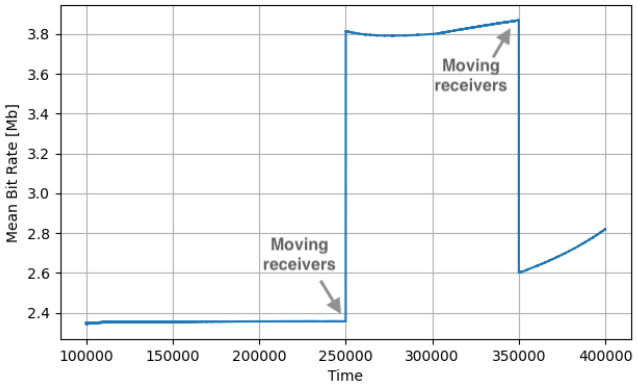

Fig. 1. Network's average bit rate.

VII. Conclusions and Future works

In this paper, the problem of controlling transmission power in wireless networks has been addressed. A single agent MDP formulation of the problem has been presented and its drawbacks highlighted. To overcome scalability and robustness issues, a MARL framework has been derived. To cope with multiple interacting agents, the power control problem has been formalized as a Markov Game. Concerning the solution algorithm, an original Distributed Reinforcement Learning algorithm for Power Control in Wireless Networks has been presented. Said algorithm leverages on a multi-agent average cost temporal difference learning algorithm which has been developed by the authors.

The most relevant aspect of the proposed approach is its distributed nature. Indeed, this allows to guarantee (i) scalability, (ii) robustness, (iii) low computational and communication costs.

A simple proof of concept has been presented to prove the effectiveness of the proposed distributed framework. As shown, the proposed solution is able to efficiently adapt the transmission power in response to the environment's variations.

Since the proposed algorithm was only tested in a simulation scenario, it is worth mentioning that in the case of a real scenario, with mobile receivers or transmitters, frequent communication connection and disconnection, and unreliable communication links between the agents (i.e., the transmitters), the dynamic consensus algorithm as proposed can issue limitations in the convergence when tracking the global average value. To avoid the time-varying network effects introduced by the behaviours mentioned above, several countermeasures can be used, resulting in a more complex algorithm but able to provide the same result.

The authors are currently working on an extension of the proposed framework to take into account personalized QoS constraints. Indeed, as mentioned in Section V, the developed mathematical model is sufficiently flexible to capture additional end-users' and network operators' features.

REFERENCES

[1] Dorri, Ali, Salil S. Kanhere, and Raja Jurdak. "Multi-agent systems: A survey." IEEE Access 6 (2018): 28573-28593.

[2] Chen, Fei, and Wei Ren. "On the control of multi-agent systems: A survey." Foundations and Trends® in Systems and Control 6.4 (2019): 339-499.

[3] Lewis, Frank L., et al. Cooperative control of multi-agent systems: optimal and adaptive design approaches. Springer Science & Business Media, 2013.

[4] Dimarogonas, Dimos V., Emilio Frazzoli, and Karl H. Johansson. "Distributed event-triggered control for multi-agent systems." IEEE Transactions on Automatic Control 57.5 (2011): 1291-1297.

[5] Zhang, Z., Zhang, W., Zeadally, S., Wang, Y., & Liu, Y. (2015). Cognitive radio spectrum sensing framework based on multi-agent architecture for 5G networks. IEEE Wireless Communications, 22(6), 34-39.

[6] Liberati, Francesco, et al. "Stochastic and exact methods for service mapping in virtualized network infrastructures." International Journal of Network Management 27.6 (2017): e1985.

[7] Ornatelli, Antonio, Andrea Tortorelli, and Francesco Liberati. "A distributed reinforcement learning approach for power control in wireless networks." 2021 IEEE World AI IoT Congress (AIIoT). IEEE, 2021.

[8] Zhou, Hao, Medhat Elsayed, and Melike Erol-Kantarci. "RAN Resource Slicing in 5G Using Multi-Agent Correlated Q-Learning." 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2021.

[9] Nair, Arun Sukumaran, et al. "Multi-agent systems for resource allocation and scheduling in a smart grid." Technology and Economics of Smart Grids and Sustainable Energy 3.1 (2018): 1-15.

[10] Park, Kui-Hong, Yong-Jae Kim, and Jong-Hwan Kim. "Modular Q-learning based multi-agent cooperation for robot soccer." Robotics and Autonomous systems 35.2 (2001): 109-122.

[11] Franchi, Enrico, and Agostino Poggi. "Multi-agent systems and social networks." Handbook of Research on Business Social Networking: Organizational, Managerial, and Technological Dimensions. IGI Global, 2012. 84-97.

[12] Sutton, Richard S., and Andrew G. Barto. "Reinforcement learning: An introduction". MIT press, 2018.

[13] Canese, Lorenzo, et al. "Multi-Agent Reinforcement Learning: A Review of Challenges and Applications." Applied Sciences 11.11 (2021): 4948.

[14] Busoniu, Lucian, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38.2 (2008): 156-172.

[15] Zhang, Kaiqing, Zhuoran Yang, and Tamer Başar. "Multi-agent reinforcement learning: A selective overview of theories and algorithms." Handbook of Reinforcement Learning and Control (2021): 321-384.

[16] Kim, Junhyeong, et al. "5G-ALLSTAR: An integrated satellite-cellular system for 5G and beyond." 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2020.

[17] A. Hassebo, M. Obaidat and M. A. Ali, "Commercial 4G LTE cellular networks for supporting emerging IoT applications," 2018 Advances in Science and Engineering Technology International Conferences (ASET), 2018, pp. 1-6, doi: 10.1109/ICASET.2018.8376832.

[18] Tse, David, and Pramod Viswanath. Fundamentals of wireless communication. Cambridge university press, 2005.

[19] Kumar, Anurag, D. Manjunath, and Joy Kuri. Wireless networking. Elsevier, 2008.

[20] Littman, Michael L. "Markov games as a framework for multi-agent reinforcement learning." Machine learning proceedings 1994. Morgan Kaufmann, 1994. 157-163.

[21] Shapley, Lloyd S. "Stochastic games." Proceedings of the national academy of sciences 39.10 (1953): 1095-1100.

[22] Van Der Wal, Johannes. Stochastic dynamic programming. Diss. Methematisch Centrum, 1980.

[23] Tsitsiklis, John N., and Benjamin Van Roy. "Average cost temporal-difference learning." Automatica 35.11 (1999): 1799-1808.

[24] Moallemi, Ciamac Cyrus, and Benjamin Van Roy. "Consensus propagation." IEEE Transactions on Information Theory 52.11 (2006): 4753-4766.

[25] Shah, Devavrat. "Gossip Algorithms." Foundations and Trends® in Networking 3.1 (2009): 1-125.

[26] Kia, Solmaz S., et al. "Tutorial on dynamic average consensus: The problem, its applications, and the algorithms." IEEE Control Systems Magazine 39.3 (2019): 40-72.

[27] Zhu, Minghui, and Sonia Martínez. "Discrete-time dynamic average consensus." Automatica 46.2 (2010): 322-329.