

## Article

# Multi-Agent Dynamic Resource Allocation in 6G in-X Subnetworks with Limited Sensing Information

Ramoni Adeogun <sup>1,†</sup> \* and Gilberto Berardinelli <sup>1,†</sup> 

<sup>1</sup> Department of Electronic Systems, Aalborg University, Aalborg, Denmark; {ra,gb}@es.aau.dk

\* Correspondence: ra@es.aau.dk; Tel.: +45-9940-7642

† These authors contributed equally to this work.

**Abstract:** In this paper, we investigate dynamic resource selection in dense deployments of a recent 6G mobile in-X subnetworks (inXSs). We cast resource selection in inXSs as a multi-objective optimization problem involving maximization of the minimum capacity per inXS while minimizing overhead from intra-subnetwork signalling. Since inXSs are expected to be autonomous, selection decisions are made by each inXS based on its local information without signalling from other inXSs. A multi-agent Q-learning (MAQL) method based on limited sensing information (SI) is then developed resulting in low intra-subnetwork SI signalling. We further propose a rule-based algorithm termed *Q-Heuristics* for performing resource selection based on similar limited information as the MAQL method. We perform simulations with focus on joint channel and transmit power selection. The results indicate that: 1) appropriate settings of Q-learning parameters leads to fast convergence of the MAQL method even with two-level quantization of the SI; 2) the proposed MAQL approach significantly better performance and is more robust to sensing and switching delays than the best baseline heuristic. The proposed Q-Heuristic shows similar performance to baseline greedy method at the 50th percentile of the per user capacity and slightly better at lower percentiles. The Q-Heuristic method shows high robustness to sensing interval, quantization threshold and switching delay.

**Keywords:** 6G; reinforcement learning; in-X subnetworks; resource allocation; Q-learning; industrial control



**Citation:** Adeogun, R.; Berardinelli, G. Multi-Agent Dynamic Resource Allocation in 6G in-X Subnetworks with Limited Sensing Information. *Sensors* **2021**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Short-range low-power in-X subnetworks (inXSs) [1–3] are receiving attention as potential radio concepts for supporting extreme communication requirements, e.g., reliability above 99.99999, up to 10 Gbps data rate and latencies below 100  $\mu$ s. Similar extreme connectivity requirements have also appeared in recent works on visions for 6th generation (6G) networks [4,5]. InXSs are expected to provide seamless support for applications such as industrial control at the sensor-actuator level, intra-vehicle control, in-body networks and intra-avionics communications even in the absence of connectivity from traditional cellular network [2,6]. Clearly, these applications represent life critical use-cases necessitating the need to guarantee specified communication requirements everywhere. Such use-cases can also lead to dense scenarios (e.g., inXSs inside a large number of vehicles at a road intersection) leading to potentially high interference levels and hence, the need for efficient interference management mechanisms.

Interference management via dynamic allocation (DA) of shared radio resources has been at the forefront of wireless communication research for several years, see e.g., [7]. Although several techniques for resource allocation have been studied, the extreme requirements as well as the expected ultra-dense deployments of inXSs makes the interference problem more challenging. This has resulted in a number of published works on resource allocation for wireless networks with uncoordinated deployment of short range subnetworks [8,9]. In [8], distributed heuristic algorithms were evaluated and compared with a

centralized graph coloring (CGC) baseline in dense deployments of inXSs. In [9], a supervised learning method for distributed channel allocation is proposed for inXSs. The works so far focus on only channel selection making their applicability to other resource selection problems such as the joint channel and power, and channel aggregation considered in this paper non-trivial. Moreover, the reliance on full sensing information (SI) by these methods imposes significant overhead on required device capabilities (and hence, cost) as well as radio resources for intra-subnetwork signalling.

To overcome these limitations, we conjecture that reinforcement learning (RL) methods [10–12] can be developed to perform resource selection, with potential performance improvement over existing approaches even with only quantized information. Moreover, a RL based method will eliminate the offline data generation requirement for the method in [9]. The idea is to equip each cell with an agent which learns to adapt resource usage to changing interference conditions.

RL based methods are becoming increasingly popular in radio resource management (RRM) due to their ability to learn complex decision problems, e.g., allocation of multi-dimensional transmission resources [13] in wireless systems. In particular, multi-agent RL (MARL) is quite popular in recent times due to its capability of achieving potentially optimal distributed intelligent management of resources. The main advantages of MARL include the ability to: 1) support heterogeneous agents with varying requirements; 2) model local interactions among agents and 3) distribute computation among agents. To this end, there has been an increase in the number of works applying MARL to RRM in different types of wireless systems, e.g., unmanned aerial vehicle (UAV) communication [11], multi-user cellular systems [14], industry4.0 device-to-device communication [15], multi-beam satellite systems [16], integrated access and backhaul networks [17], non-orthogonal multiple access [18], multi-cell networks [19], and joint scheduling of enhanced mobile broadband and URLLC in 5th generation (5G) systems [20]. Other studies have applied RL to wireless resource allocation in sensor networks for smart agriculture [21], smart ocean federated learning based IoT networks [15], and distributed antenna systems [22].

While these studies have shown the potential for learning reasonably good solutions to radio resource optimization problems, they have been predominantly based on the assumption of full environment information and some form of information exchange among the agents. These limit their applicability in practical wireless systems where the overhead associated with signalling of information is an important parameter to be kept at the minimum.

We address the problem of fully distributed and dynamic selection of radio resources for downlink transmission by inXSs operating over a finite number of shared frequency channels. Considering the practical constraints (e.g., cost, processing power, etc) associated to signalling of sensing data and channel selection decisions between devices and access points in inXSs, we restrict resources for sensing information and decision exchange (SIDE) to only a single bit per channel. The goal is then to develop a distributed learning method for resource selection based on limited sensing data. Although Deep Q-learning (DQN) [17] which relies on Deep Neural Networks (DNNs) to learn the mapping between sensing measurements and resource selection decisions have been popular owing to their relatively better scalability compared to classical *table-based Q-learning*, the simplicity of the latter makes them attractive for low-cost radio systems. We therefore focus on developing MAQL method for dynamic resource selection with lookup tables as the policy. This is reasonable in practical wireless systems since the size of actions and sensing measurements is bounded by the limited available radio resources, making scalability not much of a problem, particularly, in the case of fully distributed implementations involving only local measurements and individual action selection.

In summary, the main contributions of this paper include:

- We cast the resource selection task into a non-convex multi-objective optimization problem involving maximization of the sum capacity at each inXS subject to power, transmission bandwidth and signalling overhead constraints.

- We develop a multi-agent Q-learning (MAQL) solution to solve the problem in a fully distributed manner. To limit the overhead associated with intra-subnetwork signalling, we constrained information exchange within each inXS to a 1-bit channel and adopt a two-level (i.e., 0 and 1) quantization of the SI.
- We further develop an alternative heuristic selection method which utilizes similar quantized information as the MAQL. The algorithm termed *Q-Heuristic* involves selection of a resource (or resources) randomly either from the list of resources in level 1 or from the list of all resources in case there is no resource in level 1.
- We apply the MAQL method to the problem of joint channel and transmit power selection for mobile 6G in-XSs. We perform simulations in typical industrial factory settings to evaluate performance gains relative to baseline heuristics with full information and the proposed Q-Heuristic. Unlike existing studies on MAQL for wireless resource management, the simulations include evaluation of the impact of delayed sensing information which may be inevitable in practice. Extensive evaluation of the sensitivity of the proposed methods to main design parameters including quantization threshold and switching delay is also performed.

The remainder of this paper is organized as follows: The system and channel models as well as a description of the resource allocation problem is presented in section 2. The proposed MAQL and Q-Heuristic methods are described in section 3. This is followed by performance evaluation in section 4. Conclusions are finally drawn in section 5.

## 2. System Model and Problem Formulation

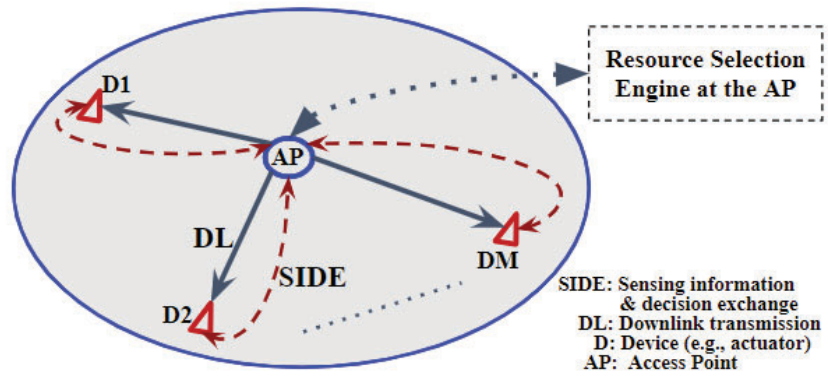
We consider the downlink (DL) of a wireless network with  $N$  independent and mobile inXSs each serving one or more devices (including sensors and actuators). The set of all inXSs in the network and the  $M_n$  devices in the  $n$ th inXS are denoted as  $\mathcal{N} = \{1, \dots, N\}$  and  $\mathcal{M}_n = \{1, \dots, M_n\}$ , respectively. As illustrated in Fig. 1, each inXS is equipped with an access point (AP) which coordinates transmissions with all associated devices. The AP is equipped with a local resource selection engine for making decisions based on local sensing data received from its associated devices via a 1-bit SIDE link as shown in Fig. 1.

The inXSs move following a specified mobility pattern which is determined by the application - e.g., inXSs deployed inside mobile robots for supporting factory operations. At any instant, transmissions within each inXS is performed over one of  $K$  ( $K \ll N$ ) shared orthogonal frequency channels denoted as  $\mathcal{K} = \{1, \dots, K\}$  with a transmit power level within the range,  $[\kappa_{\min}, \kappa_{\max}]$ , where  $\kappa_{\min}$  and  $\kappa_{\max}$  are the minimum and maximum allowed transmit power levels, respectively. To simplify the problem, we restrict the possible transmit power to a set of  $Z$  discrete levels,  $\mathcal{Z} = \{1, \dots, Z\}$ . We assume that transmissions within each inXS are orthogonal and hence, there is no intra-subnetwork interference. This assumption is reasonable since the APs can be designed to allocate orthogonal time-frequency resources to its own devices and has also been made in [1,2].

### 2.1. Channel Model and Rate Expression

The radio channel between the APs and devices in the network is characterized by three components: large scale fading, i.e., path-loss and shadowing and the small-scale effects. The path-loss on a link from node  $A$  to node  $B$  with distance,  $d_{AB}$  is defined as  $L_{AB} = c^2 d_{AB}^{-\alpha} / 16\pi^2 f^2$ , where  $c \approx 3 \times 10^8 \text{ ms}^{-1}$  is the speed of light,  $f$  is the carrier frequency and  $\alpha$  denotes the path-loss exponent. A correlated log-normal shadowing model based on a 2D Gaussian random field is considered [23]. We compute the shadowing on the link from  $A$  to  $B$  using

$$X_{AB} = \ln \left\{ \frac{1 - e^{\left(-\frac{d_{AB}}{d_c}\right)}}{\sqrt{2} \sqrt{1 + e^{\left(-\frac{d_{AB}}{d_c}\right)}}} (\mathbf{S}(A) + \mathbf{S}(B)) \right\}, \quad (1)$$



**Figure 1.** Illustration of DL transmission, and sensing information and resource selection decision exchange in a single inXS.

where  $\mathbf{S}$  is a two-dimensional Gaussian random process with exponential covariance function and  $d_c$  denotes the correlation distance. The small scale fading,  $h$  is assumed to be Rayleigh distributed. The Jake's Doppler model is utilized to capture temporal correlation of  $h$  [24].

At a given transmission instant,  $t$ , the received (or interference) power on the link between any two nodes, e.g., from A to B is computed as:

$$P_{AB}(\kappa_A(t)) = \kappa_A(t) L_{AB}(t) X_{AB}(t) |h_{AB}(t)|^2, \quad (2)$$

where  $\kappa_A(t)$  denotes the transmit power (in linear scale) of node A at time  $t$ . Assuming that the  $n$ th inXS operates over frequency channel,  $c_k : k \in \mathcal{K}$  at time  $t$ , the received signal to interference and noise ratio (SINR) from its  $m$ th device can be expressed as

$$\gamma_{nm}(c_k, \kappa^k(t)) = \frac{P_{nm}(c_k, \kappa_n^k(t))}{\sum_{i \in \mathcal{I}_k(t)} P_{ni}(c_k, \kappa_i^k(t)) + \sigma_{nm}^2(t)}, \quad (3)$$

where  $\mathcal{I}_k(t)$  and  $\kappa^k(t)$  denotes the set all devices (or APs) transmitting on channel  $c_k$  at time  $t$  and their transmit powers, respectively. The term  $\sigma_{nm}^2(t)$  is the receiver noise power calculated as  $\sigma_{nm}^2(t) = 10^{(-174 + \text{NF} + 10 \log_{10}(W_k))}$ , where  $W_k$  denotes the bandwidth of  $c_k$  and NF is the receiver noise figure. Relying on the Shannon approximation, the achieved capacity can be written as

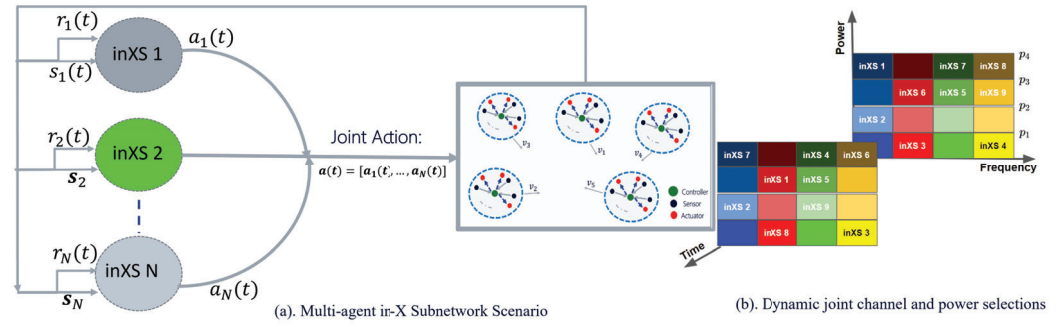
$$\zeta_{nm}(c_k, t) \approx W_k \log_2(1 + \gamma_{nm}(c_k, \kappa^k(t))). \quad (4)$$

## 2.2. Problem Formulation

In this paper, we consider a resource allocation problem involving fully distributed joint channel and power selection. This problem can be defined as multi-objective optimization tasks involving simultaneous maximization of  $N$  objective functions, one for each inXS. Taking the objective function as the lowest achieved capacity at each inXS (denoted  $\zeta_n = \min(\{\zeta_{nm}\}_{m=1}^{M_n}); \forall n \in \mathcal{N}$ ), the problem can, formally, be defined as:

$$\begin{aligned} \text{P-I: } & \max_{\mathbf{c}, \kappa} \zeta_1(c_1(t), \kappa_1(t)), \dots, \max_{\mathbf{c}, \kappa} \zeta_N(c_N(t), \kappa_N(t)) \\ \text{st: } & \kappa_{\min} \leq \kappa_n \leq \kappa_{\max} \quad \text{and} \quad \text{BW}(c_k) = W_k \quad \forall n, \end{aligned} \quad (5)$$

where  $\mathbf{c} := \{c_n | n = 1, \dots, N\}$  and  $\kappa := \{\kappa_n | n = 1, \dots, N\}$  denotes the set of channel indices and transmit powers for all inXSs, respectively. The term  $\text{BW}(c_k)$  denotes bandwidth of channel,  $c_k$ .



**Figure 2.** Illustration of the Multiagent RL scenario with N inXSs.

The problem in (5) involves joint optimization of multiple conflicting non-convex objective functions and is typically difficult to solve. The independence of the inXSs and the lack of communication coupled with the desire to minimize overhead due to intra-subnetwork signalling via quantization further aggravate the problem. We present a MAQL method with quantized SI for solving this problem in section 3. An alternative rule-based solution referred to as *Q-Heuristic* is also presented.

### 3. Resource Selection with limited Information

We cast the joint optimization problem in (5) as Multi-Agent Markov Decision Process (MMDP) [25] described as the tuple  $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$ , where  $\mathcal{S} = \mathcal{S}_1 \times \dots \times \mathcal{S}_N$  is a set of all possible states for all inXSs referred to as state space,  $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_N$  is the joint action space containing all possible actions (i.e., the set of all possible combinations of channels and power levels),  $\mathcal{R}$  denotes the reward signal and  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \Delta$  is the transition function [25], where  $\Delta$  denotes the set of probability distributions over  $\mathcal{S}$ .

In the considered MMDP, the goal of the  $n$ th agent is to find an *optimal* policy,  $\pi_n^*$  based solely on its local state and action information resulting in the so called Partially Observable MMDP (POMMDP) [26]. Typically,  $\pi_n^*$  is obtained as the policy which maximizes the total reward function [18], i.e.,

$$\pi_t^*(s) = \max_{\pi_t(s) \in \mathcal{A}} \left\{ r_t(s_t, \pi_t(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s_t, s') \pi_{t+1}^*(s') \right\}, \quad (6)$$

where  $\gamma : 0 \leq \gamma \leq 1$  denotes the discount factor. To allow mapping for all possible state-action pairs, an alternative representation,  $Q(s, a)$  referred to as the Q-function is commonly used. The Q-function for the  $n$ th agent is given as [25]

$$Q^n(s, a) = r^n(s, a) + \gamma \max_{a'} Q^n(s', a') \quad (7)$$

Since each agent has access to only local information, solving (7) results in a local maximum at each subnetwork. We assume that the local maxima on each of the  $N$  agents' Q-function is equivalent to the global maximum on the joint Q-function for the entire network, i.e.,

$$\arg \max_{\mathbf{a}} Q^\pi(\mathbf{s}, \mathbf{a}) = \begin{bmatrix} \arg \max_a Q^1(s, a) \\ \vdots \\ \arg \max_a Q^N(s, a) \end{bmatrix}. \quad (8)$$

According to (8), solution to the resource selection problem can now be obtained via local optimization at each inXS. MAQL enables solution of the  $N$  objectives via simultaneous



**Algorithm 1** Multi-Agent Resource Allocation with quantized SI: Training Procedure

---

**Input:** Simulation and environment parameters, learning rate,  $\alpha$ , discount factor,  $\gamma$ , number of episodes,  $T$ , number of steps per episode,  $N_e$ ,  $\epsilon_{\min}$ ,  $\epsilon_{\max}$   
 Start simulator, randomly drop cells and generate shadowing map  
 $t = 1; \epsilon = \epsilon_{\max}$   
 Initialize actions for all cells randomly and compute initial states,  $\{\mathbf{s}_n(1)\}_{n=1}^N$   
 Initialize Q-table,  $Q$  with zeros  
**for**  $t = 1$  **to**  $T$  **do**  
   **for**  $i = 1$  **to**  $N_e$  **do**  
   **for**  $n = 1$  **to**  $N$  **do**  
   Obtain state from SI  $\mathbf{s}_n(t)$   
   Subnetwork  $n$  select  $a_n(t)$  according to (10).  
**end for**  
 The joint resource selection of all subnetworks generate transitions into next states,  $\{\mathbf{s}_n(t+1)\}_{n=1}^N$  and immediate rewards,  $\{r_n(\mathbf{s}(t), \mathbf{a})\}_{n=1}^N$   
 Decay exploration probability using  $\epsilon = \max(\epsilon_{\min}, (\epsilon_{\max} - \epsilon_{\min})/\epsilon_{\text{step}})$   
**for**  $n = 1$  **to**  $N$  **do**  
   Update  $Q$  using  $Q(s_t, a) = (1 - \alpha)Q(s_t, a) + \alpha(r(s_t, a) + \gamma \max_{a'} Q'(s_{t+1}, a'; \pi))$   
**end for**  
**end for**  
**end for**  
**Output:** Trained Q-table,  $Q$   
 %% The table,  $Q$  is copied to all APs

---

interaction of all agents with the environment. The Q-function is iteratively estimated according to Bellman's equation as [27]

$$Q^n(s_t, a) = (1 - \alpha)Q^n(s_t, a) + \alpha(r(s_t, a) + \gamma \max_{a'} Q^n(s_{t+1}, a'; \pi)) \quad \forall n, \quad (9)$$

where  $\alpha$  denotes the learning rate and  $r_n(s_t, a)$  is the instantaneous reward received by the agent for selecting action,  $a \in \mathcal{A}$  at state  $s_t \in \mathcal{S}$ . The policy,  $\pi(s, a)$  corresponds to the conditional probability that action  $a$  is taken by an agent in state,  $s$  and must therefore satisfy  $\sum_{a \in \mathcal{A}} \pi(s, a) = 1$ .

### 3.1. MAQL Procedure for Dynamic Resource Selection

To find *optimal* estimates of the Q-functions in (9) via MAQL, we need to define the environment, state space, action space, reward signal, policy representation and training method. As described in Section 2, we consider a wireless environment with  $N$  independent inXSs each with one or more devices as illustrated in Fig. 2. The remaining components are described below.

#### 3.1.1. State and observation space

In the multi-agent scenario, the state of the environment is defined by actions of all inXSs. The achieved performance is also determined by both the *known* local characteristics of each inXS - channel gain, occupied frequency channel, transmit power level, etc, and the *unknown* information about other inXSs. We assume that each inXS has sensing capabilities for obtaining measurements of the aggregate interference power on all channels. This assumption is reasonable since each inXS device can be equipped with a transceiver that is capable of continuously performing sensing of its operational channel as well as simultaneously listening on all other channels. We denote the SI at time  $t$  as  $\mathbf{I}_n^t = [I_{n,1}^t, I_{n,2}^t, \dots, I_{n,K}^t]^T \in \mathbb{R}^{(K \times 1)}$ . To account for the effect of channel condition within each

inXS, we propose state representation based on the estimated SINR over all channels denoted for the  $n$ th inXS as  $\mathbf{s}_n^t = [s_{n,1}^t, s_{n,2}^t, \dots, s_{n,K}^t]^T$ , with  $s_{n,k} = s_d / (I_{n,k} + \sigma^2)$ , where  $s_d$  denote the received signal strength of the weakest link in the inXS. To enable Q-learning which require discrete state spaces, we perform a 2-level quantization on the SINR resulting in a state dimension of  $|\mathcal{S}| = 2^K$  comprising of all possible combinations of  $K$  channels each with two levels: 0 and 1. Denoting the SINR quantization value as  $s_{th}$ , channel,  $i$  is in state 0 if  $s_{n,i} < s_{th}$  and in state 1 otherwise.

### 3.1.2. Action space

For the joint channel and power selection task, the action space is the list of all possible combinations of available frequency channels and transmit power levels in the system. With  $K$  channels and  $Z$  discrete power levels, the action selected by inXS  $n$  at time  $t$  is from a  $KZ$ -dimensional action space comprising of all possible combinations of channel and power levels, i.e.  $a_n^t \in \mathcal{A}$ ;  $\mathcal{A} = \{\{c_1, p_1\}, \{c_1, p_2\}, \dots, \{c_K, p_Z\}\}$ .

### 3.1.3. Reward signal

Reward signal design is a crucial part of the RL design pipeline. This is typically done by considering the overall goal of the problem and how best to guide an agent towards achieving such goal. We assume that the communication metric to be maximized is the capacity of the worst link and use (4) as the reward function.

### 3.1.4. Policy Representation

The decision making component of any RL method requires a suitable framework for representing what is learnt by an agent during training. This representation is generally referred to as the policy. In this work, the policy at each inXS is represented by a  $2^K \times |\mathcal{A}|$  lookup table containing the Q-values for all state-action pairs. This has the inherent advantage of simplicity and low computation overhead since decision making is reduced to a simple look-up operation at any given time instant.

### 3.1.5. Action Selection

Resource selection decision is made by each agent via the  $\epsilon$ -greedy strategy defined as

$$a_n^t = \begin{cases} \text{a random selection} & \text{with probability, } \epsilon \\ \arg \max_{a \in \mathcal{A}(\mathbf{s}_n^t)} Q_n(\mathbf{s}_n^t, a), & \text{otherwise} \end{cases}, \quad (10)$$

where  $\epsilon$  is the exploration probability, i.e., the probability that the agent takes random action. During the training,  $\epsilon$  is decayed at each step according to

$$\epsilon = \max(\epsilon_{\min}, (\epsilon_{\max} - \epsilon_{\min}) / \epsilon_{\text{step}}), \quad (11)$$

where  $\epsilon_{\min}$  and  $\epsilon_{\max}$  denote the minimum and maximum exploration probability, respectively, and  $\epsilon_{\text{step}}$  is the number of exploration steps.

### 3.1.6. Training Procedure

Due to its better training stability and fast convergence, a *centralized training with distributed execution* framework which is popular in the multi-agent RL literature is adopted in this paper. A single Q-table is then trained by simultaneously applying it to all inXSs during the training. The procedure is described in Algorithm 1. Once the training is completed, the Q-table is copied to all inXSs for fully distributed execution.

## 3.2. Quantized Heuristic

Inspired by our initial results from the MAQL methods, we further proposed the simple *Quantized Heuristic* algorithm for resource selection based on similar 1-bit SI. The idea is to choose a channel randomly from the list of all channels in the *good* state, i.e., the

Table 1: Simulation parameters.

Deployment and system parameters	
Parameter	Value
Deployment area [m <sup>2</sup> ]	50 × 50
Number of controllers/inXSs, $N$	20
Number of devices per inXS, $M$	1
Cell radius [m]	3.0
Velocity, $v$ [m/s]	3.0
Mobility model	RRWP
Number of channels, $K$	5
Propagation and radio parameters	
Pathloss exponent, $\gamma$	2.2
Shadowing standard deviation, $\sigma_s$ [dB]	5.0
De-correlation distance, $d_c$ [m]	2
Lowest frequency [GHz]	3
Transmit power levels [dBm]	[-20 : 2 : -10]
Noise figure [dB]	10
Per channel bandwidth [MHz]	5
Q-Table and simulation settings	
Action space size, $ \mathcal{A} $	30
Discount factor, $\gamma$	0.90
Learning rate, $\alpha$	0.80
Number of training episodes/Steps per episode	3000/200
Minimum/maximum exploration probability	0.01/0.99
Number of epsilon greedy steps	$4.8 \times 10^5$

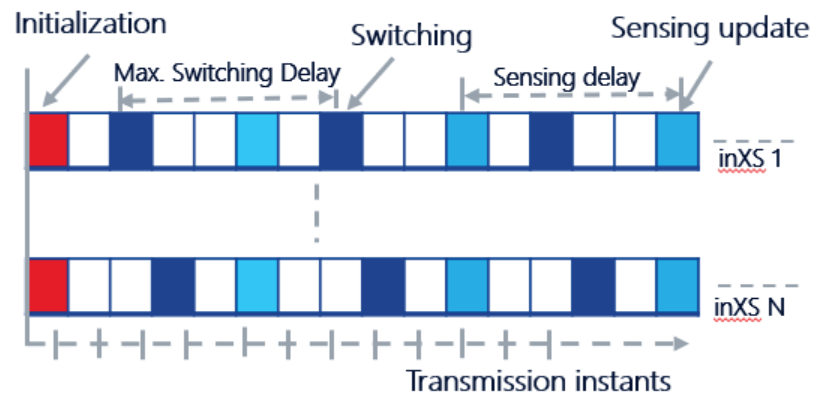
state with SINR above the quantization threshold,  $s_{th}$ . If no channel is in the good state, the channel is chosen randomly from the list of all channels.

#### 4. Performance Evaluation

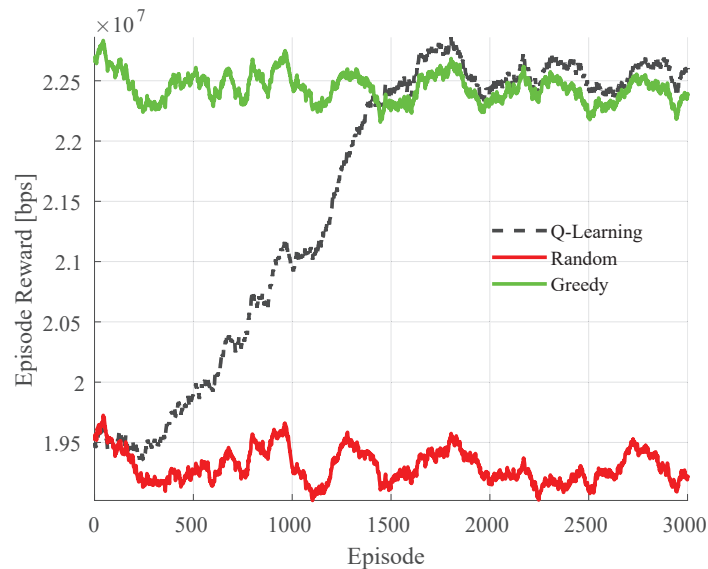
We now train and evaluate the performance of the MAQL approach and compare with fixed (i.e., random assignment at initialization without dynamic updates), greedy channel selection and Q - Heuristic using a snapshot based procedure. Except where otherwise stated, we consider a network with  $N = 20$  inXSs each with a single controller serving as the AP for a sensor-actuator pair in a 50 m × 50 m rectangular deployment area. Each inXS move in the area following a restricted random waypoint mobility (RRWP) with a constant speed,  $v = 3$  m/s. We assume that a total bandwidth  $B = 25$  MHz is available in the system and that the bandwidth is partitioned into  $K = 5$  channels. Similar to [6,8], we set the transmit power for all inXSs to -10 dBm for all algorithms except MAQL, for which we consider a total of  $Z = 6$  transmit power levels between -20 dBm and -10 dBm at intervals of 2 dB, leading to a  $30 \times 1$  action space. The power difference of  $\pm 2$  dB is used to ensure reasonable granularity in transmit power levels. Other simulation parameters are shown in Tab. 1. The deployment and system parameters are defined based on the settings used in [6,8]. The propagation model as well as its parameters are selected from 3GPP documents on channel models for industrial scenarios [28,29].

Motivated by the results in [8,9], we introduced random switching delays with a maximum value of  $\tau_{max} = 10$  transmission intervals in the simulation. This is to minimize *ping-pong* effects where multiple inXSs simultaneously switch to the same resource. Each inXS is then allowed to switch its operational resource once every 10 transmission instants. The specific time instant at which an inXS has the opportunity to update its transmit power level and/or operational frequency channel is determined by a random integer between 1 and 10. The random integer is assigned to each inXS at the beginning of each snapshot. The concept of switching delay as well as sensing interval is illustrated in Fig. 3. Except where stated otherwise, we assume perfect sensing such that measurements for making resource selection and switching decisions are up-to-date with no errors or noise. To understand the impact of imperfect information on achieved performance by the different techniques,





**Figure 3.** Sensing measurement updates and resource selection with both maximum switching delay and sensing delay equal to 5. InXSs 1 and  $N$  are assigned random switching integers 2 and 3, respectively. At initialization, all inXs perform random resource selection



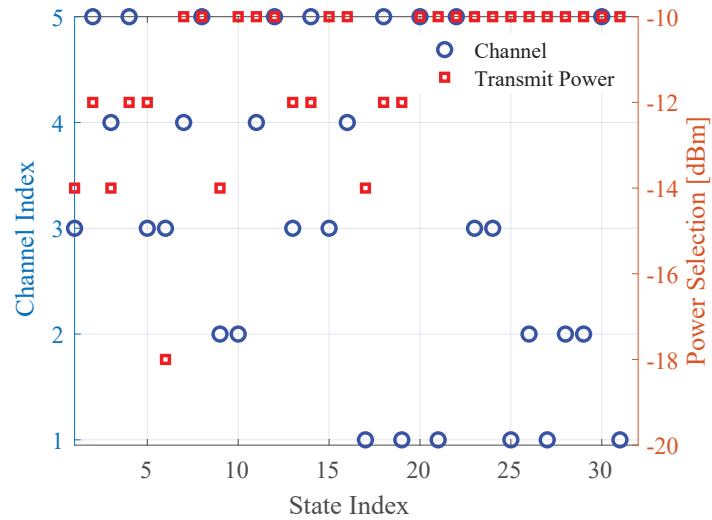
**Figure 4.** Averaged reward per episode during MAQL training for joint channel and power selection with  $s_{th} = 2$  dB.

we evaluate the algorithms with varying sensing intervals, i.e., time interval between successive update of sensing measurements at each inXS, see the illustration in Fig. 3. The results are presented in section 4.3.

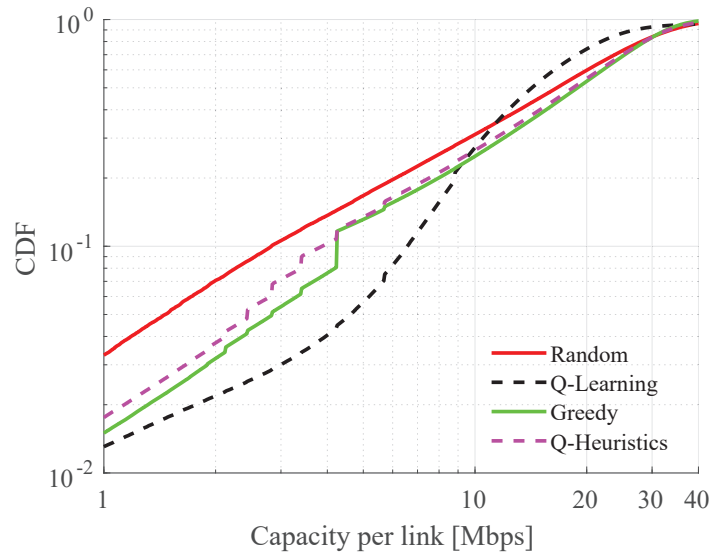
#### 4.1. Training, Convergence and Learned Policy

Fig. 4 shows the averaged reward over successive training episodes for the joint power and channel selection problem with SINR quantization threshold,  $s_{th} = 2$  dB. The averaging is performed over all steps within each episode as well as all inXSs. We benchmark the reward with those obtained from 2 heuristic algorithms viz: random and greedy channel selection. The maximum transmit power of -10 dBm is used for all inXSs in the heuristic algorithms. The figure shows that the proposed MAQL achieve convergence after approximately 1700 episodes. At convergence, the MAQL method has marginally better performance than the greedy selection baseline with full SI [8].

To understand the actions of the Q-agents, we show the learned Q-policy at convergence in Fig. 5. The policy comprises of the channel and transmit power pairs with maximum Q-value at each of the 32 ( $2^5$ ) states. The figure shows that the Q-agents converge to a channel with quantization level of 1 (i.e., with  $SINR \geq s_{th}$ ) for all states except for state



**Figure 5.** Learned Q-policy at convergence of the MAQL training for joint channel and power selection with  $s_{th} = 2$  dB.

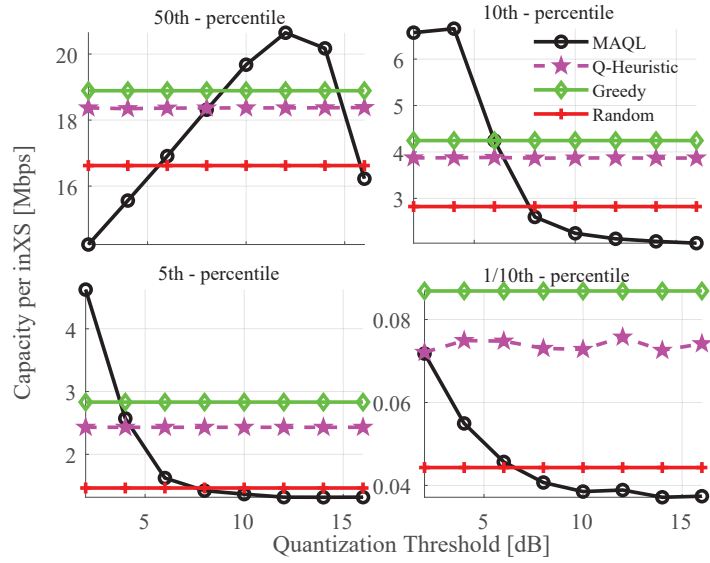


**Figure 6.** CDF of capacity per inXS with  $s_{th} = 2$  dB

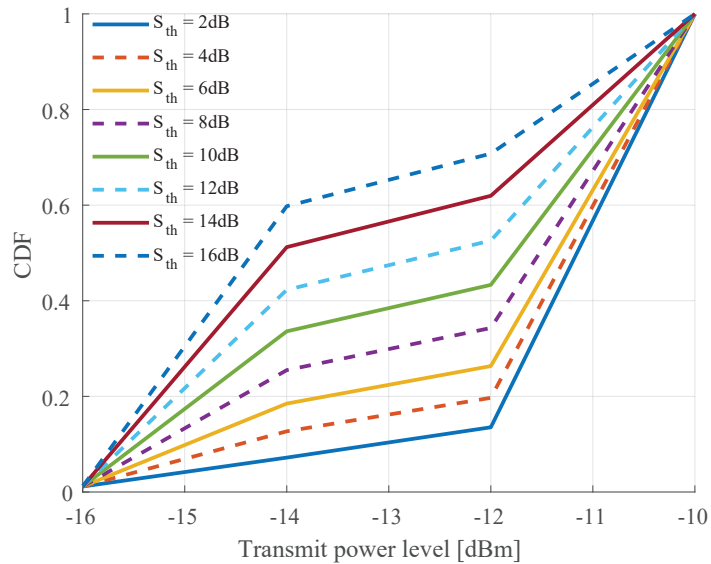
1 which has no channel in level 1. As shown in the figure, the power levels of -10 dBm, -12 dBm, -14 dBm and -18 dBm are preferred by the agents in the ratio 21:6:4:1. Two power levels viz: -20 dBm and -16 dBm are never chosen with full exploitation.

#### 4.2. Comparison with Benchmark Schemes

The trained Q-table is deployed at each inXS for distributed resource selection and performance compared with random, greedy channel selection and the proposed Q-Heuristic. Except for MAQL, all algorithms use the maximum transmit power of -10 dBm per transmission as mentioned above. Fig. 6 shows the empirical Cumulative Distribution Function (CDF) of the achieved capacity per inXS with sensing-to-action time (i.e., sensing interval) of a single time slot. The proposed MAQL method performs significantly better than simple random selection, Q-Heuristic, and greedy selection with full SI below the 30th percentile of the capacity CDF. This performance improvement appear to have been obtained at the expense of lower capacity above the same percentile. Despite using the same information as MAQL, the Q-heuristic method is only as good as the greedy baseline. A plausible explanation for the performance improvement by the MAQL is the combined effect of low SINR quantization threshold,  $s_{th}$  and utilization of different power levels.



**Figure 7.** Sensitivity to quantization threshold,  $s_{th}$ .



**Figure 8.** Distribution of selection power by the MAQL method with different  $s_{th}$  values.

#### 4.3. Sensitivity Analysis

We now present results on sensitivity of the different techniques to quantization threshold,  $s_{th}$ , sensing interval,  $\tau$  and maximum switching interval.

In Fig. 7, we plot the 50th (median), 10th, 5th and 1/10th percentiles of the capacity per inXS with test quantization thresholds between 2 dB and 16 dB using the trained policy shown in Fig. 5. Note that the training is performed with  $s_{th} = 2$  dB. The figure shows that high values of  $s_{th}$  benefit the median of per link capacity while lower values yield higher capacity at the lower percentiles. For instance, the highest 50th-, 10th- and 5th-percentiles of per inXS capacity are achieved with  $s_{th}$  values of 12 dB, 4 dB and 2 dB, respectively. Careful consideration should therefore be taken in setting the threshold based on the communication theoretic targets of the system. In Fig. 8, we evaluate the effect of  $s_{th}$  on transmit power selection. The figure indicates that increasing quantization threshold leads to higher preference of actions with lower power levels, resulting to a decrease of about  $\sim 3$  dB in the median transmit power level with change in  $s_{th}$  from 2 dB to 16 dB. A plausible explanation for this trend is that some of the 32 states becomes more likely with increasing (or decreasing) value of  $s_{th}$ .

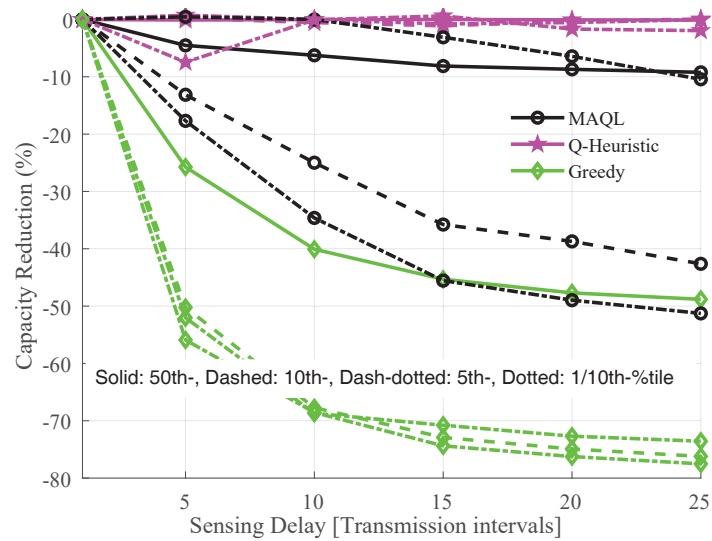


Figure 9. Sensitivity of joint channel and power selection methods to sensing delay,  $\tau$  with  $s_{th} = 2$  dB.

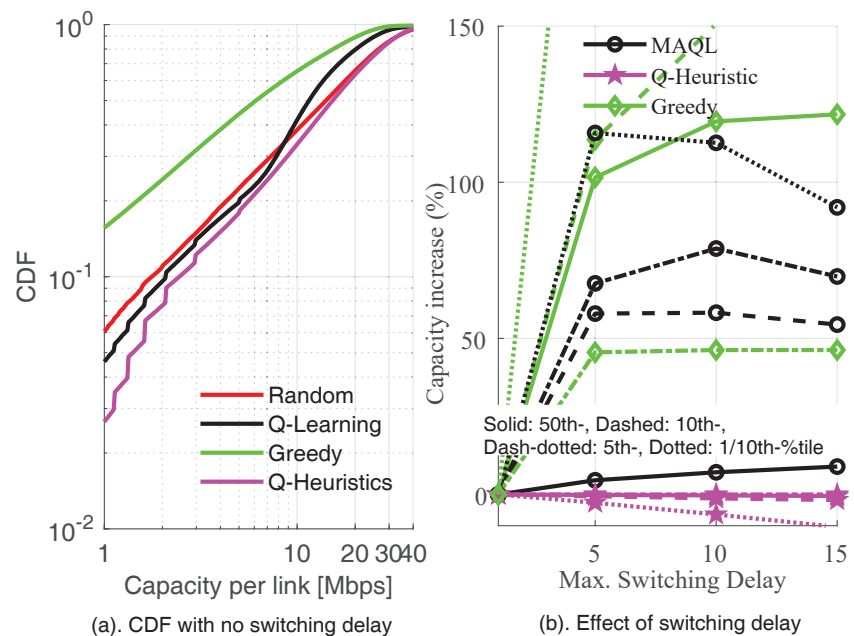


Figure 10. Different percentiles of the capacity per inXS versus maximum switching delay.

Fig. 9 shows the impact of sensing interval on performance of the MAQL, Q-Heuristic and greedy schemes. In this figure, we use the achieved capacity with perfect sensing as a baseline and plot the percentage capacity reduction with increasing sensing interval. The results show that the proposed methods with 1-bit information are in general less sensitive to sensing intervals than the greedy selection method. The Q-Heuristic method exhibits the highest robustness with little or no degradation in capacity with increasing sensing interval. Compared to greedy with up to about 80% capacity decrease, the MAQL has only 50% degradation at a delay of 25 transmission instants. This indicates that the proposed methods offer similar or better performance as the baseline but provide significant overhead reduction for SI exchange as well as better robustness to sensing intervals which may be inevitable in practice.

Finally, we study the effect of switching delay on performance of the resource selection methods in Fig. 10. In Fig. 10a, we plot the CDF of capacity per link with maximum switching delay of a single transmission interval. As a result of the simultaneous resource switching and its associated *ping-pong* effects, the greedy algorithm appear to be much worse than all other methods. This indicates that fully greedy resource selection is detri-

mental to performance in scenarios where controlled switching is not possible. Note that the performance of the MAQL is also degraded in the region below the 30th percentile when compared to Fig. 4c. To further quantify the effects of switching delay, we plot the capacity increase (in percentage) as a function of the maximum switching delay. The capacity increase at a given maximum delay value is calculated by subtracting the capacity value from its value with no delay. As shown in the figure, it is indeed beneficial to minimize *ping-pong* effects by introducing the switching delay as stated in [9]. Except for the Q-Heuristic which appears to be quite robust to switching delay, a maximum delay of 5 transmission intervals yields capacity increase above 100% for both MAQL and greedy selection methods. As seen in the figure, the greedy method is much more sensitive to switching delays than the proposed MAQL method which exhibit quite marginal sensitivity at the median of achieved capacity.

We remark here that although the performance evaluation presented in this section are based on 3GPP models for industrial environment [29], it is often useful to study the sensitivity of the new methods to variations in the wireless environment. For instance, the MAQL method can be evaluated with environment parameters, deployment density and/or configurations that are different from those used during the training leading to understanding of the ability of the proposed method to generalize to other settings. Such sensitivity analysis is however left for future work. The methods proposed in this paper also consider a single bit per channel which represents the lowest overhead for signaling information about the status of each channel within each inXS. It may then be possible to improve performance of the proposed schemes with increased number of bits per channel. Since in-XSs are expected to be low cost radio devices, we believe that the best solutions are those which require minimum signalling overhead without significant performance degradation. Another interesting avenue for further study would be to quantify the trade-off between performance and signaling overhead.

## 5. Conclusion

Multi-agent Q-learning for distributed dynamic resource selection with quantized SI can achieve better performance to the best known heuristics (i.e., greedy selection) with full information in 6G in-X subnetworks. This is particularly true for the low percentile of the capacity per link and depends on appropriate selection of the value of the SINR quantization threshold,  $s_{th}$ . With low  $s_{th}$  values (e.g., between 2 dB – 4 dB), the MAQL method performs better than both greedy and Q-Heuristic schemes at the 10th, 5th and 1/10th percentiles of per link capacity but worst at the 50th percentile. In contrast, higher  $s_{th}$  values (e.g., between 10 dB and 14 dB) benefit the 50th percentile of capacity per link but suffers the lower percentiles. Simulation results have shown that the proposed *look-up table* based MAQL method exhibits fast convergence and are more robust to sensing intervals and switching delays than greedy resource selection. A proposed alternative rule-based scheme based on similar 1-bit SI as the MAQL offer improved robustness with similar performance as the greedy selection baseline. Our ongoing work is investigating other learning based methods with capability for optimal performance while eliminating the need for controlled switching via introduction of switching delays.

**Author Contributions:** Conceptualization, R. Adeogun and G. Berrardinelli; methodology, R. Adeogun.; software, R. Adeogun; validation, R. Adeogun and G. Berrardinelli; formal analysis, R. Adeogun and G. Berrardinelli; investigation, R. Adeogun; resources, R. Adeogun and G. Berrardinelli; data curation, R. Adeogun; writing—original draft preparation, R. Adeogun; writing—review and editing, R. Adeogun and G. Berrardinelli; visualization, R. Adeogun; funding acquisition, G. Berrardinelli. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Danish Council for Independent Research, grant no. DFF 9041- 00146

**Conflicts of Interest:** "The authors declare no conflict of interest."



## References

1. Berardinelli, G.; Baracca, P.; Adeogun, R.; Khosravirad, S.; Schaich, F.; Upadhya, K.; Li, D.; Tao, T.B.; Viswanathan, H.; Mogensen, P.E. Extreme Communication in 6G: Vision and Challenges for 'in-X' Subnetworks. *IEEE OJCOM* **2021**.
2. Adeogun, R.; Berardinelli, G.; Mogensen, P.E.; Rodriguez, I.; Razzaghpour, M. Towards 6G in-X Subnetworks With Sub-Millisecond Communication Cycles and Extreme Reliability. *IEEE Access* **2020**, *8*, 110172–110188.
3. Berardinelli, G.; Mahmood, N.H.; Rodriguez, I.; Mogensen, P. Beyond 5G Wireless IRT for Industry 4.0: Design Principles and Spectrum Aspects. 2018 IEEE Globecom Workshops (GC Wkshps), 2018, pp. 1–6. doi:10.1109/GLOCOMW.2018.8644245.
4. Ziegler, V.; Viswanathan, H.; Flinck, H.; Hoffmann, M.; Räisänen, V.; Hätönen, K. 6G Architecture to Connect the Worlds. *IEEE Access* **2020**, *8*, 173508–173520.
5. Viswanathan, H.; Mogensen, P.E. Communications in the 6G Era. *IEEE Access* **2020**, *8*, 57063–57074.
6. Adeogun, R.; Berardinelli, G.; Mogensen, P.E. Enhanced Interference Management for 6G in-X Subnetworks. *IEEE Access* **2022**, *10*, 45784–45798. doi:10.1109/ACCESS.2022.3170694.
7. Hussain, F.; Hassan, S.A.; Hussain, R.; Hossain, E. Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges. *IEEE Commun. Surveys Tuts.* **2020**, *22*, 1251–1275.
8. Adeogun, R.; Berardinelli, G.; Rodriguez, I.; Mogensen, P.E. Distributed Dynamic Channel Allocation in 6G in-X Subnetworks for Industrial Automation. *IEEE Globecom Workshops*, 2020.
9. Adeogun, R.O.; Berardinelli, G.; Mogensen, P.E. Learning to Dynamically Allocate Radio Resources in Mobile 6G in-X Subnetworks. *IEEE PIMRC*, 2021.
10. Zhao, G.; Li, Y.; Xu, C.; Han, Z.; Xing, Y.; Yu, S. Joint Power Control and Channel Allocation for Interference Mitigation Based on Reinforcement Learning. *IEEE Access* **2019**, *7*, 177254–177265. doi:10.1109/ACCESS.2019.2937438.
11. Cui, J.; Liu, Y.; Nallanathan, A. Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. *IEEE Trans. Wireless Commun.* **2020**, *19*, 729–743.
12. Xiong, Z.; Zhang, Y.; Niyato, D.; Deng, R.; Wang, P.; Wang, L. Deep Reinforcement Learning for Mobile 5G and Beyond: Fundamentals, Applications, and Challenges. *IEEE Vehicular Technology Magazine* **2019**, *14*, 44–52.
13. Nguyen, T.T.; Nguyen, N.D.; Nahavandi, S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. *IEEE Transactions on Cybernetics* **2020**, *50*, 3826–3839. doi:10.1109/TCYB.2020.2977374.
14. Meng, F.; Chen, P.; Wu, L.; Cheng, J. Power Allocation in Multi-User Cellular Networks: Deep Reinforcement Learning Approaches. *IEEE Transactions on Wireless Communications* **2020**, pp. 1–1.
15. Kwon, D.; Jeon, J.; Park, S.; Kim, J.; Cho, S. Multi-Agent DDPG-based Deep Learning for Smart Ocean Federated Learning IoT Networks. *IEEE Internet of Things Journal* **2020**, pp. 1–1.
16. Liu, S.; Hu, X.; Wang, W. Deep Reinforcement Learning Based Dynamic Channel Allocation Algorithm in Multibeam Satellite Systems. *IEEE Access* **2018**, *6*, 15733–15742.
17. Lei, W.; Ye, Y.; Xiao, M. Deep Reinforcement Learning Based Spectrum Allocation in Integrated Access and Backhaul Networks. *IEEE Transactions on Cognitive Communications and Networking* **2020**.
18. He, C.; Hu, Y.; Chen, Y.; Zeng, B. Joint Power Allocation and Channel Assignment for NOMA With Deep Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2200–2210. doi: 10.1109/JSAC.2019.2933762.
19. Ahmed, K.I.; Tabassum, H.; Hossain, E. Deep Learning for Radio Resource Allocation in Multi-Cell Networks. *IEEE Network* **2019**, *33*, 188–195.
20. Li, J.; Zhang, X. Deep Reinforcement Learning Based Joint Scheduling of eMBB and URLLC in 5G Networks. *IEEE Wireless Comm. Letters* **2020**.
21. Tyagi, S.K.S.; Mukherjee, A.; Pokhrel, S.R.; Hiran, K.K. An Intelligent and Optimal Resource Allocation Approach in Sensor Networks for Smart Agri-IoT. *IEEE Sensors Journal* **2020**, pp. 1–1.
22. Liu, Y.; He, C.; Li, X.; Zhang, C.; Tian, C. Power Allocation Schemes Based on Machine Learning for Distributed Antenna Systems. *IEEE Access* **2019**, *7*, 20577–20584.
23. Lu, S.; May, J.; Haines, R.J. Effects of Correlated Shadowing Modeling on Performance Evaluation of Wireless Sensor Networks. *IEEE Vehicular Technology Conference*, 2015, pp. 1–5.
24. Jakes, W.C.; Cox, D.C. *Microwave mobile communications*; Wiley-IEEE press, 1994.

- 
25. Mukhopadhyay, S.; Jain, B. Multi-agent Markov decision processes with limited agent communication. *IEEE ISIC*, 2001, pp. 7–12. doi:10.1109/ISIC.2001.971476.
  26. Yang, Y.; Wang, J. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *CoRR* **2020**, *abs/2011.00583*.
  27. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, second ed.; The MIT Press, 2018.
  28. Study on Communication for Automation in Vertical Domains (Release 16). Technical Report 22.804 v16.2.0, 3rd Generation Partnership Project, 2018.
  29. Study on channel model for frequencies from 0.5 to 100 GHz (Release 16). Technical Report 38.901 v16.1.0, 3rd Generation Partnership Project, 2020.