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

Ramoni Adeogun, Gilberto Berardinelli

***Abstract*—In this letter, 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 per inXS sum capacities. 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 significant reduction in the overhead associated with intra-subnetwork SI exchanges. We perform simulations with focus on two similar but distinct resource allocation problems: joint channel and transmit power selection and channel selection with aggregation. The results indicate that: 1) appropriate settings of Q-learning parameters leads to fast convergence of the MAQL method even with 1-bit quantization of the SI; 2) the proposed MAQL approach offer similar performance and is more robust to sensing delays than the best baseline heuristic with full SI.**

***Index Terms*—6G; reinforcement learning; in-X subnetworks; resource allocation, Q-learning**

I. 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 μs. 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]. 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., [6], [7]. Although several techniques for resource allocation have been studied, the extreme latency requirement 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 letter 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], [11] 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 main contributions of this letter are as follows:

- 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 and transmission bandwidth 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 adopt a two-level quantization of the SI.
- We apply the MAQL learning to two related but distinct resource selection problems viz: joint channel and transmit power selection, and channel aggregation. We perform simulations in typical industrial factory settings to evaluate performance gains relative to baseline heuristics with full information.

II. 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 nth inXS are denoted as $\mathcal{N} = \{1, \cdots, N\}$ and $\mathcal{M}_n = \{1, \cdots, M_n\}$, respectively. Each inXS is equipped with an access point (AP) which coordinates transmissions with all associated devices. The cells move following a specified mobility pattern which is determined by the application. At any instant, transmissions within each cell is performed over one (or more) of K ($K << N$) shared orthogonal frequency channels denoted

The authors are with Dept. Electronic Systems, Aalborg University, Denmark. {ra,gb}@es.aau.dk. April 28, 2022.

as $\mathcal{K} = \{1, \cdots, K\}$ with a transmit power level within the range, $\kappa_{\min} \leq \kappa_{\text{tx}} \leq \kappa_{\max}$. To simplify the problem, we restrict the possible transmit power to a set of Z discrete levels, $\boldsymbol{\kappa} = \{\kappa_1, \cdots, \kappa_Z\}$. 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].

A. Channel Model and Rate Expression

The 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 $L_{\text{AB}} = c^2 d_{\text{AB}}^{-\alpha}/16\pi^2 f^2$, where $c \approx 3 \times 10^8$ ms^{-1} is the speed of light, f is the carrier frequency and α denotes the path-loss exponent. A correlated log-normal shadowing model based on a 2D Gaussian random field is considered [12]. We compute the shadowing on the link from A to B using

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

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. [13].

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:

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

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

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

where $\mathcal{I}_k(t)$ and $\boldsymbol{\kappa}^k(t)$ 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

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

B. Problem Formulation

In this letter, we consider two similar, but distinct resource selection problems: I) distributed joint channel and power selection, and II) distributed channel selection with aggregation. These problems 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 $\zeta_n = \min(\{\zeta_{nm}\}_{m=1}^{M_n}); \forall n \in \mathcal{N}$), problem I can, formally, be defined as:

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

where $\mathbf{c} \coloneqq \{c_n | n = 1, \cdots, N\}$ and $\boldsymbol{\kappa} \coloneqq \{\kappa_n | n = 1, \cdots, N\}$ denotes the set of channel indices and transmit powers for all inXSs, respectively. The term $\text{BW}(c_k)$ denotes bandwidth of the bandwidth of channel, c_k. The second problem is defined analogous to (5) with ζ_n taken as the minimum total capacity over all aggregated channels and the power levels set to an equal value for all 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 Multi-agent Q-learning (MAQL) method with quantized SI for solving these problems in section III.

III. Resource Selection with 1-bit Information

We cast the joint optimization problem in (5) as Multi-Agent Markov Decision Process (MMDP) [14] described as the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$, where $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$ is a set of all possible states for all inXSs referred to as state space, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ is the joint action space containing all possible actions (i.e., the set of all possible combinations of channels and power levels for problem I and all possible combinations of aggregated channels for problem II), $\mathcal{R}$ denotes the reward signal and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \Delta$ is the transition function [14], where Δ denotes the set of probability distributions over $\mathcal{S}$.

In the considered MMDP, the goal of the nth agent is to find an *optimal* policy, π_n^* based solely on its local state and action information resulting in the so called Partially Observable MMDP (POMMDP) [15]. Typically, π_n^* is obtained as the policy which maximizes the total reward function expressed as [16]

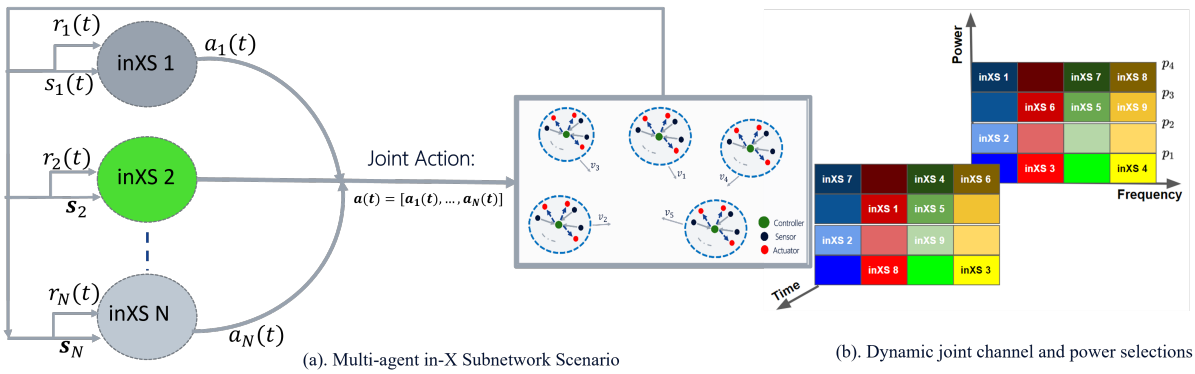
$$\pi_t^*(s) = \max_{\pi_t(s) \in \mathcal{A}} \left\{ r_t(s_t, \pi_t(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s_t, s') \pi_{t+1}^*(s') \right\}, \quad (6)$$

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 nth agent is given as [14]

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

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.,

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



(a). Multi-agent in-X Subnetwork Scenario

(b). Dynamic joint channel and power selections

Fig. 1: Illustration of the Multiagent RL scenario with N inXSs.

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 interaction of all agents with the environment. The Q-function is then iteratively estimated according to Bellman's equation as [17]

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

where α 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 $\sum_{a\in\mathcal{A}} \pi(s, a) = 1$.

A. 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 II, we consider a wireless environment with N independent inXSs each with one or more devices as illustrated in Fig. 1. The remaining components are described below.

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. We denote the SI at time t as $\mathbf{I}_n^t = [I_{n,1}^t, I_{n,2}^t, \cdots, I_{n,K}^t]^T \in \mathbb{R}^{(K\times 1)}$. To account for the effect of channel condition within each inXS, we propose state representation based on the SINR over all channels denoted for the nth inXS as $\mathbf{s}_n^t = [s_{n,1}^t, s_{n,2}^t, \cdots, s_{n,K}^t]^T$, with $s_{n,i} = s_{\mathrm{d}}/(I_{n,i} + \sigma^2)$, 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 $s_{n,i} < s_{\mathrm{th}}$ and in state 1 otherwise.

2) Action space: For the joint channel and power selection task, 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. $a_n^t \in \mathcal{A}_I$; $\mathcal{A}_I = \{\{c_1, p_1\}, \{c_1, p_2\}, \cdots, \{c_K, p_Z\}\}$. In the case of channel selection with aggregation, we consider transmission with fixed power over a maximum of 2 channels by each inXS. The action of the nth inXS at time t is then $a_n^t \in \mathcal{A}_{II}$; $\mathcal{A}_{II} = \{\{c_1\}, \cdots, \{c_K\}, \{c_1, c_2\}, \cdots, \{c_{K-1}, c_K\}\}$ with dimension $|\mathcal{A}_{II}| = K + \binom{K}{2}$.

3) Reward signal: We assume that the communication metric to be maximized is the capacity and use (4) as the reward function. In the case of channel aggregation, the reward is taken as the summation of capacity over all aggregated channels.

4) Policy Representation: The policy at each inXS is represented by a $2^K \times |\mathcal{A}_{I/II}|$ *lookup table* containing the Q-values for all state-action pairs.

5) Action Selection: Resource selection decision is made by each agent via the ϵ-greedy strategy defined as

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

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

$$\epsilon = \max\left(\epsilon_{\min}, (\epsilon_{\max} - \epsilon_{\min})/\epsilon_{\mathrm{step}}\right), \tag{11}$$

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

6) Training Procedure: A fully distributed training in which all inXSs simultaneously learn to optimized individual Q-tables is adopted in this work. The procedure is described in Algorithm 1.

IV. Numerical Analysis

We now train and evaluate the performance of the MAQL approach and compare with fixed (i.e., random assignment at

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

Input: Simulation and environment parameters, learning rate, α, discount factor, γ, 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-tables, $\{Q_n\}_{n=1}^{N}$ 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 as in (11).
    for $n = 1$ **to** N **do**
      Update Q_n using (9)
    end for
  end for
end for
Output: Trained Q-tables, $\{Q_n\}_{n=1}^{N}$

TABLE I: Simulation parameters.

Deployment and system parameters	
Parameter	Value
Deployment area [m^2]	60 × 60
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, γ	2.2
Shadowing standard deviation, σ_s [dB]	5.0
De-correlation distance, d_c [m]	2
Lowest frequency [GHz]	3
Transmit power levels [dBm]	[-10 0 10]
Noise figure [dB]	10
Target rate [bps/Hz]	0.32 – 1.60
Per channel bandwidth [MHz]	5
Q-Table and simulation settings	
Action/state space size, $\lvert\mathcal{A}\rvert/\lvert\mathcal{S}\rvert$	15/32
Discount factor, γ	0.90
Learning rate, α	0.80
Number of training episodes/Steps per episode	3000/200
Minimum/maximum exploration probability	0.01/0.99
Number of epsilon greedy steps	4.8×10^5

initialization without dynamic updates), greedy channel selection and centralized graph coloring (CGC) using a snapshot based procedure. We consider a network with $N = 20$ inXSs each with a single controller serving as the AP for a sensor-actuator pair in a 60 m × 60 m rectangular deployment area. movements in the area follows the 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. We set the transmit power for all inXSs to 10 dBm for P-II and consider a total of $Z = 3$ transmit power levels, [-10, 0, 10] dBm for P-I, leading to a 15×1 action space in both cases. Other simulation parameters are shown in Tab. I.

Fig. 2a shows the averaged reward over successive training episodes for the joint power and channel selection problem. 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 2400 episodes. Similar convergence rate was noticed for the channel aggregation problem. At convergence, the MAQL method has similar performance to greedy selection with full SI [8]. To understand the actions of the Q-agents, we show the learned Q-policy at convergence in Fig. 2b. 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 $\geq s_{\text{sh}}$) for all states except for state 1 which has no channel in level 1. As shown in the figure, the power levels of 10 dBm and 0 dBm are preferred by the agents in the ratio 27:5. The lowest power

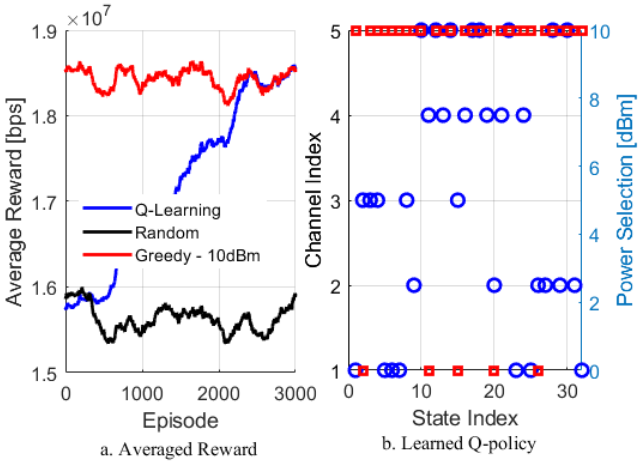


Fig. 2: Averaged reward per training episode (a) and learned policy (b).

level of -10 dBm is never chosen with full exploitation.

The trained Q-tables are deployed for distributed resource selection and performance compared with random and greedy channel selection. We use the suffix '- I' (for joint channel and power selection) and '- II' (for channel selection with aggregation) with all algorithms to indicate the corresponding problem. Thus, we have MAQL - I and MAQL - II as the proposed solution for problems I and II, respectively, and the corresponding baselines are denoted: 1). Random - I: assign channels randomly at the start of a snapshot; 2). Greedy - I: select the channel with minimum interference power; 3). Random - II: randomly assign a single or two channels for aggregation from the possible options; and 4.) Greedy - II: select the 2 least interfered channels for aggregation.

Inspired by our initial results from the MAQL methods, we further proposed a heuristic algorithm (denoted 'Q-Heuristic')

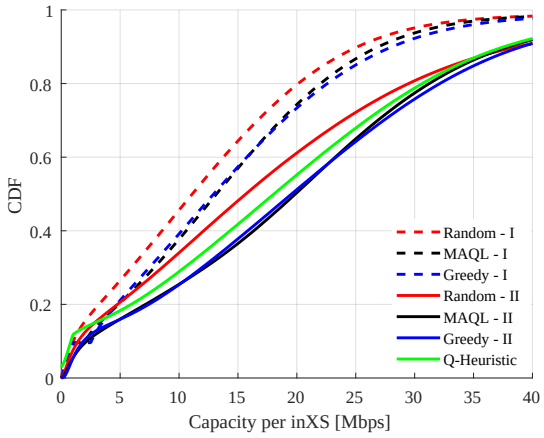


Fig. 3: Sum capacity per inXS with joint channel and power selection (I), and channel aggregation (II).

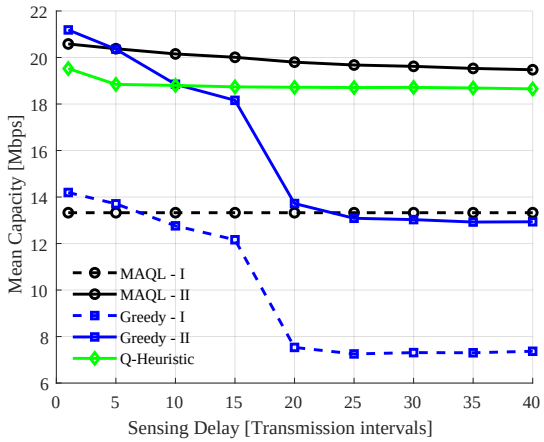


Fig. 4: Mean capacity versus sensing delay.

for resource selection based on similar quantized SI as the MAQL method. In Q-heuristic, each inX chooses 2 channels for aggregation as follows: select both channels randomly from $\mathcal{K}$ if all are in level 0; select the channel in level 1 and 1 randomly if the state has a single 1 or both randomly from the list of channel in level 1, otherwise. Except for MAQL - I, all algorithms use a transmit power of 10 dBm per transmission. Fig. 3 shows the empirical Cumulative Distribution Function (CDF) of the achieved capacity per inXS with sensing-to-action time (i.e., sensing delay) of a single time slot. The proposed MAQL method performs better than simple random selection but similar to greedy selection with full SI. Compared to Q-heuristic, the proposed MAQL method offer slightly better performance.

In Fig. 4, we study the sensitivity of the proposed methods and greedy baseline to sensing delay, i.e., the time (in number of slots) between sensing and action. While the proposed methods with 1-bit information appear robust to delays, the greedy scheme is rather quite sensitive. This indicates that the proposed methods offer similar performance as the baseline but provide significant overhead reduction for SI exchange and robustness to sensing delays which may be inevitable in practice. A similar study on sensitivity to quantization levels showed very marginal difference in the achieved capacity by the MAQL method with s_{th} in the range 2 dB – 12 dB.

V. Conclusion

Multi-agent Q-learning agents for distributed dynamic resource selection with 1-bit quantized SI can achieve similar performance to the best known heuristics (i.e., greedy selection) with full information in 6G in-X subnetworks. Simulation results have shown that the proposed MAQL methods exhibit fast convergence and are more robust to sensing delays than greedy resource selection. The proposed method is also robust to variations in SINR quantization thresholds.

Acknowledgment

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

References

[1] G. Berardinelli, P. Baracca, R. Adeogun, S. Khosravirad, F. Schaich, K. Upadhya, D. Li, T. B. Tao, H. Viswanathan, and P. E. Mogensen, "Extreme Communication in 6G: Vision and Challenges for 'in-X' Subnetworks," *IEEE OJCOM*, 2021.

[2] R. Adeogun, G. Berardinelli, P. E. Mogensen, I. Rodriguez, and M. Razzaghpour, "Towards 6G in-X Subnetworks With Sub-Millisecond Communication Cycles and Extreme Reliability," *IEEE Access*, vol. 8, pp. 110 172–110 188, 2020.

[3] G. Berardinelli, P. Mogensen, and R. O. Adeogun, "6G subnetworks for Life-Critical Communication," in *2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[4] V. Ziegler, H. Viswanathan, H. Flinck, M. Hoffmann, V. Räisänen, and K. Hätönen, "6G architecture to connect the worlds," *IEEE Access*, vol. 8, pp. 173 508–173 520, 2020.

[5] H. Viswanathan and P. E. Mogensen, "Communications in the 6G Era," *IEEE Access*, vol. 8, pp. 57 063–57 074, 2020.

[6] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2020.

[7] R. O. Adeogun, "Joint resource allocation for dual — band heterogeneous wireless network," in *IEEE WCNC*, 2018, pp. 1–5.

[8] R. Adeogun, G. Berardinelli, I. Rodriguez, and P. E. Mogensen, "Distributed Dynamic Channel Allocation in 6G in-X Subnetworks for Industrial Automation," in *IEEE Globecom Workshops*, 2020.

[9] R. O. Adeogun, G. Berardinelli, and P. E. Mogensen, "Learning to Dynamically Allocate Radio Resources in Mobile 6G in-X Subnetworks," in *IEEE PIMRC*, 2021.

[10] J. Burgueno, R. Adeogun, R. L. Bruun, C. S. M. García, I. de-la Bandera, and R. Barco, "Distributed Deep Reinforcement Learning Resource Allocation Scheme For Industry 4.0 Device-To-Device Scenarios," in *IEEE VTC-Fall)*. IEEE, 2021, pp. 1–7.

[11] J. Cui, Y. Liu, and A. Nallanathan, "Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, 2020.

[12] S. Lu, J. May, and R. J. Haines, "Effects of correlated shadowing modeling on performance evaluation of wireless sensor networks," in *IEEE Vehicular Technology Conference*, 2015, pp. 1–5.

[13] W. C. Jakes and D. C. Cox, *Microwave mobile communications*. Wiley-IEEE press, 1994.

[14] S. Mukhopadhyay and B. Jain, "Multi-agent Markov decision processes with limited agent communication," in *IEEE ISIC*, 2001, pp. 7–12.

[15] Y. Yang and J. Wang, "An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective," *CoRR*, vol. abs/2011.00583, 2020.

[16] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint Power Allocation and Channel Assignment for NOMA With Deep Reinforcement Learning," *IEEE J. Sel. Areas Commun*, vol. 37, no. 10, pp. 2200–2210, 2019.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.