Deep Reinforcement Learning Empowered OFDM Subcarrier Allocation in Integrated Sensing and Communication Networks

1st Huanyu Dong
Computer and Information Science
University of Macau
Macau, China
huanyu.dong0425@gmail.com

***Abstract*—With the development of B5G/6G networks, integrated sensing and communication (ISAC) has been considered as a key enabler for future applications such as smart vehicle networks, smart cities and smart homes. However, the high demand for faster speeds, lower latency, and greater connection capacity in the next-generation wireless networks makes the coexistence of communication and sensing should be carefully designed. Resource scheduling has been envisioned as an important way for communication and sensing to coexist in ISAC networks. Many previous works focus on subcarrier allocation problems in OFDM-based ISAC. However, did not consider when OFDM subcarrier is insufficient, which means sensing and communication have to be reused on some subcarriers. We mainly focus on the scenario that subcarrier sharing occurs between sensing and communication, and then propose a reinforcement learning empowered subcarrier allocation algorithm to maximize the system energy efficiency.**

***Index Terms*—Integrated Sensing and Communiation (ISAC), OFDM, subcarrier allocation, deep reinforcement learning**

I. Introduction

With the development of B5G/6G networks, the sensing ability of wireless networks has become a key feature that enables future applications, such as smart vehicle networks, smart cities, and smart homes [1]. The integrated sensing and communication (ISAC) networks focus on the joint design of a communication system embedded with sensing ability [2]. However, the high demand for faster speeds, lower latency, and greater connection capacity in the next-generation wireless networks makes the coexistence of communication and sensing must be carefully designed. Resource scheduling has been envisioned as an important way for communication and sensing to coexist in ISAC networks [3]. Proper design of resource allocation architecture utilizes the overall performance of communication and sensing, improves the system efficiency, and cuts down hardware costs.

Many previous works have shed light on the paradigm of resource management in ISAC networks. Depending on whether the unit resources are orthogonal to each other or not, the resource allocation in ISAC networks can be divided into two categories: orthogonal resource allocation and non-orthogonal resource allocation [2]. Typically, orthogonal resource allocation can be implemented in three ways: time-division, frequency-division, and spatial division. Compared to non-orthogonal resource allocation, orthogonal resource allocation is more easily implemented [2].

Frequency-division ISAC schedules orthogonal frequency domain resources to sensing and communication. OFDM has been widely used in 4G and 5G networks [4]. OFDM technology carrierizes channel resources and allows further allocate the subcarriers to certain services. Thus, OFDM technology is considered to be a basis of frequency-division ISAC. In [5], authors proposed an OFDM subcarrier selection and power minimization problem between sensing and communication. In [6], a mutual information maximization-oriented power allocation scheme of OFDM subcarriers. In [7], authors proposed a joint subcarrier and power allocation method to minimize the peak-to-sidelobe ratio while guaranteeing communication data rate and range resolution. However, previous works did not consider when OFDM subcarrier is insufficient, which means sensing and communication have to be reused on some subcarrier. In [8], authors derive the performance boundaries of spectrum sharing between sensing and communication. In [9], a subchannel allocation problem is proposed in integrated sensing, communication and computation networks.

In this paper, we propose an OFDM multi-carrier ISAC system and mainly focus on the scenario in which subcarriers are insufficient. We propose a reinforcement learning-empowered subcarrier allocation algorithm to maximize the system's energy efficiency. The main contribution of this paper is as follows:

- This paper focuses on the scenario that subcarriers are insufficient in an OFDM ISAC system and discovers the subcarrier reuse between sensing and communication.
- This paper proposes a subcarrier allocation algorithm based on proximal policy optimization to maximize energy efficiency in the system.
- We consider an OFDM ISAC system with multiple communication users and multiple sensing targets, which has received little attention in previous research.

II. System Model And Problem Formulation

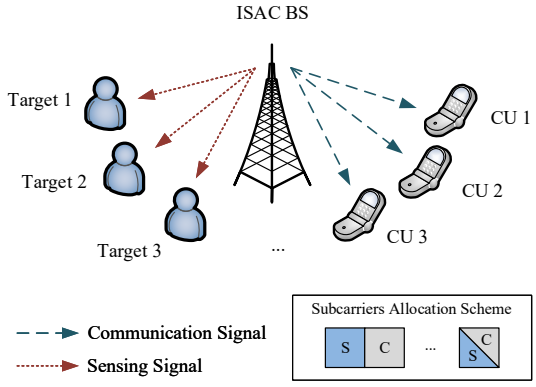


Fig. 1. Illustration of the proposed ISAC system.

As shown in Fig. 1, we consider a system of a dual-functional base station (ISAC BS), several downlink communication users (CU), and sensing targets. $\mathcal{M}_c$ and $\mathcal{M}_r$ denote the set of CUs and sensing targets. The radar sensing service and communication service share a spectrum of N subcarriers. The subcarriers are allocated to CUs for downlink communication and radar sensing service.

For simplicity, we suppose the number of sensing targets $|\mathcal{M}_r|$ and the number of CUs $|\mathcal{M}_c|$ are less than or equal to N. Each subcarrier is exclusive to one CU or one sensing target or shared by a pair of CU and sensing targets.

A. Downlink Communication Model

The received signal of CU $k \in \mathcal{M}_c$ on subcarrier n is formulated as:

$$y_{kn} = h_{kn} \sum_{q \in \mathcal{M}_c} u_{qn} \sqrt{p_n^c} s_{qn}^c + h_{kn} \sum_{l \in \mathcal{M}_r} d_{ln} \sqrt{p_n^r} s_{ln}^r + n_{kn}, \tag{1}$$

where h_{kn} denotes the downlink channel gain of CU k on subcarrier n, and p_n^c denotes the communication signal power on subcarrier n. $\{u_{kn} | \forall k \in \mathcal{M}_c, \forall n \in \{1, 2, ..., N\}\}$, and $\{d_{ln} | \forall l \in \mathcal{M}_r, \forall n \in \{1, 2, ..., N\}\}$ are binary variables denote subcarrier allocation scheme of CUs and sensing targets respectively.

If CU k is scheduled on subcarrier n, $u_{qn}, \forall q \in \mathcal{M}_c, q \neq k$ equals to zero since other CUs are not permitted to be scheduled on the same subcarrier. Thus, the signal-to-interference-and-noise ratio (SINR) of CU k on subcarrier n is simplified as:

$$\gamma_{kn}^c = \frac{u_{kn} |h_{kn}|^2 p_n^c}{\sum_{l \in \mathcal{M}_r} d_{ln} |h_{kn}|^2 p_n^r + \sigma^2}. \tag{2}$$

Then, the communication data rate of CU k is:

$$R_k^c = \sum_{n=1}^{N} \Delta f \log_2(1 + \gamma_{kn}^c). \tag{3}$$

Δf denotes the subcarrier spacing.

B. Radar Sensing Model

On subcarrier n, the transmitted radar signal of S consecutive OFDM symbols of sensing target l is:

$$s_{ln}^r(t) = d_{ln} e^{j2\pi f_n t} \sum_{m=1}^{S} \sqrt{p_n^r} c_{ln}^m e^{j2\pi k \Delta f (t - mT_s)} \cdot \mathrm{rect}\left[(t - mT_s)/T_s\right], \tag{4}$$

where p_n^r is the power of sensing signal on subcarrier n, and $c_{k,n}$ is the phase code of symbol m, T_s is the duration per OFDM symbol, f_k is the center frequency.

Similarly, the transmitted communication signal of S consecutive OFDM symbols of CU k is:

$$s_{kn}^c(t) = u_{kn} e^{j2\pi f_n t} \sum_{m=1}^{S} \sqrt{p_n^c} c_{kn}^m e^{j2\pi k \Delta f (t - mT_s)} \cdot \mathrm{rect}\left[(t - mT_s)/T_s\right]. \tag{5}$$

The received echo signal from the l-th sensing target on subcarrier n is:

$$z_{ln}(t) = \int_{-\infty}^{+\infty} \tilde{h}_{ln}(\tau) s_{ln}^r(t - \tau) d\tau + \int_{-\infty}^{+\infty} \sum_{k \in \mathcal{M}_c} \tilde{h}_{ln}(\tau) s_{kn}^c(t - \tau) d\tau + n(t), \tag{6}$$

where $n(t)$ is complex additive white Gaussian noise (AWGN) with zero mean and power spectral density $N(f)$. $\tilde{h}_{ln}$ denotes the channel gain from the ISAC BS to the l-th sensing target then to the l-th sensing target on subcarrier n.

According to [10]–[12], the widely used conditional mutual information metric is adopted in this paper. For a sensing user $l \in \mathcal{M}_r$, the conditional mutual information (MI) between the received signal and the target impulse response on subcarrier n is:

$$\begin{aligned} MI_{ln}^{\mathrm{rad}} &= I(z_n(t); \tilde{h}_{ln}(t) | s_{ln}^r(t)) \\ &= \frac{1}{2} S T_s \Delta f \log_2(1 + \gamma_{ln}^r). \end{aligned} \tag{7}$$

Thus, the total mutual information for target l is formulated as:

$$MI_l^{\mathrm{rad}} = \sum_{n=1}^{N} MI_{ln}^{\mathrm{rad}},$$

where γ_{ln}^r is:

$$\gamma_{ln}^r = \frac{d_{ln} p_n^r S T_s^2 |\tilde{H}_{ln}(f_n)|^2}{|\tilde{H}_{ln}(f_n)|^2 \sum_{q \in \mathcal{M}_c} u_{qn} p_n^c + N(f) S T_s}. \tag{8}$$

C. Problem Formulation

In an ISAC system, subcarrier sharing between downlink communication and target sensing brings about interference to both services. However, when the spectrum resource for the ISAC system is insufficient, subcarrier sharing occurs to maintain the coverage of service. Therefore, we aim to design a subcarrier scheduling algorithm to maximize the system's energy efficiency while ensuring the coverage of service.

In this section, we first define the energy efficiency of the ISAC system. Our goal is to maximize the total energy efficiency by optimizing $\{u_{kn}\}$ and $\{d_{ln}\}$.

We propose a time-domain scheduling problem. Suppose the ISAC device conduct subcarrier allocation every T_0 and the total interested time window is T^{tot}, and $T^{\text{tot}}/T_0 = C$.

For simplicity, we restrict each subcarrier is exclusive to one CU or one sensing target or shared by a pair of one CU and one sensing target:

$$0 \leq \sum_{k \in \mathcal{M}_c} \mathbf{u}(t_i)[k,n] \leq 1, \tag{9}$$

$$0 \leq \sum_{l \in \mathcal{M}_r} \mathbf{d}(t_i)[l,n] \leq 1, \ \forall 1 \leq n \leq N, \tag{10}$$

where $\mathbf{u}(t_i)$, $\mathbf{d}(t_i)$ are matrix form $\{u_{kn}\}$ and $\{d_{ln}\}$ at t_i.

We define a vector $I(t_i) \in \mathbb{N}^{|\mathcal{M}_c \cup \mathcal{M}_r|}$ to represent the total number of subcarriers allocated to each user from t_0 to t_i. The ISAC device ensures the service coverage for each user by restricting each element of $I(t_i)$ is no less than a predefined value:

$$I(t_i)[q] \geq \frac{i}{|\mathcal{M}_c \cup \mathcal{M}_r|}, \ \forall q \in \mathcal{M}_c \cup \mathcal{M}_r \tag{11}$$

The weighted energy efficiency during T^{tot} can be formulated as follows:

$$\begin{aligned} \Theta = \omega_1 \frac{\sum_{i=1}^{C} \sum_{k \in \mathcal{M}_c} R_k^c(\mathbf{u}(t_i), \mathbf{d}(t_i))}{\sum_{i=1}^{C} \sum_{n=1}^{N} \sum_{k \in \mathcal{M}_c} u_{kn}(t_i) p_n^c} \\ + \omega_2 \frac{\sum_{i=1}^{C} \sum_{l \in \mathcal{M}_r} MI^{\text{rad}}(\mathbf{u}(t_i), \mathbf{d}(t_i))}{\sum_{i=1}^{C} \sum_{n=1}^{N} \sum_{l \in \mathcal{M}_r} d_{ln}(t_i) p_n^r}, \end{aligned} \tag{12}$$

Then we formulate the average energy efficiency maximization problem.

$$\begin{aligned} \text{(EEM):} \quad & \max \ \Theta \\ \text{subject to:} \quad & 0 \leq \sum_{k \in \mathcal{M}_c} \mathbf{u}(t_i)[k,n] \leq 1, \ \forall 1 \leq n \leq N && (13) \\ & 0 \leq \sum_{l \in \mathcal{M}_r} \mathbf{d}(t_i)[l,n] \leq 1, \ \forall 1 \leq n \leq N && (14) \\ & \mathbf{u}(t_i) \in \{0,1\}^{|\mathcal{M}_c| \times N} && (15) \\ & \mathbf{d}(t_i) \in \{0,1\}^{|\mathcal{M}_r| \times N} && (16) \\ & I(t_i)[q] \geq \frac{i}{|\mathcal{M}_c \cup \mathcal{M}_r|}, \ \forall q \in \mathcal{M}_c \cup \mathcal{M}_r && (17) \\ \text{variables:} \quad & \mathbf{u}(t_i), \mathbf{d}(t_i), \ \forall 1 \leq i \leq C \end{aligned}$$

III. Algorithm Design of PPO Based Subcarrier Allocation

In this section, we first reduce the dimension of the Problem (EEM). Then we propose a subcarrier allocation algorithm based on proximal policy optimization (PPO-SA).

A. Reducing the Dimension of Problem (EEM)

Because each CU or sensing target has an allocation indicator on each subcarrier, the solution space of the Problem (EEM) is very huge. Before we formulate the problem into MDP form, we conduct dimension reduction first.

We suppose the CUs estimate channel gain through reference signals and report them to the ISAC device periodically. Thus, we suppose the ISAC device has knowledge of the downlink channel gain. Therefore, at each scheduling cycle, we first allocate subcarriers to CUs by matching the CU with the maximum channel gain on each subcarrier. Due to the lack of subcarriers, CU that is already scheduled on a subcarrier will not be chosen by another subcarrier. $\mathcal{S}$ denotes the CUs that are already scheduled.

$$k = \underset{q \in \mathcal{M}_c \setminus \mathcal{S}}{argmax} \ |h_{qn}|, \ \mathbf{u}(t_i)[k,n] = 1, \ \forall 1 \leq n \leq N. \tag{18}$$

By allocating CUs in advance, we only need to conduct subcarrier allocation for sensing service, which greatly reduces the dimension of the solution. However, the ISAC device does not have the knowledge of radar echo channels. Further design is needed for getting $\mathbf{d}(t_i)$.

Another trick we use to reduce the solution dimension is to pre-select $|\mathcal{M}_r|$ subcarrier out of the total N subcarrier. We estimate the potential interference power from each subcarrier and select $|\mathcal{M}_r|$ subcarriers with the least interference power. The potential interference power on subcarrier n is formulated as:

$$\text{IP}(t_i) = \left\{ \sum_{q \in \mathcal{M}_c} u_{qn} p_n^c, \ \forall 1 \leq n \leq N \right\}. \tag{19}$$

B. MDP Formulation

At time t_i, the RL agent take action $a(t_i)$ according to policy π_θ. After applying a_{t_i}, the agent gets a reward r_t and observes the next environment state $s(t_{i+1})$. The RL agent aims at maximizing the cumulative reward. We formulate the problem into an MDP problem.

- State: At scheduling cycle t_i, we first allocate subcarriers to CUs. Then we calculate the potential interference power on each subcarrier $\text{IP}(t_i)$. To reduce the dimension of action space, we pre-select $|\mathcal{M}_r|$ subcarrier as feasible action, denoted as feasible action set $\mathcal{N}^{\text{pre}}(t_i)$. The state is formulated as $s(t_i) = [I(t_i), \text{IP}(t_i)]$, which consists of the scheduling number record, the pre-selected subcarrier and the potential inference power on each subcarrier.
- Action: The action $a(t_i)$ is designed as a $|M_r|$-dimension vector representing the index of assigned sensing targets for each subcarrier in the pre-selected set. Each element of $a(t_i)$ has an action space of $\{0, 1, ..., |\mathcal{M}_r|\}$, where 0 denote the subcarrier is not allocated to any user. When applied to the environment, $a(t_i)$ will be mapped into binary form.
- Reward: The reward function is a crucial part of RL. The value of objective function Θ is unavailable unless an episode is ended, i.e. $t_i = t_C$. Thus, the reward function

is sparse. To tackle this problem, we introduce a reward-shaping technique. We formulated a running reward and several penalty items to promote the agent to meet the constraints. The detailed reward function design is as follows.

The reward function is formulated as $r(t_i) = r_{ee} + r_{sched} + r_{pen}$. r_{sched} and r_{pen} zero when constraint (17) and (13)(14) are satisfied. Otherwise, r_{sched} and r_{pen} are negative. r_{ee} is the weighted sum of energy efficiency calculated from t_1 to t_i, which is formulated as:

$$r_{ee}(t_i) = \omega_1 \frac{\sum_{m=1}^{i} \sum_{k\in\mathcal{M}_c} R_k^c(\mathbf{u}(t_m), \mathbf{d}(t_m))}{\sum_{m=1}^{C} \sum_{n=1}^{N} \sum_{k\in\mathcal{M}_c} u_{kn}(t_m) p_n} + \omega_2 \frac{\sum_{m=1}^{i} \sum_{l\in\mathcal{M}_r} MI^{\mathrm{rad}}(\mathbf{u}(t_i), \mathbf{d}(t_m))}{\sum_{m=1}^{i} \sum_{n=1}^{N} \sum_{l\in\mathcal{M}_r} d_{ln}(t_m) p_n}. \quad (20)$$

C. PPO Based Subcarrier Allocation

Proximal policy optimization (PPO) is an actor-critic reinforcement learning algorithm. The objective function [13] of PPO policy update is

$$L(\theta) = \hat{\mathbb{E}}_t \Big[\min(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \Big], \quad (21)$$

where $r_t(\theta) = \frac{\pi_\theta(a(t)|s(t))}{\pi_{\theta_{\mathrm{old}}}(a(t)|s(t))}$, ϵ is the hyperparameter. The clip function limit the value of $r_t(\theta)$ into $[1-\epsilon, 1+\epsilon]$. $\hat{A}_t$ is the estimation of advantage function. A well-known advantage estimation approach [14] is

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V, \quad (22)$$

where

$$\delta_t^V = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t). \quad (23)$$

The detailed update procedure of PPO-SA is as follows.

Algorithm 1 Update procedure of PPO-based subcarrier allocation.

Initialization: Initialize the network parameters θ, ϕ of policy network π_θ and value network V_ϕ. Initialize clip threshold ϵ and experienced data buffer $\mathcal{D}$.

for $t = 1, 2, ...$ **do**

  for $i = 1, 2, 3, ..., C$ **do**

    Conduct policy π_θ in environment and store $\{s(t_i), a(t_i), r(t_i)\}$ into the buffer $\mathcal{D}_k$.

  end for

  Estimate advantage $\hat{A}_t$ by (22) and (23).

  Update policy network parameter by gradient ascent on loss function (21).

  Update value network parameter by gradient descent on loss function:

  $L(\phi) = \frac{1}{|\mathcal{D}_k|T} \sum_{r\in\mathcal{D}_k} \sum_{t=0}^{T} \big(V_\phi(s_t) - \hat{R}_t\big)^2$.

end for

IV. Numerical Results

In this section, we evaluate the performance of the proposed PPO-SA algorithm.

A. Simulation Settings

Our simulations are conducted on Intel(R) Xeon(R) Gold platform with Ubuntu 20.04. We implemented the simulation using Python 3.10.4 and PyTorch 1.12.1. Training of the PPO-SA algorithm is on a GeForce RTX 3090 GPU, and the CUDA version is 11.7. Following the 3GPP standard, the carrier frequency is 6 GHz, and the subcarrier spacing is 30 kHz. The number of OFDM symbols per slot $S = 14$. The scheduling cycle T_0 is set to be 0.5 ms. And the total number of scheduling cycles during the interested time window is 100. The communication power and sensing power on each subcarrier is 0.5 W and 1 W. The power of background noise is -100 dB. The total number of subcarriers $N = 8$. And the number of CUs $|\mathcal{M}_c|$ and the number of sensing targets $|\mathcal{M}_r|$ is set to be (4,6), (7,4) and (8,2). The CUs and sensing targets are randomly distributed around the ISAC device. The variance of the range is 10 m.

The PPO-SA algorithm has 2 neural networks. The actor-network has 3 fully-connected layers, with 64, 64 and N_A neurons correspondingly. $N_A = |\mathcal{M}_r| \times (|\mathcal{M}_r| + 1)$, which depends on the number of sensing targets. The critic net is also composed of 3 fully connected layers. They have 64, 64 and 64 neurons correspondingly.

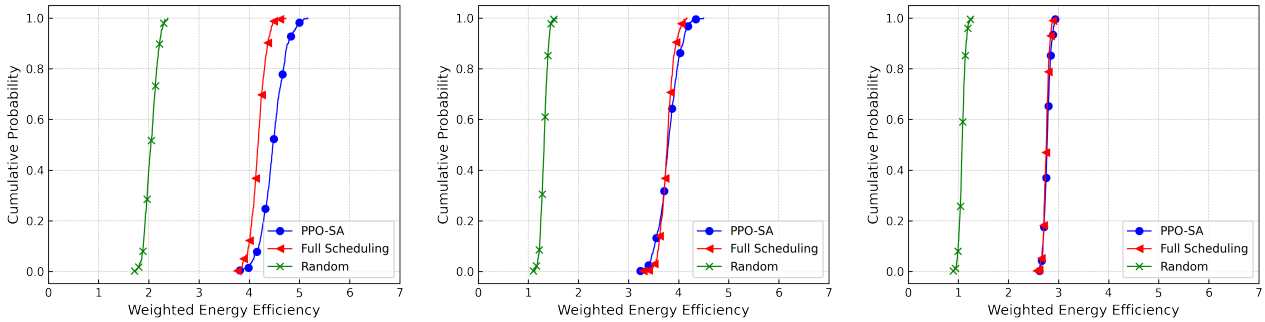
We compare the weighted energy efficiency with 2 baseline algorithms to evaluate the performance of the proposed PPO-SA.

- Random Strategy: This strategy randomly allocates each user with one subcarrier.
- Full Scheduling Strategy: This strategy needs the complete channel information. All CUs and sensing targets are scheduled during each scheduling cycle. Sensing services are scheduled to $|\mathcal{M}_r|$ subcarriers with minimum communication interference power. The rest vacant subcarriers are firstly allocated sensing targets with better radar channel quality. Sensing targets that are still not allocated will be paired with CUs that has lower communication channel quality.

B. Results

In Fig.2, we plot the cumulative probability distribution of weighted energy efficiency of Random Strategy, Full Scheduling Strategy, and PPO-SA algorithm. The data is collected via 40000 scheduling cycles in total. The proposed PPO-SA algorithm outperforms the Random Strategy and achieves performance comparable to that of the Full Scheduling Strategy, even without complete channel information, as shown in Fig.2.

As is shown in Fig. 3, we plot the average reward of PPO-SA and Full Scheduling Strategy during the neural network training. It is shown that the PPO-SA algorithm converges to maximum reward within 150 episodes. The reward of PPO-SA



(a) CDF curves of weighted energy efficiency under the setting of 8 subcarriers, 4 CUs and 6 sensing targets.

(b) CDF curves of weighted energy efficiency under the setting of 8 subcarriers, 7 CUs and 4 sensing targets.

(c) CDF curves of weighted energy efficiency under the setting of 8 subcarriers, 8 CUs and 2 sensing targets

Fig. 2. Performance comparison of three methods under different environment settings.

is lower than the Full Knowledge Strategy because the PPO-SA is a stochastic strategy, and violations of constraints may occur, resulting in negative penalties.

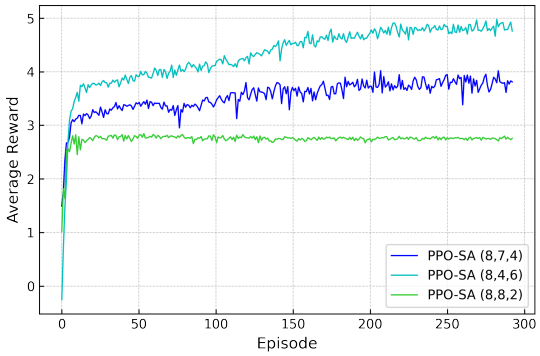


Fig. 3. Average training reward of PPO-based subcarrier allocation under three settings.

V. Conclusion

In this study, we formulated a time-domain subcarrier allocation problem for an OFDM ISAC network. The main purpose of this work is to explore the carrier allocation strategy in the case of insufficient carrier resources resulting in shared carriers for communication and sensing. To solve this problem, we designed a proximal policy optimization-based subcarrier allocation algorithm. It should be noted that in this paper we constructed an ISAC BS with a single antenna and did not exploit the spatial degree of freedom to utilize the system performance. In the future, we will move forward to work on joint beamforming and subcarrier allocation in an OFDM ISAC network.

References

[1] D.K.P. Tan, J. He, Y. Li, A. Bayesteh, Y. Chen, P. Zhu and W. Tong, "Integrated Sensing and Communication in 6G: Motivations, Use Cases, Requirements, Challenges and Future Directions", IEEE International Online Symposium on Joint Communications & Sensing (JC&S), Dresden, Germany, 2021, pp. 1-6, doi: 10.1109/JCS52304.2021.9376324.

[2] F. Liu, Y. Cui, C. Masouros, J. Xu, T.X. Han, Y.C. Eldar and S. Buzzi, "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," in IEEE Journal on Selected Areas in Communications, vol. 40, no. 6, pp. 1728-1767, June 2022, doi: 10.1109/JSAC.2022.3156632.

[3] F. Dong, F. Liu, Y. Cui, W. Wang, K. Han and Z. Wang, "Sensing as a Service in 6G Perceptive Networks: A Unified Framework for ISAC Resource Allocation," in IEEE Transactions on Wireless Communications, vol. 22, no. 5, pp. 3522-3536, May 2023, doi: 10.1109/TWC.2022.3219463.

[4] L. Han and K. Wu, "Joint wireless communication and radar sensing systems – state of the art and future prospects". IET Microw. Antennas Propag., 2013, 7: 876-885. https://doi.org/10.1049/iet-map.2012.0450

[5] C. Shi, F. Wang, S. Salous and J. Zhou, "Joint Subcarrier Assignment and Power Allocation Strategy for Integrated Radar and Communications System Based on Power Minimization," in IEEE Sensors Journal, vol. 19, no. 23, pp. 11167-11179, 1 Dec.1, 2019, doi: 10.1109/JSEN.2019.2935760.

[6] M. Bică and V. Koivunen, "Multicarrier Radar-communications Waveform Design for RF Convergence and Coexistence," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 7780-7784, doi: 10.1109/ICASSP.2019.8683655.

[7] Y. Chen, G. Liao, Y. Liu, H. Li and X. Liu, "Joint Subcarrier and Power Allocation for Integrated OFDM Waveform in RadCom Systems," in IEEE Communications Letters, vol. 27, no. 1, pp. 253-257, Jan. 2023, doi: 10.1109/LCOMM.2022.3216865.

[8] A. R. Chiriyath, B. Paul, G. M. Jacyna and D. W. Bliss, "Inner Bounds on Performance of Radar and Communications Co-Existence," in IEEE Transactions on Signal Processing, vol. 64, no. 2, pp. 464-474, Jan.15, 2016, doi: 10.1109/TSP.2015.2483485.

[9] L. Zhao, D. Wu, L. Zhou and Y. Qian, "Radio Resource Allocation for Integrated Sensing, Communication, and Computation Networks," in IEEE Transactions on Wireless Communications, vol. 21, no. 10, pp. 8675-8687, Oct. 2022, doi: 10.1109/TWC.2022.3168348.

[10] Y. Liu, G. Liao, J. Xu, Z. Yang and Y. Zhang, "Adaptive OFDM Integrated Radar and Communications Waveform Design Based on Information Theory," in IEEE Communications Letters, vol. 21, no. 10, pp. 2174-2177, Oct. 2017, doi: 10.1109/LCOMM.2017.2723890.

[11] C. Shi, F. Wang, M. Sellathurai, J. Zhou and S. Salous, "Power Minimization-Based Robust OFDM Radar Waveform Design for Radar and Communication Systems in Coexistence," in IEEE Transactions on Signal Processing, vol. 66, no. 5, pp. 1316-1330, 1 March1, 2018, doi: 10.1109/TSP.2017.2770086.

[12] T. Tian, T. Zhang, L. Kong, G. Cui and Y. Wang, "Mutual Information based Partial Band Coexistence for Joint Radar and Communication System," 2019 IEEE Radar Conference (RadarConf), Boston, MA, USA, 2019, pp. 1-5, doi: 10.1109/RADAR.2019.8835671.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, "Proximal Policy Optimization Algorithms[J]," arXiv preprint arXiv:1707.06347, 2017.

[14] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, "High-

Dimensional Continuous Control Using Generalized Advantage Estimation," arXiv preprint arXiv:1506.02438, 2015.