

# An Offline Multi-Agent Reinforcement Learning Framework for Radio Resource Management

Eslam Eldeeb and Hirley Alves

**Abstract**—Offline multi-agent reinforcement learning (MARL) addresses key limitations of online MARL, such as safety concerns, expensive data collection, extended training intervals, and high signaling overhead caused by online interactions with the environment. In this work, we propose an offline MARL algorithm for radio resource management (RRM), focusing on optimizing scheduling policies for multiple access points (APs) to jointly maximize the sum and tail rates of user equipment (UEs). We evaluate three training paradigms: centralized, independent, and centralized training with decentralized execution (CTDE). Our simulation results demonstrate that the proposed offline MARL framework outperforms conventional baseline approaches, achieving over a 15% improvement in a weighted combination of sum and tail rates. Additionally, the CTDE framework strikes an effective balance, reducing the computational complexity of centralized methods while addressing the inefficiencies of independent training. These results underscore the potential of offline MARL to deliver scalable, robust, and efficient solutions for resource management in dynamic wireless networks.

**Index Terms**—Centralized training decentralized execution, conservative Q-learning, offline multi-agent reinforcement learning, radio resource management

## I. INTRODUCTION

The road toward future intelligent wireless communication systems, such as the one envisioned by 6G, is paved with a growing interest in applying machine learning / artificial intelligence (ML/AI) to wireless systems [1], [2]. Machine learning and artificial intelligence (ML/AI) techniques have been instrumental in advancing beyond 5G systems. They are poised to play a more critical role in developing 6G networks. Given the heightened scale, complexity, and distributed nature of 6G, these technologies are essential for effectively addressing such intricate challenges [3], [4]. These challenges include (but are not limited to) radio resource management (RRM), which is often too complex to be modeled using traditional statistical methods. The RRM problem is generally a non-convex optimization problem, and its complexity increases tremendously as the network grows.

In the literature, RRM has been addressed using information theory [5], geometric programming [6], and game theory [7]. However, these algorithms may fail due to the dynamic behavior of the wireless systems. In this regard, online reinforcement learning (RL) has shown a promising contribution towards solving RRM problems [8]. Online RL

Eslam Eldeeb and Hirley Alves are with the Centre for Wireless Communications (CWC), University of Oulu, Finland. (e-mail: eslam.eldeeb@oulu.fi; hirley.alves@oulu.fi).

This work was supported by 6G Flagship (Grant Number 369116) funded by the Research Council of Finland.

involves an agent that interacts with the environment, observes its condition (state), takes a decision (action) and receives a feedback signal (reward) indicating the quality of the decision. Online RL techniques are suitable for RRM as they can solve the complex RRM problem in a model-free manner, without deployment knowledge. In addition, it benefits from the control and feedback methods in the wireless network to iteratively optimize and update the designed algorithm.

Recent advances in Online RL have witnessed the rise of two well-established RL frameworks: deep RL and multi-agent reinforcement learning (MARL). Deep RL combines robust deep neural networks (DNNs) with RL [9]. This eases optimizing complex and large-scale environments. On the other hand, MARL enables joint decision-making optimization (policy optimization) of multiple agents [10]. MARL algorithms vary from cooperative MARL [11], where various agents cooperate towards one goal, and competitive MARL [12], where multiple agents compete against each other. We focus on cooperative MARL as we aim to optimize the scheduling policy of several entities to achieve a joint goal in the system. The cooperative MARL problem itself varies according to the agents' communication rate [13]. For instance, decentralized solutions assume no communication between the agents, whereas centralized techniques allow complete communication between the agents [14]. Recent techniques propose mixed centralized and decentralized techniques [15].

Online MARL faces significant challenges when deployed to real-time wireless scenarios. First, it relies on online interaction with the environment to explore and visit the environment states. This online interaction might not be feasible, safe, timely, or costly. Second, some MARL variants, such as centralized MARL methods, enable interaction between the agents, which adds an extra layer of complexity and overhead to the environment. These problems can be addressed by optimizing the policy offline via a static dataset pre-collected using a behavioral policy. Thus, undesirable online interactions are mitigated. This opens the door for offline MARL.

### A. Offline MARL

*Offline MARL* considers an offline static dataset to be used to optimize the policies of multiple agents [16]. Offline MARL assumes that the agents can not interact with the environment during the optimization (training) phase. After training, the agents deploy the policies they have learned online. The offline dataset is usually collected using a behavioral policy, which is a policy designed using known traditional methods or even randomly. Offline MARL overcomes the safety and

cost problems accompanied by online MARL by mitigating online interaction. Moreover, it limits signaling overhead and complex communication requirements between the agents by moving policy optimization offline. In addition, since the training is performed offline, it can be easily transferred to a powerful central unit, removing computational burdens from the limited resources of wireless entities.

Adapting traditional online MARL techniques to an offline setting introduces a distributional shift between the behavior and learned policies. This shift motivates overestimation of the unseen experiences in the dataset, uncertain policies, and training degradation [17]. Several methods have suggested constraints on the difference between the behavior and learning policies, called behavior-constrained methods [18]. Another family of methods penalizes the value of out-of-distribution (OOD) actions (unseen actions in the dataset), which are called conservative methods [19]. Conservative Q-learning (CQL) is a conservative offline RL technique that uses KL-divergence as a regularization parameter to penalize the weights of OOD actions. This work proposes an offline MARL algorithm based on CQL for the RRM problem.

### B. Related Work

Over the past few years, many works in the literature have contributed to solving the RRM problem, mostly in an online fashion, for single and multi-agent reinforcement learning. Among the first to work on this problem is [20], which proposes a (single-agent) reinforcement learning approach for self-organizing networks (SONs) in small cells. Then, [21] proposes a deep RL algorithm for the spectrum sharing and resource allocation problem in cognitive radio systems. They adopt a deep RL algorithm for efficient power control so that the secondary user can share a common spectrum with the primary user. The authors in [22] propose an efficient resource and power optimization using reinforcement learning to jointly minimize the age-of-information (AoI) and transmission power of IoT sensors in unmanned aerial vehicles (UAVs) networks. In contrast, the work in [23] combines generative adversarial network (GAN) with deep RL for resource management and network slicing. A recent work in [24] solves the RRM problem using graph neural networks (GNNs). The authors formulate the problem as an unsupervised primal-dual problem. They develop a GNN architecture that parameterizes the RRM policies as a graph topology derived from the instantaneous channel conditions.

Several works have formulated MARL algorithms for the RRM problem and wireless communication. For example, the authors in [25] formulate the MARL problem for resource management in UAV networks. They present a comprehensive comparison between different MARL schemes. The authors in [26] propose an online MARL algorithm for the RRM problem to maximize both sum and tail rates. In [27], a dynamic power allocation is performed using MARL, where local observations are shared between nearby transmitters and receivers. The authors in [28] propose a distributed MARL approach for multi-cell wireless-powered communication networks to charge limited power users for efficient data collection wirelessly. In [29], the authors addressed the dynamic

resource management in X-subnetworks, where X refers to any entity such as a robot, vehicle, or module, and subnetworks refer to cells that can be part of a larger infrastructure. They propose combining MARL algorithms and attention-based layers to solve the resource management problem.

Even though offline RL and offline MARL are promising techniques, they have only recently begun to capture significant attention from the wireless communications community, e.g., [30], [31]. In [30], the authors propose an offline and distributional MARL algorithm for resource management in UAV networks. The work in [31] proposes a single-agent offline RL algorithm for the RRM problem. It has proved that a mixture of datasets of multiple behavioral policies can lead to an optimal scheduling policy. However, they assume all access points can be modeled as one agent, thus neglecting the multi-agent scenario. The work in [32] proposes an offline and distributional RL algorithm for the RRM problem that combines deep RL with distributional RL offline to overcome the uncertainties of the wireless environment. Similarly, they only focus on the single-agent case.

Most of the literature above suffers from significant drawbacks. First, the majority of these works targeted optimizing a single objective. However, the RRM problem targets multiple objectives, e.g., maximizing both sum and tail rates or AoI and pilot length. Second, some of these works considered the single-agent scenario and combined all agents in a centralized fashion. This is a notable concern as the network usually consists of many transmitters and users. Therefore, handling the RRM problem in a centralized fashion explodes the dimension and complexity of the RL problem, making it vulnerable to degrading performance. Finally, most of the existing works in the literature considered online RL or online MARL, which is unsafe, impractical, and very complex [33] due to the need for a massive online interaction with the environment, especially in the multi-agent case. These challenges heavily affect the communication network due to the need for continuous communication between the agents, leading to significant signaling overhead.

### C. Main Contributions

This work presents an offline MARL algorithm for RRM. We assume a general model and pose the RRM problem as a partially observable Markov decision process (MDP). Offline MARL proposes multi-agent optimization using only an offline static dataset without any interaction with the environment. Hence, it fits the RRM problem where multiple agents cooperate to serve the users. To illustrate our results, but without loss of generality, we model our RRM problem as a joint optimization problem that includes sum and tail rates. We aim to reach a resource management policy that maximizes a linear combination of sum and tail rates. The main contributions of this paper are summarized as follows.

- We formulate the RRM problem using a partially observable Markov decision process (MDP). In addition, we present a preliminary result using online MARL.
- We propose two offline MARL algorithms: soft actor-critic (SAC), and conservative Q-learning (CQL). We

TABLE I: List of abbreviations.

Abbreviation	Description
AI	Artificial intelligent
AoI	Age-of-information
AWGN	additive white Gaussian noise
BCQ	Behavior constrained Q-learning
CDF	Cumulative distribution function
CQL	Conservative Q-learning
AP	Access point
C-MARL	Centralized multi-agent reinforcement learning
CTDE	Centralized training and decentralized execution
DNN	Deep neural network
DQN	Deep Q-network
DRL	Distributional reinforcement learning
GAN	Generative adversarial network
GNN	Graph neural network
I-MARL	Independent training MARL
ITLinQ	Information-theoretic link scheduling
MARL	Multi-agent reinforcement learning
MDP	Markov decision process
OOD	Out-of-distribution
PF	Proportional fairness
PO-MDP	Partially-observable Markov decision process
RRM	Radio resource management
RSRP	Reference signal received power
RW	Random-walk
SAC	Soft actor-critic
SON	Self-organizing network
TDM	Time-division multiplexing
UAV	Unmanned aerial vehicles
UE	User equipment

present three variants of these offline MARL schemes using centralized learning, independent learning, and centralized training decentralized execution, respectively.

- We compare the three offline MARL schemes to four benchmarks from the literature. The proposed MARL schemes outperform the baseline models regarding both sum and tail rates.
- We demonstrate that centralized training decentralized execution approaches overcome the complexity of centralized training MARL and the inefficiency of the independent training MARL. Our algorithm surpasses existing schemes by more than 50% gain regarding the linear combination of sum and tail rates.

The rest of the paper is organized as follows: Section II introduces the RRM system model. The MARL formulation is proposed in Section III. Section IV depicts the proposed offline MARL algorithm. Simulation analysis is presented in Section V, and the paper is concluded in Section VI. Table I presents the list of abbreviations, while Table II summarizes the list of symbols and notations.

## II. SYSTEM MODEL

Consider the downlink of a cellular system as illustrated in Fig. 1. Assume an  $L \times L$  square network, and consider  $J$  user equipment (UEs) transmit their data to  $I$  access points (APs) during  $T$  discrete time intervals, forming an episode. At the beginning of each episode, APs and UEs are randomly deployed following a uniform distribution on the coordinates. During each episode, the position of each AP is fixed, while UEs move randomly within the network's borders with a fixed velocity  $v_t \in [0, 1]$  m/s. To elaborate on a practical scenario, we set three thresholds:

TABLE II: List of symbols and notations.

Symbol	Description
$\alpha$	CQL hyperparameter
$\beta$	Discount factor
$\pi_i(a^i   o^i)$	Policy of agent $i$
$a_i(t)$	Action of agent $i$ at time step $t$
$A(t)$	Joint action space
$C_j(t)$	Instantaneous rate of UE $j$
$\bar{C}_j$	Average rate of UE $j$
$\bar{C}_{\text{sum}}$	Sum rate
$\bar{C}_{5\%}$	5-percentile rate
$\bar{C}_{\text{score}}$	Score function
$d_0$	Minimum distance between two APs
$d_1$	Minimum distance between an AP and UE
$\mathcal{D}$	Offline dataset
$h_{ij}$	channel between UE $j$ and AP $i$
$I$	Number of APs
$J$	Number of UEs
$L$	Length of the network
$N$	Number of top users
$o_i(t)$	Local observations of agent $i$ at time step $t$
$PL_{ij}$	Path loss
$Q(s, a)$	Q-function
$r(t)$	Immediate reward at time step $t$
$S(t)$	Overall state space
$w_j(t)$	Weighting factor of user $j$ at time $t$

- 1) There is a minimum distance  $d_0$  between any two APs

$$d_{i'i} > d_0, \quad \forall i', i \in 1, \dots, I, \text{ and } i' \neq i, \quad (1)$$

where we keep sampling new APs positions until meeting the threshold.

- 2) There is a minimum distance  $d_1$  between each AP and each UE

$$d_{ij} > d_1, \quad \forall i \in 1, \dots, I, \text{ and } \forall j \in 1, \dots, J, \quad (2)$$

where we keep sampling new UEs positions until meeting the threshold.

- 3) All UEs are prohibited from moving outside the network's borders, where we consider a bounce back when a UE aims to move outside the borders.

The received signal of UE  $j$  from AP  $i$  at time  $t$  is

$$y_j(t) = h_{ij}(t)x_i(t) + \sum_{i' \neq i} h_{i'j}(t)x_{i'}(t) + n_j(t), \quad (3)$$

where  $n_j(t)$  is the additive white Gaussian noise (AWGN) whose power is  $\sigma^2$  and the channel between AP  $i$  and UE  $j$  is denoted as  $h_{ij}$  and comprises:

- path-loss: we follow the 3GPP indoor model [34]

$$PL_{ij} = 15.3 + 37.6 \log(d_{ij}) + 10, \quad (4)$$

- shadowing: we consider the log-normal effect with a standard deviation  $\sigma_{sh}$  [34], and
- short-term fading: we consider frequency-flat Rayleigh fading on all links in the network.

Each UE is associated with one of the APs at the beginning of each episode according to the reference signal received power (RSRP) [35]. In other words, a UE  $j$  is associated with AP  $i$  that records the max RSRP among all APs

$$i = \arg \max_{i'} \text{RSRP}_{i'j}, \quad \forall i' \in \{1, \dots, I\}. \quad (5)$$

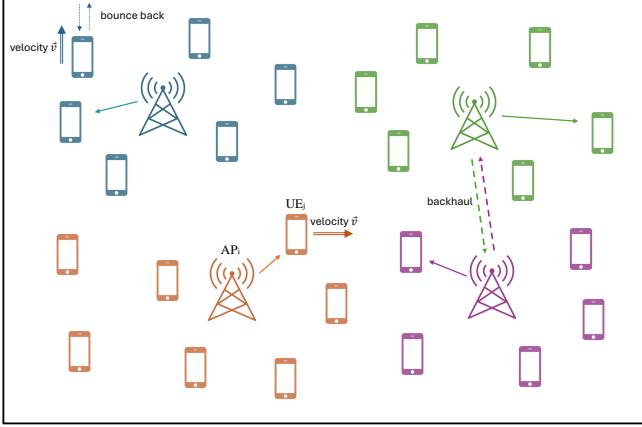


Fig. 1: A wireless environment consists of  $I$  APs and  $J$  UEs. Each UE is associated with only one AP, which chooses one of its associated UEs to serve at a time.

Then, each time  $t$ , each user selects one of its associated UEs to serve. To this end, the instantaneous rate and SINR of UE  $j$  that is associated with AP  $i$  are, respectively,

$$C_j(t) = \log_2 (1 + \gamma_j(t)), \quad (6)$$

$$\gamma_j(t) = \frac{|h_{ij}(t)|^2 p_t}{\sum_{i' \neq i} |h_{i'j}(t)|^2 p_t + \sigma^2}, \quad (7)$$

where  $p_t$  is the transmit power.

Then, in an episode, the average rate of UE  $j$ ,  $\bar{C}_j$ , is

$$\bar{C}_j = \frac{1}{T} \sum_{t=1}^T C_j(t). \quad (8)$$

To simplify the notation, time-dependent variables will be referred to without explicitly indicating the time dependency ( $t$ ), assuming it remains implicit unless specified otherwise.

#### A. Problem Formulation

The main objective in RRM problems is to find the optimal serving policy for each AP that maximizes the average rate across all users. However, simply formulating the problem to maximize (8) leads to a solution that invariably favors the user with the best SINR, thus disregarding fairness across users. To address this, the problem formulation should balance the sum and tail rate, ensuring a more equitable distribution of resources across all users.

Bearing this in mind, the sum rate is

$$C_{\text{sum}} = \sum_{j=1}^J \bar{C}_j, \quad (9)$$

whereas the tail rate, *i.e.*, the 5-percentile rate, is formulated as

$$C_{5\%} = \max C \text{ s.t. } \mathbb{P}[\bar{C}_j \geq C] \geq 0.95, \forall j \in \{1, \dots, J\}. \quad (10)$$

Next, we define the score function,  $C_{\text{score}}$ , as the linear combination of these rates

$$C_{\text{score}} = \mu_1 C_{\text{sum}} + \mu_2 C_{5\%}, \quad (11)$$

where  $\mu_1 \in \mathbb{R}_+$  and  $\mu_2 \in \mathbb{R}_+$  are user chosen parameters.

We are now ready to cast the RRM problem as

$$\mathbf{P1} : \max_A \sum_{t=1}^T C_{\text{score}}, \quad (12)$$

where  $A$  (defined in Section III) is the action space that describes the jointly serving policies of all APs.

Directly optimizing (12) imposes several challenges due to time dependency between actions that affect both sum and tail rates. In addition, the 5-percentile rate is challenging to optimize due to its instability as it can not be formulated in a closed form as a function of the system parameters. Alternatively, the authors in [26] proposed a more sophisticated approach to address this complex optimization. Consider the weighting factor  $w_j(t)$  of user  $j$  at time  $t$ , which can be recursively obtained as

$$w_j(t) = \frac{1}{\tilde{C}_j(t)}, \quad (13)$$

$$\tilde{C}_j(t) = \eta C_j(t) + (1 - \eta) \tilde{C}_j(t-1),$$

$$\tilde{C}_j(0) = C_j(0),$$

where  $\eta$  is a running average parameter and  $\tilde{C}_j$  is the long-term average rate of user  $j$  at time  $t$ . The proportional fairness (PF) ratio  $\text{PF}_j$  of user  $j$  at time  $t$  is defined as the product of (13), the weighting factor, and (6), the instantaneous rate,

$$\text{PF}_j = w_j C_j. \quad (14)$$

The PF factor indicates that if a user has low rates for a long time, its PF factor increases subsequently. Optimizing the PF factor is easier and directly influences the objective function, maximizing the score function,  $C_{\text{score}}$ . The optimization problem is now formulated as

$$\mathbf{P1} : \max_A \sum_{t=1}^T \sum_{j=1}^J (w_j)^{\lambda} \cdot C_j, \quad (15)$$

where  $\lambda \in [0, 1]$  controls the trade-off between the sum-rate and the 5-percentile rate.

Following [26] and to generalize the problem, we use the PF ratio to prioritize the associated UEs of each AP to limit the number of UEs that each AP can choose from to  $N$  users at each time  $t$ . These  $N$  UEs are the top  $N$  in PF ratios among all the associated UEs to a specific AP. This step is common in unifying the action space size among network configurations.

## III. MARL FORMULATION

In this section, we formulate the problem using a partially observable Markov decision process (PO-MDP) and present an online solution using MARL.

#### A. Partially-Observable Markov Decision Process

To solve the optimization problem in (12), we rely on MARL. In particular, we assume each AP is an individual agent contributing to his policy toward maximizing the score function. In PO-MDP, each agent  $i$  observes his local observation  $o_i$ , takes an action  $a_i$ , and receives a reward  $r$ . Jointly,

the local observations of all agents together form the state space, and the actions of all agents form the action space  $A$ . Sharing the local observations among all agents converts the problem into fully observable MDP. The PO-MDP formulation is detailed as follows:

- 1) **Local Observations:** Each agent  $i$  observes a tuple comprised of the SINR,  $\gamma_j$ , of its top  $N$  users and their weighting factor  $w_j$ . Then, each AP has a local observation  $o_i = ((\gamma_{i,1}, w_{i,1}), \dots, (\gamma_{i,N}, w_{i,N}))$  whose size is  $2 \times N$ . The state space is the concatenation of all local observations  $S = (o_1, \dots, o_I)$  whose size is  $2 \times N \times I$ , i.e., the local observations of all  $I$  APs.
  - 2) **Actions:** The action space of each agent  $a_i$  at time  $t$  comprises the UEs among its top  $N$  users chosen to be served and additional silent action (the agent turned itself off). The size of the individual action space is  $N + 1$ . The joint action space  $A = (a_1, \dots, a_I)$  has a size of  $(N + 1) \times I$ .
  - 3) **Rewards:** We use a joint centralized reward function based on the actions of all APs. The reward is
- $$r = \sum_{j=1}^J w_j^\lambda C_j. \quad (16)$$
- 4) **Policies:** Each agent's policy, denoted as  $\pi_i(o_i|a_i)$ , maps the chosen action at each visiting observation. The global policy of the environment  $\pi(S|A)$  maps the joint action to the global state. The goal is to find the optimal global policy that maximizes the rewards.

### B. Online RL

Online RL, especially deep RL, efficiently solves complex and large-scale problems, such as the presented RRM problem. In this subsection, we define the preliminaries of single-agent RL needed to better define the proposed offline MARL scheme. In this work, we choose discrete SAC (an actor-critic algorithm for environments with discrete actions) [36] as our online RL algorithm due to its stability compared to DQN, which usually sticks to local minimums and saddle points. SAC is a model-free, off-policy RL algorithm that optimizes the current policy by utilizing experiences from previous visits (across various policies). It uniquely maximizes the policy's rewards and entropy, promoting continuous and random exploration of the environment. This dual objective ensures that SAC seeks optimal actions and maintains sufficient exploration to avoid local optima. Actor-critic architectures rely on policy evaluation and policy improvement alternately [37]. SAC computes the Q-function iteratively via the policy evaluation loss

$$\mathcal{L}_{\text{eval}} = \hat{\mathbb{E}} \left[ \left( r + \beta \hat{\mathbb{E}}_{a' \sim \pi^k(a'|s')} [\hat{Q}^{(k)}(s', a') - Q(s, a)] \right)^2 \right], \quad (17)$$

where  $\hat{\mathbb{E}}[\cdot]$  is the empirical expectation over samples  $(s, a, r, s')$ ,  $\hat{Q}^{(k)}$  is the current estimate of the Q-function  $Q$  at iteration step  $k$ ,  $s'$  is the next state,  $a'$  is the next action, and the Q-function  $Q(s, a)$  is usually modeled as a neural network parameterized by weights  $\theta$ . The latter updates the

policy towards maximizing the expected Q-function through the policy improvement loss

$$\mathcal{L}_{\text{imp}} = -\hat{\mathbb{E}}_{a \sim \pi^k(s|a)} \left[ \hat{Q}^k(s, a) - \log \pi(a|s) \right], \quad (18)$$

where the term  $\log \pi(a|s)$  is the entropy regularization parameter, and the policy  $\pi(a|s)$  is usually modelled as a neural network parameterized by weights  $\phi$ .

### C. Online MARL

Multiple agents cooperate toward the joint goal in the cooperative multi-agent RL setting. There are numerous variants of the MARL problem:

1) *Centralized MARL (C-MARL):* In this MARL setting, all agents are considered as one agent with one joint observation (state space) and one joint action space. The target is to find the optimal joint policy  $\pi$  that maps the state space to the joint action space. Then, individual policies are extracted from the joint policy. Hence, the loss equations for the SAC algorithm in that case is

$$\mathcal{L}_{\text{eval}}^{\text{C}} = \hat{\mathbb{E}} \left[ \left( r + \beta \hat{\mathbb{E}}_{A' \sim \pi^k(A'|S')} \hat{Q}^{(k)}(S', A') - Q(S, A) \right)^2 \right], \quad (19)$$

where  $\hat{Q}^{(k)}$  is the current estimate of the joint Q-function at iteration step  $k$ . Similarly, the policy improvement loss for each agent is computed as follows:

$$\mathcal{L}_{\text{imp}}^{\text{C}} = -\hat{\mathbb{E}}_{A \sim \pi^k(A|S)} \left[ \hat{Q}^k(S, A) - \log \pi(S|A) \right]. \quad (20)$$

C-MARL efficiently finds the optimal policies as it utilizes all agents' information jointly but at the cost of high complexity due to the vast state and action space dimensions.

2) *Independent Training MARL (I-MARL):* In I-MARL, each agent individually optimizes its policy using a distinct neural network that maps its observations to its action space. Consequently, each agent independently implements its own SAC algorithm, relying solely on its observations, actions, and accumulated experiences to refine its strategy. Thus, the policy evaluation loss of each agent is computed as

$$\mathcal{L}_{\text{eval}}^{\text{I}} = \hat{\mathbb{E}} \left[ \left( r + \beta \hat{\mathbb{E}}_{a'_i \sim \pi_i^k(a'_i|o'_i)} \hat{Q}_i^{(k)}(o'_i, a'_i) - Q_i(o_i, a_i) \right)^2 \right], \quad (21)$$

where  $\hat{Q}_i^{(k)}$  is the current estimate of the Q-function of agent  $i$  at iteration step  $k$ . Similarly, the policy improvement loss for each agent is computed as follows:

$$\mathcal{L}_{\text{imp}}^{\text{I}} = -\hat{\mathbb{E}}_{a_i \sim \pi_i^k(o_i|a_i)} \left[ \hat{Q}_i^{(k)}(o_i, a_i) - \log \pi_i(a_i|o_i) \right]. \quad (22)$$

I-MARL overcomes the complexity of C-MARL. However, it performs worse than C-MARL due to the missing information about the observations of other agents.

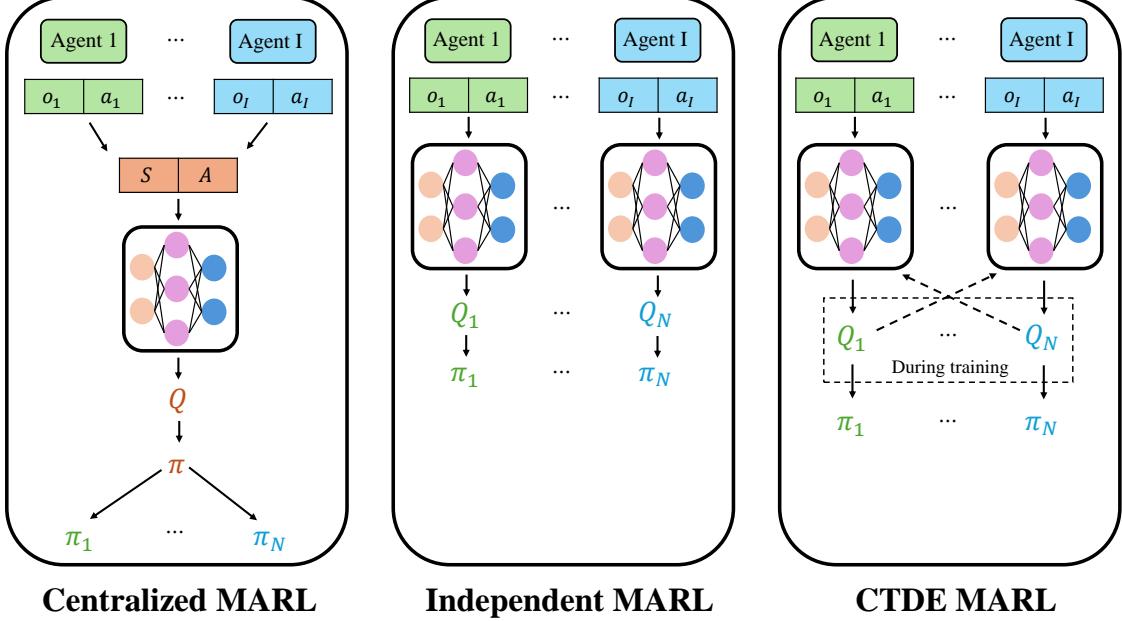


Fig. 2: An illustrative comparison between centralized MARL, independent MARL, and centralized training decentralized execution MARL. As shown, C-MARL models a joint Q-function using a single neural network, and the policies are drawn from the joint Q-function. In contrast, in I-MARL and CTDE-MARL, each agent models its Q-function as an independent neural network.

3) *Centralized training decentralized execution MARL (CTDE-MARL)*: In the CTDE setting, value decomposition [38] approximates the global value function through the sum of the individual action-value functions of each agent, *i.e.*,

$$Q_{tot}(s) = \sum_{i=1}^I \tilde{Q}^i(o_i), \quad (23)$$

where  $\tilde{Q}^i(o_i)$  represents the contribution of each agent to the global Q-function. In this framework, a joint policy evaluation loss is calculated based on the critical contributions of each agent, while each agent independently computes its policy improvement loss. The CTDE-MARL evaluation loss is

$$\begin{aligned} \mathcal{L}_{\text{eval}}^{\text{CTDE}} = \hat{\mathbb{E}} \left[ \left( r + \beta \hat{\mathbb{E}}_{a'_i \sim \pi_i^k(a'_i|o'_i)} \sum_{i=1}^I \hat{Q}_i^{(k)}(o'_i, a'_i) \right. \right. \\ \left. \left. - \sum_{i=1}^I \tilde{Q}_i(o_i, a_i) \right)^2 \right]. \end{aligned} \quad (24)$$

Then, the policy of each agent is obtained from the optimized function  $\tilde{Q}_i(o_i, a_i)$  as

$$\pi_i(a_i|o_i) = \mathbb{1} \left\{ a_i = \arg \max_{a_i} \tilde{Q}_i(o_i, a_i) \right\}. \quad (25)$$

During execution, each agent uses his policy in a decentralized fashion. CTDE-MARL overcomes the complexity of C-MARL and the inefficiency of I-MARL.

Fig. 2 illustrates the three online MARL variants: C-MARL, I-MARL, and CTDE-MARL, graphically highlighting their similarities and differences. Note that each agent calculates its action-value functions in I-MARL and CTDE-MARL. However, CTDE-MARL improves upon I-MARL by aggregating the individual action-value functions to contribute to a global

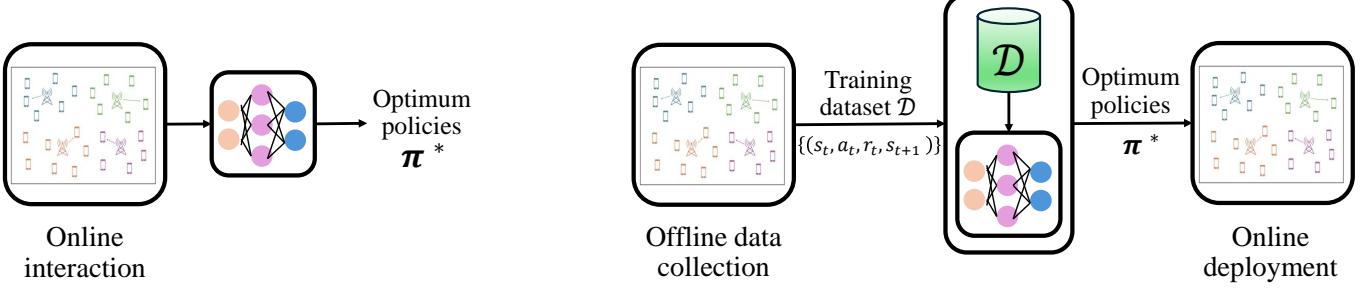
action-value function during training, as denoted in (23) and (24). This approach enhances coordination among agents by aligning their objectives toward the global objective.

#### IV. OFFLINE MARL

In the previous section, we presented online MARL and its variants. This section presents the proposed offline MARL scheme for the RRM problem. As shown in Fig. 3, offline RL / MARL utilizes static offline dataset without any reliance on online interaction with the environment [16]. This offline dataset is collected using behavioral policies from benchmark algorithms or random exploration. Since offline MARL only uses offline datasets, it removes the burden of online interaction in the RRM problem. For instance, to reach a sub-optimal policy in MARL algorithms, the agents need a large amount of online interaction with the environment. Therefore, the agents must visit as many state-action pairs as possible, which is costly regarding time and computations. Moreover, online MARL requires a high level of synchronization between the agents using a central unit that collects the agent's actions and re-distributes the calculated rewards. This creates a challenging communication overhead problem, which becomes even more cumbersome when information sharing is needed among the agents or centralized training.

Simply training the online SAC algorithm (described in Section III-B) with an offline dataset usually fails. This failure occurs because of the optimistic evaluation of the algorithm caused by the distributional shift between the learned and deployed actions, called out-of-distribution (OOD) actions.

However, recent advances in offline RL and offline MARL overcome this issue and have enabled deep RL and deep MARL algorithms to be used offline. For instance, behavior-constrained Q-learning (BCQ) [39] solved the OOD problem



## Online MARL

Fig. 3: An illustrative comparison between online MARL and offline MARL. Online MARL utilizes online interaction with the environment to optimize the policies. In contrast, offline MARL exploits offline datasets pre-collected using a behavioral policy. Offline MARL training uses the offline dataset, whereas optimum policies are used for online deployment.

**Algorithm 1:** Centralized multi-agent reinforcement learning using conservative Q-learning (C-MARL-CQL) algorithm.

```

Input: Discount factor  $\beta$ , conservative penalty constant  $\alpha$ , number of agents  $I$ , number of training iterations  $K$ , number of gradient steps  $G$ , and offline dataset  $\mathcal{D}$ 
Initialize network parameters
for iteration  $k$  in  $\{1, \dots, K\}$  do
    for gradient step  $g$  in  $\{1, \dots, G\}$  do
        Sample a batch  $\mathcal{B}$  from the dataset  $\mathcal{D}$ 
        Estimate the C-MARL-CQL loss  $\mathcal{L}_{\text{CQL}}^{\text{C}}$  in (26)
        Estimate the policy improvement loss  $\mathcal{L}_{\text{imp}}^{\text{C}}$  using (20)
        Perform a stochastic gradient step based on the estimated losses.
    end
end
```

**Algorithm 2:** Independent multi-agent reinforcement learning using conservative Q-learning (I-MARL-CQL) algorithm.

```

Input: Discount factor  $\beta$ , conservative penalty constant  $\alpha$ , number of agents  $I$ , number of training iterations  $K$ , number of gradient steps  $G$ , and offline dataset  $\mathcal{D}$ 
Initialize networks parameters
for iteration  $k$  in  $\{1, \dots, K\}$  do
    for gradient step  $g$  in  $\{1, \dots, G\}$  do
        Sample a batch  $\mathcal{B}$  from the dataset  $\mathcal{D}$ 
        for agent  $i$  in  $\{1, \dots, I\}$  do
            Estimate the I-MARL-CQL loss  $\mathcal{L}_{\text{CQL}}^{\text{I}}$  using (27)
            Estimate the policy improvement loss  $\mathcal{L}_{\text{imp}}^{\text{I}}$  using (22)
            Perform a stochastic gradient step based on the estimated losses.
        end
    end
end
```

by limiting the distance between the selected actions for the policy to those in the dataset. In contrast, conservative Q-learning (CQL) [19] adds a regularization term to the loss function to penalize large deviations between the selected actions and the actions in the dataset.

Next, we revise these algorithms in light of the RRM problem proposed in Section II and the SAC architecture discussed in Section III.

### A. Conservative Q-Learning

In this work, we choose the CQL algorithm as the offline MARL approach for solving the RRM problem due to its robust performance on various offline RL / MARL problems. In addition, we introduce a new variation of the algorithm as we build the CQL algorithm on top of SAC architecture. As in the online case, next, we assess the algorithm for three variants: centralized (C), independent (I), and CTDE.

1) *C-MARL-CQL*: To implement the CQL algorithm in the C-MARL setting, which we refer to as C-MARL-CQL, the

policy improvement loss is calculated as

$$\mathcal{L}_{\text{CQL}}^{\text{C}} = \frac{1}{2} \mathcal{L}_{\text{eval}}^{\text{C}} + \alpha \hat{\mathbb{E}} \left[ \log \left( \sum_A \exp(Q(S, A)) \right) - Q(S, A) \right], \quad (26)$$

where the term  $\alpha \hat{\mathbb{E}} \left[ \log \left( \sum_A \exp(Q(S, A)) \right) - Q(S, A) \right]$  is the regularization term (KL-divergence) and  $\alpha > 0$  is a constant. Then, the policy improvement is performed using (20) as in the online case. The C-MARL-CQL procedure is detailed in Algorithm 1.

2) *I-MARL-CQL*: To implement the CQL algorithm in the I-MARL setting, named I-MARL-CQL, each agent  $i$  computes its policy improvement loss as

$$\mathcal{L}_{\text{CQL}}^{\text{I}} = \frac{1}{2} \mathcal{L}_{\text{eval}}^{\text{I}} + \alpha \hat{\mathbb{E}} \left[ \log \left( \sum_{a_i} \exp(Q_i(o_i, a_i)) \right) - Q_i(o_i, a_i) \right], \quad (27)$$

where the term  $\alpha \hat{\mathbb{E}} \left[ \log \left( \sum_{a_i} \exp(Q_i(o_i, a_i)) \right) - Q_i(o_i, a_i) \right]$  is the regularization term for each agent. Then, the policy

---

**Algorithm 3:** Centralized training decentralized execution multi-agent reinforcement learning using conservative Q-learning (CTDE-MARL-CQL) algorithm.

---

**Input:** Discount factor  $\beta$ , conservative penalty constant  $\alpha$ , number of agents  $I$ , number of training iterations  $K$ , number of gradient steps  $G$ , and offline dataset  $\mathcal{D}$

Initialize networks parameters

**for** iteration  $k$  in  $\{1, \dots, K\}$  **do**

- for** gradient step  $g$  in  $\{1, \dots, G\}$  **do**
- Sample a batch  $\mathcal{B}$  from the dataset  $\mathcal{D}$
- Estimate the CTDE-MARL-CQL loss  $\mathcal{L}_{\text{CQL}}^{\text{CTDE}}$  using (28)
- for** agent  $i$  in  $\{1, \dots, I\}$  **do**
- Estimate the policy improvement loss  $\mathcal{L}_{\text{imp}}^i$  using (22)
- end**
- Perform a stochastic gradient step based on the estimated losses.

**end**

**end**

---

improvement is performed using (22), as in the online case. Algorithm 2 details the I-MARL-CQL algorithm.

3) *CTDE-MARL-CQL*: Finally, the CQL loss in the CTDE form, which we call CTDE-MARL-CQL, is formulated as

$$\begin{aligned} \mathcal{L}_{\text{CQL}}^{\text{CTDE}} &= \frac{1}{2} \mathcal{L}_{\text{eval}}^{\text{CTDE}} \\ &+ \alpha \hat{\mathbb{E}} \sum_{i=1}^I \left[ \log \left( \sum_{a_i} \exp(\tilde{Q}_i(o_i, a_i)) \right) - \tilde{Q}_i(o_i, a_i) \right]. \end{aligned} \quad (28)$$

The policy improvement is performed using (22) for each agent. Lastly, the CTDE-MARL-CQL algorithm is detailed in Algorithm 3.

We highlight that the key difference between each offline MARL scheme and its corresponding online MARL scheme is the carefully designed conservative term in the offline case, which pushes the learned policy close to the behavioral policy in the dataset. Therefore, Algorithms 1 to 3 can be converted to the online counterpart by replacing in the appropriate policy evaluation loss function. Note that for the C-MARL-CQL, a single neural network is used to model the Q-function (*i.e.*, the policy), whereas, in CTDE-MARL-CQL and I-MARL-CQL, each agent models its neural network.

## V. NUMERICAL RESULTS

This section presents the numerical analysis of the designed offline MARL algorithms for the RRM problem. First, we present the implementation and the baseline models, then show the experimental results of the proposed model compared to the baseline models.

### A. Implementation and Baseline Models

We consider a  $100 \text{ m} \times 100 \text{ m}$  square area with  $I = 4$  APs (agents) and  $J = 20$  UEs. At the beginning of each episode, one random environment is sampled with different AP positions and UEs' initial positions. Each episode consists of

200 time steps. We use 2 hidden layers in actor and critic with 256 neurons each. All simulations are performed on a single NVIDIA Tesla V100 GPU using the Pytorch framework. Simulation parameters are shown in Table III. First, we show

TABLE III: Simulation parameters

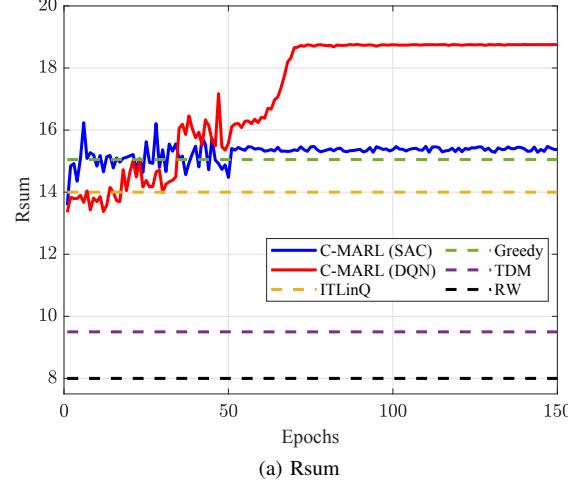
Parameter	Value	Parameter	Value
$I$	4	$J$	20
$N$	3	$L$	100
$d_0$	10 m	$d_1$	1 m
$v(t)$	1 m/s	$PL_o$	10 dB
$p_t$	10 dBm	$T$	200
$\mu_1$	$\frac{1}{M}$	$\mu_2$	3
$\lambda$	0.8	$\beta$	0.99
$\alpha$	1	Replay memory	$10^5$
Actor $lr$	$1e-5$	Critic $lr$	$1e-4$
Layers	2	Neurons	256
Optimizer	Adam	Activation	ReLU

the performance of online C-MARL (SAC) compared to the baseline models and the famous online C-MARL (DQN) as a learning-based baseline. Afterward, we compared the online performance to C-MARL-CQL (SAC). Then, we present the performance of the proposed offline MARL schemes, *i.e.*, *C-MARL-CQL (SAC)*, *I-MARL-CQL (SAC)*, and *CTDE-MARL-CQL (SAC)*. Using SAC as our deep RL framework, we always compare it to DQN. In our simulation, we use (SAC) after the name of the algorithm to refer to a scheme built on top of SAC architecture and (DQN) after the name of the algorithm to refer to a scheme built on top of the traditional DQN architecture. Finally, we show the effect of the quality and size of the dataset on the offline training.

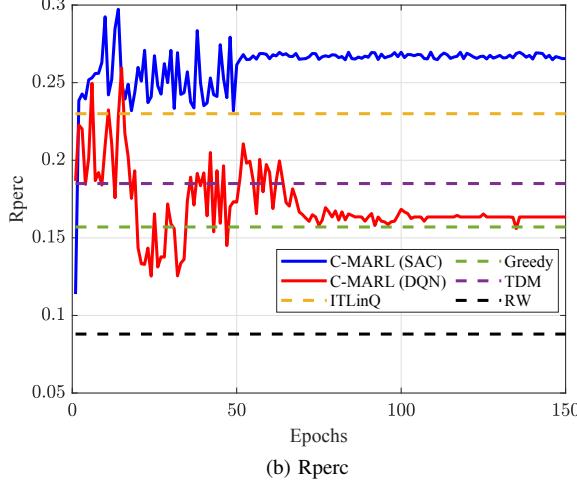
We collect different offline datasets using the experience of an C-MARL (SAC) agent and different sub-optimal benchmarks, respectively. Since the performance of the offline MARL algorithms is sensitive to the quality of the offline dataset, we adopt the online centralized MARL approach to collect good-quality data points. In addition, we test the effect of the size of the dataset on the performance of offline MARL schemes by collecting datasets with different sizes.

Besides the three developed MARL algorithms, namely C-MARL-CQL, I-MARL-CQL, and CTDE-MARL-CQL, we show the performance of four benchmarks from the literature:

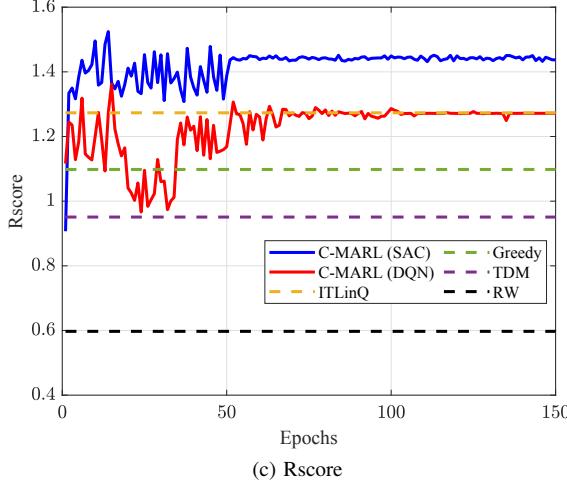
- 1) **Random-walk (RW):** At each time step  $t$ , each AP chooses randomly to serve one of its top  $N$  UEs.
- 2) **Greedy:** In the greedy method, each agent serves the user with the highest SINR among its top  $N$  users at each time step  $t$ .
- 3) **Time-division multiplexing (TDM):** At each time step  $t$ , all UEs are served equally, where each AP serves the UEs in a round-robin manner.
- 4) **Information-theoretic link scheduling (ITLinQ):** It was proved in [40], that ITLinQ algorithm reaches a sub-optimal policy. At each time step  $t$ , each AP sorts its top  $N$  UEs regarding their PF ratios. Then, each AP performs an interference tolerance check for each UE to ensure that the interference level is lower than a threshold  $MSNR_{mn}^\eta$ . This AP is turned off if no UEs have lower interference than the threshold.



(a) Rsum



(b) Rperc

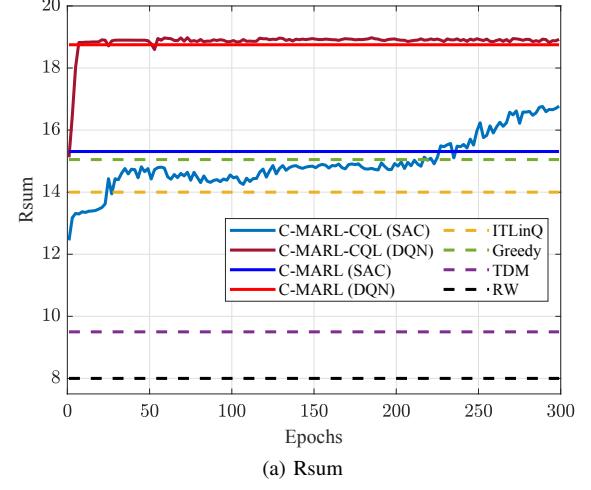


(c) Rscore

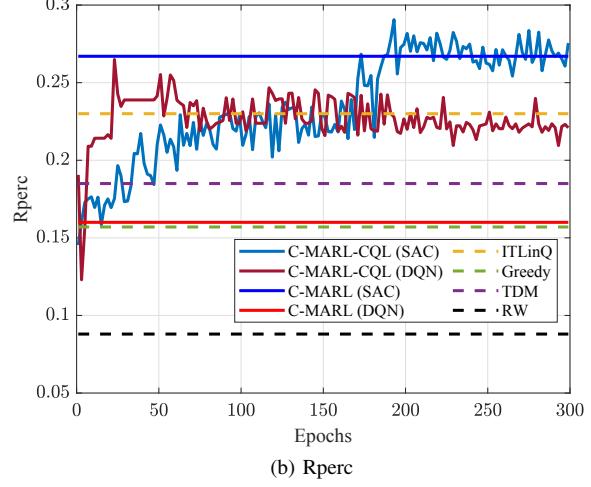
Fig. 4: The sum rate, 5-percentile rate, and Rscore reported for C-MARL algorithm built on top of both SAC and DQN compared to other benchmark schemes.

### B. Online Training and Dataset Collection

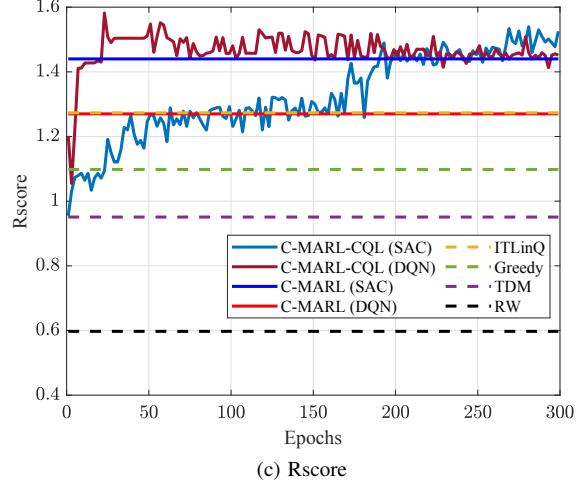
In Fig. 4, we show the training performance of a C-MARL (SAC) agent compared to the baseline schemes. In addition, we



(a) Rsum



(b) Rperc



(c) Rscore

Fig. 5: The sum rate, 5-percentile rate, and Rscore reported for the proposed C-MARL-CQL algorithm built on top of SAC and DQN compared to C-MARL and other benchmark schemes.

compare the developed C-MARL (SAC) scheme to a famous online RL algorithm, C-MARL (DQN) scheme, as a learning-based benchmark. First, the RW has the worst sum rate, 5-

percentile rate, and Rscore. The greedy algorithm maximizes the sum rate at the expense of the 5-percentile rate. In contrast, the TDM scheme prioritizes maximizing the 5-percentile rate over the sum rate. The ITLinQ benchmark has the highest Rscore among other baselines. The two online RL schemes, namely, C-MARL (SAC) and C-MARL (DQN), have the highest Rscore compared to all the baselines. We can notice in Fig. 4a that the C-MARL (DQN) agent scores the largest sum rate, whereas the C-MARL (SAC) agent maintains a relatively good sum rate compared to other benchmarks. In contrast, the 5-percentile rate of the C-MARL (DQN) agent drops noticeably compared to C-MARL (SAC), whose 5-percentile rate is around 0.28, as shown in Fig. 4b. As a result, C-MARL (SAC) has a better overall Rscore than C-MARL (DQN), as in Fig. 4c.

### C. Offline Centralized Training

In the next experiment in Fig. 5, we show the performance of the proposed C-MARL-CQL (SAC) algorithm in terms of the sum rate, 5-percentile rate, and Rscore. We compare the proposed algorithm to the traditional C-MARL-CQL (DQN), C-MARL (SAC), C-MARL (DQN) and other baselines as benchmarks to better evaluate it. In this experiment, we use the dataset, which is 16000 data points, collected from an online SAC agent. In addition, we perform centralized training, *i.e.*, C-MARL-CQL. As shown in Fig. 5a and Fig. 5b, and similar to the online case, the C-MARL-CQL (DQN) algorithm prefers to maximize the sum rate over the 5-percentile rate, where the proposed C-MARL-CQL (SAC) algorithm sacrifices the sum rate to enhance the 5-percentile rate. In Fig. 5c, the C-MARL-CQL (SAC) algorithm outperforms the C-MARL-CQL (DQN) scheme, surpassing other baselines, including online benchmarks.

### D. Offline MARL Schemes

In Fig. 6, we report the rates of the proposed offline MARL schemes, namely, C-MARL-CQL, I-MARL-CQL and CTDE-MARL-CQL built on top of SAC architecture, compared to C-MARL (SAC) and C-MARL (DQN) as two benchmark schemes. We construct the three schemes on top of SAC architecture due to its stable and high rate convergence. As in Fig. 6a, the three MARL schemes almost achieve the same sum rate, outperforming C-MARL (SAC). In contrast, C-MARL-CQL (SAC) has the highest 5-percentile rate, relying on the availability of complete observations of all agents to find the optimum policies. Comparing CTDE-MARL-CQL (SAC) to I-MARL-CQL (SAC), we observe that CTDE-MARL-CQL (SAC), due to value function sharing among agents, approaches the 5-percentile rate of C-MARL-CQL (SAC) with lower complexity, especially in execution. Hence, CTDE-MARL-CQL (SAC) outperforms C-MARL (SAC) Rscore with a tiny gap with C-MARL-CQL (SAC), as shown in Fig. 6c. This highlights the ability of the CTDE framework to overcome the complexity of centralized training and the poor performance of independent training.

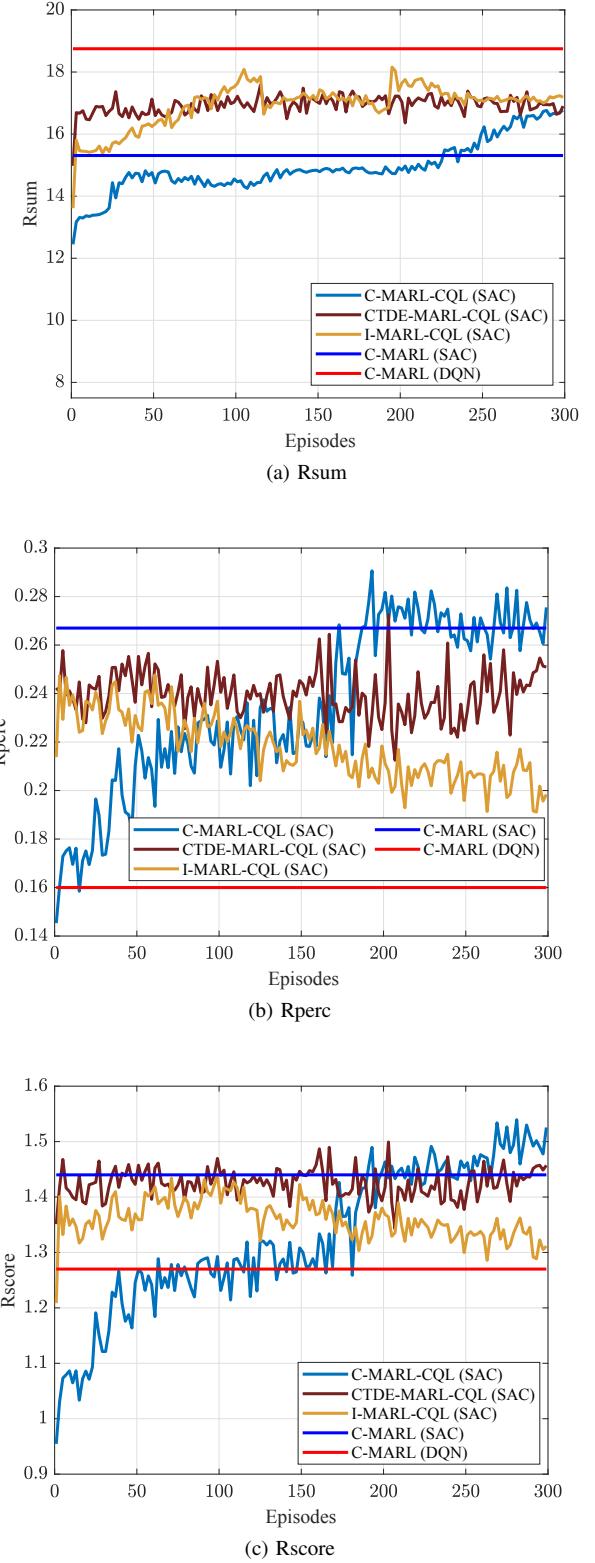


Fig. 6: The sum rate, 5-percentile rate, and Rscore reported for the proposed C-MARL-CQL, I-MARL-CQL and CTDE-MARL-CQL built on top of SAC architecture.

### E. Dataset Quality

Finally, we show the effect of the quality of the dataset and its size on the Rscore performance of the proposed CTDE-

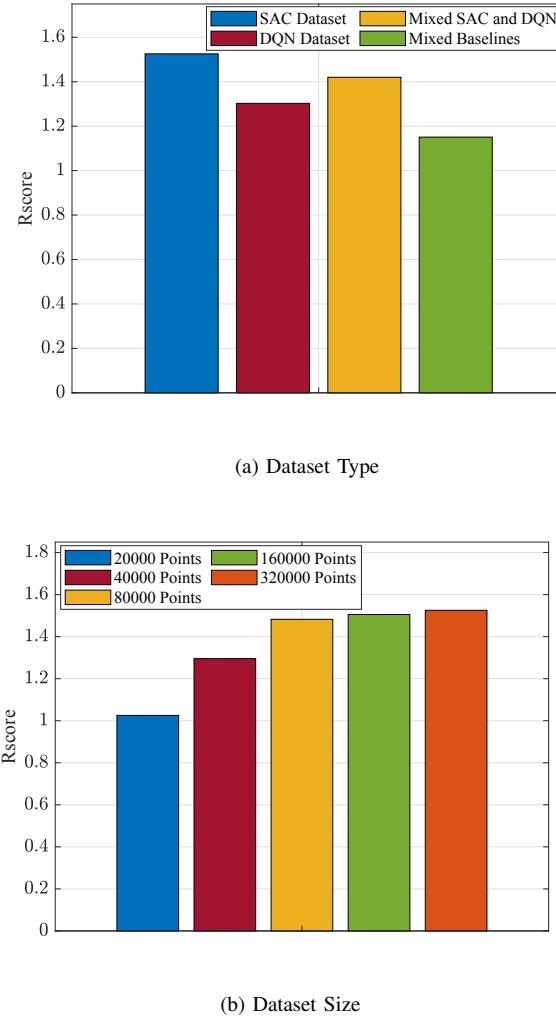


Fig. 7: The effect of the dataset on the overall performance of the proposed CTDE-MARL-CQL (SAC) scheme in terms of the achieved Rscore. Shown in (a) the effect of the quality of the collected dataset and (b) the effect of the dataset size.

MARL-CQL (SAC) presented in Fig. 7. In particular, Fig. 7a compares four sources of the offline dataset, *i.e.*, online SAC, online DQN, the mixture of online SAC and online DQN, and the mix of other baselines<sup>1</sup>. The quality of the policy used to collect the offline dataset directly reflects the achieved Rscore. A dataset collected from a good policy, such as online SAC, outperforms other datasets collected from online DQN agents and baseline policies. A mixture of SAC and DQN agents achieves a high score. This highlights that a mix of good and bad quality datasets can still be used to find a suitable policy offline [31].

Similar to Fig. 7a, Fig. 7b shows the effect of the size of the dataset on the convergence of the Rscore of the proposed CTDE-MARL-CQL (SAC) algorithm. When using a small dataset of 20000 points, the Rscore drops to 1, similar to the performance of TDM. The lack of enough experience creates optimistic uncertainty in the CQL algorithm, forcing itself to

<sup>1</sup>We only include experiences from RW, greedy, TDM, and ITLinQ benchmarks in this dataset.

converge to a saddle sub-optimal policy. When we increase the size of the dataset, the Rscore rapidly increases. Datasets with dimensions larger than 320000 data points influence the convergence stability without achieving higher Rscore values.

## VI. CONCLUSIONS

This paper presents an offline MARL framework based on the SAC architecture and the CQL algorithm for optimizing resource management in wireless networks with multiple APs serving UEs. The framework introduces three variants: C-MARL-CQL (centralized training), I-MARL-CQL (independent training), and CTDE-MARL-CQL, tailored to balance computational complexity and policy performance. Numerical results demonstrate that the offline MARL framework significantly outperforms baselines, including random-walk, greedy algorithms, TDM, and ITLinQ, regarding the Rscore metric. Among the variants, CTDE-MARL-CQL achieves the best trade-off, offering reduced computational complexity compared to C-MARL-CQL while surpassing I-MARL-CQL in policy effectiveness. Our analysis also underscores the importance of dataset quality and size in determining algorithm convergence and performance. High-quality behavioral datasets enhance rate optimization, while larger datasets contribute to stable convergence. These insights provide valuable guidance for offline MARL applications in wireless systems.

Future research will focus on extending this work to meta-offline RL and MARL, enabling dynamic adaptability to evolving environments and objectives. This direction can further enhance the scalability and robustness of offline MARL solutions in complex wireless communication scenarios.

## REFERENCES

- [1] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, “Applications of deep reinforcement learning in communications and networking: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.
- [2] N. H. Mahmood, S. Böcker, I. Moerman, O. A. López, A. Munari, K. Mikhaylov, F. Clazzer, H. Bartz, O.-S. Park, E. Mercier *et al.*, “Machine type communications: key drivers and enablers towards the 6G era,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 134, 2021.
- [3] Y. Shi, L. Lian, Y. Shi, Z. Wang, Y. Zhou, L. Fu, L. Bai, J. Zhang, and W. Zhang, “Machine learning for large-scale optimization in 6G wireless networks,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2088–2132, 2023.
- [4] E. Eldeeb, M. Shehab, A. E. Kalø r, P. Popovski, and H. Alves, “Traffic prediction and fast uplink for hidden markov IoT models,” *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17172–17184, 2022.
- [5] X. Yi and G. Caire, “ITLinQ+: An improved spectrum sharing mechanism for device-to-device communications,” in *2015 49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 1310–1314.
- [6] A. Gjendemsjo, D. Gesbert, G. E. Oien, and S. G. Kiani, “Binary power control for sum rate maximization over multiple interfering links,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, pp. 3164–3173, 2008.
- [7] L. Song, D. Niyato, Z. Han, and E. Hossain, “Game-theoretic resource allocation methods for device-to-device communication,” *IEEE Wireless Communications*, vol. 21, no. 3, pp. 136–144, 2014.
- [8] M. Zangoeei, N. Saha, M. Golkarifard, and R. Boutaba, “Reinforcement learning for radio resource management in RAN slicing: A survey,” *IEEE Communications Magazine*, vol. 61, no. 2, pp. 118–124, 2023.
- [9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

- [10] M. V. Da Silva, E. Eldeeb, M. Shehab, H. Alves, and R. D. Souza, "Distributed learning methodologies for massive machine type communication," *Authorea Preprints*, 2024.
- [11] A. Oroojlooy and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *Applied Intelligence*, vol. 53, no. 11, pp. 13 677–13 722, 2023.
- [12] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PloS one*, vol. 12, no. 4, p. e0172395, 2017.
- [13] E. Eldeeb, M. Shehab, and H. Alves, "Traffic learning and proactive UAV trajectory planning for data uplink in markovian IoT models," *IEEE Internet of Things Journal*, vol. 11, no. 8, pp. 13 496–13 508, 2024.
- [14] S. V. Albrecht, F. Christianos, and L. Schäfer, *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. [Online]. Available: <https://www.marl-book.com>
- [15] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning," 2017. [Online]. Available: <https://arxiv.org/abs/1706.05296>
- [16] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.
- [17] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit Q-learning," 2021. [Online]. Available: <https://arxiv.org/abs/2110.06169>
- [18] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.
- [19] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1179–1191. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/0d2b2061826a5df322116a5085a6052-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df322116a5085a6052-Paper.pdf)
- [20] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.
- [21] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25 463–25 473, 2018.
- [22] E. Eldeeb, J. M. de Souza Sant'Ana, D. E. Pérez, M. Shehab, N. H. Mahmood, and H. Alves, "Multi-UAV path learning for age and power optimization in IoT with UAV battery recharge," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 5356–5360, 2022.
- [23] Y. Hua, R. Li, Z. Zhao, X. Chen, and H. Zhang, "GAN-Powered Deep Distributional Reinforcement Learning for Resource Management in Network Slicing," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 2, pp. 334–349, 2020.
- [24] N. NaderiAlizadeh, M. Eisen, and A. Ribeiro, "Learning resilient radio resource management policies with graph neural networks," *IEEE Transactions on Signal Processing*, vol. 71, pp. 995–1009, 2023.
- [25] E. Eldeeb, M. Shehab, and H. Alves, "Age minimization in massive IoT via UAV swarm: A multi-agent reinforcement learning approach," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–6.
- [26] N. NaderiAlizadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3507–3523, 2021.
- [27] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.
- [28] S. Hwang, H. Kim, H. Lee, and I. Lee, "Multi-agent deep reinforcement learning for distributed resource management in wirelessly powered communication networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 14 055–14 060, 2020.
- [29] X. Du, T. Wang, Q. Feng, C. Ye, T. Tao, L. Wang, Y. Shi, and M. Chen, "Multi-agent reinforcement learning for dynamic resource management in 6G in-X subnetworks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1900–1914, 2023.
- [30] E. Eldeeb, H. Sifaou, O. Simeone, M. Shehab, and H. Alves, "Conservative and risk-aware offline multi-agent reinforcement learning for digital twins," *arXiv preprint arXiv:2402.08421*, 2024.
- [31] K. Yang, C. Shi, C. Shen, J. Yang, S.-p. Yeh, and J. J. Sydir, "Offline reinforcement learning for wireless network optimization with mixture datasets," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
- [32] E. Eldeeb and H. Alves, "Offline and distributional reinforcement learning for radio resource management," 2024. [Online]. Available: <https://arxiv.org/abs/2409.16764>
- [33] L. Meng, M. Wen, C. Le, X. Li, D. Xing, W. Zhang, Y. Wen, H. Zhang, J. Wang, Y. Yang *et al.*, "Offline pre-trained multi-agent decision transformer," *Machine Intelligence Research*, vol. 20, no. 2, pp. 233–248, 2023.
- [34] 3GPP, "Simulation assumptions and parameters for FDD HeNB RF requirements," *Tech. Rep. R4-092042*.
- [35] ———, "NR; physical layer measurements," *Technical specification (TS) 8.2.15 V18.4.0*, 2024-12.
- [36] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: <https://proceedings.mlr.press/v80/haarnoja18b.html>
- [37] ———, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: <https://proceedings.mlr.press/v80/haarnoja18b.html>
- [38] J. Su, S. Adams, and P. Beling, "Value-decomposition multi-agent actor-critics," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 13, 2021, pp. 11 352–11 360.
- [39] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 2052–2062. [Online]. Available: <https://proceedings.mlr.press/v97/fujimoto19a.html>
- [40] N. NaderiAlizadeh and A. S. Avestimehr, "ITLinQ: A new approach for spectrum sharing in device-to-device communication systems," in *2014 IEEE International Symposium on Information Theory*, 2014, pp. 1573–1577.