Received 2 April 2026; accepted 30 April 2026. Date of publication 6 May 2026; date of current version 19 May 2026.

Digital Object Identifier 10.1109/OJCOMS.2026.3690661

Dynamic Resource Optimization for a Joint Max-Min Fairness and Energy-Efficiency Problem in NOMA-Aided Underwater Optical Wireless Systems

IMENE ROMDHANE[1] (Student Member, IEEE), ZIYAUR RAHMAN[1] (Student Member, IEEE), NAHED BELHADJ MOHAMED[1], MD. ZOHEB HASSAN[2] (Member, IEEE), AND GEORGES KADDOUM[1] (Senior Member, IEEE)

[1]Department of Electrical Engineering, École de Technologie Supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada
[2]Department of Electrical and Computer Engineering, Université Laval, Quebec, QC G1V 0A6, Canada

CORRESPONDING AUTHOR: I. ROMDHANE (imene.romdhane.1@ens.etsmtl.ca)

ABSTRACT In this paper, we present a dynamic beamforming optimization framework for multi-beam underwater optical wireless communication (UOWC) systems. The UOWC transmitter concurrently transmits multiple beams, and non-orthogonal multiple access (NOMA) is applied within each beam to support multi-user communications. Our goal is to maximize the energy efficiency (EE) of a multi-beam UOWC network while guaranteeing the max-min fairness. Hence, we propose two deep deterministic policy gradient (DDPG)-based beamforming solutions that optimize beam orientations and power allocation while considering the quasi-stationarity of nodes in the underwater environment. The first solution is a single-agent DDPG-based approach, while the second solution is a multi-agent DDPG-based one. We also incorporate sequential learning capabilities into the multi-agent DDPG approach to enhance its optimality, which includes sequential learning of the beam orientation and power allocation tasks. Through extensive simulations, we show that the proposed single and multi-agent DDPG solutions achieve improved fairness and EE as compared to the equal power allocation, weighted minimum mean-square error (WMMSE)-based EE maximization with min-rate constraint (WMMSE-EE-MinRate), and QT-based EE maximization with min-rate constraint (QT-EE-MinRate) benchmarks. Specifically, the sequential multi-agent DDPG model gave at least 68% and 77% higher minimum rate and EE than benchmarks, respectively. Furthermore, the multi-agent DDPG outperforms the single-agent DDPG solution by 20% and more than 28% in terms of minimum rate and EE, respectively.

INDEX TERMS Beamforming, beam steering, energy efficiency, power allocation, rate maximization, reinforcement learning, underwater optical wireless communication.

I. INTRODUCTION

UNDERWATER optical wireless communication (UOWC) is an evolving field that seeks to address communication challenges in underwater environments and opens up new possibilities for underwater exploration, research, and monitoring. As a communication technology, UOWC involves transmitting data through the water using optical signals. Unlike traditional radio-frequency (RF) communication, which is limited in underwater environments due to the high absorption and scattering of electromagnetic waves in water, optical communication employs light to transmit information [1], [2], [3]. As compared to its traditional counterparts, RF and acoustic underwater communications, UOWC boasts several advantages, including higher bandwidth, enhanced security, lower implementation costs, and reduced time latency [1]. However, despite these benefits over RF and acoustic communications, the link range of UOWC systems is constrained by impairments in the UOWC channel - namely, absorption, scattering, and turbulence [2]. The UOWC channel can be effectively modeled by considering attenuation and fading [4]. Said simply, both attenuation and fading contribute to the signal alteration after propagation through the UOWC channel. The attenuation is primarily attributed to absorption and scattering effects [5]. Governed by the Beer-Lambert law, UOWC signal exponentially decays with distance, which makes the link range significantly lower than RF and acoustic link ranges. In addition, optical communications are based on

directional beams with a narrow field of view (FOV), which makes the system highly dependent on spatial geometry and nodes positions. Considering the dynamicity underwater, the alignment problem is further highlighted in UOWC, since nodes are considered quasi-stationary underwater. Based on these these characteristics, the UOWC is strongly dependent on nodes locations and beam orientations, which significantly differentiates the considered problem from RF and acoustic-based resource allocation frameworks.

To alleviate signal scintillation induced by turbulence and extend the UOWC range, various multiple-input–multiple-output (MIMO) systems were proposed [6], [7], [8]. In [6], a non-line of sight (NLOS) UOWC MIMO configuration was proposed to improve the system performance by reducing the path loss and improving the channel impulse response. Furthermore, in [7], the authors investigated the bit error rate (BER) performance of MIMO UOWC systems considering the generalized gamma distribution (GGD) oceanic turbulence, zero-boresight point error, and Elamassie underwater path loss. In [8], the performance of a MIMO underwater vertical wireless optical communication link was evaluated in terms of BER and outage probability, considering the presence of weak and strong oceanic turbulence, pointing errors, and attenuation losses.

The non-orthogonal multiple access (NOMA) technique was incorporated into the UOWC system to enhance spectrum utilization. A promising multi-access technology, NOMA enables efficient large-scale connectivity and perfectly aligns with the developmental requirements of UOWC [9], [10], [11], [12], [13]. In [9], the authors proposed a straightforward two stage program judgement filter (PJF) for a real-time multi-user successsive interference cancellation (SIC)-free NOMA-based uplink UOWC system. The results demonstrated that the proposed framework can decrease the BER compared to the standard SIC-based NOMA. Furthermore, in [10], NOMA was applied in conjunction with deep learning convolutional neural network (CNN) to optimize the relay selection and power allocation tasks in UOWC system. The results revealed that NOMA outperformed the orthogonal multiple access system in the studied case. Practical challenges such as attenuation and turbulence, including misalignment at the relay, were considered in [11], with a particular focus on optimizing the instantaneous time-splitting parameter to maximize the uplink achievable rate. To mitigate decoding errors caused by self-interference within NOMA paired groups, in [12], the authors proposed an alternating composite phase shift (ACPS) coding scheme based on constellation optimization. The power distribution optimization model for the UOWC system, constrained by specific light-emitting diode (LED) emission power, was established in [13] for scenarios ensuring user quality of service (QoS) for edge users and maximizing fairness for all users.

However, although communication fairness must be taken into account in nearly all resource allocation challenges in wireless networks, the fairness concern among underwater nodes was not comprehensively addressed in previous research. The absence of such consideration may result in resource starvation or redundant allocation [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. In [14], the authors highlighted open research challenges, emphasizing interconnectedness of fairness with performance, utility, optimization, and throughput on both network and node levels. The proportional fairness concept was also discussed in [15], where the authors sought to distribute resources among users to ensure a proportional and equitable share based on their individual needs and requirements. This fairness criterion sought to strike a balance between maximizing system throughput and providing a satisfactory level of service to each user. Furthermore, an optimization model was formulated in [16] and [17] to ensure that user nodes meet the fairness criterion, maximizing the minimum rate under the constraint of total transmit power. In [18], the authors addressed the challenge of meeting service requirements for different users within a single beam by categorizing them into the following two types: edge users and fair users, and introduced a power allocation algorithm designed to ensure QoS for edge users and fairness for fair users. In [19], a combined resource scheduler entitled extended weighted fair queuing with latency constraint (EWFQ/LC) was presented for packet scheduling among network slices, considering system fairness and latency constraints. Furthermore, in [20], the authors focused on the uplink of a cell-free massive MIMO system employing maximum-ratio combining (MRC) and zero-forcing (ZF) schemes and addressed a power allocation optimization problem to jointly optimize conflicting metrics, namely sum rate, and fairness. Focusing on energy efficiency (EE) and user fairness in systems with numerous antennas, the authors in [21] proposed a lexicographic-based approach for RIS-assisted mmWave systems that maximizes both EE and user fairness by optimizing power, RIS passive beamforming matrix, and analog precoders. A power management strategy was presented in [22] so as to optimize the reuse, fairness, and capacity of underwater wireless sensor networks (UWSNs). In another pertinent publication [23], the authors investigated the effect of unexpected nodes malfunctions on the performance of imperfect and energy-constrained underwater acoustic sensor networks. To this end, they proposed a semi-cooperative power allocation approach that achieves fairness-effective and robust communication.

However, previous research on fairness in UOWC systems is still very limited, which warrants further investigation. For instance, a power allocation algorithm was proposed in [13] for multiple-beam space division access-based NOMA UOWC systems to maximize max-min fairness. Meanwhile, authors in [24] aimed to maximize fairness while respecting a minimum harvested energy threshold in simultaneous lightwave information and power transfer (SLIPT)-enabled two-user NOMA UOWC systems. However, while the studies briefly reviewed above addressed fairness and EE individually, these two aspects need to be addressed simultaneously, as neglecting the fairness would lead to resource starvation

and neglecting the EE would induce energy waste. To the best of our knowledge, the joint optimization of EE and fairness in the context of UOWC networks remains largely unexplored in the literature.

Reinforcement learning (RL) techniques were also employed to overcome the dynamicity challenge within underwater environment, resulting a secure and reliable UOWC system. In [25], authors addressed the misalignment problem in point-to-point UOWC and focused on optimizing the communication between an underwater sensor and an autonomous surface vehicle that may irregularly shake above sea level. To this end, a two-step two-agent deep reinforcement learning algorithm was proposed. Furthermore, in [26], a deep reinforcement learning (DRL)-based cooperative movement scheme for multiple autonomous underwater vehicle (AUV)s-based UOWC was proposed to overcome the misalignment and positional uncertainty problems typical of underwater environments. In [27], the misalignment and power allocation problems in UOWC were evaluated, where authors jointly optimized the beam divergence and transmission power to maintain a seamless connection in P2P UOWC while minimizing the battery consumption.

While the literature birefly reviewed above demonstrates that RL can be a powerful tool for optimizing UOWC networks, most previous studies still exhibit certain research gaps. In particular, this body of work considers a NOMA-enabled multi-beam, multi-user underwater system. Due to the inherent energy constraints of optical transmitters in underwater environments, it is essential to maximize the system's EE. However, optimizing EE alone does not guarantee rate fairness among distributed nodes, which can degrade the overall quality of service. In this context, the following open challenges emerge:

1) Joint Optimization Challenge: How can beamforming and power allocation be jointly optimized in NOMA-enabled multi-user UOWC systems, while simultaneously maximizing EE and ensuring max-min fairness among users?
2) Scalable Learning-Based Optimization: How can network performance be improved through a scalable, learning-based framework that enables continuous selection of beam orientations and transmit power levels, thereby overcoming the sub-optimality introduced by discrete-action selection methods [11]?

In this work, we address these research gaps by designing a deep deterministic policy gradient(DDPG)-based multi-task DRL framework that jointly maximizes EE and fairness in NOMA-enabled multi-user UOWC networks.

Contributions: In this paper, our primary goal is to optimize the beamforming task, including beam steering and optimizing the power allocation among various receiving nodes, using NOMA in each transmitted beam. This optimization aims to maximize both the network's minimum rate and EE, while taking into consideration the quasi-stationarity of nodes underwater. The main contributions of this work can be summarized as follows:

- We propose a multi-beam UOWC network model, where each beam employs NOMA to simultaneously support multiple underwater user devices. A joint beamforming and power allocation problem is devised to jointly maximize the long-term EE and minimum rate of the network.
- A dynamic RL-based optimization framework is proposed to effectively solve the joint beamforming and power optimization problem. We propose both single-agent and sequential multi-agent DDPG models, with a comparison of their results in terms of the network's minimum rate, EE, and algorithm complexity. The proposed multi-agent model incorporates sequential decision making that enhances beam steering and power allocation tasks. Since this procedure enables training agents tailored to the specific task, these agents have a better decision making capability in dynamic environment. As shown in our simulations, compared to the single-agent approach, the multi-agent approach with sequential decision-making capability achieves better results with a slightly higher execution time.

The remainder of the paper is organized as follows. Section II outlines our system model. In section III, we formulate the problem. Section IV introduces the preliminaries of DDPG algorithm. In section V, we present the proposed DDPG-based beamforming solutions. Section VI details the simulation results. Finally, Section VII concludes. The notations used in this article are summarized in Table 1.

II. SYSTEM MODEL

In this paper, we consider a multi-beam NOMA-aided UOWC where the transmitter is equipped with multi LEDs and can simultaneously forward multi beams into different directions [13], [28], [29]. The underwater receivers are clustered into groups based on their geographical locations, where each group, also referred to as cluster, is covered by one of the transmitted beams (see Figure 1). Due to underwater dynamicity, nodes are considered quasi-stationary, where their locations randomly shift after a certain period of time within a sphere of radius ϵ centered on the estimated location of the node. The total transmitted power cannot exceed a maximum value P_{max}. Our main goal is to maximize the minimum rate and the EE of the network by optimizing the beams orientation and power allocation among the different receiving nodes. NOMA is adopted at each transmitted beam to simultaneously serve different nodes within the same cluster. Let us assume N beams are transmitted simultaneously. The transmitted signal from each beam, n, is defined as shown in (1)

$$x_n = \sum_{m'=1}^{|M_n|} \sqrt{p_{n,m'}}\, s_{n,m'}, \tag{1}$$

where $|M_n|$ is the number of nodes covered by beam n, defined as cluster n, and $p_{n,m'}$ and $s_{n,m'}$ are the power allocated and the signal transmitted to node m' from cluster n, respectively. The nodes included in cluster n are defined

TABLE 1. Table of notations.

Variable	Definition
ϵ	Uncertainty sphere radius
$\nu_{m,n}$	Thermal noise added to $y_{m,n}$
ψ	Concentrator field of view
$\psi_{n,m}$	Incident angle at the receiver
$\theta_{n,m}$	Irradiance angle at the transmitter
θ^{μ}	Online actor network
θ^{Q}	Online critic network
$\theta^{\mu'}$	Target actor network
$\theta^{Q'}$	Target critic network
$\theta_{1/2}$	Transmitter half power angle
τ	Update rate of the target networks
$\gamma_{n,i}(m)$	SINR for i^{th} user when decoding the m^{th} signal
γ	Discount factor
α_a, α_c	Learning rate
σ^2	Noise power
A_{o_n}	Continuous action set taken on orientation
A_{p_m}	Continuous action set taken on m^{th} node
A_r	Node aperture area
a_{o_n}	Orientation action
a_{p_i}	Node power action
c	Water attenuation coefficient
$c(\psi_{n,m})$	Gain of the optical concentrator
$d_{n,m}$	Distance from beam n to node m
d_e	The random moving distance of nodes
EE	Energy efficiency
$h_{n,m}$	Underwater channel attenuation function from beam n to node m
$I_{n,i}(m)$	Sum off inter and intra-interference
$L(\theta^{Q})$	Loss function
M	Total number of nodes
M_n	Number of nodes covered by beam n
N	Number of clusters/beams
N_B	Batch size
N_0	Agent discovery end
$N_{0,o}$	Orientation agent discovery end
$N_{0,p}$	Power agents discovery end
N_{ep}	Number of episodes
n_{water}	Water refraction index
$o_n^{(t)}$	Beam orientation at time t
P_f	Circuit and hardware power consumption
P_{max}	Maximum transmitted power
$p_{n,m}$	Power allocated to node m from cluster n
$p_n^{(t)}$	Power assigned to beam n
$p_{n,m}^{(t)}$	Power allocated to node m from cluster n at time t
Q	Q-function
$R_{n,i}(m)$	i^{th} user achievable rate when decoding m^{th} signal
$R_n(m)$	Achievable rate for m^{th} user from cluster n
R_{min}	Minimum rate for all NOMA users
r_t	Reward at time t
$r_o^{(t)}$	Orientation agent reward at time t
$r_p^{(t)}$	Beam power agent reward at time t
$r_{p,n}^{(t)}$	Node power agent reward at time t
S_i	Continuous state set
$S_{o,i}$	Continuous orientation state set
S_{p_m}	Continuous beam power state set
$S_{p_{n,m}}$	Continuous node power state set
$s_i^{(t)}$	Continuous state value at time t
$s_{n,m}$	Signal transmitted to node m in cluster n
t	Time slot
T	Number of steps per episode
x_n	Transmitted signal by beam n
$(x_i^{(t)}, y_i^{(t)})$	Node coordinates at time t
$(x_{0,n}, y_{0,n})$	Beam source coordinates
y_n	Received signal at node m in cluster n
$\hat{y}_{n,i}(m)$	Recovered m^{th} user signal at node m in cluster n

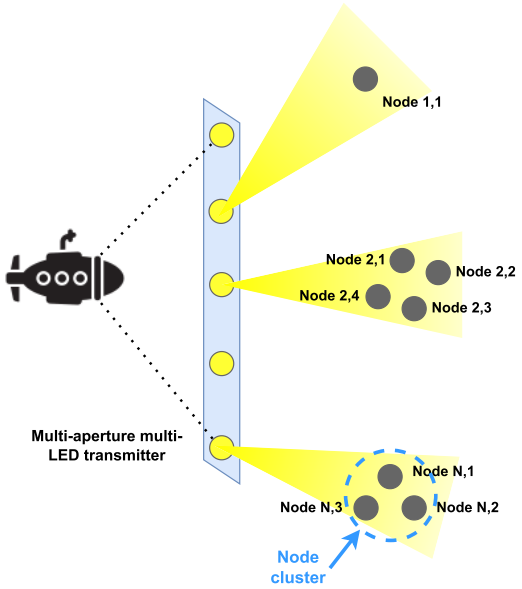


FIGURE 1. System model.

based on their geographic locations. The receiving nodes are initially decomposed into multiple clusters using standard clustering algorOithms such as the K-means clustering.[1] Each cluster is supported by one beam. The received signal at each node m covered by beam n can be expressed as shown in (2)

$$y_{n,m} = h_{n,m}x_n + \sum_{\substack{n'=1 \\ n' \neq n}}^{N} h_{n',m} \sum_{m'=1}^{|M_{n'}|} \sqrt{p_{n',m'}}\, s_{n',m'} + \nu_{m,n}, \quad (2)$$

where $\nu_{m,n} \sim \mathcal{N}(0, \sigma^2)$ is thermal noise and $h_{n,m}$ is underwater channel attenuation function from beam n to node m defined using the Beer-Lambert laww as shown in (3), as shown at the bottom of the next page,where A_r is the

[1]The proposed resource allocation is also valid for other clustering schemes.

node m's aperture areas in m^2. $\theta_{n,m}$ is the irradiance angle at the transmitter, $\psi_{n,m}$ is the incident angle at the receiver. The distance from beam n to node m is denoted by $d_{n,m}$, c is the water attenuation coefficient, and $m = -ln(2)/ln(\cos(\theta_{1/2}))$ is the Lambertian order of the transmitter where $\theta_{1/2}$ is the half power angle of the transmitter [30]. $c(\psi_{n,m})$ is the gain of the optical concentrator defined as specified in (4)

$$c(\psi_{n,m}) = \begin{cases} \frac{n_{water}^2}{\sin^2(\psi_{n,m})}), & 0 \le \psi_{n,m} \le \psi \\ 0, & \psi_{n,m} > \psi, \end{cases} \tag{4}$$

where n_{water} in the water refraction index and ψ is the concentrator field of view (FOV). The channel attenuation function captures unique characteristics of UOWC systems, where the attenuation is an exponential function of the distance, while the incident and irradiance angle directly impact the channel. This contrasts with RF communications, where path loss typically follows a polynomial decay and is less sensitive to nodes alignment. Different NOMA users within the same cluster are sorted in the ascending order of their channel attenuation function, i.e. $h_{n,1} > h_{n,2} > \ldots > h_{n,|M_n|}$. At the reception, each node retrieves its information using the SIC technique. Said differently, each node m from cluster n will first decode the signals of the users from the same cluster with worse channel attenuation functions and will then remove them from the received signal until it detects its own information. Hence, when any node i from cluster n with worse attenuation function than m (i.e. $i < m$) is decoding the m^{th} user signal, the remaining signal $\hat{y}_{n,i}(m)$ can be expressed as shown in (5) [16]

$$\begin{aligned} \hat{y}_{n,i}(m) = & \underbrace{h_{n,m}\sqrt{p_{n,m}}s_{n,m}}_{\text{Decoded signal}} \\ & + \underbrace{h_{n,m}\sum_{m'=1}^{m-1}\sqrt{p_{n,m'}}s_{n,m'}}_{\text{Intra-beam interference}} \\ & + \underbrace{\sum_{\substack{n'=1 \\ n' \neq n}}^{N} h_{n',m}\sum_{m'=1}^{|M_{n'}|}\sqrt{p_{n',m'}}s_{n',m'}}_{\text{Inter-beam interference}} + \nu_{m,i}. \end{aligned} \tag{5}$$

Consequently, the SINR for the i^{th} user when decoding the m^{th} signal can be written as shown in (6)

$$\gamma_{n,i}(m) = \frac{h_{n,m}^2 p_{n,m}}{I_{n,i}(m) + \sigma^2}, \tag{6}$$

where $I_{n,i}(m)$ is the sum of the inter and intra interference effects and can be computed as shown in (7)

$$I_{n,i}(m) = h_{n,m}^2 \sum_{m'=1}^{m-1} p_{n,m'} + \sum_{\substack{n'=1 \\ n' \neq n}}^{N} h_{n',m}^2 \sum_{m'=1}^{|M_{n'}|} p_{n',m'}. \tag{7}$$

The corresponding achievable rate can be deduced using $\gamma_{n,i}(m)$ (see (8))

$$R_{n,i}(m) = \log_2(1 + \gamma_{n,i}(m)). \tag{8}$$

Hence, the corresponding achievable rate is the minimum of these rates (see (9))

$$R_n(m) = \min_{i=1,\ldots,n} R_{n,i}(m). \tag{9}$$

The minimum rate for all NOMA users is defined as shown in (10)

$$R_{min} = \min_{n,m} R_n(m), \tag{10}$$

and the EE is defined as shown in (11) [31]

$$EE = \frac{\sum_{n=1}^{N}\sum_{m=1}^{|M_n|} R_n(m)}{\sum_{n=1}^{N}\sum_{m=1}^{|M_n|} p_{n,m} + P_f}, \tag{11}$$

where P_f is a fixed power including the circuit power consumption and hardware power consumption. We assume that devices accurately perform SIC operations. In addition, we assume that each beam has a fixed width but variable orientation, which is optimized according to the served devices locations. Finally, a time-slotted optimization is considered and we assume that the channel gains remain fixed within a given time slot and independently vary at different time slots. The assumptions taken in this paper are listed as follows.

Assumptions:

- Nodes are quasi-stationary. Their locations randomly change after a certain period of time within a sphere of radius ϵ centered on the estimated location of the node.
- The transmitter can transmit multiple optical beams simultaneously [13], [28], [29].
- The transmission power is limited to P_{max}.
- Nodes are initially clustered using a clustering technique. Each cluster is assumed to be served by one beam. The proposed framework can be applied to any clustering technique.
- Devices accurately perform SIC operations.
- Each beam has a fixed width, but variable orientation. Beams orientations are optimized according to the predefined beam clusters while avoiding beam overlapping.

$$h_{n,m}^2 = \begin{cases} \frac{(m+1)A_r}{2\pi d_{n,m}^2}\cos^m(\theta_{n,m})\cos(\psi_{n,m})c(\psi_{n,m})e^{-cd_{n,m}}, & \theta_{n,m} \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ 0, & otherwise, \end{cases} \tag{3}$$

- A quasi-static oceanic turbulence fading channel is considered where the channel gains remain fixed within a given time slot and vary independently at different time slots.

Importantly, the considered nodes are not mobile. Instead, the node's movement is only due to underwater dynamicity.

III. PROBLEM FORMULATION

In this section, we develop an optimization problem to conduct beamforming in a multi-beam NOMA technique, so as to jointly maximize the min rate and EE. Considering the exponential attenuation and directional nature of UOWC channels, the min-rate and EE largely depend on nodes positions and beam orientations. This further highlights the imbalance between receivers compared to RF systems, making minimum rate constraints particularly critical. In addition, power allocation, albeit essential, cannot fully compensate for poor channel conditions, which justifies the joint consideration of spatial and resource optimization in the proposed model. Hence, the beamforming task can be divided into the following two main sub-tasks: (1) The power allocation among different nodes and (2) the beams orientation optimization. For a multi-user network, maximizing the max-min fairness of resource allocation does not lead to optimal EE. Conversely, only maximizing EE can lead to unfair resource allocation among users. To strike a suitable balance, our ain in this study is to jointly maximize the system EE and max-min fairness of a multi-user UOWC network. To this end, we divide the overall duration into T non-overlapping time slots defined by t. Let $EE^{(t)}$ and $R_{min}^{(t)}$ be the EE and network min rate at the t-th time slot, respectively. The optimization problem can be written as shown in (12)–(15)

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \max_{p_{n,m}^{(t)}, o_n^{(t)}} \left\{ R_{min}^{(t)}, EE^{(t)} \right\} \tag{12}$$

$$\text{s.t} \quad (C1): \sum_{n,m} p_{n,m}^{(t)} \leq P_{max}, \forall t \tag{13}$$

$$(C2): p_{n,m}^{(t)} > 0 \quad \forall n, m, t \tag{14}$$

$$(C3): -\tfrac{\pi}{2} + (n-1)\tfrac{\pi}{N} \leq o_n^{(t)} \leq -\frac{\pi}{2} + n\frac{\pi}{N} \quad \forall n, t, \tag{15}$$

where constraint (C1) implies that the sum of all node powers at time t, $p_{n,m}^{(t)}$, must be below the maximum power P_{max}; constraint (C2) implies that each node's power must be greater than zero; and constraint (C3) defines the possible orientation space for each beam orientation at time t, $o_n^{(t)}$, $n \in [1, N]$. To avoid beam overlapping and interference, it divides the half space $\left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$ into N equal sub-spaces. The considered problem is non-convex, as it aims to simultaneously maximize two fractions. Also, reducing the maximization problem to only min-max rate maximization over interference channel (i.e., by removing the EE from the objective function) makes it a classic NP-hard problem [32]. Hence, the considered problem is provably NP-hard as well. Solving the considered optimization problem using conventional techniques brings up several challenges that can be listed as follows,

- Conventional optimization methods may suffer from increased complexity due to the intricate algorithms and decision-making processes involved. The complexity of these approaches can lead to longer computation times and higher resource use, making them less efficient, particularly in real-time or resource-constrained scenarios.
- The computational resources needed to perform the allocation in conventional methods can be significant.
- Traditional approaches frequently rely on precise knowledge of the users' instantaneous SINRs/CSI to make allocation decisions. This requirement for accurate and up-to-date information can be challenging, especially in dynamic and unpredictable environments, where nodes are considered quasi-stationary and their locations, and hence the CSIs, is changing constantly in a random manner.

Therefore, considering the aforementioned challenges, RL is a good candidate to solve this problem in a dynamic manner.

Remark 1: RL algorithms can learn the optimal policy directly from their own experiences. While traditional approaches heavily rely on predefined rules or heuristics and traditional machine learning techniques need ground truth, which is costly to obtain for large-scale systems, and lack adaptability in dynamic environment, RL methods can adapt and improve their performance over time through continuous interaction with the environment. This ability to learn from experience allows RL algorithms to refine their strategies and achieve higher levels of performance. Moreover, RL techniques are well-suited for dynamic environments, where the conditions or constraints may unpredictably change.

IV. DDPG PRELIMINARIES

DDPG is a a model-free, online, off-policy DRL algorithm that specifically deals with continuous states and action sets. The main elements of a DDPG and RL algorithm in general are outlined below:

- States S: It is a quantified definition of the environment encountered by the agent at certain time t. The environment represents anything that the agent can, directly or indirectly, interact with.
- Actions A: It is the decision taken by the agent considering the environment at time t.
- Policy $\pi(.)$: It is a state-action mapping. The RL agent aims to converge to the optimal policy that maximizes the cumulative reward.
- Reward: Maximizing the long-term reward function is the main objective of an RL problem. It is a quantitative feedback calculated following the action execution, as the environment moves from current state s_t to next state s_{t+1}.

At each time t, the agent at state $s_t \in S$ takes an action $a_t \in A$ following the policy $\pi(a_t/s_t)$. As a result, it receives a reward r_{t+1} and the state is updated to s_{t+1}.

The DDPG agent adopts an actor-critic approach to converge to an optimal policy π that maximizes the expected

cumulative long-term reward $R_t = \sum_{i=t+1}^{\infty} \gamma r_{t+i+1}$, where $\gamma \in (0, 1]$ is the discount factor. The DDPG agent includes the following four main networks [33]:

- An online actor network parameterized by θ^{μ}: This network takes as input the states and outputs the actions.
- An online critic network parameterized by θ^{Q}: This network takes as input the states and actions decided by the online actor network, and outputs the Q-function Q(s,a) that evaluates the given state-action tuple.
- A target actor network parameterized by $\theta^{\mu'}$: This network is a previous copy of the online actor network used for tracking the learned network to ensure stability.
- An target critic network parameterized by $\theta^{Q'}$: It is a previous copy of the online critic network.

The DDPG algorithm starts by an discovery phase where the agent discovers the environment by taking random actions and stores the corresponding resulting states transition and rewards in its memory. Afterwards, the algorithm learns the action using the actor network and adding certain exploration noise $\mathcal{N}_t$ to the action to cover more values of the continues action set. At each time t, the critic network is updated by minimizing the loss function, defined as shown in (16)

$$L(\theta^{Q}) = \frac{1}{N_B} \sum_{t} \left[Q(s_t, a_t|\theta^{Q}) - y_t \right]^2, \tag{16}$$

where N_B is the size of the batch sampled from the memory and y_t is calculated using the Bellman equation (see (17))

$$y_t = r_t + \gamma Q' \left(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'} \right). \tag{17}$$

The actor network is updated using the sampled policy gradient (see (18))

$$\nabla_{\theta^{\mu}} \approx \frac{1}{N_B} \sum_{t} \left[\nabla_a Q(s, a|\theta^{Q})|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^{\mu}} \mu(s|\theta^{\mu})|_{s=s_t} \right]. \tag{18}$$

Finally, the target networks are updated using (19a)–(19b)

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}, \tag{19a}$$
$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}, \tag{19b}$$

where τ is the update rate of the target networks. The following theorem establishes the theoretical convergence of the DDPG RL algorithm.

Theorem 1 ([34]): Let $\mathcal{T}Q(s_t, a_t) = r_t + \gamma \mathbb{E} Q(s_{t+1}, a_{t+1})$ be the Bellmann optimality operator and $\mathcal{F}$ be the class of functions that the critic network can create. The difference between the estimated Q-value and $\mathcal{T}Q$ is defined as the one-step approximation error. Considering n independent experiences (s_t, a_t, s_{t+1}, r_t) and defining $\hat{Q}$ as the solution that minimizes the loss in (16), for any $0 \le \epsilon \le 1$ and $\delta \ge 0$, the one step approximation error is upper-bounded by (20)

$$\left\| \hat{Q} - \mathcal{T}Q \right\|^2 \le (1+\epsilon) \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \left(\|f - \mathcal{T}g\|^2 \right) + C \cdot \frac{V_{\max}^2}{n\epsilon} \cdot \ln \mathcal{N}(\mathcal{F}, \delta) + C' \cdot V_{\max} \cdot \delta \tag{20}$$

where C and C' are two constants, $V_{max} = R_{max}/(1-\lambda)$ is the maximum value in the underlying Markov decision process (MDP), and $\mathcal{N}(\mathcal{F}, \delta)$ is the δ-covering number associated with the function class $\mathcal{F}$, defined as the minimum number of euclidean balls of radius δ required to cover the space spanned by $\mathcal{F}$ applied on a fixed set of inputs.

Theorem 1 confirms that, under the assumptions of an MDP and the availability of i.i.d. experiences, the convergence of DDPG can be guaranteed within a bound. Building on this theorem, in Section V we introduce single-agent and multi-agent DDPG reinforcement learning algorithms, and then validate their convergence through numerical simulations presented in Figures 3–4 of Section VI.

V. DDPG-BASED BEAMFORMING METHOD

A. SINGLE-AGENT DRL APPROACH

MDP Formulation: The considered problem can be converted into an MDP problem. The actions, states, and reward are defined as follows:

- Actions: We define actions set $A = \{A_{o_1}, ..A_{o_N}, A_{p_1}, .., A_{p_M}\}$, where $A_{o,n}$ is the continuous action set taken on the orientation of the n^{th} beam, $i \in [1, N]$, and A_{p_m} is the continuous action set taken on the power of the m^{th} node, $m \in [1, M]$, where $M = \sum_n |M_n|$ is the total number of nodes in the network. Each action takes a value between 0 and 1. The beams orientations at each time t are extracted using the first N actions using (21)

$$o_n^{(t)} = \frac{\pi}{N} a_{o_n}^{(t)} - \frac{\pi}{2} + (n-1)\frac{\pi}{N}, n \in [1, N]. \tag{21}$$

  This divides the half space $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ into N equal subspaces, so assuming that beams cannot overlap to avoid interference, each beam orientation is learned in its corresponding sub-space. The power allocation is learned using the M following actions using (22), where $n \in [1, N], m \in [1, M_n]$

$$p_{n,m}^{(t)} = \frac{a_{p_i}^{(t)} P_{max}}{M}, i = \sum_{1}^{n-1} |M_n| + m. \tag{22}$$

- States: We define states set $S = \{S_1, .., S_M\}$, where each state is a continuous value $s_i^{(t)} \in S_i$ referring to the slope of the line between node i and its corresponding beam origin defined as shown in (23)

$$s_i^{(t)} = \frac{y_i^{(t)} - y_{0,n}}{x_i^{(t)} - x_{0,n}}, \tag{23}$$

  where $(x_i^{(t)}, y_i^{(t)})$ and $(y_{0,n}, x_{0,n})$ are the corresponding coordinates of node i at time t and its corresponding beam sources, respectively.
- Reward: The reward function is defined as the minimum rate in (10) for selected actions. Such a reward function is selected after exhaustively trying all possible options of the reward design, i.e., considering the minimum rate, network EE, and the weighted sum of both functions. The minimum rate provides the most suitable outcome.

Algorithm 1 Single-agent DDPG beamforming algorithm

Input: S, A, N_{ep}, T, N_0, γ.
Initialize: Q, μ, Q', μ', B.
for $e = 0, 1, 2 \ldots N_{ep}$ **do**
  Reset $s^{(1)} \in S$
  for $t \in [1, T]$ **do**
    if $t \leq N_0$ **then**
      Randomly choose $a_t = \{a_{o_1}^{(t)}, \ldots, a_{o_N}^{(t)}, a_{p_1}^{(t)}, \ldots, a_{p_M}^{(t)}\}$
    else
      Get $a^{(t)} = \mu(s^{(t)}|\theta^{\mu}) + \mathcal{N}^{(t)}$ according to the current policy
and exploration noise
    end if
    Calculate $o_n^{(t)}, n \in [1, N]$ using (21)
    Calculate $p_{n,m}^{(t)}, n \in [1, N], m \in [1, M_n]$ using (22)
    Calculate reward $r^{(t)}$ using (10)
    Observe $s^{(t+1)}$
    Store $(s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)})$ in experience memory B
    Sample random batch of size N_B from B
    Update Q, μ, Q', μ'
    if $mod(t, 10) = 0$ **then**
      Move randomly the nodes by a distance $d_\epsilon \leq \epsilon$ from their estimated location
    end if
  end for
end for

DRL Algorithm Development: Since the defined states and actions sets are continuous, in this study, we propose a DDPG-based solution to solve the considered optimization problem. The agent defined on the multi-beam transmitter level starts by initializing the actor critic networks parameters. For each episode, the states are reset to their initial values. Then, for the first N_0 time slots in the episode, the agent takes at each time slot t random actions, and after N_0 time slots, it takes actions using the actor network and adds exploration noise $\mathcal{N}_t$ to the chosen action. Then, the agent calculates the new orientation $o_n^{(t)}$ for each beam n and power $p_{n,m}^{(t)}$ for each node m in each cluster n using (21) and (22), respectively. It also calculates reward r_t using (10) and observes next state s_{t+1}. The resulting tuple (s_t, a_t, r_t, s_{t+1}) is stored in experience memory B and a random batch of size N_B is sampled from B. Finally, the agent updates the actor and critic networks. We assume that the nodes randomly move by a distance $d_\epsilon \leq \epsilon$ from their estimated location each 10 time slots. The proposed single-agent DDPG approach is summarized in Algorithm 1. The agent tests its training each N_{test} episodes. For the testing, an algorithm similar to the training algorithm is used. Except that during the test, the agent runs only one episode and always takes actions using the actor network and without adding and exploration noise. In addition, the resulting tuples in each time slot are not stored in memory, and the actor and critic networks are not updated. Theorem 1 confirms that, under the assumptions of an MDP and the availability of i.i.d. experiences, the convergence of DDPG can be guaranteed within a bound. Building on this theorem, we introduce single-agent and multi-agent DDPG reinforcement learning algorithms in Section V, and validate their convergence through numerical simulations presented in Figure 3 and Figure 4 of Section VI.

Complexity of the single-agent DDPG algorithm is $\mathcal{O}(T.N_B.(C_{actor} + 2C_{critic}))$, where C_{actor} and C_{critic} are the cost of forward/backward pass through the actor and critic networks, respectively [35]. The cost of forward/backward pass largely depends on the architecture of the network and can be defined using the number of layers in each network and the number of input and output of each layer (see Eq. (24)).

$$C = \sum_{h=1}^{H} (n_{in}.n_{out}) + n_{out} \tag{24}$$

where H is the number of hidden layers in the network, n_{in} is the number of input units in the layer, and n_{out} is the number of output units. Consequently, computational complexity of the proposed single layer DDPG algorithm is $\mathcal{O}(T.N_B.(M(6H+1) + N(3H+1) + 3H^2 + 8H + 2))$.

This single-agent model outputs $N + M$ actions. Hence, by increasing the number of nodes or beams within the system, the actions size considerably increases, thereby slowing down the convergence of the system. To overcome this limitation, we develop a multi-agent RL approach by distributing the decision process among different agents that cooperate to reach the optimal beam orientation and power allocation in the network. Further detail on the multi-agent DRL approach is provided below.

B. MULTI-AGENT DRL APPROACH

In wireless network optimization, complex problems can be tackled effectively by breaking them down into smaller, manageable sub-problems and solving each one optimally. Inspired by this strategy, we propose a sequential multi-agent reinforcement learning approach that ensures a distributed beamforming task instead of a central one. The main task is divided into the following three main tasks: (1) optimizing beam orientation; (2) power allocation among beams; and (3) power allocation among nodes within each cluster. The different agents executing each of these sub-tasks are defined as follows:

- Beam orientation agent: This agent has M states defined as the slope values of the lines between the node i and its corresponding beam, as explained in the previous section, $S_o = \{S_{o_1}, .., S_{o_M}\}$. This variable offers an information of the elevation of node from the beam axis, which is sufficient to decide the optimal beam orientation to cover the corresponding nodes. It outputs N actions $A_o = \{A_{o_1}, ..A_{o_N}\}$. The beams orientation are calculated using (21). The reward, $r_o^{(t)}$, is defined to be the EE of the network, since the beam orientation affects the rates of all nodes, not only the minimum rate. Hence, focusing on the sum rate included in the EE equation

would give a more balanced result than focusing on the min rate only.

- Beam power allocation: This agent has M states, corresponding to the channel information between each node and its corresponding beam, $S_p = \{h^2_{n,m}, n \in [1, N], m \in [1, M_n]\} = \{S_{p_1}, .., S_{p_M}\}$. It has N actions corresponding to the power allocated to each beam $A_p = \{p_n, n \in [1, N]\} = \{A_{p_1}, .., A_{p_N}\}$. The reward function, $r_p^{(t)}$, is the minimum rate of the network. This is so because maximizing the minimum rate guarantees that all node rates are above that minimum rate. By contrast, the EE function tends to maximize the total system capacity, which can be achieved by favoring nodes with good propagation conditions—frequently at the expense of others with weaker channels. This inherently creates an imbalance among the nodes' achievable rates and may significantly degrade fairness by deteriorating the minimum rate to very low values. Hence, maximizing the minimum rate enforces the beam power agent and node power allocation agent to learn a policy that would increase data rate of all the nodes within a cluster while ensuring fairness among them.
- Node power allocation: There are N node power allocation agents, one per cluster. Each of these agents has M_n states, $n \in N$, defined as the channel information between the considered beam and its corresponding nodes. $S_{p_n} = \{h^2_{n,m}, m \in [1, M_n]\} = \{S_{p_{n,1}}, .., S_{p_{n,M_n}}\}$. It outputs M_n actions, $A_{p_n} = \{p_{n,m}, m \in [1, M_n]\} = \{A_{p_{n,1}}, .., A_{p_{n,M_n}}\}$. Following the same logic behind the choice of the beam power allocation agent's reward function, the reward function, $r_{p,n}^{(t)}$, for each agent n is the minimum rate of its corresponding cluster.

Unlike the single-agent model, this model alleviates the action size for each node, where the number of actions taken by the first and second agents is N, and the number of actions taken by the last agents equals the number of nodes in each agent. To strike a suitable balance between system EE and minimum rate, a multi-level reward function, described as follows, is selected for these agents. The orientation agent's reward is set to be the overall EE of the network, the beam power agent's reward is defined as the overall min rate of the network, while the reward of each node's power agent is the min rate of each corresponding cluster.

The proposed multi-agent DDPG-RL framework executes the following tasks sequentially.

1) The orientation agent decides the beam orientations based on the nodes elevation from the beam axis using (21).
2) Considering the channel gain based on the chosen orientation at time t, the beam power agent decides the power allocated for each beam using the (25)

$$p_n^{(t)} = \frac{a_{p_n}^{(t)} P_{max}}{N}, n \in [1, N], \tag{25}$$

where $p_n^{(t)}$ is the power assigned for beam n and $a_{p_n}^{(t)}$ is the n^{th} action taken by the second agent.

3) Once the power allocated for each beam is decided by the beam power agent at time t, each node power agent decides on the power allocated for each node within its corresponding cluster, using (26)

$$p_{n,m}^{(t)} = \frac{a_{p_{n,m}}^{(t)} p_n^{(t)}}{M_n}, n \in [1, N], m \in [1, M_m], \tag{26}$$

where $a_{p_{n,m}}^{(t)}$ is the m^{th} action taken by agent n.
4) The agents receive their rewards for their selected actions.
5) The agents update their actor and critical networks.

The structure of the proposed multi-agent model is illustrated in Figure 2. To further improve the results, we propose a sequential DDPG approach where we introduce a delay between the discovery phase of the beam orientation agent and the discovery phase of the remaining agents. In fact, the discovery phase of the power agents starts simultaneously with the learning of the orientation agent. The reason behind choosing the sequential model is that the nodes are quasi-stationary. Hence, once converged, beam orientations do not considerably change. Consequently, the sequential model focuses first on learning the beam orientation, then learns the optimal power distribution using a close to optimal orientation-based data. Since a given agent focuses on learning a particular action optimally instead of learning all the possible actions, the sequential approach shows improved capability to learn optimal solution to intricate optimization problem (12) in the complex and dynamic UOWC environment. The proposed sequential multi-agent DDPG approach is summarized in Algorithm 2. Complexity of the proposed multi-agent DDPG algorithm is defined as the sum of complexity values of all networks, hence it is $\mathcal{O}(T.N_B.(M(12H + 1) + N(3H^2 + 12H + 4) + 6H^2 + 18H + 4))$. Compared to the single-agent DDPG algorithm, the sequential multi-agent DDPG algorithm requires a higher computational complexity. This will be further demonstrated by simulations in the following section.

VI. RESULTS AND DISCUSSION

A. SIMULATION SETUP

In this section, we evaluate the performance of the proposed algorithms. The simulations were performed using Python 3.7. We consider a 25×25 m^2 square area where nodes are randomly located in each cluster. We consider three clusters with one beam allocated to each cluster, i.e. $N = 3$. The transmitted beams are separated by a distance of 10 cm vertically. We consider pure sea water with extinction coefficient $c = 0.043$ for $\lambda = 514$ nm. The node's aperture area is set to $A_r = 19.6$ mm^2 and the noise power $\sigma^2 = 5 \times 10^{-12}$. The half power angle is $\theta_{1/2} = 7.5°$ and the concentrator field of view (FOV) is $\psi = 70°$. The water refractive index is $n_{water} = 1.33$ [30], [36]. To replicate node's position uncertaintly due to underwater turbulence, we assume that each 10 steps, the nodes move from their estimated position by $\epsilon = 0.25$ m, defined as the uncertainty radius. The maximum power

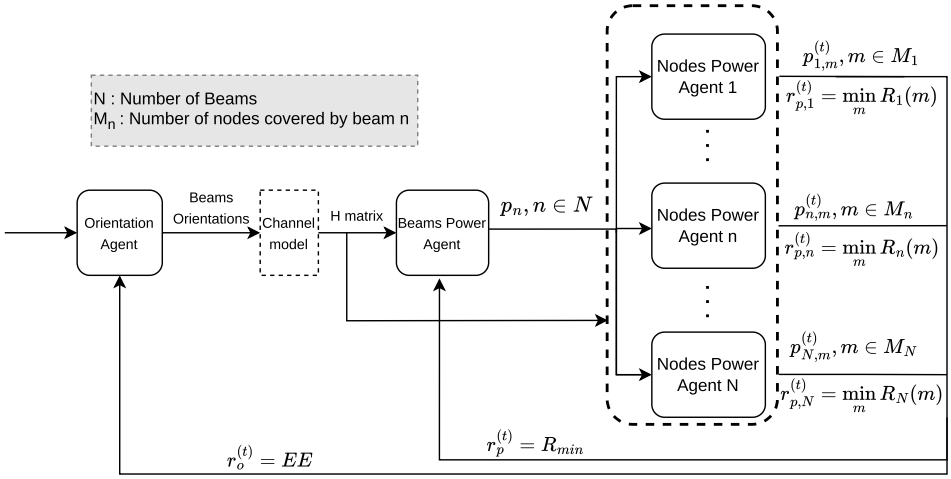


FIGURE 2. Structure of the proposed multi-agent DDPG model.

TABLE 2. System parameters.

Parameter	Value
Simulation area	$25 \times 25\ m^2$
Beam centers	$\{(0,-0.1);(0,0);(0,0.1)\}$
Number of beams	N=3
Extinction coefficient	$c = 0.043$
Wavelength	$\lambda = 514$ nm
Nodes aperture area	$A_r = 19.6\ mm^2$
Noise power	$\sigma^2 = 5 \times 10^{-12}$
Half power angle	$\theta_{1/2} = 7.5°$
FOV	$\psi = 70°$
Water refractive index	$n_{water} = 1.33$
Uncertainty radius	$\epsilon = 0.25$ m
Power budget	$P_{max} = 10$ W

TABLE 3. DDPG algorithm hyperparameters.

Parameter	Value
Experience memory capacity	8000
Batch size	$N_B = 128$
Learning rate	$\alpha_a = 10^{-4},\ \alpha_c = 10^{-3}$
Discount factor	$\gamma = 0.99$
Optimizer	Adam optimizer
Update rate	$\tau = 0.01$
Number of episodes	$N_{ep} = 80$
Number of steps per episode	$T = 50$
Orientation agent discovery end	$N_{0,o} = 10$
Power agents discovery end	$N_{0,p} = 30$

budget $P_{max} = 10$ W. The system parameters are summarized in Table 2.

The hyperparameters of the proposed DDPG algorithms are defined as follows. Each DDPG agent has four networks: online and target actor networks, denoted as μ and μ', respectively, and online and target critic networks, denoted as Q and Q', respectively. These networks have one input layer, one hidden layer, and one output layer each. The hidden layer has 100 neurons. The activation function for the input and hidden layer is the RELU function. For the actor network's output layer, we choose the tanh function to obtain an action in a defined set, i.e. between $[-1, 1]$. Conversely, for the critic network's output layer, we select the linear activation function. The experience memory capacity is set to 8000 with a batch size of $N_B = 128$. The actor networks learning rate is $\alpha_a = 10^{-4}$, while the critic networks learning rate is $\alpha_c = 10^{-3}$ and the discount factor is $\gamma = 0.99$. We use the Adam optimizer to update the target networks, and the update rate is set to $\tau = 0.01$. The number of episode is set to $N_{ep} = 80$ episodes, with each episode including $T = 50$ steps. The discovery phase of the orientation agent takes the first $N_{0,o} = 10$ episodes, while the discovery phase of the power agents takes the following 20 episodes and ends at $N_{0,p} = 30^{th}$ episode. The DDPG-related parameters are summarized in Table 3.

B. PERFORMANCE ASSESSMENT OF THE PROPOSED MODELS

We first evaluate the performance of the proposed single-agent and sequential multi-agent models. We also evaluate the effect of adding the sequential feature to the multi-agent DDPG solution by evaluating the performance of non-sequential multi-agent DDPG approach. Furthermore, we compare the results with the multi-agent DDPG approach where the reward for all agents is either set to R_{min} or to EE. We consider the case of three beams where the first and third clusters include four nodes, while the second cluster includes three nodes, i.e. $M_1 = M_3 = 4$ and $M_2 = 3$. The results of the network's minimum rate and EE for these five schemes are shown in Figure 3 and Figure 4, respectively. These figures further justify the convergence of the proposed algorithms, proven theoretically above. The network's minimum rate for the single-agent DDPG beamforming model converges to around 1.1 bits/s/Hz after

Algorithm 2 Sequential multi-agent DDPG beamforming algorithm

Input S_o, S_p, $S_{p_n}, n \in N$, A_o, A_p, $A_{p_n}, n \in N$, N_{ep}, $N_{0,o}$, $N_{0,p}$, γ.
Initialize:
- Q_o,μ_o, Q'_o, μ'_o, B_o
- Q_p, μ_p, Q'_p, μ'_p, B_p
- $Q_{p,n}$, $\mu_{p,n}$, $Q'_{p,n}$, $\mu'_{p,n}$, $B_{p,n}$, $n \in [1, N]$.

for $e = 0, 1, 2 \ldots N_{ep}$ **do**
  Reset $s_{o,1}, \in S_o$, $s_{p,1} \in S_p$, and $s_{p_n,1} \in S_{p_n}, n \in [1, N]$
  for $t \in [1, T]$ **do**
    if $e \leq N_{0,o}$ **then**
      Randomly choose $a_o^{(t)} = \{a_{o_1}^{(t)}, \ldots, a_{o_N}^{(t)}\}$
    else
      Get $a_o^{(t)} = \mu_o(s_o^{(t)}|\theta_o^{\mu_o}) + \mathcal{N}_t$ according to the current policy and exploration noise
    end if
    Calculate $o_n^{(t)}, n \in [1, N]$ using (21)
    Update $s_p^{(t)}$ according to $o_n^{(t)}$
    if $e \leq N_{0,p}$ **then**
      Randomly choose $a_p^{(t)} = \{a_{p_1}^{(t)}, \ldots, a_{p_N}^{(t)}\}$
    else
      Get $a_p^{(t)} = \mu_p(s_p^{(t)}|\theta_p^{\mu_p}) + \mathcal{N}_t$ according to the current policy and exploration noise
    end if
    Calculate $p_n^{(t)}, n \in [1, N]$ using (25)
    Update $s_{p_n}^{(t)}$ according to $p_n^{(t)}$
    if $e \leq N_{0,p}$ **then**
      Randomly choose $a_{p_n}^{(t)} = \{a_{p_{n,1}}^{(t)}, \ldots, a_{p_{n,N}}^{(t)}\}$
    else
      Get $a_{p_n}^{(t)} = \mu_{p_n}(s_{p_n}^{(t)}|\theta_{p_n}^{\mu_{p_n}}) + \mathcal{N}_t$ according to the current policy and exploration noise
    end if
    Calculate $p_{n,m}^{(t)}, n \in [1, N], m \in M_n$ using (26)
    Calculate $r_o^{(t)}$, $r_p^{(t)}$, and $r_{p,n}^{(t)}$ using (11) and (10)
    Observe $s_o^{(t+1)}$, $s_p^{(t+1)}$, $s_{p,1}^{(t+1)}$, .., $s_{p,N}^{(t+1)}$
    Store $(s_o^{(t)}, a_o^{(t)}, r_o^{(t)}, s_o^{(t+1)})$ in B_o
    Sample random batch of size N_B from B_o
    Update Q_o, μ_o, Q'_o, μ'_o
    if $e \geq N_{0,o}$ **then**
      Store $(s_p^{(t)}, a_p^{(t)}, r_p^{(t)}, s_p^{(t+1)})$ in B_p
      Sample random batch of size N_B from B_p
      Update Q_p, μ_p, Q'_p, μ'_p
      Store $(s_{p,n}^{(t)}, a_{p,n}^{(t)}, r_{p,n}^{(t)}, s_{p,n}^{(t+1)})$ in $B_{p,n}, n \in [1, N]$
      Sample random batch of size N_B from $B_{p,n}, n \in [1, N]$
      Update $Q_{p,n}, \mu_{p,n}, Q'_{p,n}, \mu'_{p,n}, n \in [1, N]$
    end if
    if $mod(t, 10) = 0$ **then**
      Move the nodes randomly by a distance $d_\epsilon \leq \epsilon$ from their estimated location.
    end if
  end for
end for

50 episodes, while the network's minimum rate for the proposed sequential multi-agent beamforming model converges to around 1.5 bits/s/Hz after 30 episodes. Furthermore, the EE for the single-agent DDPG beamforming model reaches 15 bits/J/Hz, while the EE for the sequential multi-agent DDPG approach reaches 40 bits/J/Hz. This supports our claim that, for large number of nodes in the network, the single-agent's action size becomes considerably large, thereby affecting the performance of such networks. By contrast, adopting the multi-agent model where the decision is distributed among different agents alleviates the task and ensures better and faster convergence. Furthermore, introducing the sequential feature to the multi-agent DDPG approach ensures better minimum rate and EE and faster convergence. For instance, we find that, as compared to the non-sequential approach, the sequential multi-agent DDPG approach increases the network's minimum rate and EE by 0.1 bits/s/Hz and 10 bits/J/Hz, respectively, and converges 5 episodes earlier. Furthermore, the proposed sequential multi-agent approach gives better and faster convergence than the sequential multi-agent approach with all rewards set to R_{min}. Finally, the sequential multi-agent approach with all rewards set to EE maximized the EE compared to all the other approaches at the cost of a very poor network's minimum rate around 0.1 bits/s/Hz. This is well aligned with the analytical explanation of the choice of the reward functions presented above. Based on the evidence, we conclude that the proposed sequential multi-agent DDPG beamforming method effectively balances the network's minimum rate and EE trade-off.

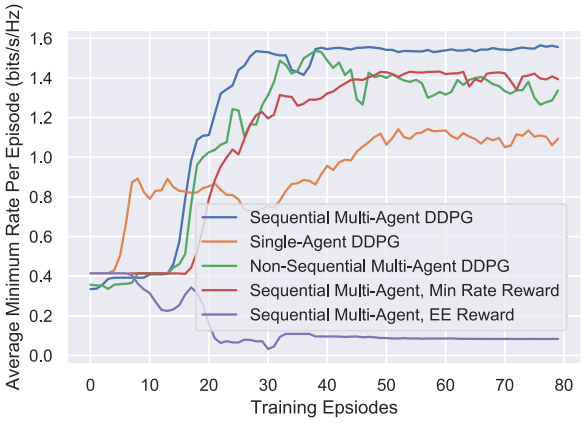


FIGURE 3. Average minimum rate per episode for $N = 3$, $M_1 = M_3 = 4$ and $M_2 = 3$, $T = 50$, $\epsilon = 0.25$, $N_0 = 30$, $N_{0,o} = 10$, and $N_{0,p} = 30$.

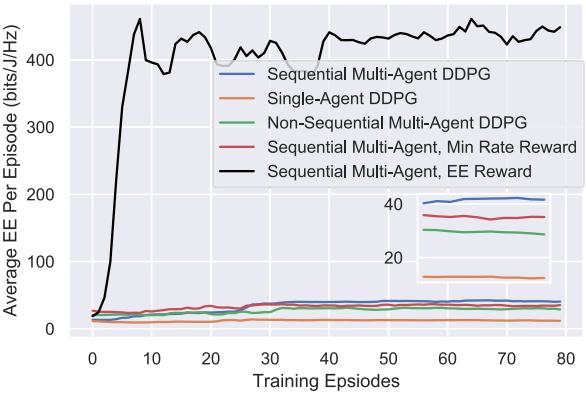


FIGURE 4. Average EE per episode for $N = 3$, $M_1 = M_3 = 4$ and $M_2 = 3$, $T = 50$, $\epsilon = 0.25$, $N_0 = 30$, $N_{0,o} = 10$, and $N_{0,p} = 30$.

C. PERFORMANCE COMPARISON WITH BENCHMARKS

Next, we evaluate the performance of the proposed single-agent and sequential multi-agent DDPG beamforming

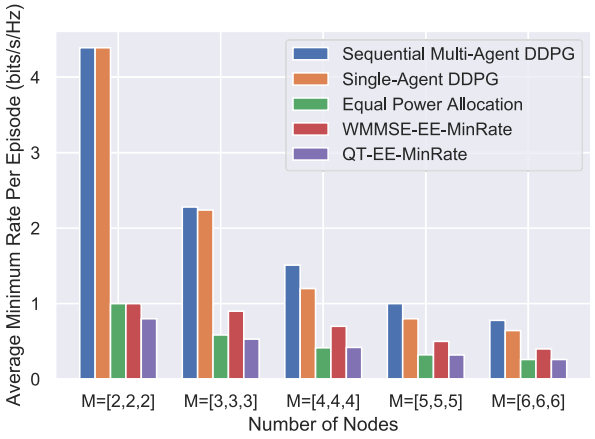


FIGURE 5. Average minimum rate per episode for different M_n values, $N = 3$, $T = 50$, $\epsilon = 0.25$, $N_0 = 30$, $N_{0,o} = 10$, and $N_{0,p} = 30$.

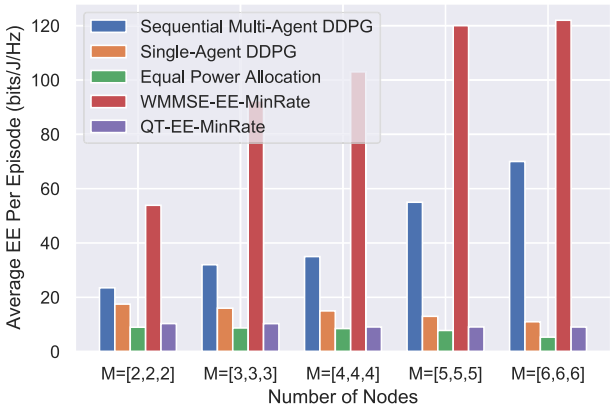


FIGURE 6. Average EE per episode for different M_n values, $N = 3$, $T = 50$, $\epsilon = 0.25$, $N_0 = 30$, $N_{0,o} = 10$, and $N_{0,p} = 30$.

methods as a function of the number of nodes per cluster and water condition. For comparison, we use the following benchmarks.

- Weighted minimum mean-square error (WMMSE)-based EE maximization with min-rate constraint (WMMSE-EE-MinRate): EE is defined as a fraction where the nominator is the sum of rates. Consequently, power allocation task is performed using fractional programming with Dinkelbach's method [37], combined with the WMMSE framework for handling the non-convex rate expressions [38]. The beam orientations are defined according to the estimated locations of nodes.
- Quadratic Transform (QT)-based EE maximization with min-rate constraint (QT-EE-MinRate): Similarly to the previous benchmark, this one uses fractional programming with Dinkelbach's method to solve the EE maximization problem, which is followed by using QT instead of WMMSE to solve the non-convex sum rate expression [39], [40]. The beam orientations are defined according to the estimated locations of nodes.
- Equal power allocation: The beam orientation task is performed using a DDPG agent as defined in the proposed multi-agent DDPG approach, while the nodes are assigned equal power, i.e. $P_{n,m} = P_{max}/M$.

The network's average minimum rate per episode is a decreasing function of the number of nodes per cluster (see Figure 5). This is because increasing the number of nodes increases the interference in the network, thereby degrading the minimum rate. For up to three nodes per beam, the proposed single and multi-agent DDPG approaches have the same performance in terms of the network's minimum rate. As the number of actions for the single-agent model is a function of the number of nodes, keeping a small number of nodes per cluster ensures a small action set size. Consequently, the performance of the two proposed approaches is similar. However, starting from four nodes per cluster, the proposed sequential multi-agent approach increases the network's minimum rate by around 0.2 bits/s/Hz as compared to the single-agent approach. At five nodes per cluster, the proposed sequential multi-agent approach achieves the best minimum rate in the range of 1 bits/s/Hz, as compared to the other approaches. Based on this evidence, we conclude that not exceeding five nodes per cluster in the considered system will ensure tolerable rates for all nodes in the network. Regarding the considered benchmarks, the resulting average minimum rate per episode is considerably lower than that of the proposed models, reaching 0.3 bits/s/Hz for the equal power allocation and QT-EE-MinRate, and 0.4 bits/s/Hz for WMMSE-EE-MinRate. This is so because the equal power allocation benchmark maximizes the power for all nodes, thus maximizing the interference within the network, and the other two benchmarks aim to maximize the network's sum rate at the cost of the minimum rate.

The EE for different numbers of nodes is illustrated in Figure 6. The proposed sequential multi-agent DDPG approach has increasing EE values by increasing the number of nodes per cluster, exceeding 50 bits/J/Hz for the case of five nodes per cluster. Conversely, the proposed single-agent approach gives decreasing EE values by increasing the number of nodes per cluster, reaching 15 bits/J/Hz at five nodes per cluster. This is due to the fact that, in the single-agent decision, only the minimum rate is considered, i.e. without taking care of the EE results; by contrast, the multi-agent approach optimizes the trade-off between network's minimum rate and EE, and achieves a balanced results for both metrics. As for the equal power allocation and QT-EE-MinRate benchmarks, the EE values are almost constant for all number of nodes per cluster, ranging about 8 bits/J/Hz for the equal power allocation and around 10 bits/J/Hz for the QT-EE-MinRate. Concerning WMMSE-EE-MinRate, the EE is an increasing function reaching 120 bits/J/Hz for 6 nodes per cluster. This increase on EE is at the expense of min-rate performance as proven by Figure 5.

Furthermore, we evaluate the effect of water condition on the performance of the proposed models. To this end, we consider three different water conditions defined by variable water attenuation coefficient values, c, as shown in Table 3. The results displayed in Figures 7- 8 reveal that the

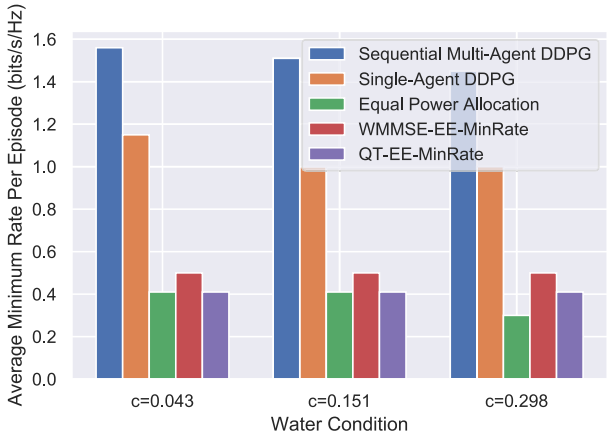


FIGURE 7. Average minimum rate per episode for different water conditions.

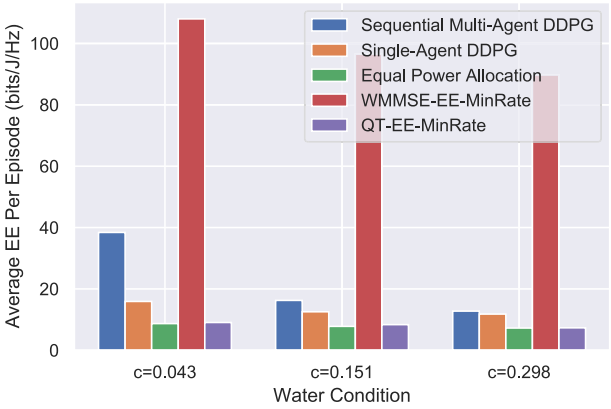


FIGURE 8. Average EE per episode for different water conditions.

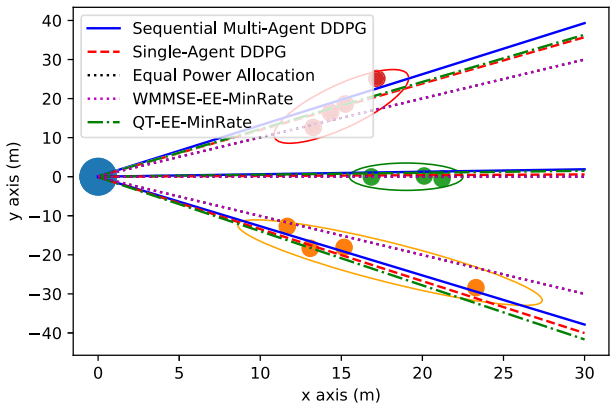


FIGURE 9. Comparison of the resulting beam axis orientations of the proposed methods and benchmarks.

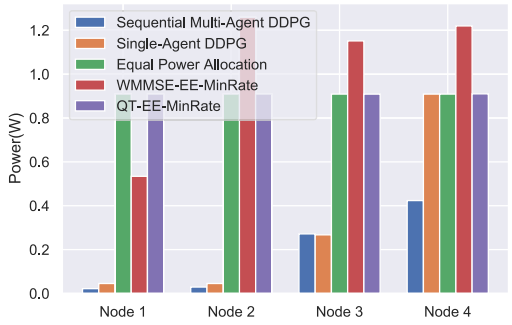
TABLE 4. Attenuation coefficient of various types of water for $\lambda = 514nm$ [41].

Water type	$c(m^{-1})$
Pure sea water	0.043
Clean ocean	0.151
Coastal ocean	0.298
Turbid harbor	2.19

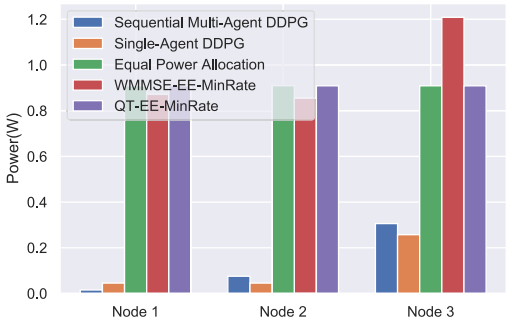
average minimum data rate and EE of the proposed models deteriorate with increasing the attenuation coefficient. This can be explained by the fact that increasing the attenuation coefficient affects the communication channel, which has an impact on the data rate and EE of the nodes. However, the proposed models outperform the benchmarks in terms of minimum data rates for the different water conditions that are considered. For instance, for $c = 0.043$, the sequential multi-agent DDPG model achieved a 74% higher minimum rate per episode as compared to the equal power allocation and QT-EE-MinRate, and a 68% higher min-rate than WMMSE-EE-MinRate. Furthermore, the single-agent DDPG model achieved a 64% higher minimum rate than the equal power allocation and QT-EE-MinRate, and a 56% higher minimum rate than WMMSE-EE-MinRate. Regarding the EE performance, the sequential multi-agent and single agent models achieved 77% and 45% higher EE than equal power allocation and QT-EE-MinRate benchmarks, respectively. In its turn, WMMSE-EE-MinRate yielded a better EE than the proposed models, which can be explained by the fact that this benchmark is maximizing EE at the expense of the min-rate, despite the constraint set on min-rate. Finally, we can conclude that the proposed DDPG frameworks adapt well to the varying underwater environment.

To further evaluate the performance of the proposed methods in comparison with considered benchmarks, we consider a specific case study with three clusters of nodes, where the first (M_1) and third (M_3) clusters have fours nodes, while the second (M_2) cluster has three nodes. The resulting beam-axis orientations and power distribution among the nodes are displayed in Figures 9- 10, respectively. The resulting beam axis for the proposed methods and the equal power benchmark positions itself along the updated node positions, unlike WMMSE-EE-MinRate and QT-EE-MinRate that consider the estimated locations of nodes. This is so because the first three methods adopt the DDPG to converge to the optimal orientation according to the updated locations of the nodes. As for the power distribution among nodes, a certain balance is guaranteed among nodes' powers for the proposed methods, with higher power allocated to further nodes from the source (see Figure 10). Regarding the benchmarks, we observe an almost equal power allocation for the QT-EE-MinRate benchmark, which explains similarity in min-rate and EE results with the equal power allocation benchmark. In our case study, the proposed sequential multi-agent DDPG framework yields improved beam orientation and power distribution among nodes. As confirmed by the results presented in Figures 6- 7, it achieves both higher EE and better rate fairness as compared to the benchmark schemes.

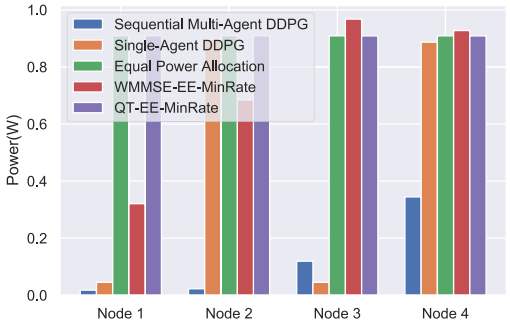
Comparison with Exhaustive Search and Branch and Bound Methods: The performance of the proposed sequential multi-agent DDPG and single-agent DDPG is also evaluated in comparison to the exhaustive search and branch and bound method. Considering the high complexity of these methods, we limit the study to a small number of nodes. Specifically, we consider the following two cases for the exhaustive search method: (a) two clusters with two nodes each, and (b) two clusters with three nodes each. We also



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

FIGURE 10. **Comparison of the resulting power distribution among nodes of the proposed methods and benchmarks.**

TABLE 5. **Comparison with the exhaustive search method.**

	Minimum rate		EE	
	M=[2,2]	M=[3,3,3]	M=[2,2]	M=[3,3]
Sequential Multi-agent DDPG	4.4	2.24	26.16	14.87
Single-agent DDPG	4.38	2.22	21,43	14.26
Exhaustive search	0.42	0.29	14.32	14.57

TABLE 6. **Comparison with the branch and bound method.**

	Minimum rate			EE		
M	[2,2,2]	[3,3,3]	[4,4,4]	[2,2,2]	[3,3,3]	[4,4,4]
Sequential Multi-agent DDPG	4.39	2.28	1.51	23.5	32	35
Single-agent DDPG	4.38	2.24	1.2	17.5	16	15
Branch and Bound	0.99	0.585	0.5	109	110	109

consider the three following cases for the branch and bound method: (a) three clusters with two nodes each, (b) three clusters with three nodes each, and (c) three clusters with four nodes each. To apply the exhaustive search, we discretize the continuous set of orientations and power to 10 equidistant values. The results are depicted in Table 5. We observe that both DDPG-based RL methods achieve a better minimum rate and EE as compared to the exhaustive search method. Of note, the computational complexity of an exhaustive search increases exponentially with the number of available beam orientations and transmit power levels. Therefore, in practice, exhaustive search is typically limited to selecting transmit power and beam orientations from a discrete set of possible actions. This limitation leads to sub-optimal solutions, despite the fact that, theoretically, exhaustive search can find the global optimum. By contrast, DDPG-based RL algorithms can select continuous action values with polynomial computational complexity. Consequently, they achieve superior performance in terms of minimum rate and EE as compared to the considered exhaustive search. Concerning the branch and bound method, we adopt the case of EE maximization with constraint on min-rate. The results show that the branch and bound method achieves higher EE results than the proposed DDPG-based RL methods, but with significantly smaller min-rate results. This outcome can be explained by the fact that, in this case, the branch and bound method tries to maximize the EE on the expense of min-rate. The branch and bound method is globally optimal for a single-objective EE maximization problem with a min-rate constraint. However, the proposed sequential multi-agent DDPG targets a fundamentally different joint objective that simultaneously maximizing both EE and min-rate fairness for which no classical global solver exists at tractable complexity. In this

TABLE 7. Comparative analysis of the proposed method with the considered benchmarks.

Criteria	Equal Power	WMMSE-EE-MinRate	QT-EE-MinRate	Exhaustive search	Branch and bound	Single-agent DDPG	Sequential multi-agent DDPG
Complexity	Low	High	Moderate	Very high	Very high	Low	Moderate
Scalability	Poor	Poor	Very poor	Very poor	Very poor	Moderate	High
Optimality	Sub-optimal	Sub-optimal	Sub-optimal	Sub-optimal	Globally optimal (for given objective)	Near-optimal	Optimal

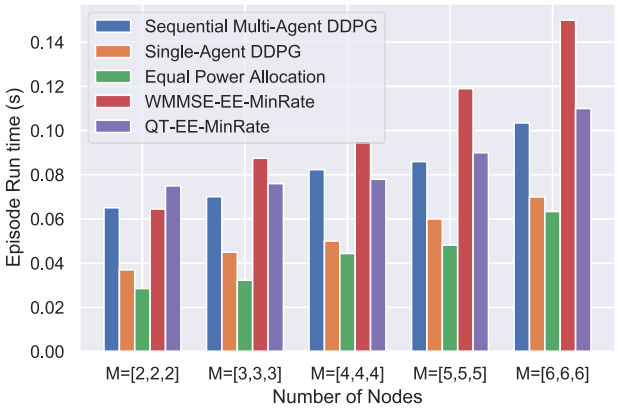


FIGURE 11. Episode run time for different M_n values, $N = 3$, $T = 50$, $\epsilon = 0.25$, $N_0 = 30$, $N_{0,o} = 10$, and $N_{0,p} = 30$.

joint-objective sense, branch and bound achieves high EE at the cost of severely degraded min-rate (0.5 – 0.99 bits/s/Hz), failing to balance both metrics. The proposed method, by contrast, achieves a superior trade-off across both objectives jointly. These results confirm that, in practice, the proposed sequential multi-agent DDPG-based RL algorithm not only converges, but also outperforms globally optimal solutions, which can be attributed to the latter's inherent implementation complexity.

D. ALGORITHM COMPLEXITY ASSESSMENT

Finally, we evaluate the complexity of the proposed algorithms. Figure 11 shows a comparison of the runtime per episode for the different considered algorithms. The single-agent algorithm has a lower runtime per episode than the sequential multi-agent allocation benchmark, reaching around 0.06 s for five nodes per cluster. Meanwhile, the sequential multi-agent algorithm's runtime per episode exceeds 0.08 s for five nodes per cluster. The proposed single-agent algorithm has one running agent, while the sequential multi-agent algorithm includes $2 + N$ agents, which makes it slower and more complex than the single-agent algorithm. Concerning the benchmarks, the equal power benchmark has a lower runtime than the proposed methods. This is so because it has one running agent that decides only on the beam orientations, which lightens the complexity of the algorithm. Meanwhile, QT-EE-MinRate has a similar run time as compared to the sequential multi-agent approach, and WMMSE-EE-MinRate exhibits the highest runtime for three or higher nodes per cluster. This behavior can be explained by the fact that these two benchmarks rely on repeated alternating updates and feasibility enforcement at every snapshot, whereas the trained DRL policy amortizes the optimization effort into an offline training stage, thereby enabling fast online decision-making. More specifically, the proposed framework's computational complexity is primarily dominated by neural network operations (i.e., multiplications and additions) and does not require any additional iterative optimization procedures.

Finally, we can conclude that the proposed single-agent and sequential multi-agent solutions achieve good results for the network's minimum rate/EE tradeoff. Compared to the single-agent model, the sequential multi-agent model enhances the network fairness and EE for higher number of nodes, with a slight increase in the run time. The sequential learning feature further improves the results of the multi-agent solution. Also, the proposed models considerably increases the network's fairness as compared to the considered benchmarks. Table 7 presents the results of a comparative analysis of the proposed models with the benchmarks.

VII. CONCLUSION

In this comprehensive exploration of the domain of UOWC, we aimed to optimize the beamforming task for a multi-beam network using the NOMA technique within each transmitted beam. A joint beam orientation and power control problem was formulated to concurrently optimize both system EE and minimum rate. To efficiently solve this problem, single-agent and multi-agent DDPG RL solutions were developed. A comparison of these solutions with the considered benchmarks-namely the equal power allocation, WMMSE-EE-MinRate and QT-EE-MinRate, proved that the proposed solution ensures more balanced results for the system's minimum rate and EE tradeoff. In addition, by distributing the task among different agents and hence alleviating the decision for each agent, the multi-agent solution improved the EE and fairness as compared to the single-agent solution for a higher number of nodes per cluster.

The proposed methods can be deployed in real-world scenarios by adding a control unit to the multi-beam transmitter. In the case of AUVs, the control unit is usually already in place, since the AUVs are remotely monitored. Importantly, the training of the DDPG agent is also performed offline. Once the model converges, only the trained model is used to perform the beamforming task. Importantly, the computationally intensive training phase is limited to only a short time (e.g., a couple of seconds as depicted by the presented results), where the convergence is achieved within

50 episodes and the runtime per each episode is less than 0.1 seconds. Thus, our proposed DDPG framework is feasible for practical UOWC systems.

The proposed methods are designed for quasi-stationary nodes, perfect SIC and CSI conditions. In our future work, we intend to investigate the case of mobile nodes where the receiving sensors are mobile AUVs traveling underwater. Under such conditions, the clustering task should be conducted dynamically. Furthermore, to alleviate the power consumption on the underwater system and realize training using more powerful equipment, the integration of digital twin that replicates the underwater propagation scenario will be investigated. In this context, we will address the case of imperfect CSI estimations and develop an appropriate RL-guided robust optimization algorithm. Finally, a prospective experimental evaluation of the proposed models will be conducted to further demonstrate their effectiveness in real-world underwater scenarios.

REFERENCES

[1] H. Kaushal and G. Kaddoum, "Underwater optical wireless communication," *IEEE Access*, vol. 4, pp. 1518–1547, 2016.

[2] Z. Zeng, S. Fu, H. Zhang, Y. Dong, and J. Cheng, "A survey of underwater optical wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 204–238, 1st Quart., 2017.

[3] A. Celik, I. Romdhane, G. Kaddoum, and A. M. Eltawil, "A top-down survey on optical wireless communications for the Internet of Things," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 1–45, 1st Quart., 2023.

[4] N. F. O. Korotkova and E. Shchepakina, "Light scintillation in oceanic turbulence," *Waves Random Complex Media*, vol. 22, no. 2, pp. 260–266, 2012.

[5] M. V. Jamali et al., "Statistical studies of fading in underwater wireless optical channels in the presence of air bubble, temperature, and salinity random variations," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4706–4723, Oct. 2018.

[6] P. Yue, X. Wang, D. Xu, and S. Xu, "Non-line-of-sight scattering channel modeling of MIMO links for underwater wireless optical communication," *Opt. Commun.*, vol. 578, Apr. 2025, Art. no. 131468.

[7] W. Wang, Y. Li, B. Hao, Y. Lv, and M. Ju, "Error performance and power allocation for MIMO-UWOC systems in composite oceanic fading scenario," *Opt. Commun.*, vol. 597, Dec. 2025, Art. no. 132301.

[8] C. S. S. Shetty, R. P. Naik, U. S. Acharya, and W.-Y. Chung, "Performance analysis of MIMO-EGC system for the underwater vertical wireless optical communication link," *IEEE Access*, vol. 11, pp. 99253–99267, 2023.

[9] X. Li et al., "Experimental demonstration of a real-time multi-user uplink UWOC system based on SIC-free NOMA," *Opt. Exp.*, vol. 31, no. 19, pp. 30146–30159, 2023.

[10] W. M. Salama, M. H. Aly, and E. S. Amer, "Underwater optical wireless communication system: Deep learning CNN with NOMA-based performance analysis," *Opt. Quantum Electron.*, vol. 55, no. 5, p. 436, May 2023.

[11] K. W. S. Palitharathna, H. A. Suraweera, R. I. Godaliyadda, V. R. Herath, and Z. Ding, "Lightwave power transfer in full-duplex NOMA underwater optical wireless communication systems," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 622–626, Mar. 2022.

[12] Y. Liang, H. Yin, L. Jing, X. Ji, and J. Wang, "Solution for self-interference of NOMA-based wireless optical communication system in underwater turbulence environment," *IEEE Access*, vol. 11, pp. 30223–30236, 2023.

[13] Y. Li, S. A. H. Mohsan, X. Chen, R. Tehseen, S. Li, and J. Wang, "Research on power allocation in multiple-beam space division access based on NOMA for underwater optical communication," *Sensors*, vol. 23, no. 3, p. 1746, Feb. 2023.

[14] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks: Issues, measures and challenges," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 5–24, 1st Quart., 2014.

[15] M. N. Dani, D. K. C. So, J. Tang, and Z. Ding, "Resource allocation for layered multicast video streaming in NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 11379–11394, Nov. 2022.

[16] R. Jiao, L. Dai, W. Wang, F. Lyu, N. Cheng, and X. Shen, "Max-min fairness for beamspace MIMO-NOMA: From single-beam to multi-beam," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 739–752, Feb. 2022.

[17] Y. Li, Y. Jiang, and J. Wang, "Research on power allocation of MIMO-NOMA in multi-beam space division multiple access for underwater optical communication," in *Proc. 14th Int. Conf. Signal Process. Syst. (ICSPS)*, Nov. 2022, pp. 645–650.

[18] J. Wang, J. Wang, Y. Li, Y. Jiang, and X. Chen, "Power allocation algorithm considering qos and max-min fairness in single-beam UVLC system," in *Proc. 14th Int. Conf. Signal Process. Syst. (ICSPS)*, Nov. 2022, pp. 628–633.

[19] W. K. G. Seah, C.-H. Lee, Y.-D. Lin, and Y.-C. Lai, "Combined communication and computing resource scheduling in sliced 5G multi-access edge computing systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3144–3154, Mar. 2022.

[20] M. Rahmani et al., "Deep reinforcement learning-based sum rate fairness trade-off for cell-free mMIMO," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6039–6055, May 2023.

[21] A. Magbool, V. Kumar, and M. F. Flanagan, "On energy efficiency and fairness maximization in RIS-assisted MU-MISO mmWave communications," in *Proc. ICC - IEEE Int. Conf. Commun.*, May 2023, pp. 5364–5369.

[22] Y. Gou, T. Zhang, T. Yang, J. Liu, S. Song, and J.-H. Cui, "A deep MARL-based power-management strategy for improving the fair reuse of UWSNs," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6507–6522, Apr. 2023.

[23] Y. Gou, T. Zhang, J. Liu, T. Yang, S. Song, and J.-H. Cui, "Achieving fair-effective communications and robustness in underwater acoustic sensor networks: A semi-cooperative approach," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 5722–5739, May 2024.

[24] A. Agarwal and I. Krikidis, "Fairness-driven optimization for NOMA-UWOC systems with energy harvesting requirements," *IEEE J. Ocean. Eng.*, vol. 50, no. 1, pp. 403–418, Jan. 2025.

[25] H. Shin, S. Baek, and Y. Song, "Multidimensional beam optimization in underwater optical wireless communication based on deep reinforcement learning," *IEEE Internet Things J.*, vol. 11, no. 17, pp. 28623–28634, Sep. 2024.

[26] M. Li, H. Luo, H. Tao, X. Li, P. Dong, and K. Wu, "Collaborative multi-AUV optical communication via deep reinforcement learning," *IEEE Sensors J.*, vol. 25, no. 1, pp. 1627–1640, Jan. 2025.

[27] H. Shin, S. M. Kim, and Y. Song, "Learning-aided joint beam divergence angle and power optimization for seamless and energy-efficient underwater optical communication," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22726–22739, Dec. 2023.

[28] Y. Jiang, Y. Li, X. Chen, S. Li, and T. Wang, "A strategy for reducing interference in underwater optical multi-beam systems based on NOMA," in *Proc. IEEE 11th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, Dec. 2023, pp. 722–727.

[29] J. Shi, C. Ma, X. Tian, H. Guo, and J. Ao, "Optimization and modeling of optical emission spatial coverage from underwater multi-faceted optical base stations," *Photonics*, vol. 12, no. 1, p. 4, Dec. 2024.

[30] K. W. S. Palitharathna, H. A. Suraweera, R. I. Godaliyadda, V. R. Herath, and J. S. Thompson, "Average rate analysis of cooperative NOMA aided underwater optical wireless systems," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2292–2310, 2021.

[31] Q. Xu, X. Li, H. Ji, and X. Du, "Energy-efficient resource allocation for heterogeneous services in OFDMA downlink networks: Systematic perspective," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2071–2082, Jun. 2014.

[32] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[33] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[34] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Proc. 2nd Conf. Learn. Dyn. Control*, Jun. 2020, pp. 486–489.

[35] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[36] J. Li et al., "A real-time, full-duplex system for underwater wireless optical communication: Hardware structure and optical link model," *IEEE Access*, vol. 8, pp. 109372–109387, 2020.

[37] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[38] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[39] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.
[40] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Found. Trends Commun. Inf. Theory*, vol. 11, nos. 3–4, pp. 185–396, Jun. 2015.
[41] C. D. Mobley, *Light and Water: Radiative Transfer in Natural Waters*. New York, NY, USA: Academic, 1994.

IMENE ROMDHANE (Student Member, IEEE) received the B.S. degree in telecommunication engineering from the École Supérieure de Communications de Tunis, Aryanah, Tunisia, in 2016, and the M.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Türkiye, in 2019. She is currently pursuing the Ph.D. degree in electrical engineering with the École de Technologie Supérieure, Université du Québec, Montreal, QC, Canada. During her studies, she has focused on optical communications, including fiber communications, wireless optical communications, and underwater optical wireless communications. Also, she has shown interest in artificial intelligent by applying machine learning techniques, including reinforcement learning, to the optical field.

ZIYAUR RAHMAN (Student Member, IEEE) received the B.Tech. and M.Tech. degrees in electronics and communication engineering from Punjab Technical University, Kapurthala, India, in 2015 and 2017, respectively, and the Ph.D. degree in wireless communications from the Birla Institute of Technology and Science, Pilani, Pilani Campus, Rajasthan, India, in 2023. He is currently a Post-Doctoral Research Fellow in the area of underwater optical wireless communications with the Department of Electrical Engineering, École de Technologie Supérieure (ETS), University of Quebec, Montreal, Canada. His research interests include optical wireless communications for terrestrial and underwater applications and machine learning for communication systems.

NAHED BELHADJ MOHAMED received the B.E. degree in telecommunication engineering from the Higher School of Communication of Tunis (SUP'COM), Ariana, Tunisia, in 2018. She is currently pursuing the Ph.D. degree in electrical engineering with the École de Technologie Supérieure (ETS), Montreal, QC, Canada. Her research interests include wireless communications, the Internet of Things, radio resource management, and the application of machine learning in physical layer communications. She is a reviewer in several prestigious conferences, such as ICC and ICMLCN.

MD. ZOHEB HASSAN (Member, IEEE) received the Ph.D. degree from the Electrical and Computer Engineering Department, The University of British Columbia, Vancouver, Canada. He is an Assistant Professor with the Department of Electrical and Computer Engineering, Université Laval, Canada. Prior to joining Université Laval, he was a Senior Post-Doctoral Research Fellow with the École de Technologie Supérieure (ETS); and a Research Assistant Professor with the ECE Department, Virginia Tech, USA. He has authored or co-authored over 30 journal articles and 15 conference papers in radio resource optimization, interference management, spectrum sharing, and optical wireless communications. He was a recipient of the NSERC Postdoctoral Fellowship Award in 2021 and the Four-Year Fellowship at The University of British Columbia in 2014. He has served/is serving as a TPC Member for various prestigious IEEE conferences, such as IEEE GLOBECOM, ICC, MILCOM, VTC, and PIMRC; and a reviewer for several major journals of the IEEE Communications Society.

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the École Nationale Supérieure de Techniques Avancées (ENSTA Bretagne), Brest, France, the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the Université de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences (INSA), University of Toulouse, Toulouse, France, in 2009. He is currently a Professor, the Research Director of the Resilient Machine Learning Institute (ReMI), and the Industrial Research Chair and the Tier 2 Canada Research Chair of the École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, Canada. He has published over more than 300 journals, conference papers, and two chapters in books; and has eight pending patents. His current research interests include wireless communication networks, tactical communications, resource allocations, and network security. He received the Best Papers Awards at 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications (WIMOB); 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC); and 2023 IEEE International Wireless Communications and Mobile Computing Conference (IWCMC). He received the IEEE Transactions on Communications Exemplary Reviewer Award, in 2015, 2017, and 2019; the Research Excellence Award of the Université du Québec, in 2018; the Research Excellence Award from ÉTS in 2019 in recognition of his outstanding research outcomes; the 2022 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence (Middle Career Researcher); and the prestigious 2023 MITACS Award for Exceptional Leadership. He served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE COMMUNICATIONS LETTERS. He is currently serving as an Area Editor for IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING and an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.