Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5042833, 12 pages
https://doi.org/10.1155/2022/5042833

Research Article

Reinforcement Learning for Distributed Energy Efficiency Optimization in Underwater Acoustic Communication Networks

Liejun Yang,[1] Hui Wang,[2] Yexian Fan,[1] Fang Luo,[1] and Wei Feng[1]

[1]*College of Information and Mechanical & Electrical Engineering, Ningde Normal University, Ningde 352000, China*
[2]*College of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China*

Correspondence should be addressed to Hui Wang; wangh0802@163.com

Received 9 December 2021; Revised 28 January 2022; Accepted 31 January 2022; Published 24 February 2022

Academic Editor: Xuebo Zhang

To solve the problems of poor quality of service and low energy efficiency of nodes in underwater multinode communication networks, a distributed power allocation algorithm based on reinforcement learning is proposed. The transmitter with reinforcement learning capability can select the power level autonomously to achieve the goal of getting higher user experience quality with lower power consumption. Firstly, we propose a distributed power optimization model based on the Markov decision process. Secondly, we further give a reward function suitable for multiobjective optimization. Finally, we present a distributed power allocation algorithm based on Q-learning and use it as an adaptive mechanism to enable each transmitter in the network to adjust the transmit power according to its own environment. The simulation results show that the proposed algorithm not only increases the total channel capacity of the system but also improves the energy efficiency of each transmitter.

1. Introduction

Marine information technology not only plays an important role in the fields of marine environment monitoring, exploration and resource development, marine disaster warning, and underwater target location tracking but also is a hot direction for information science research [1, 2]. The primary problem to be solved in the development of marine information technology is the construction of underwater sensor networks and the allocation of resources for network communication; otherwise, marine information technology is not possible [3–5]. With the increasing exploitation of underwater resources, the variety and number of communication nodes deployed underwater are becoming more and more abundant, and there will even be multiple types of underwater communication networks deployed in the same sea area. For example, in Ref. [6], a two-dimensional underwater sensing network structure was developed in which the sensor nodes were anchored to the seafloor. This means that the sensors can only detect a range of data on the seafloor. However, many other important 3D data, such as the flow rate and salinity of seawater, which are crucial for one to study the characteristics of the marine environment, are not detectable. Correspondingly, this paper proposes an autonomous underwater vehicle (AUV) to monitor and collect important 3D data, and uses different types of sensors to detect a range of data on the seafloor.

Unlike wireless electromagnetic wave communication networks, most acoustic modems in underwater acoustic communication networks (UACNs) are battery-powered, but in an underwater environment, battery replacement and charging are extremely difficult [7]. Meanwhile, there are many types of nodes deployed in UACNs, including multiple types of nodes such as master nodes, sub-nodes, AUVs, and so on. Normally, different types of nodes hope to transmit data with greater power to obtain a higher quality of service [8]. In this case, if proper interference control is not performed, there will be increased interference between nodes and a huge waste of transmit power. So it can be seen that, because of the complex underwater acoustic communication environment, the proposed resource allocation algorithm needs to have strong adaptive characteristics to counter the dynamic underwater acoustic communication environment. The low transmission rate of

orthogonal frequency-division multiplexing (OFDM) technology has obvious advantages in combating the complex communication environment of underwater acoustics. Its low transmission rate effectively reduces multipath reflection interference [9], and it is also extremely resistant to inter-code interference [10]. Motivated by previous analysis, based on the modeling of OFDM underwater heterogeneous communication networks, we consider how to find a balance between power consumption and interference level to achieve optimal system performance.

To summarize, we consider the issue of energy efficiency optimization in cooperative UACNs. Since the resource allocation process can be considered as a Markov decision process (MDP), reinforcement learning (RL) is applied to solve the above problems [11]. Specifically, RL methods are used to find the equilibrium between power consumption and interference level, i.e., to select the appropriate transmit power for each node to obtain a high quality of service within the interference allowable range. To this end, this paper seeks the global optimal strategy by constructing a global MDP. The main contributions of this work are summarized as follows:

(i) We propose a learning framework suitable for communication nodes. The framework realizes the transformation of resource allocation problem like the Markov decision model, which defines the state space and action set in the environment according to the actual problem that needs to be solved.

(ii) We propose a systematic reward function design method based on the multiobjective optimization problem and the nature of RL, which is used to guide the training method of the transmitter. The designed reward function takes into account the network environment and node energy which are uncontrollable factors, and achieves maximization of quality of service (QoS) of communication nodes with relatively small energy consumption. We further show that the proposed reward function can achieve significant improvements in energy efficiency.

(iii) We propose a resource allocation strategy for underwater transmitters based on Q-learning, which is distributed and scalable. The simulation results show that, compared with the greedy algorithm, the resource allocation strategy based on Q-learning achieves a higher system capacity and a longer life cycle.

The rest of this paper is organized as follows. Section 2 reviews the work related to resource allocation in UACNs. Section 3 introduces the multisectional cooperative communication network model and describes the problems related to resource allocation. Section 4 proposes a resource allocation strategy based on Q-learning and proves the effectiveness of the designed scheme theoretically, and Section 5 compares the proposed algorithm with the greedy algorithm. Finally, Section 6 concludes the paper.

2. Related Work

Compared to the channel bandwidth on land, the available bandwidth underwater is very narrow, only a few kilohertz. When there are more underwater communication nodes, many nodes will communicate in similar frequency bands, which will generate large interference between nodes and affect the communication quality of underwater nodes. Facing the complicated underwater communication environment, many scholars have improved the communication quality of underwater sensor networks by rationally allocating resources such as channels and power.

The problem of resource allocation has been extensively studied in UACNs. Aiming at the energy limitation and throughput problems in UACNs, the linear Gaussian relay channel (LGRC) model is used in Ref. [12] to optimize the power spectral density of the input power, effectively expanding the transmission capacity of UACNs. In a similar study, For the MQAM-OFDM underwater acoustic communication system, a joint power-rate allocation algorithm is proposed in Ref. [13], which optimizes the transmission power of the node and improves the transmission rate of the system. In Ref. [14], the authors proposed an efficient spectrum management system receiver-initiated spectrum management (RISM) for underwater acoustic cognitive networks and aimed to maximize the node channel capacity for power allocation, which effectively avoids conflicts in data transmission and improves the data transmission rate. However, the centralized optimization algorithm proposed by the abovementioned study only optimizes the transmission rate of the node, and does not consider the quality of service of the network. In order to improve its own throughput, each transmitting node usually chooses a larger transmitting power, which causes more serious network interference and further reduces the life cycle of the node. In Ref. [15], a joint frequency-power allocation-based algorithm is proposed for UACNs, which effectively extends the life cycle of nodes by setting the power level according to the distance between nodes. The disadvantage is that this algorithm is only suitable for environments with dense network nodes. Meanwhile, considering the complex underwater communication environment, it is difficult to deploy a centralized control center underwater, so the abovementioned centralized power algorithm cannot meet the strong distributed application requirements of the UACNs.

RL has been developed to continuously optimize its own strategies through continuous interaction with unknown environments, and can be used in a distributed manner to achieve better results in many scenarios [16, 17]. For example, in order to solve the multinode interference problem in UACNs, in Ref. [5], the authors converted the resource allocation problem into a Markov decision model and proposed a cooperative Q-learning optimization scheme. However, Ref. [6] did not consider the node energy consumption. Furthermore, an anti-interference relay selection scheme for deep Q network (DQN) is proposed in Ref. [18], which selects the node position based on the interference level of the node on the one hand, and adjusts the node

transmit power according to the magnitude of the BER on the other hand. The disadvantage is that the algorithm only considers a network composed of a few nodes and lacks scalability. Therefore, in order to balance node energy consumption and network interference level, and at the same time, considering the scalability of the algorithm, this paper regards the communication node as an agent, and transforms the resource allocation problem into the Q-learning algorithm model to obtain the optimized strategy result.

3. System Model and Problem Formulation

3.1. System Model. In this paper, we consider the UACNs OFDM system composed of multiple transmitter-receiver pairs. In UACNs, the transmitting nodes collect environmental information, and the receiving nodes are relay nodes or data fusion centers. According to application needs, there are many types of transmitting nodes, including sensor nodes, Autonomous Underwater Vehicle (AUV), Unmanned Underwater Vehicle (UUV), and many others. Different types of transmitter-receiver pairs have different communication requirements and priority levels. The bandwidth of the OFDM system is equally divided into $\mathbf{L}$ orthogonal sub-channels, whose set is denoted as $\mathbf{L} = [1, 2, \ldots, L]$. For convenience, we assume that the bandwidth of each sub-channel is the unit bandwidth. All orthogonal channels are shared channels that can be freely accessed by all transmitter-receiver pairs. Meanwhile, suppose that there are N pairs of sensor nodes and 1 pair of AUV pairs in the network, where $\mathbf{N} = [1, 2, \ldots, N]$ represents the index of the sensor node. The overall network configuration is shown in Figure 1. Please note that although we consider each transmitter to serve a single receiver, the proposed method can be easily adapted to serve more transmitter-receiver pairs.

From the above text, the received signal of node $n_i^R, \forall i \in \mathbf{N}$ includes interference from node $n_i^R\,(j \neq i, j \in N)$ and thermal noise; then the signal-to-interference-to-noise ratio (SINR) at node $n_i^R, \forall i \in \mathbf{N}$ can be expressed as Ref. [19]

$$\eta_i = \frac{p_i h_{ii}}{\sum_{k=1,k\neq i}^{N} p_k h_{ki} + p_j h_{ji} + \sigma^2}, \tag{1}$$

where p_j is the transmit power of AUV j; h_{ji} denotes the channel gain from the AUV j to node n_i^R; p_i indicates the transmit power of node n_i^T; h_{ii} is the channel gain from node n_i^T to node n_i^R; p_k is the transmit power of node n_k^T; and h_{ki} denotes the channel gain from node n_k^T to node n_i^R. σ^2 denotes the noise power of the underwater acoustic channel. Underwater acoustic channel noise is an important topic in the application practice of UACNs, as hydrostatic pressure effects (tides, waves, etc., caused by wind, rain, and seismic disturbances) and industrial behavior (e.g., surface sailing) remain one of the main reasons hindering the development of underwater acoustic communication [20–22]. Calculating the noise power σ^2 is a very complex challenge, because of the significant time-space-frequency variability of underwater acoustic channel noise [23, 24]. Fortunately, σ^2 can be calculated from the corresponding power spectral density [15, 25], which can be described as follows:

$$\varphi(f) = N_\tau(f) + N_w(f) + N_{th}(f) + N_t(f), \tag{2}$$

where

$$\begin{aligned} 10\log N_\tau(f) &= 40 + 20(\tau - 0.5) + 26\log_{10}(f) - 60\log_{10}(f + 3),\\ 10\log N_w(f) &= 50 + 7.5\sqrt{w} + 20\log_{10}(f) - 40\log_{10}(f + 0.4),\\ 10\log N_{\text{th}}(f) &= -15 + 20\log_{10}(f),\\ 10\log N_t(f) &= 17 - 30\log_{10}(f), \end{aligned} \tag{3}$$

where $N_\tau(f)$, $N_w(f)$, $N_{\text{th}}(f)$, and $N_t(f)$ denote ocean turbulence, ship activity, wind and waves, and thermal movement of molecules in the water, respectively. In addition, w and τ represent the influencing factor of sea surface wind speed and ship activity, respectively.

In the underwater acoustic communication system, the channel gain h can be expressed as Ref. [25]

$$h = A_0^{-1} d^{-sp} (\alpha(f))^{-d}, \tag{4}$$

where A_0 is the normalization coefficient, d denotes the transmission distance (km), f indicates the communication frequency (Hz), d^{-sp} is the expansion loss, which describes the channel characteristics of underwater acoustic propagation, sp denotes the expansion coefficient, with a value of 1.5, and $\alpha(f)$ is the absorption coefficient, which can be expressed by Thorp empirical formula as [26]

$$10\alpha(f) = \frac{0.11 f^2}{1 + f^2} + \frac{44 f^2}{4100 + f^2} + 2.75 \times 10^{-4} f^2 + 0.003. \tag{5}$$

Assume that all channel parameters are known by the transmitting node, which is consistent with previous work such as Refs. [3, 5]. In fact, this is reasonable, because the channel information can be fed back to each transmitting node through the backhaul network. Thus, the normalized capacity of any receiver can be expressed as follows:

$$C_i = \log_2(1 + \eta_i), \quad \forall i \in \mathbf{N}. \tag{6}$$

3.2. Problem Formulation. During the operation of UACNs, when the noise conditions of the underwater acoustic channel are given, each transmitter hopes to transmit data with a larger power in order to obtain a higher quality of

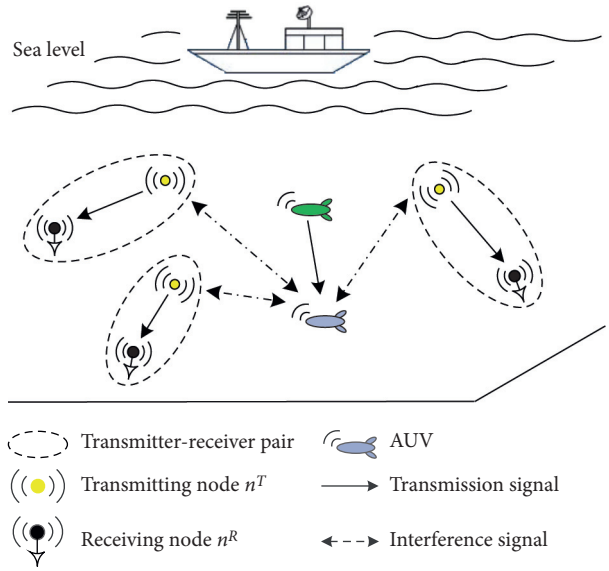

Figure 1: Underwater acoustic communication network model.

service. However, excessive transmission of power will increase the level of network interference, which will greatly reduce the communication quality. Besides this, transmitter usually uses battery power when working underwater, and excessive transmitting power will accelerate the energy consumption of the transmitter. Therefore, the main goal of our work is to solve the energy optimization problem, i.e., to maximize the service quality of the receiver with a smaller energy consumption.

As mentioned previously, if we assume that the transmitting power of the transmitting node n^T is $\mathbf{P} = [p_1, p_2, \ldots, p_N]$, then the optimization goal can be expressed as follows:

$$\max\left\{\sum_{i=1}^{N} C_i, -\sum_{i=1}^{N} p_i\right\}, \tag{7}$$

$$\text{s.t. } p_{\min} \le p_i \le p_{\max}, \quad i = 1, 2, \ldots, N, \tag{8}$$

$$\eta_i \ge \eta_{\text{th}}, \quad i = 1, 2, \ldots, N, \tag{9}$$

$$\eta_{\text{AUV}} \ge \eta_{\text{th}}', \tag{10}$$

where the objective (7) indicates the maximization of the network capacity with relatively small energy consumption. C_i denotes the information transmission capacity between the j-th transmitter-receiver pair, and p_j is the transmit power of the j-th transmitter node. The first constraint (8) denotes the power limit of the transmitting node $n_i^T, \forall i \in \mathbf{N}$. The η_{th} in (9) and η_{th}' in (10), respectively, denote the minimum SINR of node $n_i^R, \forall i \in \mathbf{N}$ and the AUV when meeting application requirements. In other words, constraints (9) and (10) ensure that all receivers have sufficient quality of service. Considering (8)–(10), it can be concluded that the optimization in (7) is not only a multiobjective optimization problem but also a nonconvex problem of UACNs. This is mainly because of the SINR expression in (1) and the optimization goal of (7). In the next section, a method based on reinforcement learning is proposed to solve the above problems.

4. Resource Allocation Based on Reinforcement Learning

4.1. Markov Decision Process. The environment that interacts with the agent is usually called a Markov Decision Process (MDP) with a finite state. We assume that S represents the discrete set of environmental states, A is the discrete set of actions that the agent can perform, r represents the reward value of the agent performing action $a, a \in A$ in state $s, s \in S$, and g be the state transition function. At each time t, the agent interacts with the environment to obtain the current state $s^t = s$, and selects an action $a^t = a$ from the action set A to execute. According to the probability distribution relation $g(s'|s, a)$, the environment is thus changed, shifting from state $s^t = s$ to $s^{(t+1)} = s'$ and generating feedback on the choice of action of the intelligence, that is, the reward value $r(s, a)$. The whole process is iterated and optimized until convergence.

The goal of the RL method is to continuously optimize the agent's decision strategy π in the iterative process. Formally, strategy π describes the mapping relationship from environmental state to action selection. The task of the intelligence is to obtain the optimal policy during the learning process so that the total expected discounted return reaches the maximum in a finite number of steps, that is

$$V^{\pi}(s) = E\left[\sum_{t=0}^{+\infty} \gamma^t r\left(s^t, \pi\left(s^t\right)\right) | s^0 = s_0\right], \tag{11}$$

where γ^t is the reward discount factor at the moment; s_0 is the initial state of the system; and r is the immediate reward obtained by executing the action strategy. $V^{\pi}(s)$ is often referred to as the value function of the intelligence at state s.

The process of RL can be described as an MDP, which has Markov properties. In other words, the state of the

environment is only related to the state of the previous moment, and not related to the state of the earlier time. Therefore, the value function can be simplified to

$$V^{\pi}(s) = E[r(s,\pi(s))] + \gamma \sum_{s' \in S} g(s' \mid s, \pi(s)) V^{\pi}(s'). \quad (12)$$

Therefore, the optimal strategy satisfies the Bellman equation as [9]

$$V^{*}(s) = \max_{a \in A} \left\{ E[r(s,a)] + \gamma \sum_{s' \in S} g(s' \mid s, a) V^{*}(s') \right\}. \quad (13)$$

However, in the actual systems, the state transition function is generally unknown. The agent cannot model the quadruple $\langle S, A, r, g \rangle$ of reinforcement learning. Therefore, it is necessary to use model-free RL algorithms. Q-learning is the most representative of these algorithms. The Q-function is defined as

$$Q^{*}(s,a) = E[r(s,a)] + \gamma \sum_{s' \in S} g(s' \mid s, a) V^{*}(s'), \quad (14)$$

where $Q^{*}(s,a)$ denotes the cumulative discount reward obtained by selecting action a at state s and choosing the optimal policy all the way through the subsequent policy selection process. Combining equations (12) and (13), the relationship between the value function and the state-action value function can be obtained as follows:

$$V^{*}(s) = \max_{a \in A} Q^{*}(s,a). \quad (15)$$

Therefore, the optimal value function $V^{*}(s)$ can be obtained from $Q^{*}(s,a)$. Then, (14) can be expressed as follows:

$$Q^{*}(s,a) = E[r(s,a)] + \gamma \sum_{s' \in S} \left\{ g(s' \mid s, a) \max_{b \in A} Q^{*}(s', b) \right\}. \quad (16)$$

From the above equation, the update rule of the predicted Q function is provided as [5]

$$Q^{t+1}(s,a) = (1 - \alpha_t) Q^{t}(s,a) + \alpha_t \left[r^{t} + \gamma \max_{b \in A} Q^{*}(s', b) \right], \quad (17)$$

where Q^{t+1} and Q^{t} denote the Q values before and after the update, respectively; $\alpha_t \in [0,1]$ is the learning rate, and a larger α_t value indicates that the update of rewards depends more on immediate rewards than on the accumulation of past experience. It can be seen that the Q value is updated using the optimal Q value of the immediate reward and the next state to which it is transferred, and the basic idea is to estimate the Q function by incrementally summing the Q values of the previous state action pairs.

4.2. Reinforcement Learning-Based Power Allocation Approach. In this paper, each emitter is considered as an intelligent body with RL capability. Next, the most important thing is to transform the resource optimization problem in UACNs into a RL algorithm model and use it to obtain optimal decision results. The existing problem scenario is modeled based on the four elements of reinforcement learning.

4.2.1. Action Space A. According to the optimization goal described in (7), the action of the agent is to select the power. Generally speaking, the Q function is stored in a look-up table. For this, we first discretize the power selection. Assuming the transmit power of the i-th agent, the selection range is $[P_{\min}, P_{\max}]$, which can be discretized as follows:

$$p_i(a_i) = p_{\min} + \frac{a_i}{Y_i}(p_{\max} - p_{\min}), \quad y_i = 0, 1, \ldots, Y_i, \quad (18)$$

where Y_i is the number of discretized powers.

4.2.2. State Space S. The state of the environment should be defined based on local observations. The key to the problem of UACNs resource allocation is to determine the level of interference around each receiver and the energy consumption of the transmitter. Therefore, at time t, we can define the state observed by transmitter i as follows:

$$s_i^{t} = (i, \psi_i, p_i(a_i)), \quad (19)$$

where $\psi_i \in \{0,1\}$ indicates whether the SINR η_i received by receiver i is greater than or lower than its threshold η_i^{*}, that is,

$$\psi_i = \begin{cases} 1, & \text{if } \eta_i(a_i, a_{-i}) \geq \eta_i^{*}, \\ 0, & \text{otherwise}, \end{cases} \quad (20)$$

where $a_{-i} = (a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N) \in A_{-i}$ represents the action vector of other receivers. In this paper, we use s_i to represent the discrete set of environmental states related to receiver i.

4.2.3. Reward Function. The reward value of the agent's RL indicates the degree of satisfaction of the agent with the strategy choice. In the current scenario, the optimization goal is to maximize the QoS of the receiver device with less power consumption, which is essentially a multiobjective optimization problem. In this paper, we transform the multiobjective problem into a single-objective problem by the weight coefficient method, and transform the optimization goal setting into the reward value, denoted as follows:

$$r_i^{t}(a_i, a_{-i}) = \frac{\beta_i}{p_i} C_i^2 C_H - \frac{1}{\beta_i} (C_H - \Gamma_{\text{th}})^2 - (C_i - \Gamma_{\text{th}}')^2. \quad (21)$$

This is based on the following points. In (21), C_H and C_i, respectively, denote the capacity of AUV and node $n_i^{R}, i \in \mathbf{N}$ at time t. Γ_{th} and Γ_{th}' are equal to $\log_2(1 + \Gamma_{\text{th}})$ and $\log_2(1 + \Gamma_{\text{th}}')$, respectively. If there is a higher SINR at the receiver, a lower bit error rate will usually be obtained, which in turn will have a higher throughput. However, an excessively high SINR requires the transmitter to transmit at a high-power level, which in turn will cause more energy consumption and increase interference to other users. To

avoid this, we consider energy efficiency, i.e., (21) select the correct number of received bits per unit of energy consumption as part of the reward function. Simultaneously, (21) also considers the deviation of AUV and node $n_i^R, i \in \mathbf{N}$ from their required capacity thresholds, that is, $(C_H - \Gamma_{\text{th}}')^2$ and $(C_i - \Gamma_{\text{th}})^2$ are reduced from (21) to decrease the value of the reward. In addition, the parameter β_i ensures the fairness of the algorithm. β_i represents the distance between the node $n_i^R, i \in \mathbf{N}$ and the AUV normalized to d_{th}. d_{th} is a constant, indicating whether the node $n_i^R, i \in \mathbf{N}$ is near an AUV. For example, if the distance between the node $n_i^R, i \in \mathbf{N}$ and the AUV is less than d_{th}, the node $n_i^R, i \in \mathbf{N}$ will be affected by the AUV more than any other transmitter with a distance greater than d_{th}. Then, the node $n_i^T, i \in \mathbf{N}$ should give less reward, which means that the first and third terms in (21) are multiplied by the inverse of β_i and β_i to reduce the reward, respectively.

Due to the independent selection of power levels by devices, different devices may interfere greatly with other devices in order to maximize their own profits. In other words, incorrect action selection may cause the SINR of some receivers to fall below its threshold, so the reward value is redefined as

$$R_i(s_i, a_i, a_{-i}) = \begin{cases} r_i(a_i, a_{-i}), & \text{if } \psi_i = 1, \\ 0, & \text{if } \psi_i = 0. \end{cases} \tag{22}$$

Specifically, if the SINR in the current channel is greater than the predefined threshold η_{th} (see (9)), i.e., the QoS is greater than the minimum requirement, the reward value is calculated from (21); otherwise, the reward value is 0. Overall, (22) is the payoff for choosing the power p_i under state s_i^t to ensure the quality of service of the transmission, as well as to achieve energy efficiency.

The convergence of the Q-learning algorithm mainly depends on the convergence of the Q-value function [27]. Next, we will analyze the convergence of the proposed algorithm.

Theorem 1. *The value of the reward function r formulated according to formula (22) is bounded in different system states.*

Proof. From (22),

$$R_i(s_i, a_i, a_{-i}) = \begin{cases} r_i(a_i, a_{-i}), & \text{if } \psi_i = 1, \\ 0, & \text{if } \psi_i = 0, \end{cases} \tag{23}$$

we need to prove that the reward function $R_i(s_i, a_i, a_{-i})$ is bounded in different system states when $\psi_i = 1$.

From (21), $r_i(a_i, a_{-i})$ consists of three components, which are the energy efficiency $\beta_i C_i^2 C_H / p_i$, the deviation of the communication capacity of the AUV from the corresponding capacity threshold $(C_H - \Gamma_{\text{th}})^2/\beta_i$, and the deviation of the communication capacity of the sensor node from the corresponding capacity threshold $(C_i - \Gamma_i)^2$. Here, β_i, $\Gamma_{\text{th}} = \log_2(1 + \eta_{\text{th}})$ and $\Gamma_{\text{th}}' = \log_2(1 + \eta_{\text{th}}')$ are constant.

Consider that the action space A defined by power discretization is a discrete finite value, i.e., $A = \{p_0, p_1, \ldots, p_{Y_i}\}$, the communication capacity C_H of the AUV and the communication capacity $C_i, i \in \mathbf{N}$ of the sensor node are bounded in any state.

Furthermore, the product form composed of the capacity value C_H, the capacity value $C_i, i \in \mathbf{N}$, and the power value p_i must also be a discrete finite value, i.e., the energy efficiency value $\beta_i C_i^2 C_H / p_i$ is bounded. Meanwhile, $(C_i - \Gamma_{\text{th}})^2$ and $(C_H - \Gamma_{\text{th}}')^2$ are bounded. So $r_i(a_i, a_{-i})$ must be bounded. □

Theorem 2. *In the iteration of the Q-value of a bounded reward function r(s,a), the learn factor $0 < \lambda \leq 1$ and satisfies*

$$\sum_{t=1}^{\infty} \lambda_t = \infty, \sum_{t=1}^{\infty} \lambda_t^2 < \infty, \quad \forall s, a. \tag{24}$$

If the optimal Q-value is denoted as $Q*(s, a)$, then when $t \longrightarrow \infty$, we have

$$\lim_{t \longrightarrow \infty} Q_t(s_t, a_t) = Q^*(s_t, a_t). \tag{25}$$

The conclusion exhibited in Theorem 2 has a detailed proof process in Ref. [28], which will not be repeated here.

4.3. Algorithm Description. Based on the above preparatory work, the Q-learning-based resource allocation algorithm for UACNs can be described as follows. Algorithm 1 first initializes the relevant parameters, and then uses the greedy method [29] to guide the behavior selection of the intelligent Q-Agent, and updates the Q-value function based on equation (17), and iterates until the Q-value function converges to make a decision on the resource allocation scheme of UACNs.

5. Numerical Results

In order to verify the effectiveness of the proposed algorithm, the next objective of this section is to evaluate the performance in two different scenarios, i.e., a sparse network consisting of four transmitter-receiver pairs and a dense network with dynamic access consisting of multiple transmitter-receiver pairs. The network model of this paper is shown in Figure 1, and the simulation parameters are set according to Refs. [19, 30]. The maximum transmit power of the transmitter $P_{\max} = 11\,\text{W}$, system bandwidth $W = 1\,\text{MHz}$, propagation coefficient $\varepsilon = 1.5$, carrier frequency $f = 20\,\text{kHz}$, noise power $\sigma^2 = 1.5 \times 10^{-7}\,\text{W}$. In addition, we consider the random nonstationary characteristics of the underwater signal, and use δ to reflect the influence of the underwater uncertainty factors on the underwater acoustic channel, where $\delta = h \times \vartheta$ and ϑ obeys the Rayleigh distribution with a mean value of 0.1. Therefore, $h + \delta$ is used for the gain of the hydroacoustic channel in the simulation.

The minimum SINR requirement for node $n_i^R, i \in \mathbf{N}$ and AUV is defined in terms of the rate required to support its corresponding receiver. In the simulation, we assume that the minimum transmission rate required to satisfy QoS for

Initialization:
(1) Set $\gamma = 0.9$, $\lambda = 0.5$.
(2) Initialize $Q(s,a) = 0, s \in S, a \in A$.
Repeated Learning: (for each episode)
(3) Looks up the Q-table and selects the state s, i.e.,
$s = \operatorname{argmax}_{s \in S} \operatorname{argmax}_{a \in A} Q(s,a)$.
(4) Execute the ε-greedy [29] method to select the action a
$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + (\varepsilon/|A(s)|), & \text{if } a = \operatorname{argmax}_a Q(s,a), \\ (\varepsilon/|A(s)|), & \text{if } a \neq \operatorname{argmax}_a Q(s,a) \end{cases}$$
(5) Calculate the reward function $r(s,a)$ based on equation (22).
(6) Calculate the current Q-value function.
(7) Update the Q-table according to equation (17).
(8) Update the state $s \longleftarrow s'$.
(9) Go back to 3 until the state s is the final state.

Algorithm 1: Q-learning-based UACNs resource allocation algorithm for node $n_i^T, i \in \mathbf{N}$.

node $n_i^T, i \in \mathbf{N}$ is 0.4 b/s/Hz, i.e. $\log_2(1 + \eta_{\text{th}}) = 0.4$ (b/s/Hz). In addition, for AUV, the minimum rate required is set to 1 b/s/Hz, i.e. $\log_2(1 + \eta'_{\text{th}}) = 1$ (b/s/Hz). It is important to note that by knowing the media access control (MAC) layer parameters, the value of the channel transmission rate can be calculated using (Ref. [21], equations (20) and (21)). The parameters associated with performing Q-learning are set as follows: learning rate $\lambda = 0.5$, discount factor $\gamma = 0.9$. e-greedy algorithm is used for the first 80% of iterations, random $e = 0.2$, and the maximum number of iterations is set to 50,000. Besides, in order to achieve noncooperative power allocation in UACNs, one of the most important issues is the definition of the receiving reward. In this paper, the concept of energy efficiency is introduced in (11), which will be used as one of the metrics for numerical evaluation.

We first consider a sparse network consisting of four transmitter-receiver pairs. Assume that the four transmitters and four receivers are randomly distributed in a region that is 1.5 km deep, 1.5 km long, and 1 km wide, and the coordinate information of the nodes is shown in Table 1. Figure 2 shows the effect of the transmit power of the AUV on the other three node $n_i^T, i \in \{1,2,3\}$. As a whole, the SINR of the three nodes $n_i^R, i \in \{1,2,3\}$ gradually decreases as the transmit power of the AUV increases and the network environment interference enhances, which makes the transmission capacity of the three nodes decrease continuously. Further, when the AUV is a certain fixed value, node n_1^R is closest to the AUV and suffers the strongest interference, i.e., the smallest SINR, and thus its acquired capacity is the smallest among the three links. Conversely, node n_3^R is farthest from the AUV and its acquired capacity is the largest.

Figure 3 shows the results of the proposed learning algorithm in this paper compared with the greedy algorithm. In order to make a fair comparison between the two algorithms, we choose energy efficiency as the evaluation index. The results are shown in Figure 3, which indicates that as the power of AUV increases, the network energy efficiency of the proposed learning algorithm, although gradually decreasing, is significantly better than that of the greedy algorithm. It should be noted that, as shown in Figure 2, the decrease in network energy efficiency is a reasonable phenomenon. In fact, in the greedy algorithm, each transmitting node always chooses the maximum power for transmission, which keeps the energy in a high consumption state, but the transmission capacity does not increase significantly.

Figure 4 illustrates the curve of AUV transmission capacity variation with transmit power. From the figure, it can be seen that the proposed algorithm can make the transmission capacity of AUV better than the greedy algorithm. This is mainly because the proposed algorithm can better balance the energy consumption and network interference level, so that the transmit power of each node in the network can be adjusted adaptively to achieve a win-win situation.

Next, we further consider a dynamic access dense network consisting of multiple transmitter-receiver pairs. Assume that the transmitting power of the AUV is 8 W, while the number of sensor nodes in the network increases continuously from 1 to 20 with random distribution. The simulation starts with one transmitter-receiver pair. After convergence, the next transmitter-receiver pair is added to the network and so on. Figure 5 shows the state of the node capacity distribution as the number of nodes in the network increases. As can be seen from the figure, under the same conditions, compared to the greedy algorithm, the learning algorithm proposed in this paper is able to maintain a better network quality of service by adaptively adjusting the node transmitting power according to the changes in the network environment. At the same time, it should be noted that as the number of nodes increases, the level of network interference increases, which makes the overall energy efficiency of nodes show a decreasing trend.

Figure 6 shows the graph of network energy efficiency with increasing number of nodes. It is obvious from the graphs that the proposed algorithm can well balance the network transmission capacity and energy consumption, which greatly improves the network service quality. In the greedy algorithm, all nodes choose the maximum transmission power for the pursuit of higher transmission

Table 1: Location information of the four transmitting receiver pairs.

Location information (km)	n_1^T	n_2^T	n_3^T	AUV
x	0.25	0.5	0.75	0.3
y	−0.2	−0.4	−0.8	−0.3
z	0	0	0	100
Location information (km)	n_1^R	n_2^R	n_3^R	AUV
x	0.25	0.5	0.75	0.3
y	−0.2	−0.4	−0.8	−0.1
z	100	100	100	100

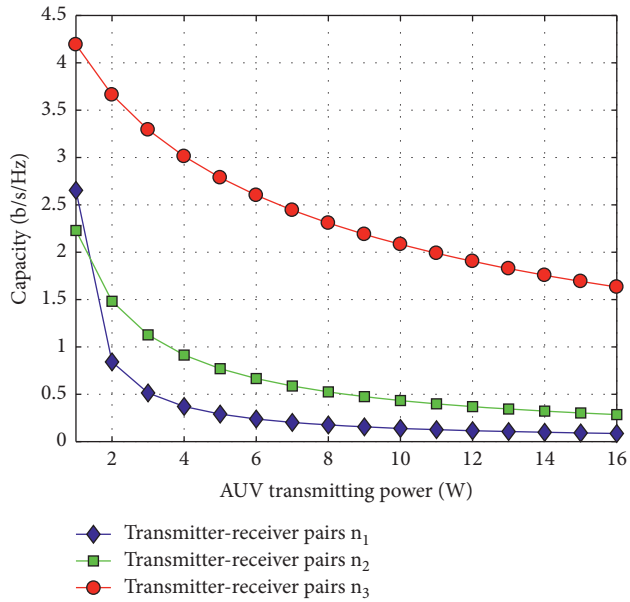

Figure 2: The graph of the change of node capacity with the transmitting power of AUV.

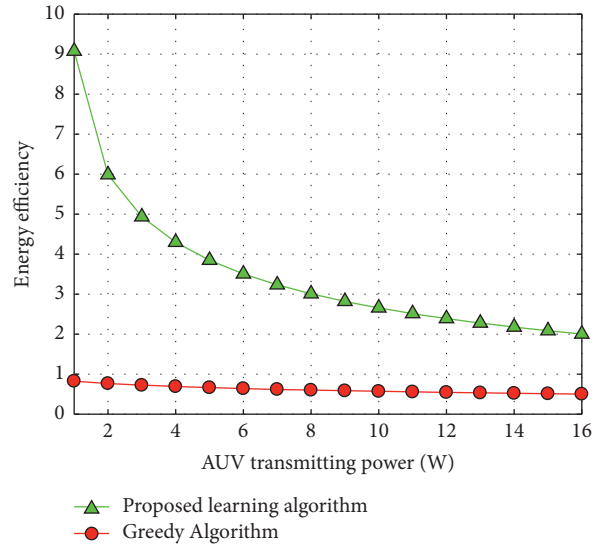

Figure 3: The graph of the change of network energy efficiency with the transmitting power of AUV.

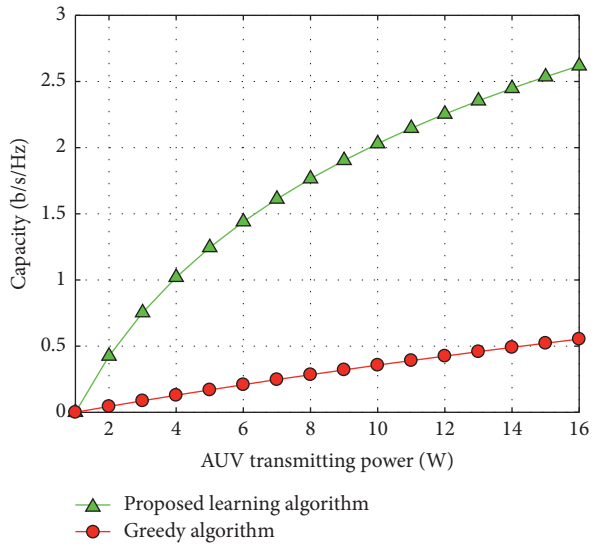

Figure 4: Curve diagram of AUV transmission capacity changing with transmitting power.

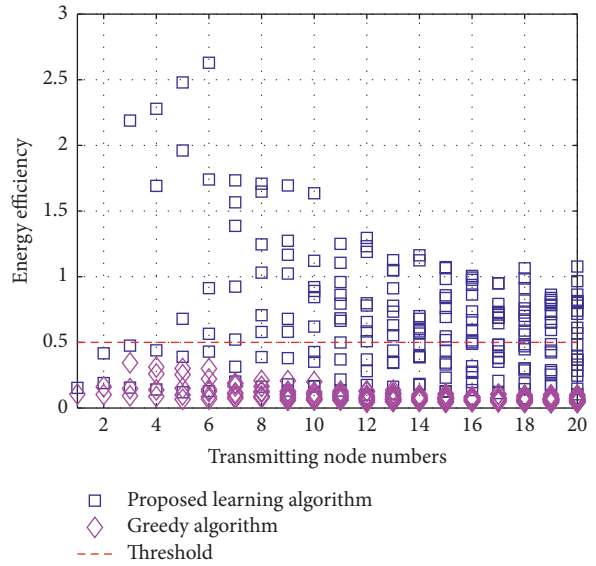

Figure 5: Distribution of energy efficiency as the number of nodes increases.

capacity, which not only causes energy waste but also enhances the interference between networks, and finally makes the network energy efficiency maintain at a low level.

Finally, we perform the convergence and complexity analysis of the algorithm. The maximum number of iterations of the proposed learning algorithm is set to 50,000, and the average number of iterations for the convergence of the algorithm in the two scenarios is shown in Figure 7. From the figure, it can be found that the proposed algorithm requires approximately equal number of iterations in the two different scenarios. In other words, the mathematical expectation and the variance of the number of iterations required for the proposed algorithm to converge are 41,200 and 35.6, respectively, in the underwater sparse scenario when the firing power of the heterogeneous nodes varies between 0 and 15, and 41236 and 49.1, respectively, in the underwater dense scenario when the number of nodes varies between 1 and 20. The stability of the proposed algorithm is thus demonstrated.

To better understand the running time of the proposed algorithm, Figure 8 shows the actual running time of the proposed algorithm on a conventional processor. Specifically, in the underwater sparse scenario, when the transmit power of the heterogeneous nodes varies between 0 and 15, the mathematical expectation and variance of the running time required for the proposed algorithm to converge are 5.65 and 0.51, respectively. In the underwater dense scenario, when the number of nodes varies between 1 and 20, the running time required for the proposed algorithm to converge gradually increases. This is mainly because when the

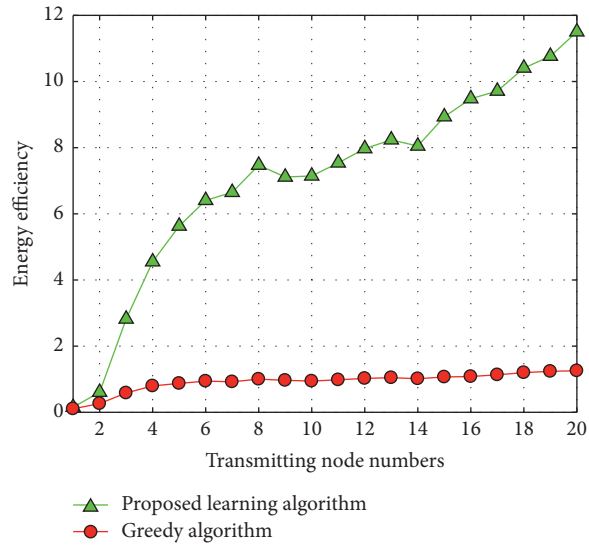

Figure 6: Distribution of energy efficiency as the number of nodes increases.

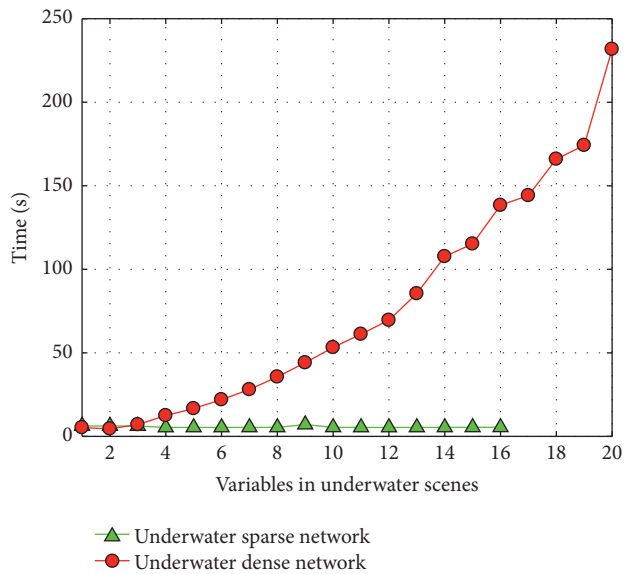

Figure 7: The average number of iterations for the algorithm to converge.

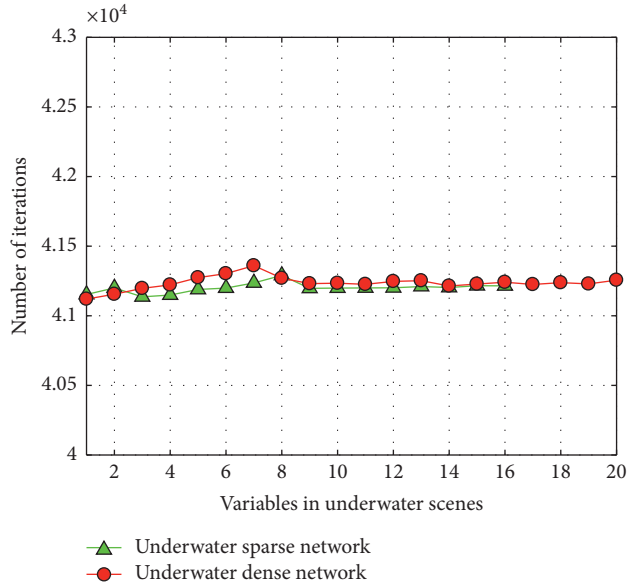

Figure 8: The average running time of algorithm convergence.

number of nodes increases, a lot of time is needed to find the equilibrium between communication capacity and energy consumption.

6. Conclusion

This paper proposes a power allocation scheme based on Q-learning. This scheme considers the interference problem in UACNs composed of multiple transmitter-receiver pairs and the energy efficiency of each transmitter, while each transmitter (sensor node, AUV) is able to train itself to select the appropriate transmit power to support its service nodes while protecting other nodes in the network. In addition, the learning algorithm proposed in this paper, as a distributed method, can solve the power optimization problem for networks with dynamic access of sensor nodes while having low complexity. The scheme is scalable and has a clear advantage in energy efficiency compared to the greedy algorithm. In future work, we design function approximators for neural networks to solve the problem of large state space and action space.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China, under grant no. 62001199, the Fujian Province Natural Science Foundation of China, under grant no. 2019J01842, the President's Fund of Minnan Normal University, under grant no. KJ2020003, the Scientific Research Fund of Fujian Provincial Education Department, under grant no. JT180596, and the Ningde Science and Technology Project, under grant nos. 20140157 and 20160044.

References

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," *Ad Hoc Networks*, vol. 3, no. 3, pp. 257–279, 2005.

[2] T. Archana, P. Rishi, and D. Sanjoy, "Localization schemes for underwater acoustic sensor networks –A review," *Computer Science Review*, vol. 37, pp. 1–18, 2020.

[3] A. Doosti-Aref and A. Ebrahimzadeh, "Adaptive relay selection and power allocation for OFDM cooperative underwater acoustic systems," *IEEE Transactions on Mobile Computing*, vol. 17, no. 1, pp. 1–15, 2018.

[4] G. Zhou, Y. Li, Y. C. He, X. Wang, and M. Yu, "Artificial fish swarm based power allocation algorithm for MIMO-OFDM relay underwater acoustic communication," *IET Communications*, vol. 12, no. 9, pp. 1079–1085, 2018.

[5] H. Wang, Y. Li, and J. Qian, "Self-adaptive resource allocation in underwater acoustic interference channel: a reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2816–2827, 2020.

[6] V. Ravelomanana, "Extremal properties of three-dimensional sensor networks with applications," *IEEE Transactions on Mobile Computing*, vol. 3, no. 3, pp. 246–257, 2004.

[7] Y. Luo, L. Pu, M. Zuba, Z. Peng, and J.-H. Cui, "Challenges and opportunities of underwater cognitive acoustic networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 2, pp. 198–211, 2014.

[8] P. Casari and M. Zorzi, "Protocol design issues in underwater acoustic networks," *Computer Communications*, vol. 34, no. 17, pp. 2013–2025, 2011.

[9] J. Zhang, H. Gharavi, and B. Hu, "Impact of cooperative space-time/frequency diversity in OFDM-based wireless

sensor systems over mobile multipath channels," *IET Wireless Sensor Systems*, vol. 6, no. 4, pp. 138–143, 2016.

[10] S. Han, X. Li, L. Yan, J. Xu, Z. Liu, and X. Guan, "Joint resource allocation in underwater acoustic communication networks: a game-based hierarchical adversarial multiplayer multiarmed bandit algorithm," *Information Sciences*, vol. 454-455, pp. 382–400, 2018.

[11] X. Shen, X. Zhang, Y. Huang, S. Chen, and Y. Wang, "Task learning over multi-day recording via internally rewarded reinforcement learning based brain machine interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 3089–3099, 2020.

[12] C. Choudhuri and U. Mitra, "Capacity bounds and power allocation for underwater acoustic relay channels with ISI," in *Proceedings of the Fourth ACM International Workshop on UnderWater Networks*, New York, NY, USA, November 2009.

[13] K. Nehra and M. Shikh-Bahaei, "Spectral efficiency of adaptive MQAM/OFDM systems with CFO over fading channels," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 1240–1247, 2011.

[14] Y. Luo, L. Pu, H. Mo, Y. Zhu, Z. Peng, and J.-H. Cui, "Receiver-initiated spectrum management for underwater cognitive acoustic network," *IEEE Transactions on Mobile Computing*, vol. 16, no. 1, pp. 198–212, 2017.

[15] J. M. Jornet, M. Stojanovic, and M. Zorzi, "On joint frequency and power allocation in a cross-layer protocol for underwater acoustic networks," *IEEE Journal of Oceanic Engineering*, vol. 35, no. 4, pp. 936–947, 2010.

[16] H. Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 5, pp. 2977–2990, 2021.

[17] X. Qiu, L. Liu, W. Chen, Z. Hong, and Z. Zheng, "Online deep reinforcement learning for computation offloading in blockchain-empowered mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8050–8062, 2019.

[18] L. Xiao, D. Jiang, Y. Chen, W. Su, and Y. Tang, "Reinforcement-learning-based relay mobility and power allocation for underwater sensor networks against jamming," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 3, pp. 1148–1156, 2020.

[19] Y. Su, Y. Zhu, H. Mo, J.-H. Cui, and Z. Jin, "A joint power control and rate adaptation MAC protocol for underwater sensor networks," *Ad Hoc Networks*, vol. 26, pp. 36–49, 2015.

[20] M. A. Chitre, J. R. Potter, and S. H. Ong, "Viterbi decoding of convolutional codes in symmetric α-stable noise," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2230–2233, 2007.

[21] A. Mahmood, M. Chitre, and M. A. Armand, "PSK communication with passband Additive symmetric α-stable noise," *IEEE Transactions on Communications*, vol. 60, no. 10, pp. 2990–3000, 2012.

[22] J. Wang, J. Li, S. Yan et al., "A novel underwater acoustic signal denoising algorithm for Gaussian/non-Gaussian impulsive noise," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 429–445, 2021.

[23] X. Zhang, W. Ying, P. Yang, and M. Sun, "Parameter estimation of underwater impulsive noise with the Class B model," *IET Radar, Sonar & Navigation*, vol. 14, no. 7, pp. 1055–1060, 2020.

[24] X. Zhang, W. W. Ying, W. Ying, and B. Yang, "Parameter estimation for class a modeled ocean ambient noise," *Journal of Engineering and Technological Sciences*, vol. 50, no. 3, pp. 330–345, 2018.

[25] M. Stojanovic, "On the relationship between capacity and distance in an underwater acoustic communication channel," *ACM SIGMOBILE-Mobile Computing and Communications Review*, vol. 11, no. 4, pp. 34–43, 2007.

[26] V. T. Vakily and M. Jannati, "A new method to improve performance of cooperative underwater acoustic wireless sensor networks via frequency controlled transmission based on length of data links," *Wireless Sensor Network*, vol. 2, Article ID 381, 2010.

[27] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[28] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-learning: a novel convergence analysis," *IEEE Transactions on Cybernetics*, vol. 47, no. 5, pp. 1224–1237, 2017.

[29] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, and P. Li, "Energy-efficient scheduling for real-time systems based on deep Q-learning model," *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 132–141, 2019.

[30] B. Wang, Z. Han, and K. J. R. Liu, "Distributed relay selection and power control for multiuser cooperative communication networks using stackelberg game," *IEEE Transactions on Mobile Computing*, vol. 8, no. 7, pp. 975–990, 2009.