remote sensing

Article

Multi-Node Joint Power Allocation Algorithm Based on Hierarchical Game Learning in Underwater Acoustic Sensor Networks

Hui Wang [1,*], Yao Huang [1], Fang Luo [2] and Liejun Yang [2]

[1] School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China
[2] School of Information and Mechanical & Electrical Engineering, Ningde Normal University, Ningde 352000, China
* Correspondence: wh1953@mnnu.edu.cn

Abstract: In order to improve the overall service quality of the network and reduce the level of network interference, power allocation has become one of the research focuses in the field of underwater acoustic communication in recent years. Aiming at the issue of power allocation when channel information is difficult to obtain in complex underwater acoustic communication networks, a completely distributed game learning algorithm is proposed that does not require any prior channel information and direct information exchange between nodes. Specifically, the power allocation problem is constructed as a multi-node multi-armed bandit (MAB) game model. Then, considering nodes as agents and multi-node networks as multi-agent networks, a power allocation algorithm based on a softmax-greedy action selection strategy is proposed. In order to improve the learning efficiency of the agent, reduce the learning cost, and mine the historical reward information, a learning algorithm based on the two-layer hierarchical game learning (HGL) strategy is further proposed. Finally, the simulation results show that the algorithm not only shows good convergence speed and stability but also can adapt to a harsh and complex network environment and has a certain tolerance for incomplete channel information acquisition.

Keywords: underwater acoustic communication; power allocation; hierarchical game learning; multi-armed bandit; distributed

Citation: Wang, H.; Huang, Y.; Yang, L.; Luo, F. Multi-Node Joint Power Allocation Algorithm Based on Hierarchical Game Learning in Underwater Acoustic Sensor Networks. *Remote Sens.* **2022**, *14*, 6215. https://doi.org/10.3390/rs14246215

Academic Editor: Silvia Liberata Ullo

Received: 24 October 2022
Accepted: 5 December 2022
Published: 8 December 2022

1. Introduction

Due to the increasing intensity of military applications and marine resource development, the research on underwater acoustic communication technology has attracted more and more attention [1,2]. In recent years, underwater acoustic communication networks (UACNs) have become more and more widely used in marine data collection, pollution monitoring, disaster prevention, and marine rescue [3–5]. However, because of the characteristics of the underwater environment, there are many disadvantages, such as limited bandwidth, limited energy supply, and prolonged propagation factors, which seriously restrict the development of underwater acoustic communication networks [6–8].

Similar to terrestrial wireless communication, resource allocation is also a key issue for UACNs to solve. Compared with radio communication, resource allocation of UACNs undoubtedly faces more challenges [9,10]. (a) The bandwidth is narrow. The available bandwidth of the underwater acoustic channel is only a few kHz to several tens of kHz; (b) the energy limitation. Most acoustic modems are battery-powered, but in the underwater environment, the replacement and charging of the battery is extremely difficult. From the perspective of overhead, the energy problem of UACNs is more valuable than RF communication on land; (c) long time delay. The propagation speed of sound waves in water is about 1500 m/s [11], which is much lower than that of electromagnetic waves in air, which makes it difficult to obtain real-time channel information in UACNs.

Moreover, only a single transmitter–receiver pair was required in the initial underwater applications, but as application requirements become more complex, multiple devices are required for simultaneous acoustic communication in the same area [12]. When there are many underwater communication users, the communication environment is more crowded, and many users have to conduct communication activities in similar frequency bands, and the communication users will form confrontations and interfere with each other, thus affecting the communication quality of underwater communication users. Therefore, it is still a challenge for UACNs to design a resource allocation method with high efficiency and low overhead for the complex underwater acoustic communication environment.

In recent years, scholars from home and abroad have made some achievements in the research of resource allocation methods. The studies reported in [13,14] regard the power allocation problem as a resource optimization problem. However, the schemes proposed in the above studies are all applied in a centralized manner. The underwater communication environment is complex, and it is difficult to deploy a centralized control center underwater. Therefore, UACNs have strong distributed requirements. Game theory with distributed characteristics is an effective means to solve the problem of multi-user resource allocation in a strong adversarial environment. In [15], game theory is used to maximize the throughput of UACNs, which improves the channel capacity and energy efficiency, but does not consider the problem of interference control between nodes in the network, so the communication service quality is relatively low, and the energy efficiency improvement is limited. In [16], an environment-friendly power allocation method is proposed, which considers the impact of transmission power on marine organisms in the power allocation algorithm based on game theory and reduces the interference of underwater communication equipment on marine organisms.

However, all the above algorithms rely on the assumption that all users can perceive the channel gain information in time. In a complex underwater network environment, channel information and interference information may not be acquired by users in a timely manner. The time delay and time variability in underwater acoustic channels will eventually lead to the result that the feedback information obtained by the transmitter cannot reflect the real-time channel environment, which seriously affects the adaptive communication system's inability to select the transmission power suitable for the real-time communication environment [17]. Therefore, the unknown channel information makes it more difficult to solve the complex multi-user joint resource allocation problem, and traditional optimization methods cannot be easily applied to UACNs.

Reinforcement learning is widely used to solve the problems of policy optimization and policy selection. In other words, communicating nodes in UACNs can interact to adapt to changes in the environment, update optimized policies, and select appropriate actions. In general, reinforcement learning can realize the self-organizing network of UACNs, that is, improve the intelligence and autonomous adaptability of communication nodes, try to minimize human intervention, and perform self-configuration, self-optimization and self-healing [18]. Multi-armed bandit (MAB) is a reinforcement learning decision theory that has shown strong applicability in dealing with problems involving unknown factors. Refs. [19] and [20], respectively, construct the routing problem and spectrum access problem in wireless networks as a MAB decision problem and use MAB theory to solve the game problem containing unknown information. However, the learning time of existing MAB-based wireless network algorithms is relatively long. According to literature statistics, it takes more than 10,000 learning times to reach the expected goal. Due to the limited energy, UACNs are not suitable for processing expensive learning algorithms [21]. Based on this, this study proposes a MAB-based resource allocation algorithm with high learning efficiency for UACNs.

In this work, we study the resource allocation problem under interference constraints for an underwater acoustic communication system consisting of multiple pairs of transmitter and receiver nodes. In the literature, some works use MAB algorithms in reinforcement learning to solve the problem of resource allocation in wireless communication systems (this will be briefly introduced later in Section 2), but the premise is that nodes must know all perfect channel gain and other policy information for the node. It is well-known that these assumptions will lead to a large amount of information exchange, and it is difficult to obtain information in real-time in complex underwater acoustic communication environments with time delays. Based on this, this paper studies the joint resource allocation problem of multi-users in the case of unknown channel information in UACNs and proposes a distributed, low-complexity, high-learning-efficiency learning game algorithm that does not rely on any prior channel information. The main contributions of this work include the following three aspects:

- We construct a multi-agent MAB game model to characterize the joint resource allocation problem. To be more specific, the node is regarded as an agent, and the multi-node network is regarded as a multi-agent network, the multi-agent action space and reward function are constructed, and the optimal response strategy of each node is solved.
- We propose a learning algorithm based on a Greedy-Softmax action selection strategy to solve this game problem. To be more specific, the proposed learning algorithm can be performed in a fully distributed manner, where nodes only need to record their own local information, without any prior information and direct information exchange, thus reducing the cost of resource allocation.
- Furthermore, in order to improve the learning efficiency and reduce the learning cost, we propose a learning algorithm based on the two-layer HGL strategy. To be more specific, the historical reward information is named virtual learning information, and the introduction of virtual learning information into the algorithm can enrich the learning information of players, thereby improving their learning ability.

The rest of this paper is organized as follows. Section 2 reviews the related work, and Section 3 expounds the UACNs model and the construction method of the multi-node MAB game model. Section 4 demonstrates the effectiveness of the constructed model. The multi-node MAB game learning algorithm is analyzed in Section 5. Section 6 verifies the effectiveness of the designed scheme from multiple perspectives, and Section 7 presents the conclusion.

2. Related Work

Power allocation plays an important role in underwater acoustic communication networks. In [22,23], the resource allocation problem in the wireless network is constructed as a goal optimization problem, and the traditional optimization algorithm is used to solve the goal. However, traditional optimization algorithms are centralized optimization algorithms. When applied in actual wireless networks, a central information processor is a must, yet it is difficult to build a CPU in an underwater environment. The centralized algorithm requires a very large information exchange overhead, which can not be borne by underwater acoustic communication.

In [15], a MAC protocol is proposed, and a distributed method is adopted to adjust the transmission power to maximize the throughput of the underwater communication network. In [13], a distributed power allocation algorithm is proposed, which considers power allocation schemes of different available power levels for different network densities to reduce energy consumption. Although the power allocation methods in the above literature are distributed, there must be an assumption that the channel information is ideal. The underwater acoustic channel is extremely complex. In different environments, the state information of the channel will change at any time. The underwater communication environment is constantly changing due to the influence of ocean currents and other factors. Communication nodes will move with the current, and the channel state information is constantly changing. The channel state information has strong uncertainty [24,25]. However, the performance of the algorithm based on ideal channel information will inevitably decline when it is applied to the underwater acoustic communication network with strong uncertainty. Therefore, it is necessary to propose a robust optimization algorithm to reduce the impact of uncertainties on resource allocation.

The nodes in the underwater acoustic communication network are independent individuals, which are very rational and selfish [26]. Game theory means that both sides of the game can maximize their own benefits in the game and can also play a great role in the resource allocation problem of wireless sensor networks. Based on the game theory, in [27], an adaptive distributed power allocation scheme is proposed to solve the power allocation problem of multiple nodes. A distributed game power allocation algorithm considering node residual energy is proposed in [28] to improve the channel capacity of network communication. Meanwhile, a robust game spectrum allocation algorithm is proposed in [29] to maximize the utility function of nodes and improve the communication quality of nodes. Although the game theory algorithm proposed in the above literature can solve the distributed optimization problem, it requires accurate channel state information for resource allocation.

When the underwater communication environment is greatly affected by ship, sunlight and other factors, the underwater acoustic communication environment is harsh, the channel state information is extremely difficult to determine, and the channel state information and interference information cannot be directly obtained. On the one hand, the traditional optimization methods can not be effectively applied to the actual communication scenarios where users vary. On the other hand, the traditional learning methods have too little learning information, resulting in very low learning efficiency. Therefore, efficient and adaptable reinforcement learning algorithms are urgently needed to solve the joint resource allocation problem in complex communication networks [30,31]. Next, we further outline the resource allocation method based on MAB theory.

In recent years, MAB has been gradually applied to solve various resource allocation problems in complex cognitive radio communication environments [32,33]. In terms of power allocation, a low-cost UCB1-based water-filling algorithm is proposed in [34] for single-user multi-channel models; in [35], a joint resource allocation method based on MAB theory is proposed for multi-channel multi-user networks. However, the problem of interference control between users in the network is not considered in its model. The above algorithms based on MAB theory all effectively overcome the unknown prior information of the channel and improve the performance of the network. However, the existing MAB-based algorithms have very long learning times due to their unique learning methods and less learning information. In general, it takes up to 10,000 learning times to reach the desired policy [36]. Undoubtedly, for a wireless network with heavy traffic, such a large learning time is unacceptable.

Inspired by the above research, this work applies the idea of MAB to the resource allocation problem of UACNs and proposes a learning algorithm based on a two-layer HGL policy to implement interference control among communication nodes in the network, so as to improve the overall service quality of the network and reduce the level of network interference.

3. System Model

3.1. UACNs Model

The UACNs model considered in this work is shown in Figure 1. The model consists of N transmitting nodes S_i, $i \in N$ and M receiving nodes R_j, $j \in M$. During the communication process, node R_j receives the signal from node S_i and forwards it to the surface base station. When various nodes in the network reuse the same frequency band, cross-layer interference and same-layer interference between different nodes inevitably occur. For this reason, this work focuses on the power allocation problem in UACNs under the interference state.

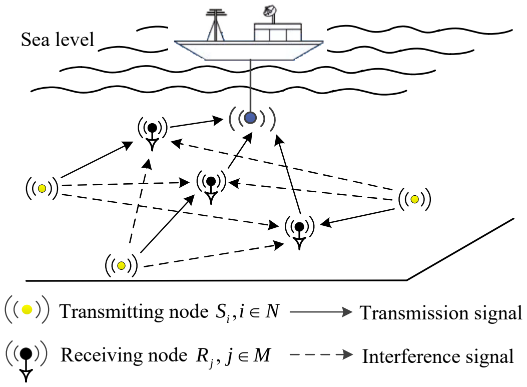

Sea level
Transmitting node $S_i, i \in N$
Transmission signal
Receiving node $R_j, j \in M$
Interference signal

Figure 1. UACNs model.

In UASNs, the signal-to-interference and noise ratio (SINR) received by the receiving node R_j can be expressed as [15]

$$\gamma_j = \frac{p_j h_{jj}}{\sum_{k\neq j,k=1}^{N} p_k h_{kj} + \sigma^2}, \tag{1}$$

where p_j denotes the transmit power of the node S_j, p_k denotes the power strategy of other transmitting nodes except node S_j, h_{kj} denotes the channel gain between the transmitting node S_k and the receiving node R_j. Further, $\sum_{k\neq j,k=1}^{N} p_k h_{kj}$ represents the interference caused by other transmitting nodes using the same frequency channel to the receiving node R_j. In addition, the channel gain in Equation (1) is expressed as [37]

$$h = A_0^{-1} d^{-sp} (\alpha(f))^{-d}, \tag{2}$$

where A_0 is the normalization coefficient, d denotes the transmission distance (km), f denotes the communication frequency (Hz), d^{-sp} is the spread loss, which describes the set characteristics of the transmission, sp denotes the expansion coefficient, which is 1.5, and $\alpha(f)$ denotes the absorption coefficient, which can be expressed by Thorp's empirical formula as [12]

$$10\alpha(f) = \frac{0.11 f^2}{1+f^2} + \frac{44 f^2}{4100+f^2} + 2.75 \times 10^{-4} f^2 + 0.003. \tag{3}$$

Furthermore, generally, the noise power of the underwater acoustic channels should be assumed to be σ^2 [13]. According to the Shannon theorem formula, if the channel parameters of all nodes are known, the channel capacity of the j channel link can be obtained as

$$C_j = \frac{B}{2} \log_2(1+\gamma_j), j = 1, 2, \cdots, M, \tag{4}$$

where B represents the channel bandwidth.

3.2. Multi-Node MAB Game Model

In the underwater transmission process, when the noise condition of the underwater acoustic channel is given, each transmitter hopes to send data with a higher power to obtain a better quality of service. If the interference between nodes is large, the excessive interference will cause nodes to increase transmission power, which will lead to more serious interference between users and increase the level of network interference, thereby reducing the quality of service for users. To this end, it is necessary to find a balance point between user service quality and network interference level so as to achieve optimal system performance.

This work uses game theory to construct a multi-user competition. In general, a strategic game $G = \left\{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{.})\}_{i\in N}\right\}$ consists of three parts. The set of agents is N transmitting nodes, the transmitting power $p_i \in \theta_i, i \in N$ of the transmitting node is the strategy of node i, and θ_i represents the strategy set of the agent. $U_i, i \in N$ is the utility of node i, which reflects the satisfaction of node i with the service obtained.

It should be noted that most studies assume that nodes can perceive the influence of external environmental factors on it in real-time [38,39]. After obtaining the external determinants, the agent can obtain the best response strategy according to the traditional optimization method. However, in the complex underwater acoustic environment, real channel perception is difficult to achieve, and the realization of other assumptions will also cause excessive information exchange, making the algorithm unable to cope with the network delay.

For each agent, the power allocation problem can be regarded as a MAB problem. In the process of solving the optimal response strategy, there is no need for any information exchange between agents. Users only learn the environment (the impact of the external environment on their utility) by analyzing their own game history data, so as to obtain the best response strategy. If all agents in the game solve their own optimal response strategies according to the above learning method, then the game can be regarded as a multi-agent MAB game problem.

In the multi-node MAB game problem in this work, the strategy of node (agent) $i \in N$ is the power allocation strategy p_i. Assuming that there are z feasible strategies for node $i \in N$, then its feasible strategy set is $P_i = \left\{p_{i,1}, p_{i,2}, \cdots, p_{i,z_i}\right\}$. In addition, the reward in the MAB problem is the utility in the game problem. Obviously, the reward of any node depends not only on the strategy chosen by itself, but also by external factors, such as the strategies of other agents and environmental noise.

4. Problem Formulation

In the multi-agent MAB game problem, the goal of each agent is to maximize its own utility (return value). In [15], the utility function constructed for agent $i \in N$ is described as follows

$$U_i(p_i) = B\log_2\left(1 + \frac{p_i h_{ii}}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \sigma^2}\right) - \alpha_i p_i, \tag{5}$$

where α_i denotes the price factor, which is set to p_{max}, which represents the maximum transmit power of the node. Obviously, the benefit part of the utility function described by Equation (5) is the node channel capacity, and the cost part is the product of the unit power price and the node transmit power. However, this utility function does not take the level of interference in the network into consideration. Since the optimal transmit powers of nodes all affect each other, if only its own power cost is considered in the utility function, while the mutual interference between nodes is not considered, then each transmitting node will try to transmit data at a higher power level. This will not only make the interference in the network more serious and reduce the quality of network service but also shorten the node's survival time.

To address the shortcomings of the above algorithm and to improve the quality of network services, this study makes the following improvements to the utility function:

$$U_i(p_i) = B\log_2\left(1 + \frac{p_i h_{ii}}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \sigma^2}\right) - \phi_i\left(\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \sigma^2\right) p_i, \tag{6}$$

where the first term represents the channel capacity when the node power is p_i and the second term is the network interference level penalty term. $\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \sigma^2$ indicates that node S_i is subject to interference from other transmitting nodes and environmental noise. ϕ_i is the adjustment coefficient. Through the above analysis, in the multi-agent MAB game problem in this work, the optimization problem for each agent i can be expressed as

$$\max_{p_i} U_i(p_i) \tag{7}$$

$$s.t. \;\; p_i \in \theta_i, \forall i \in N \tag{8}$$

$$0 \le p_i \le p_{\max}, \forall i \in N \tag{9}$$

$$p_i \in \arg\max_{p_i} \{U_j : p_i \in \theta_i\}, \forall i \in N. \tag{10}$$

The purpose of the game is to seek an optimal combination of strategies, i.e., a Nash equilibrium, which can make the strategy of each agent be the optimal response to the strategies of other agents [40]. Next, the existence and uniqueness of the newly constructed game model Nash equilibrium will be proved.

Definition 1 (Nash Equilibrium [41])**.** *Let $s = (b_1^*, b_2^*, \cdots, b_N^*)$ be a strategy combination of a game $G = \{N, \{\theta_i\}_{i\in N}, \{U_i\}_{i\in N}\}$, if all $b_i \in \theta_i$ is true for each player i: $U_i(b_i^*, b_{-i}^*) \ge U_i(b_i, b_{-i}^*)$, then strategy combination $s = (b_1^*, b_2^*, \cdots, b_N^*)$ is a Nash equilibrium of the game. b_i represents the strategy chosen by player i that is different from b_i^*, and b_{-i}^* represents the combination chosen by all participants except i in all agents.*

At the Nash equilibrium point, the power value of other users is constant, and no user can improve its utility by simply changing its power value. The Nash equilibrium point is an equilibrium point obtained in the competition that each participant is unwilling to deviate from. The following proves the existence and uniqueness of the Nash equilibrium point in the MAB game model [15,40,41].

Theorem 1. *There is a Nash Equilibrium in $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{\cdot})\}_{i\in N}\}$.*

Proof. According to Nash's theorem, $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{\cdot})\}_{i\in N}\}$ has a Nash equilibrium if the following conditions are met.

(a) θ_i is a non-empty, closed and bounded convex set of Euclidean space R^N;
(b) $U_i(p_i)$ is continuous on $[p_{\min}, p_{\max}]$ and quasi-concave on p_i.

Since the policy space θ_i of each node is defined in $[p_{\min}, p_{\max}]$ $(0 \le p_{min} \le p_{\max})$, the first condition is obviously satisfied. For condition (b), it is obvious that $U_i(p_i)$ is continuous on θ_i, so it is only necessary to prove that $U_i(p_i)$ is quasi-concave on p_i.

The first-order partial derivative of the utility function $U_i(p_i)$ with respect to p_i is:

$$\frac{\partial U_i(p_i)}{\partial p_i} = \frac{B}{\ln 2}\frac{h_{ii}}{\sum_{j=1,j\ne i}^{N} p_j h_{ji} + \delta^2 + p_i h_{ii}} - \phi_i\left(\sum_{j=1,j\ne i}^{N} p_j h_{ji} + \delta^2\right). \tag{11}$$

The second-order partial derivative of the utility function $U_i(p_i)$ with respect to p_i is:

$$\frac{\partial^2 u_i(p_i)}{\partial p_i^2} = \frac{-B}{\ln 2}\frac{h_{ii}^2}{\left(\sum_{j=1,j\ne i}^{N} p_j h_{ji} + \delta^2 + p_i h_{ii}\right)^2} < 0. \tag{12}$$

Since $\partial^2 u_i(p_i)/\partial p_i^2 < 0$, $U_i(p_i)$ is concave in p_i, a concave function is also a proposed concave function, so there is a Nash equilibrium in $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{\cdot})\}_{i\in N}\}$.

The optimal solution of $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{\cdot})\}_{i\in N}\}$ is $\arg\max_{p_i\in\theta_i} U_i(p_i)$. For continuously differentiable functions, the necessary condition for first-order optimization is $\partial u_i(p_i, p_{-i})/\partial p_i = 0$, which can be obtained from Equation (11)

$$p_i = \left[\frac{B}{\phi_i \ln 2} \frac{1}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2} - \frac{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2}{h_{ii}} \right]^{+}. \tag{13}$$

Therefore, Equation (13) is the pricing function of the power policy p_i, where $[\centerdot]^{+} = \max(\centerdot, 0)$ indicates that the transmission power is non-negative. Further, the value range of ϕ_i can be obtained

$$\phi_i \leq \frac{\ln 2}{B h_{ii}}. \tag{14}$$

That is, if the price per unit of power exceeds this range, no user can benefit from the system. The optimal solution of $G = \left\{N, \{\theta_i\}_{i\in N}, \{U_i(\centerdot)\}_{i\in N}\right\}$ is

$$p_i^{*} = \begin{cases} p_{\min} & p_i \leq p_{\min} \\ \left[\frac{B}{\phi_i \ln 2} \frac{1}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2} - \frac{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2}{h_{ii}} \right]^{+} & p_{\max} \geq p_i \geq p_{\min} \\ p_{\max} & p_i \geq p_{\max}. \end{cases} \tag{15}$$

□

Theorem 2. *The Nash equilibrium of* $G = \left\{N, \{\theta_i\}_{i\in N}, \{U_i(\centerdot)\}_{i\in N}\right\}$ *is unique.*

Proof. According to Theorem 2, there is a Nash equilibrium in $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\centerdot)\}_{i\in N}\}$, and its Nash equilibrium solution is assumed to be $\vec{P}$. According to Equation (15), the interference equation is $\vec{P} = I(p)$, where $I(p) = (I_1(p), I_2(p), \cdots, I_N(p))$. The interference equation $I(p)$ is said to be standard if the following properties are satisfied for all non-negative power vectors, (a) positive, $I(p) > 0$, (b) monotonic, (c) scalability, $\forall \alpha > 1, \alpha I(P) \geq I(\alpha P)$. The standard equation converges to a unique point. Therefore, to prove the uniqueness of the Nash equilibrium, it is only necessary to prove that $I(p)$ is a standard function, that is, it satisfies positivity, monotonicity and measurability.

Positivity. According to the value range of the pricing function, it can be guaranteed that there must be a node whose power level $p_i > 0$, then $I(p) > 0$.

Monotonicity. For $\forall i \in N$, let $p_i \geq p'_i$,

$$I_i(p_j) - I_i\left(p'_j\right) = \tag{16}$$

$$\frac{B}{\phi_i \ln 2} \sum_{j=1,j\neq i}^{N} \left(p'_j - p_j\right) h_{ji} \left(\frac{1}{\left(\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2\right)\left(\sum_{j=1,j\neq i}^{N} p'_j h_{ji} + \delta^2\right)} + \frac{1}{h_{ii}} \right) \leq 0.$$

It can be seen from Equation (16) that when $p_i \geq p'_i$, $I_i(p_j) - I_i\left(p'_j\right) \leq 0$. Therefore, $I(p)$ is a single decreasing function, and it takes the equal sign when $p_i = p'_i$.

Scalability.

$$\alpha I(p_i) - I(\alpha p_i) = (\alpha - 1)\left[\frac{B}{\phi_i \ln 2} \frac{1}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2} - \frac{\delta^2}{h_{ii}} \right]. \tag{17}$$

For the above formula, we only need to prove

$$\frac{B}{\phi_i \ln 2} \frac{1}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2} \geq \frac{\delta^2}{h_{ii}}. \tag{18}$$

when discussing $p_i \geq 0$ in the previous section, we obtain

$$\frac{B}{\phi_i \ln 2} \frac{1}{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2} \geq \frac{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2}{h_{ii}}. \tag{19}$$

From the positive condition, we can obtain $\sum_{j=1,j\neq i}^{N} p_j h_{ji} > 0$, and it is easy to obtain $\frac{\sum_{j=1,j\neq i}^{N} p_j h_{ji} + \delta^2}{h_{ii}} \geq \frac{\delta^2}{h_{ii}}$. Hence $\alpha I(p_i) - I(\alpha p_i) \geq 0$.

In summary, the interference equation $I(p)$ is a standard function, so $G = \{N, \{\theta_i\}_{i\in N}, \{U_i(\boldsymbol{\cdot})\}_{i\in N}\}$ has a unique Nash equilibrium solution. □

5. Multi-Node MAB Game Learning Algorithm

In a complex underwater environment, it is difficult for nodes to perceive channel gain information and other node state information in real-time. When there is little prior information, the game algorithm based on MAB theory enables nodes to learn the environment only by analyzing historical data, and then obtain the best response strategy without any information exchange.

5.1. Action Selection Strategy of Softmax-Greedy

In terms of action selection, we choose the softmax-greedy policy to overcome the shortcomings of greedy policy and ϵ-greedy policy. Specifically, the greedy policy selects an action under each slot to maximize the Q-value, which may achieve better performance with a limited number of actions. However, the use of this scheme may introduce incomplete exploration problems. Since it has no chance to explore other actions, the greedy scheme may not be able to explore the remaining power interval. In addition, the ϵ-greedy strategy adopts the probability ϵ to weigh the proportion of “exploration” and “exploitation” based on the greedy strategy. The agent uses the greedy strategy to select actions with a probability of 1-ϵ, that is, the action that can obtain the maximum action value function is transmitted to the receiver in each time slot; otherwise, the agent randomly selects an action from the selectable action set. However, in the ϵ-greedy strategy, there is an obvious shortcoming that when the probability is less than ϵ, the greedy scheme chooses evenly among all actions, so the probability of the best action choice may be the same as the probability of the worst action choice.

To address the problems in ϵ-greedy policy, a softmax-based action selection scheme is proposed in [41], which selects actions through exploration and utilization stages. The action probability assignment in the softmax strategy is based on the Boltzmann distribution and can be described as

$$\pi(a_j) = \frac{\exp\left(Q^{(k)}(a_i)/\tau\right)}{\sum_{j=1}^{M} \exp\left(Q^{(k)}(a_j)/\tau\right)}, \tag{20}$$

where M is the number of available action sets, $\tau > 0$ is a parameter called temperature, and $Q^{(k)}(a_i)$ represents the estimated value corresponding to action i at time k. Compared with ϵ-greedy, the softmax strategy is more “cautious” when choosing actions, and it needs to calculate the probability of each optional action before it is selected. To a certain extent, the softmax strategy will not miss any valuable actions. However, this strategy has certain limitations. Since there are a lot of calculations before each action is selected, especially in models with large state and action dimensional spaces, these large amounts of calculations will greatly affect the convergence speed of the algorithm.

In this work, we propose an adaptive action selection strategy that combines softmax and a greedy action selection strategy. In the exploration phase, the adaptive scheme selects an action and selects the action with the largest Q value with a probability of $1-\epsilon$; otherwise, it selects the action a_j with the selection probability $\pi(a_j)$. This scheme balances exploration and exploitation through ϵ and τ, which avoids the incomplete exploration problem in the greedy strategy and converges faster than the softmax strategy. The adaptive action selection scheme can be expressed as

$$\pi(\epsilon, \tau, a_i) = \begin{cases} \text{softmax strategy} & \Delta \leq \varepsilon \\ \arg\max_{a_t\in A} Q^{(k)}(a_t) & \mathit{Otherwise}, \end{cases} \tag{21}$$

where $0 < \Delta < 1$ is a uniform random number generated in each time slot, and $0 < \epsilon < 1$ is a certain value. In addition, it is known from (20) and (21) that when the temperature τ is high, all action probabilities are equal. On the contrary, when the temperature τ is very low, since the agent often chooses the action that can obtain the maximum probability, the softmax strategy case will become the same as the greedy strategy. To sum up, the execution process of game learning based on softmax-greedy is as described in Algorithm 1.

Algorithm 1 Power allocation algorithm based on softmax-greedy

```
Input: Location information of transmitting node $S_i$, $i \in N$, base station and receiving
    node $R_i$, $j \in M$, policy space $\theta_i, i \in N$, learning times $T$, exploration probability $\epsilon$;
Output: Optimal power distribution scheme $\vec{P} = [p_1, p_2, \cdots, p_N]$
Initialization: $r$ = 0, For all actions $a \in P_i$, let $Q(a) = 0, Count(a) = 0$;
for $k = 1 \longrightarrow T$ do
    if $rand() \leq \epsilon$ then
        Select action $a$ by executing the softmax policy via (20);
    else
        Select action $a$ by executing a greedy policy, i.e., $a = \arg\max_{a_t \in A} Q^{(k)}(a_t)$;
    end if
    Calculate the utility $r^{(k)} = U_i(a)$ via (6);
    Update $Q(a)$ value
        $$Q(a) \leftarrow \frac{Q(a) \times Count(a) + r^{(k)}}{Count(a) + 1} \quad (22)$$
    Update count value
        $$Count(a) \leftarrow Count(a) + 1 \quad (23)$$
end for
```

5.2. Hierarchical Optimal Feedback Strategy

In the softmax-greedy policy selection mechanism, agents (nodes) do not need any information exchange or even know any channel information. The node only needs to record its own local information, the historical strategy actually executed and the corresponding utility value. Therefore, the softmax-greedy strategy is a fully distributed algorithm. However, in one learning, an agent can only choose one to perform, that is, only one set of learning samples can be obtained in one learning. This will lead to poor learning ability of Algorithm 1, and the required learning times are too long.

From the definition of the Nash equilibrium, how much the node obtains is not only affected by the node's own strategy but also external interference, such as the strategy of other jammers and environmental noise. The node can calculate the external interference $v_i^{(k)}$ at the last moment through the channel capacity obtained at the last moment, which can be expressed as

$$v_i^{(k)} = \frac{p_i^{(k)} h_i^{(k)}}{2^{C_i^{(k)}} - 1}. \tag{24}$$

This actually only requires the transmitter to transmit a pilot signal with very weak power, and then obtain outdated interference information through Equation (1). Given the interference information $v_i^{(k)}$ at the past time n, substituting other strategies into

Equation (6) can obtain the utility of the node at the past time n if the node executes other strategies. In this work, the calculated utility information is called virtual learning information, the introduction of which can enrich the learning information, thereby improving the learning ability of the agent. Based on this, a hierarchical game learning algorithm is further proposed, as shown in Algorithm 2.

Algorithm 2 Power allocation algorithm based on hierarchical game learning (HGL)

```
Input: Location information of transmitting node $S_i$, $i \in N$, base station and receiving
    node $R_i$, $j \in M$, policy space $\theta_i, i \in N$, learning times $T$, exploration probability $\epsilon$;
Output: Optimal power distribution scheme $\vec{P} = [p_1, p_2, \cdots, p_N]$
Initialization: $r = 0$, For all actions $a \in P_i$, let $Q(a) = 0, Count(a) = 0$;
for $k = 1 \longrightarrow T$ do
    if $rand() \leq \epsilon$ then
        Select action $a$ by executing the softmax policy via (20);
    else
        Select action $a$ by executing a greedy policy, i.e., $a = \arg\max_{a_t \in A} Q^{(k)}(a_t)$;
    end if
    Calculate the utility $r^{(k)} = U_i(a)$ via (6);
    Calculate $v_i^{(t)}(a)$ via (4)
    Search for the policy $a'$ that maximizes its utility at time $k$.
        $\max U_i\left(p_i^{(k)}, p_{-i}^{(k)}\right) = C_i\left(v_i^{(k)}(a), p_i^{(k)}\right) - \beta\left(\sum_{j=1,j\neq i}^{N} \lambda_j p_j h_{ji} + \sum_{j=1,j\neq i}^{N} \lambda_j p_i h_{ij}\right)$   (25)
    Update $Q$ value
        $Q(a) \leftarrow \frac{Q(a) \times Count(a) + r^{(k)}}{Count(a) + 1}$   (26)
    Update count value
        $Count(a) \leftarrow Count(a) + 1$   (27)
end for
```

6. Scheme Analysis and Evaluation

This section verifies the superiority of the proposed algorithm by comparing five learning strategies (HGL, Softmax-greedy, ϵ-greedy, Greedy, Random). In the simulation, we assume that there are six nodes in which three transmitting nodes and three receiving nodes are randomly distributed in an area 2 km deep, 3 km long and 2 km wide. For the above scenario, we will analyze the following four aspects: a. the impact of different parameters ϵ in softmax-greedy on the algorithm. b. Comparison of four strategies such as softmax-greedy and ϵ–greedy. c. Comparison of improved HGL and softmax-greedy in terms of convergence and SINR. d. The proposed HGL is compared with the other four learning strategies in dynamic environments. Note that the size of the underwater target area is not limited when considering a dynamic environment. As shown in Figure 2, the circle indicates the transmitting node, the diamond indicates the receiving node, and the coordinate information of the nodes is shown in Table 1. During the operation of the network, the transmitting node S_i and the corresponding receiving node R_i perform information transmission. Considering the random and non-stationary characteristics of underwater signals, the influence of underwater uncertain factors on the underwater acoustic channel is reflected by the variable δ, where $\delta = h \times \gamma$, and γ obeys the Rayleigh distribution with a mean of 0.1. Therefore, in the simulation, the underwater acoustic channel gain adopts $h + \delta$. Meanwhile, set the carrier frequency to be 20 kHz, the propagation coefficient to be 1.5, the signal bandwidth to be 10 kHz, the gain per unit rate to be 0.05, and the maximum transmit power of the node to be 10 W. In addition, in order to facilitate the display of experimental results, we add a coefficient τ before the first term of Equation (6), so that $B\tau$ is equal to ln2, and at the same time, ϕ_i takes the value 0.01.

Table 1. The coordinate of transceiver nodes.

	S_1	S_2	S_3	R_1	R_2	R_3
x/km	0.10	0.50	0.70	0.50	1.00	1.20
y/km	0.10	0.10	1.1	0.50	0.50	1.00
z/km	0.10	0.11	0.10	0.12	0.10	0.10

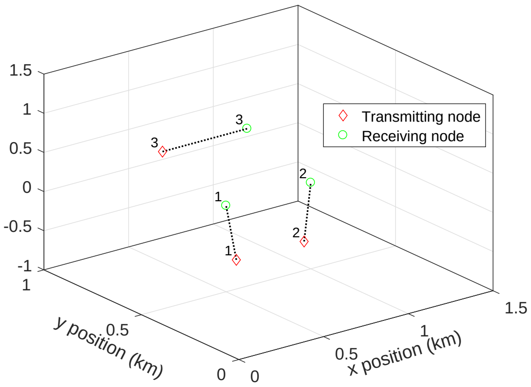


Figure 2. Distribution of communication nodes.

Figure 3 describes the impact of different ϵ on the softmax-greedy algorithm convergence when softmax-greedy action selection strategy is adopted. In this section, the values of ϵ are taken as 0.1, 0.01 and 0.001, respectively. As can be seen from the above discussion, the MAB algorithm has two stages: exploration and utilization. When the agent is in an unfamiliar environment, it needs to explore the environment more, and agents can change the proportion of exploration and exploitation by adjusting ϵ. In the exploration phase, the softmax-greedy strategy estimates the Q value corresponding to each action. From Figure 3, we can observe that within a reasonable range of values, the smaller the ϵ, the slower the convergence speed of the algorithm. The reason is that when the value of ϵ is small, the agent has a greater probability of selecting an action through the Greedy strategy, that is, it tends to select the power with the largest Q value. This ignores other actions with more value, resulting in incomplete exploration. In order to ensure the convergence of the algorithm, we make ϵ = 0.1 in the next simulation experiments.

Further, this section verifies the superiority of the softmax-greedy algorithm by comparing the performance of four learning algorithms (softmax-greedy, ϵ-greedy, Greedy, Random). Figure 4 compares the performance of four algorithms, softmax-greedy, ϵ-greedy, greedy and random, demonstrating the effectiveness of the proposed algorithm. The key to learning algorithms is to maintain an explore–exploit balance. The greedy strategy only emphasizes "exploitation", that is, the agent tends to choose the strategy that seems to have the greatest utility at present, while the random strategy only emphasizes "exploration", which always chooses an action randomly and uniformly with the same probability. The ϵ-greedy strategy balances "exploration" and "exploitation" with the help of ϵ, but during "exploitation", it still randomly selects an action with the same small probability ($P = \varepsilon/|A|$). In contrast, the softmax-greedy strategy will make the action with more utility have a greater probability of being selected, i.e., not forgetting to take advantage of it during exploration. Therefore, from Figure 4, we can observe that in terms of stability, greedy > softmax-greedy > ϵ-greedy > random, and in terms of performance, ϵ-greedy > softmax-greedy > random > greedy. Overall, the softmax-greedy strategy has the best overall performance.

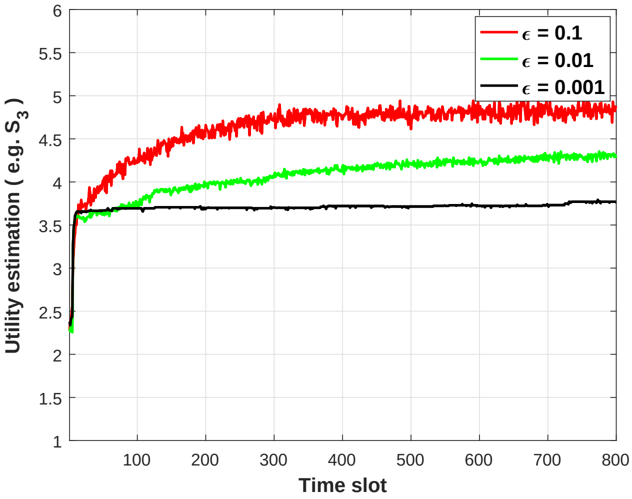


Figure 3. Influence of parameter ϵ on the convergence of softmax-greedy action selection strategy.

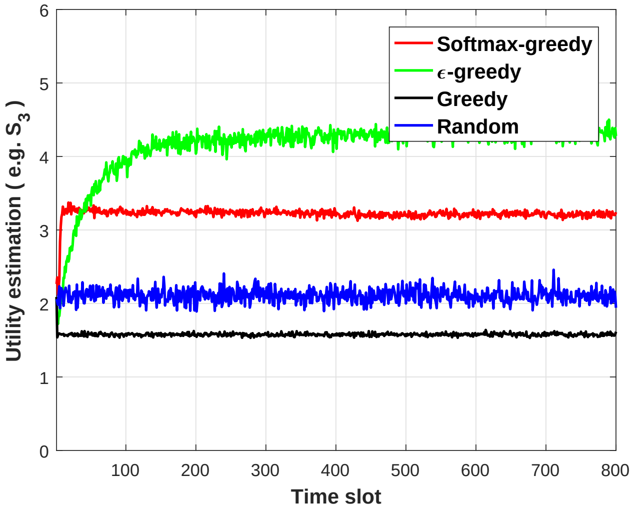


Figure 4. Comparison of utility values under different power control strategies.

It can be seen from Figure 4 that softmax-greedy is slightly lower than ϵ-greedy in terms of performance, mainly because the softmax-greedy strategy is more inclined to select actions with higher returns in exploration, and a very small amount of learning information limits the degree of exploration. Subsequently, Figure 5 shows the variation of the utility of the three nodes with time slots under the guidance of the softmax-greedy strategy and HGL strategy, respectively. Obviously, compared with the softmax-greedy strategy, the HGL strategy can make the three nodes quickly reach the convergence state after a short fluctuation. Therefore, based on the softmax-greedy strategy, the proposed HGL strategy can enrich the learning information of the agent by mining historical reward information, improve the learning efficiency and reduce the learning cost of the algorithm.

In order to verify the effectiveness of the HGL strategy, we further analyze the strategy probability distribution of node S_3 under the guidance of the two strategies. It can be seen from Figure 6 that in the initial stage, the probability of each strategy being selected is equal, which is 0.2. As the number of learning increases, the expected utility $Q(a)$ corresponding to each action will gradually change. Subsequently, the Boltzmann distribution is used as a random strategy to increase the probability of action with large $Q(a)$ being selected and, conversely, to decrease the probability. It can be seen from Figure 6 that the convergence speed and effect of the HGL strategy are better than those of the softmax-greedy strategy. In the HGL strategy, each discrete strategy can converge to a pure strategy after less iteration, while the softmax-greedy strategy can only converge to a mixed strategy. This is mainly because the number of learning times a node can obtain in one learning of the two-layer HGL strategy is one more than that of the single-layer learning algorithm. More importantly, the added virtual learning information enriches the learning information of nodes, and these virtual pieces of information are optimized information based on the actual execution information, so the two-layer HGL strategy can significantly improve the learning efficiency and reduce the number of learning times.

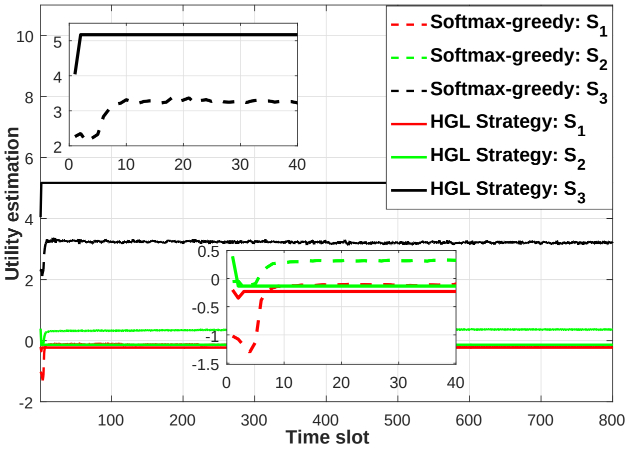
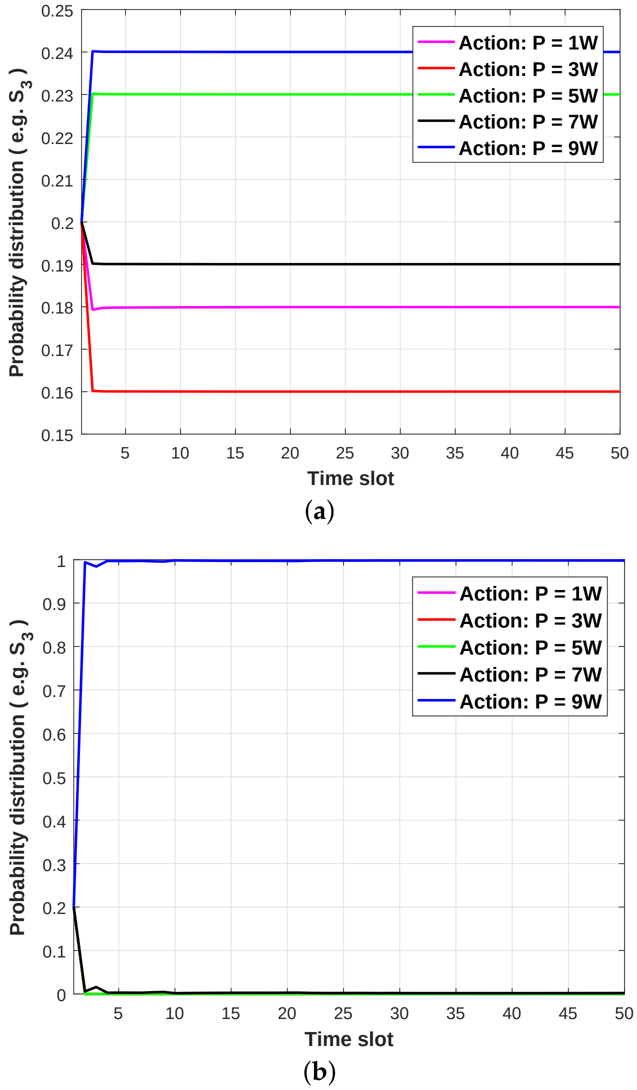


Figure 5. Convergence comparison between softmax-greedy and HGL.

In Figure 7, we compare the SINR performance of the HGL strategy and other strategies under different time slots. We can observe that the two-layer HGL strategy can obtain the best SINR performance. This is due to the efficient hierarchical learning method that greatly enhances the learning ability of the HGL learning algorithm. Specifically, in the lower-level learning, historical decision information is constructed as virtual information, and the agent learns the virtual information and mines the environmental information, thereby improving the learning efficiency.

Next, we show how the performance of five strategies varies with distance in a dynamic water environment. Assuming that the transmitter-receiver pair 1 and 3 are in the same position in the target waters, the transmitter-receiver pair 2 moves in a certain direction. In other words, the distance between the transmitter-receiver pair 2 and the transmitter-receiver pair 1 and 3 is becoming farther and farther. Figures 8 and 9 show the effect of distance d_j on the performance of the proposed scheme, d_j = [0.35, 0.5, 0.65, . . . , 2.3] km, where d_j represents the distance from transmitting node S_2 to transmitting node S_1. Two conclusions can be drawn from Figure 8. On the one hand, the utility of the node increases with the increase in d_j. This is mainly because with the increase in d_j, the interference suffered by node R_2 decreases continuously. On the other hand, the HGL strategy improves the adaptive ability of node S_2, that is, it can quickly adjust the power according to the changes in the environment. When d_j is less than 7 km, node S_2 suffers a lot of network interference and has a low utility; on the contrary, when d_j is greater than 7 km, node S_2 has a high utility. Similarly, it can be seen from Figure 9 that the SINR of the signal increases with the increase in d_j. It is further confirmed that the proposed algorithm has strong adaptive ability and can adjust the transmit power of the node according to the changes in the environment.



(**a**)

(**b**)

Figure 6. Policy probability distribution of node S_3. (**a**) Softmax-greedy strategy. (**b**) HGL strategy.

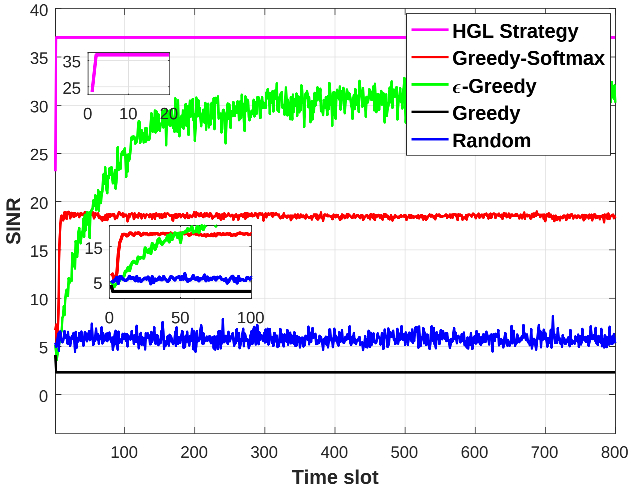


Figure 7. SINR comparison under different strategies.

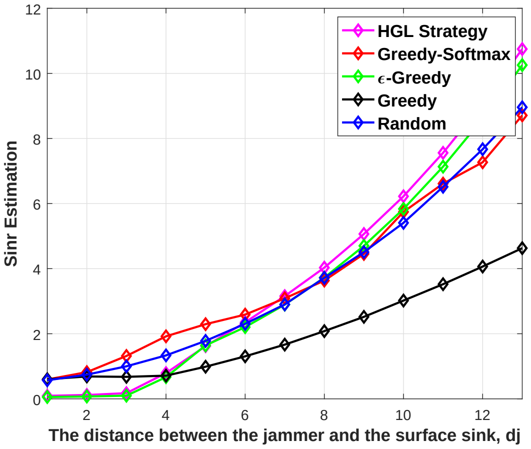


Figure 8. The utility varies with the distance d_j between node S_2 and node S_1.

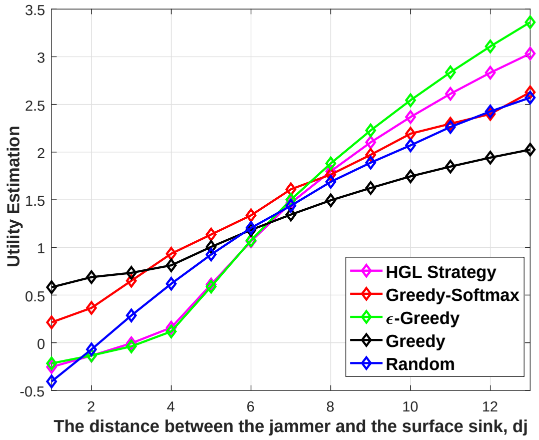


Figure 9. SINR comparison under different strategies.

7. Conclusions

This paper studies the joint resource allocation problem of multiple users in UACNs with unknown channel information and proposes a distributed, low-complexity, high-efficiency learning game algorithm that does not rely on any prior channel information. First, the joint resource allocation problem is constructed as a multi-agent MAB game model. Subsequently, a learning strategy without any prior channel gain information is designed to find the Nash equilibrium solution of the game. Two distributed learning strategies (Softmax-greedy and HGL) based on game theory are proposed to solve the above-mentioned multi-agent MAB game problem. Compared with softmax-greedy, HGL is a two-layer learning strategy. By introducing virtual learning information into the lower level learning, and learning virtual learning information, the learning efficiency of the algorithm is improved. Finally, the high learning efficiency of the proposed two-layer learning algorithm and the high efficiency of the obtained Nash equilibrium are verified by simulation tests.

Author Contributions: H.W. and Y.H. conceived and designed the whole procedure of this paper. H.W. contributed to the introduction and system model sections. L.Y. and F.L. performed and analyzed the computer simulation results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant no. 62001199), the Fujian Provincial Natural Science Fund (2019J01842, 2022J011226, 2022J011224), the project of the Education Department of Fujian Province (JT180596) and the principal fund project of Minnan Normal University (KJ2020003).

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Akyildiz, I.F.; Pompili, D.; Melodia, T. Underwater acoustic sensor networks: Research challenges. *Ad Hoc Netw.* **2005**, *3*, 257–279. [CrossRef]
2. Pompili, D.; Akyildiz, I.F. Overview of networking protocols for underwater wireless communications. *IEEE Commun. Mag.* **2009**, *47*, 97–102. [CrossRef]
3. Su, R.; Gong, Z.; Zhang, D.; Li, C.; Chen, Y.; Venkatesan, R. An adaptive asynchronous wake-up scheme for underwater acoustic sensor networks using deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2021**, *70*, 1851–1865. [CrossRef]
4. Coutinho, R.W.L.; Boukerche, A. OMUS: Efficient Opportunistic Routing in Multi-Modal Underwater Sensor Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 5642–5655. [CrossRef]
5. Du, J.; Han, G.; Lin C.; Martínez-García, M. ITrust: An Anomaly-Resilient Trust Model Based on Isolation Forest for Underwater Acoustic Sensor Networks. *IEEE Trans. Mob. Comput.* **2022**, *21*, 1684–1696. [CrossRef]
6. Liu, Y.; Wang, H.; Cai, L.; Shen, X.; Zhao, R. Fundamentals and advancements of topology discovery in underwater acoustic sensor networks: A review. *IEEE Sens. J.* **2021**, *21*, 21159–21174. [CrossRef]
7. Mezni, H.; Driss, M.; Boulila, W.; Ben Atitallah, S.; Sellami, M.; Alharbi, N. SmartWater: A service-oriented and sensor cloud-based framework for smart monitoring of water environments. *Remote Sens.* **2022**, *14*, 922. [CrossRef]
8. Villa, J.; Aaltonen, J.; Virta, S.; Koskinen, K.T. A cooperative autonomous offshore system for target detection using multi-sensor technology. *Remote Sens.* **2020**, *12*, 4106. [CrossRef]
9. Zhou, Y.; Tong, F.; Song A.; Diamant, R. Exploiting Spatial—Temporal Joint Sparsity for Underwater Acoustic Multiple-Input–Multiple-Output Communications. *IEEE J. Ocean. Eng.* **2021**, *46*, 352–369. [CrossRef]
10. Doosti-Aref, A.; Ebrahimzadeh, A. Adaptive relay selection and power allocation for OFDM cooperative underwater acoustic systems. *IEEE Trans. Mob. Comput.* **2018**, *17*, 1–15. [CrossRef]
11. Aval, Y.M.; Wilson, S.K.; Stojanovic, M. On the achievable rate of a class of acoustic channels and practical power allocation strategies for ofdm systems. *IEEE J. Ocean. Eng.* **2015**, *40*, 785–795. [CrossRef]
12. Luo, Y.; Pu, L.; Mo, H.; Zhu, Y.; Peng, Z.; Cui, J. Receiver-initiated spectrum management for underwater cognitive acoustic network. *IEEE Trans. Mob. Comput.* **2017**, *16*, 198–212. [CrossRef]
13. Jornet, J.M.; Stojanovic, M.; Zorzi, M. On joint frequency and power allocation in a cross-Layer protocol for underwater acoustic networks. *IEEE J. Ocean. Eng.* **2010**, *35*, 936–947. [CrossRef]
14. Xing, F.; Yin, H.; Ji, X.; Leung, V. C. M. Joint relay selection and power allocation for underwater cooperative optical wireless networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 251–264. [CrossRef]
15. Su, Y.; Zhu, Y.; Mo, H.; Cui, J.H.; Jin, Z. A joint power control and rate adaptation MAC protocol for underwater sensor networks. *Ad Hoc Netw.* **2015**, *26*, 36–49. [CrossRef]
16. Yang, Q.; Su, Y.; Jin Z.; Yao, G. EFPC: An environmentally friendly power control scheme for underwater sensor networks. *Sensors* **2015**, *15*, 29107–29128. [CrossRef] [PubMed]
17. Jiang W.; Tong, F. Exploiting Sparsity for Underwater Acoustic Sensor Network Under Time-Varying Channels. *IEEE Internet Things J.* **2022**, *9*, 2859–2869. [CrossRef]
18. Han, S.; Li, L.; Li, X.; Liu, Z.; Yan L.; Zhang, T. Joint Relay Selection and Power Allocation for Time-Varying Energy Harvesting-Driven UASNs: A Stratified Reinforcement Learning Approach. *IEEE Sens. J.* **2022**, *22*, 20063–20072. [CrossRef]
19. Wang, W.; Kwasinski, A.; Niyato, D.; Han, Z. Learning for Robust Routing Based on Stochastic Game in Cognitive Radio Networks. *IEEE Trans. Commun.* **2018**, *66*, 2588–2602. [CrossRef]
20. Gai, Y.; Krishnamachari, B. Distributed Stochastic Online Learning Policies for Opportunistic Spectrum Access. *IEEE Trans. Signal Process.* **2014**, *62*, 6184–6193. [CrossRef]
21. Li, X.; Liu, J.; Yan, L.; Han, S.; Guan, X. Relay Selection in Underwater Acoustic Cooperative Networks: A Contextual Bandit Approach. *IEEE Commun. Lett.* **2017**, *21*, 382–385. [CrossRef]
22. Abdelnasser, A.; Hossain, E.; Kim, D.I. Tier-Aware Resource Allocarion in OFDMA Macrocell-Small Cell Networks. *IEEE Trans. Commun.* **2015**, *63*, 695–710. [CrossRef]
23. Liu, Y.F.; Dai, Y. H.; Luo, Z.Q. Joint Power and Admission Control via Linear Programming Deflation. *IEEE Trans. Signal Process.* **2012**, *61*, 1327–1338. [CrossRef]
24. Wang, H.; Li, Y.; Qian, J. Self-adaptive resource allocation in underwater acoustic interference channel: A reinforcement learning approach. *IEEE Internet Things J.* **2020**, *7*, 2816–2827. [CrossRef]
25. Xiao, L.; Jiang, D.H.; Wan, X.; Su, W.; Tang, Y. Anti-Jamming Underwater Transmission with Mobility and Learning. *IEEE Commun. Lett.* **2018**, *22*, 542–545. [CrossRef]
26. Muhammed, D.; Anisi, M.H.; Zareei, M.; Vargas-Rosales, C.; Khan, A. Game Theory-Based Cooperation for Underwater Acoustic Sensor Networks: Taxonomy, Review, Research Challenges and Directions. *Sensors* **2018**, *18*, 425–425. [CrossRef]
27. Jing, L.; He, C.; Huang, J.; Ding, Z. Energy Management and Power Allocation for Underwater Acoustic Sensor Network. *IEEE Sens. J.* **2017**, *17*, 6451–6462. [CrossRef]
28. Fang, Z.; Wang, J.; Du, J.; Hou, X.; Ren, Y.; Han, Z. Stochastic Optimization-Aided Energy-Efficient Information Collection in Internet of Underwater Things Networks. *IEEE Internet Things J.* **2022**, *9*, 1775–1789. [CrossRef]

29. Bouabdallah, F.; Zidi, C.; Boutaba, C.R.; Mehaoua, A. Collision Avoidance Energy Efficient Multi-Channel MAC Protocol for UnderWater Acoustic Sensor Networks. *IEEE Trans. Mob. Comput.* **2019**, *18*, 2298–2314. [CrossRef]
30. Amiri, R.; Almasi, M.A.; Andrews, J.G.; Mehrpouyan, H. Reinforcement learning for self-organization and power control of two-tier heterogeneous networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 3933–3947. [CrossRef]
31. Zhao, D.; Qin, H.; Song, B.; Zhang, Y.; Du, X.; Guizani, M. A Reinforcement Learning Method for Joint Mode Selection and Power Adaptation in the V2V Communication Network in 5G. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 452–463. [CrossRef]
32. Zhang, T.; Han, G.; Yan, L.; Peng, Y. Fast Calculation of Underwater Acoustic Horizontal Range: A Guarantee for B5G Ocean Mobile Networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 2922–2933. [CrossRef]
33. Li, X.; Zhou, Y.; Yan, L.; Zhao, H.; Yan, X.; Luo, X. Optimal Node Selection for Hybrid Attack in Underwater Acoustic Sensor Networks: A Virtual Expert-Guided Bandit Algorithm. *IEEE Sens. J.* **2020**, *20*, 1679–1687. [CrossRef]
34. Gai, Y.; Krishnamachari, B. Online learning algorithms for stochastic water-filling. In Proceedings of the 2012 Information Theory and Applications Workshop, San Diego, CA, USA, 5–10 February 2012; pp. 352–356.
35. Maghsudi, S.; Stanczak, S. Joint Channel Selection and Power Control in Infrastructareless Wireless Networks: A Multiplayer Multiarmed Bandit Framework. *IEEE Trans. Veh. Technol.* **2015**, *64*, 4565–4578. [CrossRef]
36. Tong, J.; Fu, L.; Han, Z. Throughput Enhancement of Full-Duplex CSMA Networks Using Multiplayer Bandits. *IEEE Internet Things J.* **2021**, *8*, 11807–11821. [CrossRef]
37. Vandendorpe, L.; Duran, R.T.; Louveaux, J.; Zaidi, A. Power allocation for OFDM transmission with DF relaying. In Proceedings of the 2008 IEEE International Conference on Communications, Beijing, China, 19–23 May 2008; pp. 3795–3800.
38. Guruacharya, S.; Niyato, D.; Kim, D.I.; Hossain, E. Hierarchical Competition for Downlink Power Allocation in OFDMA Femtocell Networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 1543–1553. [CrossRef]
39. Shum, K.W.; Leung, K.K.; Sung, C.W. Convergence of Iterative Waterfilling Algorithm for Gaussian Interference Channels. *IEEE J. Sel. Areas Commun.* **2007**, *25*, 1091–1100. [CrossRef]
40. Xiao, L.; Li, Y.; Dai, C.; Dai, H.; Poor, H.V. Reinforcement Learning-Based NOMA Power Allocation in the Presence of Smart Jamming. *IEEE Trans. Veh. Technol.* **2018**, *67*, 3377–3389. [CrossRef]
41. Chen, X.; Zhao Z.; Zhang, H. Stochastic Power Adaptation with Multiagent Reinforcement Learning for Cognitive Wireless Mesh Networks. *IEEE Trans. Mob. Comput.* **2013**, *12*, 2155–2166. [CrossRef]