electronics

MDPI

Article

Multiple Access for Heterogeneous Wireless Networks with Imperfect Channels Based on Deep Reinforcement Learning

Yangzhou Xu [1], Jia Lou [1], Tiantian Wang [1], Junxiao Shi [1], Tao Zhang [2], Agyemang Paul [2] and Zhefu Wu [2,*]

[1] Information Communication Branch of State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310007, China; xuyangzhou_zjdl@163.com (Y.X.); loujia2015@163.com (J.L.); wangtiantian0501@126.com (T.W.); shi_junxiao@zj.sgcc.com.cn (J.S.)

[2] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310014, China; 211123030026@zjut.edu.cn (T.Z.); 221122030325@zjut.edu.cn (A.P.)

* Correspondence: wzf@zjut.edu.cn

Abstract: In heterogeneous wireless networks, when multiple nodes need to share the same wireless channel, they face the issue of multiple access, which necessitates a Medium Access Control (MAC) protocol to coordinate the data transmission of multiple nodes on the shared communication channel. This paper presents Proximal Policy Optimization-based Multiple Access (PPOMA), a novel multiple access protocol for heterogeneous wireless networks based on the Proximal Policy Optimization (PPO) algorithm from deep reinforcement learning (DRL). Specifically, we explore a network scenario where multiple nodes employ different MAC protocols to access an Access Point (AP). The novel PPOMA approach, leveraging deep reinforcement learning, adapts dynamically to coexist with other nodes. Without prior knowledge, it learns an optimal channel access strategy, aiming to maximize overall network throughput. We conduct simulation analyses using PPOMA in two scenarios: perfect channel and imperfect channel. Experimental results demonstrate that our proposed PPOMA continuously learns and refines its channel access strategy, achieving an optimal performance level in both perfect and imperfect channel scenarios. Even when faced with suboptimal channel conditions, PPOMA outperforms alternative methods by achieving higher overall network throughput and faster convergence rates. In a perfect channel scenario, PPOMA's advantage over other algorithms is primarily evident in its convergence speed, reaching convergence on average 500 iterations faster. In an imperfect channel scenario, PPOMA's advantage is mainly reflected in its higher overall network throughput, with an approximate increase of 0.04.

Keywords: medium access control; deep reinforcement learning; heterogeneous wireless network; imperfect channel

Citation: Xu, Y.; Lou, J.; Wang, T.; Shi, J.; Zhang, T.; Paul, A.; Wu, Z. Multiple Access for Heterogeneous Wireless Networks with Imperfect Channels Based on Deep Reinforcement Learning. *Electronics* **2023**, *12*, 4845. https://doi.org/10.3390/electronics12234845

Academic Editors: Hyunsoo Yoon and Namgi Kim

Received: 25 October 2023
Revised: 20 November 2023
Accepted: 29 November 2023
Published: 30 November 2023

1. Introduction

In the field of wireless communications, the concept of heterogeneous wireless networks has emerged to address the convergence of different communication technologies and protocols in order to meet the growing demands for communication. It comprises various communication technologies, including, but not limited to, Wi-Fi, cellular networks, short-range communication, IoT, LoRa, and more. The advent of heterogeneous wireless networks aims to overcome the limitations of a single technology and protocol by effectively integrating multiple technologies, providing a more flexible and efficient communication solution. While widely applied, heterogeneous wireless networks also face multiple challenges, such as spectrum resource management, mobility management, and security management [1–3].

In order to make the most of limited spectrum resources, dynamic spectrum sharing has emerged as a promising approach to enhance spectrum utilization efficiency in the context of cognitive radio technology [4]. However, to achieve effective spectrum sharing, the primary challenge is addressing the Media Access Control (MAC) problem. The MAC

protocol is a network communication protocol situated in the Data Link Layer of the OSI model. It plays a crucial role in wireless communication networks, responsible for managing and controlling terminal devices' access to the shared channel. This involves the rational allocation of spectrum resources to ensure coordination and collaboration among different nodes, thereby maximizing network throughput. Specifically, the MAC protocol establishes access rules, determining when and how each terminal device sends data. Such scheduling mechanisms are vital in avoiding data collisions and conflicting transmissions. In the context of multiple access issues, where multiple terminal devices compete for limited channel resources, the MAC protocol's task is to coordinate these devices, ensuring they transmit data effectively at the same time and in the same space.

The significance of addressing multiple access issues lies in its direct impact on network performance and efficiency. By optimizing the MAC protocol, we can enhance network throughput and reduce data packet collisions, thereby improving communication reliability and stability. Effective multiple access solutions also contribute to lowering network latency and enhancing real-time data transmission, especially in applications requiring low latency, such as VoIP [5], smart grids [6], and others.

In recent years, with the continuous advancement of artificial intelligence, deep reinforcement learning has achieved significant success in various fields. Deep reinforcement learning is an approach that combines deep learning and reinforcement learning. It aims to train an agent to perform actions in an environment to maximize cumulative rewards. This typically involves using deep neural networks to learn complex policies and value functions. It has given rise to popular algorithms like Deep Q Network (DQN) [7], which leverages deep neural networks to approximate Q-values for effective decision-making, and Proximal Policy Optimization (PPO) [8], an algorithm designed to address decision-making challenges in both continuous and discrete action spaces. These algorithms showcase remarkable optimization capabilities in addressing complex decision-making challenges, as seen in applications within gaming [9] and robot control [10]. Additionally, in the field of communications, where challenges like channel coding [11–14], spectrum allocation [15–17], and dynamic spectrum access [18,19] exist, deep reinforcement learning has emerged as a powerful tool to address these challenges, with its strengths lying in its ability to tackle complex decision optimization problems, handle large-scale datasets, and adapt to high-dimensional state and action spaces.

Deep reinforcement learning methods can optimize policies, which is highly valuable for policy selection in MAC protocols [20]. This is particularly advantageous because the multiple access problem exhibits Markovian properties, enabling the utilization of deep reinforcement learning methods to tackle decision-making challenges of this nature. The selection of strategies in MAC protocols is crucial for effectively managing and controlling terminal devices' access to shared channels. By leveraging deep reinforcement learning methods, we can optimize these strategies, thereby enhancing system performance and network throughput. This is because deep reinforcement learning excels in adapting strategies to different network conditions and variations in complex multiple access environments through continuous learning and adjustment. Additionally, the advantages of deep reinforcement learning lie in its ability to handle large-scale, high-dimensional state spaces and complex decision-making problems. In the context of MAC protocols, this implies a more comprehensive consideration of various factors, including network topology, channel states, data transmission requirements, etc., enabling the formulation of more intelligent and adaptive strategies. For example, Yu et al. [21] applied DQN as a foundational framework to formulate a novel MAC protocol, denoted as Deep-Reinforcement Learning Multiple Access (DLMA). Yu et al. [22] introduced a distributed DQN-based MAC protocol aimed at facilitating efficient and equitable spectrum sharing within heterogeneous wireless networks. References [21,22] primarily utilize the DQN algorithm from deep reinforcement learning to address the multiple access problem. However, DQN demands a substantial amount of experiential data during application and may exhibit less stable training processes. In contrast, the PPO algorithm outperforms DQN in the field of

deep reinforcement learning, particularly when dealing with highly dynamic and complex decision-making problems. It exhibits superior convergence performance, implying a quicker optimization of the channel and improvement in network performance. Moreover, the PPO algorithm can more effectively harness limited experiential data, reducing reliance on hyperparameter adjustments and typically demonstrating more consistent performance. Therefore, we have integrated the PPO algorithm with the MAC protocol to propose a novel multiple access protocol known as the Proximal Policy Optimization-based Multiple Access (PPOMA).

PPO, a type of deep reinforcement learning method, is commonly employed to address decision problems in discrete or continuous action spaces. It is based on the policy gradient method, iteratively updating policies to maximize cumulative rewards. PPOMA is a MAC protocol based on PPO, indicating that in the design of PPOMA, we draw inspiration from the PPO algorithm and apply it to protocol design at the MAC layer. This choice stems from the fact that the multiple access problem in heterogeneous wireless networks is a type of decision problem, prompting the introduction of the PPO algorithm to tackle such issues. While PPOMA borrows core ideas from PPO, adjustments to the algorithm have been made to accommodate specific requirements of wireless communication. This may include considerations for channel access, data transmission, and collision handling in wireless communication scenarios.

The primary focus of this paper is the single-channel multiple access problem in heterogeneous wireless networks. However, traditional MAC protocols suffer from issues such as low spectrum efficiency and frequent data transmission collisions. Therefore, there is a need for a more efficient MAC protocol to address the challenges of multiple access. The introduction of PPOMA aims to tackle the multiple access problem in heterogeneous wireless networks, striving to achieve an optimal channel access strategy, which inherently leads to maximizing network throughput. We demonstrated that when PPOMA protocol nodes coexist with nodes using different protocols, it is possible to achieve near-optimal total throughput without any prior knowledge. In other words, PPOMA does not require knowledge of the characteristics of the coexisting protocols; it employs deep reinforcement learning to eventually maximize network throughput in the presence of other protocols. This paper also investigates the performance of PPOMA when coexisting with other protocols in both imperfect and perfect channel conditions.

Specifically, under the assumption of a perfect channel, we disregard external interference factors like noise. In this perfect scenario, there are only two possible states during data transmission: either the data are successfully sent or a collision occurs. In the case of an imperfect channel, we consider the impact of interference factors, such as noise, on data transmission. In this scenario, a new possibility arises during data transmission: the potential loss of data due to external interference, increasing the uncertainty in data transmission. Unlike a perfect channel, the imperfect channel acknowledges the possibility of data packet loss, making the data transmission state more intricate and necessitating a more adaptable handling mechanism to deal with packet loss. Our designed PPOMA protocol demonstrates efficient channel utilization when coexisting with other protocols in both imperfect and perfect channel scenarios. In imperfect channels, the PPOMA protocol leverages the capabilities of deep reinforcement learning, gradually adapting to channel variations by continuous exploration and learning to identify optimal transmission strategies. In perfect channels, the PPOMA protocol utilizes precise channel state information to achieve optimal transmission strategies more rapidly, thereby enhancing channel utilization and increasing the overall network throughput.

The experimental results show that our PPOMA protocol, designed for heterogeneous wireless networks, improves total throughput and expedites convergence under various channel conditions, including both perfect and imperfect scenarios. This swift convergence holds particular importance in the realm of wireless networks, as it enables the algorithm to uncover superior strategies in a shorter timeframe, thereby yielding enhanced performance.

This also implies that MAC protocols can begin optimizing network resource scheduling and utilization at an earlier stage, further enhancing overall throughput and performance.

In summary, the main contributions of this paper are as follows:

1. We introduce a novel multiple access protocol, PPOMA, which is designed based on the PPO algorithm from deep reinforcement learning. PPOMA offers faster convergence, improved sampling capabilities, and enhanced exploration abilities. These features enable the effective resolution of the multiple access problem in heterogeneous wireless networks, ultimately enhancing network performance.
2. We conduct simulation experiments with the PPOMA protocol under both perfect and imperfect channel conditions, considering various real-world factors like channel interference and packet loss. This comprehensive evaluation provides a more objective assessment of PPOMA's performance and brings us closer to real-world scenarios.
3. By comparing the experimental results with existing multiple access protocols, we find that PPOMA outperforms them by achieving higher overall throughput and faster convergence in different channel conditions. This satisfies the demands of practical communication systems for efficiency and reliability. We believe that PPOMA holds great potential and broad applicability in the field of multiple access protocols.

The subsequent sections of this paper will be organized as follows: Section 2 provides an overview of related works. Section 3 offers a detailed description of the system model, encompassing crucial elements such as the communication environment. Section 4 delves into the design and specific details of the PPOMA protocol based on the PPO algorithm. This includes its operational principles and key algorithms. Section 5 presents simulation results and corresponding analyses to validate and evaluate the performance of the PPOMA protocol in various scenarios. Section 6 summarizes the experimental conclusions.

2. Related Works

The problem of dynamic spectrum access in wireless networks has garnered significant attention due to the ever-increasing demand for wireless communication resources. Efficient utilization of the available spectrum is crucial for improving network performance and accommodating the diverse needs of multiple users. In this context, Yu et al. [21] designed a MAC protocol for heterogeneous wireless networks called DLMA, utilizing an improved DQN algorithm to maximize the total throughput of heterogeneous wireless networks. However, they only verified the protocol's performance under perfect channel conditions. In contrast, our research goes further by considering scenarios with imperfect channel conditions, with a particular focus on data packet loss caused by interference. We conducted simulations of common data packet loss scenarios in real network environments, making our study more aligned with the requirements of practical communication environments.

Yu et al. [22] explored how to maximize the throughput of heterogeneous wireless networks in the presence of unreliable channels by leveraging the collaboration of multiple agents. They employed a feedback recovery mechanism to acquire accurate channel feedback information. However, the recovery of precise channel feedback post-transmission can be regarded as operating under perfect channel conditions. We do not consider this recoverable scenario; in other words, our focus is on imperfect channels rather than unreliable ones, and we investigate the protocol's performance in this context.

Kaur et al. [23] studied the problem of incomplete feedback dynamic spectrum access in multi-user wireless networks. They designed a solution based on distributed deep reinforcement learning to ensure that multiple agents collaborate to make consistent decisions to maximize network utility. This approach deals with a multi-agent problem, whereas our focus is based on using a single agent to maximize network throughput.

Naparstek et al. [24] focused on solving the spectrum access problem in multi-channel heterogeneous wireless networks. They utilized the Dueling DQN method and employed a distributed multi-user DRL algorithm, ultimately successfully maximizing network utility at a lower cost. However, their MAC protocol was designed for homogeneous wireless networks, where all users use the same MAC protocol to dynamically access multiple

wireless channels. In contrast, our designed MAC protocol is intended for heterogeneous network scenarios.

Xu et al. [25] also utilized deep reinforcement learning techniques to address dynamic spectrum access problems. Similar to the study mentioned in reference [24], they discussed multi-channel scenarios. However, a key difference is that they considered time-varying channels, as some "primary" or "legacy" users might occasionally occupy the channels. Therefore, it can also be viewed as a study on heterogeneous networks. Their approach focused on learning the channel characteristics and transmission patterns of these "primary" or "legacy" users to maximize their own throughput. In contrast, our PPOMA does not require such learning to achieve throughput maximization.

Chang et al. [26] applied deep reinforcement learning within cognitive networks, where secondary users aimed to efficiently utilize spectrum resources not in use by primary users. Specifically, secondary users used deep reinforcement learning to mitigate potential interference with primary users. In contrast, our research goes beyond the avoidance of such mutual interference and focuses on achieving the harmonious coexistence of secondary and primary users, ultimately reaching the goal of maximizing total throughput.

The MCT-DLMA protocol proposed by Zhang et al. [27] is designed to address the multi-channel transmission multiple access problem in heterogeneous wireless networks. It improves the spectrum utilization of multi-channel heterogeneous wireless networks based on considering the practical communication model with non-saturated traffic. However, our research focuses on the saturated communication traffic scenario, which is slightly different from what they considered.

The aforementioned studies have primarily employed the DQN algorithm from deep reinforcement learning, or they have introduced modifications to the DQN algorithm to address specific issues. In contrast, our research utilizes the PPO algorithm from deep reinforcement learning to design a MAC protocol known as PPOMA. PPOMA has the capability to operate cohesively with other MAC protocols in heterogeneous wireless networks, thereby enhancing the overall system throughput.

This is primarily because PPO offers several advantages over DQN:

(1) PPO employs a clipping mechanism to limit the range of policy updates, ensuring that each policy update remains within a reasonable range and avoids significant fluctuations. This stabilizes parameter updates, resulting in faster convergence.
(2) It utilizes importance sampling, estimating the value of previous policies with samples generated by the current policy. This enhances the efficiency of sampling, enabling more effective use of historical data and reducing the number of required samples, thus accelerating training.
(3) PPO exhibits stronger exploration capabilities. This is because PPO falls under the category of policy-based learning methods. It directly employs neural networks to approximate the policy function $\pi(a|s)$ for the purpose of updating policy, enabling agents to explore more flexibly during the learning process. In contrast, DQN is a value-based learning method that focuses on approximating the optimal action-value function $Q_*(s, a)$ to find the best policy, which can limit its exploration capabilities.

Therefore, we attempted to utilize the PPO algorithm for MAC protocol design.

3. System Model

The system model considered in this paper is a single-channel heterogeneous wireless network, comprising one AP node and multiple distinct transmitting nodes. This type of network structure is quite common in the context of the IoT or LoRa networks, and our network configuration can be viewed as a simplified representation of these intricate networks. It is worth noting that our research primarily focuses on LOS scenarios, where we assume our network to be a short-distance network. In this setting, multiple transmitting nodes compete on the same wireless channel to transfer data to the AP. Additionally, we assume that each node has data packets to transmit, leading to competition for channel resources among the nodes for data transmission. This scenario is particularly typical in

specific IoT applications or LoRa network environments, especially in situations where efficient short-distance communication is crucial, such as in smart homes, wireless sensor networks, and similar contexts.

The nodes employ a transmission model based on slotted ALOHA [28]. In the slotted ALOHA transmission model, time is divided into multiple frames, each containing several time slots. Each node can send data packets only at the beginning of each time slot, transmitting data packets on the shared wireless channel and completing the packet transmission before the end of the time slot. Only one node is allowed to transmit data to the AP node in each time slot; otherwise, it would result in a collision, leading to transmission failure. In the context of an imperfect communication channel, we explore a situation where data packets transmitted by nodes have the potential to be lost. Even when only a single node is transmitting in the channel without any collision, there is still a probability of data packet loss, which results in transmission failure.

Each node in the network employs different MAC protocols, including TDMA, q-ALOHA, FW-ALOHA, and EB-ALOHA. Additionally, the network has at least one PPOMA node that uses the PPOMA protocol, as shown in Figure 1. The primary task of the PPOMA node is to maximize the overall network throughput by learning and achieving an optimal channel access strategy.

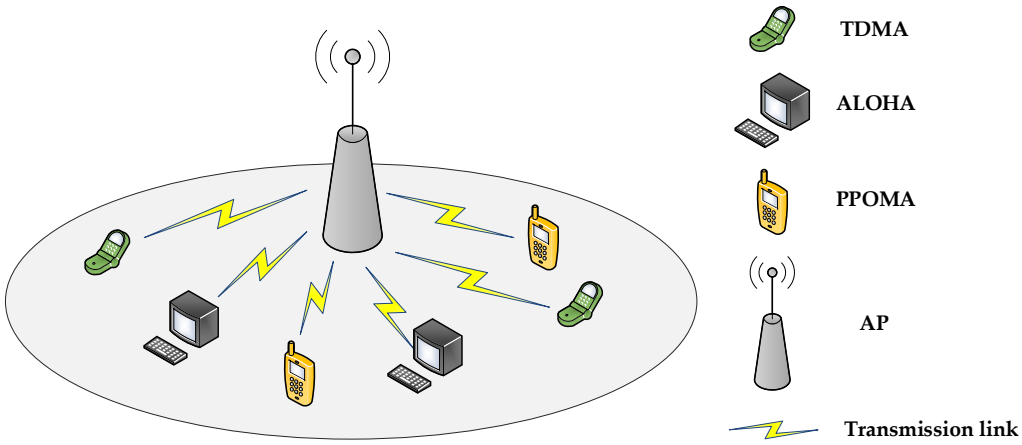


Figure 1. Heterogeneous multiple access system with multi-protocol hybrid.

The detailed descriptions of different MAC protocols used by different nodes are given below:

TDMA Protocol: TDMA nodes transmit data in X fixed time slots within each frame, while the remaining time slots are not used for data transmission. This slot allocation is ordered, not random, and offers high controllability.

q-ALOHA Protocol: The q-ALOHA protocol is a random access protocol where q-ALOHA nodes, in each time slot, decide whether to transmit data with a fixed probability q. Therefore, data transmission in q-ALOHA exhibits randomness.

FW-ALOHA Protocol: In the FW-ALOHA protocol, nodes, after completing a transmission in a specific time slot, randomly select a value w from the range of 1 to W with equal probability. Subsequently, they transmit data again in the time slot, which is w time slots ahead. W is referred to as the window size.

EB-ALOHA Protocol: The EB-ALOHA protocol introduces certain modifications to FW-ALOHA, mainly reflected in the fact that its window size is not fixed, that is, the value range of w is dynamically changed. Specifically, each collision occurrence leads to a doubling of the window size W. Continuous collisions can result in the window size reaching a maximum of $2^m W$, where m represents the "maximum backoff stage". A successful transmission is the only condition for the window size to revert to its initial value, W.

PPOMA Protocol: The PPOMA protocol employed by PPOMA nodes, as proposed in our work, offers two choices at the beginning of each time slot: (1) Send a data packet, followed by determining the success or failure of the transmission based on the reception of an ACK signal from the AP node. (2) Abstain from transmitting a data packet, and, instead, listen to the channel to obtain channel observations, gaining insights into the transmission activities of other nodes. Leveraging these observations, PPOMA nodes set learning objectives to ultimately achieve the goal of maximizing the total network throughput when coexisting with various protocols. This implies that PPO has learned the optimal channel access strategy.

4. PPOMA Protocol Design

4.1. Overview of PPO Algorithm

PPO is a high-performance deep reinforcement learning algorithm [29] that utilizes the actor–critic architecture [30]. It demonstrates excellent performance in both discrete and continuous action spaces. The algorithm's framework is illustrated in Figure 2. Its neural network structure primarily consists of two parts: (1) Actor Network (Policy Network): The Actor network approximates the policy, denoted as $\pi(a|s;\theta) \approx \pi(a|s)$. It takes the current state s_t as input and outputs a probability distribution $\pi(a_t|s_t;\theta)$ over the possible actions a_t. The actor network's main role is to learn and generate the policy for taking actions in different states. (2) Critic Network (Value Network): The Critic network approximates the state-value function, denoted as $V_\pi(s_t;w) \approx V_\pi(s_t)$. It takes the state s_t as an input and provides the state value function $V_\pi(s_t;w)$ as an output. The critic network's primary function is to evaluate the quality of states and assist the actor network in adjusting action policies. Specifically, the critic network can be used to calculate the advantage function, which estimates the advantage of each action relative to the average action and guides policy improvements. In the PPO algorithm, these two networks work together to help the agent learn better policies, ultimately achieving the effect described in Equation (1).

$$\max_{\theta}\left\{J(\theta) \triangleq \mathbb{E}_S[V_\pi(S)]\right\} \tag{1}$$

where $V_\pi(S)$ represents the state-value function, $\mathbb{E}_S[V_\pi(S)]$ denotes the expectation over all possible states s_t, θ represents the parameters of the neural network, and $J(\theta)$ represents the performance metric under the parameter θ, which in this context is the expectation of the state value.

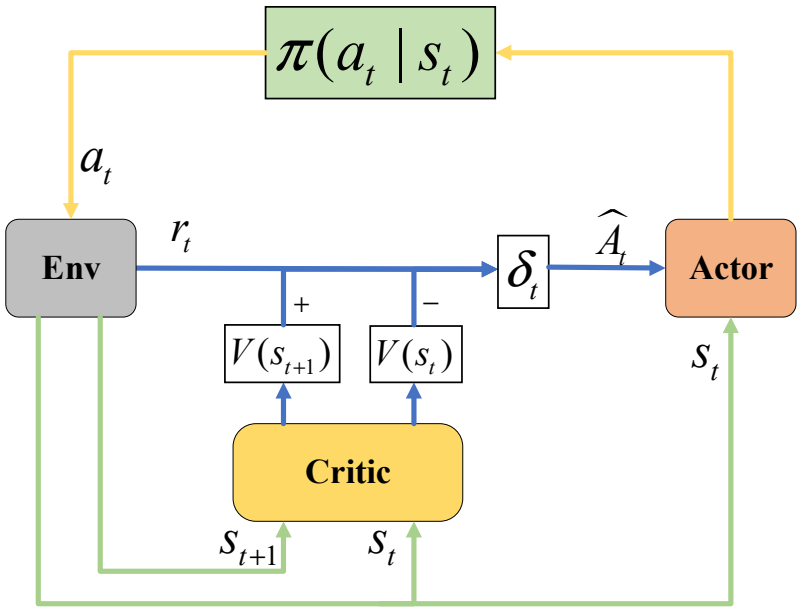


Figure 2. The framework of the PPO algorithm.

PPO can be regarded as an improvement over the Trust Region Policy Optimization (TRPO) algorithm [31]. The primary motivation for this enhancement is that TRPO involves highly complex computations, demanding substantial computational resources for every policy update. The core idea of both algorithms is to limit the magnitude of policy updates during the training process, ensuring that each update stays within an acceptable range. The key difference is that TRPO uses the KL divergence to impose this constraint, while PPO employs a clipping function to bind the policy update magnitude.

TRPO approximates the objective function $J(\theta)$ in Equation (1) with a loss function $L(\theta)$. The definition of $L(\theta)$ is as follows:

$$L(\theta) = \hat{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \cdot \hat{A}_t\right] = \hat{E}_t\left[r_t(\theta) \cdot \hat{A}_t\right] \tag{2}$$

$$\text{s.t. } \hat{E}_t\left\{KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right]\right\} \leq \Delta \tag{3}$$

where $\hat{E}_t[\cdot]$ represents the expectation over multiple samples, $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$ is the probability ratio of the new policy to the old policy. $\pi_\theta(a_t|s_t)$ is the new policy, while $\pi_{\theta_{old}}(a_t|s_t)$ is the old policy. $KL[\cdot]$ computes the divergence between the new and old policy probability distributions, with smaller divergence values indicating less difference between new and old policies. Δ represents the confidence region, and having a divergence below the confidence region signifies that the difference between the new and old policies is small.

In contrast, in PPOMA, the loss function $L^{CLIP}(\theta)$ is defined as

$$L^{CLIP}(\theta) = \hat{E}_t\left[min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)\right] \tag{4}$$

where the clip function is a truncation function that restricts the range of changes between the new and old policies to $[1-\varepsilon, 1+\varepsilon]$. Compared to TRPO, which uses KL divergence to constrain the new and old policies, using the clip function is notably simpler and reduces computational complexity. $\hat{A}_t$ represents the advantage function used to assess the advantage of taking action a in state s relative to the average behavior. It is generally defined as follows:

$$\hat{A}_t = Q_\pi(s_t, a_t) - V_\pi(s_t) \tag{5}$$

where $Q_\pi(s_t, a_t)$ is the action value function, representing the expected total return of taking action a_t in state s_t, and $V_\pi(s_t)$ is the state value function, representing the expected total return in state s_t.

However, in practical applications, the Generalized Advantage Estimation (GAE) is often used. In this case, the definition of the advantage function $\hat{A}_t$ is updated as follows:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \ldots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{6}$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{7}$$

where δ_t is the TD error at time t, r_t is the reward obtained after taking an action, γ is the discount factor, λ is the hyperparameter for GAE, controlling the weighting of future rewards, and T represents the total number of time steps.

The importance of employing the GAE algorithm lies in its weighted averaging of future TD errors, with the weights being determined by the value of $\gamma\lambda$. This allows the algorithm to better consider future rewards when assessing the advantage of the current action. It improves the estimation of the advantage function, stabilizes the learning process, and enhances the performance of the PPO algorithm.

Finally, the policy and value network parameters in PPO are updated using gradient ascent. The updated method is as follows:

$$w_{t+1} = w_t + \alpha^w \delta_t \nabla_w V(s_t, w_t) \tag{8}$$

$$\delta_t = r_t + \gamma V(s_{t+1}, w_{t+1}) - V(s_t, w_t) \tag{9}$$

$$\theta_{t+1} = \theta_t + \alpha^{\theta} \delta_t \nabla_{\theta} lg \pi_{\theta_t}(a_t | s_t; \theta) \tag{10}$$

where w_t and θ_t represent the parameters of the policy network and value network at the current time step, respectively. w_{t+1} and θ_{t+1} represent the updated network parameters. α^w and α^{θ} are the learning rates for the networks. $\nabla_w V(s_t, w_t)$ denotes the gradient with respect to w. $\nabla_{\theta} lg \pi_{\theta_t}(a_t | s_t; \theta)$ represents the gradient with respect to θ.

When applying the PPO algorithm to the MAC protocol, we leverage its outstanding performance in deep reinforcement learning, particularly its strong adaptability and effectiveness in both discrete and continuous action spaces. In MAC protocol design, we model the problem as a Markov Decision Process, where nodes need to make optimal decisions to maximize network performance when competing for limited communication resources, ultimately resulting in the design of the PPOMA protocol. PPOMA achieves a near-optimal channel access strategy without prior knowledge.

Specifically, in the initial phase, PPOMA nodes lack sufficient experience samples, leading to significant randomness in the actions chosen at the beginning of each time slot. The node might choose to send data or not, and simultaneous data transmission by other nodes may lead to collisions, reducing network throughput and increasing latency. In such instances, a penalty is assigned to the PPOMA node's intelligent agent. After accumulating experience samples over time, the agent simultaneously chooses actions based on the policy and optimizes the strategy based on the acquired experience samples.

For instance, having learned from experience samples with penalties, in similar situations, the agent might choose not to send data to avoid penalties. The agent's ultimate goal is to maximize the cumulative reward value, prompting it to avoid actions that incur penalties. This minimizes collisions in the network, maximizing throughput. After each time slot, regardless of whether the channel environment is perfect or imperfect, an experience sample is generated, allowing PPOMA to continuously learn and refine its channel access strategy.

Algorithm 1 provides pseudocode for the PPO-based multiple access algorithm, offering a clear illustration of our specific process in MAC protocol design.

Algorithm 1: PPOMA

Initialize s_0, ρ, γ, F
Initialize the parameter of actor as θ, the parameter of critic as w, the parameter of actor–target as θ^-
For $t = t_0, t_1, t_2, \ldots$ **do**
  Input s_t into actor and output $\pi(a|s_t,\ \theta)$
  Sample action a_t based on the probability $\pi(a|s_t,\ \theta)$
  Observe z_t, r_t
  Compute s_{t+1} from s_t, a_t, z_t
  Store (s_t, a_t, r_t, s_{t+1})
  If Remainder $(t/F == 0)$ **then**
    Update θ^- by setting $\theta^- = \theta$
  End if
  Input s_t into actor–target and output $\pi_{old}(a_{old}|s_t,\ \theta^-)$
  Input s_t into critic and output $V(s_t, w)$
  Calculate $\hat{A}_t, \delta_t, r_t(\theta)$
  Calculate $L^{CLIP}(\theta) = \hat{E}_t\left[min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)\right]$
  Update θ and w by using gradient descent
  Minimize $L(\theta)$
End for

4.2. Action, Observation, State, and Reward

The PPOMA node selects an action $a_t \in \{a_0, a_1\}$ at the beginning of each time slot, where a_0 indicates that the node does not transmit data, and a_1 signifies that the node

sends data. If it chooses not to transmit data, PPOMA will engage in channel sensing to understand the transmission status of other nodes. If it opts to send data, it will determine the success of the transmission based on whether it receives an ACK signal from the AP node. Ultimately, the PPOMA node determines the channel observation value z_t based on the received ACK signal from the AP and channel sensing.

In PPOMA node, $z_t \in \left\{z_{yes}, z_{failed}, z_{no}\right\}$ represents the channel observation value after taking action, where z_{yes} indicates that in the current time slot, only one node has successfully transmitted data, meaning that only one node has chosen to send data, while the rest of the nodes have chosen not to transmit data. z_{failed} signifies data transmission failure, indicating that in the current time slot, at least two or more nodes have chosen to send data, leading to a collision in the channel, resulting in data transmission failure. z_{no} represents an idle channel, indicating that there is no data transmission by any node in the current time slot. Based on these observations, the PPOMA node establishes learning objectives to ultimately achieve the goal of maximizing the total network throughput when coexisting with different protocols. This implies that PPO learns the optimal channel access strategy.

States are the foundation for making policy decisions in deep reinforcement learning. Well-designed states should capture critical information about the environment to enable intelligent decision-making by the agent. We define the action–observation pair at time $t+1$ as $c_{t+1} \triangleq (a_t, z_t)$. Where c_{t+1} primarily records the action a_t taken by the agent at time t and the channel observation value z_t. As previously discussed, there are a total of six permutations of these two values. However, the combination $\{a_1, z_n\}$ is impossible, representing a situation where the PPOMA node sends data, but the channel is idle. The remaining five possible combinations for c_{t+1} are $\{a_1, z_y\}$,$\left\{a_1, z_f\right\}$, $\{a_0, z_y\}$,$\{a_0, z_n\}$, and $\left\{a_0, z_f\right\}$. It is important to note that, for a PPOMA node coexisting with another protocol node, the combination $\left\{a_0, z_f\right\}$ is also impossible.

We define the state at time $t+1$ as $s_{t+1} \triangleq [c_{t-M+2}, \ldots, c_t, c_{t+1}]$, where the state s_{t+1} is composed of the actions and channel observation values from the previous M time steps. The parameter M represents the length of historical states that need to be recorded. By combining multiple past action–observation pairs into a sequence of states, PPOMA can consider more historical information, leading to better decision-making. Intuitively, a longer historical state length M will lead to better decisions, but it will also require a larger state space, increasing computational complexity and time. Therefore, it is necessary to choose a reasonable value for M.

After PPOMA takes action a_t, it transitions from state s_t to state s_{t+1} and generates a reward value r_t. The reward value is a feedback mechanism for the agent to understand the quality of its actions within the environment. It reflects whether the actions taken are, to some extent, good or bad. The ideal strategy is to accumulate the maximum reward value at each step, where each obtained reward is the best possible reward.

We define r_t as follows:

$$r_t = \begin{cases} 1, & z_t = z_{yes} \\ 0, & z_t = z_{failed} \text{ or } z_{no} \end{cases} \tag{11}$$

In other words, when data transmission is successful on the channel, the reward value is one. When there is a collision in data transmission or when the channel is idle, the reward value is zero.

5. Performance Evaluation

In this section, we perform detailed experiments to evaluate the performance of our proposed PPOMA protocol. We use the PPOMA protocol in two different environments: a perfect channel and an imperfect channel. Specifically, we consider the following three sce-

narios: (1) coexistence with the TDMA protocol; (2) coexistence with three different ALOHA protocols separately; and (3) coexistence with the TDMA protocol and ALOHA protocol.

5.1. Simulation Setup

We conducted our simulation program on a computer equipped with an Intel(R) Core(TM) i5-11260H CPU and an NVIDIA GeForce RTX 3050 Laptop GPU. We used Python 3.6 and the Keras deep learning platform [32], along with the Adam optimizer [33] for training the neural networks. We employed fully connected neural networks with *ReLU* activation functions for the neurons. It is worth noting that the output layer of the actor network utilized the Softmax function, while the critic network's output layer had no activation function. These settings were chosen to accurately model and evaluate the performance of the PPOMA protocol. Table 1 provides the hyperparameter settings used in PPOMA.

Table 1. PPOMA hyperparameter settings.

Hyperparameters	Value
state history length M	20
learning rate in Adam optimizer	0.001
target network update frequency F	10
target network update weight w	0.9
discount factor γ	0.99
value in the Clip function ε	0.2
number of training batches B	10

5.2. Baseline and Performance Metrics

We compared the PPOMA protocol with the DLMA protocol, using theoretical values computed from benchmark tests under sensor nodes as a reference.

DLMA protocol [21]: DLMA nodes are wireless nodes using the DLMA protocol. DLMA is a multiple access protocol designed based on DQN, which approximates the maximization of total throughput by using deep reinforcement learning and recording historical state information.

Benchmark tests [34]: These tests involve coexistence with a sensing node and the calculation of the theoretically optimal total throughput. Sensing nodes are different from PPOMA nodes in principle. Sensing nodes have prior knowledge and are aware of the information and patterns of coexisting MAC protocols. In contrast, PPOMA lacks prior knowledge and can only achieve optimal throughput through deep reinforcement learning, without prior knowledge of the MAC protocols used by the coexisting nodes.

Specifically, when a sensing node coexists with TDMA nodes, it fully understands the number of time slots used by the TDMA protocol and can occupy the unsent time slots of TDMA to maximize total throughput. In this case, the theoretical optimal throughput is one because it sends data packets in every time slot. When a sensing node coexists with q-ALOHA nodes, it knows the sending probability q, allowing it to send less data when q is high and more data when q is low to achieve optimal throughput. When sensing nodes coexist with FW-ALOHA and EB-ALOHA nodes, they are aware of the window size W and the rules for sending data packets in advance.

This study employs throughput as the performance evaluation metric, defined as the average number of successfully transmitted data packets in each timeslot, averaged over the results of N timeslots. The throughput of nodes is calculated using the following normalization formula:

$$T = \sum_{\tau=t-N+1}^{t} \frac{n_\tau}{N} \tag{12}$$

where $\sum_{\tau=t-N+1}^{t} n_\tau$ represents the number of successfully transmitted time slots within N time slots, and N represents the total number of time slots.

For the calculation of short-term throughput, we set the total number of time slots N to 1000, with each time slot lasting 1 millisecond. This results in short-term throughput reflect the performance of network nodes in the past 1 s. If we wish to calculate long-term throughput, N can be set to t, providing the average throughput over t time slots.

5.3. Comparison of Total Throughput under Perfect Channel Conditions

5.3.1. Coexistence of PPOMA Nodes with TDMA Nodes

We first conducted a simulation analysis of the coexistence of PPOMA nodes with a single TDMA node.

As shown in Figure 3a, when the total number of time slots is 10 and the number of time slots allocated to TDMA is given by $X = \{2, 3, 5, 7, 8\}$, the theoretical optimal value calculated by the sensor node is one. It can be observed that as the number of time slots allocated to TDMA X increases, the throughput of TDMA nodes in the network also increases. This is because the PPOMA node learns to occupy the idle time slots of TDMA to send data packets, thereby achieving optimal total throughput. The experimental results show that for this relatively simple scenario of coexistence with TDMA nodes, both protocols eventually achieve optimal total throughput levels. However, the difference between the two protocols lies in their convergence speed.

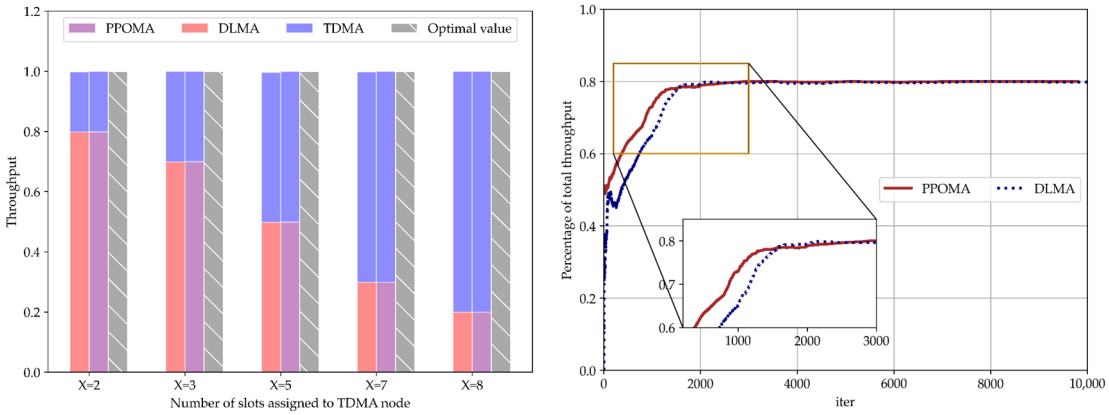


Figure 3. Coexistence of PPOMA nodes with TDMA nodes. (**a**) Throughput under different numbers of time slots X; (**b**) convergence speed when $X = 2$.

Figure 3b shows a comparison of the convergence speed between the two protocols, with a specific focus on the scenario where $X = 2$. Figure 3b shows that both protocols converge to a value near 0.8 within about 2000 iterations. Nevertheless, PPOMA reaches convergence approximately 200 iterations faster than DLMA, achieving the same level of performance with fewer iterations. Furthermore, after convergence, the throughput curves exhibit very little fluctuation. This is primarily because the TDMA protocol itself is highly controllable, and once the PPOMA node fully understands the number of time slots it is using, it can avoid collisions.

5.3.2. Coexistence of PPOMA Nodes with ALOHA Nodes

Next, we consider the coexistence of PPOMA nodes with a single ALOHA node. Here, we study three variants of ALOHA protocols: q-ALOHA, FW-ALOHA, and EB-ALOHA.

In Figure 4a, we investigate the coexistence of PPOMA nodes with q-ALOHA protocol nodes, where q represents the probability of ALOHA sending data packets at the beginning of each time slot. We set q to various values, such as $q = \{0.2, 0.3, 0.5, 0.7, 0.8\}$, and compare the results with the theoretically optimal values calculated by the sensor node. The optimal values for the five different q values are 0.8, 0.7, 0.5, 0.7, and 0.8, respectively. The results show that when q is small (e.g., $q = 0.2$ and $q = 0.3$), the throughput of the q-ALOHA

nodes is nearly zero. This is because when the transmission probability is low, PPOMA nodes, aiming to maximize the total network throughput, will choose to send data in every time slot, leading to frequent collisions with the q-ALOHA nodes, resulting in almost zero throughput. As q increases (e.g., $q = 0.7$ and $q = 0.8$), the situation is reversed, and our nodes choose to send data less frequently to avoid collisions with the q-ALOHA nodes. In this case, most of the total network throughput comes from the q-ALOHA nodes. Overall, in the case of these five different q values, using the PPOMA protocol resulted in a total network throughput that was 0.013, 0.014, 0.004, 0.016, and 0.015 higher compared to DLMA, respectively. This demonstrates that our PPOMA protocol can achieve results closer to the optimal throughput.

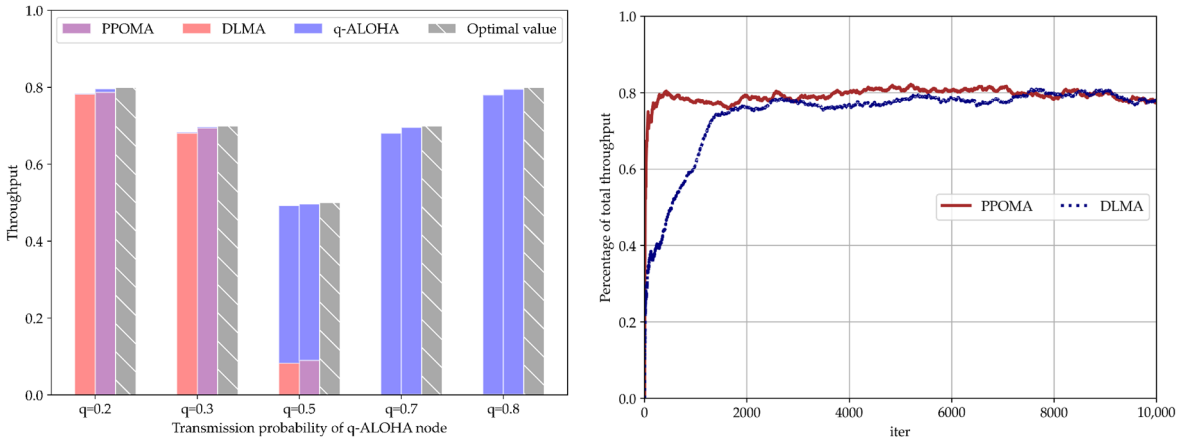


Figure 4. Coexistence of PPOMA nodes with q-ALOHA nodes. (**a**) Throughput under different probability q; (**b**) convergence speed for $q = 0.2$.

To illustrate the convergence speed of both protocols, we only selected the case with $q = 0.2$ for comparison. As shown in Figure 4b, PPOMA nodes exhibit significantly faster convergence compared to DLMA, reaching the convergence state about 800 iterations earlier. This indicates that our PPOMA protocol enhances the performance of the communication system.

Figure 5a illustrates the coexistence of PPOMA nodes with FW-ALOHA protocol nodes, where FW-ALOHA protocol's window size varies ($W = \{2, 3, 4, 5, 6\}$). According to the theoretical values calculated by the sensor node, the optimal throughput values for the five scenarios are 0.667, 0.667, 0.7, 0.733, and 0.762, respectively.

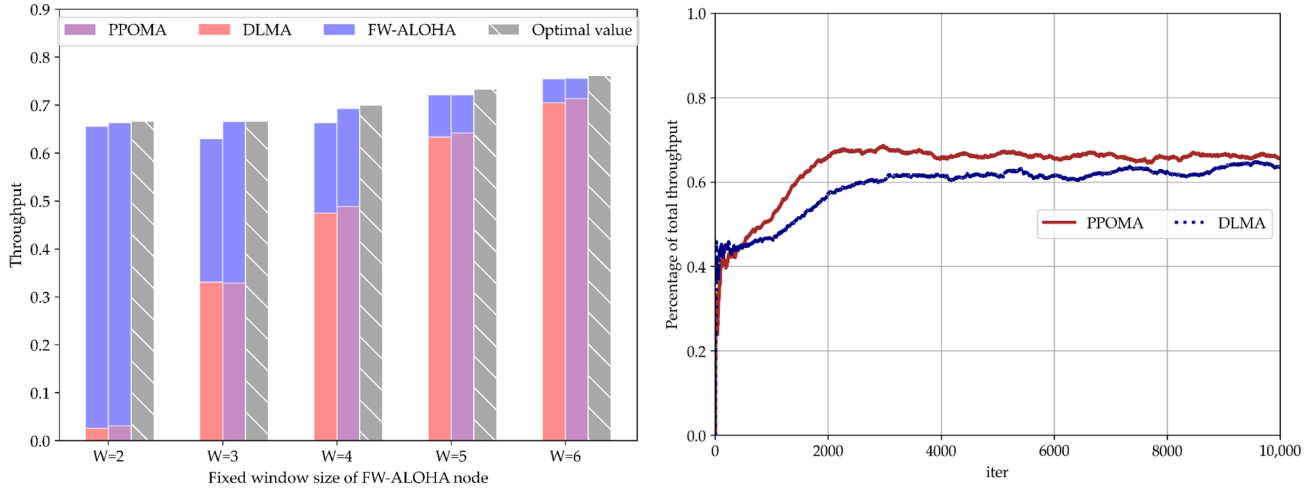


Figure 5. Coexistence of PPOMA Nodes with FW-ALOHA Nodes. (**a**) Throughput for different W; (**b**) convergence Speed for $W = 5$.

The window size, W, governs the data packet transmission frequency in the FW-ALOHA protocol. For instance, when $W = 2$, the FW-ALOHA protocol will attempt to retransmit data packets in the next one or two time slots after a collision occurs. As W increases, FW-ALOHA nodes wait for a greater number of time slots. When $W = 6$, a collision results in FW-ALOHA randomly selecting one of the next one to six time slots for retransmission. Smaller values of W lead to more frequent data packet transmissions by FW-ALOHA nodes, resulting in higher throughput. Conversely, larger W values cause occasional data packet transmissions, leading to numerous idle time slots in the channel. PPOMA nodes can seize these idle time slots to maximize the network's total throughput. Notably, a significant throughput improvement is observed only in the cases of $W = 3$ and $W = 4$, with throughputs exceeding DLMA by 0.036 and 0.03, respectively. In the other scenarios, the total throughput remains relatively similar. Experimental results indicate that PPOMA nodes can bring the network's total throughput closer to the optimal value compared to DLMA nodes.

Figure 5b displays a comparison of the convergence speed of the two types of nodes for the case of $W = 5$. It is evident that PPOMA roughly achieves convergence around the 2000th iteration, nearly 900 iterations faster than DLMA. It is worth noting that FW-ALOHA's selection of waiting time slots, w, is random, which makes collisions unavoidable, resulting in greater fluctuations in its throughput curve.

Figure 6a depicts the coexistence of PPOMA nodes with EB-ALOHA protocol nodes, with theoretical optimal values of 0.785, 0.846, 0.882, 0.905, and 0.92 for different window sizes (W = 2, 3, 4, 5, and 6). Here, the EB-ALOHA nodes are configured with a maximum backoff level $m_{max} = 2$, meaning the window size can increase to up to four times its original value. The window sizes W are employed to regulate the frequency of data packet transmissions in the EB-ALOHA protocol, similar to FW-ALOHA in some respects. Experimental results demonstrate that when coexisting with EB-ALOHA nodes, regardless of the chosen W value, the throughput of EB-ALOHA nodes remains close to zero, and the overall system's throughput is predominantly determined by the PPOMA nodes. In this scenario, the throughput of PPOMA nodes significantly outperforms that of DLMA nodes.

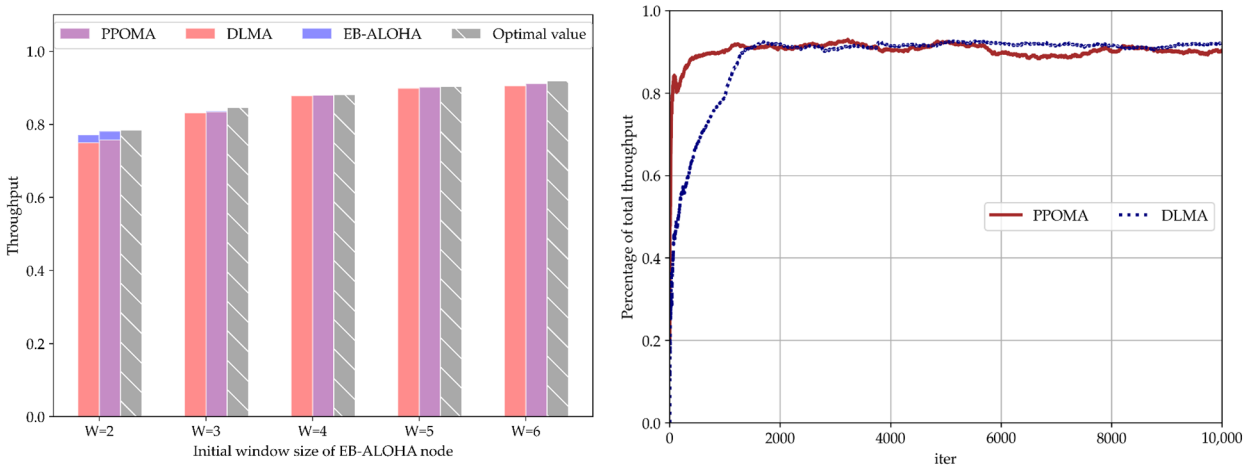


Figure 6. Coexistence of PPOMA Nodes with EB-ALOHA Nodes. (**a**) Throughput for different W; (**b**) convergence Speed at $W = 6$.

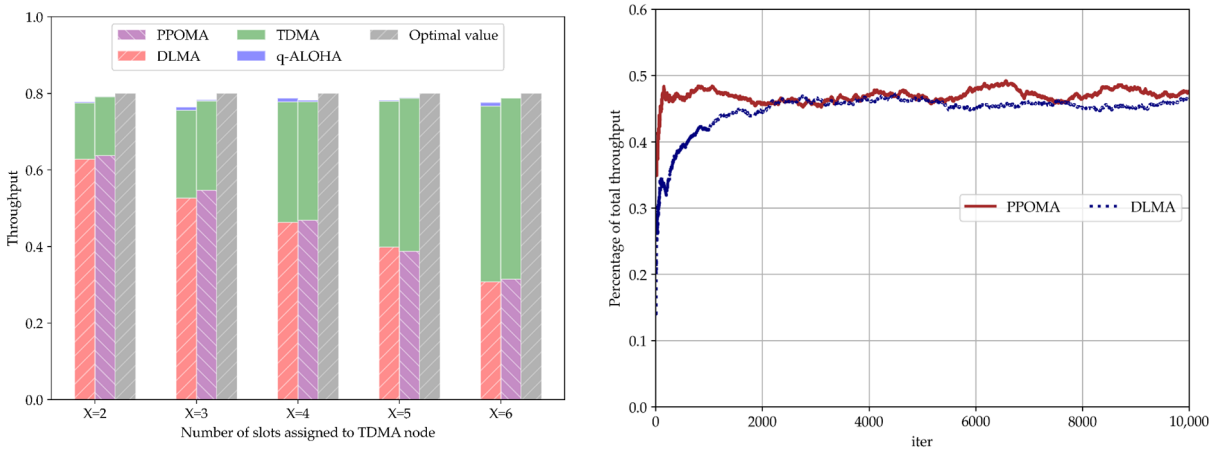
This discrepancy primarily arises from the consideration of the backoff level m in EB-ALOHA. By comparing Figure 6a with Figure 5a, we observe that the presence of m results in a considerably lower number of data packet transmissions by EB-ALOHA nodes for the same window size. Specifically, when $W = 2$, if EB-ALOHA experiences two consecutive collisions, its maximum window size would increase to $4W = 8$, which is equivalent to setting $W = 8$ in FW-ALOHA (indicating a significantly lower data packet transmission frequency). Therefore, even if EB-ALOHA and FW-ALOHA both use the same W value, the final throughput results differ significantly. Figure 6b displays the comparison of convergence speed between the two types of nodes when $W = 6$. In this case, PPOMA reaches convergence approximately 300 iterations earlier than DLMA.

5.3.3. Coexistence of PPOMA Nodes with TDMA and ALOHA Nodes

Finally, we consider the coexistence of PPOMA nodes with both a TDMA node and a q-ALOHA node and conduct simulation analysis for two specific scenarios: (1) The transmission probability q of the q-ALOHA protocol remains constant at 0.2, while the number of time slots allocated to TDMA transmission, $X = \{2, 3, 4, 5, 6\}$, varies. In this case, we primarily investigate the impact of different TDMA time slot numbers X on the total throughput when all three coexist. (2) The number of time slots allocated to TDMA transmission remains constant at three, but the transmission probability q for the q-ALOHA protocol varies as $q = \{0.1, 0.2, 0.5, 0.7, 0.8\}$. In this scenario, we examine the effect of varying q values in q-ALOHA on the total throughput when all three coexist. Similar to the previous sections, we use the theoretically derived maximum throughput values from the sensor node for comparison. The theoretical maximum values for the two scenarios are as follows: (1) 0.8, 0.8, 0.8, 0.8, 0.8; (2) 0.9, 0.8, 0.5, 0.58, 0.62.

Figure 7a addresses the first scenario in which we compared the throughput of PPOMA and DLMA at different X values. In this scenario, at different X values, the throughput of PPOMA is higher than that of DLMA by 0.01, 0.021, 0.006, 0.011, and 0.007, respectively. As can be observed from the figure, it is evident that the variation in X values impacts the total throughput. As X increases, more time slots become available for TDMA data transmission. Consequently, PPOMA considers not only the possibility of q-ALOHA sending data within a time slot but also that TDMA will transmit data at fixed time slots. Taking these factors into account, PPOMA ultimately achieves throughput levels close to the optimum. Since we set a low transmission probability q for q-ALOHA in this scenario, even if it chooses to send, it often results in collisions due to the other two nodes also transmitting in the same time slot, making q-ALOHA's throughput almost imperceptible in the graph. Figure 7b illustrates a comparison of the convergence speeds between the two nodes when $q = 0.2$ and

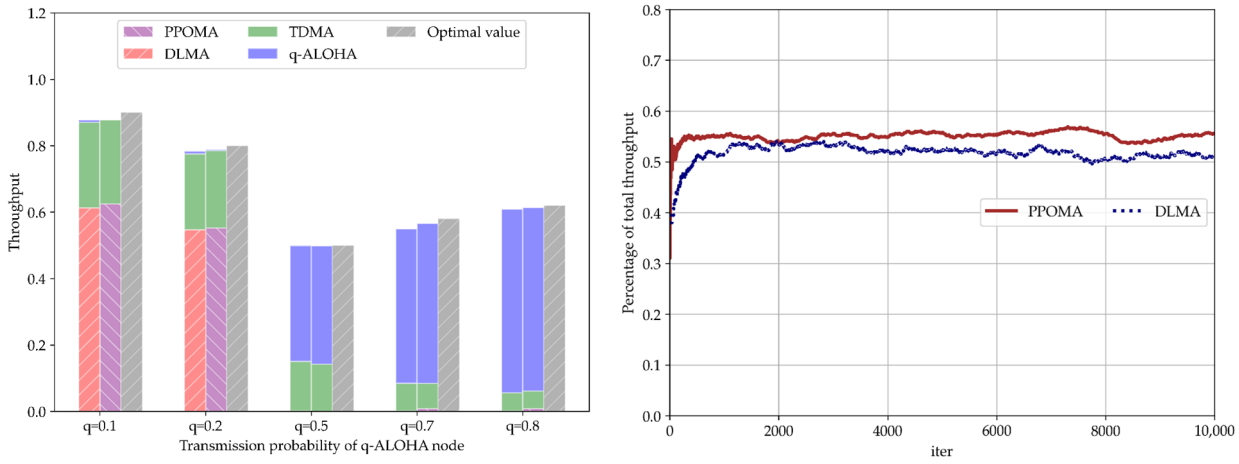
$X = 4$. It is apparent from the graph that PPOMA reaches convergence in approximately 400 iterations, while DLMA requires around 2000 iterations to converge. In a multi-node mixed scenario, PPOMA achieves faster convergence compared to DLMA.



(**a**) (**b**)

Figure 7. Coexistence of PPOMA Nodes with TDMA and q-ALOHA Nodes (Scenario 1). (**a**) $q = 0.2$, $X = \{2,\ 3,\ 4,\ 5,\ 6\}$; (**b**) $q = 0.2$, $X = 4$ convergence speed.

In Figure 8a, the second scenario is investigated. Here, when q is set to 0.1 and 0.2, PPOMA outperforms DLMA by 0.012 and 0.005, respectively. As can be observed from the figure, it is evident that changes in the q value have an impact on the total throughput. An increase in q implies a higher probability of q-ALOHA transmitting during each time slot. Similar to previous discussions, when q is relatively small (as in the cases of $q = 0.1$ and $q = 0.2$), the throughput contributed by q-ALOHA is close to zero. It is only when q values increase that q-ALOHA starts to exhibit higher throughput.



(**a**) (**b**)

Figure 8. Coexistence of PPOMA Nodes with TDMA and q-ALOHA Nodes (Scenario 2). (**a**) $X = 3$, $q = \{0.1,\ 0.2,\ 0.5,\ 0.7,\ 0.8\}$; (**b**) $X = 3$, $q = 0.2$ convergence speed.

Additionally, increasing q values also affects the throughput of TDMA nodes. Specifically, by comparing the two scenarios with larger q values (0.7 and 0.8) to those with smaller q values (0.1 and 0.2) in Figure 8a, it can be observed that higher q values lead to a reduction in TDMA node throughput. This is because when q-ALOHA has a higher q value, it is more likely to send data regardless of whether the current time slot is occupied by TDMA. If the current time slot is not occupied by TDMA, the higher transmission probability increases

the network throughput without affecting the TDMA throughput. However, if the current time slot is occupied by TDMA, the higher transmission probability increases the chances of collision between two nodes, ultimately decreasing TDMA throughput.

Figure 8b depicts the comparison of convergence speeds between PPOMA and DLMA under the conditions of $X = 3$ and $q = 0.2$. Here, PPOMA achieves convergence in around 400 iterations, while DLMA requires approximately 1200 iterations to reach convergence.

Overall, in both of these scenarios, the PPOMA protocol approaches the optimal throughput more closely than the DLMA protocol and achieves convergence faster.

5.4. Comparison of Total Throughput under Imperfect Channel Conditions

In the previous section, our analysis was based on an idealized scenario, assuming communication under perfect channel conditions. However, in real communication environments, various interference factors can lead to data loss. Therefore, we need a more realistic model for the multiple access problem.

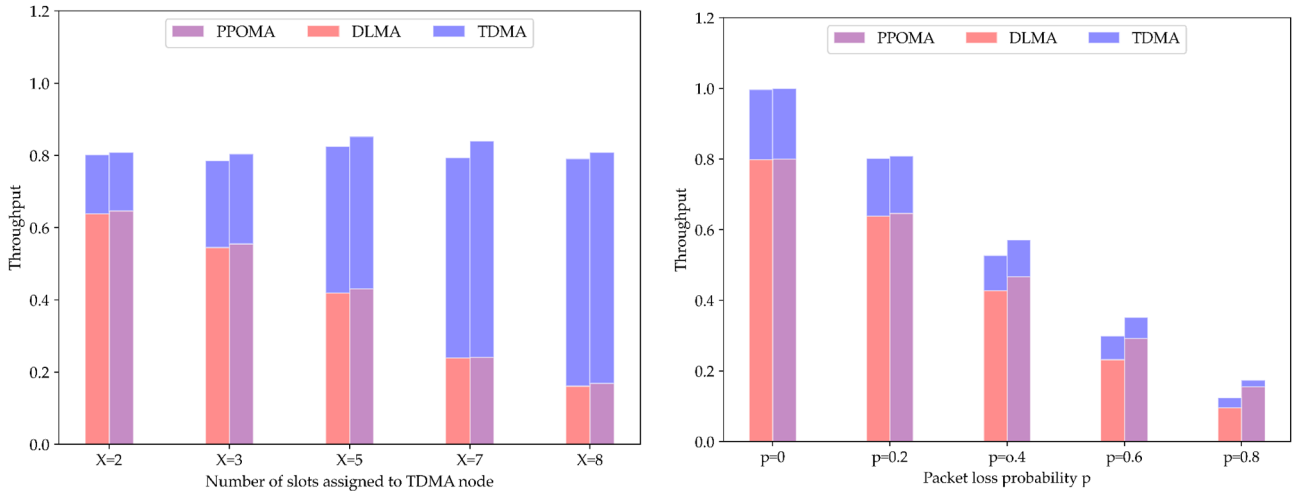
In this section, we specifically focus on the multiple access problem in an imperfect channel environment. In the context of imperfect channels, we have examined the potential impact of interference factors, such as noise, on data transmission. Unlike perfect channels, imperfect channel conditions introduce not only the possibilities of successful data transmission and collision-induced failures, but also the probability of data loss. This introduces an added layer of complexity to the data transmission status in imperfect channel scenarios. Specifically, regardless of whether PPOMA nodes coexist with any other nodes, we introduce a new variable, "p", to represent the probability of data loss when sending packets. This is because in real-world communication, due to the presence of noise and interference, there is a certain probability, p, that data transmission may fail. We set this probability p to various values, including 0%, 20%, 40%, 60%, and 80%, where 0% represents the ideal scenario of a perfect channel condition, while the other values represent different degrees of imperfections in the channel environment.

Figure 9 provides a detailed exploration of how PPOMA and TDMA nodes coexist in an imperfect channel environment. In Figure 9a, we observed the total throughput for TDMA with varying slot numbers X, while keeping the data loss probability p fixed at 0.2. As the number of slots allocated to TDMA, denoted as X, increases, the throughput of TDMA nodes also increases, while the throughput of PPOMA nodes decreases. This happens because PPOMA nodes have learned a channel access strategy to send data packets when TDMA slots are idle, optimizing the overall throughput. However, due to the probabilistic nature of data transmission, the total throughput cannot reach the optional value of one, as seen in perfect channel conditions.

Figure 9b illustrates the scenario with $X = 2$, while varying data loss probability p. From the graph, it is evident that there is a difference in throughput between PPOMA and DLMA. At $p = 0$, representing a perfect channel condition, both types of nodes maximize the total network throughput. However, as the channel environment deteriorates, with increasing data loss probability p, our PPOMA protocol performs better in achieving maximum network throughput. Specifically, when p = 0.2, 0.4, 0.6, and 0.8, PPOMA's throughput is higher than DLMA by 0.008, 0.04, 0.06, and 0.059, respectively. These results emphasize the superior performance of PPOMA in an imperfect channel environment, particularly in scenarios with high data loss probabilities, where its throughput outperforms DLMA.

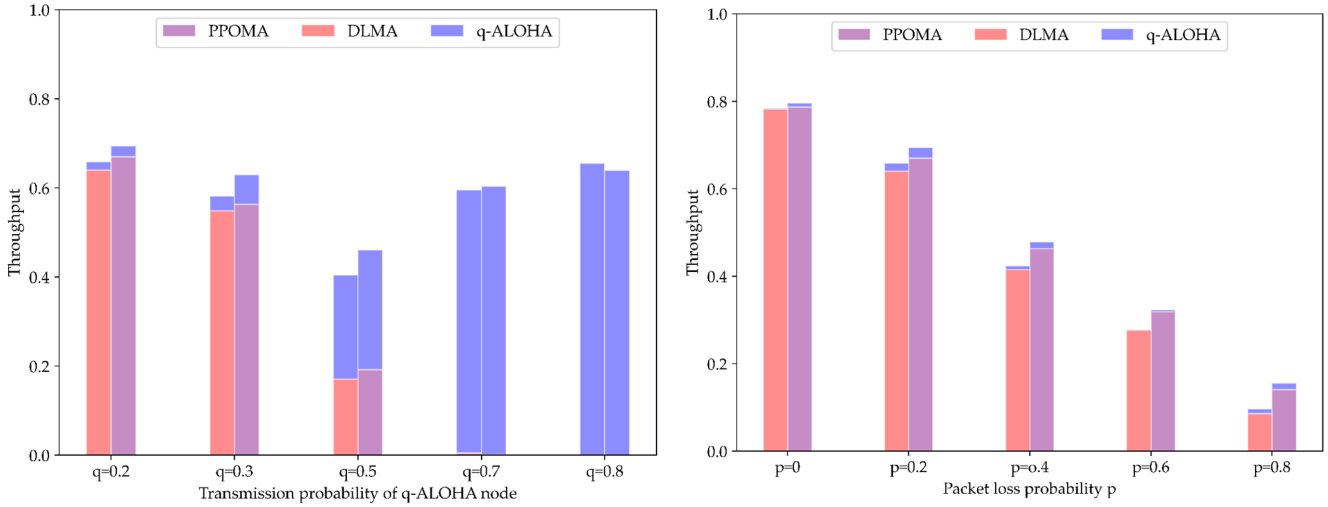
Figure 10 explores the throughput results of the coexistence of PPOMA and q-ALOHA nodes in an imperfect channel environment. In Figure 10a, when the data loss probability is fixed at $p = 0.2$, we compare the total throughput under different sending probabilities q for q-ALOHA. These observations are similar to the coexistence scenario of PPOMA and q-ALOHA nodes under perfect channel conditions. Specifically, when the sending probability q is low, PPOMA nodes tend to send as many data packets as possible to fully utilize each time slot, maximizing the total throughput. However, when the sending probability q is high, PPOMA nodes choose to avoid sending data packets as much as possible to reduce conflicts with q-ALOHA nodes, aiming to maximize the total throughput. Nevertheless,

due to the impact of data loss probability, the total throughput cannot reach the optimal level observed under perfect channel conditions.



(**a**) (**b**)

Figure 9. Coexistence of PPOMA and TDMA nodes in an imperfect channel. (**a**) When $p = 0.2$, $X = \{2, 3, 5, 7, 8\}$; (**b**) when $X = 2$, $p = \{0, 0.2, 0.4, 0.6, 0.8\}$.



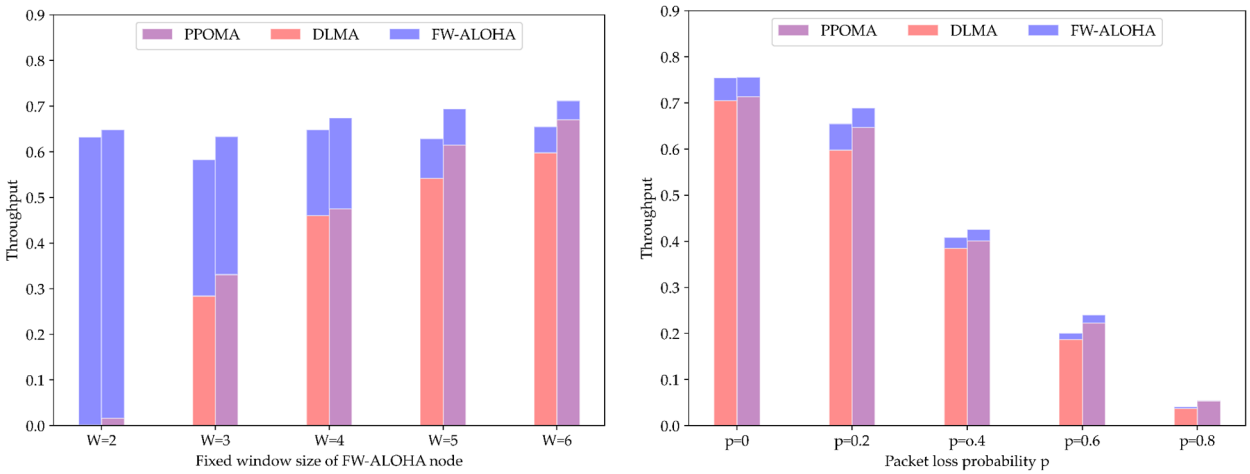
(**a**) (**b**)

Figure 10. Coexistence of PPOMA and q-ALOHA nodes in an imperfect channel. (**a**) When $p = 0.2$, $q = \{0.2, 0.3, 0.5, 0.7, 0.8\}$; (**b**) when $q = 0.2$, $p = \{0, 0.2, 0.4, 0.6, 0.8\}$.

In Figure 10b, we keep the sending probability q fixed at 0.2 and study the impact of different data loss probabilities p on the total throughput. It is evident from the graph that, when facing an imperfect channel environment, PPOMA outperforms DLMA. Under data loss probabilities of p = 0.2, 0.4, 0.6, and 0.8, PPOMA's total throughput is higher than DLMA by 0.03, 0.048, 0.042, and 0.055, respectively. These results highlight the superior performance of PPOMA in an imperfect channel environment, especially in scenarios with high data loss probabilities, where its throughput significantly exceeds that of DLMA.

Figure 11 presents the throughput performance of PPOMA and FW-ALOHA nodes in an imperfect channel environment. In Figure 11a, we study the total throughput of FW-ALOHA under different window sizes W when the data loss probability is fixed at $p = 0.2$. It is worth noting that when focusing on the throughput of PPOMA under different W values, it is clear that PPOMA achieves higher throughput compared to DLMA. Particularly,

in the cases of $W = 5$ and $W = 6$, PPOMA's throughput is higher than DLMA by 0.053 and 0.049, respectively.

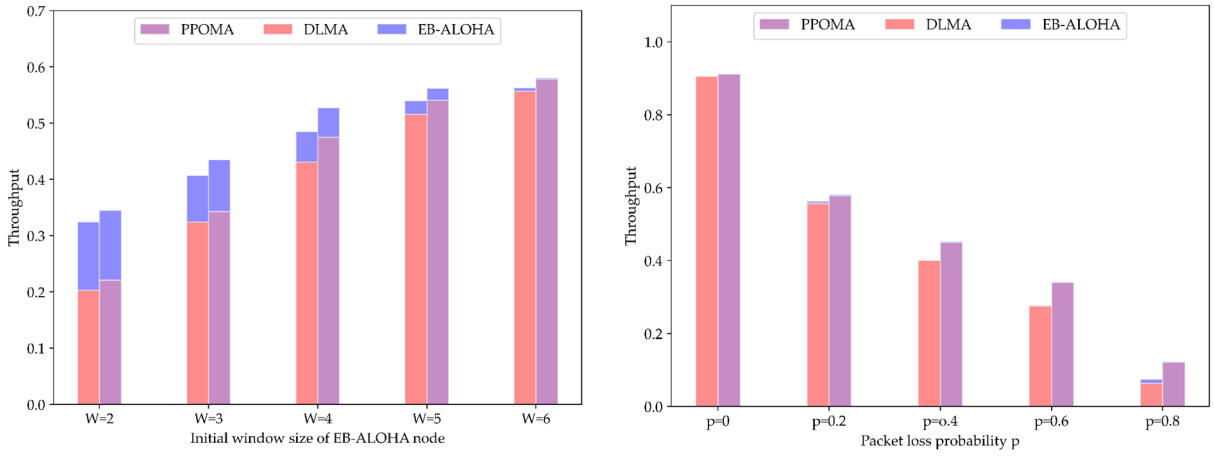


(**a**) (**b**)

Figure 11. Coexistence of PPOMA and FW-ALOHA nodes in an imperfect channel. (**a**) When $p = 0.2$, $W = \{2, 3, 4, 5, 6\}$; (**b**) when $W = 5$, $p = \{0, 0.2, 0.4, 0.6, 0.8\}$.

Figure 11b further demonstrates the impact of different data loss probabilities p on throughput when W is fixed at five. In a perfect channel environment, the total throughput of both types of nodes is similar. However, once the imperfect channel condition is considered, PPOMA's superiority becomes evident. As the data loss probability p increases, the total throughput decreases, but our PPOMA achieves higher total throughput. For instance, at $p = 0.2$, PPOMA outperforms DLMA by 0.049, and the gap increases to 0.036 at $p = 0.6$. These results emphasize the outstanding performance of PPOMA in an imperfect channel, especially in scenarios with high data loss probabilities, where its throughput significantly exceeds that of DLMA.

Figure 12 investigates the coexistence of PPOMA nodes with EB-ALOHA nodes in an imperfect channel environment. In Figure 12a, when the data loss probability is fixed at $p = 0.2$, we observe the cases with different window sizes W for EB-ALOHA, specifically {2, 3, 4, 5, 6}. The results show that as W increases, the total network throughput also increases. However, at the same time, the proportion of throughput contributed by EB-ALOHA gradually decreases and even approaches zero.

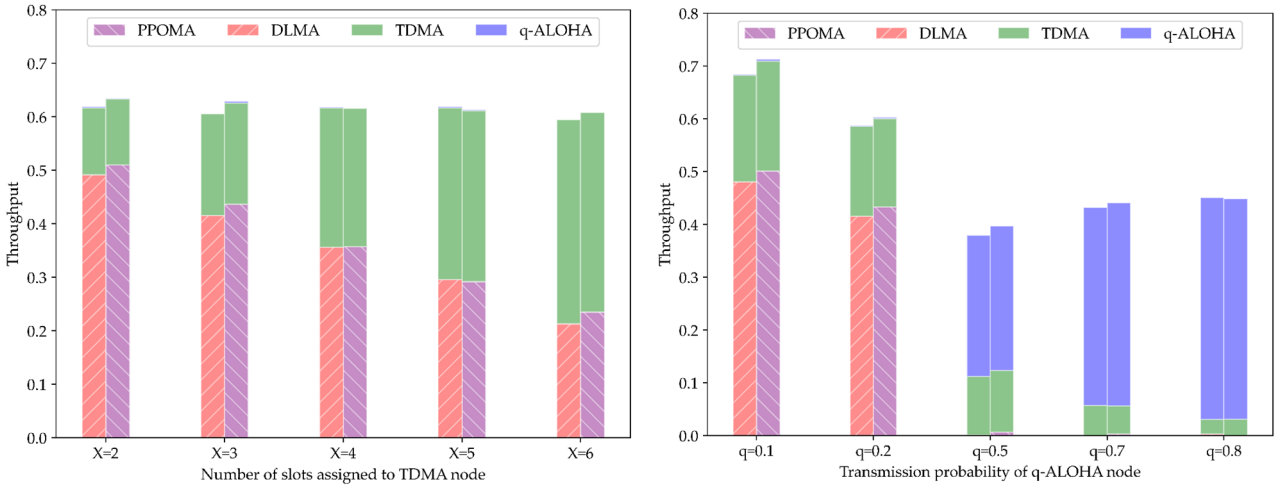


(**a**) (**b**)

Figure 12. Coexistence of PPOMA and EB-ALOHA nodes in an imperfect channel. (**a**) When $p = 0.2$, $W = \{2, 3, 4, 5, 6\}$; (**b**) when $W = 6$, $p = \{0, 0.2, 0.4, 0.6, 0.8\}$.

In Figure 12b, we study the impact of different data loss probabilities p on the total network throughput when W is fixed at six. It can be observed that as the data loss probability p increases, the total network throughput sharply decreases. Nevertheless, the PPOMA protocol is capable of maintaining a relatively high total throughput. This indicates that in an imperfect channel condition, PPOMA outperforms DLMA. These series of observations emphasize the superiority of PPOMA in dealing with imperfect channel environments.

Figure 13 explores the coexistence of PPOMA nodes with a TDMA node and a q-ALOHA node in an imperfect channel environment. With a data loss rate of $p = 0.2$, different q and X values were set for comparison. Specifically, in Figure 13a, the sending probability q for q-ALOHA was fixed at 0.2, primarily examining the impact of the number of TDMA slots X on throughput. The experimental results indicate that as X increases, the proportion of TDMA in the total throughput gradually rises, while the proportion of PPOMA decreases. Additionally, the throughput of q-ALOHA is almost imperceptible, primarily due to the very low q value, which leads to collisions even if q-ALOHA attempts to transmit data, resulting in transmission failures.



(**a**) (**b**)

Figure 13. Coexistence of PPOMA nodes with TDMA and q-ALOHA nodes in an imperfect channel. (**a**) When $q = 0.2$, $X = \{2, 3, 4, 5, 6\}$; (**b**) when $X = 3$, $q = \{0.1, 0.2, 0.5, 0.7, 0.8\}$.

In Figure 13b, with X fixed at three, the impact of different q values on throughput is considered. When q is low, the total network throughput is primarily contributed by TDMA and PPOMA. When q is high, the total network throughput is mainly divided between TDMA and q-ALOHA. This is because, with a high q value, PPOMA opts to reduce its data transmission to maximize the overall network throughput.

In summary, the results from Figure 13 demonstrate that in an imperfect channel environment, different q and X values have varying impacts on throughput allocation. However, our PPOMA consistently outperforms DLMA under all conditions.

To delve deeper into the discussion of how different data loss rates p affect the throughput when three types of nodes coexist, we chose to further investigate the case with $q = 0.2$ and $X = 3$, as depicted in Figure 14. In this graph, we display the total network throughput under various p values. By comparing the data in the graph, it is evident that the total network throughput decreases as the data loss rate p increases.

However, it is worth noting that even under high data loss rates, PPOMA continues to exhibit better performance than DLMA. Specifically, in scenarios with p = 0.2, 0.4, 0.6, and 0.8, PPOMA's throughput is higher than DLMA by 0.021, 0.034, 0.035, and 0.029, respectively. These results underscore the superior performance of PPOMA when facing

different data loss rates, especially in high data loss rate scenarios, where its throughput significantly outperforms DLMA.

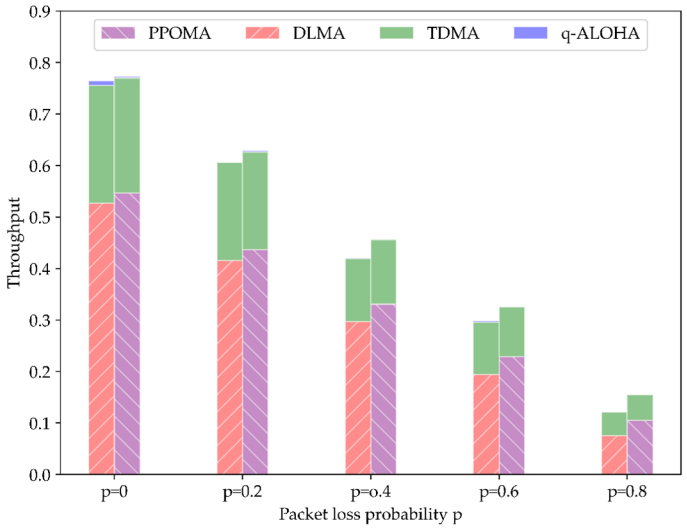


Figure 14. When $q = 0.2$, $X = 3$, $p = \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8\}$.

6. Conclusions

In this research, we propose an innovative multiple access protocol, namely PPOMA, which leverages the advantages of the PPO algorithm from the field of deep reinforcement learning. This protocol excels not only in accelerating convergence but also in possessing robust sampling and exploration capabilities. These remarkable features collectively position PPOMA as an outstanding solution to address the various challenges posed by multiple access in heterogeneous wireless networks. These challenges include navigating complex communication environments, handling high-dimensional state spaces, and addressing intricate decision-making scenarios. The cumulative impact of these advantages significantly elevates the overall performance level of the entire network. Through a series of experiments, we rigorously evaluate the performance of the PPOMA protocol under diverse and realistic channel conditions, which encompass channel interference and packet loss. In a perfect channel scenario, PPOMA's advantage over other algorithms is primarily evident in its convergence speed, reaching convergence on average 500 iterations faster. In an imperfect channel scenario, PPOMA's advantage is mainly reflected in its higher overall network throughput, with an approximate increase of 0.04. The comparative analysis between PPOMA and existing multiple access protocols indicates that, under various channel conditions, PPOMA achieves higher overall throughput and faster convergence. This is attributed to the superior performance of the underlying PPO algorithm employed by PPOMA compared to the DQN algorithm. These findings strongly align with the essential requirements of modern communication systems, where efficiency and reliability are paramount. Consequently, we firmly believe that PPOMA possesses significant potential and wide-ranging applicability within the realm of multiple access protocols. This research thus opens doors to advancements in wireless network performance, promising a more efficient and dependable communication environment. Nevertheless, there are limitations to our approach. While PPOMA attains almost optimal throughput, it comes at the cost of fairness among nodes. In future research, we are keen to utilize PPOMA to specifically tackle the crucial issue of fairness among network nodes.

Author Contributions: Conceptualization, Y.X. and Z.W.; methodology, J.L.; software, J.L.; validation, J.S. and T.Z.; formal analysis, T.W.; investigation, Z.W.; resources, Y.X.; data curation, T.W. and T.Z.; writing—original draft preparation, T.W. and A.P.; writing—review and editing, Y.X. and A.P.; visualization, T.W. and J.S.; supervision, Z.W.; project administration, Y.X.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Project of State Grid Zhejiang Electronic Power Co., Ltd., grant number B311XT230019.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Xu, Y., Lou, J., Wang, T. and Shi, J were employed by the company State Grid Zhejiang Electronic Power Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Agiwal, M.; Roy, A.; Saxena, N. Next Generation 5G Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1617–1655. [CrossRef]
2. Odarchenko, R.; Iavich, M.; Iashvili, G.; Fedushko, S.; Syerov, Y. Assessment of Security KPIs for 5G Network Slices for Special Groups of Subscribers. *BDCC* **2023**, *7*, 169. [CrossRef]
3. Patel, N.J.; Jadhav, A. A Systematic Review of Privacy Preservation Models in Wireless Networks. *Int. J. Wirel. Microw. Technol.* **2023**, *13*, 7–22. [CrossRef]
4. Peha, J.M. Sharing Spectrum Through Spectrum Policy Reform and Cognitive Radio. *Proc. IEEE* **2009**, *97*, 708–719. [CrossRef]
5. Ali, Z.; Naz, F.; Javed; Qurban, M.; Yasir, M.; Jehangir, S. Analysis of VoIP over Wired & Wireless Network with Implementation of QoS CBWFQ & 802.11e. *Int. J. Comput. Netw. Inf. Secur.* **2020**, *12*, 43–49. [CrossRef]
6. Hao, H.; Wang, Y.; Shi, Y.; Li, Z.; Wu, Y.; Li, C. IoT-G: A Low-Latency and High-Reliability Private Power Wireless Communication Architecture for Smart Grid. In Proceedings of the 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Beijing, China, 21–23 October 2019; pp. 1–6. [CrossRef]
7. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
8. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
9. Lample, G.; Chaplot, D.S. Playing FPS Games with Deep Reinforcement Learning. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 10827. [CrossRef]
10. Zhu, P.; Dai, W.; Yao, W.; Ma, J.; Zeng, Z.; Lu, H. Multi-Robot Flocking Control Based on Deep Reinforcement Learning. *IEEE Access* **2020**, *8*, 150397–150406. [CrossRef]
11. Tung, T.-Y.; Kobus, S.; Roig, J.P.; Gunduz, D. Effective Communications: A Joint Learning and Communication Framework for Multi-Agent Reinforcement Learning Over Noisy Channels. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2590–2603. [CrossRef]
12. Zhang, L.; Tan, J.; Liang, Y.-C.; Feng, G.; Niyato, D. Deep Reinforcement Learning-Based Modulation and Coding Scheme Selection in Cognitive Heterogeneous Networks. *IEEE Trans. Wirel. Commun.* **2018**, *18*, 3281–3294. [CrossRef]
13. Mota, M.P.; Araujo, D.C.; Costa Neto, F.H.; De Almeida, A.L.F.; Cavalcanti, F.R. Adaptive Modulation and Coding Based on Reinforcement Learning for 5G Networks. In Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6. [CrossRef]
14. Zhou, M.; Wei, X.; Kwong, S.; Jia, W.; Fang, B. Rate Control Method Based on Deep Reinforcement Learning for Dynamic Video Sequences in HEVC. *IEEE Trans. Multimed.* **2021**, *23*, 1106–1121. [CrossRef]
15. He, C.; Hu, Y.; Chen, Y.; Zeng, B. Joint Power Allocation and Channel Assignment for NOMA with Deep Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2200–2210. [CrossRef]
16. Lei, W.; Ye, Y.; Xiao, M. Deep Reinforcement Learning-Based Spectrum Allocation in Integrated Access and Backhaul Networks. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 970–979. [CrossRef]
17. Xiong, X.; Zheng, K.; Lei, L.; Hou, L. Resource Allocation Based on Deep Reinforcement Learning in IoT Edge Computing. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1133–1146. [CrossRef]
18. Huang, J.; Yang, Y.; He, G.; Xiao, Y.; Liu, J. Deep Reinforcement Learning-Based Dynamic Spectrum Access for D2D Communication Underlay Cellular Networks. *IEEE Commun. Lett.* **2021**, *25*, 2614–2618. [CrossRef]
19. Wang, Y.; Li, X.; Wan, P.; Shao, R. Intelligent Dynamic Spectrum Access Using Deep Reinforcement Learning for VANETs. *IEEE Sens. J.* **2021**, *21*, 15554–15563. [CrossRef]
20. Zheng, Z.; Jiang, S.; Feng, R.; Ge, L.; Gu, C. Survey of Reinforcement-Learning-Based MAC Protocols for Wireless Ad Hoc Networks with a MAC Reference Model. *Entropy* **2023**, *25*, 101. [CrossRef]
21. Yu, Y.; Wang, T.; Liew, S. Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1277–1290. [CrossRef]
22. Yu, Y.; Liew, S.C.; Wang, T. Multi-Agent Deep Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks with Imperfect Channels. *IEEE Trans. Mob. Comput.* **2022**, *21*, 3718–3730. [CrossRef]

23. Kaur, A.; Thakur, J.; Thakur, M.; Kumar, K.; Prakash, A.; Tripathi, R. Deep Recurrent Reinforcement Learning-Based Distributed Dynamic Spectrum Access in Multichannel Wireless Networks with Imperfect Feedback. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 281–292. [CrossRef]
24. Naparstek, O.; Cohen, K. Deep Multi-User Reinforcement Learning for Dynamic Spectrum Access in Multichannel Wireless Networks. In Proceedings of the GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–7. [CrossRef]
25. Xu, Y.; Yu, J.; Headley, W.C.; Buehrer, R.M. Deep Reinforcement Learning for Dynamic Spectrum Access in Wireless Networks. In Proceedings of the MILCOM 2018—2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, USA, 29–31 October 2018; pp. 207–212. [CrossRef]
26. Chang, H.-H.; Song, H.; Yi, Y.; Zhang, J.; He, H.; Liu, L. Distributive Dynausingmic Spectrum Access Through Deep Reinforcement Learning: A Reservoir Computing-Based Approach. *IEEE Internet Things J.* **2018**, *6*, 1938–1948. [CrossRef]
27. Zhang, X.; Chen, P.; Yu, G.; Wang, S. Deep Reinforcement Learning Heterogeneous Channels for Poisson Multiple Access. *Mathematics* **2023**, *11*, 992. [CrossRef]
28. Ma, R.T.B.; Misra, V.; Rubenstein, D. An Analysis of Generalized Slotted-Aloha Protocols. *IEEEACM Trans. Netw.* **2009**, *17*, 936–949. [CrossRef]
29. Sutton, R.S.; Barto, A.G. *Reinforcement Learning, Second Edition: An Introduction*; MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0-262-35270-3.
30. Grondman, I.; Busoniu, L.; Lopes, G.A.D.; Babuska, R. A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2012**, *42*, 1291–1307. [CrossRef]
31. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust Region Policy Optimization. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 1 June 2015; pp. 1889–1897. [CrossRef]
32. Keras: Deep Learning for Humans. Available online: https://keras.io/ (accessed on 19 October 2023).
33. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
34. Yu, Y.; Wang, T.; Liew, S.C. Model-Aware Nodes in Heterogeneous Networks. Available online: https://github.com/YidingYu/DLMA/blob/master/DLMA-benchmark.pdf (accessed on 15 October 2023).