Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 19(2): 2056–2094.
DOI: 10.3934/mbe.2022097
Received: 31 October 2021
Accepted: 16 December 2021
Published: 27 December 2021

Research article

Online machine learning algorithms to optimize performances of complex wireless communication systems

Koji Oshima[1,2], **Daisuke Yamamoto**[2], **Atsuhiro Yumoto**[2], **Song-Ju Kim**[3], **Yusuke Ito**[2] **and Mikio Hasegawa**[2,*]

[1] Innovation Design Initiative, National Institute of Information and Communications Technology, Koganei, Tokyo, Japan

[2] Department of Electrical Engineering, Tokyo University of Science, Katsushika, Tokyo, Japan

[3] SOBIN Institute LLC, Kawanishi, Hyogo, Japan

* **Correspondence:** Email: hasegawa@ee.kagu.tus.ac.jp.

Abstract: Data-driven and feedback cycle-based approaches are necessary to optimize the performance of modern complex wireless communication systems. Machine learning technologies can provide solutions for these requirements. This study shows a comprehensive framework of optimizing wireless communication systems and proposes two optimal decision schemes that have not been well-investigated in existing research. The first one is supervised learning modeling and optimal decision making by optimization, and the second is a simple and implementable reinforcement learning algorithm. The proposed schemes were verified through real-world experiments and computer simulations, which revealed the necessity and validity of this research.

Keywords: wireless communication systems; machine learning; cross layer optimization; optimization algorithm; cognitive radio; complex systems; multi-armed bandit problem; reinforcement learning

1. Introduction

Recent advancements in wireless communication technologies have led to widespread deployment and use of technologies worldwide. Along with widespread usage, the requirements and use cases are becoming higher and broader, leading to highly complex wireless communication systems.

Today's wireless communication systems are becoming large-scale networks, where numerous wireless devices are deployed in a rather small area, such as massive IoT, and where the structures of networks are becoming more heterogeneous: for example, multiple radio access technologies are simultaneously used in a mobile terminal, multiple radio transmission ranges are deployed, and various requirements for application traffic are required. The complexity of wireless communication systems comes from not only the interaction among the functions inside wireless devices but also the radio interferences among wireless devices. Moreover, the number of parameters to be controlled in a device is large because of the multiple layered structure of radio systems, which prevents the operators or systems from finding the optimal values in a straightforward manner.

One of the sources of the complexity is the layered structure of wireless devices. The physical layer encodes and decodes wireless signals and handles the wireless link through transmission power control. The MAC layer deals with the management of wireless links by choosing an appropriate modulation and coding scheme (MCS). It also manages access to wireless resources through frequency domain duplex (FDD), time domain duplex (TDD), or carrier-sensing multiple access with collision avoidance (CSMA/CA). The network layer provides routing management and labelling of network addresses, typically the internet protocol (IP). The transport layer handles the transportation of data on communication networks through a 2-way protocol such as the transport control protocol (TCP) or 1-way protocol such as the user datagram protocol (UDP). The session, presentation, and application layers are based on user demand for the application of communication networks. These structures, namely the open systems interconnection model (OSI model) of seven abstraction layers, are designed as flexible communication systems to manage and meet various communication needs and demands with scalability. In these structures, the functions of different layers are designed to be more independent and less combined to maintain flexibility and scalability. It provides flexible wireless communication systems such as IEEE 802.11 wireless local area network (WLAN), 3GPP long-term evolution (LTE), or 5G, which can use common TCP/IP protocols on the Internet.

In contrast, because each layer is designed independently, it is important but difficult to optimize end-to-end communication performance as a whole system. It can be easily understood if you consider that the number of wireless parameters is large: modulation scheme of radio signals, selection of radio frequency and bandwidth, transmission power, and access scheme to wireless resources. In addition, as mentioned above, higher-layer protocols should be included to optimize the user experience. This means that the conventional optimization of the performance of wireless communication systems can no longer be applied because it is based on a mathematical formulation that is too complex to be realized. It is essential to solve this issue to optimize the performance of future wireless systems such as Beyond 5G/6G, which is becoming increasingly complex.

The aim of this study is to develop a novel engineering scheme to overcome this problem by introducing machine learning technologies. Machine learning technologies are data-driven modeling approaches that have been recently developed in various fields of society. In the field of wireless communications, some applications of machine learning technologies have been researched recently, such as supervised learning for signal processing and deep learning to predict the traffic demand and emphasize the quality of experience (QoE). It is expected that as more data are obtained and more parameters can be controlled, intelligent wireless communication systems using machine learning can perform better and achieve superior performance in various environments and use cases. Deep learning technologies [1] are novel applications and superior tools for data-driven modeling of various systems and for various purposes. It is an application of supervised learning and recently became famous for image recognition and Google AlphaGo [2], in which a combination of deep learning and Q-learning with deep reinforcement learning (DRL) was used. While many studies have recently applied DRL for the wireless communication field, there are still challenges and open issues.

In this study, we discuss the application of machine learning technologies to wireless communication systems in terms of performance optimization. Existing research is classified along with the viewpoint of the amount of information available and the method of decision for optimal action. Then, we propose a new scheme for optimizing the performance of complex wireless communication systems using machine learning technologies. Another scheme, to provide a feasible solution for mobile wireless devices that have rather limited resources, is proposed as a simple reinforcement learning-based optimal decision.

The rest of the paper is organized as follows. Starting with the indication of issues of current and future wireless communication systems in Section 2, various approaches for optimal decision making by machine learning-based modeling are reviewed in Section 3, and the proposed schemes are introduced. In Section 4, one of the proposed schemes, supervised learning modeling and optimization algorithm, is elaborated, and some experimental results are presented. In Section 5, another proposed scheme, a simple reinforcement learning-based optimal decision making method using an MAB algorithm, is elaborated, and some experimental results are presented. Finally, in Section 6, the conclusion and some remarks are described.

2. Issue of current and future wireless systems

Owing to the advancements in radio and communication technologies, today's wireless communication systems provide high-speed data transfer and a wide area of communication links, while they have become very complex systems. Wireless systems, including wired systems, are generally composed of multiple layers of multiple functions, such as the physical layer, MAC layer, network layer, transport layer, and application layer. Each layer has different protocols. It leads to a difficulty in modeling wireless systems.

5G systems [3] have been deployed in several regions in the world very recently. Compared to 4G/LTE or 3G, the major difference in 5G systems is that there is no single technology to develop such as orthogonal frequency division multiple access(OFDMA) in 4G/LTE or code division multiplexing (CDMA) in 3G. A 5G system network consists of a radio access network (5G NR) and a core network (5G CN). For 5G NR, physical transmission technologies such as OFDM(A) in the higher frequency band such as 60 GHz are proposed, and massive multiple-input-multiple-output (MIMO) with beam foaming. For 5G CN, several network management technologies are required such as network function virtualization (NVF), software defined network (SDN), edge computing, and network slicing. Along with these, the 5G system is more comprehensive: it can include a legacy 4G/LTE system and licensed assisted access (LAA) / licensed shared access (LSA) assuming the usage of the industrial, scientific, medical (ISM) band frequency which requires a certain channel access scheme for sharing spectra with other wireless systems like listen before talk (LBT).

From the viewpoint of performance requirements, there are three major requirements for 5G systems: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low latency communications (URRC). Fundamentally, these are different aspects of requirements. Indeed, as use cases, various independent scenarios are proposed based on these requirements: a rapid download of large data such as movie data of some gigabytes (GB) by satisfying eMBB, management of massive amounts of mobile sensors such as IoT devices by satisfying the mMTC, and telemedicine with extreme low latency wireless network by satisfying URRC.

To satisfy these extreme requirements of 5G systems, high-performance devices/equipment are needed in wireless communication systems which can allow management of low-end devices, such as IoT. All these requirements and use cases indicate that current and future wireless communication systems are becoming more and more complex and heterogeneous. Authors of [4] indicate that these viewpoints are applied in 6G, and propose communication networks using artificial intelligence (AI).

2.1. Research challenge to optimize future wireless communication system

Numerous studies have focused on seeking the optimal decision of wireless communication systems through mathematical and theoretic formulation of wireless channel and transmission power control [5–10], modulation and coding, the behavior of MAC protocol [11] or higher layer protocols. The common approach of these studies is to define the mathematical model of the function of wireless communication system, to formulate the maximization or minimization problem, and to obtain the optimal solution by solving the problem.

An advantage of these "classical" approaches is that the theoretical optimal solutions or parameters, or at least their upper- or lower-bounds, can be obtained under the assumption of the continuity of the function described.

In contrast, in terms of the modeling of wireless communication systems, the classical approach generally focuses on a certain layer performance such as channel capacity in physical layer or throughput in MAC layer, and cannot cover whole system modeling. Indeed, current and future complex wireless communication systems are hard to be described mathematically as a whole system. Moreover, the mathematical formulation of wireless systems cannot always be applied to time-varying situations. Wireless systems such as Wi-Fi or Bluetooth are operated autonomously and interact with each other over time, which is beyond the description of mathematical modeling. These indicate that the classical approach faces fundamental difficulty in applying the optimization of modern wireless communication systems.

In short, there are two issues in optimizing the performance of current and future complex wireless communication systems.

- Issue 1. Classical one-way optimal decisions are not suitable for complex time-varying situations.
- Issue 2. Classical modeling by mathematical formulation cannot be applied to multi-layer and multiprotocol complex wireless systems.

2.2. Cognitive radio technology

Cognitive radio [12, 13] is a fundamental concept in wireless systems. It learns the environment and behavior of wireless communication nodes and performs trial-and-error to seek the best action to improve the performance. It can adapt to various changes in the environment, such as sudden increases or decreases in traffic demand, variations in wireless channels, and contention among wireless transmitters. Figure 1 shows the concept of cognition cycle [12]. The key idea of cognitive radio is the cognitive cycle [13]: learning and action, and their feedback cycle.

The concept of cognitive radio gives insights on facing the issues mentioned above. Important points are feedback cycle and learning. In the next section, we discuss those by applying machine learning technologies.

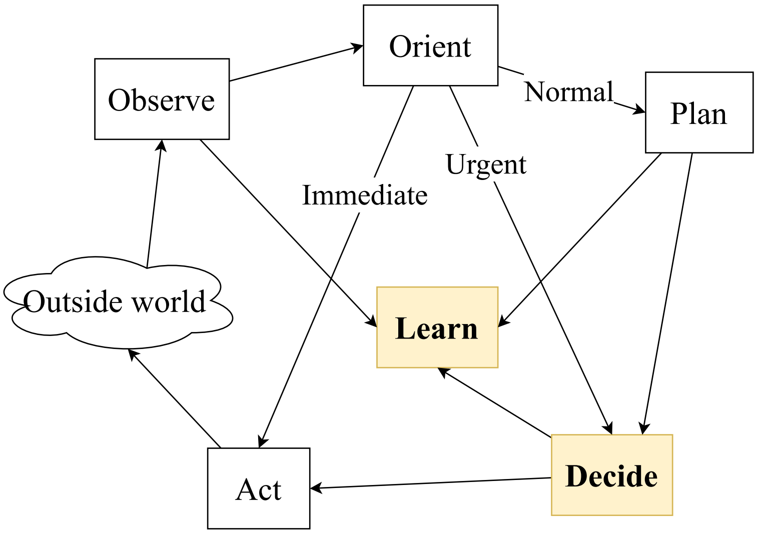


Figure 1. Concept of cognitive radio.

3. Machine learning for wireless system

Machine learning technologies are becoming a more and more important solution for various issues in current society. They provide us with a strong tool to build a whole system model using a large collection of data in a wireless communication system and wireless communication network. Machine learning technologies, including deep learning, have been increasingly researched in the field of communication technology recently [14–17]. Several studies indicate that the management of 5G/Beyond 5G systems requires machine learning technologies [18, 19].

In reference [4], AI-enabled intelligent architecture for 6G is proposed. Their proposed architecture is divided into four layers: intelligent sensing layer, data mining and analytics layer, intelligent control layer and smart application layer. Between them, intelligent control layer consists of learning, optimization, and decision-making. They indicate that significantly dynamic and complex network in 6G cannot be optimized through traditional mathematical algorithms. Our research is based on the same viewpoint, and proposes schemes to optimize such complex networks by using machine learning technologies.

There are two points of view when using machine learning to optimize the decisions and actions of wireless communication systems. One point is the amount of data. Supervised learning, especially deep learning and its related methods, can deal with a large amount of data to extract the characteristics of the system from which the data are collected. If the amount of data is limited, the reinforcement learning approach would be more suitable. It allows one to make the optimal decision through the iteration of trial-and-error cycles under the environment of limited information and parameters to be controlled.

Another point is to how to decide the optimal action to achieve higher performance. There are two strategies: decision by learning scheme or by optimization algorithm, namely maximization or minimization of a formula. If changes in the environment surrounding wireless communication systems are relatively slow, and if the relation between parameters and performance is continuous, not discrete, then an optimization algorithm would be suitable. Furthermore, notably, if the relation between parameters and performance is discrete, the reinforcement learning approach would be suitable.

Figure 2 shows the relation of these approaches and examples of application of optimizing wireless communication systems. In this study, we investigate two panels of this figure. The figure on the lower-right pane is based on modelling by supervised machine learning and decisioning through an optimization algorithm. The one on the upper-left pane is simple reinforcement learning. We focus on a multi-armed bandit (MAB) problem formulation for this approach.

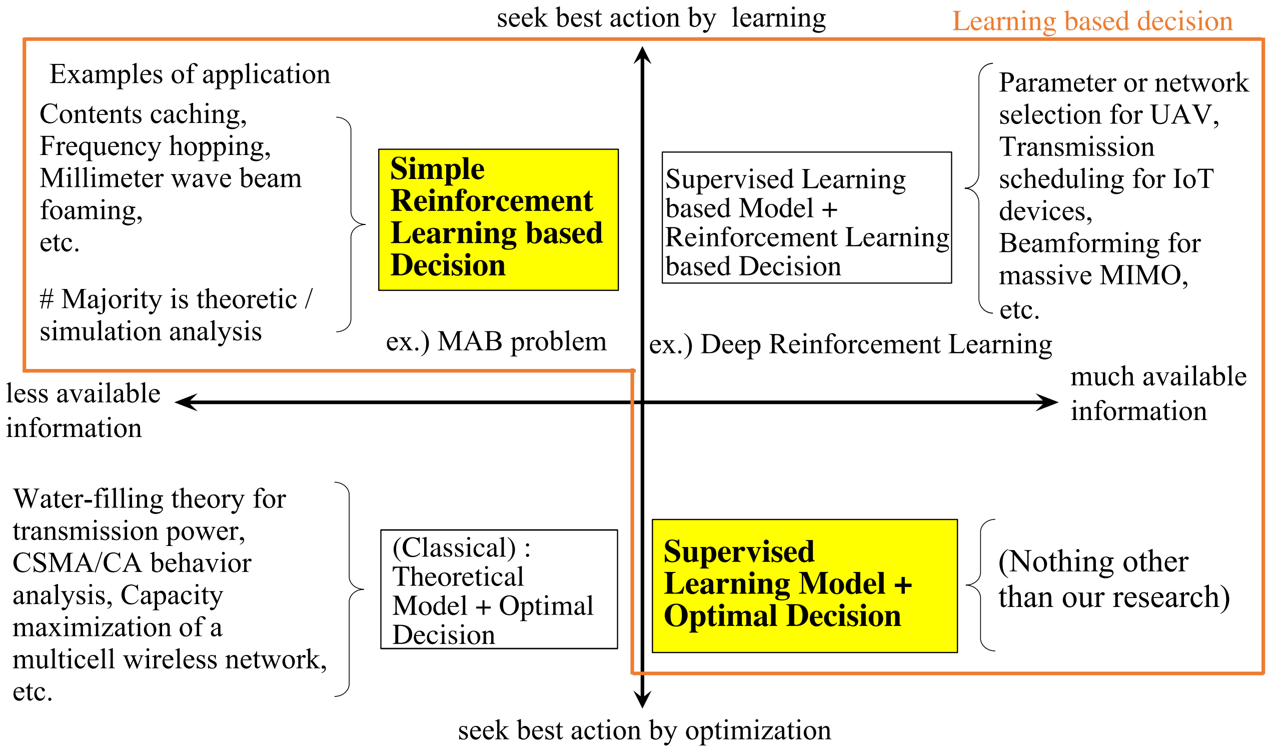

seek best action by learning
Learning based decision
Examples of application
Contents caching, Frequency hopping, Millimeter wave beam foaming, etc.
# Majority is theoretic / simulation analysis
Simple Reinforcement Learning based Decision
ex.) MAB problem
Supervised Learning based Model + Reinforcement Learning based Decision
Parameter or network selection for UAV, Transmission scheduling for IoT devices, Beamforming for massive MIMO, etc.
ex.) Deep Reinforcement Learning
less available information
much available information
Water-filling theory for transmission power, CSMA/CA behavior analysis, Capacity maximization of a multicell wireless network, etc.
(Classical) : Theoretical Model + Optimal Decision
Supervised Learning Model + Optimal Decision
(Nothing other than our research)
seek best action by optimization

Figure 2. Optimization strategy of wireless communication systems using machine learning. Examples of existing research topics are classified.

3.1. Classical theoretical model and optimal decision

The lower left panel in Figure 2 indicates the “classical” theoretical model and optimization. It uses a mathematical formulation of wireless communication systems, usually focusing on a certain layer (or two layers, such as the physical layer and MAC layer). Then, the formulation is converted to the optimization problem: maximization or minimization of the formulated equation, whose solution yields the optimal parameters of the system. The obtained optimal parameters, usually under the assumption of continuity of the function expressed in the formulated equation, are the theoretical optimal values. There are many good examples of this type of research, such as the water-filling optimization of transmission power [20].

3.2. Deep-learning–based model and reinforcement-learning–based decision

If a large amount of data of wireless communication systems is expected to be obtained, and to seek the best action by learning, the approach could be a combination of modelling by supervised learning and the decision making by reinforcement learning, as shown in the upper-right panel of Figure 2. Deep reinforcement learning (DRL) is a typical example of this approach.

Deep learning (DL) is a newly developed and rapidly spreading technique in various fields. It is an advanced form of an artificial neural network and a type of supervised learning. The first achievement of the DL was in the field of computer vision. This technique has been introduced in various layers of communication systems [21]. The early applications of DL were for the estimation of parameters of the propagation channel [22, 23] and device location estimation [24–27].

DRL is a combination of deep learning and reinforcement learning, as shown in Figure 3. DRL has been applied very recently in the field of wireless communication systems [28–34]. It can be seen as an implementation of the cognitive cycle: it learns the relationship among environment, parameter, action, and performance of the wireless communication node through deep learning. Decisioning trial-and-error: it seeks better actions through reinforcement learning. The major strong point of DRL is to build a performance model by deep learning in an online manner and to utilize it to predict the performance of the systems when certain parameters are deployed. Reinforcement learning, usually Q-learning based algorithms, are applied to seek better action by evaluating the results, updating its network, and choosing the predicted parameters better using deep learning. Note that DRL always requires the information of the state of wireless communication systems, which might be unrealistic in the real world.

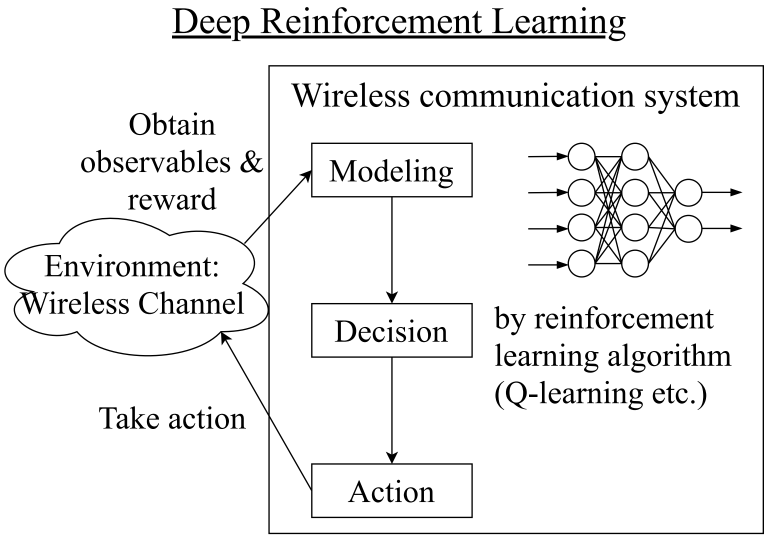

Deep Reinforcement Learning
Wireless communication system
Obtain observables & reward
Modeling
Environment: Wireless Channel
Decision
by reinforcement learning algorithm (Q-learning etc.)
Take action
Action

Figure 3. The concept of deep reinforcement learning.

3.3. Supervised-learning–based model and optimal decision

DRL is a strong technique for seeking better decisions for wireless communication systems, as described in the previous section. However, there are some assumptions and limitations. First, it requires a large amount of data owing to training in deep learning. It leads to some temporal training overhead in the real world implementation. Second, it uses reinforcement learning: even if the variables are continuous and can be optimized by minimization or maximization of continuous functions, it is forced to do trial-and-error processes. It may suffer from insufficient performance because the number of trials is finite in the real world. In other words, the DRL approach can underperform compared to the classical mathematical formulation and optimization approaches. It is because the classical scheme brings about theoretically determined parameters which will achieve maximum or at least some sort of upper-bound performance of wireless communication systems. It cannot be assured in general that the DRL approach will reach the maximum theoretical performance.

This discussion provides insights into a better solution for using machine learning technologies to optimize wireless communication systems. What if some sort of mathematical optimization can be applied to seek the best parameters while using machine learning as a tool to build the performance model of wireless communication systems? The answer proposed in this paper is the wireless communication system optimization method based on cognitive cycles using machine learning and an optimization algorithm. It uses a supervised learning algorithm to build the performance model of wireless communication systems, and defines the optimization problem using this performance model as a function of variables, observables, and performance of the systems. Then, by solving that optimization problem, the optimal parameters are obtained. After taking actions according to the optimum parameters, wireless communication systems observe the results and then update the performance model by machine learning. This feedback loop is an implementation of a cognitive cycle based on machine leaning, taking advantage of classical mathematical formulation approaches. The details of the proposed scheme are elaborated in the following section. Note that the word "optimization" used here means mathematical optimization, i.e., the maximization or minimization of certain functions and not in the strict meaning of a simple optimization problem.

3.4. Simple reinforcement learning decision

The assumption of the proposed cognitive cycle approach described in the previous section is that the communication nodes in wireless communication systems to be optimized can obtain and control various observables and parameters. However, wireless equipment in the real world, especially simple devices such as IoT sensors, cannot deal with such multiple parameters in general. Indeed, there are one or a few parameters to control, as well as observables, in typical IoT devices. In addition, the computational resources and communication frequencies are limited in such devices. Therefore, more simple, light-weight algorithms must be developed.

The multi-armed bandit (MAB) problem [35] is a simple machine learning problem, which can achieve good performance under the limitation of a finite number of trials. It is used in some areas of wireless communication systems, like a channel selection in a cognitive radio [36, 37], or resource allocation in 5G small cells [38], but the numbers of studies have been limited. In addition, very a few or none of them have performed experimental validation of the research by implementing on wireless devices.

In this study, a light-weight and high-performance algorithm is proposed using the Tug-of-War (TOW) algorithm, which was developed recently. Its performance is similar to that of well-known algorithms such as UCB-1 tuned, while the implementation is very simple [39]. Notably, TOW does not require any information on the current state of the system.

3.5. Contribution of this study

In summary, we propose two novel approaches that are different from classical mathematical optimization or current state-of-the-art deep reinforcement learning. The first proposed approach is a wireless communication system optimization method based on cognitive cycle using machine learning. It uses machine learning to build the performance model of the systems, obtain the optimal parameters by the optimization formulation, and update the performance model in an online manner. The second approach is a simple reinforcement learning, MAB problem formulation. Using a light-weight algorithm, it is more feasible in terms of the implementation and the operation of wireless devices such as IoT. Through these proposals and discussions, this paper provides a new understanding that is useful for building strategies to optimize the performance of complex wireless communication systems.

4. Optimal decision making through the cognitive cycle using the supervised-learning–based model

Wireless systems have recently become increasingly complex, which makes it difficult to build a cross-layer model, as previously mentioned. The relations between radio variables and system performance are further complicated. Consequently, the optimization of whole wireless systems through cross-layer modelling cannot be realized.

As discussed in Section 3, if sufficient data are available, using supervised machine learning technology can overcome the issue of modelling. The relationship between action and performance is learned by increasing the number of samples. Thus, the complex relations among various radio parameters and network performance can be obtained, which improves the precision of decision-making for the best performance.

This section proposes a wireless system optimization method based on the cognitive cycle using machine learning. Our approach uses an optimization algorithm to seek optimal action, in contrast to reinforcement learning in DRL, as indicated in the right-hand side of Figure 2 and Section 3.3.

4.1. Related works

Technologies for next-generation wireless networks, such as 5G, are a major topic in the field of wireless communication today. In reference [18], the possibilities of machine learning technologies for next-generation 5G networks are discussed. Supervised learning techniques can be used to support channel state estimation in MIMO systems. Unsupervised learning for cell clustering, especially in heterogeneous networks, and reinforcement learning for the decision-making process of mobile users have also been suggested. The authors of [40] discussed autonomic communications in future software-driven networks. In particular, they suggested the potential of machine learning in network optimization and the need to redesign more decentralized concepts.

In the next-generation wireless networks, networks become heterogeneous. A discussion of licensed shared access (LSA) was as follows [41–43]: 5G network nodes can use not only licensed spectra but also unlicensed bands. S. Haykin discussed the comprehensive function of a cognitive dynamic system to organize communications using both licensed and unlicensed bands [44]. The need for dynamic spectrum management using a cognitive dynamic system in 5G was discussed. In reference [45], the authors analyzed the performance optimization of heterogeneous cognitive wireless networks. A typical optimization problem of load balancing was analyzed in both the centralized and decentralized cases. In reference [46], the authors introduced machine learning in mobile terminals to optimize the aggregation method for IEEE 1900.4 [47] heterogeneous wireless networks and maximize throughput.

4.2. Cross layer modeling of wireless system

Several studies have attempted to understand the relationships between various variables and performance to optimize wireless networks [6, 10, 48–50]. These studies generally focused on the performance of a certain layer, such as channel capacity in the physical layer and throughput in the MAC layer, and do not cover higher layer application throughputs. For an example of the optimization of wireless network capacity, we refer to the resource allocation problem in [6]. In principle, assuming ideal link adaptation, the formulation of the sum capacity of a multicell wireless network is expressed as

$$C(\boldsymbol{U}, \boldsymbol{P}) = \frac{1}{N} \sum_{n=1}^{N} \log(1 + \Gamma([\boldsymbol{U}]_n, \boldsymbol{P})),$$

where N is the total number of cells, Γ is the signal-to-interference and noise ratio (SINR) at the receiver, $\boldsymbol{U}$ is the set of users simultaneously scheduled across all cells, $[\boldsymbol{U}]_n$ is the users in cell n, and $\boldsymbol{P}$ is the transmit power of the scheduled users. Then, the capacity optimization problem by resource allocation is formulated as

$$\arg \max_{\boldsymbol{U}, \boldsymbol{P}} C(\boldsymbol{U}, \boldsymbol{P}). \tag{4.1}$$

As referred to in [6], this problem is nonconvex, so the solution is not straightforward; however, this equation represents the fundamental relations among radio variables and system performance.

In another example, in [10], the optimization problem of cooperative sensing in cognitive radio networks was analyzed. This is a sensing-throughput tradeoff problem: a strict sensing policy minimizes the possibility of interference to the primary user, although the opportunities to gain more throughput would be missed, and *vice versa*. The achievable MAC layer throughputs of the secondary users R can be given as

$$R(\tau, k, \epsilon) = C_0 P(H_0) \left(1 - \frac{\tau}{T}\right) (1 - \mathbb{P}_f(\tau, k, \epsilon)),$$

where τ is the sensing time, T is the total frame time (including sensing time τ), k is the number of sensing results of sensor nodes ($1 \leq k \leq N$, N is the total number of sensor nodes), and ϵ is the threshold parameter of the energy detector at the sensor node. C_0 is the ideal throughput of secondary users if the primary user is always absent, $P(H_0)$ is the probability that the primary user is absent in the channel, and $\mathbb{P}_f$ is the probability of a false alarm. Focusing on the maximization of the secondary users' throughput, that is, the minimum probability of detection of the primary user is assumed, the sensing threshold ϵ can be given by the function of τ, k, and the received signal-to-noise ratio (SNR). Under this condition, the optimization of the throughput of secondary users is formulated as

$$\arg \max_{\tau, k} R(\tau, k). \tag{4.2}$$

Because the throughput depends on the probabilities of false alarm and detection, which depend on the SNR, Eq (4.2) can be expressed as a function of τ, k, and SNR. This formulation was examined by computer simulation, and the optimal values of τ and k for a given SNR were obtained.

The formulations of the optimization problems (4.1) and (4.2) can be generalized as follows. Let the radio parameters be p (such as $\boldsymbol{U}, \boldsymbol{P}$, or τ, k), the observed radio environment be z (such as SINR or SNR), and the system performance be y (such as capacity or throughput). Then, they can be formulated as $y = f(p, z)$, where f represents the relations among the radio parameters, environment, and performance. Then, the optimization problem is formulated as

$$\arg\max_{\boldsymbol{p}} E(\boldsymbol{y}) = \arg\max_{\boldsymbol{p}} E(f(\boldsymbol{p}, \boldsymbol{z})), \tag{4.3}$$

where $E(\boldsymbol{y})$ is the utility function of throughputs, for example, the summation of the expected throughput of each node. By solving the above equation, the optimal set of parameters ($\boldsymbol{p}$) required to maximize the network performance is obtained. This can be achieved if the relation between the inputs and output is mathematically described.

In recent wireless systems, however, the situation has become more complicated. As mentioned above, modern wireless systems are equipped with various technologies for each layer. Some systems transmit signals on a single carrier with frequency hopping and others on a multicarrier with OFDM. The channel access of one protocol is TDMA, and that of the others is CSMA/CA. In general, wireless communication applications use higher-layer protocols such as IP, TCP, or UDP. Therefore, we need to consider various observables $\boldsymbol{z}$ and parameters $\boldsymbol{p}$. Moreover, the relations among these variables and network performance are hardly known. Consequently, the mathematical formulation of function f cannot be realized.

Machine learning technologies, which have the fundamental characteristics of data-driven modeling, aid in this difficulty. By using them, the hidden and complex relations among various wireless observables and parameters and network performance can be obtained. We propose a generalized cross-layer modeling of wireless system performance using machine learning. In the proposed model, $E(\boldsymbol{y})$ can denote the utility of the whole system performance, including the application. $\boldsymbol{p}$ denotes various layer parameters, and $\boldsymbol{z}$ denotes various observables. The optimization method using the proposed modeling is described in the next subsection.

4.3. Cognitive cycle and optimization for complex wireless systems

Cognitive radio is the concept of an intelligent radio that can learn from its past experience and autonomously decide its actions suitable for radio environments and needs for communication. The cognitive cycle [12] is a feedback cycle of observation, learning, decision making, and action. Haykin proposed a more concrete process of cognitive radio in [13] from an engineering perspective. He addressed the following fundamental tasks for a cognitive radio: radio-scene analysis, channel-state estimation, transmit-power control, and dynamic spectrum management. Wireless network nodes can change the radio parameters of transmission and reception to avoid interference among users and improve communication quality.

In general, wireless communication requires learning to establish wireless links and satisfy communication qualities. For example, a radio frequency (RF) module controls the coding rate based on the received signal strength indicator (RSSI) to reduce the error probability of wireless links. This means that the RF module learns the relationship between the inputs (RSSI, coding rate) and output (link quality). In cognitive radio networks, the cognitive engine should determine and coordinate the actions of the cognitive radio based on the learning of the environment. The relationship between inputs and outputs becomes more complicated in cognitive radio networks owing to its flexibility, such as software-defined radios. Cognitive radio can control various parameters such as frequency, channel, coding rate, and transmission power. The relationship between these parameters and the performance of wireless communication is hardly formulated. Thus, machine learning technologies, which can learn the complex, non-linear relationships among various information, would be the solution. By combining the concept of cognitive cycle and optimal decisioning, we propose the concept of the entire system, as depicted in Figure 4.

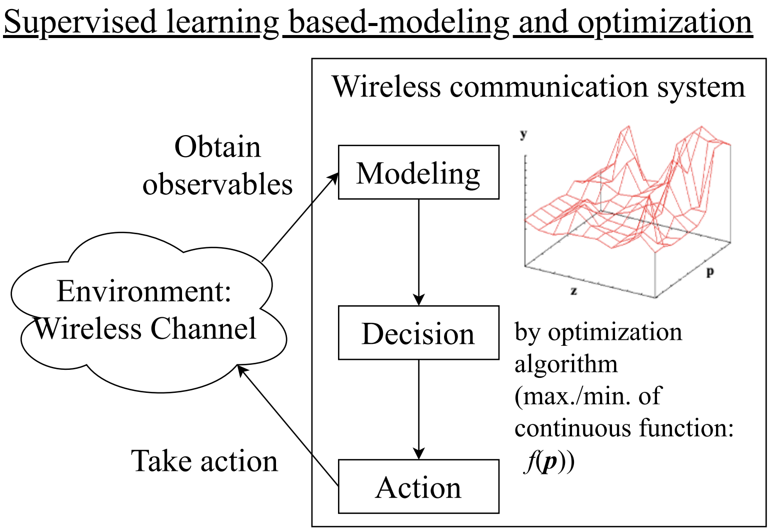

Supervised learning based-modeling and optimization
Wireless communication system
Obtain observables
Modeling
y
z
p
Environment: Wireless Channel
Decision
by optimization algorithm (max./min. of continuous function: f(p))
Take action
Action

Figure 4. The proposed supervised-learning–based modeling and optimization-based decision making scheme.

Figure 5 elaborates the proposed wireless system optimization method using machine learning. The observables of environment z are collected, which include not only the radio status but also MAC statistics, or higher layer statistics. For $\boldsymbol{p}$, various parameters of the wireless node or network were considered. Besides these variables, network performance y is observed. They are a set of samples, S, for a machine-learning algorithm:

$$S = \{(\boldsymbol{p_1}, z_1, y_1), (\boldsymbol{p_2}, z_2, y_2), ..., (\boldsymbol{p_n}, z_n, y_n)\}.$$

Using S, the cognitive engine builds and updates the model f by machine learning:

$$y = f(\boldsymbol{p}, z). \tag{4.4}$$

The updating method depends on the type of algorithm. For supervised learning, it uses S as the training data, and for unsupervised learning, it uses S for clustering or dimension reduction.

By solving the optimization problem (4.3), a cognitive engine decides the optimal action to adopt the current situation. The solution of Eq (4.3), $\boldsymbol{p}^*$, yields the optimal parameters for communication entities. After deciding the optimal action, to use parameter $\boldsymbol{p}^*$, the cognitive engine starts to reconfigure the wireless network. The necessary information is sent to communication entities.

In the following subsections, we describe examples of wireless communication systems used to evaluate our proposed method.

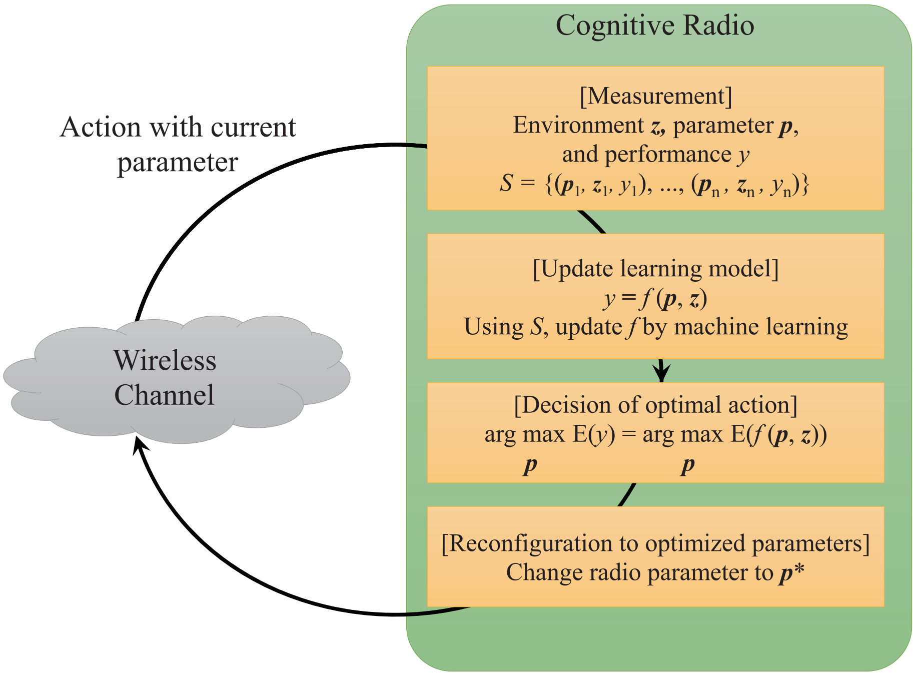

Cognitive Radio
Action with current parameter
Wireless Channel
[Measurement]
Environment z, parameter $\boldsymbol{p}$, and performance y
$S = \{(\boldsymbol{p}_1, z_1, y_1), ..., (\boldsymbol{p}_n, z_n, y_n)\}$
[Update learning model]
$y = f(\boldsymbol{p}, z)$
Using S, update f by machine learning
[Decision of optimal action]
$\arg\max_{\boldsymbol{p}} \mathrm{E}(y) = \arg\max_{\boldsymbol{p}} \mathrm{E}(f(\boldsymbol{p}, z))$
[Reconfiguration to optimized parameters]
Change radio parameter to $\boldsymbol{p}^*$

Figure 5. The proposed method based on cognitive cycle using machine learning.

4.4. Application to IEEE 802.11 WLAN

We evaluate the proposed optimization method by applying it to the IEEE 802.11 WLAN. As an example of an optimization scenario, we consider the parameter optimization of the IEEE 802.11 stations (STAs) operated in the infrastructure mode. Each STA and cognitive controller connected to access points (APs) has functions of a cognitive engine described in the previous section and runs the cognitive cycle, as mentioned below.

Each STA measures the wireless environment z, obtains the current radio parameter $\boldsymbol{p}$ and performance y, and then, adds a sample to S. z includes the radio status at the STA, such as the RSSI and wireless link quality. $\boldsymbol{p}$ includes wireless parameters such as the transmit power, operating channel, and address of the connecting AP. The uplink or downlink throughput is considered as the performance index y.

The cognitive engine in the STA updates the learning model f using S. We consider supervised learning for the evaluation. The cognitive engine builds a model that represents the relations among z, $\boldsymbol{p}$, and y from training samples s and then sends the information of the model to the cognitive controller.

The cognitive controller solves the optimization problem (4.3) using the information of the model from STAs and obtains the optimal parameters $\boldsymbol{p}^*$ of STAs. The information of the optimal parameters is sent to STAs to reconfigure the network. The STA changes its wireless parameters according to $\boldsymbol{p}^*$ and then continues the cycle starting from the measurement of the environment and performance.

4.4.1. Implementation of learning and optimization

We use support vector regression (SVR) as a learning algorithm, similar to [46]. SVR is an analog output version of support vector machines (SVMs) [51]. In SVR, the estimation function f can be expressed as [52]

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i' - \alpha_i)K(\boldsymbol{x}, \boldsymbol{x}_i) + b, \tag{4.5}$$

where l is the number of training samples, $\boldsymbol{x}_i$ is the input of the training samples ($\boldsymbol{p}$ and z), $\boldsymbol{x}$ is an unknown input set for the learning algorithm, and K is a kernel function. α_i, α_i', and b are unknown parameters obtained by the optimization technique proposed in [52], using training samples $\boldsymbol{p}$, z, and y, respectively.

We formulate the optimization problem (4.3) in the evaluation as follows:

$$\arg\max_{\boldsymbol{p}} \sum_{n=1}^{N} \log\left(1 + f(\boldsymbol{p}_n, z_n)\right), \tag{4.6}$$

where N is the number of STAs, $\boldsymbol{p}_n$ is the possible parameter set for the STA-n, z_n is the current measured quality of the radio environment at STA-n, and $f(\boldsymbol{p}_n, z_n)$ is the estimated throughput of STA-n obtained using the throughput model described above. Here, we use the logarithmic utility function of throughput considering fairness among STAs, where STAs with lower throughput have relatively larger gains for the objective function than those with higher throughput.

4.4.2. Experiments using IEEE 802.11 devices

We implement the method for the IEEE 802.11 WLAN devices. The experiments were coordinated in our university laboratory working space [53,54].

The IEEE 802.11 WLAN APs and STAs are operated in the 2.4 GHz ISM band. Laptop PCs with Ubuntu 14.04 are used as both STAs and APs. In each cognitive cycle, the STA observes the delay and packet loss ratio through pinging, RSSI from its connecting AP using the iwconfig command, the number of packets around the STA using the tcpdump command as the link quality (z), and the throughput (y) using the TCP iPerf command. The STA sets the transmission power, channel number (from 1 to 13), and data rate at the physical layer (from 6 to 54 Mb/s) for the current wireless parameters ($\boldsymbol{p}$).

The STA then builds the throughput model through SVR and sends information regarding the SVR model to its connecting AP. The AP sends it to the cognitive controller. We set up one of the APs as the cognitive controller, which calculates the optimal set of STA parameters $\boldsymbol{p}^*$ and returns the result to the AP, and then, the STA obtains the result from its connecting AP. To reduce the calculation costs for solving the optimization problem, we use the particle swarm optimization (PSO) algorithm [55,56] at the cognitive controller.

In the experiment, three APs and nine STAs are operated in channels 1, 6, and 11 in IEEE 802.11 g. The operating channel is fixed for each AP. The locations of all APs and STAs are fixed during the experiment. We use uplink TCP throughputs to evaluate the performance since uplink traffic generally makes radio resource usage more competitive in CSMA/CA. We also add background UDP traffic of approximately 8 Mb/s on channel 11. To verify the performance of the proposed system, the uplink throughput performance is compared with that of other algorithms, focusing on the selection of the connecting AP at the STA as follows: (A) selection by RSSI, (B) random selection, (C) selection by radio resource utilization, and (D) selection of the number of STAs as equally as possible among channels. In algorithm (A) using RSSI, the STA selects an AP with the highest RSSI. This seems to be a popular method for devices in the market. In algorithm (C), the STA selects the AP of a channel where the minimum number of packets is observed in each cycle. In each algorithm, each cycle runs for 30 s. All STAs start iPerf traffic of 2 s at the same time in each cycle. Before starting the proposed method, the STA observes the radio environment in each channel for 1 h and utilizes it as training data.

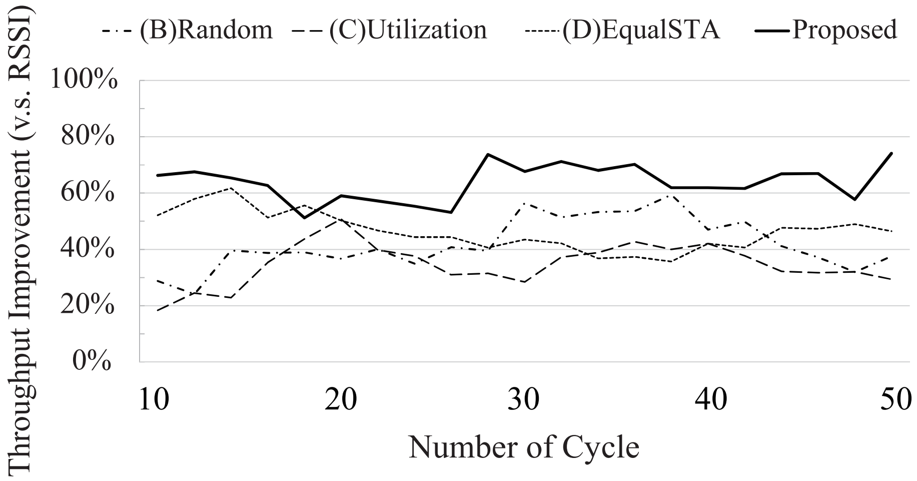


Figure 6. Improvement in throughput relative to (A) RSSI over time for each algorithm.

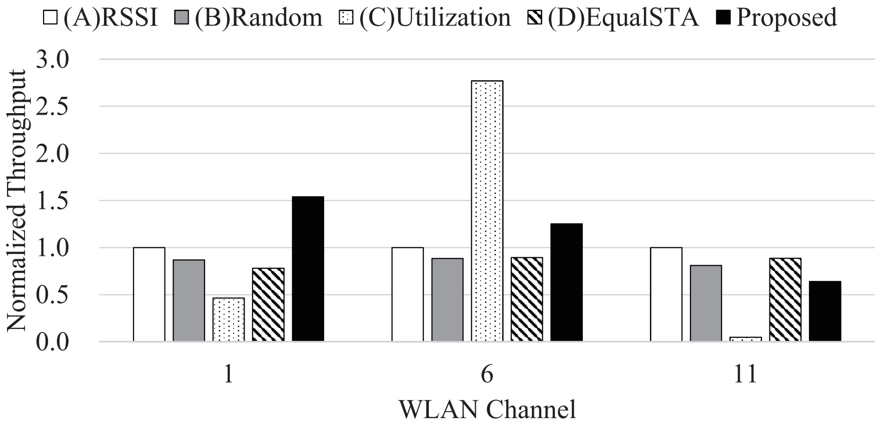


Figure 7. Average throughput for each channel. The throughput was normalized by those in (A) RSSI.

Figure 6 shows the moving average throughput by time for each algorithm. Throughputs are normalized by those in RSSI. The time is expressed as the number of cognitive cycles, and the throughput is averaged every 10 cycles (5 min). The proposed method shows a greater throughput than the other

Table 1. Simulation settings.

Parameter	Value
Area size	20 m x 20 m
Number of ieteration	10 times
Pathloss model	Free space decay
Channel model	Additive white gaussian noise (AWGN)
Traffic in proposed system	TCP of 1.4 Mbps in 6 STAs TCP of 0.7 Mbps in 15 STAs

algorithms, indicating that the STAs can select APs effectively.

Figure 7 compares the average throughput per channel among the algorithms. The utilization-based algorithm (C) shows a higher throughput at channel 6, where it is detected as the most vacant channel. However, the throughputs at the other channels are much lower. This algorithm is based on observations of a wireless environment, but neither learns nor optimizes the entire system.

In contrast, the proposed method, which has a function of learning and optimization, shows higher throughput at channels 1 and 6 and lower throughput at channel 11, which has a higher background traffic. As a whole, the proposed method can improve the network performance. These results indicate that the proposed method can build an appropriate throughput model through learning and can select the optimized wireless parameters that improve the overall network performance.

4.4.3. Evaluation by computer simulation

We also conducted a computer simulation for an extended evaluation of the proposed method [54]. The basic implementation is the same as that in the experiments already shown. The binary programs of learning and optimization were the same as those in the experimental devices. The network simulator QualNet 7.4 [57] was used as the platform for computer simulation. The number of STAs was 21, and that of APs was 3; the operating channels were 1, 6, and 11. The variables of the learning sample $(\boldsymbol{p}, z, y)$ were the same as those in the experiment conducted in the laboratory. The STAs in the proposed system send uplink TCP traffic of two types of offered loads. The background traffic is generated by constant bit rate (CBR) traffic. The offered load of channel 6 is smaller than that of channels 1 and 11. The detailed settings of simulation are shown in Table 1. As background communication nodes, three APs in each channel, five STAs in channel 1, one STA in channel 6, and seven STAs in channel 11 are set. Background CBR traffic adds 500 Kbps/STA in channel 1, 100 Kbps in channel 6, 500 Kbps/STA in channel 11.

The main difference in settings from those of the experiments is the offered load variation of the STAs in the proposed system. Similar to the experimental results, the computer simulation results show an improvement by introducing the proposed method, as shown in Figures 8 and 9. From Figure 9, the proposed cognitive cycle using machine learning can optimize the choice of channel according to the formulation of Eq (4.6).

4.5. Application for space communication

The proposed supervised learning-based optimization scheme is applied to space communication. Figure 10 shows a communication system in the space, inspired by the use case indicated in [58].

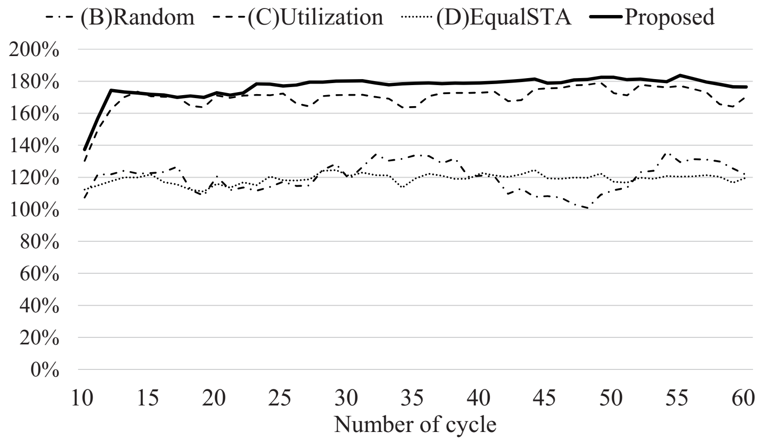


Figure 8. Improvement in throughput relative to (A) RSSI by time in each algorithm in the computer simulation.

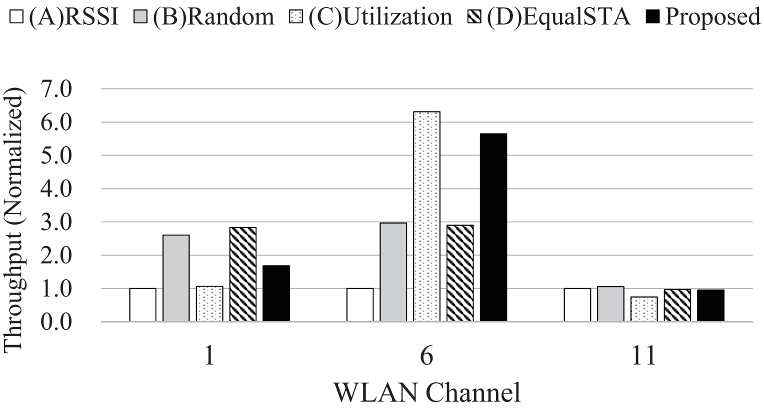


Figure 9. Average throughput at each channel in computer simulation. The throughput was normalized by those in (A) RSSI.

One of the characteristics of space communication different from terrestrial wireless communication is the large communication delay. The main component of communication delay is a propagation delay. The distance from the Earth to the Moon, for example, is 384,400 km in average, where the bidirectional wireless communication experiences at least more than 2.56 seconds. This delay scale is to be handled at higher layer than physical or MAC layers, namely, transport layer with such as TCP. However, the physical and MAC parameters have also to be taken into consideration: MCS, transmission power, etc. This is a similar situation shown in the previous subsection in the application for IEEE 802.11 devices. Therefore, in this subsection, the application of proposed supervised learning-based optimization scheme to the space communication is examined.

4.5.1. System model

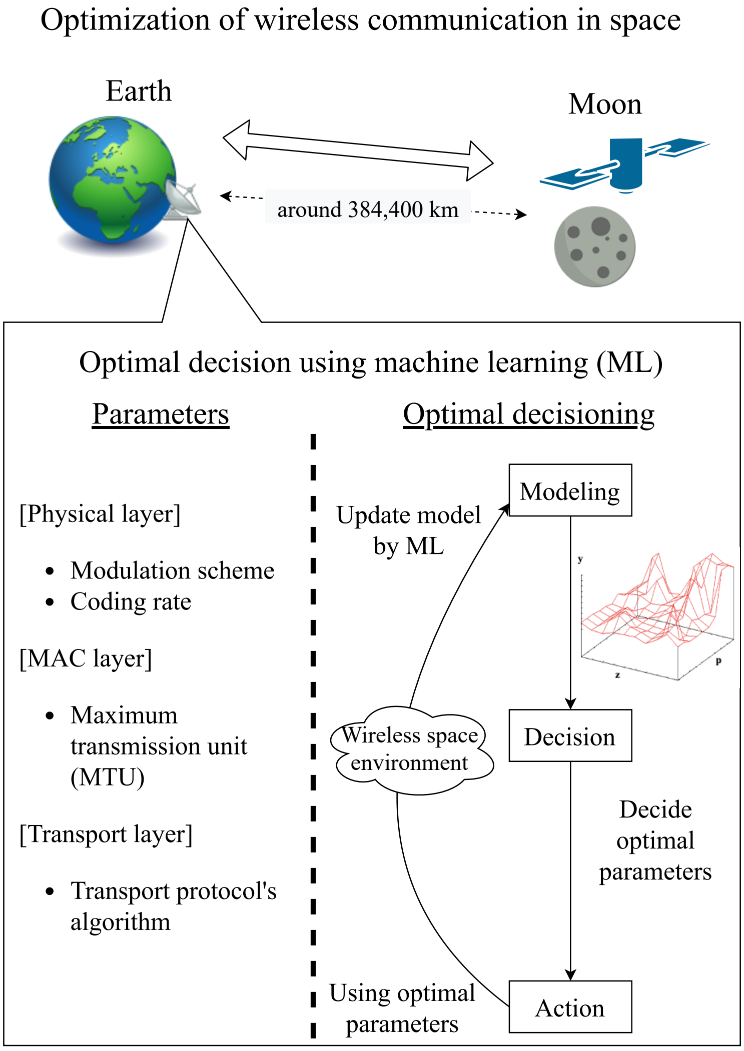

Optimization of wireless communication in space
Earth
Moon
around 384,400 km
Optimal decision using machine learning (ML)
Parameters
Optimal decisioning
[Physical layer]
• Modulation scheme
• Coding rate
[MAC layer]
• Maximum transmission unit (MTU)
[Transport layer]
• Transport protocol's algorithm
Update model by ML
Modeling
Wireless space environment
Decision
Decide optimal parameters
Using optimal parameters
Action

Figure 10. A system model to verify the proposed scheme in the future wireless communication between the Earth and the Moon as an example of proposed scheme.

Figure 10 shows the system model of evaluation. The wireless communication network is simplified to a direct communication between the ground station on the Earth and the satellite near the Moon. The ground station runs the proposed scheme: the optimal decision of communication parameters from the physical layer to the transport layer. Here, the specifications of the physical and the MAC layer are abstract models: wireless communication nodes, including the base station on the Earth and a satellite near the Moon, transmit signals using some MCSs. The set of available MCSs of the evaluation is (QPSK:1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 2/3, 3/4, 8PSK: 2/3, 3/4, 16QAM: 2/3, 3/4, 32QAM: 2/3, 3/4). Other parameters, Maximum Transmission Unit (MTU) and TCP algorithms: MTU is from 500 to 1500 Byte, and TCP algorithms are Reno, cubic, and Bottleneck Bandwidth and Round-trip propagation time (BBR) [60].

4.5.2. Implementation and evaluation

Table 2 shows the algorithms for evaluation. These correspond with the right quadrants of Figure 2, i.e., the amount of available information is large and the decision is made by a learning or optimization algorithm.

Table 2. Algorithms of modeling and decisioning using supervised learning.

ID	Modeling	Decisioning
(a)	Deep learning	Reinforcement learning: MAB with ϵ-Greedy
(b)	Support vector regression	Reinforcement learning: MAB with ϵ-Greedy
(c)	Support vector regression	Optimization (using PSO)

Desktop computer running Ubuntu 17.10 is used to emulate space communication by adding communication delay with the tc command. For the application traffic, an image file of 10 MByte is transfered by a socket program. The duration of transferring is monitored to obtain the throughput. Parameters such as delay caused by the radio propagation in space and other processing factors and packet error rate are variable. In order to train each algorithm, several pre-training with random sampling parameters was conducted before the experiments and used for each algorithm to build the model. Table 3 shows other settings of parameters.

As a reference algorithm, a deep reinforcement learning, as previously shown in Figure 3 is examined through experiments. The training data for all algorithms is obtained through 500 rounds before the experiment. The training data is composed of the time for file transfer of 10 MB, observed round trip time (RTT), selected TCP algorithm, MTU, and MCS. Parameters are randomly selected for generating the training data. The neural network is composed of three fully-connected layers, including two hidden layers and 7 and 50 neurons each, as described in the reference research [59].

Figure 11 shows the throughput results of those algorithms with a communication distance of 390,000 km. It roughly corresponds to the distance between the Earth and the Moon (384,400 km), where the throughput performances show the superiority of the proposed algorithm. The percentage values show the relative throughput increase or decrease to that of the deep reinforcement learning (DRL) algorithm (deep learning with the ϵ-Greedy algorithm). The proposed algorithm(c), using support vector regression (SVR) modeling and optimization algorithm (PSO), shows an 18% increase in throughput.

In contrast, the throughput of supervised learning modeling and reinforcement learning decision(b), i.e., SVR and MAB with the ϵ-Greedy algorithm, shows lower throughput than those of DRL(a) and proposed(c). Figure 12 shows the parameter selection of TCP and MTU in each algorithm. It suggests that the proposed algorithm selects proper TCP algorithm BBR and MTU values around 1200, while other algorithms do not select them, which brings the difference of throughputs.

Table 3. Parameters for experiments.

Parameter	Value
iterations	10 times
commnication delay	propagation delay in space
pathloss model	free space attenuation
center frequency	14.25 GHz

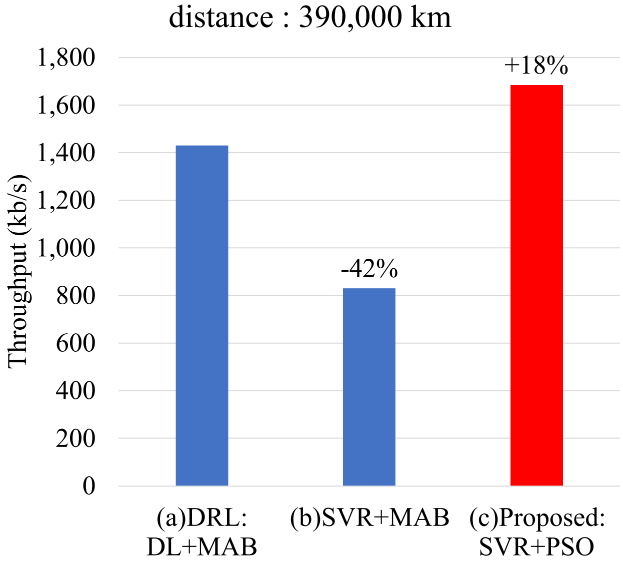


Figure 11. Throughput of each algorithm with communication distance 390,000 km.

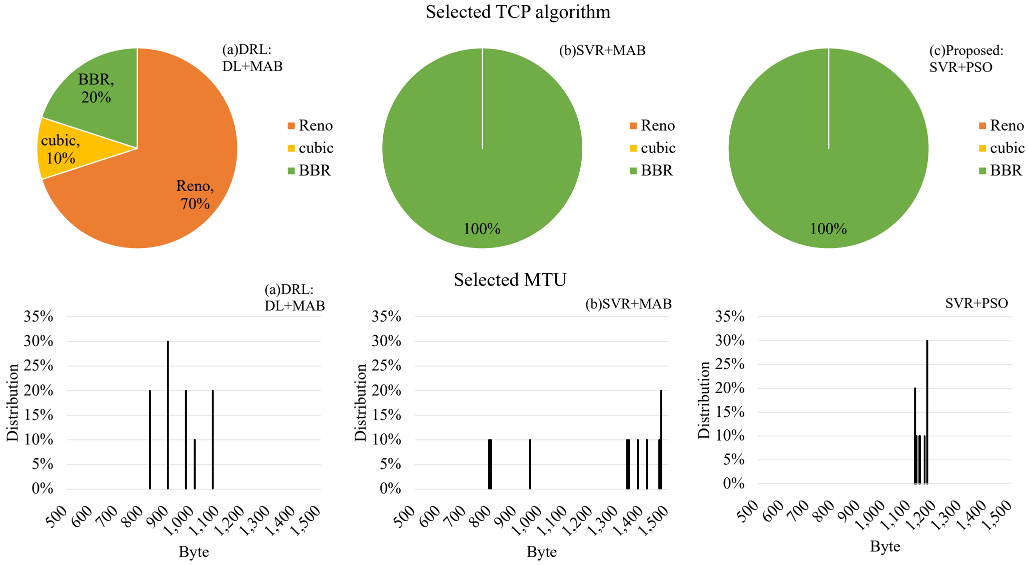


Figure 12. Parameter selection of each algorithm.

In the context of complexity and real-time issues, the authors of [61] propose modified deep reinforcement learning-based optimization for beamforming. They introduce post-decision state learning to improve learning speed. Regarding our proposed scheme, SVR can be deployed with a rather limited amount of data in general compared to deep learning. On the other hand, complexity for solving the optimization problem becomes large as the number of parameters to optimize increases in the proposed scheme. Therefore, using some mathematical optimization algorithms, such as PSO in this example, is recommended to reduce calculation costs when implemented on real devices.

5. Simple reinforcement-learning–based decision making

This section elaborates on the proposed approach of simple reinforcement-learning-based decision making, as previously indicated in the top left of Figure 2 and Section 3.4. In this section, we formulate the problem of seeking better action with simple reinforcement learning as a multi-armed bandit (MAB) problem and demonstrate the effectiveness of a novel MAB algorithm called the TOW. We examined our approach in two cases. The first is network selection in heterogeneous wireless networks [62], and the second is channel selection in massive IoT. Experimental results are shown through the implementation on real devices or through computer simulations.

5.1. Wireless system optimization as an MAB problem

The MAB problem [35] is a simple machine learning problem in which a player attempts to obtain the maximum reward from multiple slot machines. The aim of the MAB is to decide which slot machine should be selected to obtain the maximum reward through finite trials. The assumption is that the player does not have any prior information on each slot machine. The player starts to gather information on each slot machine by trying as many slot machines as possible. Then, the player estimates which slot machine has the highest expected reward and selects that slot machine to play. Through this process, the player gets more rewards. There is a trade-off between exploitation and exploration. If the player takes a long time for estimation, the player can estimate the reward more precisely, although the time for playing the selected slot machine becomes short. If the player takes only a short time for estimation, the player can take a long time to play the selected slot machine, although the reward of that slot machine might be low. Figure 13 shows the concept of the optimal decision making for the MAB problem in wireless communication systems.

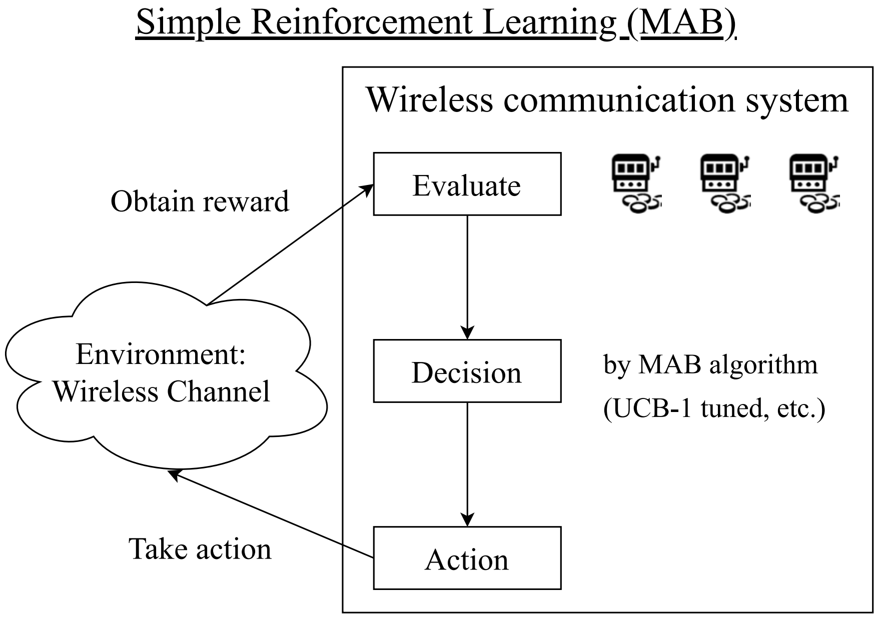


Figure 13. The concept of optimal decision making for the MAB problem.

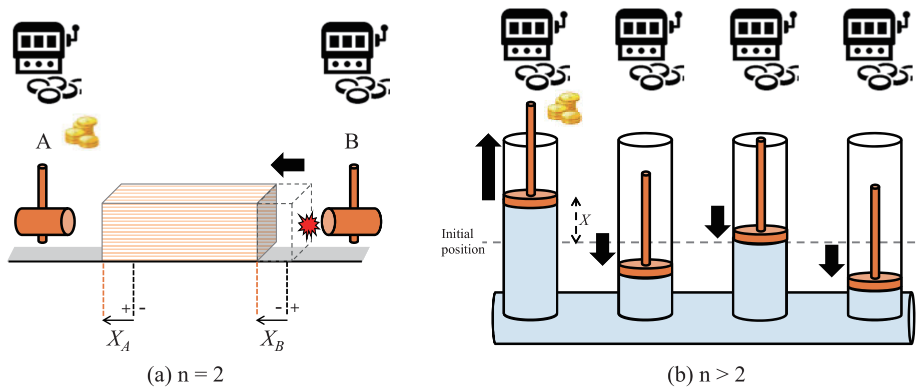
5.2. Multi-armed bandit algorithm and tug-of-war model

Several algorithms have been proposed to solve MAB problems [63–65], such as ϵ-greedy algorithm, softmax algorithm, and UCB1-tuned algorithm. Although the UCB1-tuned algorithm is known as the best algorithm among parameter-free algorithms, the TOW model [39,66–68] has approximately the same performance as the UCB1-tuned algorithm. The calculation in the TOW model does not require a variance, such as in UCB-1 tuned algorithm, therefore it is suitable for mobile devices such as IoT [69]. The TOW model can adapt the variable environment where the reward probability changes dynamically. The authors of [70] discusses an issue of adapting uncertain and dynamic environment by using reinforcement learning. It also discusses the complexity of reinforcement learning algorithms. Unlike the Q-learning algorithm or others, the TOW model does not require state information, which reduces complexity. Indeed, the TOW model is confirmed to work well in a real-time manner with IoT testbed under such an environment [69]. Therefore, the TOW model is fitted to solve the problem of decision-making in cognitive terminals.

The TOW model is a multi-armed bandit algorithm inspired by the behavior of amoeboid organisms. Unlike other algorithms for estimating the reward probability of each slot machine, TOW dynamics use a unique learning method that is equivalent to updating all machine estimates simultaneously based on the volume conservation law. In the TOW model, the decision is made according to the displacements of the imaginary volume-conserving objects, which increase or decrease along with rewards, as shown in Figure 14. The TOW models of the two machines are formulated as below. Imagine that the player plays a slot machine A or B at a time. When playing machine A, if the player receives rewards, then 1 is added to an estimator Q_A; otherwise, ω is decreased (punishment). After playing time step t, the displacement of machine A, X_A $(= -X_B)$, is expressed as follows:

$$\begin{aligned} X_A(t+1) &= Q_A(t) - Q_B(t) + \delta(t), && (5.1) \\ Q_A(t) &= N_A(t) - (1+\omega)L_A(t), && (5.2) \end{aligned}$$

where $\delta(t)$ is a fluctuation, $N_i(t)$ $(i \in \{A, B\})$ is the number of times that machine i has been played until time $t, andL_i(t)$ counts the number of punishments when playing machine i until time t. ω is a weighting parameter described below.



(a) n = 2

(b) n > 2

Figure 14. (a) TOW model with two slot machines. The solid bar maintained its shape constant. (b) TOW model with several slot machines. A branched cylinder is filled with an uncompressed fluid.

Let the probabilities of providing rewards in machines i be P_i $(i \in A, B)$. Considering the ideal situation where the sum of the reward probabilities $\gamma = P_A + P_B$ is known to the player, the expected reward Q'_i $(i \in A, B)$ is given as

$$\begin{aligned} Q'_A &= N_A \frac{N_A - L_A}{N_A} + N_B \left(\gamma - \frac{N_B - L_B}{N_B}\right) \\ &= N_A - L_A + (\gamma - 1) N_B + L_B, \qquad (5.3) \\ Q'_B &= N_A \left(\gamma - \frac{N_A - L_A}{N_A}\right) + N_B \frac{N_B - L_B}{N_B} \\ &= N_B - L_B + (\gamma - 1) N_A + L_A. \qquad (5.4) \end{aligned}$$

If we define $Q''_j = Q'_j/(2-\gamma)$, we can obtain the difference in estimates in an ideal situation as

$$Q''_A - Q''_B = (N_A - N_B) - \frac{2}{2-\gamma}(L_A - L_B)\,. \qquad (5.5)$$

In contrast, the difference between Q_A and Q_B from Eq (5.2) is given by

$$Q_A - Q_B = (N_A - N_B) - (1 + \omega)(L_A - L_B)\,. \qquad (5.6)$$

From the above two equations, we can obtain the nearly optimal weighting parameter ω in terms of γ as

$$\omega = \frac{\gamma}{2-\gamma}. \qquad (5.7)$$

If the number of machines is n $(n > 2)$, $\omega = \gamma/(2-\gamma)$ is given by $\gamma = P_1 + P_2$, where P_1 and P_2 are the first- and second-highest reward probabilities, respectively, [67]. Then, Eq.(5.1) can be expressed as follows:

$$X_i(t) = Q_i(t) - \frac{1}{n-1} \sum_{k=1,\neq i}^{n} Q_k(t) + \zeta_i(t), \qquad (5.8)$$

where ζ is the fluctuation in each slot machine. The player selects the machine which has the highest $X_i(t)$. We use the following $\zeta_i(t)$ in this paper:

$$\zeta_i(t) = A \cos\left(\frac{2\pi t}{n} + \frac{2(i-1)\pi}{n}\right), \qquad (5.9)$$

where A is the amplitude of the fluctuation.

5.3. Application of the MAB algorithm to wireless network selection

For the first example, we apply the MAB problem and TOW algorithm to wireless network selection in a cognitive mobile terminal [62]. Figure 15 shows the concept of reward in the MAB problem in this model. In this example, three networks, Wi-Fi 1, Wi-Fi 2, and LTE, are available at a cognitive mobile terminal. The mobile terminal evaluates each network based on the reward corresponding to the performance of each network, such as throughput.

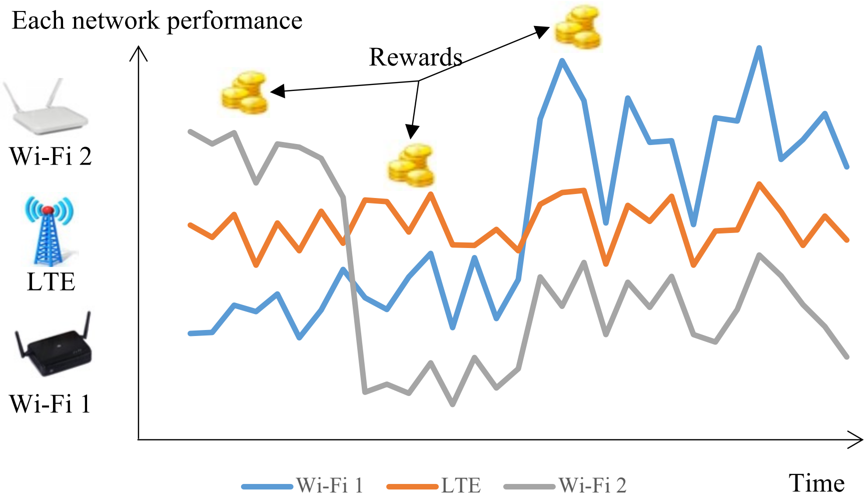


Figure 15. Concept of reward in network selection in cognitive radio as a multi-armed bandit problem.

5.3.1. Background

Recent wireless devices are equipped with multiple wireless interfaces like smartphones. Various wireless networks, such as 3G, LTE, and Wi-Fi, are available. Moreover, similar to access points in Wi-Fi, a mobile terminal can choose from multiple access networks. Ideally, in such a heterogeneous network, a user may choose the best wireless network by gathering information from each network. Several studies have focused on heterogeneous wireless network selection. There are two types of approaches: network-centric and user-centric. In the network-centric approach [71–73], the selection decision is performed using a central controller. It assumes that detailed information on each network status is available at the central controller. However, it is difficult to exchange information among various wireless networks that operate independently. Therefore, in this section, we focus on the selection of wireless networks at mobile terminals.

Several studies have investigated wireless network selection at the mobile terminal [74–77]. Most of them require considerable information of networks or computational capacity. However, it is not always possible for mobile devices to gather information from all networks or to spare battery resources for complex calculations. It is important for the heterogeneous network selection to seek the best solution as much as and as fast as possible using the limited information about the networks. Simultaneously, it is also important for mobile devices to suppress the complexity of calculations for making decisions. These constraints and requirements are similar to those in the MAB problem. In the MAB problem, a player of slot machines attempts to obtain the maximum reward through finite trials. Therefore, the MAB problem approach can help in heterogeneous network selection.

The major challenges in the selection of wireless networks at mobile terminals in heterogeneous environments are as follows.

- Efficient decisions must be made under the situation in which a small amount of information regarding each network is available.
- A practical algorithm that can be implemented on resource-constraint mobile devices is required.

To overcome these issues, several studies investigated algorithms and their performances. In reference [74], a non-cooperative game formulation and analysis were provided for Wi-Fi and cellular network selection on a mobile terminal. Results of computer simulations showed that the game can converge to Nash equilibria. However, the assumption that the mobile terminal can obtain information from other mobile users is not always possible. In reference [77], a reinforcement learning solution and simulation analysis were provided for heterogeneous cellular networks. Although the simulation results showed the convergence speed and the suppression of overheads, it requires feedback information from the networks, which is only feasible in cellular networks. It is important for the mobile terminal in a heterogeneous network to select a better network without coordination from the networks. At the same time, it is important for mobile devices to suppress the complexity of calculations for making decisions. These constraints and requirements are closely similar to those in the MAB problem.

5.3.2. Heterogeneous network selection as an MAB problem

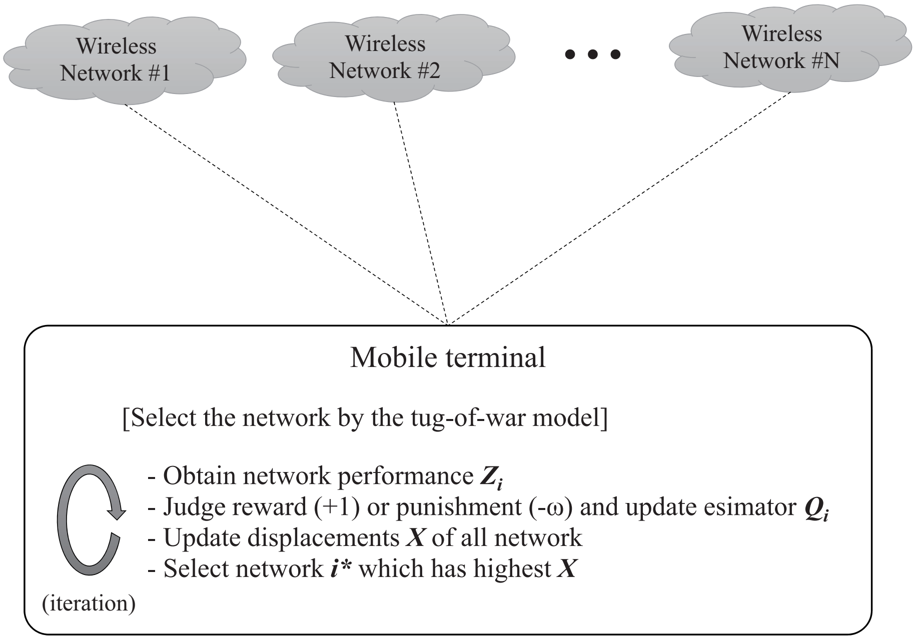

Wireless Network #1
Wireless Network #2
• • •
Wireless Network #N
Mobile terminal
[Select the network by the tug-of-war model]
- Obtain network performance Z_i
- Judge reward (+1) or punishment (-ω) and update esimator Q_i
- Update displacements X of all network
- Select network $i*$ which has highest X
(iteration)

Figure 16. System model of the proposed wireless system which selects the wireless network based on the MAB algorithm.

We propose a wireless network selection technique based on the MAB problem. As described in the previous section, we used the TOW algorithm to solve the MAB problem. Figure 16 shows the system model of the proposed wireless network selection. The mobile terminal, capable of connecting various types of wireless networks S_i ($i \in \{1, 2, ...N\}$), functions as a TOW algorithm. It observes the performance of network Z_i, where i is the selected network. Z_i can be any index of performance, such as throughput, delay, or other metrics of wireless networks. The TOW algorithm then judges whether the reward or punishment is to be added by evaluating the performance of network i. If the reward is given, it updates the estimator as $Q_i + 1$; otherwise, it updates the estimator as $Q_i - \omega$. Then, X, the displacement of each network, is updated as shown in Eq (5.8). Note that all X_j ($j \in \{1, 2, ..., i, ..., N\}$, and not only the selected network i, are updated here. The algorithm in the mobile terminal is as follows:

1) Start observing the performance of each wireless network i. All the networks are monitored at least once.
2) Update Q_i and all X based on the observation of network i by Eqs (5.2) and (5.8).
3) Select a wireless network i^* with highest X_i.
4) Observe the performance of the selected network and decide whether a reward or punishment is to be given.
5) Back to 2 and continue.

5.3.3. Implementation of the proposed scheme

To validate the proposed method in a heterogeneous wireless network environment, we implemented the proposed algorithm on a wireless device and performed experiments. The TOW algorithm was installed on Ubuntu Linux on a laptop PC as a cognitive mobile terminal, equipped with both 802.11n/ac (2.4 GHz and 5 GHz) and LTE communication module. To judge the reward or punishment (1 or $-\omega$) in Eq (5.2), we used the average throughput of the wireless networks as a threshold. If the observed throughput of wireless network i is larger than the average, then the reward is given for Q_i; otherwise, a punishment is given. We use the first- and second-highest reward probabilities among the networks as γ in Eq (5.7). The experiments were conducted in and around the university building. The mobile terminal selects the wireless network from two Wi-Fi networks (2.4 GHz and 5 GHz) of the university infrastructure and LTE to communicate with the server.

5.3.4. Experimental setup

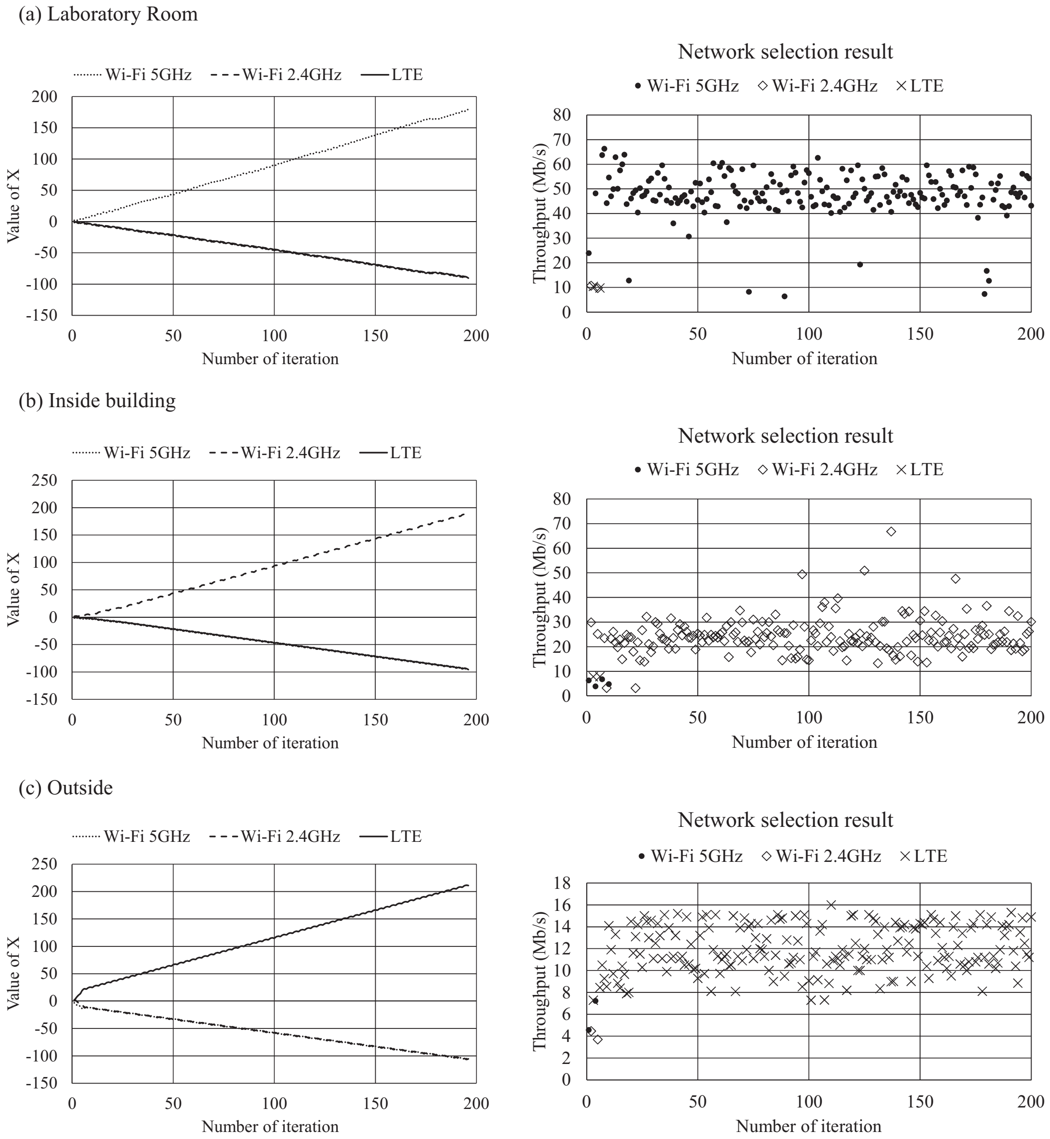
We used the iperf command to observe the throughput. The parameter A of the fluctuation in Eq (5.9) is set to 5. Each iteration cycle required about 2 s. The locations of the experiments are (a) the laboratory room, (b) inside the building, and (c) outside. The average received signal strength indicator (RSSI) of Wi-Fi is listed in Table 4.

Table 4. Average RSSI of Wi-Fi

Location	Wi-Fi 2.4 GHz	Wi-Fi 5 GHz
(a) Laboratory room	−36.0	−37.0
(b) Inside the building	−39.0	−76.0
(c) Outside	−73.0	−81.0

5.3.5. Verification of the proposed scheme

Figure 17 shows an example of network selections of the proposed algorithm. The values of X for the TOW algorithm in Eq (5.8) are also shown. In each case, after the initial trial of all wireless networks, the selection of the wireless network converges to the highest performance network. Note that the value X of the unselected networks is also updated (decreased) through iterations. Although the estimators Q of the unselected networks in Eq (5.2) are not updated, the displacements X of all networks are updated in Eq (5.8). This is a unique characteristic of the TOW model, as described in the previous section. In cases (a) and (b), the values X of Wi-Fi 5 GHz and 2.4 GHz, which have higher RSSI and throughput in the laboratory room and inside the building, become larger according to the number of iterations, whereas the value of X for unselected Wi-Fi and LTE becomes smaller. As a result, the selection converges to Wi-Fi at 5 GHz and 2.4 GHz. In case (c), where the signal strength from the Wi-Fi access point becomes much lower, the value X of LTE increases according to the number of iterations, whereas X of unselected Wi-Fi and LTE becomes smaller. As a result, the selection of the networks is converges to LTE.


(a) Laboratory Room

(b) Inside building

(c) Outside

Figure 17. Examples of value X in the TOW algorithm (left) and the results of network selection (right).

5.3.6. Evaluation of the throughput performance of the proposed scheme

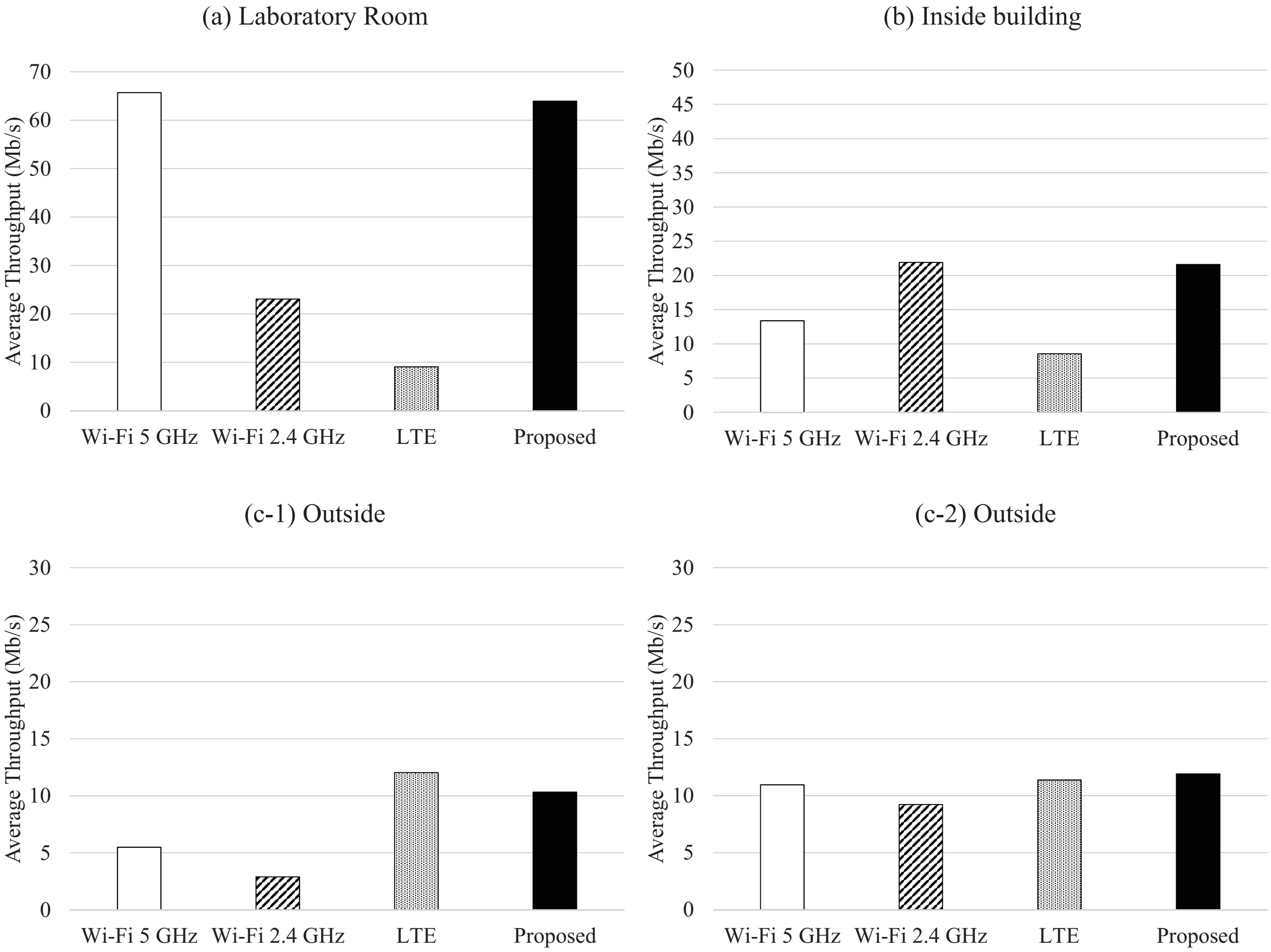


Figure 18. Average throughputs of each wireless network and the proposed TOW algorithm in different environments.

To verify the performance in heterogeneous environments, we examined the throughput at each location. Figure 18 shows the average throughputs of each wireless network and the proposed TOW algorithm. Each experiment was repeated three times, and the throughputs shown here are the average values. The locations (c-1) and (c-2) are both outside the building but different in the Wi-Fi traffic situation: (c-1) is more crowded. It is shown that the proposed TOW algorithm achieves a high average throughput among other wireless networks. In the laboratory room (a) and inside the building (b), the throughput of the proposed system is as high as that of Wi-Fi at 5 GHz and 2.4 GHz, respectively. In contrast, outside the building at (c-1) and (c-2), where the signal strength from the Wi-Fi access point becomes much lower, the throughput of the proposed system is as high as that of LTE. Moreover, as shown in case (c-2), the proposed algorithm can achieve an average performance which is as high as that of LTE, where the differences in performance among all networks are rather small. The results show that the proposed algorithm can accurately estimate the probabilities of rewards among wireless networks.

5.4. Application of the MAB algorithm to dynamic channel selection in IoT devices

In this subsection, another example of the application of a simple reinforcement learning-based optimal decision, the dynamic channel selection in IoT devices, is presented. The major challenge in the selection of wireless channels autonomously at the mobile terminal in massive IoT use cases is making efficient decisions in situations where little or no information of each mobile node is available. It is also challenging to find a practical algorithm that can be implemented on resource-constrained mobile devices of the IoT. Lai et al. modeled a cognitive radio as a multi-armed bandit (MAB) problem [36,37]. In their model, the channel selection of a cognitive radio is defined as an MAB problem under the assumption of a probabilistic vacancy of each channel. In a previous work [78], TOW applications for channel selection in wireless LANs were proposed. It gave an efficient dynamic spectrum-sharing for cognitive radio. In this subsection, we show another application of TOW in massive IoT. For the evaluation of this application, computer simulation experiments were conducted.

5.4.1. System model

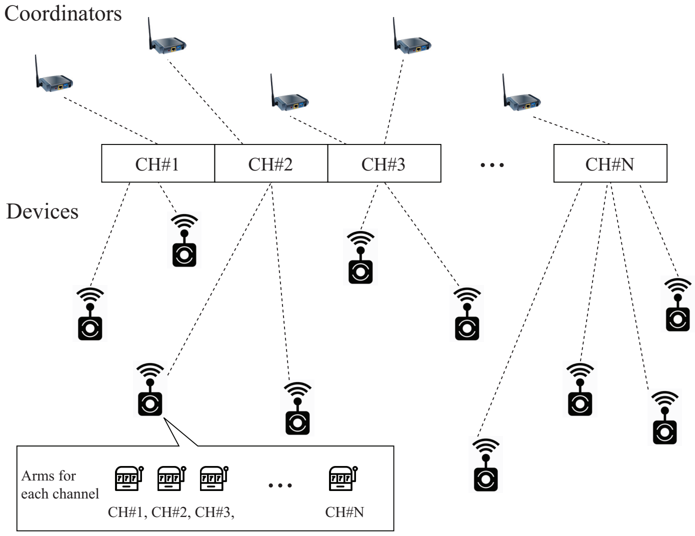


Figure 19. The proposed wireless channel selection based on the MAB algorithm in massive IoT.

Figure 19 shows the proposed wireless channel selection in the massive IoT scenario. The devices use one of the wireless channels from $CH\#i$ ($i \in \{1, 2, ..., N\}$), which is decided by the TOW algorithm in each device. The acknowledgement frame (ACK) received upon successful communication is the reward of the TOW algorithm.

5.4.2. Simulation setup

To verify the proposed simple reinforcement algorithm in a massive IoT scenario, network simulation using ns-3 [79] was conducted. For the simulation, the LR-WPAN model was used. The settings of the simulation are shown in Table 5. Two simulation scenarios, A and B, were considered, as described in Table 6. Note that the association between the device and coordinator is simplified in the simulation (message passings are omitted).

Table 5. Simulation settings.

Parameter	Value
Area size	100 m x 100 m
Simulation duration	90 s
Simulation iterations	10 times
Number of repetition	10 times with different random seeds
Pathloss model	Free space decay
Propagation channel model	Additive white gaussian noise (AWGN)
Wireless standard	IEEE 802.15.4
Radio frequency	2.4 GHz
Modulation scheme	O-QPSK
Transmission power	1 mW
Radio communication range	around 100 m
Mac protocol	CSMA/CA
Number of channels	4
Traffic	UDP of 100 Byte/0.2 s
Number of coordinators	10
Number of devices	40
Node placement	normalized random distribution
Mobility	All nodes are stationary

Table 6. Two scenarios of channel selection by simple reinforcement-learning–based optimal decision in massive IoT.

Case	Channel on coordinator	Channel on device
A	Fixed manually	Selected by MAB algorithms
B	Selected by MAB algorithms	Selected by MAB algorithms

The performance index of this simulation is set as the frame success rate (FSR), which is the ratio number of received packets to the total number of transmitted packets, as shown in Eq (5.10).

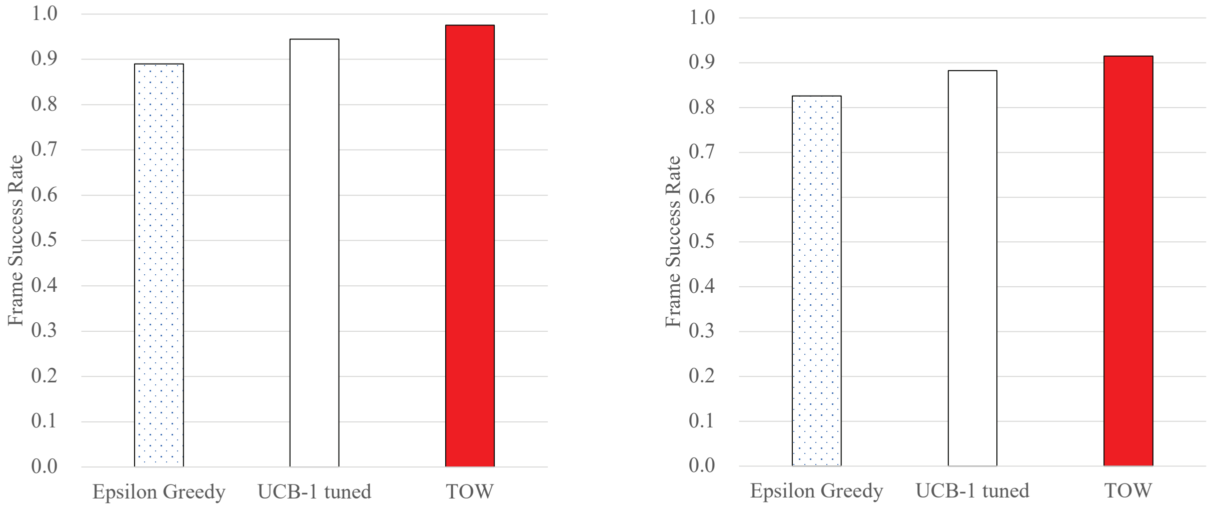
$$FSR = \frac{\sum_{i=1}^{M} \sum_{j=1}^{K} r_j^i}{\sum_{i=1}^{M} \sum_{j=1}^{K} n_j^i}, \tag{5.10}$$

where M is the number of devices, K is the number of available channels, and r_i^j is the number of counts receiving reward (ACK) when node i uses channel j, and n_i^j is the number of counts node i attempts to send using channel j.

5.4.3. Verification of the proposed scheme

Figure 20(a) shows the FSR of nodes in the scenario A, where coordinators are operated in fixed channels. The FSR when devices select their operational channels by TOW algorithm outperforms those by ϵ-Greedy algorithm or by UCB-1 tuned algorithm. It indicates that TOW algorithm can efficiently select operational channels among other MAB algorithms.

Scenario B, where not only devices but also coordinators select their operational channels autonomously, is more difficult to select optimal channels. This scenario is an example of a highly distributed decision network. The result of this scenario is shown in Figure 20(b). When devices select their operational channels by the TOW algorithm, FSR outperforms ϵ-Greedy algorithm or by UCB-1 tuned algorithm. It indicates that the TOW algorithm can select operational channels efficiently among other MAB algorithms, even in the more distributed situation.



(a) Frame success rate of each algorithm in scenario A

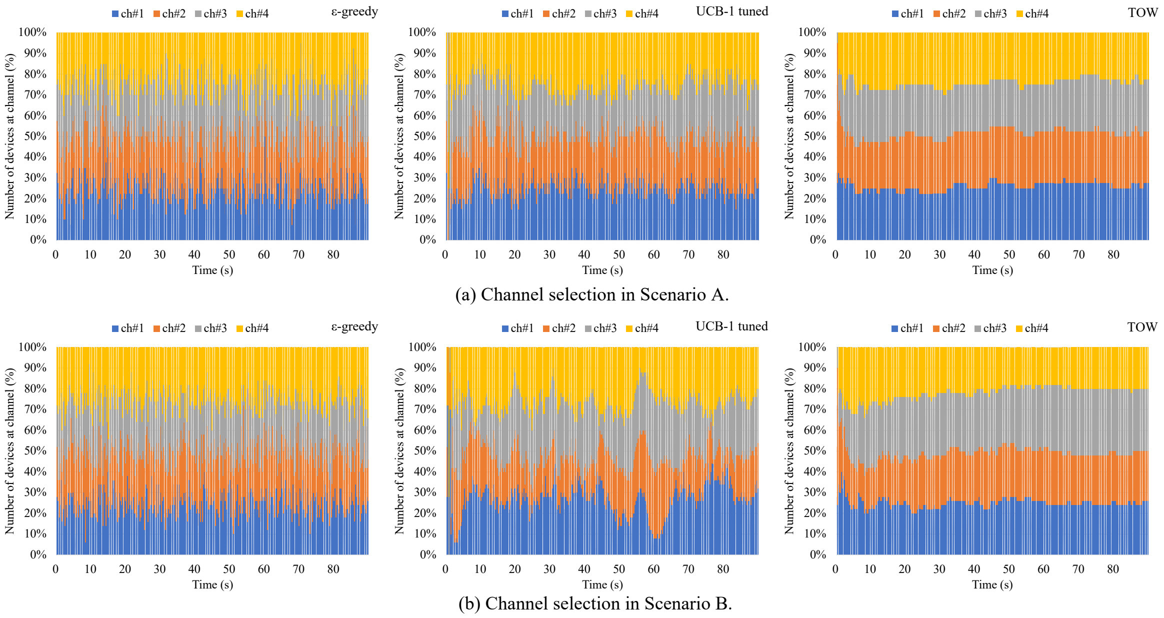
(b) Frame success rate of each algorithm in scenario B

Figure 20. Average frame success rate of wireless nodes.

Figure 21 shows examples of channel selection results in each algorithm in scenario A (top figures), where the channels of coordinators are fixed, and in scenario B (bottom figures), where the channels of coordinators are also selected by the same algorithm as the devices. In both scenarios, the channel selections of the TOW algorithm show convergence, while those of UCB-1 and ϵ-Greedy show fluctuations. The selected channels show convergence in the TOW algorithm, while those of UCB-1 also show convergence but slightly fluctuate over time. Compared to other algorithms, the selected channels of ϵ-Greedy are unstable and fluctuate, because of the randomness of ”greedy” exploration of the algorithm. The results show that the proposed algorithm can efficiently select channels in a distributed manner.

6. Conclusions

Advancements in wireless communication technologies have led to enormous positive changes globally. Along with this, wireless systems have become increasingly complex, not only in a single communication node, but also as a communication system. From the viewpoint of exploiting its capability, two simple questions rise. One question is how to build models for complex wireless systems of the present and the future. Another is how to decide optimal action using models of wireless communication systems.



(a) Channel selection in Scenario A.

(b) Channel selection in Scenario B.

Figure 21. Channel selection results with ϵ-Greedy, UCB-1 tuned, and TOW in scenarios A (a) and B (b).

A classical mathematical formulation-based optimization scheme cannot be applied to today's complex wireless communication systems because the complexity of the systems prevents building mathematical models. It opens the window for applications of machine learning technologies to optimize the performance of wireless communication systems. Data-driven modeling, by machine learning, is an aid for these issues. Deep reinforcement learning, which combines deep learning-based modeling with reinforcement learning-based decision making for action, is a state-of-the-art scheme in recent research fields. Various studies have applied it in the field of wireless communication. However, there still exist future works to be focused on. One point is to seek alternatives for modeling to realize continuous function-based modeling to obtain better solutions for continuous systems. Another point is to seek a feasible yet effective scheme if the amount of available information is rather small, similar to IoT systems. In this study, to provide solutions for these points, two novel schemes that are different not only from classical mathematical optimization but also from the current state-of-the-art deep reinforcement-learning approach are proposed.

One of the proposed schemes is supervised learning based modeling and optimization. It uses a certain amount of information to build a model of the wireless communication system and obtain optimal parameters using an optimization algorithm. This is based on a cognitive cycle using machine learning. It uses a supervised machine learning algorithm to build the performance model of the systems, obtains the optimal parameters by solving the optimization problem, takes action according to the decision, and updates the performance model in an online manner. Through both real-world experiments and computer simulations of the application of IEEE 802.11 WLAN, the validity of the proposed scheme was confirmed.

Another proposed scheme is simple yet easily implementable reinforcement learning by the MAB problem formulation. By using a novel, lightweight, and distributed TOW algorithm, adaptive learning wireless communication systems with limited software and/or hardware capabilities, such as IoT, can be realised. Two applications are shown: heterogeneous network selection and channel selection in massive IoT. Both real-world experiments and computer simulations demonstrated the validity of the proposed scheme.

These results show the effectiveness and feasibility of the proposed schemes. Various applications based on the proposed schemes are currently being developed [69, 80–82], proving that this research have opened a new field of application of online machine learning technologies to optimize the complex wireless communication systems of the present and the future.

The major achievements of this study are as follows:

1) We have provided a comprehensive framework for optimizing complex wireless communication systems from the viewpoint of application of machine learning technologies.
2) We proposed a scheme using supervised learning and optimization that is a better alternative to deep reinforcement learning especially when parameters are continuous.
3) We have proposed a scheme using simple reinforcement learning based on TOW, a lightweight MAB algorithm, which provides a feasible solution to increase performances of wireless communication systems when the amount of available information is small, similar to IoT.

There are some future works, which is to confirm the effectiveness of this research for more complex real-world networks and applications. The proposed schemes are based on the fundamental framework to optimize complex wireless communication systems. Therefore, they can be applied to various applications including 5G/6G. Mobile edge computing (MEC), or multi-access edge computing, is one of the hot topics in the 5G/6G networks. The proposed schemes can be applied to optimize communication and computing in MEC. Network orchestration is also an important topic in heterogeneous 5G/6G networks. End-to-end optimization of communications in heterogeneous networks, ether in a distributed manner or centralized manner, is to be examined with the proposed schemes. Non-terrestrial networks add vertical communication in 5G/6G networks. Further applications such as a relay network in space and multiple interfaces for radio and optical communication can be optimized through the proposed schemes. Through those verifications, a more concrete figure of a future communication network using machine learning will be revealed.

Copyright notice

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444. doi: 10.1038/nature14539.
2. Google AI Blog, *AlphaGo: Mastering the ancient game of Go with Machine Learning*, 2016. Available from: https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.
3. The 3rd Generation Partnership Project (3GPP). Available from: https://www.3gpp.org/specifications/specifications.
4. H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, K. Wu, Artificial-intelligence-enabled intelligent 6G networks, *IEEE Network*, **34** (2020), 272–280. doi: 10.1109/MNET.011.2000195.
5. A. Goldsmith, S. Chua, Adaptive coded modulation for fading channels, *IEEE Trans. Commun.*, **46** (1998), 595–602. doi: 10.1109/26.668727.
6. D. Gesbert, S. Kiani, A. Gjendemsjo, G. Oien, Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks, *Proc. IEEE*, **95** (2007), 2393–2409. doi: 10.1109/JPROC.2007.907125.
7. J. Lu, T. T. Tjhung, F. Adachi, C. L. Huang, BER performance of OFDM-MDPSK system in frequency-selective Rician fading with diversity reception, *IEEE Trans. Veh. Technol.*, **49** (2000), 1216–1225. doi: 10.1109/25.875231.
8. S. T. Chung, A. J. Goldsmith, Degrees of freedom in adaptive modulation: a unified view, *IEEE Trans. Commun.*, **49** (2001), 1561–1571. doi: 10.1109/VETECS.2001.944588.
9. D. Qiao, S. Choi, K. G. Shin, Goodput analysis and link adaptation for IEEE 802.11a wireless LANs, *IEEE Trans. Mobile Comput.*, **1** (2002), 278–291. doi: 10.1109/TMC.2002.1175541
10. E. Peh, Y. Liang, Y. Guan, Y. Zeng, Optimization of cooperative sensing in cognitive radio networks: A sensing-throughput tradeoff view, *IEEE Trans. Veh. Technol.*, **58** (2009), 5294–5299. doi: 10.1109/TVT.2009.2028030.
11. G. Bianchi, Performance analysis of the IEEE 802.11 distributed coordination function, *IEEE J. Sel. Areas Commun.*, **18** (2000), 535–547. doi: 10.1109/49.840210.
12. J. Mitola, G. Q. Maguire, Cognitive radio: making software radios more personal, *IEEE Pers. Commun.*, **6** (1999), 13–18. doi: 10.1109/98.788210.
13. S. Haykin, Cognitive radio: Brain-empowered wireless communications, *IEEE J. Sel. Areas Commun.*, **23** (2005), 201–220. doi: 10.1109/JSAC.2004.839380.
14. M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, F. Kojima, Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks, *IEEE Access*, **6** (2018), 32328–32338. doi: 10.1109/ACCESS.2018.2837692.
15. M. Chen, U. Challita, W. Saad, C. Yin, M. Debbah, Artificial neural networks-based machine learning for wireless networks: A tutorial, *IEEE Commun. Surv. Tutor.*, **21** (2019), 3039–3071. doi: 10.1109/COMST.2019.2926625.
16. J. Wang, C. Jiang, H. Zhang, Y. Ren, K. -C. Chen, L. Hanzo, Thirty years of machine learning: The road to Pareto-optimal wireless networks, *IEEE Commun. Surv. Tutor.*, **22** (2020), 1472–1514. doi: 10.1109/COMST.2020.2965856.

17. M. Kulin, T. Kazaz, I. Moerman, E. De Poorter, End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications, *IEEE Access*, **6** (2018), 18484–18501. doi: 10.1109/ACCESS.2018.2818794.

18. C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, L. Hanzo, Machine learning paradigms for next-generation wireless networks, *IEEE Wireless Commun.*, **24** (2017), 98–105. doi: 10.1109/MWC.2016.1500356WC.

19. R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, *et al.*, Intelligent 5G: When cellular networks meet artificial intelligence, *IEEE Wireless Commun.*, **24** (2017), 175–183. doi: 10.1109/MWC.2017.1600304WC.

20. T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

21. Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, et al., State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems, *IEEE Commun. Surv. Tutor.*, **19** (2017), 2432–2455. doi: 10.1109/COMST.2017.2707140.

22. H. Ye, G. Y. Li, B. Juang, Power of deep learning for channel estimation and signal detection in OFDM systems, *IEEE Wireless Commun. Lett.*, **7** (2018), 114–117. doi: 10.1109/LWC.2017.2757490.

23. H. Huang, J. Yang, H. Huang, Y. Song, G. Gui, Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system, *IEEE Trans. Veh. Technol.*, **67** (2018), 8549–8560. doi: 10.1109/TVT.2018.2851783.

24. R. W. Ouyang, A. K. Wong, C. Lea, M. Chiang, Indoor location estimation with reduced calibration exploiting unlabeled data via hybrid generative/discriminative learning, *IEEE Trans. Mob. Comput.*, **11** (2012), 1613–1626. doi: 10.1109/TMC.2011.193.

25. Y. Chen, Q. Yang, J. Yin, X. Chai, Power-efficient access-point selection for indoor location estimation, *IEEE Trans. Knowl. Data. Eng.*, **18** (2006), 877–888. doi: 10.1109/TKDE.2006.112.

26. S. Marano, W. M. Gifford, H. Wymeersch, M. Z. Win, NLOS identification and mitigation for localization based on UWB experimental data, *IEEE J. Sel. Areas Commun.*, **28** (2010), 1026–1035. doi: 10.1109/JSAC.2010.100907.

27. B. Mager, P. Lundrigan, N. Patwari, Fingerprint-based device-free localization performance in changing environments, *IEEE J. Sel. Areas Commun.*, **33** (2015), 2429–2438. doi: 10.1109/JSAC.2015.2430515.

28. L. U. Khan, I. Yaqoob, M. Imran, Z. Han, C. S. Hong, 6G wireless systems: A vision, architectural elements, and future directions, *IEEE Access*, **8** (2020), 147029–147044. doi: 10.1109/ACCESS.2020.3015289.

29. Y. He, C. Liang, F. R. Yu, Z. Han, Trust-based social networks with computing, caching and communications: A deep reinforcement learning approach, *IEEE Trans. Network Sci. Eng.*, **7** (2020), 66–79. doi: 10.1109/TNSE.2018.2865183.

30. L. Li, Y. Xu, J. Yin, W. Liang, X. Li, W. Chen, Z. Han, Deep reinforcement learning approaches for content caching in cache-enabled D2D networks, *IEEE Internet Things J.*, **7** (2020), 544–557. doi: 10.1109/JIOT.2019.2951509.

31. F. B. Mismar, B. L. Evans, A. Alkhateeb, Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination, *IEEE Trans. Commun.*, **68** (2020), 1581–1592. doi: 10.1109/TCOMM.2019.2961332.

32. X. Fu, F. R. Yu, J. Wang, Q. Qi, J. Liao, Dynamic service function chain embedding for NFV-enabled IoT: A deep reinforcement learning approach, *IEEE Trans. Wireless Commun.*, **19** (2020), 507–519. doi: 10.1109/TWC.2019.2946797.

33. H. Huang, Y. Yang, H. Wang, Z. Ding, H. Sari, F. Adachi, Deep reinforcement learning for UAV navigation through massive MIMO technique, *IEEE trans. Veh. Technol.*, **69** (2020), 1117–1121. doi: 10.1109/TVT.2019.2952549.

34. Z. Zhang, H. Chen, M. Hua, C. Li, Y. Huang, L. Yang, Double coded caching in ultra dense networks: Caching and multicast scheduling via deep reinforcement learning, *IEEE Trans. Commun.*, **68** (2020), 1071–1086. doi: 10.1109/TCOMM.2019.2955490.

35. H. Robbins, Some aspects of the sequential design of experiments, *Bull. Amer. Math. Soc.*, **58** (1952), 527–535.

36. L. Lai, H. E. Jiang, H. V. Poor, Medium access in cognitive radio networks: A competitive multi-armed bandit frame work, in *Processing of IEEE 42th Asilomar Conference on Signals, System, and Computer* (2008), 98–102. doi: 10.1109/ACSSC.2008.5074370.

37. L. Lai, H. E. Jiang, H. V. Poor, Cognitive medium access: Exploration, exploitation, and competition, *IEEE Trans. Mobile Comput.*, **10** (2011), 239–253. doi: 10.1109/TMC.2010.65.

38. S. Maghsudi, E. Hossain, Multi-armed bandits with application to 5G small cells, *IEEE Wireless Commun.*, **23** (2016), 64–73. doi: 10.1109/MWC.2016.7498076.

39. S. J. Kim, M. Aono, M. Hara, Tug-of-war model for the two-bandit problem: Nonlocally-correlated parallel exploration via resource conservation, *Biosystems*, **101** (2010), 29–36. doi: 10.1016/j.biosystems.2010.04.002.

40. Z. Zhao, E. Schiller, E. Kalogeiton, T. Braun, B. Stiller, M. T. Garip, et al., Autonomic communications in software-driven networks, *IEEE J. Sel. Areas Commun.*, **35** (2017), 2431–2445. doi: 10.1109/JSAC.2017.2760354.

41. Radio Spectrum Policy Group, *Report on Collective Use of Spectrum and Other Sharing Approaches*, 2011. Available from: https://rspg-spectrum.eu/wp-content/uploads/2013/05/rspg11_392_report_CUS_other_approaches_final.pdf.

42. CEPT Electronic Communications Committee, *ECC Report 205–Licensed Shared Access (LSA)*, 2014. Available from: http://spectrum.welter.fr/international/cept/ecc-reports/ecc-report-205-LSA-2300-MHz-2400-MHz.pdf.

43. Reconfigurable Radio Systems (RRS), *Information Elements and Protocols for the Interface between the LSA Controller (LC) and LSA Repository (LR) for the Operation of Licensed Shared Access (LSA) in the 2300 MHz–2400 MHz Band*, 2017. Available from: https://www.etsi.org/deliver/etsi_ts/103300_103399/103379/01.01.01_60/ts_103379v010101p.pdf

44. S. Haykin, P. Setoodeh, S. Feng, D. Findlay, Cognitive dynamic system as the brain of complex networks, *IEEE J. Sel. Areas Commun.*, **34** (2016), 2791–2800. doi: 10.1109/JSAC.2016.2605240.

45. M. Hasegawa, H. Hirai, K. Nagano, H. Harada, K. Aihara, Optimization for centralized and decentralized cognitive radio networks, *Proc. IEEE*, **102** (2014), 574–584. doi: 10.1109/JPROC.2014.2306255.

46. Y. Kon, K. Hashiguchi, M. Ito, M. Hasegawa, K. Ishizu, H. Murakami, et al., Autonomous throughput improvement scheme using machine learning algorithms for heterogeneous wireless networks aggregation, *IEICE Trans. Commun.*, **95** (2012), 1143–1151. doi: 10.1587/transcom.E95.B.1143.

47. IEEE, *IEEE Standard for Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks*, 2009. Available from: https://ieeexplore.ieee.org/document/4798288.

48. R. Combes, A. Proutiere, Dynamic rate and channel selection in cognitive radio systems, *IEEE J. Sel. Areas Commun.*, **33** (2015), 910–921. doi: 10.1109/JSAC.2014.2361084.

49. D. Shiung, Y. Yang, Rate enhancement for cognitive radios using the relationship between transmission rate and signal-to-interference ratio statistics, *IET Commun.*, **7** (2013). doi: 10.1049/iet-com.2013.0047.

50. J. Lehtomaki, M. Benitez, K. Umebayashi, M. Juntti, Improved channel occupancy rate estimation, *IEEE Trans. Commun.*, **63** (2015), 643–654. doi: 10.1109/TCOMM.2015.2402195.

51. B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the fifth annual workshop on Computational learning theory*, (1992), 144–152. doi: 10.1145/130385.130401.

52. V. Vapnik, S. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, *Adv. Neural Inf. Process. Syst.*, *1997* (1997), 281–287.

53. K. Oshima, T. Kobayashi, Y. Taenaka, K. Kuroda, M. Hasegawa, Wireless network optimization method based on cognitive cycle using machine learning, *IEICE Commun. Express*, **7** (2018), 278–283. doi: 10.1587/comex.2018XBL0061.

54. K. Oshima, T. Kobayashi, Y. Taenaka, K. Kuroda, M. Hasegawa, Autonomous wireless system optimization method based on cross-layer modeling using machine learning, in *IEEE International Conference on Ubiquitous and Future Networks*, (2019), 239–244. doi: 10.1109/ICUFN.2019.8806031.

55. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of ICNN'95-International Conference on Neural Networks*, **4** (1995), 1942–1948. doi: 10.1109/ICNN.1995.488968.

56. V. Kadirkamanathan, K. Selvarajah, P. J. Fleming, Stability analysis of the particle dynamics in particle swarm optimizer, *IEEE Trans. Evol. Comput.*, **10** (2006), 245–255. doi: 10.1109/TEVC.2005.857077.

57. Scalable Network Technologies, Inc., Available from: https://www.scalable-networks.com.

58. Beyond 5G/6G White Paper, *National Institute of Information and Communications Technology*, 2021. Available from: https://beyond5g.nict.go.jp/en/download/index.html.

59. P. V. R. Ferreira, R. Paffenroth, A. M. Wyglinski, T. M. Hackett, S. G. Bilen, R. C. Reinhart, et al., Multiobjective reinforcement learning for cognitive satellite communications using deep neural network ensembles, *IEEE J. Sel. Areas Commun.*, **36** (2018). doi: 10.1109/JSAC.2018.2832820.

60. N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, V. Jacobson, BBR: congestion-based congestion control, *ACM Queue*, **14** (2016), 20–53.

61. H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, L. Xiao, Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications, *IEEE Trans. Wireless Commun.*, **20** (2021), 375–388. doi: 10.1109/TWC.2020.3024860.

62. K. Oshima, T. Onishi, S. J. Kim, J. Ma, M. Hasegawa, Efficient wireless network selection by using multi-armed bandit algorithm for mobile terminals, *Nonlinear Theory Its Appl. IEICE*, **11** (2020), 68–77. doi: 10.1587/nolta.11.68.

63. R. Sutton, A. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, 1998 .

64. N. D. Daw, J. P. O'doherty, P. Dayan, B. Seymour, R. J. Dolan, Cortical substrates for exploratory decisions in humans, *Nature*, **441** (2006), 876–879. doi: 10.1038/nature04766.

65. P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, *Mach. Learn.*, **47** (2002), 235–256. doi: 10.1023/A:1013689704352.

66. S. J. Kim, M. Aono, Amoeba-inspired algorithm for cognitive medium access, *NOLTA*, **5** (2014), 198–209. doi: 10.1587/nolta.5.198.

67. S. J. Kim, M. Aono, E. Nameda, Efficient decision-making by volume-conserving physical object, *New J. Phys.*, **17** (2015). doi: 10.1088/1367-2630/17/8/083023.

68. S. J. Kim, M. Aono, E. Nameda, Decision maker based on atomic switches, *AIMS Mater. Sci.*, **3** (2016), 245–259. doi: 10.3934/matersci.2016.1.245.

69. J. Ma, S. Hasegawa, S. J. Kim, M. Hasegawa, A reinforcement-learning-based distributed resource selection algorithm for massive IoT, *Appl. Sci.*, **9** (2019). doi: 10.3390/app9183730.

70. H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor,et al., Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach, *IEEE Trans. Wireless Commun.*, **20** (2021), 1963–1974. doi: 10.1109/TWC.2020.3037767.

71. S. Singh, J. G. Andrews, Joint resource partitioning and offloading in heterogeneous cellular networks, *IEEE Trans. Wireless Commun.*, **13** (2014), 888–901. doi: 10.1109/TWC.2013.120713.130548.

72. C. Liu, M. Li, S. V. Hanly, P. Whiting, Joint downlink user association and interference management in two-tier HetNets with dynamic resource partitioning, *IEEE Trans. Veh. Technol.*, **66** (2017), 1365–1378. doi: 10.1109/TVT.2016.2565538.

73. V. Sagar, R. Chandramouli, K. P. Subbalakshmi, Software defined access for HetNets, *IEEE Commun. Mag.*, **54** (2016), 84–89. doi: 10.1109/MCOM.2016.7378430.

74. A. Keshavarz-Haddad, E. Aryafar, M. Wang, M. Chiang, HetNets selection by clients: Convergence efficiency and practicality, *IEEE/ACM Trans. Netw.*, **25** (2017), 406–419. doi: 10.1109/TNET.2016.2587622.

75. E. Aryafar, A. Keshavarz-Haddad, M. Wang, M. Chiang, RAT selection games in HetNets, in *2013 Proceedings IEEE INFOCOM*, (2013), 998–1006. doi: 10.1109/INFCOM.2013.6566889.

76. X. Wang, J. Li, L. Wang, C. Yang, Z. Han, Intelligent user-centric network selection: A model-driven reinforcement learning framework, *IEEE Access*, **7** (2019), 21645–21661. doi: 10.1109/ACCESS.2019.2898205.

77. D. D. Nguyen, H. X. Nguyen, L. B. White, Reinforcement learning with network-assisted feedback for heterogeneous RAT selection, *IEEE Trans. Wireless Commun.*, **19** (2017). doi: 10.1109/TWC.2017.2718526.

78. K. Kuroda, H. Kato, S. J. Kim, M. Naruse, M. Hasegawa, Improving throughput using multi-armed bandit algorithm for wireless LANs, *NOLTA*, **9** (2018), 74–81. doi: 10.1587/nolta.9.74.

79. ns-3, *ns-3: A Discrete-event Network Simulator for Internet Systems*, Available from: https://www.nsnam.org/.

80. S. Takeuchi, M. Hasegawa, K. Kanno, A. Uchida, N. Chauvet, M. Naruse, Dynamic channel selection in wireless communications via a multi-armed bandit algorithm using laser chaos time series, *Sci. Rep.*, **10** (2020). doi: 10.1038/s41598-020-58541-2.

81. H. Kanemasa, A. Li, M. Naruse, N. Chauvet, M. Hasegawa, Dynamic channel bonding using laser chaos decision maker in WLANs, *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, (2021), 078–082. doi: 10.1109/ICAIIC51459.2021.9415227.

82. Z. Duan, N. Okada, A. Li, M. Naruse, N. Chauvet, M. Hasegawa, High-speed optimization of user pairing in NOMA system using laser chaos based MAB algorithm, *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, (2021), 073–077. doi: 10.1109/ICAIIC51459.2021.9415234.

AIMS Press