# Resource Allocation for Satellite Network Access: A Combinatorial Bandit Learning Approach

***Abstract*—With the advent of the satellite Internet constructed based on the broadband low-orbit satellite network, providing better network services has become an increasing concern. Under limited network resources constraints, the traditional allocation mechanism of satellite service provisions, e.g., the quality of service (QoS) provisions to regions, is manual and over-provisioned, bringing in inflexible configuration and resource waste. To address this issue, we investigate an artificial intelligence-enabled QoS provision allocation scheme and algorithm for the satellite network operator (SNO) to better match the user demands of the service quality. In the provision allocation scheme, we design new functions and entities for SNO, enabling it capable of observing the service region features, estimating and predicting the latent revenue with respect to a QoS provision, and adaptively generating strategies for allocating QoS provisions to service regions. Through this process, the provision allocation scheme can provide predictive and proactive services to service regions. Moreover, considering the fact that the relationship between QoS provisions to service regions and latent revenues is unknown to SNO, we formulate the problem as a combination of unknown parameter learning and combinatorial optimization problems. A learning-based online provision allocation algorithm derived from the multi-armed bandit and combinatorial optimization theories has been proposed. The algorithm is theoretically proven to be sub-linear with service time and validated by simulations.**

***Index Terms*—multi-armed bandit, combinatorial optimization, satellite network access, QoS provision, resource allocation.**

## I. Introduction

With the ability of global coverage, the combination of satellite and terrestrial wireless networks has been considered as a promising approach to improve the delivery of communication services [1]. In order to efficiently utilize the limited satellite network resources, and provide satisfying service provision, e.g, the QoS provision like bandwidth and transmit power, it is time for the satellite network operator (SNO) to revisit its path to the future development [2].

However, three limitations exist in most of existing work and industry. First, the existing QoS provision allocation scheme adopts a manual and static manner of establishment of satellite network service. The most straightforward way to adjust the QoS provision is to over-provision the provision manually and statically manage using a pre-determined Service Level Agreement (SLA) [3]. The considerable cost for the maintenance and provision modifications is inevitable.

Second, the existing QoS provision allocation scheme does not consider the diversities and characteristics of service regions but provides the identical QoS guarantee. The relationship between QoS provisions provided by SNO, service quality of the service region, and obtaining the revenue by SNO is not strongly coupled and fully explored. Moreover, the existing service providing adopts all-IP (Internet Protocol) broadband access based on a reactive mechanism, which results in poor utilization efficiency [4]. It is expected that the proactive and predictive delivery of QoS provisions to service regions leads to a better service experience.

Lastly, the relationship between the provided QoS provision and the revenue obtained by SNO will also be affected by other network operators. In specific, with the increasing developments for the future system, the co-existence of multiple SNOs and terrestrial network operators may form a competitive situation striving for users, traffic, and latent revenue. Since the QoS provision information is not shared by operators, the competitive situation makes the relationship between the provided QoS provision and the obtained revenue more uncertain to SNO.

In conclusion, the main concern above is how to allocate QoS provisions to various service areas for the maximal revenue by SNO. The design of frameworks and algorithms for efficient resource sharing and management is still an open issue. One of the most promising solutions up-to-date is being through allocating satellite resources to service regions with artificial intelligence [5]. In specific, SNO can adopt the emerging techniques, e.g., software-defined networking (SDN), network function virtualization (NFV), and network slicing to decouple the hardware and software resources, formulate isolated QoS provisions with various QoS parameters, e.g., the bandwidth, the transmit power of the satellite, the actual throughput, the time service delay, and the packet loss probability, and allocate QoS provisions to service regions [6]. However, the relationship between the QoS provisions allocated to service areas and the latent revenues by SNO is not certain [7] and there has no relevant studies. It is expected that the big analytic and artificial intelligence (AI) which are referred to as context-aware resource allocations to be one of the most promising solutions [8].

In this paper, we consider the artificial intelligence-enabled allocation of QoS provisions to various service regions for the satellite Internet established on the broadband low-orbit satellite network. To provide services with various QoS guarantees, e.g., the spectrum width and the transmit power, SNO can logically divide network resources into various QoS provisions through the SDN/NFV technique and allocate QoS provisions to regions through the illuminated beams (i.e., by transponders). At the meantime, there co-exist multiple heterogeneous wireless networks operated by other operators adopting different technical specifications and providing wireless network access service with a specific QoS guarantee. In such an en-

vironment, each wireless network infrastructure continuously broadcasts the NSI containing QoS provisions information. The user end can receive the broadcast information and select the appropriate network in accordance with the QoS criteria. If the QoS provision provided by SNO outperforms other operators, the satellite network will be chosen by the user with a high probability, and the revenue obtained by SNO will be increased. Three highlights in this work are presented as follows.

- We propose an improved QoS provision allocation scheme on the basis of [5] for SNO to provide predictive and proactive services for various regions and users. Additional functions and entities are designed. In specific, we design observation and re-provision functions for the satellite, where the satellite can observe the contextual information on the service region, e.g., the service time and the user information through the observation function and adjust the onboard parameters, e.g., the transmit power and bandwidth through the re-provision function. Moreover, we design an entity network provision controller (NRC) with three functions: compound contextual information observation, virtual network embedding, and strategy generation, where SNO can obtain the service information from the satellite through the compound contextual information observation function, can generate various QoS provisions through the SDN/NFV through the virtual network embedding, and makes the decision to allocate QoS provisions to regions through the strategy generation function. In our scheme, SNO can generate QoS provisions and allocate appropriate QoS provisions to various regions.
- To establish the relationship between the QoS provisions allocated to regions and obtained revenues, a probabilistic prediction model derived from [9] is established to predict the revenue of a region with respect to a QoS provision. Two parameters, the maximal latent obtainable revenue (MLOR) and the revenue obtained probability (ROP), are defined. In specific, MLOR represents the maximal revenue that can be obtained by SNO from the region, which is only related to the gross domestic product (GDP) and can be obtained with the scientific methods [9]. It is a constant value assumed to be known *a priori* by SNO. ROP represents the probability that the revenue can be obtained from the region with respect to a specific QoS provision. The underlying meaning of ROP is that when a region is allocated to a QoS provision, the network service is selected by users with a probability. It is an unknown parameter and must be estimated through long-term observations by SNO. Based on the two parameters above, the revenue of a region with respect to a QoS provision can be predicted.
- On the basis of the proposed QoS provision allocation scheme and the probabilistic prediction model, we design an artificial intelligence-enabled algorithm for SNO to implement the resource allocation. Since the relationship between the QoS provisions with respect to areas and obtained revenues is uncertain, we formulate the provision allocation problem as a combination of contextual multi-armed bandit (MAB) and combinatorial optimization problems, where allocating $N$ QoS provisions to $N$ regions is regarded as a provision combination set, and each provision combination set is regarded as an arm in MAB. The aim of SNO is to find the optimal provision combination set that can bring in the largest revenue through long-term online learning. Through the theoretical analysis and simulations, the effectiveness of the proposed algorithm has been verified.

The rest of the work is organized as follows. Section II introduces related works from several parts containing the satellite Internet architecture, the resource management in the satellite network, the MAB, and the combinatorial optimization problem. The model and problem formulation have been depicted in Section III. Section IV introduces the provision allocation scheme, and Section V presents the online provision allocation algorithm along with theoretical analysis. In Section VI and Section VII, simulations and conclusions are presented, respectively.

*Notation:* Variables, vectors, and matrices are written as italic letters $x$, bold italic letters $\boldsymbol{x}$, and bold capital italic letters $\boldsymbol{X}$, respectively. The operators $\mathbb{E}$, $(\cdot)^{\mathrm{T}}$ denote the expectation with respect to all the randomness, the transpose, and the inverse, respectively. Define $\mathcal{I}_N = \{1, 2, \ldots, N\}$ as a shorthand as the index set. $\mathbb{P}[\mathcal{A}]$ denoted the probability of the event $\mathcal{A}$. $\mathbb{I}[\cdot]$ represents an indicator function. Given two functions $h_1(n)$ and $h_2(n)$, $h_1(n) = O(h_2(n))$ means that there exists a constant $c$ and a fixed $n_0$, such that $h_1(n) \leq c \cdot h_2(n)$. For a positive definite matrix $\boldsymbol{D}$, denote $\|\boldsymbol{x}\|_{\boldsymbol{D}}$ as $\sqrt{\boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{D} \cdot \boldsymbol{x}}$.

## II. Related Works

In this section, we introduce the related studies from the following three aspects: the satellite Internet architecture, the resource management in the satellite network, and the MAB and combinatorial optimization problem.

### A. The Satellite Internet Infrastructure

The traditional satellite Internet consists of three segments: the access subsystem, the core subsystem, and the control and management subsystem [3].

The access subsystem is composed of satellite terminals and gateways connected through wireless channels (transponders) of a satellite. The satellite terminals contain handheld devices, portable stations, and various communication terminals. The gateways are responsible for the call handling, switching, and interface to the ground communications network. The core subsystem is an aggregation network connecting different gateways located at the same or different teleport and the network nodes located in some Points of Presence (PoPs) with the external networks. The control and management subsystem comprises the Network Control Center (NCC) and Network Management Center (NMC). The NCC is responsible for the control of the satellite network, such as maintaining,

monitoring, and managing the orbital position and altitude of the satellite, while the NMC takes control of the management of system elements of the satellite network, such as handling users registration, identity verification, billing, and management functions.

With the advent of SDN and NFV, there have been several seminal studies about improvements in the satellite network architecture. In [5], a flexible and re-configurable broadband satellite network architecture embedded with SDN/NFV has been proposed. Different from the traditional architecture, a logically centralized network reprovision controller (NRC) has been deployed, which substitutes part of traditional control and management logic inside gateways and satellites, and extends the function of allocating QoS provisions for QoS guarantee. In specific, the satellite is embedded with the function of collecting local resource status, e.g., the spectrum resource availability and capacity, and reporting to the regional resource status collection function inside an associated gateway. By collecting such information, the NCC/NRC can perform the mission planning, virtual network embedding, strategy generation, strategy evaluation, and global resource status collection. In such an architecture, the resource management and provision process can be flexible and intelligent.

### B. The Resource Management in Satellite Network

The resource management in satellite networks can be discussed from four sections: time, frequency, power, and space [10]. Many of the existing works are made to improve the service quality through QoS guarantee [11]. To improve the low resource utilization efficiency brought out by distributed spatial and temporal of differential service demands, many dynamic resource allocation algorithms for bandwidth, power, and beam hopping are proposed in [12]–[14]. In [12], authors investigate how to flexibly adjust satellite network resources to satisfy different conditions with respect to the latent uncertainty of differentiated service requirements and the non-uniform spatial distribution of capacity requests. To address this issue, a method based on multi-objective deep reinforcement learning dynamically allocating beams to match the system capacity demand has been proposed. To provide the service with the QoS guarantee, authors in [15] investigate a scenario where each beam is embedded with non-orthogonal multiple access (NOMA), and the transmit power is optimized subject to the constraints of limited network resources.

As the SDN/NFV technologies arise, many works investigate the resource management to improve service quality. In [3], [16], authors study how to improve the service quality through the SDN/NFV-based network slicing. Several scenarios and use cases containing dynamically allocating bandwidth are discussed. In [6], authors consider an low-orbit satellite system based on the SDN/NFV technique providing QoS guarantee for the complicated multi-rate traffic environment. Two policies referred to as the fixed channel reservation and the threshold call admission, have been analyzed under an analytic framework.

### C. MAB and Combinatorial Optimization Problems

MAB is an efficient and well-studied tool modeling the sequential-decision making scenario. Compared with other models in reinforcement learning, the MAB can give the anytime optimal solutions with rigorous theoretical proof in both finite and asymptotic time, which is instructive for mission-critical applications [4], [17]. As one of the most critical aspects of reinforcement learning, it has been widely applied in the problem of statistical learning in unknown environments [18].

For the classic MAB model, the user faces up with $K$ arms. At each time, the user can choose one arm and obtain a reward reflecting the stationary reward probability distribution of this chosen arm. The goal of the user is to maximize the reward during a successive finite or infinite time. In this case, the user faces a dilemma: it can decide to select the arm based on the historical observations or select other arms to continue estimating the arm reward probability. The first process is referred to as exploration, and the second process is referred to as exploitation.

The MAB can be efficiently adopted to the unknown environment with uncertain information and limited feedback. The model is established to regard the unknown parameters or the parameter combinations as arms. Taking the state of parameters as i.i.d. process, e.g., the Bernoulli distribution or Markovian process, the user can learn the information and choose the optimal selections [19]–[21].

Apart from the classical MAB, some variants with more practical applications have been investigated. The contextual multi-armed bandit, which takes the QoE-related factors like throughput and frame rate as the contextual information of the wireless environment, has been discussed in [4]. Another emerging research region is the combination of MAB and combinatorial optimization problem, which can be seen as a statistical online learning problem where at each step, the user selects a subset of ground items subject to combinatorial constraints, and then receives stochastic rewards of those items [22], [23]. Since many combinatorial optimization problems have linear objectives, the combination of MAB, and combinatorial optimization can be widely applied in resource allocation, shortest paths, and recommendations [22]. In the further, several works are also studied for the combination of contextual MAB and combinatorial optimization problem [24], where algorithms with $O(\sqrt{T})$ regret have been proposed.

## III. Model Description and Problem Formulation

In this section, we first introduce our model. Two parameters, the maximal latent obtainable revenue and the revenue obtained probability are established to depict the relationship between the QoS provisions allocated to regions and obtained revenues by SNO. Lastly, we formulate the problem.

### A. Model Description

In this work, we consider $N$ fixed regions on the surface of the earth, which are denoted as $\{A_1, A_2, ..., A_N\}$. SNO has $K$ ($K \geq N$) QoS provisions, denoted as $\{C_1, C_2, ..., C_K\}$.

We assume that provisions can be generated by SDN/NFV introduced in [3], [5].

Due to the continuous movement of the satellite in space, there exists a finite and fixed time duration for a satellite serving a region on the earth. In this work, we refer to such a service time duration as the period. At the beginning of each period, SNO selects $N$ from $K$ QoS provisions and allocates them to $N$ illuminated regions. In this way, there exist $\binom{K}{N}N!$ permutations, each of which is termed provision combination set in this work.

When SNO allocates a QoS provision to a region within a period, SNO obtains revenue from users in the service region. For various regions, such as city and countryside, mountain and plain, ocean and land, we assume that there exist various service demands and QoS requirements. Two parameters termed the maximal latent obtainable revenue and the revenue obtained probability derived from [9] are defined. In specific, the maximal latent obtainable revenue $r_{A_j}$ for region $A_j$, $j \in \mathcal{I}_N$, represents the maximal revenue only related to the GDP and GDP per capita [9]. It is reasonable to believe that the maximal latent obtainable revenue obtained of the region with a higher GPD is more significant than that of a region with a lower GDP. This parameter can be obtained through the scientific method introduced in [9], and assumed to be the constant and known *a priori* by SNO in our work. The revenue obtained probability $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, represents the probability that the revenue can be obtained by the region allocated to a specific QoS provision. The underlying reason of the revenue obtained probability is that when a region is allocated to a QoS provision, the network service is selected by users with a probability since there may co-exist other network operators. Another reason is that users only need network services at certain times, enabling SNO to obtain revenues at a period with a certain probability. In this sense, the revenue obtained probability is related to the quality of the service provided by SNO, the service provided by other network operators, and the region information, like the user number and the service requirement [9]. It is assumed to be an unknown parameter and the statistical significance must be observed over an adequate long period of time. Moreover, since the revenue obtained probability represents the probability, there exists an instantaneous revenue obtained state in each period denoted as $\tilde{\phi}_{k,j}(t)$, $j \in \mathcal{I}_N$ and $k \in \mathcal{I}_K$, which is to depict the instantaneous state that the satellite Internet service is selected in period $t$. In this work, we assume that the instantaneous revenue obtained state $\tilde{\phi}_{k,j}(t)$ experiences the unknown distributions with the mean value of $\phi_{k,j}(t)$, where

$$\mathbb{E}[\tilde{\phi}_{k,j}(t)] = \phi_{k,j}(t), \ \ k \in \mathcal{I}_K, j \in \mathcal{I}_N. \tag{1}$$

Both $\tilde{\phi}_{k,j}(t)$ and $\phi_{k,j}(t)$ are assumed constant in one period.

From the defined maximal latent obtainable revenue and the revenue obtained probability, the instantaneous revenue $r_{k,j}(t)$ for region $A_j$, $j \in \mathcal{I}_N$ with respect to QoS provision $C_k$, $k \in \mathcal{I}_K$ in period $t$ can be represented as

$$r_{k,j}(t) = \tilde{\phi}_{k,j}(t) \cdot r_{A_j}. \tag{2}$$

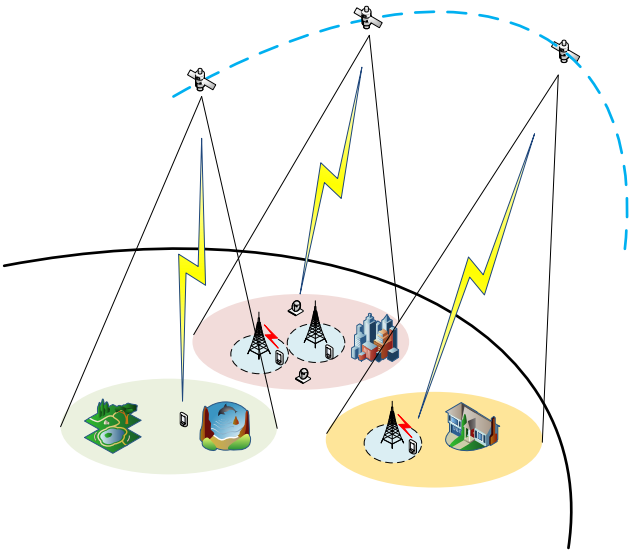


Fig. 1. A depiction of the service region. The region with green color represents the natural environment like lake and ocean, the region with red color represents the city region and the region with yellow color represents the suburb region. Various colors of regions also describe various service demands and QoS requirements.

The expected revenue that SNO can obtain in $T$ periods for $N$ service regions can be

$$R^{\pi}(T) = \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{p=1}^{\binom{K}{N}N!}\sum_{(k,j)\in\mathcal{C}_p}\tilde{\phi}_{k,j}(t)\cdot r_{A_j}\cdot\mathbb{I}[\pi(t)=\mathcal{C}_p]\right\} \tag{3}$$

where $R^{\pi}(T)$ is the expected cumulative revenue obtained by the allocation scheme $\pi$ within $T$ periods. $\mathcal{C}_p$, $p \in \mathcal{I}_{\binom{N}{M}N!}$ represents the provision combination set. Each element of the set represents the allocation relationship of $N$ QoS provisions to $N$ regions. $(k,j)$ is the provision allocation relationship between provision $\mathcal{C}_k$, $k \in \mathcal{I}_K$ and region $A_j$, $j \in \mathcal{I}_N$, i.e., SNO configures provision $\mathcal{C}_k$ to region $A_j$. $\pi(t)$ represents the provision combination set chosen in period $t$.

### *B. Problem Formulation*

The goal of SNO is to design an algorithm $\pi$ selecting $N$ from $K$ QoS provisions and allocating $N$ selected QoS provisions to $N$ service regions to obtain the expected cumulative revenue. To address this issue, the relationship between the QoS provisions and revenue must be known by SNO.

In our model, the relationship between the QoS provisions and revenue is depicted by the revenue obtained probability, which is unknown to SNO. SNO must estimate it through the long-term observation. Since the revenue obtained probability is related to the region information and the QoS provision, we first define the compound contextual information $\boldsymbol{x}_{k,j}(t)$ of region $j \in \mathcal{I}_N$ with respect to QoS provision $k \in \mathcal{I}_K$ as

$$\boldsymbol{x}_{k,j}(t) = [C_k, A_j, t, \mathcal{N}_j(t)]^{\mathrm{T}}. \tag{4}$$

where $\mathcal{N}_j(t)$ represents the specific information of users in the region containing the latent user number and service type in period $t$.

Then, a linear model is adopted to depict the revenue obtained probability of $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ as

$$\phi_{k,j}(t) = \boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \boldsymbol{\theta}_{k,j} \tag{5}$$

where $\boldsymbol{\theta}_{k,j}$ represents the unknown parameters that should be estimated and learned by SNO.

The task of SNO is to estimate $\hat{\boldsymbol{\theta}}_{k,j}(t)$ of $\boldsymbol{\theta}_{k,j}$ and $\hat{\phi}_{k,j}(t)$ of $\phi_{k,j}(t)$ for $j \in \mathcal{I}_N$ and $k \in \mathcal{I}_K$ as

$$\hat{\phi}_{k,j}(t) = \boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \hat{\boldsymbol{\theta}}_{k,j}(t). \tag{6}$$

Based on the estimated information, SNO adopts the provision allocation algorithm to select $N$ from $K$ QoS provisions and allocate $N$ selected QoS provisions to $N$ service regions.

Among all QoS provision allocation algorithms, there exists a so-called genie-aided scheme, which can maximize the expected cumulative revenue with $\phi_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ known *a priori*. Denoting the genie-aided algorithm as $\pi^H$, the goal by SNO is to find the optimal provision allocation algorithm that

$$\pi^* = \arg\min_{\pi} \mathrm{REG}^{\pi}(T) = \arg\min_{\pi} R^{\pi^H}(T) - R^{\pi}(T) \tag{7}$$

where

$$\mathrm{REG}^{\pi}(T) = R^{\pi^H}(T) - R^{\pi}(T) \tag{8}$$

represents the expected cumulative revenue gap between $\pi^H$ and $\pi$, which is also termed regret in this work.

## IV. Provision Allocation Scheme

The existing service-providing mode of SNO is based on a reactive mechanism, which leads to poor resource utilization efficiency shown in [4]. To address this issue, we propose a predictive and proactive service providing scheme where SNO can predict the latent revenue of the service region and in accordance with which allocate the provision. To realize this scheme, we design the additional functions and entities for SNO on the basis of [5].

In specific, we design two functions for the satellite, i.e., the satellite observation and the satellite reprovision. Through the satellite observation function, the satellite can observe the service region and collect the related information about the region, e.g., the service period $t$, region index $A_j$, and information of users $\mathcal{N}_j(t)$, $j \in \mathcal{I}_N$. The satellite reprovision function can adjust its onboard payloads and the QoS provision (e.g., the transmit power and the bandwidth) by receiving the controlling information from the associated gateway and the NCC.

Moreover, we design an entity referred to as network reprovision controller (NRC), which takes control of the resource management for SNO. In this way, the NCC/NRC can adjust the service strategy for SNO through three functions: compound contextual information observation, virtual network embedding, and strategy generation. In specific, the NCC/NRC first collects the compound contextual information re-transmitted by gateways from the satellite. Then, the virtual network embedding function generates multiple virtual sub-networks, and the strategy generation function derives the provision allocation strategy in accordance with the compound contextual information. Finally, the NCC/NRC sends the signal to gateways and the corresponding satellites for the provision allocation.

The overall provision allocation process is described as follows. At the beginning of a period, the satellite observes the region and obtains the contextual information, i.e., $A_j$, $t$, and $\mathcal{N}_j(t)$ for $j \in \mathcal{I}_N$. Then, the satellite transmits the contextual information to the associated gateway, and the gateway re-transmits the information to the NCC/NRC. At that time, the NCC/NRC obtains the compound contextual information $\boldsymbol{x}_{k,j}(t) = [C_k, A_j, t, \mathcal{N}_j(t)]^{\mathrm{T}}$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$, and can estimate the revenue obtained probability $\hat{\phi}_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ in accordance with (6). Based on the estimated revenue obtained probability $\hat{\phi}_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ and the maximal latent obtainable revenue $r_{A_j}$ for $j \in \mathcal{I}_N$, the NCC/NRC can predict the latent revenue of the region, establish the virtual network with QoS guarantee, and allocate provisions to service regions by transmitting signals to the satellite. At the end of period $t$, SNO obtains revenue $r_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$.

Noting that the revenue obtained probability $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ is unknown to SNO, which must learn and estimate the information through long-term observations. The online learning algorithm is designed in the following.

## V. Provision Allocation Algorithm

In this section, we introduce how to allocate QoS provisions to regions to obtain the maximal expected cumulative revenue by SNO. In the first, we introduce the optimal provision allocation algorithm where the revenue obtained probability is known *a priori* by SNO. In this case, allocating $N$ from $K$ provisions to $N$ regions is formulated as a maximal weighted bipartite matching problem. Then, we consider that the revenue obtained probability is unknown to SNO. SNO must learn and estimate the revenue obtained probability information through historical observations. By formulating the problem as a combination of MAB and optimization problems, we propose an online provision allocation algorithm.

### A. The Optimal Allocation Algorithm with Known Information

We first analyze the optimal provision allocation algorithm with the assumption that the revenue obtained probability $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, is deterministic and known *a priori* by SNO. In this case, SNO can predict the revenue of a service region when it allocates a specific QoS provision to the region. The problem is cast on how to select $N$ from $K$ QoS provisions and allocate to $N$ service regions to obtain the maximal expected cumulative revenue.

In fact, this provision allocation problem can be attributed to a maximum weighted bipartite matching problem where provisions and regions can be taken as two node sets. The

**Algorithm 1** Optimal provision Allocation Algorithm

1: **for** period $t$ until $T$ **do**
2: NCC/NRC obtains the compound contextual information $\boldsymbol{x}_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$.
3: Calculates $\phi_{k,j}(t) \cdot r_{A_j}$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$.
4: Solves $\mathcal{O}racle(\{\phi_{k,j}(t) \cdot r_{A_j}\})$, for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$, obtains the optimal provision combination set solution $\mathcal{C}^*(t)$, and in accordance with which allocates provisions to $N$ regions.
5: **end for**

edge connecting any two nodes from two sets is the latent revenue $\phi_{k,j}(t) \cdot r_{A_j}$ for region $A_j$, $j \in \mathcal{I}_N$ with respect to QoS provision $C_k$, $k \in \mathcal{I}_K$, in period $t$. The solution to the problem is to find a combination of $N$ edges that can maximize the revenue in each period. The well-known Hungarian Algorithm is adopted to solve the provision allocation problem in our work.

The algorithm is as follows. SNO first observes the service region and obtains the compound contextual information $\boldsymbol{x}_{k,j}(t)$ and revenue obtained probability $\phi_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ in period $t$. Then, SNO predicts the instantaneous revenue $r_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$ in period $t$. To maximize the revenue, SNO solves a maximum weighted bipartite matching problem and obtains the optimal provision combination set solution $\mathcal{C}^*(t)$ for period $t$. Finally, SNO allocates such provisions to regions. The overall algorithm $\pi^H$ is depicted in Algorithm 1.

*Remark 1*: In line 4 of $\pi^H$, we adopt the Hungarian algorithm to solve the maximum weighted bipartite matching problem. The solver is denoted as $\mathcal{O}racle(\{.\})$.

### *B. The Online provision Allocation Algorithm*

We subsequently discuss the scenario where the revenue obtained probability is unknown to SNO. SNO must estimate and learn the revenue obtained probability $\phi_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$. An online provision allocation algorithm combining the contextual MAB with the optimization problem has been designed. The main idea of the designed algorithm is to develop the estimates of $\phi_{k,j}(t)$, and the upper estimates of the $r_{k,j}(t)$ with particular confidence bound for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$. Based on the upper estimates, SNO then solves a maximum weighted bipartite matching problem to allocate provisions to the regions.

For clarity, we first introduce how to develop the upper estimates of $\phi_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$, and then describe the overall algorithm.

Recall that

$$\hat{\phi}_{k,j}(t) = \boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \hat{\boldsymbol{\theta}}_{k,j}(t). \tag{9}$$

To estimate $\hat{\phi}_{k,j}(t)$, SNO should first estimate $\hat{\boldsymbol{\theta}}_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$. In this work, we adopt the regression analysis formulated as

$$(P1): \min_{\boldsymbol{\theta}_{k,j}} \sum_{t' \in t_{k,j}(T)} \left(r_{k,A_j}(t') - \boldsymbol{x}_{k,j}^{\mathrm{T}}(t') \cdot \boldsymbol{\theta}_{k,j} \cdot r_{A_j}\right)^2 \tag{10}$$

where $t_{k,j}(T)$ represents the period sets that the region $A_j$, $j \in \mathcal{I}_N$ is allocated to the provision $C_k$, $k \in \mathcal{I}_K$ in $T$ periods.

The solution to (10) is

$$\hat{\boldsymbol{\theta}}_{k,j}(T) = \boldsymbol{D}_{k,j}^{-1}(T) \cdot \sum_{t' \in t_{k,j}(T)} \boldsymbol{x}_{k,j}(t') \cdot \tilde{\phi}_{k,j}(t') \tag{11}$$

where

$$\boldsymbol{D}_{k,j}(T) = \sum_{t' \in t_{k,j}(T)} \boldsymbol{x}_{k,j}(t') \cdot \boldsymbol{x}_{k,j}^{\mathrm{T}}(t'). \tag{12}$$

From (11), we can see that $\boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \hat{\boldsymbol{\theta}}_{k,j}(t)$ is the unbiased estimation of $\phi_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$. And we conclude a conclusion from [25] as follows.

*Lemma 1:* With the probability at least $1 - \delta$, we have

$$\begin{aligned} &\left|\boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \hat{\boldsymbol{\theta}}_{k,j}(t) - \phi_{k,j}(t)\right| \\ &\leq \left(1 + \sqrt{\ln(2/\delta)/2}\right) \cdot \|\boldsymbol{x}_{k,j}(t)\|_{\boldsymbol{D}_{k,j}^{-1}(t)}. \end{aligned} \tag{13}$$

Lemma 1 presents a theoretical guarantee for the estimation of parameter $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ with $\boldsymbol{x}_{k,j}^{\mathrm{T}}(t) \cdot \hat{\boldsymbol{\theta}}_{k,j}(t)$. Based on this fact, we develop the upper estimates of the revenue for a region allocated to a provision, which is shown as follows:

$$g_{k,j}(t) = \hat{\phi}_{k,j}(t) \cdot r_{A_j} + \alpha_j \cdot \|\boldsymbol{x}_{k,j}(t)\|_{\boldsymbol{D}_{k,j}^{-1}(t)} \cdot \sqrt{\ln t} \tag{14}$$

where $g_{k,j}(t)$ represents the upper estimates of $r_{k,j}(t)$ with a particular confidence bound. $\hat{\phi}_{k,j}(t) \cdot r_{A_j}$ represents the estimation of $r_{k,j}(t)$ in period $t$. $\alpha_j$ is a hyper-parameter and

$$\alpha_j = \sqrt{1+N} \cdot r_{A_j}. \tag{15}$$

Then, based on the upper estimates $g_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, we design the online provision allocation algorithm depicted as follows. In the initialization process, SNO allocates provisions to regions once, obtains $\boldsymbol{x}_{k,j}(t)$ and the instantaneous revenue $r_{k,j}(t)$ for $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$. Then, for period $t$, SNO first observes the compound contextual features of the region, estimates the unknown parameters $\hat{\boldsymbol{\theta}}_{k,j}(t)$ in (11) and $\hat{\phi}_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ in (9). Then, for each region $A_j$, $j \in \mathcal{I}_N$, and each provision $C_k$, $k \in \mathcal{I}_K$, SNO calculates the upper estimates of revenue $g_{k,j}(t)$. By solving the maximum weighted bipartite matching problem with $g_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_M$, SNO obtains the optimal provision combination set solution and accordingly allocates provisions to regions. At the end the period, SNO obtains the revenue and updates the information on $\hat{\boldsymbol{\theta}}_{k,j}(t)$ and $\hat{\phi}_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$. The overall algorithm $\pi$ is presented in Algorithm 2.

In the following, we evaluate the algorithm performance by regret in (8). Since the regret comes from the fact that SNO does not choose the optimal provision combination set, the regret of our algorithm can be bounded as

$$\mathrm{REG}^{\pi}(T) = R^H(T) - R^{\pi}(T) \leq \Delta_{\max} \cdot \sum_{p=1}^{\binom{K}{N}N!} \tilde{t}_{\mathcal{C}_p}(T), \tag{16}$$

**Algorithm 2** Online provision Allocation Algorithm

Initialization: NCC/NRC observes the $\boldsymbol{x}_{k,j}(t)$ and obtains the revenue $r_{k,j}(t)$, for the region $A_j$, $j \in \mathcal{I}_N$ and provision $C_k$, $k \in \mathcal{I}_K$.
**for** period $t$ until $T$ **do**
  **for** $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ **do**
    Obtains the compound contextual information $\boldsymbol{x}_{k,j}(t)$
    Computes the matrix $\boldsymbol{D}_{k,j}(t)$ in (12)
    Computes the parameters $\hat{\boldsymbol{\theta}}_{k,j}(t)$ in (11)
    Computes $\hat{\phi}_{k,j}(t)$ in (9)
    Computes $g_{k,j}(t)$ in (14)
  **end for**
  Solves $\mathcal{O}racle(\{g_{k,j}(t)\})$, $k \in \mathcal{I}_K$ and $j \in \mathcal{I}_N$, obtains the provision combination set $\mathcal{C}_p(t)$.
  Allocates provisions to regions in accordance with $\mathcal{C}_p(t)$.
  $t_{k,j}(t) = t_{k,j}(t) + 1$, for $(k,j) \in \mathcal{C}_p(t)$.
**end for**

where

$$\Delta_{\max} = \max_{t \in \mathcal{I}_T} \left\{ \sum_{(k,j) \in \mathcal{C}^*(t)} r_{k,j}(t) - \sum_{(k,j) \in \mathcal{C}_p(t)} r_{k,j}(t) \right\}, \tag{17}$$

and $\tilde{t}_{\mathcal{C}_p}(T)$ represents the periods that the provision combination set $\mathcal{C}_p$, $p \in \mathcal{I}_{\binom{K}{N}N!}$ is not optimal but chosen by SNO in $T$ periods.

Then, we analyze $\tilde{t}_{\mathcal{C}_p}(T)$ for $p \in \mathcal{I}_{\binom{K}{N}N!}$ and $\mathrm{REG}^{\pi}(T)$. The conclusions are presented as follows.

*Lemma 2:* For $\tilde{t}_{\mathcal{C}_p}(T)$, we have

$$\mathbb{E}\left\{\tilde{t}_{\mathcal{C}_p}(T)\right\} \leq \frac{4(1+N)N^3 \cdot r_{\max}^2 \ln T}{\Delta_{\min}^2} + N + \frac{N^2\pi^2}{3} \tag{18}$$

where

$$\Delta_{\min} = \min_{t \in \mathcal{I}_T} \mathbb{E} \left\{ \sum_{(k,j) \in \mathcal{C}^*(t)} r_{k,j}(t) - \sum_{(k,j) \in \mathcal{C}_p(t)} r_{k,j}(t) \right\}. \tag{19}$$

*Theorem 1:* For algorithm $\pi$, the order of the regret is $O(\ln T)$.

*Proof:* The analysis process can refer to []. For the sake of brevity, the proof is omitted in this work. ■

It can be implied from Theorem 1 that since the regret order is $O(\ln T)$, the revenue loss of the proposed algorithm is bounded by $\omega_1 \cdot \ln T + \omega_2$ for constants $\omega_1$ and $\omega_2$. The intuitive meaning is that the revenue loss has a non-linear relationship with period $T$. From the analysis, the proposed algorithm is asymptotically bounded and has a performance guarantee for finite periods.

## VI. Performance Evaluation

In this section, we present simulations to evaluate the proposed provision allocation algorithm. We first describe the setup of simulation parameters and then the simulation results.

TABLE I
Compound contextual information for $0 < t < t_1 = 24$.

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|
| $C_1$ | $[1,1,1.4,.84]^{\mathrm{T}}$ | $[1,2,2.98,.32]^{\mathrm{T}}$ | $[1,3,4.34,.12]^{\mathrm{T}}$ | $[1,4,2.46,.22]^{\mathrm{T}}$ |
| $C_2$ | $[2,1,1.4,.84]^{\mathrm{T}}$ | $[2,2,2.98,.32]^{\mathrm{T}}$ | $[2,3,4.34,.12]^{\mathrm{T}}$ | $[2,4,2.46,.22]^{\mathrm{T}}$ |
| $C_3$ | $[3,1,1.4,.84]^{\mathrm{T}}$ | $[3,2,2.98,.32]^{\mathrm{T}}$ | $[3,3,4.34,.12]^{\mathrm{T}}$ | $[3,4,2.46,.22]^{\mathrm{T}}$ |
| $C_4$ | $[4,1,1.4,.84]^{\mathrm{T}}$ | $[4,2,2.98,.32]^{\mathrm{T}}$ | $[4,3,4.34,.12]^{\mathrm{T}}$ | $[4,4,2.46,.22]^{\mathrm{T}}$ |
| $C_5$ | $[5,1,1.4,.84]^{\mathrm{T}}$ | $[5,2,2.98,.32]^{\mathrm{T}}$ | $[5,3,4.34,.12]^{\mathrm{T}}$ | $[5,4,2.46,.22]^{\mathrm{T}}$ |
| $C_6$ | $[6,1,1.4,.84]^{\mathrm{T}}$ | $[6,2,2.98,.32]^{\mathrm{T}}$ | $[6,3,4.34,.12]^{\mathrm{T}}$ | $[6,4,2.46,.22]^{\mathrm{T}}$ |

### A. Simulation Parameters

The compound contextual information in our simulation is given as

$$\boldsymbol{x}_{k,j}(t) = [k, j, g(t), \mathcal{N}_j(t)]^{\mathrm{T}}, k \in \mathcal{I}_K, j \in \mathcal{I}_N. \tag{20}$$

The simulation parameters are clarified as follows. The QoS provisions and regions are denoted by indexes $k \in \{1,2,...,K\}$ and $j \in \{1,2,...,N\}$, respectively. In our simulation, we choose $K = 6$ and $N = 4$. The revenue obtained probability is related to the service time [9], we define a piecewise function taking service time as the variable:

$$g(t) = \begin{cases} \omega_3, & t \in [0, t_1] \\ \omega_4, & t \in [t_1, t_2] \end{cases} \tag{21}$$

where $\omega_3$ and $\omega_4$ are constants.[1] $\mathcal{N}_j(t)$ represents the user information containing the user type, (i.e., the Internet of Things device, the emergence type, the ship and mobile equipment, and the normal user type) and user number, which implies the latent traffic volume. In the simulation, we use the real number to represent the latent volume. The larger number represents the larger volume of the region [9]. The parameters are shown in Table I. Taking $\boldsymbol{x}_{1,1}(t) = [1,1,1.4,.84]^{\mathrm{T}}$ as an example, the first 1 represents provision $C_1$, the second 1 represents region $A_1$, 1.4 represents the service time function value, .84 (i.e., 0.84) is the user information representing the user type and latent traffic volume of the region.

To represent $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, we randomly generate $\boldsymbol{\theta}_{k,j}$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ and present them in Table II.

The information on $\phi_{k,j}$ and $r_{A_j}$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ is presented in Table III. Taking region $A_1$ with respect to provision $C_1$ as the example, the revenue obtained probability is $\phi_{1,1}(t) = \boldsymbol{x}_{1,1}^{\mathrm{T}}(t) \cdot \boldsymbol{\theta}_{1,1} = 0.1900$, and the maximal latent obtainable revenue is $r_{A_1} = \phi_{1,1} \cdot r_{A_1} = 0.6$.

In our simulations, we consider an extreme case that $\tilde{\phi}_{k,j}(t)$ is sampled following the Bernoulli distribution with mean value of $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, i.e., $\tilde{\phi}_{k,j}(t)$ equals to 1 with the probability of $\phi_{k,j}(t)$, and equals to 0 with the probability of $1 - \phi_{k,j}(t)$.

[1] In [9], it shows that the network selection by users is time-dependent, i.e., in the daytime, the service demand can be large while at other times, the service demand may be relatively small.

TABLE II
THE UNKNOWN PARAMETERS.

| $\boldsymbol{\theta}_{k,j}$ | unknown parameters |
|---|---|
| $\boldsymbol{\theta}_{1,1}$ | $[0.06507891, 0.03039441, 0.05278145, 0.02464874]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{1,2}$ | $[0.04261938, 0.01770781, 0.03382688, 0.08396560]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{1,3}$ | $[0.08016707, 0.04108847, 0.01159379, 0.04744703]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{1,4}$ | $[0.00298659, 0.00915860, 0.01366620, 0.08344974]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{2,1}$ | $[0.06478504, 0.08318699, 0.03126685, 0.06256880]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{2,2}$ | $[0.00209509, 0.05704747, 0.08118368, 0.03002668]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{2,3}$ | $[0.04157633, 0.06689239, 0.06445171, 0.02704707]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{2,4}$ | $[0.04696715, 0.09112315, 0.08046196, 0.06149898]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{3,1}$ | $[0.01098605, 0.04204004, 0.01445178, 0.07725486]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{3,2}$ | $[0.00563915, 0.00512468, 0.04327764, 0.09163134]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{3,3}$ | $[0.00165764, 0.01851609, 0.07193018, 0.02092529]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{3,4}$ | $[0.04089278, 0.05328067, 0.08539708, 0.00557047]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{4,1}$ | $[0.04320833, 0.07453994, 0.04339369, 0.03524397]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{4,2}$ | $[0.05602383, 0.00564825, 0.00982798, 0.04059740]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{4,3}$ | $[0.07011584, 0.08055636, 0.0139874, 0.07647559]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{4,4}$ | $[0.00696762, 0.0460277, 0.02930147, 0.07980496]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{5,1}$ | $[0.01909760, 0.07523230, 0.03621137, 0.05185359]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{5,2}$ | $[0.09122279, 0.05127172, 0.06749957, 0.04916657]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{5,3}$ | $[0.04966760, 0.01090616, 0.04823315, 0.04062562]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{5,4}$ | $[0.00245316, 0.06403389, 0.07570841, 0.07672658]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{6,1}$ | $[0.07320331, 0.04413893, 0.07958211, 0.08787199]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{6,2}$ | $[0.09058194, 0.07460044, 0.04842352, 0.03319661]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{6,3}$ | $[0.06796988, 0.07016280, 0.02078242, 0.05374153]^{\mathrm{T}}$ |
| $\boldsymbol{\theta}_{6,4}$ | $[0.08629245, 0.03201999, 0.03588407, 0.04191287]^{\mathrm{T}}$ |

TABLE III
THE RELATIONSHIP BETWEEN THE PROVISION AND REVENUE.

| | $\phi_{k,1}(t), r_{A_1}$ | $\phi_{k,2}(t), r_{A_2}$ | $\phi_{k,3}(t), r_{A_3}$ | $\phi_{k,4}(t), r_{A_4}$ |
|---|---|---|---|---|
| $C_1$ | 0.1900, .6 | 0.2057, .8 | 0.2594, .3 | 0.0915, .4 |
| $C_2$ | 0.3090, .6 | 0.3698, .8 | 0.5667, .3 | 0.6698, .4 |
| $C_3$ | 0.1601, .6 | 0.1854, .8 | 0.3752, .3 | 0.5471, .4 |
| $C_4$ | 0.3377, .6 | 0.2776, .8 | 0.5920, .3 | 0.3016, .4 |
| $C_5$ | 0.2649, .6 | 0.7755, .8 | 0.4952, .3 | 0.4715, .4 |
| $C_6$ | 0.6685, .6 | 0.4255, .8 | 0.7149, .3 | 0.7433, .4 |

### B. Simulation Results

Firstly, we consider the method $\pi^H$ where the parameters $\phi_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, are known *a priori* by SNO. By Algorithm 1, the optimal provision allocation set is to allocate provision $C_5$ to $A_2$, $C_6$ to $A_1$, $C_4$ to $A_3$ and $C_2$ to $A_4$. The optimal revenue in a period is $\phi_{5,2}(t) * r_{A_2} + \phi_{6,1}(t) * r_{A_1} + \phi_{4,3}(t) * r_{A_3} + \phi_{2,4}(t) * r_{A_4} = 0.7755 * 0.8 + 0.6685 * 0.6 + 0.5920 * 0.3 + 0.6698 * 0.4 = 1.46702$.

Next, we perform simulations on our proposed online provision allocation algorithm and present simulation results in Figures 2(a) and 2(b). The x-axis represents the period. The y-axes of Figures 2(a) and 2(b) represent the cumulative revenue and regret. It can be concluded from Figure 2(a) that the cumulative revenue of the proposed algorithm is less than that of the optimal algorithm. As can be seen from Figure 2(b), the revenue gap between the optimal algorithm and the proposed algorithm has a logarithmic relationship with the period. Finally, the regret converges to a stable value, which means that SNO finds the optimal provision combination set and continuously allocates the optimal provisions to regions.

TABLE IV
NUMBER OF PROVISIONS ALLOCATED TO REGIONS WITH $t = 40000$.

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|
| $C_1$ | 60 | 17 | 18 | 9 |
| $C_2$ | 31 | 34 | 370 | 39522 |
| $C_3$ | 1001 | 15 | 68 | 172 |
| $C_4$ | 148 | 22 | 38510 | 17 |
| $C_5$ | 24 | 39874 | 29 | 76 |
| $C_6$ | 38741 | 43 | 1010 | 209 |

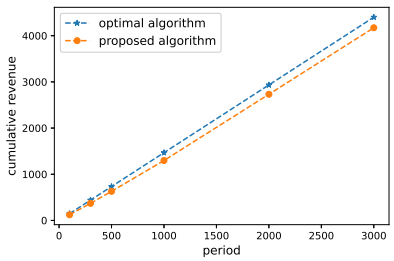
Setting the period to $40000$, we present the expected number of provisions allocated to regions in Table IV. As can be seen from the table, the provision relationships $C_5$ to $A_2$, $C_6$ to $A_1$, $C_4$ to $A_3$, and $C_2$ to $A_4$ are selected most times. The allocation result is in line with the optimal provision allocation of Algorithm 1.

Lastly, we further evaluate the proposed algorithm in comparison with other existing algorithms: random allocation algorithm, $\varepsilon-$greedy algorithm, Q-learning algorithm, and SARSA. In random allocation algorithm, SNO randomly allocates $N$ provisions to $N$ regions in each period. In $\varepsilon-$greedy algorithm, SNO first estimates the revenue of each provision allocated to each region, then solves a combinatorial optimization problem, and allocates provisions to regions in accordance with the solution. The overall process is performed with the probability of $1-\varepsilon$. With the probability of $\varepsilon$, SNO randomly allocates provisions to regions. In Q-learning and SARSA, the learning rate is set $0.05$, and the discount parameter is set $0.9$. SNO updates the Q-table in accordance with the revenue of provisions allocated to the regions. The simulation results are shown in Figure 2(c).

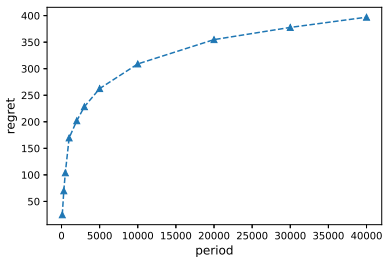
In Figure 2(c), the x-axis represents the period, the y-axis represents the regret. The random algorithm incurs the largest regret. The regrets of $\varepsilon-$greedy algorithms with $\varepsilon = 0.1$ and $\varepsilon = 0.05$ have linear relationship with the period. The regrets of the Q-learning algorithm, SARSA, and the proposed algorithm have a sub-linear relationship with the period. It can be observed that the proposed algorithm incurs the minimum regret and finally converges to the optimal provision allocation combination. The latent meaning is that SNO finally learns the information of the revenue obtained probability and finds the optimal provision allocation combination.

## VII. CONCLUSION

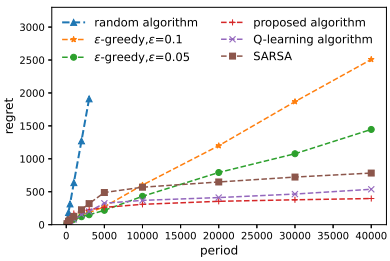
In this paper, we consider the allocation of QoS provisions to various regions for a satellite network. There co-exist multiple heterogeneous wireless networks operated by multiple operators adopting different technical specifications and pro-




(a) The comparison with other algorithms.




(b) The comparison with other algorithms.




(c) The comparison with other algorithms.

Fig. 2. The simulation results.

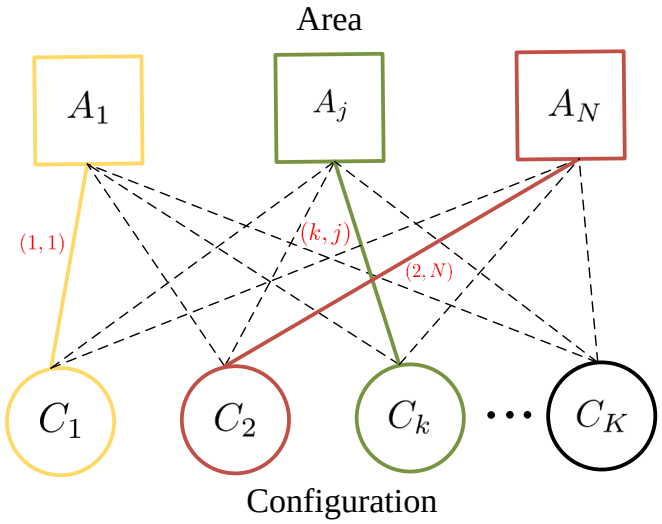



Fig. 3. The problem reformulation.

viding wireless Internet service with a specific QoS guarantee. Each wireless network infrastructure continuously broadcasts the NSI containing QoS provisions information. The user end can receive the broadcast information and select the appropriate network in accordance with the QoS criteria. To provide the predictive and proactive service to the region, a provision allocation scheme for SNO is first proposed, where SNO can allocate provisions to regions in accordance with the region information. To maximize the expected cumulative revenue, the provision allocation problem has been formulated as a combination of multi-armed bandit and combinatorial optimization problems. An online provision allocation algorithm has been proposed. The algorithm performance has been theoretically analyzed and validated by simulations.

## Appendix A
## Proof of Theorem 1

Allocating provisions to regions can be regarded as a maximal weighted bipartite matching problem. The problem is reformulated in the form of a graph in Figure 3.

As can be seen from Figure 3, there are regions $A_1, A_j, A_N$ and provisions $C_1, C_2, C_k, C_K$. We assume that the optimal provision combination is allocating provision $C_1$ to $A_1$, provision $C_2$ to $A_N$, and provision $C_k$ to $A_j$ in period $t$. For simplicity, we use the edge $(1,1)$, $(k,j)$ and $(2,N)$ to represent the provision allocation relationship, and we have $\mathcal{C}_p^*(t) = \{(1,1),(k,j),(2,N)\}$ in period $t$.

We subsequently clarify the notations used in our analysis. Denote $\tilde{t}_{k,j}(T)$ as the counter of the periods that the edge $(k,j)$ not belonging to the optimal provision combination set but has been chosen in $T$ periods after the initialization. For period $t \in \mathcal{I}_T$, we design the following counter rule: if non-optimal provision combination set $\mathcal{C}_p(t)$ is chosen, then the counters of all edges $(k,j) \in \mathcal{C}_p(t)$ increase 1. If optimal provision set $\mathcal{C}_p^*(t)$ is chosen, the counters of edges will not change. Following the counter rule we make, we have the following relationship as

$$\sum_{p=1}^{\binom{K}{N}N!} \tilde{t}_{\mathcal{C}_p}(T) \leq \sum_{p=1}^{\binom{K}{N}N!} \sum_{(k,j)\in\mathcal{C}_p} \tilde{t}_{k,j}(T) \tag{22}$$

where $\tilde{t}_{\mathcal{C}_p}(T)$ is the number of periods that the provision set $\mathcal{C}_p$ is not optimal but chosen in $T$ periods.

We then analyze the number $\tilde{t}_{k,j}(T)$ that the edge $(k,j) \notin \mathcal{C}_p^*(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$, $t \in \mathcal{I}_T$ has been selected in $T$ periods in (23)-(28).

Eq. (23) describes the period number that the edge $(k,j) \notin \mathcal{C}_p^*(t)$ but has been chosen in $T$ periods.

The parameter $l$ in (24) represents a positive integer. $\mathcal{A}(t)$ represents the event that there exists an edge $(k,j) \notin \mathcal{C}_p^*(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ has been chosen in period $t$.

Choosing the non-optimal edge means that a non-optimal provision combination set is chosen. Since $g_{k,j}(t)$, $k \in \mathcal{I}_K$, $j \in \mathcal{I}_N$ are taken as inputs to the maximal weighted bipartite matching problem for the solution, we can see there exists a non-optimal provision combination set $\mathcal{C}_p(t)$, and $\sum_{(k,j)\in\mathcal{C}_p(t)} g_{k,j}(t)$ is larger than $\sum_{(m,n)\in\mathcal{C}_p^*(t)} g_{m,n}(t)$. Based on this fact, we obtain (25)-(28).

In the following, we analyze

$$\sum_{i=1}^{N} g_{m_i,n_i}(t_{m_i,n_i}(t)) \leq \sum_{i=1}^{N} g_{k_i,j_i}(t_{k_i,j_i}(t)) \tag{29}$$

$$\tilde{t}_{k,j}(T) = \sum_{t=1}^{T} \mathbb{I}[(k,j) \in \mathcal{C}_p(t), \mathcal{C}_p(t) \neq \mathcal{C}_p^*(t)] \tag{23}$$

$$\leq l + \sum_{t=1}^{T} \mathbb{I}[\mathcal{A}(t), \tilde{t}_{k,j}(t) \geq l] \tag{24}$$

$$\leq l + \sum_{t=1}^{T} \mathbb{I}\left[\sum_{(k,j)\in\mathcal{C}_p(t)} g_{k,j}(t) \geq \sum_{(m,n)\in\mathcal{C}_p^*(t)} g_{m,n}(t), \tilde{t}_{k,j}(t) \geq l\right] \tag{25}$$

$$\leq l + \sum_{t=1}^{T} \mathbb{I}\left[g_{k_1,j_1}(t) + g_{k_2,j_2}(t) + \cdots + g_{k_N,j_N}(t) \geq g_{m_1,n_1}(t) + g_{m_2,n_2}(t) + \cdots + g_{m_N,n_N}(t), \tilde{t}_{k,j}(t) \geq l\right] \tag{26}$$

$$\leq l + \sum_{t=1}^{T} \mathbb{I}\left[\min_{1\leq t_{m_1,n_1}(t),\ldots,t_{m_N,n_N}(t)\leq t} \sum_{i=1}^{N} g_{m_i,n_i}(t_{m_i,n_i}(t)) \leq \max_{l\leq t_{k_1,j_1}(t),\ldots,t_{k_N,j_N}(t)\leq t} \sum_{i=1}^{N} g_{k_i,j_i}(t_{k_i,j_i}(t))\right] \tag{27}$$

$$\leq l + \sum_{t=1}^{T} \sum_{t_{m_1,n_1}(t)=1}^{t} \cdots \sum_{t_{m_n,n_N}(t)=1}^{t} \sum_{t_{k_1,j_1}(t)=l}^{t} \cdots \sum_{t_{k_N,j_N}(t)=l}^{t} \mathbb{I}\left[\sum_{i=1}^{N} g_{m_i,n_i}(t_{m_i,n_i}(t)) \leq \sum_{i=1}^{N} g_{k_i,j_i}(t_{k_i,j_i}(t))\right] \tag{28}$$

from (28). Combining (14), (29) can be expanded as

$$\begin{aligned}
&\sum_{i=1}^{N} \hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) \cdot r_{A_{n_i}} \\
&\quad + \sum_{i=1}^{N} \alpha_{n_i} \cdot \|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \cdot \sqrt{\ln t} \\
&\leq \sum_{i=1}^{N} \hat{\phi}_{k_i,j_i}(t_{k_i,j_i}(t)) \cdot r_{A_{j_i}} \\
&\quad + \sum_{i=1}^{N} \alpha_{j_i} \cdot \|\boldsymbol{x}_{k_i,j_i}(t_{k_i,j_i}(t))\|_{\boldsymbol{D}^{-1}_{k_i,j_i}(t_{k_i,j_i}(t))} \cdot \sqrt{\ln t}
\end{aligned} \tag{30}$$

which implies at least one of the three inequalities holds

$$\begin{aligned}
&-\sum_{i=1}^{N} \alpha_{n_i} \cdot \|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \cdot \sqrt{\ln t} \\
&+ \sum_{i=1}^{N} \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}} \geq \sum_{i=1}^{N} \hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) \cdot r_{A_{n_i}}
\end{aligned} \tag{31}$$

$$\begin{aligned}
&\sum_{i=1}^{N} \alpha_{j_i} \cdot \|\boldsymbol{x}_{k_i,j_i}(t_{k_i,j_i}(t))\|_{\boldsymbol{D}^{-1}_{k_i,j_i}(t_{k_i,j_i}(t))} \cdot \sqrt{\ln t} \\
&+ \sum_{i=1}^{N} \phi_{k_i,j_i}(t) \cdot r_{A_{j_i}} \leq \sum_{i=1}^{N} \hat{\phi}_{k_i,j_i}(t_{k_i,j_i}(t)) \cdot r_{A_{j_i}}
\end{aligned} \tag{32}$$

$$\begin{aligned}
&\sum_{i=1}^{N} \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}} < \sum_{i=1}^{N} \phi_{k_i,j_i}(t) \cdot r_{A_{j_i}} \\
&+ 2\sum_{i=1}^{N} \alpha_{j_i} \cdot \|\boldsymbol{x}_{k_i,j_i}(t_{k_i,j_i}(t))\|_{\boldsymbol{D}^{-1}_{k_i,j_i}(t_{k_i,j_i}(t))} \cdot \sqrt{\ln t}
\end{aligned} \tag{33}$$

We subsequently separately analyze (31), (32), and (33) as follows.

By union bound, the probability of (31) holding is upper bounded in (35)-(39).

Eq. (37) comes from the fact that we expand $\alpha_{n_i}$ in accordance with (15).

Eq. (38) comes from the fact from [**?**] that

$$\|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \leq \frac{1}{\sqrt{t_{m_i,n_i}(t)}}. \tag{34}$$

By Hoeffding inequality, we obtain the final conclusion in (39).

The same conclusion holds for (32).

Then, we analyze (33) as

$$\begin{aligned}
0 < &2\sum_{i=1}^{N} \alpha_{j_i} \cdot \|\boldsymbol{x}_{k_i,j_i}(t_{k_i,j_i}(t))\|_{\boldsymbol{D}^{-1}_{k_i,j_i}(t_{k_i,j_i}(t))} \cdot \sqrt{\ln t} \\
&+ \sum_{i=1}^{N} \phi_{k_i,j_i}(t) \cdot r_{A_{j_i}} - \sum_{i=1}^{N} \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}}
\end{aligned} \tag{40}$$

$$\begin{aligned}
&\leq \sum_{i=1}^{N} \phi_{k_i,j_i}(t) \cdot r_{A_{j_i}} - \sum_{i=1}^{N} \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}} \\
&\quad + \sum_{i=1}^{N} r_{A_{j_i}} \cdot \sqrt{\frac{4(1+N)\ln t}{t_{k_i,j_i}(t)}}
\end{aligned} \tag{41}$$

$$\begin{aligned}
&\leq \sum_{i=1}^{N} \phi_{k_i,j_i}(t) \cdot r_{A_{j_i}} - \sum_{i=1}^{N} \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}} \\
&\quad + N \cdot r_{\max} \cdot \sqrt{\frac{4(1+N)\ln t}{l}}
\end{aligned} \tag{42}$$

where (41) comes from (34). $l \in \{1, ..., t_{k_i,j_i}(t)\}$ in (42) is an integer, and $r_{\max}$ is

$$r_{\max} = \max_{i\in\mathcal{I}_N} r_{A_{j_i}}. \tag{43}$$

$$\mathbb{P}\left[(31)\text{ holds}\right] \leq \sum_{i=1}^{N} \mathbb{P}\left[\hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) \cdot r_{A_{n_i}} + \alpha_{n_i} \cdot \|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \cdot \sqrt{\ln t} \leq \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}}\right] \tag{35}$$

$$= \sum_{i=1}^{N} \mathbb{P}\left[\hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) \cdot r_{A_{n_i}} - \phi_{m_i,n_i}(t) \cdot r_{A_{n_i}} \leq -\alpha_{n_i} \cdot \|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \cdot \sqrt{\ln t}\right] \tag{36}$$

$$= \sum_{i=1}^{N} \mathbb{P}\left[\hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) - \phi_{m_i,n_i}(t) \leq -\sqrt{(N+1)} \cdot \|\boldsymbol{x}_{m_i,n_i}(t_{m_i,n_i}(t))\|_{\boldsymbol{D}^{-1}_{m_i,n_i}(t_{m_i,n_i}(t))} \cdot \sqrt{\ln t}\right] \tag{37}$$

$$= \sum_{i=1}^{N} \mathbb{P}\left[\hat{\phi}_{m_i,n_i}(t_{m_i,n_i}(t)) - \phi_{m_i,n_i}(t) \leq -\sqrt{(N+1)} \cdot \frac{1}{\sqrt{t_{m_i,n_i}(t)}} \sqrt{\ln t}\right] \tag{38}$$

$$\leq \sum_{i=1}^{N} e^{-2t_{m_i,n_i}(t) \cdot (N+1) \cdot \frac{1}{t_{m_i,n_i}(t)} \ln t} = N \cdot t^{-2(N+1)} \tag{39}$$

---

$$\mathbb{E}\{\tilde{t}_{k,j}(T)\} \leq l + \sum_{t=1}^{T} \sum_{t_{m_1,n_1}(t)=1}^{t} \cdots \sum_{t_{m_n,n_N}(t)=1}^{t} \sum_{t_{k_1,j_1}(t)=l}^{t} \cdots \sum_{t_{k_N,j_N}(t)=l}^{t} \mathbb{E}\left\{\mathbb{I}\left[\sum_{i=1}^{N} g_{m_i,n_i}(t_{m_i,n_i}(t)) \leq \sum_{i=1}^{N} g_{k_i,j_i}(t_{k_i,j_i}(t))\right]\right\} \tag{45}$$

$$\leq l + \sum_{t=1}^{T} \sum_{t_{m_1,n_1}(t)=1}^{t} \cdots \sum_{t_{m_n,n_N}(t)=1}^{t} \sum_{t_{k_1,j_1}(t)=l}^{t} \cdots \sum_{t_{k_N,j_N}(t)=l}^{t} \mathbb{P}[(31)\text{ holds}] + \mathbb{P}[(32)\text{ holds}] + \mathbb{P}[(33)\text{ holds}] \tag{46}$$

$$\leq l + \sum_{t=1}^{T} \sum_{t_{m_1,n_1}(t)=1}^{t} \cdots \sum_{t_{m_n,n_N}(t)=1}^{t} \sum_{t_{k_1,j_1}(t)=l}^{t} \cdots \sum_{t_{k_N,j_N}(t)=l}^{t} \left(N \cdot t^{-2(N+1)} + N \cdot t^{-2(N+1)} + \mathbb{P}((33)\text{ holds})\right) \tag{47}$$

$$\leq l + \sum_{t=1}^{T} t^{2N}(2N \cdot t^{-2(N+1)} + \mathbb{P}[(33)\text{ holds}]) \tag{48}$$

$$\leq l + \sum_{t=1}^{\infty}(2N \cdot t^{-2} + t^{2N}\mathbb{P}[(33)\text{ holds}]) \tag{49}$$

$$\leq \frac{4(1+N)N^2 \cdot r_{\max}^2 \ln T}{\Delta_{\min}^2} + 1 + \frac{N\pi^2}{3} \tag{50}$$

---

When $l \geq \left\lceil \frac{4(1+N)N^2 \cdot r_{\max}^2 \ln t}{\Delta_{\min}^2} \right\rceil$, where

$$\Delta_{\min} = \min_{t \in \mathcal{I}_T} \mathbb{E}\left\{\sum_{(k,j) \in \mathcal{C}^*(t)} r_{k,j}(t) - \sum_{(k,j) \in \mathcal{C}_p(t)} r_{k,j}(t)\right\}, \tag{44}$$

Eq. (42) does not hold and $\mathbb{P}[(33)\text{ holds}] = 0$.

Thus, the expectation of (28) can be given in (45)-(50), where (46) comes from the analysis (29)-(33), (47) comes from (39), (50) holds for $l \geq \left\lceil \frac{4(1+N)N^2 \cdot r_{\max}^2 \ln t}{\Delta_{\min}^2} \right\rceil$.

Hence,

$$\mathbb{E}\left\{\tilde{t}_{\mathcal{C}_p}(T)\right\} \leq \sum_{(k,j) \in \mathcal{C}_p} \mathbb{E}\left\{\tilde{t}_{k,j}(T)\right\} \tag{51}$$

$$\leq \frac{4(1+N)N^3 \cdot r_{\max}^2 \ln T}{\Delta_{\min}^2} + N + \frac{N^2\pi^2}{3} \tag{52}$$

Therefore, the expected cumulative regret can be expressed as

$$\text{REG}^{\pi}(T) \leq \Delta_{\max} \cdot \sum_{p=1}^{\binom{K}{N}N!} \mathbb{E}\left\{\tilde{t}_{\mathcal{C}_p}(T)\right\} \tag{53}$$

$$\leq \Delta_{\max} \cdot \binom{K}{N} N! \cdot \left(\frac{4(1+N)N^3 \cdot r_{\max}^2 \ln T}{\Delta_{\min}^2} + N + \frac{N^2\pi^2}{3}\right) \tag{54}$$

where

$$\Delta_{\max} = \max_{t \in \mathcal{I}_T} \left\{\sum_{(k,j) \in \mathcal{C}^*(t)} r_{k,j}(t) - \sum_{(k,j) \in \mathcal{C}_p(t)} r_{k,j}(t)\right\}, \tag{55}$$

Finally, we obtain the conclusion that the regret order of our proposed algorithm is $O(\ln T)$, the proof is included. ■

## REFERENCES


[1] H. Yao, L. Wang, X. Wang, Z. Lu and Y. Liu, "The Space-Terrestrial Integrated Network: An Overview," IEEE Communications Magazine, vol. 56, no. 9, pp. 178-185, Sept. 2018.

[2] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio and R. Prasad, "Satellite Communications Supporting Internet of Remote Things," IEEE Internet of Things Journal, vol. 3, no. 1, pp. 113-123, Feb. 2016.

[3] R. Ferrus, H. Koumaras, O. Sallent, G. Agapiou, T. Rasheed, M. A. Kourtis, C. Boustie, P. Gélard, and T. Ahmed, "SDN/NFV-enabled satellite communications networks: Opportunities, scenarios and challenges," Physical Communication, vol. 18, 2015.

[4] P. Zhou, J. Xu, W. Wang, C. Jiang, K. Wang, and J. Hu, "Human-Behavior and QoE-Aware Dynamic Channel Allocation for 5G Networks: A Latent Contextual Bandit Learning Approach," IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 436-451, Jan. 2020.

[5] M. Sheng, Y. Wang, J. Li, R. Liu, D. Zhou and L. He, "Toward a Flexible and Reconfigurable Broadband Satellite Network: Resource Management Architecture and Strategies," IEEE Wireless Communications, vol. 24, no. 4, pp. 127-133, Aug. 2017.

[6] I. D. Moscholios, V. G. Vassilakis, N. C. Sagias and M. D. Logothetis, "On Channel Sharing Policies in LEO Mobile Satellite Systems," IEEE Transactions on Aerospace and Electronic Systems, vol. 54, no. 4, pp. 1628-1640, Aug. 2018.

[7] C. C. González, E. F. Pupo, L. Atzori and M. Murroni, "Dynamic Radio Access Selection and Slice Allocation for Differentiated Traffic Management on Future Mobile Networks," *IEEE Transactions on Network and Service Management*, 2022.

[8] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," IEEE Transactions on Vehicular Technology, vol. 64, no. 6, pp. 2561–2576, Jun. 2015.

[9] Y. F. Hu, R. E. Sheriff, E. D. Re, R. Fantacci, and G. Giambene, "Satellite-UMTS Traffic Dimensioning and Resource Management Technique Analysis," IEEE Transactions on Vehicular Technology, vol. 47, no. 4, pp. 1329-1341, Nov. 1998.

[10] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO Satellite Communications: Architectures and Key Technologies," IEEE Wireless Communications, vol. 26, pp. 55-61, 2019.

[11] W. Huang, T. Song and J. An, "QA2: QoS-Guaranteed Access Assistance for Space-Air-Ground Internet of Vehicle Networks," IEEE Internet of Things Journal, 2021.

[12] X. Hu, Y. Zhang, X. Liao, Z. Liu, W. Wang, F. M. Ghannouchi, "Dynamic Beam Hopping Method Based on Multi-Objective Deep Reinforcement Learning for Next Generation Satellite Broadband Systems," IEEE Transactions on Broadcasting, vol. 66, no. 3, Sep. 2020.

[13] G. Cocco, T. de Cola, M. Angelone, Z. Katona, and S. Erl, "Radio resource management optimization of flexible satellite payloads for DVB-S2 systems," IEEE Trans. Broadcast., vol. 64, no. 2, pp. 266–280, 2018.

[14] J. Lei and M. Á. Vázquez-Castro, "Multibeam satellite frequency/time duality study and capacity optimization," Journal of Communications and Networks, vol. 13, no. 5, pp. 472-480, Oct. 2011.

[15] X. Liu, X. B. Zhai, W. Lu and C. Wu, "QoS-Guarantee Resource Allocation for Multibeam Satellite Industrial Internet of Things With NOMA," IEEE Transactions on Industrial Informatics, vol. 17, no. 3, pp. 2052-2061, Mar. 2021.

[16] L. Bertaux, .S Medjiah, P. Berthou, S. Abdellatif, A. Hakiri, P. Gelard, F. Patrick, and M. Fabrice, "Software Defined Networking and Virtualization for Broadband Satellite Networks," IEEE Communications Magazine, vol. 53, pp. 54-60, Mar. 2015.

[17] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P . Wang, Y . C. Liang, D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," IEEE Communications Surveys and Tutorials, vol. 21, no. 4, pp. 3133-3174, 2019.

[18] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. -C. Chen and L. Hanzo, "Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472-1514, 2020.

[19] Z. Zhang, H. Jiang, P. Tan and J. Slevinsky, "Channel exploration and exploitation with imperfect spectrum sensing in cognitive radio networks", *IEEE J. Sel. regions Commun.*, vol. 31, no. 3, pp. 429-441, 2013.

[20] Z. Xu, Z. Zhang, S. Wang, A. Jolfaei, A. K. Bashir, Y. Yan, and S. Mumtaz, "Decentralized Opportunistic Channel Access in CRNs Using Big-Data Driven Learning Algorithm," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 1, pp. 57-69, 2021.

[21] Lifeng Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive Medium Access-Exploration Exploitation and Competition," *IEEE Trans on Mobile Computing*, vol. 10, no. 2, pp.239-253, 2011.

[22] Y. Gai, B. Krishnamachari and R. Jain, "Combinatorial Network Optimization With Unknown Variables: Multi-Armed Bandits With Linear Rewards and Individual Observations," IEEE/ACM Transactions on Networking, vol. 20, no. 5, pp. 1466-1478, 2012.

[23] B. Kveton, Z. Wen, A. Ashkan and C. Szepesvari, "Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits," in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, PMLR, vol. 38, pp. 535-543, 2015.

[24] S. Li, B. Wang, S. Zhang and W. Chen, "Contextual Combinatorial Cascading Bandits," in Proceedings of the 33 rd International Conference on Machine Learning, New York, JMLR, vol. 48, pp. NY, USA, 2016.

[25] W. Chu, L. Li, L. Reyzin, R. E. Schapire, "Contextual bandits with linear Payoff functions," Journal of Machine Learning Research, vol. 15, pp. 208-214, 2011.