Machine Learning-Based Resource Allocation in Satellite Networks Supporting Internet of Remote Things

Di Zhou, *Member, IEEE,* Min Sheng, *Senior Member, IEEE*, Yixin Wang, Jiandong Li, *Fellow, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*— Satellite networks have been regarded as a promising architecture for supporting the Internet of remote things (IoRT) due to their advantages of wide coverage and high communication capacity in remote areas, which further promotes the development of the satellites for IoRT networks (SIoRTNs). The effectiveness of multi-dimensional resource collaboration has significant impacts on the IoRT data downloading performance. However, the environment's dynamics, e.g., channel conditions and solar infeed process, are unknown in practical scenarios, which poses daunting challenges in making efficient utilization of multi-dimensional resources. Motivated by this fact, we model the joint resource scheduling and IoRT data scheduling problem with the aim of maximizing the amount of the IoRT data of the overall network by applying the model-free reinforcement learning framework. To overcome the limitations of traditional reinforcement learning algorithms, we propose several feature functions by investigating the natural attributes of the multi-dimensional resources of the SIoRTNs, and further exploit the concept of function approximation to approximate the expected downloaded IoRT data given the network state. Furthermore, we propose a state-action-reward-state-action (SARSA) based actor-critic reinforcement learning (SACRL) resource allocation strategy to achieve the optimal resource allocation and IoRT data scheduling with casual information at LEO satellites. Simulations validate the convergence property and the effectiveness of the proposed SACRL algorithm in terms of the amount of the downloaded IoRT data. Particularly, we investigate the impact of typical network parameters on network performance to further provide guidance for future SIoRTN system design.**

***Index Terms*— Satellites for Internet of remote things networks, resource allocation, stochastic fluctuations in charge, casual knowledge, actor-critic reinforcement learning.**

I. Introduction

THE Internet of remote things (IoRT) for terrestrial deployments becomes of paramount importance for the next-generation communication system [1]. However, the sensors for obtaining the IoRT data of interest are always located in wide and remote areas in many application scenarios (e.g., monitoring of remote areas, vegetation extent monitoring, and spatiotemporal continuous information of PM2.5 monitoring) [2], [3], which may cause the IoRT data cannot be served by terrestrial access networks. As a consequence, satellite networks that consist of Low Earth orbit (LEO) satellites and geostationary Earth orbit (GEO) satellites can provide a promising alternative solution by exploiting the satellite networks to offload the IoRT data due to their advantages of wide coverage and high communication capacity in remote areas. Notably, due to the power limitations of the sensor equipments, the obtained IoRT data should be first transmitted to the LEO satellite [4]. Afterwards, the way the IoRT data is transferred to the database center depends on whether there is a direct LEO satellite-to-ground link. In addition, GEO satellites can provide a seamless service for LEO satellites to solve the limited communication resource problem of the direct LEO satellite-to-ground station links [5], [6]. Consequently, satellites for IoRT networks (SIoRTNs) have been regarded as a new infrastructure of the next-generation wireless communication system [7], [8].

The mass and size limitations of the LEO satellites have severely limited energy supply and buffer capacity to collect, store, and transmit the obtained IoRT data in SIoRTNs. Furthermore, LEO satellites may spend over 30% of their time during eclipse phase, where the large-scale discharging process may lead to the fact that the battery possibly ends up in a critically low state of charge (SoC), which further restricts the IoRT data acquisition and transmission, thereby deteriorating network performance [9], [10]. Therefore, the efficient multi-dimensional resources allocation strategy is the core element to improve the network performance. However, only causal knowledge about the channel fading status and amount of fluctuating solar infeed is available, which may further impose

Manuscript received October 25, 2020; revised February 20, 2021; accepted April 18, 2021. Date of publication April 30, 2021; date of current version October 11, 2021. This work was supported in part by the Natural Science Foundation of China under Grant U19B2025, Grant 61725103, and Grant 62001347; in part by the China Postdoctoral Science Foundation under Grant 2019TQ0241 and Grant 2020M673344; in part by the Fundamental Research Funds for the Central Universities under Grant XJS200117; and in part by NSF under Grant EARS-1839818, Grant CNS1717454, Grant CNS-1731424, and Grant CNS-1702850. The associate editor coordinating the review of this article and approving it for publication was X. Cheng. *(Corresponding author: Min Sheng.)*

Di Zhou, Min Sheng, Yixin Wang, and Jiandong Li are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: zhoudi@xidian.edu.cn; msheng@mail.xidian.edu.cn; yixinw@stu.xidian.edu.cn; jdli@mail.xidian.edu.cn;).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: zhan2@uh.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3075289.

Digital Object Identifier 10.1109/TWC.2021.3075289

1536-1276 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

great technical challenges on resource allocation design during the IoRT data downloading process, and can be specifically shown as the following two aspects:

- **Time-variant channel fading process**. The time-variant channel fading process in SIoRTNs concludes two aspects, i.e., the direct LEO satellite to ground station downlinks (SGDLs) and inter-satellite links (ISLs). Notably, due to the atmospheric precipitation (rain, fog, clouds, etc.) where rain attenuation is the most dominant external impairment factor, SGDLs may suffer from the severe signal attenuation. Especially at the Ku- and Ka-frequency bands, reduction of the link capacity (sometimes even complete outages for non-negligible periods of time) can be caused by the atmospheric attenuation [11]. With regard to ISLs, their communication performance is mainly affected by the external free space loss, where the distance change caused by satellite orbit motion is the most dominant factor.
- **Stochastically fluctuating solar infeed process**. The energy demands of LEOs for the IoRT data acquisition and transmission depend on the on-board battery storage and the harvested solar energy during the Sun phase. However, due to solar panel loss and cloud occlusion, the solar energy harvesting process is dynamic, which causes the amount of harvested energy in sunlight varies over time [12], [13]. To efficiently utilize the harvested energy for the IoRT data acquisition and transmission, the stochastically fluctuating solar infeed process should be considered together with the dynamic channel fading process.

Motivated by the aforementioned practical problems, some existing efforts have focused on the energy and channel aware resource allocation strategies design [14]–[18], where perfect non-causal knowledge about either the solar infeed process or the channel fading process of the SGDLs is assumed to be available in satellites. To deal with the dynamic satellite environment issues with only causal knowledge at satellites, other works pay attention to the stochastic optimization and learning based resource allocation strategies [19]–[23]. However, these existing works mainly utilize the RL and Markov decision process (MDP) based algorithms to solve finite state problems and neglect the coupling effects of the aforementioned two stochastic processes on resource allocation strategy. Therefore, the efficient resource allocation policies should be carefully designed with joint consideration of the time-variant channel fading process and the stochastically fluctuating solar infeed process. In particular, power allocation in battery management and buffer management during the IoRT data acquisition and transmission scheduling process play a pivotal role in enhancing the service efficiency for IoRT over satellites.

In this paper, we formulate the aforementioned two practical and dynamic processes into the resource allocation for IoRT data scheduling (RAIDS) problem. Considering the dynamic environments, the proposed RAIDS problem is formulated with the model-free reinforcement learning (RL) framework aiming at maximizing the amount of the long-term downloaded IoRT data. Furthermore, we propose a state-action-reward-state-action (SARSA) based actor-critic reinforcement learning (SACRL) resource allocation strategy to accommodate highly dynamic environments and achieve the optimal power allocation policy and IoRT data scheduling policy including data acquisition and transmission at the LEO satellite. In particular, by excavating the resource characteristics of satellites, we design several key feature functions to approximate the action-value function. In a nutshell, artificial intelligence technology is introduced to solve the dynamic resource allocation and IoRT data scheduling problem with joint consideration of the dynamic channel fading and solar infeed processes in SIoRTNs.

The main contributions of this paper are summarized as follows:

- *Practical SIoRTN system*: We formulate the online RAIDS problem by exploiting the model-free RL framework to adapt to the realistic issues of dynamic environments. Notably, it is not necessary to know completely time-variant channel fading and stochastically fluctuating solar infeed states, as well as their transition probabilities and rewards in the proposed formulation framework.
- *Feature functions considering satellite characteristics*: By investigating the natural attributes of the multi-dimensional resources of the SIoRTNs, several key feature functions are defined to provide an efficient model of the effects of possible actions on the on-board transponder at the LEO and possible transmission power values on the state of the LEO's transmitter. Notably, the action-value mapping relationship can be effectively described by using the proposed feature functions instead of the precise mathematical action-value relationship which is always absent in practical satellite systems.
- *RL-based dynamic multi-dimensional resources allocation algorithm*: By exploiting the proposed feature functions, we design an actor-critic resource scheduling decision-making architecture. Furthermore, a SACRL algorithm is proposed according to casual knowledge to obtain the optimal resource allocation policy and IoRT data scheduling policy to accommodate highly dynamic environments and achieve long-term IoRT data downloading performance.
- *Verification*: Extensive simulations have been conducted to evaluate the developed strategy and validate the effectiveness and efficiency of the proposed SACRL algorithm in terms of learning effectiveness and the amount of downloaded IoRT data. We also investigate the impact of several typical network parameters on the power allocation and IoRT data scheduling strategies, which can provide a guidance for system design for future SIoRTNs.

The remainder of the paper is organized as follows. We first overview the related works in Section II. Then, we elaborate the system model in Section III and formulate the RAIDS problem by exploiting the model free RL framework in Section IV. Next, we propose an SACRL resource allocation algorithm to adapt to dynamic environmental changes in Section V. In Section VI, we conduct numerous simulations and evaluations to verify the effectiveness of the proposed algorithm. Finally, we draw a conclusion in Section VII.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

II. Related Works

In this section, we elaborate on the recent research on resource allocation in satellite networks, which has also been in full swing and can be classified into two categories, i.e., the energy and channel aware resource allocation strategies [14]–[18] and the stochastic optimization and learning based resource allocation strategies [19]–[23].

A. Energy and Channel Aware Resource Allocation

Efforts on various resource allocation strategy design have been made for efficient data downloading in satellite networks [14]–[18]. Specifically, a collaborative data delivery strategy was developed in [14] to achieve optimal throughput by optimally scheduling the communication resources of the ISLs. To further consider the impact of the channel fading process on data transmission process, we investigated the joint power allocation and data scheduling problem in broadband data relay satellite networks and further proposed a two-stage algorithm to achieve high network performance in [15]. By exploiting the high-capacity LEO-based backhaul, the recent work [16] proposed two matching algorithms to solve the joint user scheduling and backhaul resource allocation problem to achieve efficient data downloading. In addition, to formulate the dynamic energy consumption into the data transmission process, an extended time-expanded graph for energy flows was proposed in [17] to characterize the impact of energy consumption and harvesting processes on data transmission process in satellite networks. Furthermore, a primal decomposition based multi-resources allocation scheme was proposed in [18] to efficiently solve the downloaded data maximization problem with energy constraints.

However, the researches mentioned above assume that perfect non-causal knowledge about either the solar infeed process or the channel fading process of the SGDLs can be obtained on the satellite, which may result in that they cannot be directly applied to the network scenarios where only causal knowledge is available.

B. Stochastic Optimization and Learning Based Resource Allocation

Due to the practical stochastic fluctuations in charge and stochastic property of channel conditions of SGDLs and ISLs, the sequential resource allocation and data scheduling decisions in satellite networks belong to stochastic optimization problems. Therefore, some efforts concentrate on designing stochastic optimization and learning based resource allocation algorithms to overcome the requirements of the offline setting in energy and channel aware resource allocation strategies [19]–[23]. Specifically, a distributionally robust two-stage stochastic optimization framework was proposed in [19] to achieve the efficient long-term data downloading performance with considering the uncertain distribution of the long-term random data arrival. Besides, by considering the practical random solar infeed, we exploited the MDP to formulate the dynamic resource allocation problem in our recent work [20] and further proposed a backward induction based strategy to maximize the long-term network utility. Notably, the transition probability of the stochastically harvested solar energy is assumed available at the satellite in [20]. However, a perfect and complete model of the environment is always absent in satellite environments. Therefore, the recent works in [21] and [22] focus on RL framework for dynamic resource allocation in satellite networks. Besides, a RL-based capacity management strategy was investigated in [23] to maximize the long-term network utility in the three-layer heterogeneous satellite network. However, since these existing dynamic resource allocation strategies either neglect the stochastically fluctuating solar infeed process or assume that the transition probability of the time-variant solar energy during the Sun phase is available, they cannot be directly applied to solve the RAIDS problem in SIoRTNs with joint consideration of the causal information of the time-variant channel fading process and the stochastically fluctuating solar infeed process.

In light of the existing works, the joint resource allocation and IoRT data scheduling strategies in SIoRTNs are still worthy of further study to improve the overall and long-term network performance. In contrast to most of existing works that use the RL and MDP based algorithms to solve finite state problems, this paper significantly extends the aforementioned work in [20] and takes the research a step further by applying the model-free RL to formulate the RAIDS problem with the joint consideration of the time-variant channel fading process and the stochastically fluctuating solar infeed process. Furthermore, we design a SACRL algorithm to acquire the optimal resource scheduling and data scheduling policies with continuous state for SIoRTNs.

III. System Model

In this section, we first introduce the considered SIoRTN in this paper. Then, we present two models, i.e., the dynamic channel model and the dynamic on-board energy model, which have significant impacts on the resource allocation during the IoRT data scheduling process in the practical SIoRTNs. The key notations and abbreviations are summarized in Table I.

A. Network Configuration

We consider a SIoRTN as shown in Fig. 1, which consists of a LEO, several isolated ground stations denoted by $\mathcal{GS} = \{GS_1, \cdots GS_J\}$, and several GEO relay satellites denoted by $\mathcal{RS} = \{GEO_1, \cdots GEO_L\}$ which can provide a seamless service coverage for the LEO. Specifically, the LEO first obtains IoRT data by turning on its on-board transponders. Then, the obtained IoRT data can be stored in the LEO for further transmission (i.e., either through the direct SGDLs or through the ISLs). Eventually, all IoRT data will be aggregated to the database server for further data analysis in scientific experiments and climate analysis, etc. [4].[1]

Due to the periodic orbital motion characteristics of satellites, the network topology (i.e., the connection relationship

[1] In the practical SIoRTN scenario, once the ground stations and relay satellites receive the IoRT data, the IoRT data can be immediately transmitted to the database server through the terrestrial fixed links and the fixed GEO relay satellite to the dedicated ground stations links, respectively.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

TABLE I
THE KEY NOTATIONS AND ABBREVIATIONS

Notations/Abbreviations	Physical meaning/Full name
$\tau_{i,e}$, $\tau_{i,c}$	The effective time to collect solar energy during its sunlight phase and the effective available time of the network communication links during time slot i.
C_{isr}^{l}, C_{isd}^{j}	The achievable data rates (in bps) of an ISL and a SGDL between the LEO and the ground station GS_j in slot i.
E_i^c, E_{it}^c, E_{ia}	The energy consumption of the LEO in time slot i; The energy consumption for IoRT data transmission during slot i; The actual energy available in time slot i.
P_{da}, P_n, P_i	The constant reception power for data acquisition; The nominal operation power for normal operations in the LEO; The selected transmission power in slot i.
B_i, D_i, $B_{\max}$, $D_{\max}$	The available energy in the battery and the data buffer level of the LEO at the beginning of time slot i; The maximum capacity of the battery; The buffer size.
S_i, A_i, A_{S_i}, $\Theta^{\Pi}(S_i, A_i)$	The state of the system, the system action, the feasible action set, and an action-value function in time slot i.
$f_z(S_i, A_i)$, ϖ_{critic}, ϖ_{actor}, $\hat{\Theta}_i^{\Pi}(S_i, A_i, \varpi_{\text{actor}})$	The z-th ($z = 1, \cdots, Z$) feature function for state-action pair (S_i, A_i); The weight matrix in the critic network; The weight matrix in the actor network; The approximated action-value function with a given state action pair (S_i, A_i).
$\mathcal{L}(\varpi_{\text{actor}})$, ϵ_i, β_i	The loss function based on mean squared error; The ϵ parameter used in the ϵ-greedy policy and the learning rate β during time slot i.
T_{train}, T_{copy}, $T_{learning}$	The update intervals for the actor network; The update intervals for the critic network; The total number of iterations.
LEO, GEO	Low Earth orbit; Geostationary Earth orbit.
IoRT, SIoRTNs	The Internet of remote things; Satellites for IoRT networks.
SGDLs, ISLs	The direct LEO satellite to ground station downlinks; inter-satellite links.
RAIDS, SARSA, SACRL	Resource allocation during IoRT data scheduling; State-action-reward-state-action; SARSA based actor-critic reinforcement learning.
MDP, CDF	Markov decision process; Cumulative distribution function.
A-ADID-E	the average amount of downloaded IoRT data during an episode period.

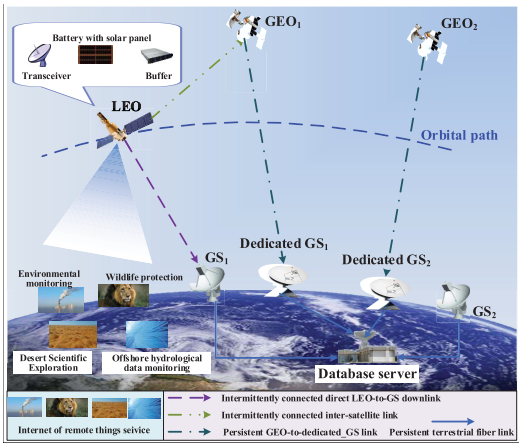


Fig. 1. Network model.

between the LEO and the ground stations as well as GEO (i.e., relay satellites) exhibits periodic changes. A network topology period is defined as an episode and the time duration of an episode is denoted as $\mathcal{T}$.[2] Besides, we assume that the whole satellite system operates in a slot-by-slot fashion and the network topology remains unchanged within a time slot [24]. An episode is discretized into several time slots with slot length τ. To more accurately depict the available network resources during time slot i, we denote the effective available time of the network communication links as $\tau_{i,c}$ and the effective time for the LEO to collect solar energy during its sunlight phase as $\tau_{i,e}$. Note that the specific values of $\tau_{i,c}$ and $\tau_{i,e}$ can be obtained by the Satellite Tool Kit (STK). Since the LEO is equipped with a limited number of on-board transponders, the LEO can only establish one communication link for data transmission in a time slot (i.e., either with a ground station or with a GEO relay satellite). Generally speaking, due to a relatively large antenna equipped by the ground station, the achievable data-rate of a SGDL is higher than that of an ISL [13]. Consequently, to make an efficient use of shorter but faster direct SGDL resources, the LEO satellite prefers to establish a communication link with a ground station when a SGDL exists.

[2] $\mathcal{T}$ is the least common multiple of the period of the LEO's orbit around the Earth and the period of the Earth's rotation.

B. Dynamic Channel Model

In this subsection, we focus on the channel dynamics of SGDLs and ISLs. First, we elaborate on the channel model of the ISLs affected by the distance change caused by satellite orbit motion. Then, since rain attenuation dominates propagation loss at the Ka band and above, we depict the cumulative distribution function (CDF) of the rain attenuation at the location of the specific ground station to characterize the random channel characteristics of SGDLs.

Dynamic ISLs channel model: Following recent work [25], the achievable data rate (in bps) of ISLs in slot i denoted by C_{isr}^{l} is shown as:

$$C_{isr}^{l} = \frac{P_{tsr}^{i} G_{tsr} G_{rsr} L_{fl}^{i}}{k U_s \cdot (E_b/N_0)_{req} \cdot A}. \tag{1}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

In (1), P_{tsr}^{i} is the transmission power (in W) of the LEO in time slot i. G_{tsr} and G_{rsr} are the transmitting antenna gain of the LEO and receiving antenna gain of the GEO relay satellite, respectively. Besides, k is the Boltzmann constant (in JK^{-1}) and U_s is the total system noise temperature (in K). $\left(E_b/N_0\right)_{req}$ is the required ratio of received energy-per-bit to noise-density and A is the link margin. L_{fl}^{i} is the free space loss of the ISL between the LEO and the GEO relay satellite, which can be expressed as:

$$L_{fl}^{i}=\left(\frac{c}{4\pi\cdot SR_i\cdot f}\right)^2. \tag{2}$$

Here, c is the speed of light (in km/s) and SR_i is the slant range (in km) in slot i. f is the communications center frequency (in Hz) of ISLs.

Dynamic SGDLs channel model: Due to the differentiated atmospheric environment of the geographical location of different ground stations, the channel statuses of different ground station locations are different. We first obtain the CDFs of rain attenuation at different ground station locations by collecting statistics on rainfall at the location of the ground station according to ITU-R P.837. Specifically, we collect the location and weather conditions of each ground station to obtain its probability of rainfall intensity according to the International Telecommunication Union (ITU) model in [26]. With the probability, the CDF of the rain attenuation at each ground station can be characterized, which shows the percentage of time a certain attenuation value is exceeded at a given frequency band. Following Recommendation ITU-R P.618-13, the predicted rain attenuation exceeded for $p\%$ of an average year for ground station location GS_j denoted by B_p^j can be expressed as:

$$B_p^j=\psi_A\cdot L_E^j\left(\frac{p}{0.01}\right)^{-\Lambda}. \tag{3}$$

Here, ψ_A (in dB/km) denotes the specific attenuation per kilometer which can be obtained according to Recommendation ITU-R P.838-3. L_E^j (in km) refers to the equivalent effective slant-path length at the ground station location GS_j, which can be obtained according to Recommendation ITU-R P.618-12. Besides, $\left(\psi_A\cdot L_E^j\right)$ is the predicted attenuation exceeded for 0.01% of an average year, which can be obtained according to Recommendation ITU-R P.837-7.

On the basis of the mentioned above parameters, we can calculate Λ according to Recommendation ITU-R P.618-13 as follows:

$$\Lambda=0.655+0.033\ln p-0.045\ln\left(\psi_A^i\cdot L_E\right)-p\left(1-p\right)\sin\phi. \tag{4}$$

Until now, we can obtain the CDF of the rain attenuation at each ground station location. At each time slot, the atmospheric attenuation at a specific ground station location is assumed to be independent and identically distributed (i.i.d.) and the specific rain attenuation can be obtained by sampling on the corresponding CDFs. We assume that only causal knowledge is available at the LEO, which means that during time slot i, the LEO has knowledge about its satellite downlink channel status at current time slot.

To further obtain the achievable data rate of a specific satellite downlink, we first calculate its corresponding SNR. For example, if the satellite downlink between the LEO and the ground station GS_j exists, the specific SNR of this satellite downlink in slot i is expressed as follows [27]:

$$SNR_i^j=\frac{P_{tsd}^{i}G_{tsd}G_{rsd}L_{fj}^{i}L_{rj}^{i}}{N}. \tag{5}$$

In (5), P_{tsd}^{i} and G_{tsd} are the transmission power (in W) and the transmitting antenna gain of the LEO in slot i, respectively. G_{rsd} is the receiving antenna gain of the ground station. L_{fj}^{i} and L_{rj}^{i} are the free space loss and atmospheric attenuation between the LEO and the ground station GS_j, which can be calculated according to (2) and (3), respectively.

On the basis of the mentioned above parameters, the achievable data rate (in bps) of a satellite downlink between the LEO and the ground station GS_j in time slot i if it exists, denoted by C_{isd}^{j}, can be obtained according to the Shannon formula [28] as follows:

$$C_{isd}^{j}=B_c\cdot\log_2\left(1+SNR_i^j\right). \tag{6}$$

Here, B_c is the available satellite downlink channel bandwidth.

C. Dynamic On-Board Energy Model

Following [10], we adopt a state-of-the-art formal battery model which is applied to the Iridium satellite system and consists of the following two aspects.

Dynamic energy consumption model: To maintain the normal operations of the LEO and satisfy the necessary IoRT data scheduling requirements, the LEO has the necessary energy consumption, which is mainly composed of two parts, i.e., the static energy consumption (such as energy consumption of the thermal control subsystem and satellite service subsystem) which is the daily mode and the dynamic energy consumption (such as energy consumption for IoRT data acquisition and transmission) which is the on-demand mode [9]. The nominal operation power for normal operations in the LEO denoted by P_n is assumed to be a constant parameter which means that the energy consumption for the normal operation is daily rather than sudden on demand.

We denote E_i^c as the energy consumption of the LEO in time slot i, which can be expressed as:

$$E_i^c=P_{da}\cdot\tau\cdot\varsigma_i+E_{it}^c+P_n\cdot\tau. \tag{7}$$

Here, P_{da} is the constant reception power for IoRT data acquisition. ς_i is a binary variable which denotes that whether the on-board transponder of the LEO is turned on to collect IoRT data during time slot i. Thus, the first item and the third item in the righthand of (7) are the energy consumption for IoRT data acquisition and for normal operation during time slot i, respectively. E_{it}^c in (7) is the energy consumption for IoRT data transmission during slot i. Note that due to the constrained available on-board energy, the actual energy that can be consumed must be less than the actual energy available denoted by E_{ia} in time slot i. Therefore, the specific E_{it}^c can be calculated by

$$E_{it}^c=\min\left\{P_i\cdot\left(\min\left\{\frac{D_i}{C_i(H_i,P_i)},\tau_{i,c}\right\}\right),E_{ia}\right\}. \tag{8}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

In (8), P_i and $C_i(H_i, P_i)$ refer to the selected transmission power and the achievable data rate corresponding to this transmission power in slot i, respectively, which means that if the LEO establishes a communication link with a ground station during time slot i, $P_i = P^i_{tsd}$ and $C_i(H_i, P_i) = C^j_{isd}$, otherwise $P_i = P^i_{tsr}$ and $C_i(H_i, P_i) = C^l_{isr}$. H_i denotes the channel coefficients. Specifically, when the LEO establishes a communication link with a ground station during time slot i, $H_i = G_{tsd}G_{rsd}L^i_{fj}L^i_{rj}/N$. Otherwise, $H_i = (G_{tsr}G_{rsr}L^i_{fl})/(kU_s \cdot (E_b/N_0)_{req} \cdot A)$, which means that LEO selects a GEO relay satellite to relay its stored IoRT data to the ground in time slot i. Notably, due to the limitation of the amount of IoRT data to be transmitted, the actual data transmission time of the LEO in time slot i is $\min\{D_i/C_i(H_i, P_i), \tau_{i,c}\}$, where D_i refers to the data buffer level of the LEO, i.e., the amount of data to be transmitted in the buffer of the LEO at the beginning of time slot i. Besides, if $P_i = 0$, $C_i(H_i, P_i) = 0$ which means that the LEO does not transmit IoRT data during time slot i. Besides, if the battery level, i.e., the available energy in the battery at the beginning of time slot i denoted by B_i is less than a specific threshold, i.e., $B_i \le B_{\min} + P_n\tau$, there is actually no extra energy available for data acquisition and transmission in this time slot. Thus, the actual energy available in time slot i, i.e., E_{ia} is represented as

$$E_{ia} = \max\{B_i - B_{\min} - P_n \cdot \tau, 0\}. \tag{9}$$

Here, $B_{\min}$ is a safety threshold. Once B_i drops to or below this safety threshold $B_{\min}$, B_i is assumed to be at a critically low battery level. Specifically, $B_{\min} = (1-\chi) \cdot B_{\max} \ge 0$ can be utilized to prevent battery over-stress, wherein χ is the maximum depth of discharge and $B_{\max}$ is the battery capacity.

Dynamic Energy Harvesting Model: The on-board battery of the LEO is recharged only if the LEO is currently exposed to the Sun. Furthermore, since the solar rays and the loss of solar panels are dynamic, the solar energy that can be harvested by batteries in the Sun phase fluctuates over time [12]. Therefore, we assume that when the LEO is in the Sun phase, the battery recharge rate in slot i denoted by E^h_i, is an i.i.d. process across time slots. After $\left(E^h_i \cdot \tau_{i,e}\right)$ is harvested, it is stored in a rechargeable finite battery with the maximum capacity $B_{\max}$. Beside, in practical satellite systems, the on-board batteries are causal, which means that the energy harvested in the current time slot can only be stored for use in the forthcoming time slots. In other words, the on-board energy that can be used in the current time slot can only be the solar energy stored before this time slot. Due to the constrained battery capacity, once the available energy in the battery hits $B_{\max}$ during a recharging slot, it will remain at $B_{\max}$ for the remainder of that slot. Note that the battery recharge rate should be zero, i.e., $E^h_i = 0$ when the LEO is on eclipse.

IV. Reinforcement Learning Based Problem Formulation

In this section, we formulate the process of resource consumption during the IoRT data scheduling process as the process of resource status transfer by utilizing the model-free RL framework. We first define some key elements in RL including the network state, resource scheduling action, and the network immediate reward [29]. Then, we present the action-value function according to the defined system state, action, and network immediate reward.

1) State: The state of the system in slot i is denoted as

$$S_i = \left\{D_i, B_i, E^h_i, H_i\right\}. \tag{10}$$

Note that D_i, B_i, E^h_i, and H_i can take any value within their respective feasible continuous range. Consequently, set S_i contains an infinite number of possible states.

2) Action: In the resource allocation problem during IoRT data scheduling process, the LEO needs to decide whether to turn on its on-board transponder for data acquisition, whether to transmit data, and the power level to transmit data in each time slot. Notably, the available transmission power in real satellite systems is discrete in several modes. Therefore, the action in slot i can be expressed as $A_i = \{\varsigma_i, P_i\}$, where, $\varsigma_i \in \{0, 1\}$ and $\varsigma_i = 1$ means that the LEO chooses to turn on its on-board transponder to collect IoRT data. P_i denotes the selected transmission power by the LEO. Note that $P_i = 0$ means that the LEO does not transmit IoRT data during time slot i. Besides, we denote $P^{\max}_{tsd}$ and $P^{\max}_{tsr}$ are the maximum values of the transmission power for SGDLs and ISLs, respectively. Notably, $P^{\max}_{tsd}$ and $P^{\max}_{tsr}$ can be discretized into several intervals by step sizes δ_{tsd} and δ_{tsr}, which means that δ_{tsd} and δ_{tsr} are two quantization levels. Furthermore, we denote $\{0 : \delta_{tsd} : P^{\max}_{tsd}\}$ and $\{0 : \delta_{tsr} : P^{\max}_{tsr}\}$ as the set of the transmission power for SGDLs and ISLs, respectively. Thus, the action space is a finite set of all the possible actions. Furthermore, since the available energy is limited, the feasible action set in time slot i denoted by A_{S_i} is represented as:

$$A_{S_i} = \begin{cases} \{0, 0\} & B_i \le B_{th}, \forall D_i \\ \{\times, 0\} & B_i > B_{th} \& D_i = 0, \\ \{\times, \times\} & B_i > B_{th} \& D_i > 0. \end{cases} \tag{11}$$

Here, B_{th} is a key battery level threshold and $B_{th} = B_{\min} + P_n \cdot \tau$, which means that sufficient energy plays an important role in IoRT data collection and transmission. It can be noted from (11) that when there is IoRT data to be transmitted in the data buffer at the beginning of the time slot, it is necessary to decide whether to perform data transmission operation and the transmission power level in this time slot. Notably, $\times$ in (11) denotes the resource scheduling decisions to be determined and it should satisfy the available resource limitations of battery and data buffer, which can be specifically represented as:

$$\begin{cases} \tau_{i,c} \cdot P_i + P_{da} \cdot \tau \cdot \varsigma_i \le \max(B_i - B_{th}, 0), \\ \tau_{i,c} \cdot C_i(H_i, P_i) \le D_i, \end{cases} \quad \forall A_i \in A_{S_i}. \tag{12}$$

Note that once the resource scheduling strategy is determined, combined with the stochastic network environment, the evolution relationship of the data buffer level state and the battery level state between two consecutive time slots can be obtained. Hereinafter, we first present the evolution

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

relationship of the data buffer level state of the LEO as follows:

$$D_{i+1} = \min\left\{D_{\max}, D_i + D_i^R - D_i^T\right\}. \tag{13}$$

Here, $D_i^R = \varsigma_i \cdot r \cdot \tau$ denotes the obtained IoRT data during time slot i, where r is the collection data rate. D_i^T refers to the practical amount of data that can be transmitted during slot i, which should be less than the amount of data to be transmitted stored in the LEO at the beginning of slot i (i.e., $D_i^T = C_i(H_i, P_i) \cdot \tau_{i,c} \leq D_i$).[3] Furthermore, to avoid data buffer overflow, D_{i+1} should be less than the data buffer capacity denoted by $D_{\max}$, which is shown in (13). Similarly, the evolution relationship of the battery level state of the LEO can be represented as:

$$B_{i+1} = \min\left\{B_{\max}, B_i + E_i^h \cdot \tau_{i,e} - E_i^c\right\}. \tag{14}$$

3) Reward function: The reward function plays a pivotal role in providing feedback to a learning model about the performance of the previous actions. Besides, it can provide a correct guidance for the learning process, which can further accordingly help to take the best resource scheduling action in a particular network state. Therefore, in order to ensure the effectiveness of the resource scheduling decision-making system, we define the reward as the amount of successfully downloaded IoRT data. Specifically, given the network state S_i and the resource scheduling action A_i in slot i, the reward function at this time slot can be represented as $R_i(S_i, A_i) = D_i^T$. Notably, $R_i(S_i, A_i)$ depends on S_i and A_i.

Until now, we model the multi-dimensional resource scheduling problem during the IoRT data scheduling process as a MDP to find the joint power allocation and the transponder scheduling policies aiming at maximizing the amount of successfully downloaded IoRT data during an episode. We further define Π as a resource scheduling policy (including power and transponder scheduling) which is a mapping from a given S_i to the A_i that should be selected, i.e., $\Pi : S_i \mapsto A_i$. However, the future data buffer levels, the specific amount of energy that will be harvested, and the future satellite downlink channels are not known in advance. Consequently, the trade-off between utilizing and saving the energy stored in its on-board battery in the current time slot should be considered carefully by the LEO to avoid battery and data buffer overflows or to save the energy for the forthcoming time slots which might or might not have better channel conditions.

Therefore, considering this uncertainty, our goal is to learn an optimal resource scheduling policy to maximize the cumulative expected downloaded IoRT data rewards starting from the initial network state. We further define an action-value function denoted by $\Theta^{\Pi}(S_i, A_i)$ which represents the cumulative expected downloaded IoRT data reward in the future starting from state S_i by selecting action A_i and following Π thereafter. Specifically, $\Theta^{\Pi}(S_i, A_i)$ can be represented as

$$\begin{aligned}\Theta^{\Pi}(S_i, A_i) &= \mathbb{E}\left[R_i + \Upsilon R_{i+1} + \Upsilon^2 R_{i+2} + \cdots | S = S_i A = A_i, \Pi\right] \\ &= \mathbb{E}\left[\sum_{m=0}^{\infty} \Upsilon^m \cdot R_{i+m} | S = S_i, A = A_i, \Pi\right]. \end{aligned} \tag{15}$$

[3]The on-board transmitters at the LEO are assumed to be causal, which means that the IoRT data obtained in the current time slot can only be stored for transmission in the forthcoming time slots.

Here, $\mathbb{E}[\cdot]$ denotes the expectation function. Besides, we introduce a discount factor Υ with $0 \leq \Upsilon \leq 1$ to account for the preference of higher amount of successfully downloaded IoRT data values in the current time slot, which means that the nearest downloaded IoRT data rewards are worthier than the rewards in the further future. Notably, the optimal policy Π^* is the resource scheduling policy whose action-value function is greater than or equal to any other resource scheduling policy for every state. We further define Θ^* as the corresponding action-value function for the optimal policy Π^*. Consequently, $\Theta^{\Pi^*}(S_i, A_i)$ can be expressed as

$$\Theta^{\Pi^*}(S_i, A_i) = \mathbb{E}_{S'}\left[R_i + \Upsilon \max_{S'} \Theta^{\Pi^*}(S', A') | S_i, A_i\right]. \tag{16}$$

Note that when Θ^* is known, determining Π^* is straightforward. Unfortunately, the action-value function Θ^{Π} is unknown due to the fact that only causal knowledge is available at the LEO. Therefore, we propose a SACRL scheme in the next section to build an estimate of the action-value function from the states that are visited and the earned rewards. During each time slot i, the LEO selects a resource scheduling action A_i according to its current network state S_i. Then, a network reward R_i is generated according to the selected resource scheduling action A_i. Subsequently, the network is in state S_{i+1} and a new resource scheduling action A_{i+1} is selected based on this new state S_{i+1}. Θ^{Π} is updated according to considering S_i, A_i, R_i, S_{i+1}, A_{i+1}.

V. Reinforcement Learning Framework for Resource Allocation

In the resource scheduling decision-making system in SIoRTNs, the number of states is infinite, which can inevitably result in the fact that solving (16) directly is challenging. Therefore, we first explore the network resource characteristics and further design a linear function approximation scheme based on the resource characteristics to approximate the action-value function, i.e., $\Theta^{\Pi}(S_i, A_i)$. Then, we propose an actor-critic resource scheduling decision-making architecture and further design a SACRL algorithm to allow the LEO to quickly obtain the resource scheduling strategy to adapt to dynamic environmental changes and maximize the amount of long-term successfully downloaded IoRT data performance.

A. Linear Function Approximation

To effectively approximate the action-value function, we first define several key feature functions to provide an efficient model of the effects of possible actions on the on-board transponder at the LEO and possible transmit power values on the state of the LEO's transmitter. Specifically, by considering the limited buffer capacity and battery capacity, the data causality constraint in buffer as well as energy causality constraint in battery, and the unknown energy harvesting process during the Sun phase, we propose several feature functions to approximate the action-value function by utilizing a linear combination of the proposed feature functions which can map the state-action pair into a feature value.

We denote $f_z(S_i, A_i)$ as the z-th ($z = 1, \cdots, Z$) feature function for state-action pair (S_i, A_i), which means that

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

$f_z(S_i, A_i)$ can map (S_i, A_i) into a specific feature value. Specifically, we utilize normalized value (i.e., $[0, 1]$) to indicate the quality of the resource scheduling action A_i performed during time slot i according to the current network state S_i. The specific definitions of several feature functions are shown as follows.

1) Considering the limitation of the on-board battery capacity, the first feature function is dedicated to judging whether the energy consumed by performing a feasible resource scheduling action A_i can eliminate the battery overflow caused by the absorption of solar energy in Sun phase. Thus, the feature function denoted by $f_1(S_i, A_i)$ can be represented as

$$f_1(S_i, A_i) = \begin{cases} 1, & if\ B_i + E_i^h \cdot \tau_{i,e} - E_i^c \le B_{\max}, \\ 0, & else. \end{cases} \tag{17}$$

Here, E_i^c can be calculated according to (7). Note that $f_1(S_i, A_i)$ will be equal to 1 if no overflow is caused by the use of resource scheduling strategy A_i and the feasibility action condition shown in (11) and (12) is fulfilled.

2) The second feature function indicates whether the resource scheduling action considers the capacity limitation of the on-board data buffer. Specifically, whether the amount of data transmitted during time slot i can eliminate the potential data buffer overflow caused by the stored data. Therefore, except for the special case where the on-board data buffer size is extremely small, the second feature function denoted by $f_2(S_i, A_i)$ can be expressed as

$$f_2(S_i, A_i) = \begin{cases} 1, if\ r \cdot \tau < D_{\max} \wedge D_i + D_i^R - D_i^T \le D_{\max}, \\ 0, else. \end{cases} \tag{18}$$

Similar to (17), we design the feature function $f_2(S_i, A_i)$ to indicate whether data buffer overflow situations can be avoided by selecting a given resource scheduling action A_i. Note that $f_2(S_i, A_i) = 1$ if the data buffer overflow can be avoided under certain circumstances, otherwise, $f_2(S_i, A_i) = 0$.

3) To address the power allocation problem of the LEO by using SGDLs or ISLs for downloading IoRT data, we propose the third feature function, denoted by $f_3(S_i, A_i)$. We adopt the idea of water-filling to allocate power between the current channel and the estimated channel by using the past channel realizations to estimate the mean value $\bar{H}_i$ of the channel gain, which can be expressed as $\bar{H}_i = \frac{1}{m}\sum_{m=1}^{i} H_m$. Due to the orbital movement of the LEO, SGDLs and relay ISLs may be utilized for downloading the IoRT data. Consequently, there are four types of link switching between the SGDLs and relay ISLs. We first focus on the switch between two consecutive SGDLs. By introducing a Lagrangian multiplier and applying the Karush–Kuhn–Tucker (KKT) conditions, we can derive the allocated transmission power in slot i denoted by P_i^{dd} as

$$P_i^{dd} = \max\left(\frac{1}{\tau_{i,c} + \tau_{i+1,c}}\left(B_{[i,i+1]} + \frac{\tau_{i+1,c}}{\bar{H}_i} + \frac{\tau_{i,c}}{H_i}\right) - \frac{1}{H_i}, 0\right). \tag{19}$$

Here, $B_{[i,i+1]}$ denotes the maximum available energy of the battery during time slots i to $(i+1)$, which can be represented as in (20), shown at the bottom of the page. We further define P_i^{dr}, P_i^{rr}, and P_i^{rd} as the calculated power values by using the idea of water-filling for switchings from an SGDL to a relay ISL, between two consecutive relay ISLs, and from a relay ISL to a SGDL, respectively. The detailed derivations of P_i^{dr}, P_i^{rr}, and P_i^{rd} will be described below.

Since power P_i should be selected from a discrete set, the optimal power allocation value in time slot i denoted by P_i^{opt} is rounded so that $P_i^{opt} \in A_{S_i}$ holds. In addition, due to the available energy limitation, P_i^{opt} is represented as follows:

$$P_i^{opt} = \min\left\{P_i^{\max}, \left\lfloor P_i^{swt}/\delta \right\rfloor \cdot \delta\right\}. \tag{21}$$

Note that P_i^{swt} denotes the calculated power value by using the idea of water-filling for four switch cases, which means that P_i^{swt} is equal to P_i^{dd} or P_i^{dr} or P_i^{rr} or P_i^{rd} for the mentioned above four link switching situations, respectively. Besides, P_i^{opt} equals P_{idd}^{opt} or P_{idr}^{opt} or P_{irr}^{opt} or P_{ird}^{opt} for four switch cases, respectively. Here, $\lfloor a \rfloor$ refers to the rounding operation to the nearest integer less than or equal to a. Besides, δ equals the δ_{tsd} and δ_{tsr} for SGDLs and relay ISLs, respectively. Different actions $\left(\varsigma_i, P_i^{opt}\right)$ lead to different amounts of downloaded IoRT data during two consecutive slots, which are denoted by $D_{[i,i+1]}^{Tdd}$, $D_{[i,i+1]}^{Tdr}$, $D_{[i,i+1]}^{Trr}$, and $D_{[i,i+1]}^{Trd}$ for the mentioned above four switch cases, respectively. The specific expression of $D_{[i,i+1]}^{Tdd}$ is shown in (22), at the bottom of the page. $P_{i+1}^{\max}$ in (22) refers to the maximum feasible power in time slot $(i+1)$ when action during time slot i is selected as $A_i = \left(\varsigma_i, P_i^{opt}\right)$.

Similarly, when the link switching occurs from a SGDL to a relay ISL, the allocated transmission power P_i^{dr} can be derived using the KKT conditions as follows:

$$P_i^{dr} = \max\left(0, \frac{B_c}{\bar{H}_i \cdot \ln 2} - \frac{1}{H_i}\right). \tag{23}$$

Following the restriction of (1), $D_{[i,i+1]}^{Tdr}$ can be derived directly and calculated by

$$D_{[i,i+1]}^{Tdr} = \tau_{i,c} \cdot C_i\left(H_i, P_{idr}^{opt}\right) + \tau_{i+1,c} \cdot C_{i+1}\left(\bar{H}_i, P_{i+1}^{\max}\right), \forall \varsigma_i \in A_{S_i}. \tag{24}$$

$$B_{[i,i+1]} = \max(0, B_i - B_{th} - P_{da} \cdot \tau \cdot \varsigma_i) + \max(0, \min(B_i - P_{da} \cdot \tau \cdot \varsigma_i - P_n \cdot \tau - \max(0, B_i - B_{th} - P_{da} \cdot \tau \cdot \varsigma_i) + E_i^h \cdot \tau_{i,e}, B_{\max}) - B_{th}). \tag{20}$$

$$D_{[i,i+1]}^{Tdd} = \tau_{i,c} \cdot C_i\left(H_i, P_i^{opt}\right) + \tau_{i+1,c} \cdot C_{i+1}\left(\bar{H}_i, \min\left(P_{i+1}^{\max}, \left\lfloor P_i^{dd}/\delta \right\rfloor \cdot \delta\right)\right)\ \forall \varsigma_i \in A_{S_i}. \tag{22}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

Besides, for the case of switch between two consecutive relay ISLs, the transmission power of the LEO in slot i P_i^{rr} can be expressed as

$$P_i^{rr}=\begin{cases}\max\left(0,\min\left(\dfrac{B_i'}{\tau_{i,c}},\dfrac{D_i'}{\tau_{i,c}\cdot H_i}\right)\right) & if\ H_i<\bar{H}_i,\\ P_i^{\max} & else.\end{cases} \tag{25}$$

Here, B_i' and D_i' represent the available energy resource for IoRT data transmission and the amount of IoRT data to be transmitted during time slot i, respectively, when $H_i<\bar{H}_i$ and energy resource are preferentially used in slot $(i+1)$ for IoRT data transmission. It means that the power allocation and IoRT data scheduling strategies are first obtained and then B_i' and D_i' can be calculated by an inverse operation. Therefore, the specific B_i' and D_i' can be expressed as

$$B_i'=\begin{cases}B_{[i,i+1]}-P_{i+1}^{\max 0}\cdot\tau_{i+1,c} & if B_{[i,i+1]}+B_{th}<B_{\max},\\ \min\left(B_i,B_{[i,i+1]}+B_{th}-B_{\max}\right) & else.\end{cases} \tag{26}$$

$$D_i'=\begin{cases}D_i+D_i^R-C_{i+1}\left(\bar{H}_i,P_{i+1}^{\max 0}\right)\cdot\tau_{i+1,c} & if\ D_i+D_i^R<D_{\max},\\ \min\left(D_i,D_i+D_i^R-D_{\max}\right) & else.\end{cases} \tag{27}$$

Here, $P_{i+1}^{\max 0}$ denotes the maximum transmission power during slot $(i+1)$ when transmission power during time slot i is zero. Furthermore, by substituting P_i^{rr} to (21), P_{idd}^{opt} can be obtained and $D_{[i,i+1]}^{Trr}$ can be further derived as shown in (28), at the bottom of the page.

In addition, for the case of switching from a relay ISL to a SGDL, the allocated transmission power of the LEO in slot $(i+1)$ denoted by $P_{i+1,rd}^{opt}$, can be calculated by

$$P_{i+1,rd}^{opt}=\max\left(0,\min\left(\left\lfloor\left(\frac{B_c}{H_i\cdot\ln 2}-\frac{1}{\bar{H}_i}\right)/\delta\right\rfloor\cdot\delta,P_{i+1}^{\max 0}\right)\right). \tag{29}$$

Similar to the case of switch between two consecutive relay ISLs, we can replace $P_{i+1}^{\max 0}$ with $P_{i+1,rd}^{opt}$ in (26) and (27) to obtain P_i^{rd}, which can be calculated by:

$$P_i^{rd}=\max\left(0,\min\left(\frac{B_i'}{\tau_{i,c}},\frac{D_i'}{\tau_{i,c}\cdot H_i}\right)\right). \tag{30}$$

By exploiting (21) to obtain the feasible value of P_i^{rd}, i.e., P_{ird}^{opt}, $D_{[i,i+1]}^{Trd}$ can be further given as

$$D_{[i,i+1]}^{Trd}=C_i\left(H_i,P_{ird}^{opt}\right)\cdot\tau_{i,c}+C_{i+1}\left(\bar{H}_i,P_{i+1}^{opt}\right)\cdot\tau_{i+1,c},\forall\varsigma_i\in A_{S_i}. \tag{31}$$

Our goal is to find the optimal power allocation policy to maximize $D_{[i,i+1]}^T$ in this feature, where $D_{[i,i+1]}^T$ is equal to $D_{[i,i+1]}^{Tdd}$ or $D_{[i,i+1]}^{Tdr}$ or $D_{[i,i+1]}^{Trr}$ or $D_{[i,i+1]}^{Trd}$ for the corresponding switch case, respectively. Consequently, the feature function $f_3(S_i,A_i)$ can be given as

$$f_3(S_i,A_i)=\begin{cases}1, & if\ \{\varsigma_i,P_i\}=\arg\max\limits_{\{\varsigma_i,P_i\}} D_{[i,i+1]}^T,\\ 0, & else.\end{cases} \tag{32}$$

4) To avoid wasting energy, the LEO tends to allocate energy as much as possible when the solar energy acquired in current time slot i is abundant. Therefore, the feature function $f_4(S_i,A_i)$ can be represented as

$$f_4(S_i,A_i)=\begin{cases}1, & if\ E_i^h\cdot\tau_{i,e}>B_{\max}-B_i+E_i^{c\max}\wedge\\ & \quad E_i^c=E_i^{c\max},\\ 0, & else.\end{cases} \tag{33}$$

Here, $\wedge$ refers to the logical conjunction operation. $E_i^{c\max}$ denotes the maximum amount of energy that the LEO can consume, which is constrained by the available energy in slot i and the amount of the IoRT data to be transmitted and is represented as

$$E_i^{c\max}=\max\left\{P_i\cdot\tau_{i,c}+P_{da}\cdot\tau\cdot\varsigma_i+P_n\cdot\tau,\forall(\varsigma_i,P_i)\in A_{S_i}\right\}. \tag{34}$$

5) If the data buffer overflows, partial energy will be wasted. However, when the solar energy acquired in this time slot is abundant, this kind of waste can be ignored. Therefore, the fifth feature function denoted as $f_5(S_i,A_i)$ is an exceptional case of the second feature function and can be represented as

$$f_5(S_i,A_i)=\begin{cases}1,\ if\ B_i+E_i^h\cdot\tau_{i,e}-E_i^c>B_{\max}\wedge\\ \quad r\cdot\tau<D_{\max}\wedge D_i+D_i^R-D_i^T>D_{\max},\\ 0,\ else.\end{cases} \tag{35}$$

6) To represent the effectiveness of collecting IoRT data, we further define the sixth feature function $f_6(S_i,A_i)$ as follows

$$f_6(S_i,A_i)=\frac{\min\left(D_{\max}-D_i+D_i^T,D_i^R\right)}{D_{\max}}. \tag{36}$$

On the basis of the above feature functions, the action-value function can be approximated as follows:

$$\Theta^{\Pi}(S_i,A_i)\approx\hat{\Theta}^{\Pi}(S_i,A_i,\boldsymbol{\varpi}(A_i))=\mathbf{f}^T(S_i,A_i)\cdot\boldsymbol{\varpi}(A_i). \tag{37}$$

Here, $\mathbf{f}(S_i,A_i)=[f_1(S_i,A_i);\cdots;f_Z(S_i,A_i)]\in Z\times 1$ denotes the feature vector which can be obtained with a given state action pair (S_i,A_i) according to (17)-(36) and $\mathbf{f}^T$ represents the transpose of the vector $\mathbf{f}$. $\boldsymbol{\varpi}(A_i)\in Z\times 1$

$$D_{[i,i+1]}^{Trr}=\begin{cases}C_i\left(H_i,P_{irr}^{opt}\right)\cdot\tau_{i,c}+C_{i+1}\left(\bar{H}_i,P_{i+1}^{\max 0}\right)\cdot\tau_{i+1,c} & if\ H_i<\bar{H}_i\\ C_i\left(H_i,P_{irr}^{opt}\right)\cdot\tau_{i,c}+C_{i+1}\left(\bar{H}_i,P_{i+1}^{\max}\right)\cdot\tau_{i+1,c} & else\end{cases}\quad\forall\varsigma_i\in A_{S_i}. \tag{28}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

is a weight vector which indicates the contribution of each feature. Notably, there is a corresponding weight vector for each resource scheduling action. Besides, we define $\varpi = [\varpi\left(A^1\right), \cdots, \varpi\left(A^{|\mathcal{A}|}\right)] \in Z\times \mid A \mid$ as the weight matrix for all potential resource scheduling actions, where $\mid A \mid$ is the number of potential resource scheduling actions and A^1 refers to the first resource scheduling action scheme. Since these weight vectors can control the contribution of each feature function on $\hat{\Theta}^{\Pi}(S_i, A_i, \varpi(A_i))$, our goal is to learn the optimal weight vector for each action to further effectively approximate $\Theta^{\Pi}(S_i, A_i)$. Once the optimal ϖ is obtained, in practical applications, we can select the resource scheduling action which can maximize $\hat{\Theta}^{\Pi}(S_i, A_i, \varpi(A_i))$ for a given network state S_i as the current resource scheduling strategy.

B. Actor-Critic Decision-Making Architecture and Algorithm Design

In this sub-section, we first design an actor-critic resource scheduling decision-making architecture and further propose a SACRL algorithm to effectively update the weight matrix ϖ for all potential resource scheduling actions and learn the optimal resource scheduling strategy.

An actor-critic resource scheduling decision-making architecture is shown in Fig. 2. Specifically, the proposed actor-critic resource scheduling decision-making architecture consists of two important components, i.e., an actor network and a critic network. At each resource scheduling decision in time slot i, the feature vectors for all feasible actions can be obtained through feature extraction operation in Fig. 2 according to (17)-(36). Then, the actor network takes system state S_i as input and produces a resource scheduling action used in exploration or exploitation which can be denoted as A_i. Afterwards, the LEO adopts the resource scheduling scheme A_i to obtain the downloaded IoRT data reward R_i and the corresponding feature vector $\mathbf{f}_i$, and then moves to the next network state S_{i+1}. According to the current values of action-value function and the policy Π, the resource scheduling in the next time slot A_{i+1} can be selected and the corresponding $\mathbf{f}_{i+1}$ can be further obtained. Then, we store $(\mathbf{f}_i,\ A_i,\ R_i,\ \mathbf{f}_{i+1},\ A_{i+1})$ in the replay memory. The actor network utilizes the experience replay during training and determines the loss function based on the proposed SACRL algorithm. Specifically, we randomly select a small batch of transfer samples from the replay memory buffer to update the parameters (i.e., the weight matrix in the actor network denoted by ϖ_{actor}) in the actor network utilizing the stochastic gradient descent (SGD) algorithm and the critic network towards maximizing the long-term data downloaded IoRT data rewards, i.e., $\hat{\Theta}^{\Pi}(S_i, A_i, \varpi_{\text{critic}})$, where ϖ_{critic} is the weight matrix in the critic network.

The approximated action-value function from the actor network with a given state action pair (S_i, A_i) in time slot i, i.e., $\hat{\Theta}_i^{\Pi}(S_i, A_i, \varpi_{\text{actor}})$, can be represented as:

$$\hat{\Theta}_i^{\Pi}(S_i, A_i, \varpi_{\text{actor}}) = \mathbf{f}^T(S_i, A_i) \cdot \varpi_{\text{actor}}. \tag{38}$$

Similarly, in the next time slot $(i+1)$, we denote the approximated action-value function from the critic network as $\hat{\Theta}_i'^{\Pi}(S_{i+1}, A_{i+1}, \varpi_{\text{critic}})$ under the given network state S_{i+1} and selected resource scheduling action A_{i+1}, which can be expressed as

$$\hat{\Theta}_i'^{\Pi}(S_{i+1}, A_{i+1}, \varpi_{\text{critic}}) = \left(\mathbf{f}'(S_{i+1}, A_{i+1})\right)^T \cdot \varpi_{\text{critic}}. \tag{39}$$

Note that the selected A_i and A_{i+1} are obtained by using the exploration or exploitation strategy, i.e., the ϵ-greedy resource scheduling strategy which means that the LEO acts greedily with respect to its corresponding action-value function with a probability of $(1-\epsilon)$. In other words, the LEO will randomly select a resource scheduling strategy from the feasible action set A_{S_i}. Specifically, following the recent work [30], the ϵ-greedy resource scheduling strategy can be represented as:

$$\Pr\left(A_i = \arg\max_{A_i \in A_{S_i}} \hat{\Theta}^{\Pi}(S_i, A_i, \varpi_{\text{actor}})\right) = 1-\epsilon_i,\ 0<\epsilon_i \leq 1. \tag{40}$$

It can be noted from [31] and [32] that this resource scheduling scheme provides a trade-off between the exploration of new resource scheduling strategies and the exploitation of the known ones.

Suppose that the actor will fetch several transfer samples at once from replay memory for training. The number of samples fetched at one time is called the batch size, which is denoted as $\mathcal{M}_{batch}$. Therefore, during training phase, following the recent work [33] the loss function based on mean squared error can be expressed in (41), shown at the bottom of the next page, Here, (S_n, A_n, R_n) and (S_{n+1}, A_{n+1}) are the current state, action, reward and the next state and action of the n-th memory sampled from the experience replay memory, respectively. Besides, $\left(R_n + \Upsilon \cdot \hat{\Theta}_n'^{\Pi}(S_{n+1}, A_{n+1}, \varpi_{\text{critic}})\right)$ is the target action value. We adopt a separate network, i.e., the critic network to generate the target action value so that the action value of the network will shift at every training step. Using only one network may incur uncontrolled estimation of the action value. In this case, the network will become more and more unstable based on the feedback loop structure of reinforcement learning [34]. Notably, the actor depends on the experience the critic network learns shown in (41).

By exploiting the SGD for ϖ_{actor} in (41), we can update ϖ_{actor} according to [35] as follows:

$$\varpi_{\text{actor}} = \varpi_{\text{actor}} - \Delta\varpi_{\text{actor}}. \tag{42}$$

Here $\Delta\varpi_{\text{actor}}$ can be shown in (43), at the bottom of the next page. Note that $\nabla_{\varpi_{\text{actor}}}\hat{\Theta}_n^{\Pi}(S_n, A_n, \varpi_{\text{actor}})$ in (43) can be derived according to (38) as follows:

$$\nabla_{\varpi_{\text{actor}}}\hat{\Theta}_n^{\Pi}(S_n, A_n, \varpi_{\text{actor}}) = \mathbf{f}^T(S_n, A_n). \tag{44}$$

Until now, the update of the weight matrix ϖ_{actor} in the actor network is completed. Then, we can update the critic network by copying the weight matrix ϖ_{actor} in the actor network to the critic network after a certain number of iterations, i.e., $\varpi_{\text{critic}} = \varpi_{\text{actor}}$. Notably, the update of the actor network is generally more frequent than the update of the critic network.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

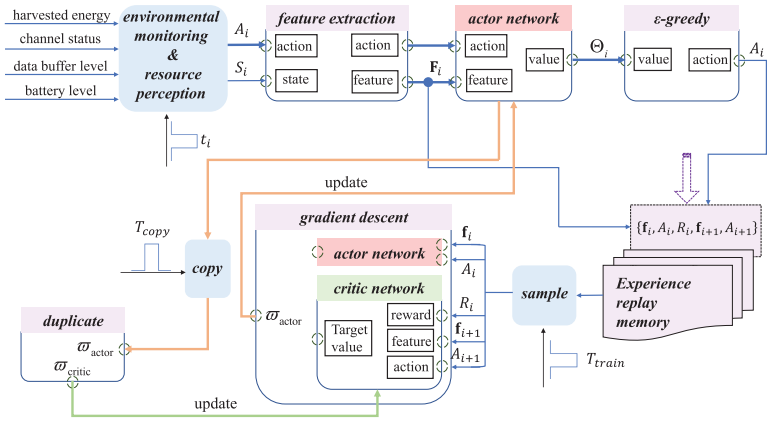


Fig. 2. An actor-critic resource scheduling decision-making architecture.

Regarding the convergence properties of the proposed SACRL algorithm, the ϵ parameter used in the ϵ-greedy policy and the learning rate β during time slot i are denoted by ϵ_i and β_i, which should be non-ascending in each time slot. Consequently, we set

$$\epsilon_i = \varrho_1^{\lfloor i/\Gamma \rfloor} \quad \text{and} \quad \beta_i = \frac{\varrho_2}{\lfloor i/\Gamma \rfloor + 1}. \tag{45}$$

Here, $\varrho_1\,(0 < \varrho_1 < 1)$ and $\varrho_2\,(0 < \varrho_2 < 1)$ are two reference values and $\Gamma = \mathcal{T}/\tau$ refers to the number of time slots during an episode. It can be demonstrated according to [36] that the mentioned above definition of β_i can guarantee the convergence of the proposed approximate SACRL algorithm. Specifically, the proposed algorithm will converge to a bounded region with probability one, which means that it does not diverge.[4]

Remark 1: Since the channel model and feasible transmission power action for SGDLs and ISLs are different, two independent SACRL networks are utilized to learn the optimal $\boldsymbol{\varpi}_{\text{actor}}$ and $\boldsymbol{\varpi}_{\text{critic}}$ for SGDLs and ISLs, respectively. Notably, the update processes of two weight matrices (i.e., $\boldsymbol{\varpi}_{\text{actor}}$ and $\boldsymbol{\varpi}_{\text{critic}}$) for SGDLs and ISLs are actually coupled. Due to the orbital movements of the LEO, there exists a switch between the SGDLs and the ISLs for data backhaul. For example, the resource scheduling action decision during satellite downlink phase will have an impact on the network state and further affect the feature vector corresponding to this network state in the forthcoming ISL phase, which ultimately has an impact on the respective action-value and weight matrices (i.e., $\boldsymbol{\varpi}_{\text{actor}}$ and $\boldsymbol{\varpi}_{\text{critic}}$) for ISLs, vice versa.

On the basis of the above description, we can summarize the proposed SACRL algorithm as shown in Algorithm 1. Firstly, we initialize some trainable parameters (line 3), such as the weight matrices $\boldsymbol{\varpi}_{\text{actor}}$ and $\boldsymbol{\varpi}_{\text{critic}}$ in the actor network and critic network, batch size $\mathcal{M}_{batch}$, discount factor Υ, learning rate β, etc. Next, the initial network state including the resource status of the LEO and the channel status can be obtained by the on-board sensor system (line 4). Afterwards, through continuous iterations (lines 5-24), the weight matrices $\boldsymbol{\varpi}_{\text{critic}}$ and $\boldsymbol{\varpi}_{\text{actor}}$ can be updated (lines 14-16 and lines 17-21, respectively) by using the ϵ-greedy strategy to select a feasible resource scheduling action in the actor network (line 5 and lines 11-12), storing the samples into the experience replay memory (line 13), and computing the loss function and its gradient of the actor network (line 19-20). Finally, the resource scheduling policy can be obtained by learning the optimal $\boldsymbol{\varpi}_{\text{critic}}$ which can be utilized to select the optimal resource scheduling and data scheduling action with given current network state.

In the practical applications, the LEO first observes the current state and calculates all the feasible actions according to (11) and (12). Then, the LEO computes the feature vectors based on the current state and feasible actions. Finally, the LEO can select the optimal action from feasible resource scheduling actions that maximizes the approximated action-value function according to (37) for the current time slot. Consequently, the optimal feasible resource scheduling action can be obtained immediately in each time slot according

[4]Notably, proving the convergence of the proposed algorithm plays a significant role in the learning phase, which means that a dynamic resource scheduling strategy can be trained and obtained by the proposed algorithm. In practical applications, the trained resource scheduling strategy can adapt to the dynamic environment in SIoRTNs.

$$\mathcal{L}\left(\boldsymbol{\varpi}_{\text{actor}}\right) = \frac{1}{2\mathcal{M}_{batch}} \sum_{n=1}^{\mathcal{M}_{batch}} \left(R_n + \Upsilon \cdot \hat{\Theta}_n'^{\Pi}(S_{n+1}, A_{n+1}, \boldsymbol{\varpi}_{\text{critic}}) - \hat{\Theta}_n^{\Pi}(S_n, A_n, \boldsymbol{\varpi}_{\text{actor}})\right)^2. \tag{41}$$

$$\Delta\boldsymbol{\varpi}_{\text{actor}} = -\frac{\beta_i}{\mathcal{M}_{batch}} \sum_{n=1}^{\mathcal{M}_{batch}} \left[\left(R_n + \Upsilon \cdot \hat{\Theta}_n'^{\Pi}(S_{n+1}, A_{n+1}, \boldsymbol{\varpi}_{\text{critic}}) - \hat{\Theta}_n^{\Pi}(S_n, A_n, \boldsymbol{\varpi}_{\text{actor}})\right) \times \nabla_{\boldsymbol{\varpi}_{\text{actor}}} \hat{\Theta}_n^{\Pi}(S_n, A_n, \boldsymbol{\varpi}_{\text{actor}})\right]. \tag{43}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 SACRL Algorithm

Input: Network topology obtained by STK, number of time slots Γ during an episode, and reward function R;
Output: ϖ_{critic} in the critic network to obtain the resource scheduling policy;
Initialization: Set $\varpi_{\text{actor}} = \varpi_{\text{critic}} = \varpi_0$, ε in ε-greedy strategy, initialize batch size $\mathcal{M}_{batch}$, discount factor Υ, learning rate β, total number of iterations $T_{learning}$, the update intervals for the actor network and critic network, i.e., T_{train} and T_{copy};
Observe the LEO resource status information from the on-board sensor system and the channel status, i.e., the network state S_i;
Utilize the ϵ-greedy strategy to select a resource scheduling action A_i in the feasible action set according to (11);
for Learning episode $t = 1, 2, \cdots T_{learning}$ **do**
  Obtain the initial resource status from the on-board sensor system;
  for Time slot during an episode $l = 1, 2, \cdots \Gamma$ **do**
    Time slot during training $i = (t-1) \cdot \Gamma + l$;
    Execute the action A_i, compute the corresponding reward $R_i(S_i, A_i)$ and f_i, and observe the next new state S_i';
    Update ϵ and β according to (45);
    Utilize the ϵ-greedy strategy to select a resource scheduling action A_i' for state S_i' in the feasible action set according to (11) and calculate f_i';
    Store $(f_i, A_i, R_i, f_i', A_i')$ into the experience replay memory;
    if $i \% T_{copy} = 0$ **then**
      Update ϖ_{critic} for the critic network by copying the weight matrix ϖ_{actor} in the actor network, i.e., $\varpi_{\text{critic}} = \varpi_{\text{actor}}$;
    end if
    if $i \% T_{train} = 0$ **then**
      Randomly select a batch size of $\mathcal{M}_{batch}$ transitions samples from experience replay memory as a mini-batch training data for the actor network, which can be expressed as $(f_n, A_n, R_n, f_{n+1}, A_{n+1})$;
      Calculate the loss function according to (41);
      Perform the SGD algorithm for the loss function in (41) to calculate the value gradient of the actor network and update ϖ_{actor} according to (42);
    end if
    $f_{i+1} = f_i'$, $A_{i+1} = A_i'$;
  end for
end for

to the current network state, which facilitates its practical applications for dynamic satellite systems.

VI. Simulation Results and Analysis

In this section, we conduct numerical simulations by utilizing the real-world satellite parameters supplied from the STK and further adopt the Matlab and Python simulators to evaluate the proposed SACRL algorithm. Firstly, we present the simulation scenario and parameter settings. Then, we verify the performance of the proposed SACRL algorithm from the aspects of convergence and effectiveness.

A. Simulation Setup

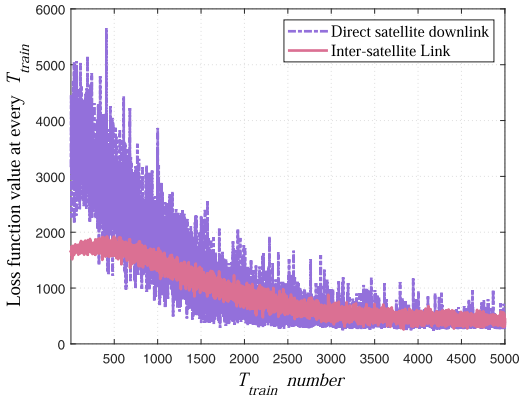
We consider a SIoRTN consisting of a LEO satellite with an inclination of 97.86° at a height of 554.8km, three GEO relay satellites which are distributed at nominal longitudes of 16.7°E, 77.0°E, and 176.7°E, and six ground stations which are located at Kashi (39.5°N, 76°E), Xi'an (34°N, 108.9°E), Beijing (40°N, 116°E), Qingdao (36°N, 120°E), Nanjing (32°N, 118.8°E), and Sanya (18°N, 109.5°E). Note that the time duration of each episode in the proposed SIoRTN scenario is about 24hours. We further adopt STK to obtain the Sun phases, potential contacts and geographical locations of satellites from 15 Feb. 2020 04:00:00 to 16 Feb. 2020 04:00:00 by STK. Besides, E_i^h is taken from a uniform distribution in the interval $\left[E^{h\min}, E^{h\max}\right]$, where $E^{h\min}$ and $E^{h\max}$ are the minimum and the maximum battery recharge rate and $E^{h\min} = 0.3 \cdot E^{h\max}$. The specific H_i for SGDLs is i.i.d. and can be obtained according to the CDF of the atmospheric attenuation at each ground station location in (3). In addition, the specific H_i for ISLs can be obtained according to (2). In the simulations, we set $\tau = 300$s, $P_{da} = 30$W, $P_n = 10$W, $P_{tsd}^{\max} = 80$W, $P_{tsr}^{\max} = 70$W, $\delta_{tsd} = \delta_{tsr} = 1$W, $r = 100$Mbps, $T_{train} = 2$ episodes and $T_{copy} = 3$ episodes, $\mathcal{M}_{batch} = 20$, $\Upsilon = 0.9$, $\chi = 0.4$, $\varrho_1 = 0.9994$, and $\varrho_2 = 0.00001$ [25].

B. Performance Evaluation

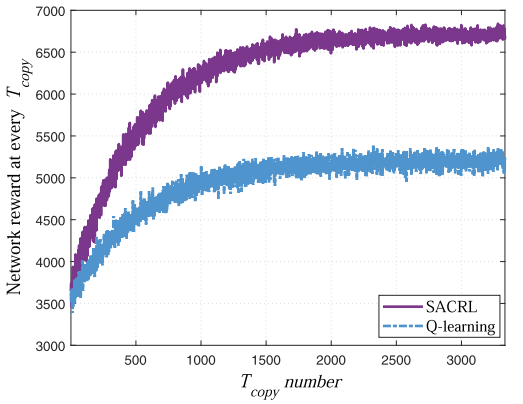
To evaluate the performance of the proposed SACRL algorithm from the performance of both the convergence and effectiveness of the algorithm, we mainly consider the following three baseline strategies (i.e., the Q-learning based resource scheduling strategy, the Myopia-greedy resource scheduling scheme, and the random resource scheduling strategy) and further evaluate their performance under the same conditions.

- Q-learning: The Q-learning algorithm is the off-policy temporal-difference RL approach and the network state, e.g., D_i and B_i, should be discretized in this method.
- Myopia-greedy: The Myopia-greedy algorithm selects the feasible resource scheduling action which can maximize the instantaneous IoRT data reward during the current time slot. In other words, future dynamic resource status including the dynamic channel and dynamic harvested energy during the Sun phase are completely ignored in the Myopia-greedy scheme.
- Random: This scheme randomly selects the resource scheduling action from the obtained feasible action set with the same probability. Therefore, it is a blind resource scheduling strategy for collecting and transmitting IoRT data without any consideration of resource status and the impact of the current decisions on the future.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.



(a) Loss function value at every T_{train} interval *v.s.* T_{train} number



(b) Network reward at every T_{copy} interval *v.s.* T_{copy} number

Fig. 3. Loss function value and network reward performance ($B_{\max} = 200$KJ, $D_{\max} = 100$Gbits, $E^{h\,\max} = 150$W). (a) The loss function value at every T_{train} interval during training process. (b) Network reward at every T_{copy} interval performance comparison for SACRL and Q-learning algorithms.

Convergence: To demonstrate the convergence of the proposed SACRL algorithm, we plot the loss function value at every T_{train} interval during training for direct SGDLs and ISLs in Fig. 3a, respectively. Observe from Fig. 3a that, with the guidance of SGD strategy during training process, loss function values for SGDLs and ISLs initially have large fluctuations but follow a downward trend as a whole and eventually become stabilized with small fluctuations, which indicates that the approximation error first decreases and then gradually stabilizes. Therefore, the proposed SACRL algorithm can effectively converge to a stable resource allocation and IoRT data scheduling strategy model during training.

To further investigate the convergence and the effectiveness of the proposed SACRL algorithm, we plot the network reward performance at every T_{copy} interval during training for the proposed SACRL and Q-learning algorithms in Fig. 3b. As shown in Fig. 3b, the network rewards for two learning algorithms gradually stabilize at their corresponding optimal values, which means that the optimal resource allocation strategy and IoRT data scheduling strategy have been learned to maximize long-term network rewards. Besides, it can be seen that the network reward performance of the proposed

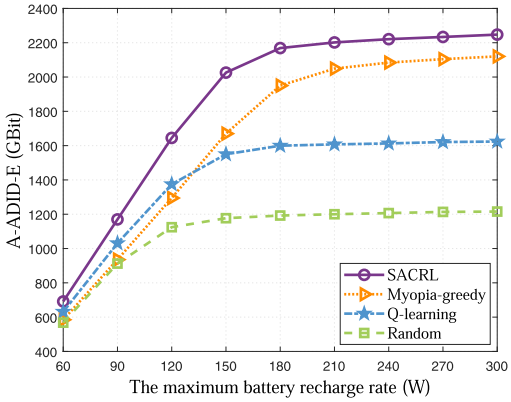


Fig. 4. A-ADID-E *v.s.* The maximum battery recharge rate ($B_{\max} = 200$KJ, $D_{\max} = 100$Gbits).

SACRL strategy is superior to the Q-learning algorithm. The underlying reason is that the discretization operation in network state in the Q-learning algorithm replace the real state value with an approximation from a finite number of discrete state values, which can impede this algorithm to find the true optimal resource allocation and IoRT data scheduling policies.

Effectiveness: To verify the effectiveness of the proposed SACRL strategy compared with other three baseline schemes during application process, we define the average amount of downloaded IoRT data during an episode period (A-ADID-E), i.e., the average episode reward as a performance metric. Besides, to avoid the randomness of the simulation results brought by the dynamic environment, e.g., channel and solar infeed during the Sun phase, the following simulation results are all obtained by averaging 100 times.

We investigate the effect of the maximum battery recharge rate $E^{h\,\max}$ on A-ADID-E performance for four strategies in Fig. 4. As expected, the A-ADID-E performance follows an upward trend for all strategies as $E^{h\,\max}$ increases. Notably, the increasing of $E^{h\,\max}$ can significantly facilitate the improvement in A-ADID-E performance for small $E_i^{h\,\max}$. The reason for this is that more abundant energy supply can be stored and used for IoRT data acquisition and transmission. However, when $E_i^{h\,\max}$ is sufficiently large, the A-ADID-E performance shows a slow growth and finally tends to be stable as $E_i^{h\,\max}$ increases. This behavior is explained by the fact that other network resources, e.g., the limited communication, buffer, and energy resources have become bottlenecks in this case. It can be further noted that, the proposed SACRL algorithm performs better than the other strategies. Specifically, the A-ADID-E in SACRL has increased by 5.96%, 38.78% and 85.44% compared with that in Myopia-greedy, Q-learning and Random algorithms at their maximum A-ADID-E performance values, respectively, which is attributed to jointly considering the time-variant channel fading and stochastically fluctuating solar infeed processes as well as the impact of the resource scheduling and data scheduling policies in current slot on the future.

Subsequently, we study the impact of battery capacity $B_{\max}$ on A-ADID-E performance for four approaches in Fig. 5.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

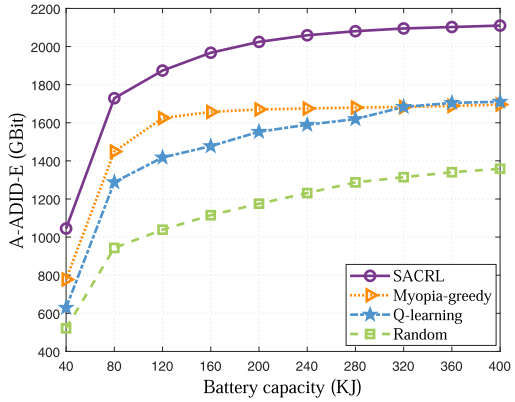


Fig. 5. A-ADID-E *v.s.* Battery capacity ($E^{h\max} = 150$W, $D_{\max} = 100$Gbits).

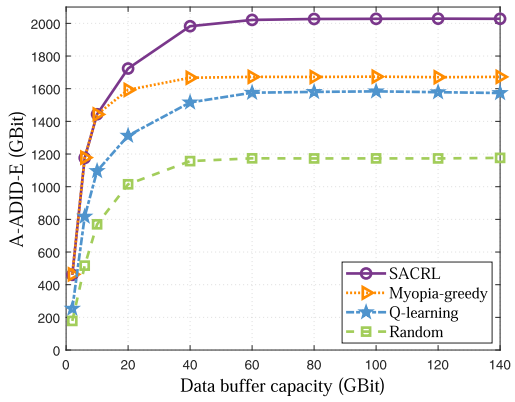


Fig. 6. A-ADID-E *v.s.* Data buffer capacity ($B_{\max} = 200$KJ, $E^{h\max} = 150$W).

As expected, increasing battery capacity can improve the A-ADID-E performance for all algorithms. The reason lies in that under this circumstance, more harvested energy can be stored for future utilization. Once $B_{\max}$ increases to a certain value, the A-ADID-E performance tends to bottleneck for four strategies as $B_{\max}$ further increases due to the limited network resources, e.g., the harvested energy resource, communication resource, and data buffer resource. Additionally, due to the neglection of the impact of the current resource allocation and IoRT data scheduling strategies on the future in Myopia-greedy and Random schemes as well as the discretization operation in network state in the Q-learning algorithm, the proposed SACRL algorithm outperforms to other compared strategies. Notably, compared with the best-performing Q-learning algorithm among the three comparison schemes, the A-ADID-E performance value in SACRL algorithm has increased by 23.50% at their corresponding maximum A-ADID-E values.

We further investigate the impact of the data buffer capacity $D_{\max}$ on A-ADID-E performance for four strategies in Fig. 6. It can be seen from Fig. 6 that the A-ADID-E performance shows a non-declining trend with the increase of $D_{\max}$ for four strategies. This trend is reasonable because of the fact that larger $D_{\max}$ can store more IoRT data to increase its potential possibility of successful transmission. Furthermore, the proposed SACRL algorithm and Myopia-greedy algorithm have similar A-ADID-E performance when $D_{\max}$ is in an extremely low value. This behavior is explained by the fact that under this condition, all the stored data has to be transmitted using the maximum power to avoid data buffer overflow and make efficient utilization of the energy resource because in this instance, the energy supply is relatively sufficient since the small $D_{\max}$ constrains the stored IoRT data to be transmitted and further decreases the demand for energy consumption. Another observation is that, when $D_{\max}$ falls in a certain area with relatively large value, increasing $D_{\max}$ will not bring significant gains to A-ADID-E performance due to the bottlenecks of other network resources, e.g., the battery resource, communication resource, and the harvested energy resource in the Sun phase. Notably, since the proposed approach SACRL exploits the casual information and designs an efficient approximation action-value function to avoid discretization in network state, its A-ADID-E performance is the best among all comparison algorithms. Specifically, the 21.44%, 28.52%, and 73.01% increasements in A-ADID-E are achieved by the proposed SACRL scheme in comparison with other three strategies, i.e., the Myopia-greedy, the Q-learning and the Random strategies, at their corresponding maximum A-ADID-E values, respectively.

VII. Conclusion

In this paper, we formulated the RAIDS problem by using model-free RL framework with joint consideration of the time-variant channel fading and stochastically fluctuating solar infeed processes. Considering the fact that the network state space is continuous and infinite, we designed several feature functions corresponding to the natural attributes of the multi-dimensional resources to approximate action-value function by exploiting a linear function. Furthermore, an actor-critic resource scheduling decision-making architecture was constructed to learn the optimal resource scheduling and IoRT data scheduling strategies. Simulation results were presented to illustrate that the proposed SACRL strategy could efficiently improve the A-ADID-E performance compared with other baseline algorithms based on casual information at the LEO in SIoRTNs. Additionally, we investigated the impact of some network parameters, e.g., battery capacity and data buffer capacity on network performance, which can provide guidance for the future SIoRTN system design.

References

[1] I. F. Akyildiz and A. Kak, "The Internet of space things/CubeSats," *IEEE Netw.*, vol. 33, no. 5, pp. 212–218, Sep. 2019.

[2] A. Campbell and Y. Wang, "Examining the influence of tidal stage on salt marsh mapping using high-spatial-resolution satellite remote sensing and topobathymetric LiDAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5169–5176, Sep. 2018.

[3] Y. Sun, Q. Zeng, B. Geng, X. Lin, B. Sude, and L. Chen, "Deep learning architecture for estimating hourly ground-level $PM_{2.5}$ using satellite remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1343–1347, Sep. 2019.

[4] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio, and R. Prasad, "Satellite communications supporting Internet of remote things," *IEEE Internet Things J.*, vol. 3, no. 1, pp. 113–123, Feb. 2016.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

[5] Y. Bi *et al.*, "Software defined space-terrestrial integrated networks: Architecture, challenges, and solutions," *IEEE Netw.*, vol. 33, no. 1, pp. 22–28, Jan. 2019.
[6] I. Sanad and D. G. Michelson, "A framework for heterogeneous satellite constellation design for rapid response earth observations," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, Mar. 2019, pp. 1–10.
[7] M. Bacco *et al.*, "IoT applications and services in space information networks," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 31–37, Apr. 2019.
[8] T. Hong, W. Zhao, R. Liu, and M. Kadoch, "Space-air-ground IoT network and related key technologies," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 96–104, Apr. 2020.
[9] J.-W. Lee, Y. K. Anguchamy, and B. N. Popov, "Simulation of charge–discharge cycling of lithium-ion batteries under low-earth-orbit conditions," *J. Power Sources*, vol. 162, no. 2, pp. 1395–1400, Sep. 2006.
[10] Y. Yang, M. Xu, D. Wang, and Y. Wang, "Towards energy-efficient routing in satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3869–3886, Dec. 2016.
[11] C.-S. Lu *et al.*, "A new rain attenuation prediction model for the earth-space links," *IEEE Trans. Antennas Propag.*, vol. 66, no. 10, pp. 5432–5442, Oct. 2018.
[12] H. Hermanns, J. Krčál, and G. Nies, "How is your satellite doing? battery kinetics with recharging and uncertainty," *Leibniz Trans. Embedded Syst.*, vol. 4, no. 1, pp. 04:1–04:28, Jan. 2017.
[13] J. A. Fraire, G. Nies, C. Gerstacker, H. Hermanns, K. Bay, and M. Bisgaard, "Battery-aware contact plan design for LEO satellite constellations: The Ulloriaq case study," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 236–245, Mar. 2020.
[14] X. Jia, T. Lv, F. He, and H. Huang, "Collaborative data downloading by using inter-satellite links in LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1523–1532, Mar. 2017.
[15] D. Zhou, M. Sheng, R. Liu, Y. Wang, and J. Li, "Channel-aware mission scheduling in broadband data relay satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1052–1064, May 2018.
[16] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.
[17] R. Liu, M. Sheng, K.-S. Lui, X. Wang, Y. Wang, and D. Zhou, "An analytical framework for resource-limited small satellite networks," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 388–391, Feb. 2016.
[18] D. Zhou, M. Sheng, X. Wang, C. Xu, R. Liu, and J. Li, "Mission aware contact plan design in resource-limited small satellite networks," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2451–2466, Jun. 2017.
[19] D. Zhou, M. Sheng, B. Li, J. Li, and Z. Han, "Distributionally robust planning for data delivery in distributed satellite cluster network," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3642–3657, Jul. 2019.
[20] D. Zhou, M. Sheng, Y. Zhu, J. Li, and Z. Han, "Mission QoS and satellite service lifetime tradeoff in remote sensing satellite networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 990–994, Jul. 2020.
[21] B. Deng, C. Jiang, H. Yao, S. Guo, and S. Zhao, "The next generation heterogeneous satellite communication networks: Integration of resource management and deep reinforcement learning," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 105–111, Apr. 2020.
[22] P. V. R. Ferreira *et al.*, "Reinforcement learning for satellite communications: From LEO to deep space operations," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 70–75, May 2019.
[23] C. Jiang and X. Zhu, "Reinforcement learning based capacity management in multi-layer satellite networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4685–4699, Jul. 2020.
[24] J. Du, C. Jiang, J. Wang, Y. Ren, S. Yu, and Z. Han, "Resource allocation in space multiaccess systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 598–618, Apr. 2017.
[25] A. Golkar and I. L. I. Cruz, "The federated satellite systems paradigm: Concept and business case evaluation," *Acta Astronautica*, vol. 111, pp. 230–248, Jun. 2015.
[26] I. del Portillo. (2017). *ITU-Rpy: A Python Implementation of the ITU-R P. Recommendations to Compute Atmospheric Attenuation in Slant and Horizontal Paths*. [Online]. Available: https://github.com/iportillo/ITU-Rpy/
[27] H. Tsuchida *et al.*, "Efficient power control for satellite-borne batteries using Q-learning in low-earth-orbit satellite constellations," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 809–812, Jun. 2020.
[28] S. Fu, J. Gao, and L. Zhao, "Integrated resource management for terrestrial-satellite systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3256–3266, Mar. 2020.
[29] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.
[30] B. Gu, X. Zhang, Z. Lin, and M. Alazab, "Deep multiagent reinforcement-learning-based resource allocation for Internet of controllable things," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3066–3074, Mar. 2021.
[31] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd Ed. Upper Saddle River, NJ: Prentice-Hall, 2010.
[32] 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 Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.
[33] X. Du, H. Van Nguyen, C. Jiang, Y. Li, F. R. Yu, and Z. Han, "Virtual relay selection in LTE-V: A deep reinforcement learning approach to heterogeneous data," *IEEE Access*, vol. 8, pp. 102477–102492, May 2020.
[34] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled IoT using natural actor–critic deep reinforcement learning," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2061–2073, Apr. 2019.
[35] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting decode-and-forward two-hop communications," *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 3, pp. 309–319, Sep. 2017.
[36] G. J. Gordon, "Reinforcement learning with function approximation-converges to a region," in *Proc. Adv. Neural Inform. Process. Syst.* Cambridge, MA, USA: MIT Press, 2001, pp. 1040–1046.

Di Zhou (Member, IEEE) received the B.E. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 2013 and 2019, respectively. She was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Houston, from 2017 to 2018. Since 2019, she has been with the Broadband Wireless Communications Laboratory, School of Telecommunications Engineering, Xidian University, where she currently holds a faculty postdoctoral position. Her research interests include routing, dynamic resource allocation, and mission planning in space-terrestrial integration networks.

Min Sheng (Senior Member, IEEE) joined Xidian University in 2000, where she is currently a Full Professor and the Director of the State Key Laboratory of Integrated Services Networks. She has published over 200 refereed articles in international leading journals and key conferences in the area of wireless communications and networking. Her current research interests include space-terrestrial integration networks, intelligent wireless networks, and mobile ad hoc networks. She received the China National Funds for Distinguished Young Scientists in 2018. She is the Vice Chair of the IEEE Xi'an Section. She is an Editor of IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.

Yixin Wang received the B.Eng. degree in spatial information and digital technology from Xidian University, Xi'an, China, in 2020, where she is currently pursuing the M.Eng. degree in communication and information systems. Her research interests include mission planning and resource allocation in space information networks.

Jiandong Li (Fellow, IEEE) received the B.E., M.S., and Ph.D. degrees in communications engineering from Xidian University, Xi'an, China, in 1982, 1985, and 1991, respectively. He has been a Faculty Member with the School of Telecommunications Engineering, Xidian University, since 1985, where he is currently a Professor and the Vice Director of the Academic Committee with the State Key Laboratory of Integrated Service Networks. He was a Visiting Professor with the Department of Electrical and Computer Engineering, Cornell University, from 2002 to 2003. He served as the General Vice Chair for ChinaCom 2009 and the TPC Chair of the IEEE ICCC 2013. He was awarded as a Distinguished Young Researcher from NSFC and the Changjiang Scholar from the Ministry of Education, China. His major research interests include wireless communication theory, cognitive radio, and signal processing.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1999 and 2003, respectively. From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department and with the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He has been an AAAS fellow since 2019 and an ACM Distinguished Member since 2019. He is 1% highly cited researcher since 2017 according to Web of Science. He received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, IEEE Leonard G. Abraham Prize in the field of communications systems (Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018. He is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation for contributions to game theory and distributed management of autonomous communication networks.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 10,2023 at 09:49:22 UTC from IEEE Xplore. Restrictions apply.