

Fairness Designs for Load Balancing Optimization in Satellite-Cell-Free Massive MIMO Systems

Trinh Van Chien, Ngo Tran Anh Thu, Nguyen Hoang Lam, Hien Quoc Ngo, *Fellow, IEEE*, Symeon Chatzinotas, *Fellow, IEEE*, and Huynh Thi Thanh Binh, *Member, IEEE*

Abstract—Space-ground communication systems are important in providing ubiquitous services in a large area. This paper considers the fairness designs under a load-balancing framework with heterogeneous receivers comprising access points (APs) and a satellite. We derive an ergodic throughput of each user in the uplink data transmission for an arbitrary association pattern and imperfect channel state information, followed by a closed-form expression with the maximum-ratio combining and rich scattering environments. We further formulate a generic fairness optimization problem, subject to the optimal association patterns for all the users. Despite the combinatorial structure, the global optimal solution to the association patterns can be obtained by an exhaustive search for small-scale networks with several APs and users. We design a low computational complexity algorithm for large-scale networks based on evolutionary computation that obtains good patterns in polynomial time. Specifically, the genetic algorithm (GA) is adapted to the discrete feasible region and the concrete fairness metrics. We extensively observe the fairness design problem by incorporating transmit power control and propose a hybrid genetic algorithm to address the problem. Numerical results demonstrate that the association pattern to each user has a significant impact on the network throughput. Moreover, the proposed GA-based algorithm offers the same performance as an exhaustive search for small-scale networks, while it unveils interesting practical association patterns as the network dimensions go large. The load-balancing approach, combined with power control factors, significantly enhances system performance compared to conventional schemes and configurations with fixed factors.

Index Terms—Satellite-Cell-Free Massive MIMO, load balancing, fairness design, genetic algorithm.

I. INTRODUCTION

The transition to the sixth generation (6G) of terrestrial-space communications marks the forthcoming frontier in wireless network technologies, set to succeed the ubiquitous 5G. These new 6G networks are still in the initial stages of research and development, but they are expected to bring new ways of connecting people with faster data speeds, reduced latency, and increased capacity compared to their predecessors [1]–[4]. In particular, 6G may enable data communication at unprecedented terabit per second speeds, achieve sub-millisecond latency, and offer the capacity to facilitate an extraordinary amount of connected devices within a compact square kilometer area [5]. A technological breakthrough supporting 6G

is the implementation of advanced Massive MIMO systems, designed with numerous antennas utilizing linear beamforming techniques to manage connections within highly concentrated networks of devices [6]. Moreover, the advent of cell-free Massive MIMO technology is expected to further advance spectral and energy efficiency by overcoming the conventional limitations imposed by cellular networks. The design of these networks is fundamentally load balancing, characterized by distributed transmitters cooperatively operating to provide dedicated service to individual users, thereby intensifying the focus on creating a customized network experience [7]. This emphasis on load balancing Cell-free Massive MIMO network propels the user experience to new heights, making possible a range of disruptive technological applications, especially in the domain of immersive extended reality; however, Cell-free Massive MIMO is most effective within restricted coverage areas such as urban [8].

Because of its potential to offer widespread connectivity to numerous users over extensive areas, the field of satellite communications has experienced renewed interest [9], [10]. While geostationary (GEO) satellites can cover large regions, they come with high costs, shared bandwidth, and significant latency [10]. These drawbacks have prompted the development of non-geostationary (NGSO) satellites, such as those in low-Earth orbit (LEO), which are poised to revolutionize radio systems with the expectation of integration into future 6G networks [11]. NGSO satellites, orbiting at lower altitudes than GEO ones, provide benefits like reduced latency and tailored coverage for particular uses or isolated locations [12]. Integrating satellites with cell-free Massive MIMO is recognized as a network architecture for advancing wireless communication systems [7]. Nonetheless, to the best of our knowledge, there is a lack of research exploring a load balancing approach for hybrid space-ground systems, particularly under the framework of coherent signal processing and resource management.

Evolutionary algorithms (EAs) continue to draw interest because of their demonstrated capabilities in solving complex optimization problems that are prevalent in real-world scenarios. As future networks should be multi-layered systems with various integrated technologies, evolutionary strategies like the genetic algorithm (GA) have shown promise in radio resource management, attributed to their effectiveness in effective system designs [13]–[15]. The GA operates on evolutionary biology principles, utilizing selection, mutation, and crossover processes to evolve populations toward high-quality solutions [16]. The strength and adaptability of GAs make them particularly suitable for navigating complex configurations, allowing them to search through large solution spaces to identify superior solutions. Integrating diverse cutting-edge

Trinh Van Chien, Ngo Tran Anh Thu, Nguyen Hoang Lam, and Huynh Thi Thanh Binh are with the School of Information and Communication Technology (SoICT), Hanoi University of Science and Technology (HUST), Vietnam (email: chientv@soict.hust.edu.vn, ngotrananhthu1632001@gmail.com, hoanglam02qh@gmail.com, binhht@soict.hust.edu.vn). Hien Quoc Ngo is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, United Kingdom (email: hien.ngo@qub.ac.uk). Symeon Chatzinotas is with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, L-1855 Luxembourg, Luxembourg (email: symeon.chatzinotas@uni.lu). (Corresponding author: Trinh Van Chien).

technologies in 6G networks introduces significantly complex optimization problems, especially in extensive network environments [17]. In addition, the ability of GAs to scale makes them ideal for deployment in networks that must support an enormous number of devices with a wide array of service needs [18]. Notably, the user association problem is a form of combinatorial optimization, which complicates the direct application of the standard GA.

In this paper, a load balancing network formed by cell-free Massive MIMO and satellite is studied to improve spectral efficiency (SE) with different fairness criteria. Our main contributions are summarized as follows

- We derive the uplink ergodic SE of each user under arbitrary association patterns. The closed-form ergodic SE expression is further obtained as the maximum ratio combining (MRC) locally applied at the APs and satellite.
- We formulate a generic fairness problem to seek the optimal association between each user and the ground and space receivers. For small-scale networks, an exhaustive search through the combinatorial structure can obtain the globally optimal association.
- We propose an efficient algorithm for large-scale networks by exploiting the GA adapted to optimize the association sets under a load balancing framework. The main concept is that binary encoding can be easily adapted to represent the user association. To further enhance system performance, we integrate the power control into the load balancing optimization problem. We then develop a GA framework incorporating hybrid variables tailored to the optimization structure.
- Numerical results demonstrate the critical roles of the satellite and APs in enhancing the ergodic SE. The GA-based association designs offer the same performance as an exhaustive search for small-scale networks and further efficiently optimize for large-scale networks. The transmit power control leads to improved system performance by approximately 20% compared to alternative schemes.

The rest of this paper is organized as follows: Section II presents the integrated satellite-cell-free massive MIMO system along with the channel estimation process. In Section III, the uplink data transmission is described, and then the ergodic SE for every user is obtained in closed form. We formulate the generic fairness optimization problem under the load balancing topology and exploit an exhaustive search to obtain the global optimum in Section IV; besides, for networks with numerous users and APs, we propose GA-based association designs with high-quality solutions. Section V extends the fairness optimization framework by incorporating power control mechanisms and proposes a hybrid genetic algorithm to efficiently solve the problem. Section VI provides numerical results to validate the analytical SE and efficiency of our proposals. Finally, Section VII draws main conclusions.

Notation: Matrices and vectors are denoted by the capital and lower bold letters, respectively. The matrix transpose and Hermitian are denoted by $(\cdot)^T$ and $(\cdot)^H$, respectively. The Euclidean norm is denoted by $\|\cdot\|$. The expectation of a random variable is denoted by $\mathbb{E}\{\cdot\}$, and $\mathcal{CN}(\cdot, \cdot)$, $\mathcal{U}(\cdot, \cdot)$, and

$\mathcal{B}(\cdot, \cdot)$ represent the circularly symmetric Gaussian, uniform, and Bernoulli distributions, respectively. Notations $\wedge, \vee, \neg, \oplus$ are logical operators, specifically conjunction (AND), disjunction (OR), negation (NOT), and exclusive OR (XOR).

II. LOAD BALANCING INTEGRATED SYSTEM, PILOT TRAINING, AND CHANNEL ESTIMATION

This section delves into space-ground communications, taking the practical challenges posed by imperfect channel state information and limited associations.

A. System and Channel Models

We consider an integrated space-ground network where an LEO satellite with M antennas cooperates with N single-antenna APs to serve K single-antenna users. For load balancing purposes which enable the high spectral and energy efficiency, we assume that users can flexibly associate with the satellite and APs. To obtain a good spectral efficiency for each terrestrial user, it should hold that $M + N \gg K$. As illustrated in Fig.1, the APs utilize optical fronthaul links, whereas the satellite connects to the ground station via a radio downlink (feeder link). The ground station then relays the uplinks signals from users to the CPU. We assume that both the optical fronthaul links and the feeder link has imperfect channel gains modeled by a complex Gaussian distribution, affecting the performance of both the pilot training and data transmission phases.¹ Although the propagation channels vary over the time and frequency plane, we adopt the quasi-static fading model where the propagation links are static and frequency-flat over each coherence intervals of τ_c symbols. We assume the network operates in a fast-fading environment, and K symbols in each coherence interval are dedicated to the pilot training phase. The remaining $\tau_c - K$ symbols are used for the uplink data transmission.

Due to sharing the same time and frequency resource, each AP and satellite can simultaneously receive the transmitted signal from all the users. Coherent processing allows us to define the terrestrial and/or non-terrestrial mode by optimizing the load balancing. For the load balancing design, let us denote $\tilde{\alpha}_k \in \{0, 1\}$ the binary variable that establishes the association between user k and the satellite. If $\tilde{\alpha}_k = 1$, user k is served by the satellite. Otherwise, it is not served by the satellite. If user k is served by the satellite, the space channel $\mathbf{h}_k \in \mathbb{C}^M$ is distributed as $\mathbf{h}_k \sim \mathcal{CN}(\bar{\mathbf{h}}_k, \mathbf{R}_k)$. Here, $\bar{\mathbf{h}}_k \in \mathbb{C}^M$ consists of the line of sight (LoS) components, while $\mathbf{R}_k \in \mathbb{C}^{M \times M}$ denotes the spatial correlation. For the user-APs association, we introduce the binary variables $\alpha_k \in \{0, 1\}$, where $\alpha_k = 1$ implies APs provide service to user k , otherwise, APs do not serve user k . Besides, if user k is served by AP n , $g_{nk} \in \mathbb{C}$

¹By utilizing the optical links, the fronthaul enables ideal gains and therefore the exact CPU's position is not necessarily stated due to the assumption on the Gaussian distribution. However, extending our framework to incorporate fronthaul-aware pilot allocation and user association strategies represents a valuable direct for future work.

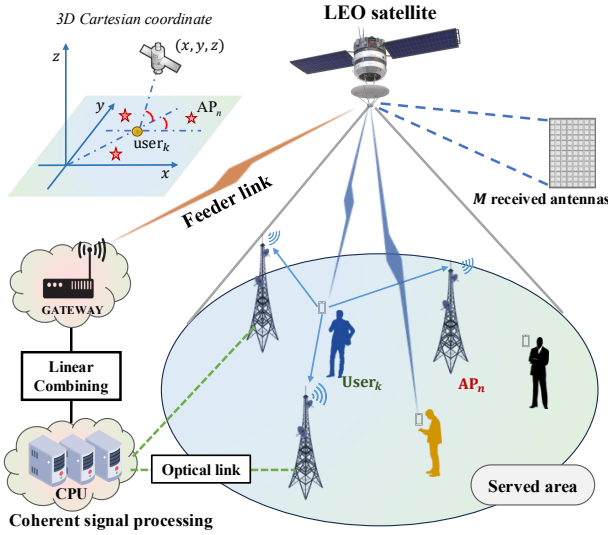


Fig. 1: Illustration of a Satellite-Cell-Free MIMO network.

denotes the ground channel between AP n and user k , which is distributed as $g_{nk} \sim \mathcal{CN}(0, \beta_{nk})$.²

B. Uplink Pilot Training Phase

During the pilot training phase in the uplink, all the K users simultaneously transmit their pilot sequences to the receivers. More specifically, user k transmits its pilot sequence $\psi_k \in \mathbb{C}^K$ with $\|\psi_k\|_2 = 1$ and $\psi_k^H \psi_{k'} = 0, \forall k \neq k'$. Then, the pilot signal received at AP n , denoted by $\mathbf{y}_{pn} \in \mathbb{C}^K$, and at the gateway, denoted by $\mathbf{Y}_p$, are respectively formulated as

$$\mathbf{y}_{pn} = \sum_{k=1}^K \sqrt{pK} g_{nk} \psi_k^H + \mathbf{n}_{pn}, \quad (1)$$

$$\mathbf{Y}_p = \sum_{k=1}^K \sqrt{pK} \mathbf{h}_k \psi_k^H + \mathbf{N}_p, \quad (2)$$

where $p \geq 0$ is the transmit power allocated to each pilot symbol. The additive noise at AP n is denoted by $\mathbf{n}_{pn} \in \mathbb{C}^K$ and distributed as $\mathbf{n}_{pn} \sim \mathcal{CN}(\mathbf{0}, \sigma_a^2 \mathbf{I}_K)$. Besides, $\mathbf{N}_p \in \mathbb{C}^{M \times K}$ denotes the additive noise at the satellite whose elements are i.i.d. $\mathcal{CN}(0, \sigma_s^2)$ random variables. From (1) and (2), the channel estimates and estimation errors are obtained by the minimum mean square error (MMSE) estimation [19].

Lemma 1. *If the MMSE estimation is exploited to estimate the channel g_{nk} , then the channel estimate $\hat{g}_{nk}$ is*

$$\hat{g}_{nk} = \frac{pK\beta_{nk}}{pK\beta_{nk} + \sigma_a^2} \mathbf{y}_{pn}^H \psi_k, \quad (3)$$

distributed as $\hat{g}_{nk} \sim \mathcal{CN}(0, \varrho_{nk})$ and the variance ϱ_{nk} as

$$\varrho_{nk} = \frac{pK\beta_{nk}^2}{pK\beta_{nk} + \sigma_a^2}. \quad (4)$$

The estimation error, defined as $e_{mk} = g_{mk} - \hat{g}_{mk}$, is distributed as $e_{mk} \sim \mathcal{CN}(0, \beta_{mk} - \varrho_{mk})$. Note that the channel estimate $\hat{g}_{mk}$ and the estimation error e_{mk} are independent.

²Uncorrelated Rayleigh fading is considered for terrestrial links where APs are deployed in rich scattering environments that are aligned with measurement data. Inclusion of correlated Rayleigh to describe spatial correlation or Rician fading to characterize the LoS components are also of interest for particular propagation environments. These potential extensions are left for future work.

If the MMSE estimation is exploited to estimate the channel $\mathbf{h}_k$, the channel estimate $\hat{\mathbf{h}}_k$ is formulated as

$$\hat{\mathbf{h}}_k = \bar{\mathbf{h}}_k + \sqrt{pK} \mathbf{R}_k \mathbf{\Psi}_k (\mathbf{Y}_p \mathbf{\Psi}_k - \sqrt{pK} \bar{\mathbf{h}}_k), \quad (5)$$

where $\mathbf{\Psi}_k = (pK\mathbf{R}_k + \sigma_s^2 \mathbf{I}_M)^{-1}$. The channel estimate $\hat{\mathbf{h}}_k$ distributes as $\hat{\mathbf{h}}_k \sim \mathcal{CN}(\bar{\mathbf{h}}_k, pK\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k)$. Let us define the estimation error of the space link $\mathbf{e}_k = \mathbf{h}_k - \hat{\mathbf{h}}_k$, then it distributes as $\mathbf{e}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k - pK\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k)$.

Proof. To estimate the channel g_{nk} , we first project the received pilot signal $\mathbf{y}_{pn}^H$ in (1) into ψ_k and perform the MMSE estimation as $\hat{g}_{nk} = \mathbb{E}\{g_{nk} | \mathbf{y}_{pn}^H \psi_k\}$ by noting that, in our case, the MMSE estimation and the linear MMSE estimation are the same. A similar methodology can be applied to the space link to obtain the channel estimate $\hat{\mathbf{h}}_k = \mathbb{E}\{\mathbf{h}_k | \mathbf{Y}_p \mathbf{\Psi}_k, \bar{\mathbf{h}}_k\}$ by the MMSE estimation. $\square$

The channel estimates and estimation errors are analytically obtained in Lemma 1 as a function of the space and ground links. One can improve the channel estimation quality by carefully selecting the satellite and APs. The spatial correlation $\mathbf{R}_k$ shows the contributions to the channel estimates of the space links. The channel estimates in Lemma 1 are applicable for the fast-fading models, whose channel errors are negligible under the limited pilot power and finite coherence time.

III. UPLINK DATA TRANSMISSION AND ERGODIC THROUGHPUT ANALYSIS

A. Uplink Data Transmission

In the uplink data transmission phase, all the K users transmit signals to its served APs and/or the satellite. In particular, user k transmits a symbol s_k with $\mathbb{E}\{|s_k|^2\} = 1$ and the received signal at AP n , denoted by $y_n \in \mathbb{C}$, is

$$y_n = \sum_{k=1}^K \sqrt{p_k} g_{nk} s_k + n_n, \quad (6)$$

where p_k is the transmit power that user k and $n_n \sim \mathcal{CN}(0, \sigma_n^2)$ is the additive noise. Similarly, the received signal at the gateway of the satellite, denoted by $\mathbf{y} \in \mathbb{C}^M$, is

$$\mathbf{y} = \sum_{k=1}^K \sqrt{p_k} \mathbf{h}_k s_k + \mathbf{n}, \quad (7)$$

where $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_s^2 \mathbf{I}_M)$ is the additive noise during the uplink data transmission. At the CPU, the signal sent from user k is decoded by the following combination

$$\hat{s}_k = \mathbf{w}_k^H \tilde{\alpha}_k \mathbf{y} + \alpha_k \sum_{n=1}^N w_{nk}^* y_n, \quad (8)$$

where $\mathbf{w}_k \in \mathbb{C}^M$ and $w_{nk} \in \mathbb{C}$ are the detection vectors utilized at the gateway and AP n , $\forall n$, respectively.

B. Ergodic Throughput Analysis

In order to decode the transmitted data symbol s_k sent by user k , let us use the formulations of the received signals in (6) and (7) into (8) and obtain

$$\begin{aligned} \hat{s}_k = & \sum_{k'=1}^K \sqrt{p_{k'}} \left(\tilde{\alpha}_{k'} \mathbf{w}_k^H \mathbf{h}_{k'} + \sum_{n=1}^N \tilde{\alpha}_{k'} w_{nk}^* g_{nk'} \right) s_{k'} \\ & + \mathbf{w}_k^H \mathbf{n} + \alpha_k \sum_{n=1}^N w_{nk}^* n_n, \end{aligned} \quad (9)$$

$$\text{SINR}_k = \frac{p_k |\mathbb{E}\{o_{kk}\}|^2}{\sum_{k'=1}^K p_{k'} \mathbb{E}\{|o_{kk'}|^2\} - p_k |\mathbb{E}\{o_{kk}\}|^2 + \mathbb{E}\{\|\mathbf{w}_k\|^2\} \sigma_s^2 + \sum_{n=1}^N \mathbb{E}\{|w_{nk}|^2\} \sigma_n^2}. \quad (14)$$

which is the superposition of the transmitted signals associated with the several receivers and additive noise. In (9), we denote that a proper load balancing approach will mitigate mutual interference, while additive noise is accumulated from the satellite and all the APs. Let us introduce a new variable

$$o_{kk'} = \tilde{\alpha}_{k'} \mathbf{w}_k^H \mathbf{h}_{k'} + \sum_{n=1}^N \tilde{\alpha}_{k'} w_{nk}^* g_{nk'}, \quad (10)$$

which stands for the overall channel and taking the association into account. Besides, let us denote the aggregated noise as

$$\tilde{n}_k = \mathbf{w}_k^H \mathbf{n} + \sum_{n=1}^N w_{nk}^* n_n. \quad (11)$$

After that, by utilizing (10) and (11) into (9), the decoded signal of user k performed at the CPU is equivalent to as

$$\hat{s}_k = \sum_{k'=1}^K \sqrt{p_{k'}} o_{kk'} s_{k'} + \tilde{n}_k = \sqrt{p_k} \mathbb{E}\{o_{kk}\} s_k + \sqrt{p_k} (o_{kk} - \mathbb{E}\{o_{kk}\}) s_k + \sum_{k'=1, k' \neq k}^K \sqrt{p_{k'}} o_{kk'} s_{k'} + \tilde{n}_k, \quad (12)$$

where the first part in the second equation of (12) represents the desired signal from user k with a deterministic channel gain; the second part denotes the beamforming uncertainty; and the remaining parts contain mutual interference and noise. By exploiting the use-and-then-forget channel capacity bounding technique, the uplink ergodic throughput of user k is

$$R_k = B (1 - K/\tau_c) \log_2(1 + \text{SINR}_k), \quad [\text{Mbps}], \quad (13)$$

where B [MHz] is the system bandwidth and the effective signal-to-interference-and-noise ratio (SINR) is given as in (14). The uplink ergodic throughput in (13) can be applied for an arbitrary channel model and detection vectors by numerically evaluating several expectations in the numerator and denominator of (14) over many different realizations of small-scale fading coefficients. However, these expectations are costly for large-scale networks with many APs and users. By exploiting the MRC technique, one can compute (13) in closed form as shown in Theorem 1.³

Theorem 1. *If the MRC technique is deployed at the gateway and APs, the uplink ergodic throughput of user k is*

$$R_k^{\text{mrc}} = B (1 - K/\tau_c) \log_2(1 + \text{SINR}_k^{\text{mrc}}), \quad [\text{Mbps}], \quad (15)$$

where the effective SINR is computed as

$$\text{SINR}_k^{\text{mrc}} = \frac{p_k \left(\tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 + \tilde{\alpha}_k p K \text{tr}(\mathbf{\Theta}_k) + \alpha_k \sum_{n=1}^N \varrho_{nk} \right)^2}{\text{MI}_k + \text{NO}_k} \quad (16)$$

³In this paper, we can derive the exact closed-form solution of the uplink ergodic rate for MRC detection for an arbitrary M , N , and K . Furthermore, MRC is a scalable framework to expand the network with many APs and users. Developing a framework to approximately derive a closed-form expression of the uplink ergodic rate as partial MMSE detection can, in fact, be developed. However, the approach would be different because an assumption $(M + N)/K \rightarrow \infty$ at a fixed rate should be used to align the closed-form expression of the uplink ergodic rate and Monte-Carlo simulations. Since this is a different approach, we would prefer to leave this interesting issue for our future research.

The mutual interference, denoted as MI_k , and the noise, denoted as NO_k , are respectively defined as follows

$$\begin{aligned} \text{MI}_k = & \tilde{\alpha}_k \tilde{\alpha}_{k'} \left(\sum_{k'=1, k' \neq k}^K p_{k'} |\bar{\mathbf{h}}_k^H \bar{\mathbf{h}}_{k'}|^2 + p K \sum_{k'=1}^K p_{k'} \bar{\mathbf{h}}_{k'}^H \mathbf{\Theta}_k \bar{\mathbf{h}}_{k'} \right. \\ & + \sum_{k'=1}^K p_{k'} \bar{\mathbf{h}}_k^H \mathbf{R}_{k'} \bar{\mathbf{h}}_k + p K \sum_{k'=1}^K p_{k'} \text{tr}(\mathbf{R}_{k'} \mathbf{\Theta}_k) \Big) \\ & + \alpha_k \alpha_{k'} \sum_{k'=1}^K \sum_{n=1}^N p_{k'} \varrho_{nk} \beta_{nk'}, \end{aligned} \quad (17)$$

$$\text{NO}_k = \tilde{\alpha}_k \left(\sigma_s^2 \|\bar{\mathbf{h}}_k\|^2 + p K \sigma_s^2 \text{tr}(\mathbf{\Theta}_k) \right) + \alpha_k \sigma_a^2 \sum_{n=1}^N \varrho_{nk}. \quad (18)$$

Proof. The proof is based on computing the expectations in (14) with the channel estimates and estimation errors in Lemma 1. The detailed proof is available in Appendix A. $\square$

The numerator of the SINR in (16) demonstrates the contributions of the space link with both the LoS and NLoS components. The array gain from the satellite antennas and coherent combination scales up with the order of $O((M + N)^2)$ as shown in the numerator of (16). The mutual interference due to multiple access is expressed in (17), which scales up with the order of $O(KM + NK)$. Moreover, the strength of additive noise is added from the APs and satellite with the order of $O(M + N)$. Consequently, the optimal association obtained for each user will effectively reduce mutual interference and noise, leading to improved data throughput. In our network setting, the coherence time is sufficient long to deploy the orthogonal pilot signals, so the system does not suffer from pilot contamination. For the scenarios with the short coherent time and the pilot reuse, large-scale fading decoding can efficiently mitigate both coherent and non-coherent interference thanks to sharing the signaling through the backhaul among the satellite and APs. This potential extension will be left for future work.

IV. FAIRNESS OPTIMIZATION FOR INTEGRATED SYSTEM

This section presents and analyzes a trio of widely considered optimization problems that highlight the benefits of a cooperation between space and ground links.

A. Problem Formulation

Due to the intricate interactions among user utility functions, simultaneous optimization seems elusive. Our primary objective is to optimize the aggregate utility function $f(\alpha_k, \tilde{\alpha}_k)$ to match the requirements of user k within the system, where $\alpha_k, \forall k$, denote the association rules for ground links, and

$\tilde{\alpha}_k, \forall k$, represent those for space links⁴. With the associated notations, the generic fairness optimization problem is

$$\text{maximize } f(\{\alpha_k\}, \{\tilde{\alpha}_k\}) \quad (19a)$$

$$\text{subject to } \alpha_k \in \{0, 1\}, \forall k, \quad (19b)$$

$$\tilde{\alpha}_k \in \{0, 1\}, \forall k. \quad (19c)$$

In this paper, the three popular choices are typically considered for the system utility function $f(\{\alpha_k\}, \{\tilde{\alpha}_k\})$ as

i) Arithmetic mean utility:

$$f(\{\alpha_k\}, \{\tilde{\alpha}_k\}) = \frac{1}{K} \sum_{k=1}^K R_k(\{\alpha_k\}, \{\tilde{\alpha}_k\}). \quad (20)$$

ii) Geometric mean utility:

$$f(\{\alpha_k\}, \{\tilde{\alpha}_k\}) = \left(\prod_{k=1}^K R_k(\{\alpha_k\}, \{\tilde{\alpha}_k\}) \right)^{\frac{1}{K}}. \quad (21)$$

iii) Max-min fairness utility:

$$f(\{\alpha_k\}, \{\tilde{\alpha}_k\}) = \min_{1 \leq k \leq K} R_k(\{\alpha_k\}, \{\tilde{\alpha}_k\}), \quad (22)$$

with $R_k(\{\alpha_k\}, \{\tilde{\alpha}_k\})$ being defined in (13). We emphasize that the max-min fairness optimization provides an equal performance across the users. Contrarily, the arithmetic throughput approach disregards the experience of each individual, prioritizing the overall system utility instead. The geometric mean utility function provides a middle ground between the sum-rate and max-min utilities. It achieves a balance by ensuring that the system maintains a high overall spectral efficiency and provides a reasonable quality of service for each user. The transmit power of each user is fixed so that the framework totally focuses on examining association patterns and their implications on system-level spectral efficiency and fairness. Despite this assumption, the optimization problem is NP-hard due to the binary nature of the association variables. Adding power control would significantly increase the complexity of the algorithm design since it is transformed into a mixed-integer non-convex problem and require more sophisticated approaches to seek for the solution. Jointly optimizing association and transmit power is, therefore, left for future work.

B. Globally Optimal Solution to the Fairness Design Problem

The globally optimal solution to the fairness optimization problem in (19) can be obtained through an exhaustive search, as outlined in Algorithm 1. This method involves generating all possible solutions and evaluating their corresponding fairness levels, which are defined by the objective function. Once all possibilities have been generated, the globally optimal solution that maximizes the objective function is selected. However, the computational complexity of an exhaustive search grows exponentially with the number of available users and receivers, including APs and the satellite, resulting in a complexity of $O(4^K)$. As a result, Algorithm 1 becomes impractical for large-scale networks due to the limited computing power of practical hardware configurations. While an exhaustive search may still be feasible for small-scale networks with a limited

⁴To maintain analytical tractability and focus on the influence of channel conditions and interference, our model assumes an infinite number of connections per AP and satellite.

Algorithm 1 Exhaustive Search Algorithm

Ensure: The optimal solution

- 1: Initialize the current maximum solution as $\text{maxVal} \leftarrow -\infty$.
 - 2: Set parameter $D = 2K$.
 - 3: **for** $k \leftarrow 2^D - 1$ **down to** 1 **do**
 - 4: Convert the index k to its D - dimension binary representation, then convert to a vector with $2K$ dimensions similar to [IV-C1].
 - 5: Using the created vector, evaluate the fitness value F .
 - 6: **if** $F > \text{maxVal}$ **then**
 - 7: Update $\text{maxVal} \leftarrow F$.
 - 8: **end if**
 - 9: **end for**
 - 10: **return** maxVal as the best solution found.
-

number of users and APs, alternative methods that are more efficient in solving fairness design problems must be explored for large-scale networks.

C. Binary-Coded Genetic Algorithm (BCGA)

We introduce a low computational complexity algorithm based on GA associated with an individual encoding to address the three fairness optimization problems.

Algorithm 2 Binary-Coded Genetic Algorithm

Require: Large-scale fading coefficient, bandwidth, number of subcarriers in an (OFDM) symbol, carrier frequency.

Ensure: The sub-optimal solution.

- ```

\\ Initialization
1: Generate initial population S_p [IV-C1].
2: Set max generations S_{MAX} , initialize $S \leftarrow 1$.
3: Set parameters: crossover rate p_c , mutation rate p_m , offspring count $n_c = 2\lfloor p_c \frac{Q}{2} \rfloor$, mutants count $n_m = \lfloor p_m Q \rfloor$.
4: while $S \leq S_{MAX}$ do
5: Initialize next generation $S_{p'} = \emptyset$.
 \\ Crossover
6: for $k_c = 1$ to $n_c/2$ do
7: Select two distinct parents from S_p .
8: Generate two offspring via [IV-C2], add to S_c .
9: end for
10: for $k_m = 1$ to n_m do
11: Select an individual from S_c .
12: Generate mutant via [IV-C3], add to S_m .
13: end for
 \\ Selection
14: Evaluate fitness of $S_{p'} = S_c \cup S_m$.
15: Select Q individuals from $S_p \cup S_{p'}$ for next S_p .
16: $S \leftarrow S + 1$.
17: end while
18: return Best solution found.

```
- 

1) *Solution representation and Initialization:* The population  $S_p$  in the BCGA represents a set of  $Q$  potential solutions. The  $i$ -solution,  $\phi_i = \{\phi_{i1}, \phi_{i2}, \dots, \phi_{iK}\}$  with  $i \in \{1, \dots, Q\}$ , represents a vector that is made up of  $K$  segments, each

having 2 dimensions. Thus, the total dimensions of the vector are  $2K$ . Each component of these vectors, referred to as genes, are binary, meaning they can be either 0 or 1. Specifically, the vector segment  $\phi_{ij}$ , being two-dimensional denoted by  $\phi_{ij} = \{\phi_{ij1}, \phi_{ij2}\}$  as  $z = 1$ ,  $\phi_{ijz}$  indicates the connectivity of user  $j$  and APs, and  $z = 2$ ,  $\phi_{ijz}$  indicates the connectivity of user  $j$  and the satellite. In detail, with a value of one signifying a connection and zero indicating no connection. During the initialization phase of the algorithm, a special individual is created with all the  $2K$  genes set to one, representing a scenario where the users are served by both the satellite and  $N$  APs. The remaining  $Q - 1$  individuals are randomly assigned values as  $\phi_{ijz} = 0$  if  $u < 0.5$  with  $i \in \{1, \dots, Q - 1\}$ ,  $j \in \{1, \dots, K\}$ ,  $z \in \{1, 2\}$ , and  $u \sim \mathcal{U}([0, 1])$ . Otherwise,  $\phi_{ijz} = 1$ .

2) *Binary crossover operator*: Instead of limiting to one type of the crossover method, we incorporate three different crossover techniques, which are applied based on dynamically adjusted probabilities  $\varepsilon_1$  and  $\varepsilon_2$  that reflect their effectiveness during the search process. We create two children solutions  $\phi_{c^1}$  and  $\phi_{c^2}$  to form crossover offspring population  $S_c$  from the two parent solutions  $\phi_{p^1}$  and  $\phi_{p^2}$  as

$$\begin{cases} \phi_{c^1} = (\mathbf{c}_{\text{mask}} \wedge \phi_{p^1}) \vee (\neg \mathbf{c}_{\text{mask}} \wedge \phi_{p^2}), \\ \phi_{c^2} = (\mathbf{c}_{\text{mask}} \wedge \phi_{p^2}) \vee (\neg \mathbf{c}_{\text{mask}} \wedge \phi_{p^1}), \end{cases} \quad (23)$$

and  $\mathbf{c}_{\text{mask}}$  is especially designed for particular crossover. In particular, it holds that

- One-point mask crossover with the probability of  $\varepsilon_1$ :  $\mathbf{c}_{\text{mask}}[i] = 0$ , if  $i < \text{cp}$ , and  $\mathbf{c}_{\text{mask}}[i] = 1$ , otherwise, where  $\text{cp}$  is the crossover point and  $\text{cp} \sim \mathcal{U}(K(N + 1))$ .
- Two-point mask crossover with the probability of  $\varepsilon_2$ :  $\mathbf{c}_{\text{mask}}[i] = 0$ , if  $\text{cp}_1 \leq i < \text{cp}_2$ , and  $\mathbf{c}_{\text{mask}}[i] = 1$ , otherwise, where  $\text{cp}_1$  and  $\text{cp}_2$  are two pivotal crossover points. These points are dependently selected:  $\text{cp}_1 \sim \mathcal{U}(1, N(K + 1) - 1)$  and  $\text{cp}_2 \sim \mathcal{U}(\text{cp}_1 + 1, N(K + 1))$ .
- Uniform mask crossover with the probability of  $1 - \varepsilon_1 - \varepsilon_2$ :  $\mathbf{c}_{\text{mask}} \sim \mathcal{B}(2K, 0.5)$ . Each  $\text{mask}[i]$  is either 0 or 1, chosen randomly with a probability of 0.5. The mask  $\mathbf{c}_{\text{mask}}$  in uniform crossover is not determined by specific crossover points. Instead, it is generated randomly for each gene position across the entire length of the chromosome.

3) *Bitwise mutation operator*: Following the mutation process shown in pseudocode of Algorithm 2, based on the probability  $p_m$ , the parent vector  $\phi_c$  is selected from the current population  $S_p$  and undergoes mutation to produce the trial solution  $\phi_m$  to form mutated offspring population  $S_m$  as

$$\phi_m = \phi_c \oplus \mathbf{M}(n_{\text{mutate}}, 2K), \quad (24)$$

where  $n_{\text{mutate}}$  is a set of indices randomly chosen from 1 to  $2K$  and create a binary mask  $\mathbf{M}(n_{\text{mutate}}, 2K)$  of length  $2K$  where the selected indices are set to 1 (indicating mutation), and others are 0. Besides, the mutation probability  $p_m$  is a critical parameter in this process. It determines the probability of each solution undergoing mutation. Setting a higher mutation probability can increase the frequency of mutations, allowing the algorithm to explore a broader range of genetic variations. This

can be advantageous for escaping local optima and enhancing the genetic diversity within the population. However, if the mutation probability is too high, it might lead to premature convergence, where the algorithm settles too quickly on less optimal solutions. A low mutation probability may slow down the evolution, potentially causing the algorithm to stagnate in local optima due to inadequate exploration.

4) *Survival selection operator*: After generating the offspring population  $S_{p'} = S_c \cup S_m$ , using the above crossover and mutation, the parent population  $S_p$  is combined with to form the merged population  $S_{\text{merged}}$  as:

$$S_{\text{merged}} = S_p \cup S_{p'} = S_p \cup S_c \cup S_m, \quad (25)$$

then sort the merged population  $S_{\text{merged}}$  based on the fitness value and select the best  $Q$  individuals to the new generation as

$$S_p = \text{argsort}(\text{value}(S_{\text{merged}})[1 : Q]), \quad (26)$$

the best solution is selected for the next generation, and the equation (26) guarantees that the objective function of (19) does not decrease across generations.

5) *Convergence and Computational Complexity*: To analyze the computational complexity of the proposal described in Algorithm 2 to solve the problem (19), we break down the computational effort required by its different components. An initialization of the population requires the complexity of the order of  $O(QK)$ . For each generation, if  $\bar{\varphi}$  denotes the count of chromosomes involved in crossovers, the computational load for the crossover operation amounts to the order of  $O(\bar{\varphi}K)$ . The mutation process operates at a basic computational complexity  $O(1)$ , and the selection phase requires  $O(Q \log(Q))$ . Combining these costs, the total computational complexity of Algorithm 2 is in the order of  $O(QK + S(\bar{\varphi}K + Q \log(Q)))$ , where  $Q$  is the size of the population,  $S$  represents the number of generations, and  $K$  denotes the number of users. These complexity formulas show that the algorithm's performance depends on the number of users  $K$  and the GA parameters  $Q$  and  $S$ , but not on the number of antennas  $M$ , which is an important advantage of our approach. Regarding the convergence, we define the expected first hitting time as the average number of generations needed for Algorithm 2 to achieve a fixed solution as shown in Theorem 2.

**Theorem 2.** *From an initial population, solving the fairness design in problem (19) using the BCGA with bitwise mutation at a specified probability  $p_m$ , and assuming the population size matches the solution size, the expected first hitting time is constrained by the following bound:*

$$\mathbb{E}\{\tau\} \geq c(1 - p_m)^{-2K} K^{-1}, \quad (27)$$

where  $\tau$  signifies the number of generations until the best solution is reached,  $c$  is a positive constant, and  $p_m \in (0, 0.5]$  in our experiments.

*Proof.* The proof is completed by formulating the procedure of the BCGA as a Markov chain and investigating its features. The detailed proof is available in Appendix C.  $\square$

Based on Theorem 2, our BCGA algorithm guarantees convergence to the optimum from an initial population as the number of generations increases, ensuring the reliability of our optimization framework even in complex network scenarios.

**Remark 1.** *The optimization problem (19) is NP-hard and complicated due to the network setting with both the terrestrial and nonterrestrial links. The combinatorial complexity makes it challenging to solve, especially for large-scale networks. That is the main reason why intelligent computation algorithms as GA are proposed since the quality of the solution is updated without computing the first and second derivatives of the objective function and constraints. The extension of our framework to support multi-antennas APs is straightforward but non-trivial, which may change the mathematical analysis of network performance and optimization structure. We therefore leave this potential extension for future work.*

## V. FAIRNESS OPTIMIZATION WITH JOINT DATA POWER CONTROL AND LOAD BALANCING

### A. Problem Formulation

To further enhance the network performance and ensure fairness among users, we extend the association optimization framework by incorporating transmit power control. In this enhanced formulation, each user's transmit power is constrained within a practical range  $[0, P_{\max,k}]$ , where  $P_{\max,k}$  denotes the maximum allowable transmit power of user  $k$ . The problem (19) with power control optimization is given by

$$\underset{\{\alpha_k\}, \{\tilde{\alpha}_k\}}{\text{maximize}} \quad f(\{\alpha_k\}, \{\tilde{\alpha}_k\}) \quad (28a)$$

$$\text{subject to} \quad \alpha_k, \tilde{\alpha}_k \in \{0, 1\}, \forall k, \quad (28b)$$

$$0 \leq p_k \leq P_{\max,k}, \forall k. \quad (28c)$$

To normalize the power allocation variable, we define a new variable  $\xi_k$  such that  $p_k = \xi_k P_{\max,k}$ , where  $0 \leq \xi_k \leq 1$ . Consequently, the problem (28) becomes

$$\underset{\{\alpha_k\}, \{\tilde{\alpha}_k\}, \{\xi_k\}}{\text{maximize}} \quad f(\{\alpha_k\}, \{\tilde{\alpha}_k\}, \{\xi_k\}) \quad (29a)$$

$$\text{subject to} \quad \alpha_k, \tilde{\alpha}_k \in \{0, 1\}, \forall k, \quad (29b)$$

$$\xi_k \in [0, 1], \forall k. \quad (29c)$$

which makes the optimization problem more tractable, especially when employing optimization algorithms.

**Lemma 2.** *Let  $\{p_k\}_{k=1}^K$  denote the original power allocation variables with individual constraints  $0 \leq p_k \leq P_{\max,k}$ . Define the normalized variables  $\xi_k \triangleq \frac{p_k}{P_{\max,k}}$  such that  $\xi_k \in [0, 1]$  for all  $k$ . Suppose the system performance metric is modeled by a utility function  $f(\mathbf{p}) = f(p_1, \dots, p_K)$ , where  $f$  is quasi-concave [7] over the box-constrained domain  $\mathcal{P} = \prod_{k=1}^K [0, P_{\max,k}]$ . Define the transformed utility function in the normalized domain as  $\tilde{f}(\boldsymbol{\xi}) \triangleq f(P_{\max,1}\xi_1, \dots, P_{\max,K}\xi_K)$ . Then,  $\tilde{f}(\boldsymbol{\xi})$  is quasi-concave over the unit cube  $\mathcal{B} = [0, 1]^K$ . Furthermore, if  $f$  is strictly quasi-concave and continuously differentiable, then so is  $\tilde{f}$ , and the set of global maximizers of  $f$  and  $\tilde{f}$  correspond via a one-to-one affine mapping.*

*Proof.* Let  $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)} \in [0, 1]^K$  and consider any  $\lambda \in [0, 1]$ . Define the convex combination  $\boldsymbol{\xi}^{(\lambda)} = \lambda \boldsymbol{\xi}^{(1)} + (1 - \lambda) \boldsymbol{\xi}^{(2)}$ . Under the affine transformation  $p_k = \xi_k P_{\max,k}$ , this implies  $\mathbf{p}^{(\lambda)} = \lambda \mathbf{p}^{(1)} + (1 - \lambda) \mathbf{p}^{(2)}$ , where  $\mathbf{p}^{(i)} = (P_{\max,1}\xi_1^{(i)}, \dots, P_{\max,K}\xi_K^{(i)})$  for  $i = 1, 2$ . Since  $f(\mathbf{p})$  is quasi-concave, we have:  $f(\mathbf{p}^{(\lambda)}) \geq \min\{f(\mathbf{p}^{(1)}), f(\mathbf{p}^{(2)})\}$ . Therefore,  $\tilde{f}(\boldsymbol{\xi}^{(\lambda)}) = f(\mathbf{p}^{(\lambda)}) \geq \min\{\tilde{f}(\boldsymbol{\xi}^{(1)}), \tilde{f}(\boldsymbol{\xi}^{(2)})\}$ , which proves that  $\tilde{f}$  is quasi-concave on  $[0, 1]^K$ . When  $\tilde{f}$  is strictly quasi-concave and continuously differentiable,  $\tilde{f}$  similarly exhibits these traits through the one-to-one affine transformation. The global maximizers correspond directly due to the invertibility of the mapping  $p_k = P_{\max,k}\xi_k$ .  $\square$

### B. Hybrid Genetic Algorithm (HGA)

We propose a low-complexity Hybrid Genetic Algorithm (HGA) framework, incorporating a mixed-variable encoding scheme, to efficiently tackle the fairness-driven power control and user association optimization problems. In detail, the updated  $i$ -solution, denoted as  $\boldsymbol{\gamma}_i = \{\boldsymbol{\gamma}_{i1}, \boldsymbol{\gamma}_{i2}, \dots, \boldsymbol{\gamma}_{iK}\}$  for  $i \in \{1, \dots, Q\}$ , is structured as a composite vector consisting of  $K$  segments. Each segment comprises three core dimensions, augmented by an additional dimension to incorporate power control constraints. The current representation is hybrid in nature: one component comprises binary-coded vectors initialized as described in IV-C1, while the other consists of real-valued vectors. For the binary components, the optimization procedure follows Algorithm 2, which addresses discrete association variables. Meanwhile, the real-valued segments are optimized using a Real-Coded Genetic Algorithm framework, following the methodologies i.e. Simulated Binary Crossover and Parameter-based Mutation based on [20], [21].

1) *Simulated Binary Crossover (SBX)*: For the real-valued power allocation variables  $\{\xi_k\}$ , we employ SBX to generate offspring solutions. Given two parent solutions  $\xi_k^{p1}$  and  $\xi_k^{p2}$  for user  $k$ , SBX produces two offspring  $\xi_k^{c1}$  and  $\xi_k^{c2}$  through the following procedure: i) Compute boundary parameters within the normalized bounds  $[0, 1]$ :  $\varepsilon_1 = 1 + \frac{2\xi_k^{p1}}{\xi_k^{p2} - \xi_k^{p1}}$ , and  $\varepsilon_2 = 1 + \frac{2(1 - \xi_k^{p2})}{\xi_k^{p2} - \xi_k^{p1}}$ , with assuming  $\xi_k^{p1} < \xi_k^{p2}$ . ii) Then calculate:  $\vartheta = 2 - \varepsilon^{-(\eta_c + 1)}$ . The distribution index  $\eta_c$  controls the proximity of offspring to their parents. Based on a random parameter  $\mu \sim \mathcal{U}(0, 1)$ , determine the spread factor  $\bar{\varepsilon}$  using polynomial probability distribution as  $\bar{\varepsilon} = (\vartheta\mu)^{\frac{1}{(\eta_c + 1)}}$ , if  $\mu \leq \frac{1}{\vartheta}$ , and otherwise  $\bar{\varepsilon} = \left(\frac{1}{2 - \vartheta\mu}\right)^{\frac{1}{(\eta_c + 1)}}$ . iii) Generate offspring power allocation variables

$$\xi_k^{c1} = 0.5 \left( (\xi_k^{p1} + \xi_k^{p2}) - \bar{\varepsilon} |\xi_k^{p2} - \xi_k^{p1}| \right), \quad (30)$$

$$\xi_k^{c2} = 0.5 \left( (\xi_k^{p1} + \xi_k^{p2}) + \bar{\varepsilon} |\xi_k^{p2} - \xi_k^{p1}| \right). \quad (31)$$

2) *Parameter-based Mutation*: For the normalized power variables  $\xi_k \in [0, 1]$ , polynomial mutation generates a perturbed solution  $\xi_k'$  in the neighborhood of the original  $\xi_k$  as follows: i) generate random parameter  $\mu \sim \mathcal{U}(0, 1)$ ; ii) compute the perturbation factor  $\bar{\delta} = (2\mu + (1 - 2\mu)(1 - \delta_k)^{\eta_{m+1}})^{\frac{1}{\eta_{m+1}}} - 1$ , if  $\mu \leq 0.5$  and  $\bar{\delta} = 1 -$

$(2(1 - \mu) + 2(\mu - 0.5)(1 - \delta_k)^{\eta_m+1})^{\frac{1}{\eta_m+1}}$ , otherwise. Where  $\delta_k = \min[\xi_k, (1 - \xi_k)]$  ensures the mutated solution remains within  $[0, 1]$  and the distribution index  $\eta_m$  controls mutation intensity; and *iii*) generate the mutated power allocation

$$\xi'_k = \xi_k + \bar{\delta}. \quad (32)$$

### C. Convergence and Computational Complexity for HGA

To analyze the computational complexity of the proposed HGA solving problem (29), we examine its hybrid components. Population initialization requires  $O(QK)$  complexity. For each generation, binary crossover operations require  $O(\bar{\varphi}_b K)$ , real-valued crossover (SBX) requires  $O(\bar{\varphi}_r K)$ , real-valued mutation requires  $O(\bar{v}_r K)$ , and selection requires  $O(Q \log(Q))$ . The total computational complexity is  $O(QK + S((\bar{\varphi}_b + \bar{\varphi}_r + \bar{v}_r)K + Q \log(Q)))$ , where  $\bar{\varphi}_b$ ,  $\bar{\varphi}_r$ , and  $\bar{v}_r$  represent the number of individuals participating in binary crossover, real-number crossover, and real-number mutation per generation, respectively. For convergence analysis, we must account for both discrete (association variables) and continuous (power control variables) components.

**Theorem 3.** Consider the HGA solving problem (29) with hybrid solution space  $\mathcal{Y} = \{0, 1\}^{2K} \times [0, 1]^K$ , where the optimal solution set is  $\mathcal{Y}^* \subset \mathcal{Y}$ . Let  $p_m, \eta_m \in (0, 0.5]$  be the binary mutation probability and the polynomial mutation distribution index, respectively. The expected of EFHT  $\tau$  satisfies

$$\mathbb{E}\{\tau\} \geq \tilde{c}(1 - p_m)^{-2K}(1 - \eta_m)^{-K}K^{-1}, \quad (33)$$

where  $\tilde{c} > 0$  is a constant, and  $\tau$  is the number of generations until the optimum is found.

*Proof.* The proof is completed by formulating the procedure of the HGA as a Markov chain and investigating its features. The detailed proof is available in Appendix D.  $\square$

**Remark 2.** Theorem 3 reveals that HGA convergence is limited by the slower of two components binary and real values.

## VI. NUMERICAL RESULTS

This section provides numerical results to demonstrate the system's performance and the effectiveness of two proposed algorithms in comparison to state-of-the-art benchmarks. We simulate a network involving up to 50 APs and 70 users. These users are uniformly dispersed within a 15 square kilometer area, visually represented within a three-dimensional Cartesian coordinate system  $(x, y, z)$ , as illustrated in Fig. 1. An LEO satellite, positioned at the coordinate (300, 350, 400) [km], is equipped with 100 antennas and features an antenna gain of 26.9 [dBi], contrasting with the 10.0 [dBi] gain of the ground devices [22]. The system operates on a bandwidth of 100 [MHz] with a carrier frequency of 20 [GHz], employing a coherence block of 10000 OFDM subcarriers for data transmission. Each data symbol is transmitted at 20 dBW [7]. The noise figures are set at 6 [dB] and 1.3 [dB] for the APs and satellite, respectively. The parameter settings adhere to 3GPP standards outlined in [23], which consider both large and small fading effects in the propagation channels, e.g., for a rural area

### Algorithm 3 Hybrid Genetic Algorithm (HGA)

**Require:** Large-scale fading coefficient, bandwidth, number of subcarriers in an (OFDM) symbol, carrier frequency.  
**Ensure:** The sub-optimal solution.

---

```

\\ Initialization
1: Generate initial population $\mathcal{G}_p$ with hybrid individuals.
2: Set max generations S_{MAX} , initialize $S \leftarrow 1$.
3: Set crossover and mutation parameters: p_c, p_m, η_n, η_m .
4: while $S \leq S_{MAX}$ do
5: Initialize next generation $\mathcal{G}_{p'} = \emptyset$.
 \\ Crossover
6: for $k_c = 1$ to $2\lfloor(p_c + \eta_c)Q/4\rfloor$ do
7: Select two parents from $\mathcal{G}_p$.
8: Generate two offspring. Binary vector follows as in [IV-C2]; meanwhile, the real-valued vector follows as in [V-B1], add to $\mathcal{G}_c$.
9: end for
 \\ Mutation
10: for $k_m = 1$ to $\lfloor(p_m + \eta_m)Q/2\rfloor$ do
11: Select an individual from $\mathcal{G}_c$.
12: Generate mutant. Binary vector follows as in [IV-C3]; meanwhile, the real-valued vector follows as in [V-B2], add to $\mathcal{G}_m$.
13: end for
 \\ Selection
14: Evaluate fitness of each individual in $\mathcal{G}_{p'} = \mathcal{G}_b \cup \mathcal{G}_m$.
15: Select Q best individuals from $\mathcal{G}_p \cup \mathcal{G}_{p'}$.
16: $S \leftarrow S + 1$.
17: end while
18: return Best solution found.
```

---

as  $\beta_{nk} = H_m + H_k - 8.50 - 20 \log_{10}(f_c) - 38.63 \log_{10}(r_{nk}) + \eta_{nk}$ , where  $H_m$  and  $H_k$  are represented the antenna gains at AP  $n$  and user  $k$ , respectively, whereas  $f_c$  is the carrier frequency. The distance between this user and AP  $n$  is denoted as  $r_{nk}$ , and shadow fading,  $\eta_{nk}$ , is modelled as a random variation following a log-normal distribution with a standard deviation of 7 [dB]. The large-scale fading coefficient between user  $k$  and the satellite is determined using one of the models proposed in [24] as

$$\beta_k = H + H_k + \tilde{H}_k - 32.45 - 20 \log_{10}(f_c r_k) + \eta_k, \quad (34)$$

where  $H$  is the receiver antenna gain at the satellite, and its normalized beam pattern is [22], as

$$\tilde{H}_k = \begin{cases} 4 \left| J_1 \left( \frac{2\pi}{\lambda} \alpha \sin(\phi_k) \right) / \left( \frac{2\pi}{\lambda} \alpha \sin(\phi_k) \right) \right|^2, & \text{if } 0 \leq \phi_k \leq \frac{\pi}{2}, \\ 0, & \text{if } \phi_k = 0, \end{cases} \quad (35)$$

where  $\alpha$  denotes the radius of the antenna's circular aperture;  $\lambda$  is the wavelength; and  $\phi_k$  is the angle between user  $k$  and its beam center. In (34), the shadow fading  $\eta_k$  is determined by a log-normal distribution with the standard derivation depending on the carrier frequency, channel condition, and the elevation angle [24]. Besides,  $r_k$  [m] represents the distance between the satellite and user  $k$ , defined as  $z_k = \sqrt{R^2 \sin^2(\theta_k) + z_0^2 + 2z_0 R - R \sin(\phi_k)}$ , where  $R$  is the

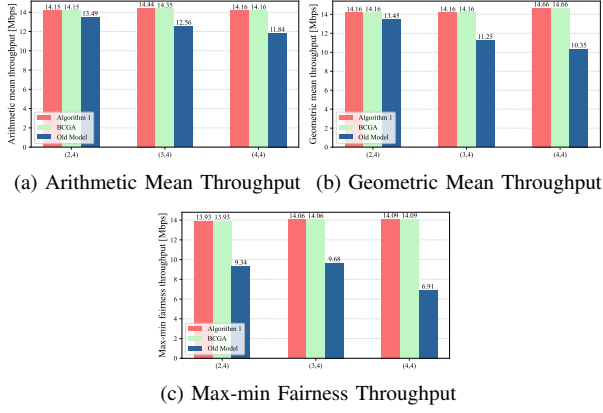


Fig. 2: The optimal results of the proposed algorithms compared the fully connected model.

Earth's radius and  $z_0$  is the satellite altitude. The system model and proposed algorithms were evaluated through a MATLAB environment installed in a personal desktop with an Intel(R) Core(TM) i7-12700 2.80 [GHz] and 64 [GB] RAM. The following benchmarks compare the performances of the proposed algorithms (Exhaustive search in Algorithm 1 and Binary-coded genetic algorithm in Algorithm 2) to the full connected model in [7], another evolutionary algorithms called differential evolutionary (DE) built in [25] and real-coded genetic algorithm (RCGA) built-in [26].

In a small-scale network with four users and the numbers of APs is in  $\{2, 3, 4\}$  together with the values of fitness functions under different benchmarks are presented in Fig. 2. The results demonstrate that Algorithm 2 yields solutions that are nearly identical to the globally optimal solution obtained through an exhaustive search algorithm 1. Additionally, we compare the performance of our proposed algorithms with the full association approach. When considering the Arithmetic mean throughput, our proposed algorithms consistently outperform the full association, with fitness values surpassing it by 10% to 20%. The superiority of our proposed algorithms becomes even more evident when evaluating the Geometric mean throughput and the Max-min fairness throughput, where improvements range from 30% to an impressive 50%.

To observe the convergence of the proposed algorithm (BCGA) and the impact of the problem size, we compare its performance with DE and RCGA. Firstly, within a communication system of 15 users and 15 APs, the performance of all these considered benchmarks is evaluated in terms of the number of generations, as displayed in Fig 4a. Across all three throughput metrics, Arithmetic mean throughput, Geometric mean throughput, and Max-min fairness throughput, the BCGA algorithm consistently outperforms both DE and RCGA. In the first utility, arithmetic mean, BCGA converges to a higher value compared to DE and RCGA, showcasing its superiority in maximizing the overall system throughput. In detail, the optimal result provided by BCGA is slightly better than that by DE and RCGA. Meanwhile, for the remaining utilities, geometric mean and max-min fairness, BCGA excels over another counterpart by reaching notable optimal values. These results highlight the potential of BCGA in effectively

|                           | Arithmetic Utility | Geometric Utility | Max-min Utility |
|---------------------------|--------------------|-------------------|-----------------|
| <b>Total Throughput</b>   | <b>231.145</b>     | 225.927           | 225.365         |
| <b>Minimum Throughput</b> | <b>0.0085</b>      | 5.1648            | <b>6.3699</b>   |

TABLE I: The total throughput and minimum throughput in optimized utility systems.

|                                  | Arithmetic Utility | Geometric Utility | Max-min Utility |
|----------------------------------|--------------------|-------------------|-----------------|
| <b>Satellite only</b>            | 56.80              | 68.80             | <b>74.40</b>    |
| <b>APs only</b>                  | <b>40.40</b>       | 25.20             | 22.40           |
| <b>Both of Satellite and APs</b> | 1.20               | <b>6.00</b>       | 3.20            |

TABLE II: The proportion (%) of user connections between satellite and APs for different optimization utilities.

addressing various utility functions and its robustness in delivering improved network performance compared to other evolutionary algorithms such as DE and RCGA. Furthermore, while keeping the number of APs at 5 and gradually increasing the number of users, we observe a decreasing trend in the throughput of the objective functions, as shown in Fig. 4b, clearly demonstrating the impact of user density on system performance. It is explained that when the number of users is sufficient for the number of APs, the system can still provide sufficiently good throughput for each objective function; however, as the number of users becomes too large, the throughput of each objective function gradually decreases. This trend is attributed to the fixed serving resources of the system. To investigate this phenomenon further, we analyze the results of the three algorithms for each utility. With the Arithmetic Mean, there is no significant difference among the three algorithms when the number of users exceeds 40. For the Geometric Mean and Max-min fairness utilities, even when the number of users surpasses 50, there remains a small discrepancy between the algorithms. Nevertheless, BCGA consistently delivers the best results across all scenarios. On the contrary, a monotonically increasing trend is observed by fixing the number of users at 20 and continuously increasing the number of APs. Regarding the rate of improvement, for the Arithmetic Mean and Geometric Mean, the optimal value achieved by the proposed algorithms exhibits a significant enhancement compared to smaller values of  $N$ . Conversely, for the Max-min fairness objective, the optimal value delivered by the algorithms is exhibited as a slight climb. The monotonically increasing trend is a direct consequence of the independent nature of the connections between APs and users. Adding more APs can improve the data throughput of each individual user.

On the other hand, to evaluate the effectiveness of the benchmarks, Table I presents the total throughput and minimum throughput for the different fitness functions using through tests problem set of 70 users and 50 APs. It compares the system-level throughput and user-level fairness across different optimization approaches. The total throughput indicates the overall performance of the system. The minimum throughput represents the throughput experienced by the worst-off user, which reflects fairness. In the Arithmetic Throughput utility, the overall system performance reaches its peak with a total throughput of 231.145 [Mbps]. However, some users are neglected, with throughput as low as 0.0085 [Mbps]. Conversely, in the Max-min fairness utility, the performance of each individual user improves significantly, even though

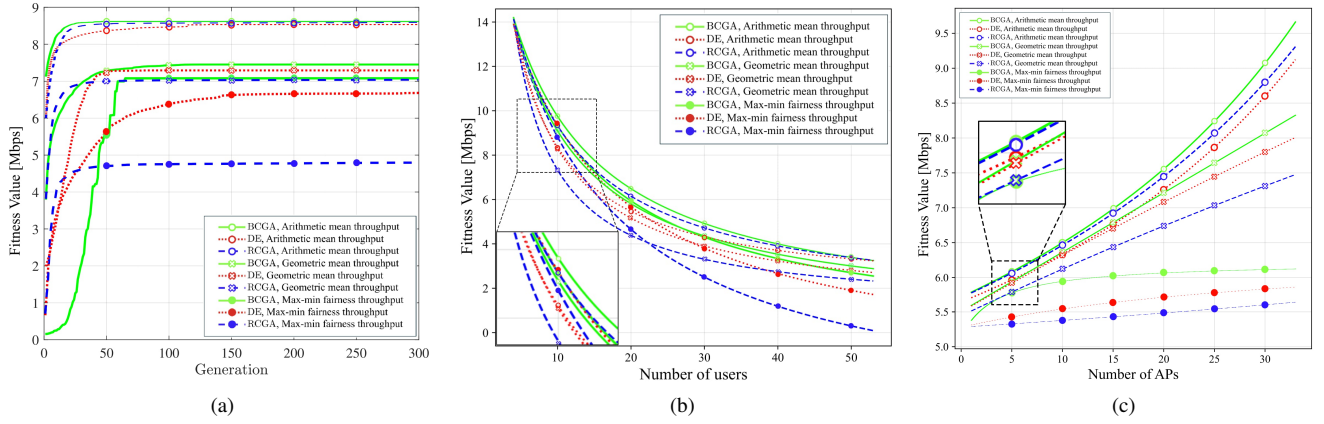


Fig. 3: The system performances of different evolutionary algorithms within model in Problem (19): (a) The convergence rate as the number of generations increases; (b) The evaluation of the gap between the optimal values of different algorithms as the test sets vary; (c) The evaluation of the gap between the optimal values of different algorithms as the test sets vary.

the overall system performance is not as high, with a total throughput of 225.365 [Mbps] and a minimum throughput of 6.3699 [Mbps]. The geometric mean throughput utility offers a balanced trade-off between these two utilities, as previously mentioned. It achieves a total throughput of 225.927 [Mbps], which is slightly lower than the arithmetic throughput utility but higher than the max-min fairness utility. Additionally, the minimum throughput in the Geometric Mean Throughput utility is 5.1648 [Mbps], which is a significant improvement compared to the Arithmetic Throughput utility and only slightly lower than the Max-min fairness utility. Therefore, the Geometric Mean Throughput utility balances maximizing total system performance and ensuring fairness among users. This makes it a proper optimization target that considers the demands of all users.

The roles of the space and ground links in enhancing the uplink spectral efficiency are captured by optimizing the connection of the satellite, APs, and users. We compared the differences between the ergodic throughput of three distinct network architectures: spacial satellite-only (Mode S), terrestrial APs only (Mode A), and our proposed cell-free-terrestrial integrated model in Fig. 4. The box plots demonstrate that our proposed model achieves superior performance metrics compared to both baseline approaches. Using only terrestrial APs might reduce throughput in areas with high user density. Similarly, satellite-only deployment yields reduced user throughput, albeit with minimal variance observed across experimental trials. This indicates a hybrid approach, providing users with the benefits of the joint connectivity where both APs and satellite connections are optimized for different types of traffic or service levels. Moreover, our analysis of the model optimized by Algorithm 2 reveals preferred connections to either satellite or APs, with priority given to the satellite due to the LoS links. Table II shows the percentage of users served only from either the single link or from both. Furthermore, the preference for satellite connectivity is most pronounced in the max-min fairness that optimizes the connection quality for the weakest users. Specifically, this objective shows the highest percentage of users up to 74.40% served only by the

satellite, suggesting a prioritization of ensuring that each user achieves at least a minimum level of service quality, even if it might reduce the overall system performance. This observation highlights reliance on satellite connections, which may provide more consistent coverage and bandwidth compared to terrestrial networks. Integrating a satellite enhances coverage, especially in areas where terrestrial APs struggle. Without the satellite, terrestrial networks would offer limited coverage. Conversely, the sum throughput utility shows a percentage of users (56.80%) served by the satellite only. This objective distributes resources more evenly across all the users, balancing the load between the satellite and APs. In all the realizations, there are 40.40% served by the only APs. This approach can lead to a compromise in individual performance to enhance the total throughput, making it suitable for scenarios where the overall network performance is critical. The geometric throughput, which balances between the total network and per-user throughput, shows intermediate values in both categories, with an enhanced percentage of users about 6.00% connecting through both the satellite and APs.

We benchmark the system performance with and without the incorporation of power efficiency constraints across three utility optimization models in Fig. 5. The numerical evaluations consistently reveal that integrating power efficiency constraints yields a system throughput gain ranging from 10% to 20%. These findings confirm that although this enhancement entails a moderate increase in computational complexity, it leads to a notably more energy-efficient and high-performing network design.

## VII. CONCLUSION

This paper has investigated the applications of adapted GA to solve the user association in integrated space-ground cell-free Massive MIMO systems, where a load balancing approach is utilized to enhance spectral efficiency. Moreover, we formulated and solved the APs and satellite cooperation problem in an effective manner by adjusting the Binary-coded genetic algorithm. Numerical results demonstrated that the proposed algorithm BCGA yields the fairness level closed

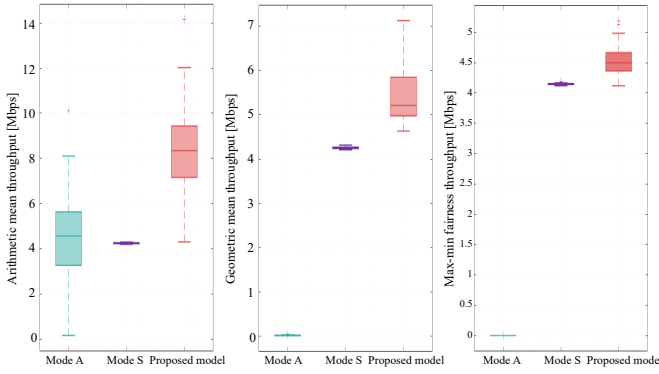


Fig. 4: The comparison of ergodic throughput across different network architectures.

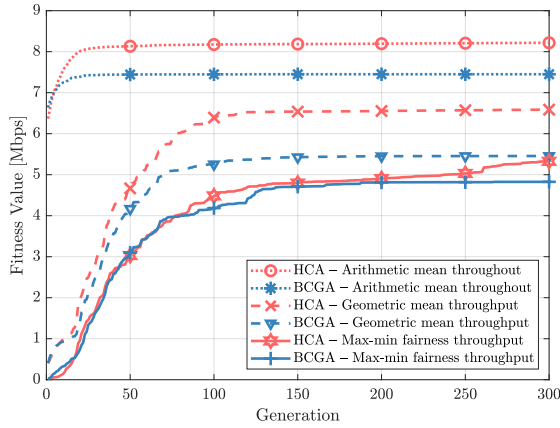


Fig. 5: The system performance comparison between the fixed and optimized power allocation systems based on the optimal results obtained by BCGA and HGA.

to the global optimum for small-scale networks. Besides, it provides a superior solution compared with the full association benchmarks for large-scale networks. We also highlighted the critical roles of the satellite and APs in improving the throughput of each user. Furthermore, the inclusion of power control strategies into the proposed framework enables more flexible resource allocation, which significantly boosts system performance under varying user demands and channel conditions.

## APPENDIX

### A. Proof of Theorem 1

The closed-form expression of the desired signal gain is

$$\begin{aligned} \mathbb{E}\{o_{kk}\} &= \tilde{\alpha}_k \mathbb{E}\{\hat{\mathbf{h}}_k^H \mathbf{h}_k\} + \sum_{n=1}^N \tilde{\alpha}_k \mathbb{E}\{\hat{g}_{nk}^* g_{nk}\} \\ &\stackrel{(a)}{=} \tilde{\alpha}_k \mathbb{E}\{\|\hat{\mathbf{h}}_k\|^2\} + \sum_{n=1}^N \tilde{\alpha}_k \mathbb{E}\{|\hat{g}_{nk}|^2\} \\ &\stackrel{(b)}{=} \tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 + pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k) + \alpha_k \sum_{n=1}^N \varrho_{nk}, \end{aligned} \quad (36)$$

where (a) is obtained by the independence of the channel estimate and its estimation error from user  $k$  to the satellite and APs. Meanwhile, (b) is obtained by the channel statistics in Lemma 1. Consequently, the numerator of (14) is

$$p_k |\mathbb{E}\{o_{kk}\}|^2 = p_k \left| \tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 + pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k) + \alpha_k \sum_{n=1}^N \varrho_{nk} \right|^2. \quad (37)$$

The first term in the denominator of (14) is reformulated as

$$\sum_{k'=1}^K p_{k'} \mathbb{E}\{|o_{kk'}|^2\} = \sum_{k'=1, k' \neq k}^K p_{k'} \mathbb{E}\{|o_{kk'}|^2\} + p_k \mathbb{E}\{|o_{kk}|^2\}. \quad (38)$$

Each expectation of the mutual interference in (38) is computed for  $k \neq k'$  as follows

$$\begin{aligned} \mathbb{E}\{|o_{kk'}|^2\} &= \tilde{\alpha}_{k'} \mathbb{E}\{|\hat{\mathbf{h}}_k^H \mathbf{h}_{k'}|^2\} + \tilde{\alpha}_{k'} \sum_{n=1}^N \mathbb{E}\{|\hat{g}_{nk} g_{nk'}|^2\} = \\ &\tilde{\alpha}_{k'} \text{tr}(\mathbf{R}_{k'} \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_{k'}) + \tilde{\alpha}_{k'} \bar{\mathbf{h}}_k^H \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \bar{\mathbf{h}}_{k'} + \tilde{\alpha}_{k'} \alpha_{k'} \sum_{n=1}^N \varrho_{nk} \beta_{nk'}, \end{aligned} \quad (39)$$

where the first expectation on the right-hand side of (39) is computed in closed form by decomposing the space link into the LoS and NLoS components. The second expectation on the right-hand side of (39) is obtained by the independence of the channel estimate and the estimation error. For  $k' = k$ , the second expectation on the right-hand side of (38) is driven as

$$\begin{aligned} \mathbb{E}\{|o_{kk}|^2\} &= \mathbb{E}\left\{\left|\|\hat{\mathbf{h}}_k\|^2 + \hat{\mathbf{h}}_k^H \mathbf{e}_k + \sum_{n=1}^N |\hat{g}_{nk}|^2 + \sum_{n=1}^N \hat{g}_{nk}^* e_{nk}\right|^2\right\} \\ &= \mathbb{E}\{\|\hat{\mathbf{h}}_k\|^4\} + \mathbb{E}\{|\hat{\mathbf{h}}_k^H \mathbf{e}_k|^2\} + \mathbb{E}\left\{\left|\sum_{n=1}^N |\hat{g}_{nk}|^2\right|^2\right\} \\ &\quad + \mathbb{E}\left\{\left|\sum_{n=1}^N \hat{g}_{nk}^* e_{nk}\right|^2\right\} + 2\mathbb{E}\left\{\|\hat{\mathbf{h}}_k\|^2 \sum_{n=1}^N |\hat{g}_{nk}|^2\right\}, \end{aligned} \quad (40)$$

where the remaining expectations in the last equation of (40) are disappeared since the zero mean of the additive noise. The first expectation in the last equation of (40) is computed as

$$\begin{aligned} \mathbb{E}\{\|\hat{\mathbf{h}}_k\|^4\} &= \left(\tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 + 2pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k)\right)^2 + \\ &2pK\tilde{\alpha}_k \bar{\mathbf{h}}_k^H \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \bar{\mathbf{h}}_k + p^2 K^2 \tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k), \end{aligned} \quad (41)$$

which is obtained based on the channel estimate in (5) and the Boolean property  $\tilde{\alpha}_k^2 = \tilde{\alpha}_k$ . The second expectation in the last equation of (40) is computed as

$$\begin{aligned} \mathbb{E}\{|\hat{\mathbf{h}}_k^H \mathbf{e}_k|^2\} &= \tilde{\alpha}_k \bar{\mathbf{h}}_k^H \mathbf{R}_k \bar{\mathbf{h}}_k - pK\tilde{\alpha}_k \bar{\mathbf{h}}_k^H \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \bar{\mathbf{h}}_k + \\ &pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k) - p^2 K^2 \tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k), \end{aligned} \quad (42)$$

where the channel estimate is given in (5) and the channel estimation error together with its statistics are defined in Lemma 1. The third expectation in the last equation of (40) is computed as

$$\mathbb{E}\left\{\left|\sum_{n=1}^N |\hat{g}_{nk}|^2\right|^2\right\} = \sum_{n=1}^N \varrho_{nk}^2 + \left(\sum_{n=1}^N \varrho_{nk}\right)^2, \quad (43)$$

where the channel estimate  $\hat{g}_{nk}$  is defined in (3) and its variance given in (4). The fourth expectation in the last equation of (40) is computed as

$$\mathbb{E}\left\{\left|\sum_{n=1}^N \hat{g}_{nk}^* e_{nk}\right|^2\right\} = \sum_{n=1}^N \varrho_{nk} (\beta_{nk} - \varrho_{nk}), \quad (44)$$

which is based on the independence of  $\hat{g}_{nk}$  and  $e_{nk}$ ,  $\forall n, k$ . The last expectation in the last equation of (40) is computed as

$$\mathbb{E} \left\{ \|\hat{\mathbf{h}}_k\|^2 \sum_{n=1}^N |\hat{g}_{nk}|^2 \right\} = \left( pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k) + \tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 \right) \sum_{n=1}^N \varrho_{nk}, \quad (45)$$

thanks to the independent of the space and ground channels. Substituting, (41)-(45) into (40), obtaining the closed-form expression

$$\mathbb{E}\{|o_{kk}|^2\} = \left( \tilde{\alpha}_k \|\bar{\mathbf{h}}_k\|^2 + pK\tilde{\alpha}_k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k) \right)^2 + 2pK\tilde{\alpha}_k \bar{\mathbf{h}}_k^H \mathbf{\Psi}_k \bar{\mathbf{h}}_k + p^2 K^2 \tilde{\alpha}_k k \text{tr}(\mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k \mathbf{R}_k \mathbf{\Psi}_k \mathbf{R}_k), \quad (46)$$

after some algebra. The other expectations are computed in a similar manner and we obtain the result as in the theorem.

### B. Useful Definition and Lemmas of Theorem 2

**Definition 1** ([27], [28]). *In the initial stage, we shall introduce essential notations and definitions used for the proof. Let us consider a population space denoted by  $X$ , and  $X^*$  represent a collection of all the optimal solutions. We define a Markov chain as a sequence  $\{\zeta_t\}_{t=0}^\infty$ , where each  $\zeta_t$  is a state at time instance  $t$  ( $t = 0, 1, \dots$ ). The notation  $\mu_t$  denotes the probability of  $\zeta_t$  belonging to  $X^*$ , which is*

$$\mu_t = \sum_{x \in X^*} \Pr(\zeta_t = x), \quad (47)$$

where  $\Pr(\cdot)$  is the probability of an event. We note that  $\{\zeta_t\}_{t=0}^\infty$  is said to converge to  $X^*$  if  $\lim_{t \rightarrow \infty} \mu_t = 1$  and the convergence rate is measured by  $1 - \mu_t$  at time instance  $t$ . In this paper, absorbing the Markov chain will be utilized to model the procedure that obtain a solution to problem (19) due to its desirable theoretical attributes and their feasibility in practical applications. We recall that  $\{\zeta_t\}_{t=0}^\infty$  is an absorbing chain if  $\forall t \in \{0, 1, \dots\} : \Pr(\zeta_{t+1} \notin X^* \mid \zeta_t \in X^*) = 0$ . Let's introduce a random variable  $\tau$  representing the events

$$\tau = 0 : \zeta_0 \in X^* \text{ and } \tau = t : \zeta_t \in X^* \wedge \zeta_i \notin X^* \quad (\forall i \in \{0, 1, \dots, t-1\}), \forall t \geq 1. \quad (48)$$

The expectation  $\mathbb{E}\{\tau\}$ , is called the expected first hitting time of the above Markov chain.

**Lemma 3** ([29]). *Given an absorbing Markov chain  $\{\zeta_t\}_{t=0}^\infty$  with  $\zeta_t \in X$  and a target subspace  $X^* \subset X$ , if two sequences  $\{A_t\}_{t=0}^\infty$  and  $\{B_t\}_{t=0}^\infty$  satisfy*

$$\prod_{t=0}^\infty (1 - A_t) = 0, \quad (49)$$

$$B_t \geq \sum_{x \notin X^*} P(\zeta_{t+1} \in X^* \mid \zeta_t = x) \frac{P(\zeta_t = x)}{1 - \mu_t} \geq A_t, \quad (50)$$

then the Markov chain converges to  $X^*$  and the convergence rate  $(1 - \mu_t)$  is bounded by

$$(1 - \mu_0) \prod_{i=0}^{t-1} (1 - A_i) \geq 1 - \mu_t \geq (1 - \mu_0) \prod_{i=0}^{t-1} (1 - B_i). \quad (51)$$

**Lemma 4** ([29]). *Let  $m$  and  $n$  represent two discrete random variables that take on non-negative integer values with the*

*finite expectations. We denote  $F_m(\cdot)$  and  $F_n(\cdot)$  their respective cumulative distribution functions*

$$F_m(t) = \Pr(m \leq t) = \sum_{i=0}^t \Pr(m = i), \quad (52)$$

$$F_n(t) = \Pr(n \leq t) = \sum_{i=0}^t \Pr(n = i). \quad (53)$$

If  $F_m(t) \geq F_n(t) (\forall t = 0, 1, \dots)$ , then the expectations satisfy

$$\mathbb{E}\{m\} \leq \mathbb{E}\{n\}, \quad (54)$$

with  $\mathbb{E}\{u\} = \sum_t t \Pr(u = t)$ ,  $\mathbb{E}\{v\} = \sum_t t \Pr(v = t)$ .

### C. Proof of Theorem 2

By utilizing Lemma 3 with (50), one obtains

$$1 - \mu_t \leq (1 - \mu_0) \prod_{i=0}^{t-1} (1 - A_i). \quad (55)$$

Note that  $\mu_t$  expresses the distribution of  $\tau$ , i.e.,  $\mu_t = F_\tau(t)$ , we can get the lower bound of  $F_\tau(t)$  as

$$F_\tau(t) \geq \begin{cases} \mu_0, & t = 0, \\ 1 - (1 - \mu_0) \prod_{i=0}^{t-1} (1 - A_i), & t \geq 1. \end{cases} \quad (56)$$

Let's denote a virtual random variable  $\eta$  with its cumulative distribution function equal the lower bound of  $F_\tau(t)$ . After that, the expectation of  $\eta$  is

$$\begin{aligned} \mathbb{E}\{\eta\} &= 0\mu_0 + 1(1 - (1 - A_0)(1 - \mu_0) - \mu_0) + \\ &\sum_{t=2}^{+\infty} t \left( (1 - \mu_0) \prod_{i=0}^{t-2} (1 - A_i) - (1 - \mu_0) \prod_{i=0}^{t-1} (1 - A_i) \right) \\ &= \left( A_0 + \sum_{t=2}^{+\infty} t A_{t-1} \prod_{i=0}^{t-2} (1 - A_i) \right) (1 - \mu_0). \end{aligned} \quad (57)$$

Since  $D_\tau(\tau) \geq D_\eta(\tau)$ , according to Lemma 4,  $\mathbb{E}\{\tau\} \leq \mathbb{E}\{\eta\}$ . Thus, the upper bound of the expected first hitting time is

$$\mathbb{E}\{\tau\} \leq \left( A_0 + \sum_{t=2}^{+\infty} t A_{t-1} \prod_{i=0}^{t-2} (1 - A_i) \right) (1 - \mu_0). \quad (58)$$

We have assumed that Algorithm 2 begins with a non-optimal solution, which implies  $\mu_0 = 0$ . In a similar manner, the lower bound of the expected first hitting time can be derived as

$$\mathbb{E}\{\tau\} \geq \left( B_0 + \sum_{t=2}^{+\infty} t B_{t-1} \prod_{i=0}^{t-2} (1 - B_i) \right) (1 - \mu_0). \quad (59)$$

By mutation, the max probability of a solution being mutated to be the optimal solution is  $p_m(1 - p_m)^{2K-1}$ . Consequently, the maximum probability of a mutated population to become an optimal population is  $1 - (1 - p_m(1 - p_m)^{2K-1})^{2K}$ . Due to the fact that

$$\Pr(\zeta_{n+1} \in X^* \mid \zeta_t = x) \leq 1 - (1 - p_m(1 - p_m)^{2K-1})^{2K}, \quad (60)$$

we obtain the following upper bound

$$\begin{aligned} &\sum_{x \notin X^*} \Pr(\zeta_{n+1} \in X^* \mid \zeta_t = x) \frac{\Pr(\zeta_t = x)}{1 - \mu_t} \\ &\leq \sum_{x \notin X^*} (1 - (1 - p_m(1 - p_m)^{2K-1})^{2K}) \frac{\Pr(\zeta_t = x)}{1 - \mu_t} \\ &= (1 - (1 - p_m(1 - p_m)^{2K-1})^{2K}) \frac{\sum_{x \notin X^*} \Pr(\zeta_t = x)}{1 - \mu_t} \\ &\stackrel{(a)}{\approx} 1 - (1 - p_m(1 - p_m)^{2K-1})^{2K}, \end{aligned} \quad (61)$$

where (a) is obtained by the definition  $\mu_t = \sum_{x \in \mathcal{X}^*} \Pr(\zeta_t = x)$  as shown in Definition 1. Hence,  $1 - \mu_t = \sum_{x \notin \mathcal{X}^*} \Pr(\zeta_t = x)$ . Let  $B_t = 2Kp_m(1 - p_m)^{2K-1}$  and by utilizing (59), we obtain

$$\begin{aligned} \mathbb{E}\{\tau\} &\geq B_o + \sum_{t=2}^{\infty} tB_{t-1} \prod_{i=0}^{t-2} (1 - B_i) = \frac{1}{n} \frac{1}{p_m} \left( \frac{1}{1 - p_m} \right)^{2K-1} \\ &= \frac{(1 - p_m)^{-2K}}{2Kp_m}. \end{aligned} \quad (62)$$

In fact,  $p_m$  is predetermined and therefore,  $\mathbb{E}\{\tau\}$  is obtained as in the theorem.

#### D. Proof of Theorem 3

We extend Theorem (2) by modeling HGA as a Markov chain over the hybrid state space. Each solution is a vector  $\gamma = (\phi, \xi)$ , where i)  $\phi \in \{0, 1\}^{2K}$  (binary associations), ii)  $\xi \in \{0, \frac{1}{L}, \frac{2}{L}, \dots, 1\}^K$  (discretized power levels). The entire state space has size  $2^{2K} \times (L+1)^K$ . Let analysis the mutation probability: 1. Probability of binary-mutating to the optimal  $\phi^*$  is  $\leq 1 - (1 - p_m)^{2K}$  (as in Theorem 2). 2. Probability of Real-mutating to the optimal  $\xi^*$  is  $\leq 1 - (1 - \eta_m)^K$ , where  $\eta_m$  is the minimal probability to jump to a specific discretized level. Real-valued power variables  $\xi_k$  are discretized into  $L$  levels (e.g., fixed-point precision). Then, the minimal mutation probability to reach the optimal discretized power level is  $\eta_m = \Theta(1/L)$ . In worst-case transition probability, the probability to generate the optimal solution from any suboptimal state is bounded below by  $\Pr(p_{\text{hybrid}}) \leq 1 - \underbrace{(1 - p_m)^{2K}}_{\text{binary}} \times \underbrace{(1 - \eta_m)^K}_{\text{real}}$ .

Using the same Markov chain absorption and framework as (61) and (59), the EFHT satisfies

$$\mathbb{E}\{\tau\} \geq \Omega \left( \frac{1}{(1 - p_m)^{2K} (1 - \eta_m)^K |S|} \right). \quad (63)$$

Accounting for  $|S| = \Theta(K)$  (as in Theorem 2), we obtain the bound  $\mathbb{E}\{\tau\} \geq \tilde{c}(1 - p_m)^{-2K} (1 - \eta_m)^{-K} K^{-1}$ .

#### REFERENCES

- [1] B. Smida, A. Sabharwal, G. Fodor, G. C. Alexandropoulos, H. A. Suraweera, and C.-B. Chae, "Full-duplex wireless for 6g: Progress brings new opportunities and challenges," *IEEE Journal on Selected Areas in Communications*, 2023.
- [2] M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, and V. F. Fusco, "The road to 6g: Ten physical layer challenges for communications engineers," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 64–69, 2021.
- [3] H. H. H. Mahmoud, A. A. Amer, and T. Ismail, "6g: A comprehensive survey on technologies, applications, challenges, and research problems," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 4, p. e4233, 2021.
- [4] R. Liu, M. Hua, K. Guan, X. Wang, L. Zhang, T. Mao, D. Zhang, Q. Wu, and A. Jamalipour, "6g enabled advanced transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [5] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, and Y. C. Eldar, "6g wireless communications: From far-field beam steering to near-field beam focusing," *IEEE Communications Magazine*, 2023.
- [6] Y. Zhang, W. Xia, H. Zhao, G. Zheng, S. Lambbotharan, and L. Yang, "Performance analysis of ris-assisted cell-free massive mimo systems with transceiver hardware impairments," *IEEE Transactions on Communications*, 2023.
- [7] T. Van Chien, E. Lagunas, T. M. Hoang, S. Chatzinotas, B. Ottersten, and L. Hanzo, "Space-terrestrial cooperation over spatially correlated channels relying on imperfect channel estimates: Uplink performance analysis and optimization," *IEEE Transactions on Communications*, vol. 71, no. 2, pp. 773–791, 2022.
- [8] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-centric cell-free massive mimo networks: A survey of opportunities, challenges and solutions," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 611–652, 2021.
- [9] R. T. Schwarz, T. Delamotte, K.-U. Storek, and A. Knopp, "Mimo applications for multibeam satellites," *IEEE Transactions on Broadcasting*, vol. 65, no. 4, pp. 664–681, 2019.
- [10] T. Van Chien, E. Lagunas, T. H. Ta, S. Chatzinotas, and B. Ottersten, "User scheduling and power allocation for precoded multi-beam high throughput satellite systems with individual quality of service constraints," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 907–923, 2022.
- [11] H. Zhao, K. Ren, T. Yue, C. Zhang, and S. Yuan, "Transfg: A cross-view geo-localization of satellite and uavs imagery pipeline using transformer-based feature aggregation and gradient guidance," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [12] H. Jia, Z. Ni, C. Jiang, L. Kuang, and J. Lu, "Uplink interference and performance analysis for megasatellite constellation," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4318–4329, 2021.
- [13] T. Van Chien, N. T. A. Thu, N. Hoang Lam, N. Binh, and H. Binh, "On the performance of user association in space-ground communications with integer-coded genetic algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2024, pp. 1373–1380.
- [14] T. V. Chien, B. T. Duc, H. V. D. Luong, H. T. T. Binh, H. Q. Ngo, and S. Chatzinotas, "Active and passive beamforming designs for ser minimization in ris-assisted mimo systems," *IEEE Transactions on Wireless Communications*, 2024.
- [15] N. T. M. Binh, H. T. T. Binh, H. V. D. Luong, N. T. Long, and T. Van Chien, "An efficient exact method with polynomial time-complexity to achieve k-strong barrier coverage in heterogeneous wireless multimedia sensor networks," *Journal of Network and Computer Applications*, vol. 231, p. 103985, 2024.
- [16] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.
- [17] C. A. C. Coello, *Evolutionary algorithms for solving multi-objective problems*. Springer, 2007.
- [18] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6g: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 42–50, 2019.
- [19] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.
- [20] K. Deb, R. B. Agrawal *et al.*, "Simulated binary crossover for continuous search space," *Complex systems*, vol. 9, no. 2, pp. 115–148, 1995.
- [21] K. Deb, M. Goyal *et al.*, "A combined genetic adaptive search (geneas) for engineering design," *Computer Science and informatics*, vol. 26, pp. 30–45, 1996.
- [22] A. Bisognin, D. Titz, C. Luxey, G. Jacquemod, R. Pilard, F. Gianesello, D. Gloria, C. Laporte, H. Ezzedine, J. R. Costa *et al.*, "Millimeter-wave antenna-in-package solutions for wigig and backhaul applications," in *2015 International Workshop on Antenna Technology (iWAT)*. IEEE, 2015, pp. 52–55.
- [23] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostafaei *et al.*, "Evolution of non-terrestrial networks from 5g to 6g: A survey," *IEEE communications surveys & tutorials*, vol. 24, no. 4, pp. 2633–2672, 2022.
- [24] M. Giordani and M. Zorzi, "Non-terrestrial networks in the 6g era: Challenges and opportunities," *IEEE network*, vol. 35, no. 2, pp. 244–251, 2020.
- [25] T. Van Chien, C. V. Le, H. T. T. Binh, H. Q. Ngo, and S. Chatzinotas, "Phase shift design for ris-aided cell-free massive mimo with improved differential evolution," *IEEE Wireless Communications Letters*, 2023.
- [26] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer methods in applied mechanics and engineering*, vol. 186, no. 2–4, pp. 311–338, 2000.
- [27] J. He and X. Yao, "Drift analysis and average time complexity of evolutionary algorithms," *Artificial intelligence*, vol. 127, no. 1, pp. 57–85, 2001.
- [28] —, "Towards an analytic framework for analysing the computation time of evolutionary algorithms," *Artificial Intelligence*, vol. 145, no. 1–2, pp. 59–97, 2003.
- [29] Y. Yu and Z.-H. Zhou, "A new approach to estimating the expected first hitting time of evolutionary algorithms," *Artificial Intelligence*, vol. 172, no. 15, pp. 1809–1832, 2008.