Efficient Energy Efficiency Optimization Method for Cell-Free Massive MIMO-Enabled URLLC Downlink Systems

Bin Yan, Zheng Wang, *Senior Member, IEEE*, Amin Sakzad, *Member, IEEE*, Chuan Zhang, *Senior Member, IEEE*, Yongming Huang, *Fellow, IEEE*, and Derrick Wing Kwan Ng, *Fellow, IEEE*

***Abstract*—This paper investigates the downlink energy efficiency (EE) optimization for cell-free massive multiple input multiple output (CF-mMIMO) systems subject to ultra-reliable and low-latency communication (URLLC) requirements. To achieve superior performance, we jointly consider the impacts of power allocation, access point (AP)-user association, and AP sleep modes under the finite blocklength (FBL) regime, leading to a challenging mixed-integer (MI) non-convex optimization problem. Utilizing a sequential convex approximation (SCA) framework, we first propose the SCA-Relaxation algorithm to convert the original problem into a series of second-order cone programming (SOCP) sub-problems, which can be efficiently addressed via modern convex programming solvers. Moreover, for further reducing computational complexity, we approximate the original problem as a continuous-variable optimization and tackle it via a combination of the Dinkelbach transformation, penalty functions, as well as an accelerated proximal gradient method with adaptive momentum, resulting in the proposed low complexity EE maximization (LCEE-max) algorithm. Besides, the related convergence and complexity analysis of these two algorithms are also presented in detail. Simulation results demonstrate that compared to the state-of-the-art baseline algorithm, the proposed two algorithms achieve the EE improvements of approximately 40% and 30%, respectively, along with a substantial reduction in complexity, thereby enabling efficient and fast resource allocation in CF-mMIMO-enabled URLLC scenarios.**

***Index Terms*—Energy efficiency, ultra-reliable and low-latency communication, mixed-integer non-convex optimization, successive convex approximation, proximal gradient.**

I. Introduction

THE beyond fifth-generation (B5G) and 6G mobile communications place stringent demands on decoding accuracy and transmission latency, which draws significant attention to ultra-reliable and low-latency communication (URLLC) [1]–[10]. To overcome the limitations of traditional Shannon channel capacity, which assumes infinite blocklength and is not suitable for URLLC scenarios, finite blocklength (FBL) information theory has been developed [11]. Meanwhile, benefiting from spatial diversity and macro-diversity, the well-established cell-free massive multiple input multiple output (CF-mMIMO) has been envisioned as a viable paradigm for redeploying URLLC [1]. However, incorporating FBL constraints often complicates resource allocation (RA) for CF-mMIMO-enabled URLLC, especially when energy efficiency (EE) is considered as a critical optimization criterion [3], [7], [12]–[25].

Existing studies have explored various optimization approaches to tackle this challenging problem. Specifically, the performance upper bound of EE optimization was investigated in [19] and [20], and optimal algorithms based on the branch-reduce-and-bound (BRB) strategy were designed. However, due to the exponentially growing computational complexity, their practical applicability is limited, especially in large-scale systems [18]. To this end, numerous optimization schemes based on the sequential convex approximation (SCA) framework have been proposed as an alternative for engineering application. Specifically, by transforming the non-convex problem into a series of tractable convex problems, stationary sub-optimal solutions can be obtained in iterative way [3], [18]–[23]. Although the SCA framework provides a powerful approach to tackle non-convex optimization problems in communication fields, solving the resulting convex sub-problems by standard convex solvers often involves large Hessian matrices storage [17]. To address this issue, the first-order proximal gradient (PG) algorithm from machine learning (ML) also has been introduced into RA optimization, which achieves lower complexity than these SCA methods, while maintaining comparable performance [24]–[27], [29]–[32].

Despite significant research efforts dedicated to EE optimization, existing problem formulations for CF-mMIMO remain some insufficient investigations. In particular, the EE optimization problem for CF-mMIMO downlink system was first formulated in [18], but with the assumption that all access points (APs) are activated and simultaneously serve all users (i.e., a fully connected architecture). Intuitively, the objective function stemming from such an assumption is rather ideal, regardless of the impact of fronthaul overhead on the power consumption. Similarly, [24] also investigates the problem of EE optimization under a fully connected CF-mMIMO architecture, where a monotone accelerated proximal gradient (APG) method is employed to reduce the computational complexity of power allocation in large-scale systems. Notably, it

This work was supported in part by the National Natural Science Foundation of China under Grant 62371124, 61720106003, 62225107, in part by the National Key R&D Program of China under Grants No.2023YFC2205501, in part by the Jiangsu Provincial NSF under BG2024004, with Southeast University, Nanjing 211189, China; and also with Purple Mountain Laboratories, Nanjing 211189, China. (*Corresponding author: Zheng Wang*)

B. Yan, Z. Wang, C. Zhang and Y. Huang are with School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mails: wznuaa@gmail.com). A. Sakzad is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: amin.sakzad@monash.edu). D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

TABLE I
A BRIEF COMPARISON OF THE RELATED LITERATURE.
HERE, M AND K REPRESENT THE NUMBERS OF AP AND USER, RESPECTIVELY.
I_P, I_{APG} AND I_{SCA} DENOTE THE NUMBERS OF PENALTY, APG AND SCA ITERATIONS, RESPECTIVELY.

Literature	Power Allocation	AP-user Association	AP Sleep Modes	URLLC	Limited fronthaul capacity	Computational Complexity
[24]	✓					$\mathcal{O}\left(I_P I_{APG} M K^2\right)$
[18]	✓	✓				$\mathcal{O}\left(I_{SCA}\sqrt{M+K}M^3K^4\right)$
[34]	✓		✓			$\mathcal{O}\left(2^M M^{3.5} K^{3.5}\right)$
[1]	✓			✓		$\mathcal{O}\left(I_{SCA}\left(K^3\left(M+K\right)\right)\right)$
[3]	✓	✓		✓		$\mathcal{O}\left(I_{SCA}2^{MK}\left(M^2K^2\left(M+K\right)\right)\right)$
[19], [28]	✓	✓	✓		✓	Uncertain exponential complexity[1]
SCA-Relaxation algorithm **(This work)**	✓	✓	✓	✓	✓	$\mathcal{O}\left(I_{SCA}I_P\sqrt{MK+M+K}M^3K^3\right)$
LCEE-max algorithm **(This work)**	✓	✓	✓	✓	✓	$\mathcal{O}\left(I_{DT}I_P I_{APG} M^2 K^2\right)$

[1] The complexity of global optimal algorithm is affected by various factors, including the initialization point, the adopted convex optimization solvers, and the specific BRB strategy. Neither [19] nor [28] provides an explicit complexity analysis, but both indicate that the algorithm suffers from an indeterminate exponential complexity. As shown in Section VI, the global optimal algorithm is only applicable for small-scale systems.

has been widely recognized that the AP-user association in CF-mMIMO systems tends to be sparse, which facilitates efficient RA [12], [13], [33]. However, [24] does not explicitly optimize the AP-user association during problem formulation. As a result, some AP-user pairs with extremely poor channel quality may still be allocated near-zero transmit power, which leads to unnecessary fronthaul overhead and inefficient use of communication resources. To improve EE performance, the work in [3] jointly optimized both power allocation and AP-user association for CF-mMIMO-enabled URLLC systems. However, this study neglects the impact of fronthaul traffic, which remains unaddressed. Besides, in EE optimization problems, the sleep modes of APs should be explicitly considered in the formulation, which is overlooked in [3], [18], [24]. This is essential because APs exhibit different power consumption levels depending on whether they are in active or sleep modes. Due to the oversight of AP sleep modes, overestimations of system power consumption would most likely be made [33], [34]. Furthermore, the sleep mode coefficients are discrete value of $\{0, 1\}$ and also tightly coupled with power allocation and AP-user association, complicating optimization algorithm design. Although [19], [28], [34] have investigated the AP sleep control issue, their proposed algorithms rely heavily on general-purpose convex optimization solvers, which inevitably result in relatively high computational complexity.

In particular, EE optimization problems in CF-mMIMO-enabled URLLC scenarios face even more stringent challenges [1], [3], [7]. Specifically, under the FBL regime, the achievable rate expression must incorporate a non-convex channel dispersion function, which significantly increases the complexity of algorithm design [11]. It is worth noting that the aforementioned optimization works have largely overlooked the critical constraint of limited fronthaul capacity. Consequently, the feasibility of achievable rate characterization under this constraint remains an open problem. Furthermore, the optimization algorithms in [1]–[10] are all based on the SCA framework, where feasible solutions are obtained iteratively leveraging modern convex solvers. However, in practical latency-sensitive communication scenarios, low-complexity first-order algorithms may offer a more favorable solution, yet this approach has remained largely unexplored in prior research.

In this work, to improve the achievable gains of EE optimization while strictly satisfying the required system constraints, we conduct a detailed investigation for CF-mMIMO-enabled URLLC downlink systems. In a nutshell, the specific contributions of this paper are summarized as follows:

- First, under imperfect statistical CSI, pilot contamination, and limited fronthaul link capacity, we formulate a sophisticated mixed-integer (MI) non-convex EE optimization problem for CF-mMIMO-enabled URLLC downlink systems, which includes the factors of power allocation, AP-user association and AP sleep modes. While previous studies have individually considered subsets of these factors, our work comprehensively investigates power allocation, AP-user association, and AP sleep control within a unified optimization framework under fronthaul constraint for CF-mMIMO-enabled URLLC systems.
- Second, to efficiently solve the above MI non-convex EE optimization problem, the SCA-Relaxation algorithm is proposed. In particular, to handle the fronthaul link capacity constraint, a concave lower bound for channel dispersion function is derived, which transforms the non-convex constraint into a tractable form. By using compact surrogate functions, we transform the original MI non-convex problem into a series of second-order cone programming (SOCP) problems, which can be solved

iteratively with convex programming solvers. More importantly, the iterations are guaranteed to at least converge to a *Fritz John* point, which ensures necessary optimality conditions in nonlinear programming [21].

- Third, we propose a low complexity EE maximization (LCEE-max) algorithm to further reduce the complexity burden. Specifically, we exploit the coupling between optimization variables to approximate the original MI non-convex problem as a continuous non-convex one. Then, the *Dinkelbach* transformation (DT) is applied to convert it into a parametric subtracting form [23], where the penalty function method is then adopted to turn it into an unconstrained optimization sub-problem. Built on this, we adopt the accelerated proximal gradient with adaptive momentum (APG-AM) method to address it. Moreover, based on the *Lipschitz* smoothness of the objective and proper step size, the LCEE-max algorithm converges to a stationary point with significantly reduced complexity.

The two proposed algorithms exhibit a trade-off between performance and complexity: the former emphasizes performance enhancement, while the latter focuses on lower computational complexity. To sum up, we present a clear comparison of our contributions to the related literature in Table I. Here, M and K represent the numbers of AP and user, respectively. I_P, I_{APG} and I_{SCA} denote the numbers of penalty, APG and SCA iterations, respectively.

The rest of paper is organized as follows. Section II briefly introduces the model of CF-mMIMO-enabled URLLC downlink systems. In Section III, we establish the detailed EE optimization problem, which leads to a complex MI non-convex challenge. Section IV introduces the proposed SCA-Relaxation algorithm, followed by the related complexity analysis. In Section V, the LCEE-max algorithm is proposed to further reduce the complexity. Section VI presents the simulation results of the proposed two algorithms for CF-mMIMO-enabled URLLC downlink systems. Finally, Section VII concludes the paper.

Notation: Bold lower and upper case letters represent vectors and matrices. The superscripts $(\cdot)^T$ and $(\cdot)^H$ stand for the transpose and conjugate-transpose, respectively. The notations $\|\cdot\|$, $\langle\cdot,\cdot\rangle_F$, $\odot$ and $\mathbb{E}\{\cdot\}$ describe the ℓ_2-norm, Frobenius inner product, Hadamard product and expectation, respectively. Besides, $\mathbf{X}\sim\mathcal{CN}(\mathbf{0},\mathbf{I}_N)$ denotes a N-dimensional complex vector, each element of which is independent and obeys the zero mean and unit variance symmetric complex Gaussian random variable. $\mathbf{X}\in\mathbb{R}^{M\times K}(\mathbb{C}^{M\times K})$ means that $\mathbf{X}$ is a real (complex) matrix with M rows and N columns. x_{mk} is the (m,k)-th entry of matrix $\mathbf{X}$. $[a,b]$ represents the concatenation of two elements into a row vector, while $[a;b]$ represents their concatenation into a column vector and $\mathrm{diag}(a,b)$ form a square matrix with diagonal elements a,b.

II. SYSTEM MODEL

Consider an user-centric CF mMIMO downlink system with M multi-antenna APs and K single-antenna users, while each AP is equipped with N antennas. The channel vector $\mathbf{g}_{mk}\in\mathbb{C}^N$ between the m-th AP, $\forall m\in\{1,\cdots,M\}$ and k-th user, $\forall k\in\{1,\cdots,K\}$ is modeled as [1]

$$\mathbf{g}_{mk}=\sqrt{\beta_{mk}}\mathbf{h}_{mk}. \tag{1}$$

Here, β_{mk} represents the large-scale fading coefficient (LSFC), which includes both path loss and shadow fading, $\mathbf{h}_{mk}$ is the small-scale fading vector following $\mathcal{CN}(\mathbf{0},\mathbf{I}_N)$. Similar to most of previous works, it is assumed that the system operates in time division duplex (TDD) mode, which avoids the downlink channel estimation by channel reciprocity. The entire integral coherence interval T is divided into two parts: τ_u is allocated for uplink channel estimation and the remaining duration $\tau_d\triangleq T-\tau_u$ is for downlink data transmission. As illustrated in Fig. 1, the large-scale fading information remains stable over several coherence intervals, and once it changes, each AP is either activated or put into sleeping mode according to the dynamically changing channel gains.

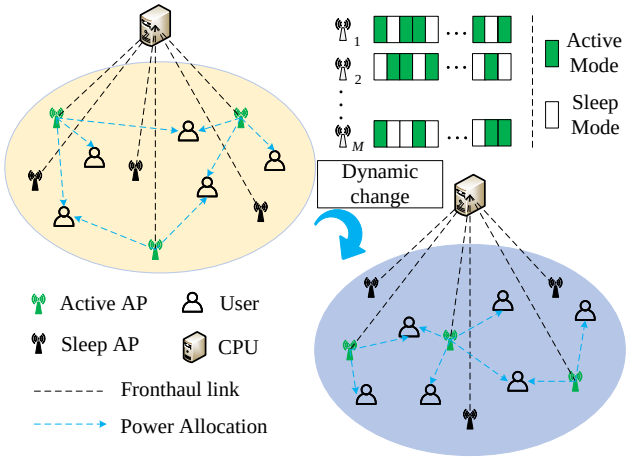


Fig. 1. Illustration of the CF-mMIMO-enabled URLLC systems.

A. Uplink Channel Estimation

In the uplink channel estimation phase, all users synchronously transmit their pilot sequences to APs. The pilot signal sent by the k-th user is $\sqrt{\rho_u\tau_u}\boldsymbol{\psi}_k$, where ρ_u is the normalized uplink transmit power with pilot sequence $\boldsymbol{\psi}_k\in\mathbb{C}^{\tau_u}$. Then, the signal received at the m-th AP is given by [35]

$$\mathbf{Y}_m=\sqrt{\rho_u\tau_u}\sum_{k=1}^{K}\mathbf{g}_{mk}\boldsymbol{\psi}_k^H+\mathbf{N}_m, \tag{2}$$

and $\mathbf{N}_m\in\mathbb{C}^{N\times\tau_u}$ is a normalized unit Gaussian noise matrix for analytical tractability [34]. According to minimum mean square error estimation (MMSE) criterion, the channel estimate $\hat{\mathbf{g}}_{mk}\in\mathbb{C}^N$ is derived as [24]

$$\hat{\mathbf{g}}_{mk}=\frac{\sqrt{\rho_u\tau_u}\beta_{mk}}{\rho_u\tau_u\sum_{k'=1}^{K}\beta_{mk'}|\boldsymbol{\psi}_{k'}^H\boldsymbol{\psi}_k|^2+1}\mathbf{Y}_m\boldsymbol{\psi}_k. \tag{3}$$

Eventually, the estimated value of the channel vector follows a distribution of $\mathcal{CN}(\mathbf{0}, \gamma_{mk}\mathbf{I}_N)$, which leads to the following mean-square of the estimate

$$\gamma_{mk} = \frac{\rho_u \tau_u \beta_{mk}^2}{\rho_u \tau_u \sum_{k'=1}^{K} \beta_{mk'} |\boldsymbol{\psi}_{k'}^H \boldsymbol{\psi}_k|^2 + 1}. \tag{4}$$

Note that when the k-th user and the k'-th user are assigned the same pilot sequence, it has $|\boldsymbol{\psi}_{k'}^H \boldsymbol{\psi}_k|^2 = 1$, i.e. there exists pilot contamination, otherwise $|\boldsymbol{\psi}_{k'}^H \boldsymbol{\psi}_k|^2 = 0$.

B. Downlink Data Transmission

In the downlink data transmission phase, AP adopts the precoding technique and controls the power allocation. Typically, the downlink transmitted signal $\mathbf{x}_m^d \in \mathbb{C}^N$ from the m-th AP is given by [18]

$$\mathbf{x}_m^d = \sqrt{\rho_d} \sum_{k=1}^{K} \sqrt{\eta_{mk}} \mathbf{w}_{mk} q_k. \tag{5}$$

Here, ρ_d represents the normalized downlink transmit power, η_{mk} is the non-negative power allocation coefficient of the m-th AP for the k-th user, q_k is the symbol intended for the k-th user satisfying $\mathbb{E}\{|q_k|^2\} = 1, \forall k$, $\mathbb{E}\{q_i^* q_k\} = 0, \forall k, \forall i \neq k$, and $\mathbf{w}_{mk} \in \mathbb{C}^N$ represents the precoding vector. To strike a balance between performance and computational complexity, we adopt the widely used full-pilot zero-forcing (FZF) precoding scheme [3], [6] As a scalable procoding scheme, FZF precoding can provide an array gain of $(N - \tau_u)$, without increasing the fronthaul overhead [35]. To be specific, the form of the FZF precoding is as follows

$$\mathbf{w}_{mk}^{\text{FZF}} = \frac{\overline{\mathbf{G}}_m (\overline{\mathbf{G}}_m^H \overline{\mathbf{G}}_m)^{-1} \mathbf{e}_k}{\sqrt{\mathbb{E}\{\|\overline{\mathbf{G}}_m (\overline{\mathbf{G}}_m^H \overline{\mathbf{G}}_m)^{-1} \mathbf{e}_k\|^2\}}}, \text{ s.t. } N \geq \tau_u + 1. \tag{6}$$

Here, $\mathbf{e}_k \in \mathbb{C}^{\tau_u \times 1}$ denote the k-th column of the identity matrix $\mathbf{I}_{\tau_u}$, where the index k corresponds to the pilot sequence assigned to the k-th user. Meanwhile, due to pilot contamination, the local channel information $\overline{\mathbf{G}}_m$ corresponds to the processed full-rank estimation matrix, i.e. [31]

$$\overline{\mathbf{G}}_m = \mathbf{Y}_m \boldsymbol{\Psi} \in \mathbb{C}^{N \times \tau_u}, \tag{7}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{\tau_u}] \in \mathbb{C}^{\tau_u \times \tau_u}$ is the pilot-book matrix. Based on the *use-and-then-forget* (UatF) principle [35], the signal-to-interference-plus-noise ratio (SINR) of the k-th user is given in (8).

Here, to simplify the representation of subsequent optimization algorithms, the following definitions are given: $\theta_{mk} \triangleq \sqrt{\eta_{mk}}$ and $\boldsymbol{\theta}_k \triangleq [\theta_{1k}, \theta_{2k}, \ldots, \theta_{Mk}]^T \in \mathbb{R}^M$, the array gain $a_g \triangleq N - \tau_u$. $\boldsymbol{\gamma}_{kk'} \triangleq |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_{k'}| [\sqrt{\gamma_{1k}}, \ldots, \sqrt{\gamma_{MK}}]^T \in \mathbb{R}^M$, and $\boldsymbol{\kappa}_k \triangleq [\sqrt{\beta_{1k} - \gamma_{1k}}, \ldots, \sqrt{\beta_{Mk} - \gamma_{Mk}}]^T \in \mathbb{R}^M$. Then, the SINR expression in (8) can be simplified as (9), which is parameterized by $\boldsymbol{\theta}$ rather than $\boldsymbol{\eta}$. Here, $\text{CP}_k(\boldsymbol{\theta}) \triangleq \rho_d a_g (\boldsymbol{\theta}_k^T \boldsymbol{\gamma}_{kk})^2$ denotes the coherent precoding gain, $\text{NI}_k(\boldsymbol{\theta}) \triangleq \rho_d a_g \sum_{k' \neq k}^{K} (\boldsymbol{\theta}_{k'}^T \boldsymbol{\gamma}_{kk'})^2 + \rho_d \sum_{k'=1}^{K} \|\boldsymbol{\theta}_{k'} \odot \boldsymbol{\kappa}_k\|^2$ represents the equivalent noise interference.

C. Spectral Efficiency (SE) Under Finite Blocklength

Unlike Shannon capacity under the assumption of infinite coding blocklength, the influence of decoding error probability on SE must be considered in URLLC. Specifically, based on (9), the FBL spectral efficiency (SE) of the k-th user with the acceptable decoding error probability $\epsilon_k > 0$ is given by [11]

$$\text{SE}_k(\boldsymbol{\theta}) = \tilde{\tau}_d \log_2(1 + \text{SINR}_k(\boldsymbol{\theta})) - \sqrt{\frac{\tilde{\tau}_d V_k(\boldsymbol{\theta})}{T}} Q^{-1}(\epsilon_k), \tag{10}$$

where $\tilde{\tau}_d = \frac{\tau_d}{T}$, $V_k(\boldsymbol{\theta}) = \left[1 - \frac{1}{(1+\text{SINR}_k(\boldsymbol{\theta}))^2}\right] \log_2^2(e)$ denotes the channel dispersion, $Q^{-1}(x)$ is the inverse function of the tail distribution function $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} \text{d}t$.

D. User-centric Power Consumption Model

Generally, the total system power consumption mainly consists of two parts: AP operation and fronthaul link [18], [33].

1) Power consumption of AP operation: After precoding and power allocation, each AP transmits signals and operates its circuitry, both of which contribute to power consumption. To reduce power consumption, each AP can operate either in active mode or sleep mode, with corresponding changes in power consumption. In this paper, a binary variable b_m is defined to represent the operational status of the m-th AP. Specifically, this consumption for the m-th AP is given by [3]

$$P_m^{\text{A}}(\boldsymbol{\theta}_m, b_m) = \frac{\rho_d N_0 \|\boldsymbol{\theta}_m\|^2}{\alpha_m^{\text{A}}} + b_m N P_m^{\text{A,chain}} + P_m^{\text{A,fix}}. \tag{11}$$

Here, N_0 is the noise power, $\boldsymbol{\theta}_m \in \mathbb{R}^{1 \times K}$ corresponds to the square root of all the power allocation coefficients for the m-th AP, $\rho_d \|\boldsymbol{\theta}_m\|^2 = \mathbb{E}\{\|\mathbf{x}_m^d\|^2\}$ is the m-th AP transmitted power. The quantity $0 < \alpha_m^{\text{A}} \leq 1$ is the power amplifier efficiency, and $P_m^{\text{A,fix}}$ is the fixed AP power consumption. Besides, $P_m^{\text{A,chain}}$ represents the power consumption of the circuitry associated with each RF chain. When the m-th AP is in sleep mode, i.e., $b_m = 0$, the related power consumption disappears.

2) Fronthaul link power consumption: Both APs and the central processing unit (CPU) require fronthaul link for data transfer, which also incur power consumption. This power consumption chiefly depends on the association status between APs and users [18]. Similarly, a binary variable a_{mk} is introduced to represent the association status between the m-th AP and the k-th user. To be specific, the power consumption generated by the fronthaul link is [33], [34]

$$P_m^{\text{F}}(\boldsymbol{\theta}_m, \mathbf{a}_m) = B \xi_m^{\text{F}} \sum\nolimits_{k=1}^{K} a_{mk} \text{SE}_k(\boldsymbol{\theta}) + P_m^{\text{F,fix}}, \tag{12}$$

where B is the system bandwidth, ξ_m^{F} represents the traffic-dependent power (in W/bps), and $P_m^{\text{F,fix}}$ is the fixed fronthaul power consumption.

Overall, the total downlink systems power consumption can be expressed as

$$P_{\text{total}}(\boldsymbol{\theta}, \mathbf{A}, \mathbf{b}) = \sum_{m=1}^{M} \left(P_m^{\text{A}}(\boldsymbol{\theta}_m, b_m) + P_m^{\text{F}}(\boldsymbol{\theta}_m, \mathbf{a}_m) \right). \tag{13}$$

III. Problem Formulation of Joint Optimization

We now consider jointly optimizing the power allocation coefficients, the service association between APs and users, and the sleep modes of each AP, with the aim of maximizing the system's EE under practical constraints, which corresponds to the following problem formulation:

$$\mathcal{P}_1: \max_{\boldsymbol{\theta},\mathbf{A},\mathbf{b},\mathbf{V}} \quad \mathrm{EE}(\boldsymbol{\theta},\mathbf{A},\mathbf{b}) \triangleq \frac{B\sum_{k=1}^{K}\mathrm{SE}_k(\boldsymbol{\theta})}{P_{\text{total}}(\boldsymbol{\theta},\mathbf{A},\mathbf{b})} \tag{14a}$$

$$\text{s.t.} \quad \mathrm{SE}_k(\boldsymbol{\theta}) \geq \mathrm{SE}_{th}, \quad \forall k, \tag{14b}$$

$$\sum\nolimits_{k=1}^{K} a_{mk}\mathrm{SE}_k(\boldsymbol{\theta}) \leq C_{\max}, \quad \forall m, \tag{14c}$$

$$\theta_{mk} \geq 0, \quad \forall m,\forall k, \tag{14d}$$

$$\|\boldsymbol{\theta}_m\|^2 \leq b_m, \quad \forall m, \tag{14e}$$

$$\theta_{mk}^2 \leq a_{mk}v_{mk}, \quad \forall m,\forall k, \tag{14f}$$

$$v_{mk} \leq a_{mk}, \quad \forall m,\forall k, \tag{14g}$$

$$a_{mk} \leq b_m,\forall k; b_m \leq \sum_{k=1}^{K} a_{mk},\forall m, \tag{14h}$$

$$a_{mk} \in \{0,1\}, b_m \in \{0,1\},\forall m,\forall k. \tag{14i}$$

Specifically, SE_{th} is the given SE threshold, (14b) ensures that each user experiences reliable quality-of-service (QoS). In (14c), $C_{\max}$ represents the finite per-fronthaul capacity, which is a crucial constraint in distributed systems [17], [28]. Furthermore, (14e) represents the limited transmission power of each AP [35]. Nevertheless, it should be noted that there is a coupling relationship between power allocation and AP sleep modes. When the m-th AP is in sleep mode, the corresponding power allocation coefficients for the K users is forced to be zero (i.e. $b_m = 0 \Leftrightarrow \{\theta_{m1} = \cdots = \theta_{mK} = 0\}$), and vice versa. Clearly, the binary vector $\mathbf{b}$ provides an explicit link between AP's sleep mode and power allocation.

On the other hand, there also exists a coupling relationship between power allocation and AP-user service connection, i.e. $\theta_{mk} = 0 \Leftrightarrow a_{mk} = 0$ and $\theta_{mk} \neq 0 \Leftrightarrow a_{mk} = 1$, which leads to (14f) and (14g). These two constraints highlight the limiting effect that enforces $\theta_{mk} = 0$ when $a_{mk} = 0$. In general, when $a_{mk} = 1$, θ_{mk} is unlikely to be optimized to zero, which ensures the validity of the coupling relationship between them. This coupling is both reasonable and necessary for the optimization process. Clearly, $\boldsymbol{\theta}$ exhibits a step mapping relationship with both $\mathbf{A}$ and $\mathbf{b}$. Intuitively, (14f) can be equivalently transformed into the following second-order cone

$$\frac{a_{mk}+v_{mk}}{2} \geq \left\| \left[\frac{a_{mk}-v_{mk}}{2}, \theta_{mk} \right]^T \right\|, \quad \forall m,\forall k. \tag{15}$$

Note that the introduced auxiliary variable v_{mk} represents the soft power level from the m-th AP to the k-th user, and a proper optimization of it could improve the accuracy of the solution process [28]. Additionally, constraint (14h) also characterizes the coupling relationship between AP-user association and AP sleep modes.

In summary, $\mathcal{P}_1$ involves coupling relationships between several discrete and continuous variables, which belongs to an MI non-convex optimization problem.

IV. SCA-Relaxation Algorithm for the MI Non-Convex Optimization Problem

In this section, we propose the SCA-Relaxation algorithm for solving the problem $\mathcal{P}_1$. Firstly, by relaxation and transformation, we approximate $\mathcal{P}_1$ by a new problem $\mathcal{P}_2$. Subsequently, by applying the concept of the SCA framework, $\mathcal{P}_2$ is transformed into a series of second-order cone programming (SOCP) problems $\mathcal{P}_3$ that can be solved directly.

A. Continuous Relaxation and Equivalent Transformation

To start with, regarding to the discrete optimization matirx $\mathbf{A}$ and vector $\mathbf{b}$ in $\mathcal{P}_1$, we consider the following continuous relaxation of the discrete variables [19]

$$x \in \{0,1\} \Leftrightarrow x^2 - x \geq 0, x \in [0,1]. \tag{16}$$

According to it, (14i) becomes

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\left(a_{mk}^2 - a_{mk}\right) \geq 0, \tag{17}$$

$$a_{mk} \in [0,1], \quad b_m \in [0,1], \quad \forall m,\forall k. \tag{18}$$

Note that when $\mathbf{A}$ takes discrete values, $\mathbf{b}$ will automatically be adjusted as discrete according to constraint (14h). Furthermore, to address the non-convexity of (17), we adopt the SCA framework. Let t denote iteration index and $\mathbf{A}^{(t)}$ represent $\mathbf{A}$ after t times iterations of SCA, then (17) can be approximated by invoking the first-order *Taylor* expansion as

$$Q(\mathbf{A};\mathbf{A}^{(t)}) \triangleq \sum_{m=1}^{M}\sum_{k=1}^{K}\left(2a_{mk}a_{mk}^{(t)} - a_{mk} - \left(a_{mk}^{(t)}\right)^2\right) \geq 0. \tag{19}$$

Typically, when both $\mathbf{A}$ and $\mathbf{A}^{(t)}$ values are non-binary, the summation on the left-hand side (LHS) of (19) becomes negative. Inspired by [19], we multiply this summation by a coefficient ξ and incorporate it into the objective function as a penalty term for violating the binary constraint. Since this summation is an affine function, it would not affect the convexity of the objective function. Besides, we also introduce some necessary auxiliary variables $\{\mathbf{u} \in \mathbb{R}^K, \mathbf{z} \in \mathbb{R}^M, \tilde{\mathbf{u}} \in$

$$\mathrm{SINR}_k(\boldsymbol{\eta}) = \frac{\rho_d(N-\tau_u)(\sum_{m=1}^{M}\sqrt{\eta_{mk}\gamma_{mk}})^2}{1+\rho_d(N-\tau_u)\sum_{k'\neq k}^{K}(\sum_{m=1}^{M}\sqrt{\eta_{mk'}\gamma_{mk}})^2|\boldsymbol{\psi}_k^H\boldsymbol{\psi}_{k'}|^2 + \rho_d\sum_{k'=1}^{K}\sum_{m=1}^{M}\eta_{mk'}(\beta_{mk}-\gamma_{mk})}. \tag{8}$$

$$\mathrm{SINR}_k(\boldsymbol{\theta}) = \frac{\rho_d a_g(\boldsymbol{\theta}_k^T\boldsymbol{\gamma}_{kk})^2}{1+\rho_d a_g\sum_{k'\neq k}^{K}(\boldsymbol{\theta}_{k'}^T\boldsymbol{\gamma}_{kk'})^2 + \rho_d\sum_{k'=1}^{K}\|\boldsymbol{\theta}_{k'}\odot\boldsymbol{\kappa}_k\|^2} \triangleq \frac{\mathrm{CP}_k(\boldsymbol{\theta})}{1+\mathrm{NI}_k(\boldsymbol{\theta})}. \tag{9}$$

$\mathbb{R}^K, \eta \in \mathbb{R}, p \in \mathbb{R}\}$ into $\mathcal{P}_1$, which leads to the problem $\mathcal{P}_2$ shown below

$$\mathcal{P}_2: \max_{\substack{\boldsymbol{\theta},\mathbf{A},\mathbf{b},\mathbf{V},\\ \mathbf{u},\eta,p,\tilde{\mathbf{u}},\mathbf{z}}} \quad F_\xi(\eta,\mathbf{A};\mathbf{A}^{(t)}) \triangleq \eta + \xi Q(\mathbf{A};\mathbf{A}^{(t)})$$

$$\text{s.t.} \quad \eta p \le B\sum_{k=1}^{K} u_k, \tag{20a}$$

$$p \ge \sum_{m=1}^{M}\left(\frac{\rho_d N_0 \|\boldsymbol{\theta}_m\|^2}{\alpha_m^{\mathrm{A}}}\right) + \tilde{P}_c(\mathbf{b},\mathbf{z}), \tag{20b}$$

$$\mathrm{SE}_k(\boldsymbol{\theta}) \ge u_k, \quad \forall k, \tag{20c}$$

$$u_k \ge \mathrm{SE}_{th}, \quad \forall k, \tag{20d}$$

$$\sum\nolimits_{k=1}^{K} a_{mk}\tilde{u}_k \le z_m, \quad \forall m, \tag{20e}$$

$$\tilde{u}_k \ge \mathrm{SE}_k(\boldsymbol{\theta}), \quad \forall k, \tag{20f}$$

$$0 \le z_m \le C_{\max}, \quad \forall m, \tag{20g}$$

$$(14d), (14e), (14g), (14h), (15), (18) \tag{20h}$$

with $\tilde{P}_c(\mathbf{b},\mathbf{z}) \triangleq P^{\mathrm{Fix}} + \sum_{m=1}^{M}(b_m N P_m^{\mathrm{A,chain}} + B\xi_m^{\mathrm{F}} z_m)$ and $P^{\mathrm{Fix}} \triangleq \sum_{m=1}^{M}(P_m^{\mathrm{A,fix}} + P_m^{\mathrm{F,fix}})$.

Overall, problem $\mathcal{P}_2$ is more tractable than $\mathcal{P}_1$, which enables us to approximate the non-convex constraints (20a), (20c), (20e) and (20f) with suitable surrogate functions. Here, we highlight the distinction between the auxiliary variables $\mathbf{u}$ and $\tilde{\mathbf{u}}$, which represent the lower and upper bounds of SE respectively. These two optimization variables gradually become the same during iteration process, ensuring the convergence of the following SCA-Relaxation algorithm.

B. The Proposed SCA-Relaxation Algorithm

Firstly, note that the LHS of constraints (20a) and (20e) involve the product of optimization variables, which can be equivalently rewritten as

$$(\eta+p)^2 \le \|[\eta;p]\|^2 + 2B\sum\nolimits_{k=1}^{K} u_k, \tag{21}$$

$$\sum\nolimits_{k=1}^{K}(a_{mk}+\tilde{u}_k)^2 \le \|[\mathbf{a}_m^T;\tilde{\mathbf{u}}]\|^2 + 2z_m, \forall m \tag{22}$$

with $\mathbf{a}_m = [a_{m1}, a_{m2}, \cdots, a_{mK}]$ denoting the m-th row of $\mathbf{A}$. Intuitively, both the right-hand side (RHS) of (21) and (22) contain convex squared ℓ_2-norm terms. When these components appear on the RHS of inequalities, their affine lower bounds can be derived using the first-order *Taylor* expansion. This allows the application of SCA by utilizing the results of the previous iterations. Specifically, (21) and (22) can be further approximated by

$$(\eta+p)^2 \le 2[\eta^{(t)},p^{(t)}][\eta;p] - \|[\eta^{(t)};p^{(t)}]\|^2 + 2B\sum\nolimits_{k=1}^{K} u_k, \tag{23}$$

$$\begin{aligned}\sum\nolimits_{k=1}^{K}(a_{mk}+\tilde{u}_k)^2 \le{}& 2[(\mathbf{a}_m^{(t)})^T,\tilde{\mathbf{u}}^{(t)}]^T[\mathbf{a}_m^T;\tilde{\mathbf{u}}] \\ &- \|[(\mathbf{a}_m^{(t)})^T;\tilde{\mathbf{u}}^{(t)}]\|^2 + 2z_m, \forall m,\end{aligned} \tag{24}$$

which belong to the standard SOC constraints.

Next, as for the constraint (20c), it is challenging to find a tight concave surrogate function for the SE expression under FBL communications. Fortunately, the following lower bound can be employed [3]–[5]

$$\mathrm{SE}_k(\boldsymbol{\theta}) \ge \frac{\tilde{\tau}_d}{\ln 2} R_k\left(\frac{1}{\mathrm{SINR}_k(\boldsymbol{\theta})}\right), \tag{25}$$

where $R_k(x) \triangleq \ln(1+\frac{1}{x}) - \frac{Q^{-1}(\epsilon_k)}{\sqrt{T\tilde{\tau}_d}}\sqrt{\frac{2x+1}{(x+1)^2}}$ is a convex function. Based on the decreasing monotonicity of $R_k(x)$[1], we introduce auxiliary variable $\mathbf{g} \in \mathbb{R}_+^K$ to transfer (20c) as

$$\frac{\mathrm{CP}_k(\boldsymbol{\theta})}{g_k} \ge 1 + \mathrm{NI}_k(\boldsymbol{\theta}), \quad \forall k, \tag{26}$$

$$R_k\left(\frac{1}{g_k}\right) \ge \frac{\ln 2}{\tilde{\tau}_d} u_k, \quad \forall k, \tag{27}$$

where $\mathrm{CP}_k(\boldsymbol{\theta})$ and $\mathrm{NI}_k(\boldsymbol{\theta})$ are defined in (9). Since the LHS of (26) is a convex quadratic-over-linear function, it can be approximated by its first-order *Taylor* expansion to obtain a tight lower bound. Therefore, (26) can be replaced by the following SOC constraint

$$G_1(\boldsymbol{\theta}, g_k; \boldsymbol{\theta}^{(t)}, g_k^{(t)}) \ge 1 + \mathrm{NI}_k(\boldsymbol{\theta}), \forall k \tag{28}$$

with

$$\begin{aligned}G_1(\boldsymbol{\theta}, g_k; \boldsymbol{\theta}^{(t)}, g_k^{(t)}) \triangleq{}& \frac{\mathrm{CP}_k(\boldsymbol{\theta}^{(t)})}{g_k^{(t)}} - \frac{\mathrm{CP}_k(\boldsymbol{\theta}^{(t)})}{g_k^{(t)}}\left(\frac{g_k - g_k^{(t)}}{g_k^{(t)}}\right) \\ &+ 2\frac{\rho_d a_g}{g_k^{(t)}}(\boldsymbol{\theta}_k^{(t)})^T(\boldsymbol{\gamma}_{kk}\boldsymbol{\gamma}_{kk}^T)(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{(t)}).\end{aligned} \tag{29}$$

Similarly, [3, Lemma 1] provides a lower bound for the LHS of (27), allowing us to reformulate (27) as the following tractable constraint

$$\lambda_1 + \lambda_2\left(1 - \frac{g_k^{(t)}}{g_k}\right) \ge \frac{\ln 2}{\tilde{\tau}_d} u_k, \quad \forall k \tag{30}$$

with $\lambda_1 \triangleq \ln g_k^{(t)} - \frac{Q^{-1}(\epsilon_k)}{\sqrt{T\tilde{\tau}_d}}\sqrt{1-\frac{1}{(g_k^{(t)})^2}}$ and $\lambda_2 \triangleq 1 - \frac{Q^{-1}(\epsilon_k)}{\sqrt{T\tilde{\tau}_d}}\left(\left(g_k^{(t)}\right)^2\sqrt{1-\frac{1}{(g_k^{(t)})^2}}\right)^{-1}$. Furthermore, the above constraint can be rearranged into a rotated SOC as follows

$$\begin{aligned}&\frac{g_k + (\lambda_1+\lambda_2) - \frac{\ln 2}{\tilde{\tau}_d}u_k}{2} \\ &\ge \left\|\left[\sqrt{\lambda_2 g_k^{(t)}}, \frac{g_k - (\lambda_1+\lambda_2) + \frac{\ln 2}{\tilde{\tau}_d}u_k}{2}\right]^T\right\|, \forall k.\end{aligned} \tag{31}$$

Regarding constraint (20f), to tackle its non-convexity, we introduce another auxiliary variable $\mathbf{c} \in \mathbb{R}_+^K$, which transform the original constraint into the following form

$$\begin{aligned}\tilde{u}_k \ge{}& \tilde{\tau}_d \log_2(1+c_k) \\ &- \sqrt{\frac{\tilde{\tau}_d}{T}}\log_2(e)Q^{-1}(\epsilon_k)\sqrt{1-\frac{1}{(1+c_k)^2}}, \quad \forall k,\end{aligned} \tag{32}$$

$$c_k \ge \mathrm{SINR}_k(\boldsymbol{\theta}) = \frac{\mathrm{CP}_k(\boldsymbol{\theta})}{1+\mathrm{NI}_k(\boldsymbol{\theta})}, \quad \forall k. \tag{33}$$

[1]The monotonicity and convexity of $R_k(x)$ exist within a certain interval, which is typically satisfied in standard systems [4, Lemma 1].

Moreover, in order to transform (32) into a convex constraint, we need to derive a convex upper bound function for the RHS, where the most challenging part is to obtain the concave lower bound function for $\sqrt{1-\frac{1}{(1+c_k)^2}}$. To address this issue, we present the following Lemma.

Lemma 1. *For given $\tilde{x} > 0$ and $\forall x > 0$, the following tight concave lower bound inequality holds:*

$$F_1(x) \triangleq \sqrt{1-\frac{1}{(1+x)^2}} \geq F_2(x;\tilde{x}) \triangleq \sqrt{1-\frac{1}{(1+\tilde{x})^2}} \times \left[1+\frac{1}{2}f_1(x;\tilde{x})-\frac{1}{2}f_2(x;\tilde{x})-\frac{1}{2}\ln\left(1-\frac{1}{(1+\tilde{x})^2}\right)\right], \tag{34}$$

where

$$f_1(x;\tilde{x}) \triangleq \ln(2\tilde{x}+\tilde{x}^2)+2-\frac{\tilde{x}}{x}-\frac{\tilde{x}+2}{x+2}, \tag{35}$$

$$f_2(x;\tilde{x}) \triangleq \ln\left[(\tilde{x}+1)^2\right]+\frac{(x+1)^2}{(\tilde{x}+1)^2}-1. \tag{36}$$

Proof: Please see Appendix A. ∎

Based on Lemma 1 and *Taylor* expansion of the logarithmic function, the fronthaul link capacity non-convex constraint (32) can be approximated as

$$\tilde{u}_k \geq \frac{\tilde{\tau}_d}{\ln 2}\left(\ln(1+c_k^{(t)})+\frac{c_k-c_k^{(t)}}{1+c_k^{(t)}}\right) - \sqrt{\frac{\tilde{\tau}_d}{T}}\log_2(e)Q^{-1}(\epsilon_k)F_2(c_k;c_k^{(t)}), \quad \forall k. \tag{37}$$

In addition, (33) can be represented as $1+\mathrm{NI}_k(\boldsymbol{\theta}) \geq \frac{\mathrm{CP}_k(\boldsymbol{\theta})}{c_k}, \forall k$, where both sides of the inequality are convex functions. Similar to (28), by applying *Taylor* expansion to the LHS, we can convert it into the standard SOC form, where the following convex constraint can be obtained

$$\frac{G_2(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})+c_k}{2} \geq \left\| \left[\frac{G_2(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})-c_k}{2}, \sqrt{\mathrm{CP}_k(\boldsymbol{\theta})}\right]^T \right\|, \forall k \tag{38}$$

with

$$G_2(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) \triangleq 1+\mathrm{NI}_k(\boldsymbol{\theta}^{(t)})+\sum_{k'=1}^{K} 2\rho_d(\boldsymbol{\theta}_{k'}^{(t)})^T\tilde{\boldsymbol{\kappa}}_k(\boldsymbol{\theta}_{k'}-\boldsymbol{\theta}_{k'}^{(t)}) + \sum_{k'\neq k}^{K} 2\rho_d a_g(\boldsymbol{\theta}_{k'}^{(t)})^T(\boldsymbol{\gamma}_{kk'}\boldsymbol{\gamma}_{kk'}^T)(\boldsymbol{\theta}_{k'}-\boldsymbol{\theta}_{k'}^{(t)}), \tag{39}$$

and $\tilde{\boldsymbol{\kappa}}_k \triangleq \mathrm{diag}(\beta_{1k}-\gamma_{1k},\ldots,\beta_{Mk}-\gamma_{Mk}) \in \mathbb{R}_+^{M\times M}$.

Finally, since all non-convex constraints have been relaxed, $\mathcal{P}_2$ can be reformulated as the following SOCP problem, which is named as $\mathcal{P}_3$

$$\mathcal{P}_3: \max_{\substack{\boldsymbol{\theta},\mathbf{A},\mathbf{b},\mathbf{V},\mathbf{u},\\ \eta,p,\mathbf{z},\tilde{\mathbf{u}},\mathbf{g},\mathbf{c}}} \quad F_\xi(\eta,\mathbf{A};\mathbf{A}^{(t)}) \quad \text{s.t.} \quad (20b),(20d),(20g),(20h),(23),(24),(28),(31),(37),(38).$$

Clearly, $\mathcal{P}_3$ can be directly solved by modern SOCP solvers, such as Mosek and Gurobi [28]. To summarize, the proposed SCA-Relaxation algorithm is outlined in Algorithm 1, where the initial point can be obtained by using a low-complexity heuristic approach based on the channel conditions, i.e. $\theta_{mk}^{(0)} = \frac{\gamma_{mk}}{\sum_{k'=1}^{K}\gamma_{mk'}}, \forall m, \forall k$ [35]. Specifically, this approach allocates the initial power based on the weights derived from the estimated LSFC, and then computes $\mathbf{A}^{(0)}$ by exploiting the step mapping relationship among $\mathbf{A}$ and $\boldsymbol{\theta}$. The initial values of the remaining variables $\eta^{(0)}, p^{(0)}, \tilde{\mathbf{u}}^{(0)}, \mathbf{g}^{(0)}, \mathbf{c}^{(0)}$ can be readily obtained by setting the constraints in problem $\mathcal{P}_2$ to equality. As for the convergence, following a similar analytical approach in [21, Appendix B], the proposed SCA-Relaxation algorithm is expected to converge to at least a *Fritz John* point of $\mathcal{P}_2$, where the proof is omitted here.

Algorithm 1: The SCA-Relaxation algorithm for $\mathcal{P}_2$

Input : $\{\eta^{(0)}, p^{(0)}, \mathbf{A}^{(0)}, \tilde{\mathbf{u}}^{(0)}, \boldsymbol{\theta}^{(0)}, \mathbf{g}^{(0)}, \mathbf{c}^{(0)}\}$, initial penalty coefficient: $\xi = 2$, maximum penalty coefficient: $\xi_{\max} = 128$, tolerance: $\varepsilon = 10^{-3}$

repeat // Penalty loop
  repeat // SCA loop
    Solve $\mathcal{P}_3$ with SOCP solvers to obtain $\{\eta^\star, p^\star, \mathbf{A}^\star, \mathbf{u}^\star, \boldsymbol{\theta}^\star, \tilde{\mathbf{u}}^\star, \mathbf{g}^\star, \mathbf{b}^\star, \mathbf{V}^\star, \mathbf{z}^\star, \mathbf{c}^\star\}$;
    Update $\eta^{(t)}, p^{(t)}, \mathbf{A}^{(t)}, \tilde{\mathbf{u}}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{g}^{(t)}, \mathbf{c}^{(t)}$;
    $t = t+1$;
  until $|F_\xi(\eta^{(t)},\mathbf{A}^{(t)};\mathbf{A}^{(t-1)}) - F_\xi(\eta^{(t-1)},\mathbf{A}^{(t-1)};\mathbf{A}^{(t-2)})| \leq \varepsilon$;
  Update $\xi = \min\{\xi_{\max}; 2\times\xi\}$;
until $|Q(\mathbf{A}^{(t)};\mathbf{A}^{(t-1)}) - Q(\mathbf{A}^{(t-1)};\mathbf{A}^{(t-2)})| \leq \varepsilon$;

Output: Stationary solutions: $\boldsymbol{\theta}^\star, \mathbf{A}^\star, \mathbf{b}^\star$

We now go through the complexity of the proposed SCA-Relaxation algorithm. Specifically, $\mathcal{P}_3$ consists of $A_v = 3KM + 4K + 2M + 2$ real-valued variables, $A_l = 4KM + M + 2K$ linear constraints and $A_s = KM + 2M + 3K + 2$ SOC constraints. Therefore, the computational complexity per iteration for solving $\mathcal{P}_3$ by the modern SOCP solvers is $\mathcal{O}\left(\sqrt{A_l+A_s}\left(A_v+A_l+A_s\right)A_v^2\right) \approx \mathcal{O}\left(\sqrt{MK+M+K}M^3K^3\right)$. Overall, compared to the state-of-the-art EE optimization algorithms designed for CF-mMIMO-enabled URLLC downlink systems [3], the proposed SCA-Relaxation significantly reduces the computational complexity through continuous relaxation, thereby avoiding the use of MI-SOCP solvers (complexity order $\mathcal{O}\left(2^{MK}\left(M^2K^2\left(M+K\right)\right)\right)$). More importantly, by explicitly considering the AP sleep mode impacts and fronthaul overhead on EE, which are neglected in [3], the proposed SCA-Relaxation algorithm reduces operational power consumption and achieves notable improvements in EE.

V. LCEE-max Algorithm for Reduced Computational Complexity

In this section, by eliminating the discreteness through variables coupling relationship and utilizing first-order information, LCEE-max algorithm is proposed, which achieves much lower complexity than SCA-Relaxation algorithm.

A. Establishment of Approximated Continuous Problem

As shown in (14), the original optimization problem $\mathcal{P}_1$ involves coupled variables, i.e. $\boldsymbol{\theta}, \mathbf{A}$ and $\mathbf{b}$, which present a step mapping relationship. This interdependence not only introduces the challenging joint non-convexity to the problem but also brings additional constraints (14e)-(14h).

To efficiently address the joint non-convexity problem, we leverage the coupling relationship between optimization variables and apply the following smooth characterization function

$$\sigma(\mathbf{x}) = \frac{(\|\mathbf{x}\|)^{\varpi}}{(\|\mathbf{x}\|)^{\varpi} + \delta}, \quad \mathbf{x} \in \mathbb{R}^n, \tag{40}$$

where $\varpi \geq 1$ controls the smoothness of $\sigma(\mathbf{x})$, and δ is a small positive number. For a better understanding, the impacts of ϖ and δ upon the function in (40) is illustrated in Fig. 2 [2]. Clearly, $\sigma(\mathbf{x})$ serves as a smoothed version of the step function, approximating an output that nearly converges to either 0 or 1 based on the magnitude of the input variable ℓ_2-norm. Therefore, $\mathbf{A}$ and $\mathbf{b}$ can be approximated by leveraging the element-wise correspondence of $\boldsymbol{\theta}$ and the norm-based relationship of the associated vectors, respectively

$$a_{mk} \approx \sigma_1(\theta_{mk}) = \frac{(\theta_{mk})^2}{(\theta_{mk})^2 + \delta_1}, \quad \forall m, \forall k, \ \delta_1 > 0 \tag{41}$$

$$b_m \approx \sigma_2(\boldsymbol{\theta}_m) = \frac{(\|\boldsymbol{\theta}_m\|)^2}{(\|\boldsymbol{\theta}_m\|)^2 + \delta_2}, \quad \forall m, \ \delta_2 > 0. \tag{42}$$

From them, the power consumption can be approximated as

$$\begin{aligned} P_{\text{total}}(\boldsymbol{\theta}, \mathbf{A}, \mathbf{b}) \approx P_{\text{alt}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{\rho_d N_0 \|\boldsymbol{\theta}_m\|^2}{\alpha_m^{\text{A}}} + P^{\text{Fix}} \\ + \sum_{m=1}^{M} \left(\sigma_2(\boldsymbol{\theta}_m) N P_m^{\text{A,chain}} + B \xi_m^{\text{F}} \sum_{k=1}^{K} \sigma_1(\theta_{mk}) \text{SE}_k(\boldsymbol{\theta}) \right) \end{aligned} \tag{43}$$

which leads to the following optimization problem $\mathcal{P}_4$

$$\mathcal{P}_4: \quad \max_{\boldsymbol{\theta}} \quad \overline{\text{EE}}(\boldsymbol{\theta}) \triangleq \frac{B \sum_{k=1}^{K} \text{SE}_k(\boldsymbol{\theta})}{P_{\text{alt}}(\boldsymbol{\theta})} \tag{44a}$$

$$\text{s.t.} \quad \sum\nolimits_{k=1}^{K} \sigma_1(\theta_{mk}) \text{SE}_k(\boldsymbol{\theta}) \leq C_{\max}, \forall m, \tag{44b}$$

$$\|\boldsymbol{\theta}_m\|^2 \leq 1, \quad \forall m, \tag{44c}$$

$$(14b), (14d). \tag{44d}$$

Related to $\mathcal{P}_1$, $\mathcal{P}_4$ eliminates both of the discrete variables and the coupling constraints. Nevertheless, it still belongs to a challenging non-convex fractional optimization problem.

Next, to seek for a feasible solution form, we employ the classical *Dinkelbach* transformation (DT) to transfer $\mathcal{P}_4$ to $\mathcal{P}_5$

$$\mathcal{P}_5: \quad \max_{\boldsymbol{\theta}} \quad \Omega_{\vartheta^{(i)}}(\boldsymbol{\theta}) \triangleq B \sum_{k=1}^{K} \text{SE}_k(\boldsymbol{\theta}) - \vartheta^{(i)} P_{\text{alt}}(\boldsymbol{\theta}) \tag{45a}$$

$$\text{s.t.} \quad (44b), (44c), (44d). \tag{45b}$$

During the entire iterative process, $\vartheta^{(i)}$ will be iteratively updated. Specifically, by letting $\boldsymbol{\theta}^{(i)}$ represent the obtained solution during the i-th DT iteration of $\mathcal{P}_5$, $\vartheta^{(i+1)}$ is updated according to

$$\vartheta^{(i+1)} = \frac{B \sum_{k=1}^{K} \text{SE}_k(\boldsymbol{\theta}^{(i)})}{P_{\text{alt}}(\boldsymbol{\theta}^{(i)})}. \tag{46}$$

Theorem 1. *For given $\vartheta^{(i)}$, if the optimized objective function of $\mathcal{P}_5$ is zero, i.e.,*

$$\max_{\boldsymbol{\theta} \in (45b)} \Omega_{\vartheta^{(i)}}(\boldsymbol{\theta}) = 0, \tag{47}$$

then $\vartheta^{(i)}$ satisfies the following equation

$$\vartheta^{(i)} = \frac{B \sum_{k=1}^{K} \text{SE}_k(\boldsymbol{\theta}^{\star})}{P_{\text{alt}}(\boldsymbol{\theta}^{\star})}, \tag{48}$$

where $\boldsymbol{\theta}^{\star}$ is the optimal solution of $\mathcal{P}_4$.

Proof: Please refer to [23, Appendix B]. ■

The above theorem indicates that the optimal solution of the univariate equation (47) coincides with the optimal solution of $\mathcal{P}_4$, providing the foundation and termination criterion for the iterative solution of $\mathcal{P}_4$. Additionally, one of the main advantages of adopting DT is that it allows for an easy conversion of $\mathcal{P}_5$ into a weighted sum optimization problem between throughput and power consumption. As a result, the subsequent proposed algorithm can be conveniently extended to solve either the weighted sum rate maximization problem or the power minimization problem [6], [34].

B. The Proposed LCEE-max Algorithm

We now consider employing the APG-AM algorithm to solve $\mathcal{P}_5$, which is designed to cope with the unconstrained optimization problem [37]. Among all the constraints in $\mathcal{P}_5$, (14b) and (44b) are non-convex, whereas (14d) and (44c) are convex. Therefore, we represent the former using quadratic penalty function and equivalently transform the latter by an non-smooth indicator function, which results in the following auxiliary function

$$Q_1(\boldsymbol{\theta}) \triangleq \sum_{k=1}^{K} \left[\max\left(0, \text{SE}_{th} - \text{SE}_k(\boldsymbol{\theta})\right) \right]^2, \tag{49}$$

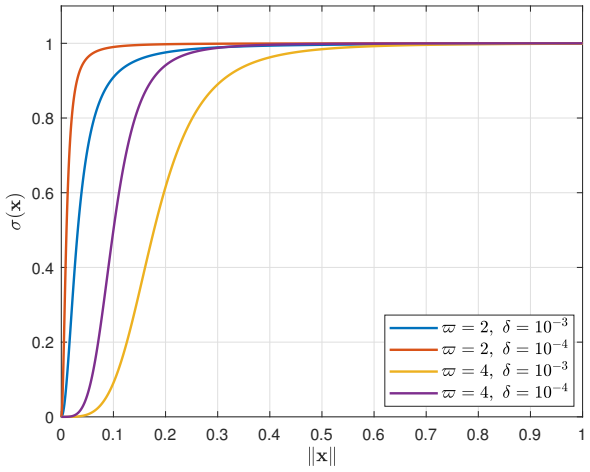


Fig. 2. The proposed smooth characterization function.

[2] As δ decreases, the curve fitting improves but leads to a large gradient near 0, similar to gradient explosion in ML, which may impact the APG-AM algorithm. The choice of δ will be discussed in Section VI. Besides, for the convenience of gradient computation within appropriate approximation, we use $\varpi = 2$ to complete the subsequent algorithm.

$$Q_2(\boldsymbol{\theta}) \triangleq \sum_{m=1}^{M}\left[\max(0, \sum\nolimits_{k=1}^{K}\sigma_1(\theta_{mk})\mathrm{SE}_k(\boldsymbol{\theta}) - C_{\max}\right]^2, \quad (50)$$

$$g(\boldsymbol{\theta}) \triangleq \begin{cases} 0, & \text{if} \quad \boldsymbol{\theta} \in \mathcal{C} \triangleq \{(14\text{d}), (44\text{c})\} \\ -\infty, & \text{otherwise.} \end{cases} \quad (51)$$

In this way, we can reformulate $\mathcal{P}_5$ into the following unconstrained optimization problem $\mathcal{P}_6$

$$\mathcal{P}_6: \quad \max_{\boldsymbol{\theta}\in\mathbb{R}^n} F_{\vartheta^{(i)},\chi}(\mathbf{x}) \triangleq f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}) + g(\boldsymbol{\theta}) \quad (52)$$

with

$$f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}) \triangleq \Omega_{\vartheta^{(i)}}(\boldsymbol{\theta}) - \chi[\mu_1 Q_1(\boldsymbol{\theta}) + \mu_2 Q_2(\boldsymbol{\theta})]. \quad (53)$$

Here, χ is the penalty coefficient, μ_1 and μ_2 are the given scaling coefficients to ensure that the penalty function and the objective function have proper magnitudes [32]. Undoubtedly, the choice of χ is crucial for solving $\mathcal{P}_6$. If χ is too small, the feasibility of the problem is not guaranteed. Conversely, if χ is too large, the optimization problem may be numerically ill-conditioned [21]. To this end, we employ a dynamical strategy, namely, when the penalty function exceeds a given tolerance, χ will be increased and then the problem $\mathcal{P}_6$ is re-solved.

For the transformed problem $\mathcal{P}_6$, we now describe the solution process of the APG-AM algorithm in detail. Specifically, let t denote the current APG-AM iteration index. In addition to the optimization variable $\boldsymbol{\theta}$ that we aim to obtain, the APG-AM algorithm also introduces auxiliary variables $\mathbf{Z} \in \mathbb{R}^{M\times K}$, and updates the optimization variables according to the following criteria

$$\boldsymbol{\theta}^{(t)} = \mathrm{prox}_{\alpha^{(t)},g}\left(\mathbf{Z}^{(t)} + \alpha^{(t)}\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})\right), \quad (54)$$

where $\alpha^{(t)}$ represents the step size at the t-th iteration and $\mathrm{prox}_{\alpha^{(t)},g}$ is the so-called proximal operator defined as [3]

$$\mathrm{prox}_{\alpha^{(t)},g}(\mathbf{X}) = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^{M\times K}}\left(g(\boldsymbol{\theta}) - \frac{1}{2\alpha^{(t)}}\|\boldsymbol{\theta}-\mathbf{X}\|^2\right). \quad (55)$$

In particular, the proximal operator here differs from the definitions in [29] and [37] by a negative sign, since we consider a maximization problem in this paper. However, they are essentially equivalent. Combining (54) into the above calculation, the following equivalent form can be derived

$$\begin{aligned}\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^{M\times K}}\Bigg\{&\left\langle\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}), \left(\boldsymbol{\theta}-\mathbf{Z}^{(t)}\right)\right\rangle_F \\ &+ g(\boldsymbol{\theta}) + f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}) - \frac{1}{2\alpha^{(t)}}\|\boldsymbol{\theta}-\mathbf{Z}^{(t)}\|^2\Bigg\}.\end{aligned} \quad (56)$$

Clearly, the RHS of (56) approximates the smooth term $f_{\vartheta^{(i)},\chi}(\cdot)$ at point $\mathbf{z}^{(t)}$ and keeps the non-smooth term $g(\cdot)$ unchanged. Let L_f denote the constant *Lipschitz* of $\nabla f_{\vartheta^{(i)},\chi}(\cdot)$, when $\alpha^{(t)} \in (0, \frac{1}{L_f}]$, the maximized term on the RHS of (56) is a *Lipschitz* lower bound of the objective function in (52), such that the APG-AM algorithm obeys the convergence criteria of the Majorization-Minimization (MM) framework [29], [39]. In particular, by utilizing the defined indicator function $g(\cdot)$, the proximal operator can be simplified into the following projection operation

$$\begin{aligned}\mathrm{prox}_{\alpha^{(t)},g}(\mathbf{X}) &= \arg\max_{\boldsymbol{\theta}\in\mathcal{C}}\left(-\frac{1}{2\alpha^{(t)}}\|\boldsymbol{\theta}-\mathbf{X}\|^2\right) \\ &= \arg\min_{\boldsymbol{\theta}\in\mathcal{C}}\|\boldsymbol{\theta}-\mathbf{X}\|^2,\end{aligned} \quad (57)$$

which is a standard convex optimization problem. Besides, it can be decomposed into M sub-problems, which are then solved in a parallel manner [26]. More specifically, the closed-form solution for each sub-problem is as follows [24], [31]

$$\boldsymbol{\theta}_m = \frac{1}{\max\left(\|[\mathbf{x}_m]_+\|, 1\right)}[\mathbf{x}_m]_+, \quad \forall m, \quad (58)$$

where $[\mathbf{x}_m]_+ \triangleq [\max(0, x_{m1}), \ldots, \max(0, x_{mK})] \in \mathbb{R}^{1\times K}$.

To further accelerate the convergence, APG-AM adjusts the optimization variables as follows[4]

$$\mathbf{V}^{(t)} = \boldsymbol{\theta}^{(t)} + \beta\left(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right), \quad (59)$$

$$\begin{cases}\text{If } f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) \geq f_{\vartheta^{(i)},\chi}(\mathbf{V}^{(t)}) \\ \qquad \mathbf{Z}^{(t+1)} = \boldsymbol{\theta}^{(t)}, \beta = \zeta_1 \cdot \beta \\ \text{If } f_{\vartheta^{(i)},\chi}(\mathbf{V}^{(t)}) > f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) \\ \qquad \mathbf{Z}^{(t+1)} = \mathbf{V}^{(t)}, \beta = \min\{\frac{\beta}{\zeta_1}, 1\}.\end{cases} \quad (60)$$

Here, β denotes the momentum parameter and $\zeta_1 \in (0,1)$ is the scaling factor. Besides, $\mathbf{V} \in \mathbb{R}^{M\times K}$ is an extrapolated variable. When its corresponding function value is larger, i.e. $f_{\vartheta^{(i)},\chi}(\mathbf{V}^{(t)}) > f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)})$, the momentum is increased to further exploit the opportunity for accelerating.

In a nutshell, we transform the fractional problem $\mathcal{P}_4$ into a more tractable affine form and then handle the non-convex constraints using the penalty function method. Given fixed $\vartheta^{(i)}$ and χ, APG-AM algorithm is employed to efficiently solve the optimization problem involving a non-smooth indicator function. To summarize, the proposed LCEE-max algorithm for solving the approximated continuous problem $\mathcal{P}_4$ is outlined in Algorithm 2. Moreover, as shown in Algorithm 2, the computation of the gradient information $f_{\vartheta^{(i)},\chi}(\cdot)$ is required.

Proposition 1. *The closed-form expression of $\nabla f_{\vartheta^{(i)},\chi}(\cdot)$ can be efficiently obtained by the chain rule and the gradient rule for multi-variable function, which has $\mathcal{O}(M^2K^2)$ computational complexity order.*

Proof: The computation process of $\nabla f_{\vartheta^{(i)},\chi}(\cdot)$ and its complexity analysis are detailed in Appendix B. ■

[3]Strictly speaking, the norm in this expression should be the matrix Frobenius norm. For simplicity, this paper does not distinguish between the Frobenius norm and the ℓ_2-norm, as the Frobenius norm can be equivalently transformed into the ℓ_2-norm by vectorizing the matrix.

[4]The adopted APG-AM algorithm differs significantly from the PG methods employed in [24], [30]. Specifically, [24] requires two proximal operations during per iteration, which may be redundant in some RA problems and thus lead to unnecessary computational overhead. On the other hand, the PG algorithms in [30] does not incorporate any acceleration strategies, which also results in increased overall complexity.

Algorithm 2: The LCEE-max algorithm for $\mathcal{P}_4$

Input : $\vartheta^{(0)}=0, \mathbf{z}^{(1)}=\boldsymbol{\theta}^{(0)}, \mu_1=10^3, \mu_2=10, \beta=2, \zeta_1=0.5, \chi=1, \zeta_2=10, i=0, t=1, \varepsilon=10^{-3}, L_0=10^{-1}$

repeat // DT loop
  repeat // Penalty loop
    repeat // APG-AM loop
      Update $\boldsymbol{\theta}^{(t)}$ via (54) and $\mathbf{V}^{(t)}$ via (59);
      Update $\mathbf{Z}^{(t+1)}$ and β via (60);
      if $|F_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) - F_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)})| > \varepsilon$ **then**
        $t=t+1$;
      end
    until $|F_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) - F_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)})| \le \varepsilon$;
    Increase the penalty coefficient $\chi=\zeta_2\cdot\chi$;
  until $|Q_1(\boldsymbol{\theta}^{(t)}) - Q_1(\boldsymbol{\theta}^{(t-1)})| \le \varepsilon$ as well as $|Q_2(\boldsymbol{\theta}^{(t)}) - Q_2(\boldsymbol{\theta}^{(t-1)})| \le \varepsilon$;
  $\boldsymbol{\theta}^{(i)}=\boldsymbol{\theta}^{(t)}$;
  Update $\vartheta^{(i+1)}$ via (46) and $i=i+1$;
until $|\Omega_{\vartheta^{(i)}}(\boldsymbol{\theta}^{(i)}) - \Omega_{\vartheta^{(i-1)}}(\boldsymbol{\theta}^{(i-1)})| \le \varepsilon$;
Output: Stationary power allocation coefficients: $\boldsymbol{\theta}^{\star}$

C. Convergence Analysis

After obtaining the closed-form solution of $\nabla f_{\vartheta^{(t)},\chi}(\cdot)$, it is necessary to verify whether $f_{\vartheta^{(t)},\chi}(\cdot)$ possesses the *Lipschitz* smoothness property. This property ensures that the gradient of the differentiable function does not exhibit excessive fluctuations, which is the core for the convergence of LCEE-max algorithm. Specifically, we arrive at the following lemma.

Lemma 2. *For* $\mathrm{SINR}_k(\cdot) \neq 0, \forall k$, $f_{\vartheta^{(i)},\chi}(\cdot)$ *achieves the* L_f *Lipschitz smoothness property, i.e. there exists a constant* $L_f>0$ *such that the following inequality holds*

$$\|\nabla f_{\vartheta^{(i)},\chi}(\mathbf{X}) - \nabla f_{\vartheta^{(i)},\chi}(\mathbf{Y})\| \le L_f\|\mathbf{X}-\mathbf{Y}\|,\ \forall \mathbf{X},\mathbf{Y}\in\mathcal{C}. \tag{61}$$

Proof: Please see Appendix C. ∎

The significance of Lemma 2 lies in providing the theoretical foundation for the convergence of the APG-AM algorithm. In practical scenarios, computing the *Lipschitz* constant remains a challenging task due to the need for a large number of eigenvalues. Therefore, we adopt a backtracking approach to dynamically adjust the step size based on Armijo-Goldstein condition [30]. Specifically, let $\alpha_0>0$ denote the initial step size. In contrast to the explicit step size given in (54), we replace the proximal operation with

$$\boldsymbol{\theta}^{(t)} = \mathrm{prox}_{\alpha_0\rho_L^{\kappa},g}\left(\mathbf{Z}^{(t)} + \alpha_0\rho_L^{\kappa}\nabla f(\mathbf{Z}^{(t)})\right), \tag{62}$$

where $\rho_L\in(0,1)$ and κ is the smallest non-negative integer satisfying the following condition

$$f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) \ge f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)}) + \delta_L\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t-1)}\|^2. \tag{63}$$

Here $\delta_L>0$ is a positive small constant to make sure sufficient ascent [26]. When the solution $\boldsymbol{\theta}^{(t)}$ generated by (62) does not satisfy the sufficient ascent condition in (63), the index κ is incremented, and (62) is repeatedly solved until (63) is fulfilled. As shown in Lemma 2, since $f_{\vartheta^{(i)},\chi}(\cdot)$ has L_f smoothness property, this step size search process would terminate in a finite number of steps [24]. By substituting the step 4 of Algorithm 2 with (62) and (63), further convergence improvement can be attained, which also avoids the complicated calculation process of the L_f constant.

Next, based on the Lemma 2, we proceed with the convergence analysis of the proposed LCEE-max algorithm, which contains three-layer iterative process. As for the innermost APG-AM loop, for a given DT parameter $\vartheta^{(i)}$ and penalty coefficient χ, we have the following convergence result.

Theorem 2. *Given the Lipschitz smoothness of* $f_{\vartheta^{(i)},\chi}(\cdot)$ *and* $\alpha^{(t)}<\frac{1}{L_f}$*, the APG-AM loop in Algorithm 2 is guaranteed to converge to a stationary point of* $\mathcal{P}_6$.

Proof: Please see Appendix D. ∎

Then, we analyze the convergence of the intermediate loop that uses the penalty method. Specifically, this convergence analysis follows the approach in [21]. When $\mathcal{P}_5$ is feasible, as χ increases, it imposes a more significant penalty gradient on $f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta})$, such that both $Q_1(\boldsymbol{\theta})$ and $Q_2(\boldsymbol{\theta})$ will approach zero. Meanwhile, $\mathcal{P}_5$ and $\mathcal{P}_6$ satisfy strong duality, i.e. $\max\limits_{\boldsymbol{\theta}\in(45b)} \Omega_{\vartheta^{(i)}}(\boldsymbol{\theta}) = \sup\limits_{\chi\ge 0}\max\limits_{\boldsymbol{\theta}\in\mathbb{R}^n} F_{\vartheta^{(i)},\chi}(\boldsymbol{\theta})$. The detailed proof can be found in [21, Appendix A]. Furthermore, the convergence of the outermost DT is guaranteed by Theorem 1. To sum up, since LCEE-max algorithm will terminate for finite $\vartheta^{(i)}$ and χ when a pre-determined error tolerance ε is met, it ultimately converges to an approximate stationary solution for $\mathcal{P}_4$.

Finally, we briefly analyze the complexity of the proposed LCEE-max algorithm. Similar to the algorithms proposed in [24]–[26], the total computational complexity of the proposed LCEE-max algorithm primarily stems from the computation of $\nabla f_{\vartheta^{(i)},\chi}(\cdot)$, which has $\mathcal{O}(M^2K^2)$ computational complexity order as shown in Proposition 1. Given its three-loop structure, the overall complexity order is $\mathcal{O}\left(I_{DT}I_PI_{APG}M^2K^2\right)$.

VI. Numerical Results

In this section, we numerically evaluate the performance of the proposed algorithms. Consider a $1\times 1\mathrm{km}^2$ area, where $M=40$ APs and $K=12$ users are randomly distributed, and the wrapped-around technique is used to imitate an infinite network [18]. The channel path loss follows the three-slope model as in [17], [24]. Specifically, the decoding error probability and channel blocklength follow the settings from 3GPP Release 15 to support scenarios such as Tactile Internet and smart factories, thus the decoding error probability ϵ_k is set to 10^{-7} and the physical layer latency $t_c=0.01\mathrm{ms}$. Therefore, with the operating bandwidth $B=20\mathrm{MHz}$, the total channel FBL is given by $T=B\times t_c=200$ symbols [3]. We assume $\tau_d=190$ and hence $\tau_u=10$. The QoS requirement is setting to $\mathrm{SE}_{th}=1\mathrm{bit/s/Hz}$ and the maximum fronthaul capacity is $C_{\max}=40\mathrm{bit/s/Hz}$. To ensure that the array gain of the FZF precoding is greater than 0, i.e., $N\ge\tau_u+1$, each AP is equipped with 14 antennas [35]. Additionally, in our simulation, the downlink transmit power is set to 1W, and the power consumption parameters is consistent with [33]: $\{\alpha_m^{\mathrm{A}}\}_{\forall m}=0.4, \{P_m^{\mathrm{A,chain}}\}_{\forall m}=0.3\mathrm{W}, \{P_m^{\mathrm{A,fix}}\}_{\forall m}=$

$0.2\text{W}, \{\xi_m^{\text{F}}\}_{\forall m} = 0.6\text{W/Gbps}, \{P_m^{\text{F,fix}}\}_{\forall m} = 0.825\text{W}$. Unless otherwise specified, all simulations use the above parameters. Specifically, the benchmark algorithms considered for comparison are as follows:

- Baseline 1 [19]: The global optimal solution of $\mathcal{P}_1$ is obtained through the BRB strategy.
- Baseline 2 [3]: Employing the MI-SOCP solver to jointly optimize power allocation and AP-user association for EE maximization.
- Baseline 3 [24]: Optimizing EE by using the monotone APG method for fully connected CF-mMIMO systems.
- Baseline 4 [34]: The MI-SOCP solver is employed to jointly optimize power allocation and AP sleep mode, with the objective of minimizing power consumption.
- Baseline 5 [32]: Using non-monotone APG method to jointly optimize power allocation and AP-user association for maximizing sum SE.
- Baseline 6 [3], [18]: First, we adopt SOCP solver to optimize the power allocation coefficients to maximize EE. Then, we compute the total power allocated to each AP and sort these values in descending order. Subsequently, we select the top M_0 APs to remain active, while for the remaining $M - M_0$ APs, we set the corresponding estimated LSFC to zero (i.e., $\gamma_{mk} = 0, \forall m = M - M_0 + 1, \ldots, M, \forall k = 1, \ldots, K$) and then re-solve the optimization problem.

In Fig. 3, we show the convergence behavior of the proposed SCA-Relaxation algorithm under different system scales. Obviously, this algorithm converges to a high performance suboptimal solution within a few iterations. Moreover, since the optimal BRB strategy is only feasible for small-scale systems, we compare the SCA-Relaxation algorithm with Baseline 1 under $M = 8, K = 2$ scenario. It can be observed that the SCA-Relaxation algorithm converges to the near optimal point within approximately eight iterations. This behavior proves that the proposed algorithm is fast convergent and effective method. Furthermore, it can be observed that with a fixed number of APs, an increase in the number of users also improves the system EE. This indicates that in URLLC scenarios, when sufficient service resources are available, accommodating more devices would achieve better green communication paradigm.

Similarly, Fig. 4 illustrates the convergence behavior of the LCEE-max algorithm, where the iteration index represents the update steps of the outermost DT cycle. Given the lower complexity of the LCEE-max algorithm, we extend the comparison to larger system scales (i.e., $M = 80, K = 20$). Note that in large-scale systems, the number of iterations required by the algorithm may increase. This is because the non-convexity of the problem becomes more pronounced, and the LCEE-max algorithm may require more momentum-based extrapolation to explore high-performance solutions, which corresponds to the jump observed in the blue curve around index 16→17 in Fig. 4. The results show that the LCEE-max algorithm converges rapidly to an stationary point across all considered system sizes. Obviously, since this algorithm solves the continuously approximated problem $\mathcal{P}_4$, its performance is slightly compromised and cannot achieve the near optimum

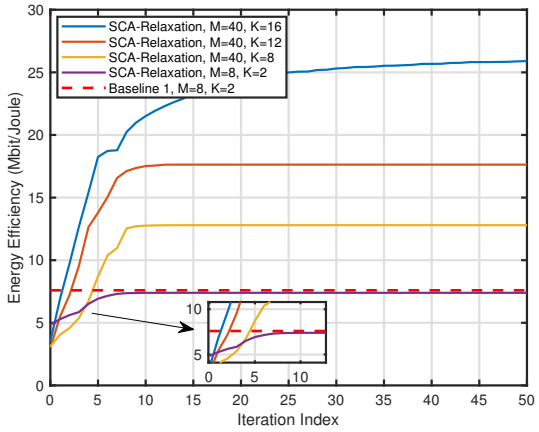


Fig. 3. The iterative convergence behaviour of the SCA-Relaxation algorithm.

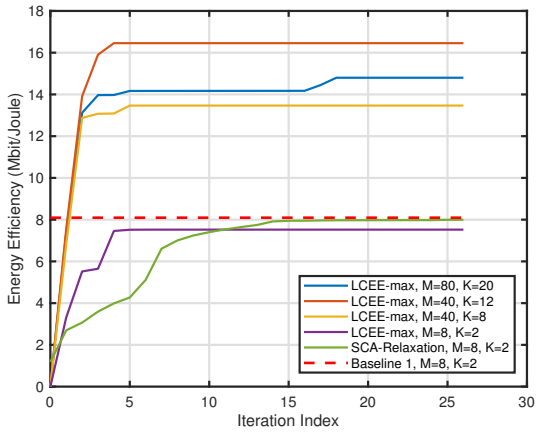


Fig. 4. The iterative convergence behavior of the LCEE-max algorithm.

in small-scale systems. However, as shown in Fig. 5, this performance loss is compensated by a significant reduction in algorithm average run time. Compared to traditional convex optimization solvers based on SOCP or MI-SOCP, the first-order APG-based algorithm can reduce run time by two orders of magnitude. Particularly for URLLC communication systems, where latency is a dominant factor, the advantage of this low-complexity algorithm becomes more pronounced.

Fig. 6 presents the variation of EE with the acceptable decoding error probability ϵ_k. It is evident that as ϵ_k increases, the value of $Q^{-1}(\epsilon_k)$ decreases, which allows the system to enhance throughout under relatively relaxed reliability requirements, resulting in a gradual increase in EE. Moreover, compared to the latest EE optimization algorithm in CF-mMIMO-enabled URLLC systems, the proposed two algorithms achieve significant average gains of 40.2% and 29.3%, respectively.

The foundation of this significant gain is further illustrated in Fig. 7, which compares the sum SE and power consumption components. Obviously, compared to Baseline 2 and Baseline 5, the proposed two algorithms have only slight losses in achievable sum SE. While significantly reduced AP operation power consumption that is approximately 45%

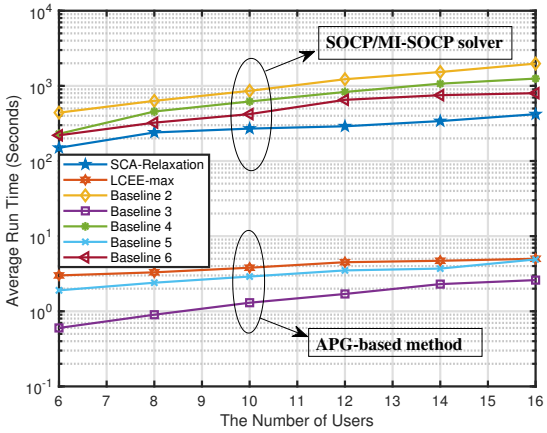


Fig. 5. Average run time versus the number of users.

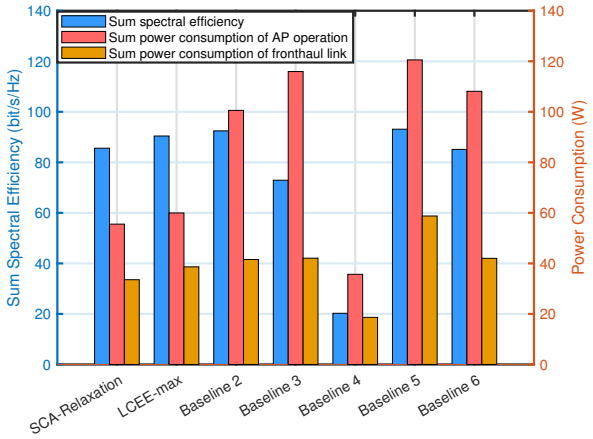


Fig. 7. Comparison of sum SE and specific power consumption components under different RA optimization algorithms.

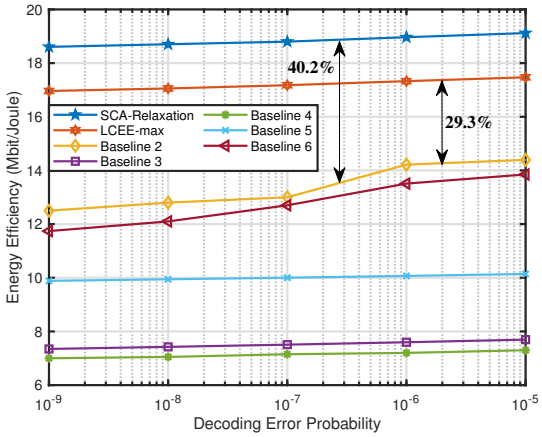


Fig. 6. EE versus different decoding error probability.

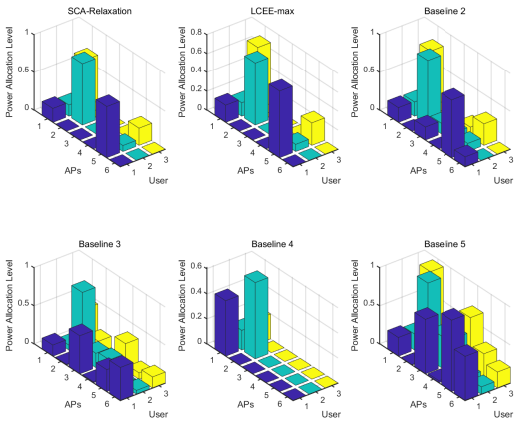


Fig. 8. Association status between APs and Users.

and 54%, respectively. This improvement is attributed to the explicit consideration of the AP sleep mode and more practical objective function. Except for Baseline 4, none of the other comparison baseline algorithms explicitly account for the impact of AP sleep mode. Although Baseline 6 heuristically optimizes AP sleep control by computing the total power allocated to each AP and deactivating those with lower allocated power, this mapping-based strategy suffers from two key limitations. First, it incurs high computational overhead, as it requires solving the optimization problem twice. Second, and more critically, its performance falls significantly short of our proposed algorithm. This performance gap stems from the heuristic nature of the approach: the initial power allocation coefficients are merely locally optimal, which can lead to the erroneous deactivation of APs that should remain active. Consequently, this results in substantial spectral efficiency degradation, as demonstrated in Fig. 7. Furthermore, since Baseline 3 does not consider AP-user association, APs cannot effectively allocate communication resources by selecting users when the fronthaul link capacity is limited, resulting in lower sum SE. In fact, since the EE optimization problem is a complex non-convex problem, without explicitly defining the AP sleep and AP-user association variables, the obtained solutions are typically prone to converge to a relatively poor local optimum.

To intuitively illustrate the communication patterns between APs and users, Fig. 8 depicts the AP-user association in a small-scale system based on the magnitude of the power allocation coefficients. As previously mentioned, the SCA-Relaxation and LCEE-max algorithms exhibit similar performance in small-scale systems, leading to nearly identical optimization strategies. Both these two algorithms deactivate $\{3, 4, 6\}$ APs, retaining only three active APs to provide service. Meanwhile, although Baseline 2 disables some AP-user association, it does not explicitly consider AP sleep modes in its objective function, resulting in no APs being turned off and causing unnecessary activation power consumption. Notably, Baseline 4 activates only two APs since its objective is to minimize total system power consumption. However, as shown in Fig. 7, this comes at the cost of significantly reduced user achievable SE. Therefore, the proposed two algorithms achieve a well-balanced trade-off between SE and system power consumption, leading to higher EE.

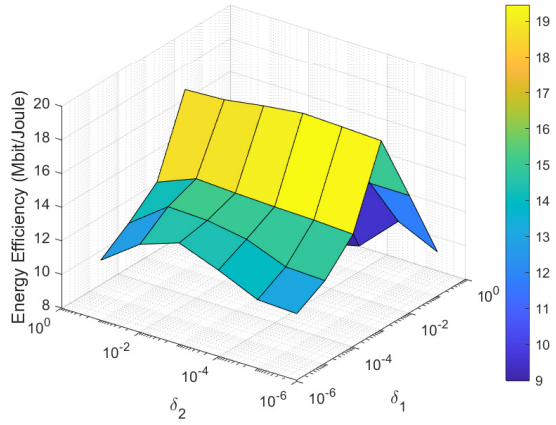


Fig. 9. EE versus the fitting parameters value of the smooth characterization function.

It is also important to emphasize that AP activation and sleep-mode transitions inevitably incur additional power consumption. However, since the optimization problem in this paper is only re-solved when the LSFC changes, each AP remains in a fixed operational state for relatively long periods [23]. As a result, we do not explicitly account for this overhead in our formulation. In contrast, such activation-related power consumption becomes non-negligible in beamforming problems capitalize instantaneous CSI, where frequent reconfigurations occur. This aspect warrants further investigation in future studies.

Fig. 9 illustrates the impact of fitting parameters in the adopted characterization function on the optimized EE. It is evident that with different fitting parameters, the optimized EE varies from 9.3Mbit/Joule to 19.4Mbit/Joule. Obviously, the values of δ_1 and δ_2 are most suitable around 10^{-3}, as this not only ensures good approximation accuracy of the characterization function while also maintaining the *Lipschitz* constant of the objective function within a reasonable range. This differs significantly from the parameter settings in [38], which arises from the sensitivity of convex optimization and the PG algorithm to gradients. Therefore, when exploiting first-order methods with approximation fitting, parameter selection requires careful consideration. In summary, the fitting parameters δ_1 and δ_2 are empirically selected to balance approximation sharpness and numerical stability, as commonly practiced in surrogate-based optimization. Exploring ML techniques to determine optimal fitting parameters could be a promising research direction for similar studies.

Finally, we highlight the trade-off between the SCA-Relaxation and LCEE-max algorithms. The former operates within SCA framework and solves each sub-problem via modern convex solvers (i.e. Gurobi), which is guaranteed to converge to at least a *Fritz John* point of $\mathcal{P}_1$, thus satisfying the second-order optimality in nonlinear programming. In contrast, LCEE-max approximates system power consumption and reformulates $\mathcal{P}_1$ as an upper-bounded problem $\mathcal{P}_4$, which is then solved by the APG-AM algorithm with convergence only to a stationary point. In practical URLLC systems, the choice depends on the system scale and latency sensitivity: SCA-Relaxation offers higher solution quality, while LCEE-max provides faster and more efficient convergence.

VII. Conclusions

In this paper, we propose two efficient algorithms to maximize downlink EE for CF-mMIMO-enabled URLLC systems. To tackle the MI non-convexity of the EE optimization problem, the SCA-Relaxation algorithm is established under SCA framework, where surrogate functions approximate non-convex constraints with convex ones, leading to a series of solvable SOCP problems. To further reduce the complexity cost of the SCA-Relaxation algorithm, we approximate the problem as a continuous-variable non-convex optimization by leveraging the proper characterization function. The *Lipschitz* smoothness of the objective is verified and a closed-form gradient expression is derived, followed by APG-AM method to obtain the final solution. Notably, the proposed LCEE-max algorithm can be extended to other RA problems, balancing communication rate and power consumption. Simulations comprehensively compare our algorithms with existing RA schemes, highlighting their superior performance and complexity trade-off in EE optimization for URLLC scenarios.

Appendix A
Proof of Lemma 1

According to the monomial inequality [39, Eq (17)], we can obtain the lower bound of $F_1(x), \forall x>0, \forall \tilde{x}>0$ as

$$F_1(x) \triangleq \sqrt{1-\frac{1}{(1+x)^2}} \geq \sqrt{1-\frac{1}{(1+\tilde{x})^2}} \times \left[1+\frac{1}{2}\ln\left(1-\frac{1}{(1+x)^2}\right)-\frac{1}{2}\ln\left(1-\frac{1}{(1+\tilde{x})^2}\right)\right]. \tag{64}$$

Based on the properties of logarithmic functions, the term containing the variable x in the above equation can be decomposed as follows

$$\ln\left(1-\frac{1}{(1+x)^2}\right) = \underbrace{\ln(x)+\ln(x+2)}_{\tilde{f}_1(x)} - \underbrace{\ln(1+2x+x^2)}_{\tilde{f}_2(x)}. \tag{65}$$

It is worth noting that although $\tilde{f}_1(x)$ is a concave function, existing convex optimization solvers actually introduce high-complexity exponential cones when handling logarithmic constraints, which makes them impractical for large-scale systems [18]. Therefore, it is necessary to derive a tight lower bound for $\tilde{f}_1(x)$ while preserving its concavity. To this end, we establish the following (in)equality relationships $\forall x>0, \forall \tilde{x}>0$

$$\begin{aligned}\tilde{f}_1(x) &= \ln\left(\frac{x}{\tilde{x}}\tilde{x}\right)+\ln\left(\frac{x+2}{\tilde{x}+2}(\tilde{x}+2)\right)\\ &= \ln(\tilde{x})+\ln\left(\frac{x}{\tilde{x}}\right)+\ln\left(\frac{x+2}{\tilde{x}+2}\right)+\ln(\tilde{x}+2)\\ &\overset{(a)}{\geq} \ln(2\tilde{x}+\tilde{x}^2)+2-\frac{\tilde{x}}{x}-\frac{\tilde{x}+2}{x+2},\end{aligned} \tag{66}$$

where (a) follows from the well-known inequality $\ln(x)\geq 1-\frac{1}{x}, \forall x>0$. In addition, the upper bound of $\tilde{f}_2$ can be obtained

based on the conclusion given in [32, Eq (36)], leading to equation (36). Therefore, we can ultimately obtain the concave lower bound of $F_1(x)$, as presented in (34). Furthermore, since the surrogate functions in the above parts are all tight, this implies that $F_1(\tilde{x}) = F_2(\tilde{x};\tilde{x})$ and $F_1'(\tilde{x}) = F_2'(\tilde{x};\tilde{x})$, which also supports the convergence of the SCA framework [39]. This completes the proof.

Appendix B
Proof of Proposition 1

Based on formulae (10), (43), (45a), and (53), the gradient of $f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta})$ can be expressed as

$$\begin{aligned}\nabla f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}) = B\sum_{k=1}^{K}\nabla \mathrm{SE}_k(\boldsymbol{\theta}) - \vartheta^{(i)}\nabla P_{\text{alt}}(\boldsymbol{\theta}) \\ - \chi[\mu_1\nabla Q_1(\boldsymbol{\theta}) + \mu_2\nabla Q_2(\boldsymbol{\theta})] \in \mathbb{R}^{M\times K},\end{aligned} \tag{67}$$

the following calculations are performed for each of these components separately.

1) The calculation of $\nabla \mathrm{SE}_k(\boldsymbol{\theta})$: according to the chain rule, the gradient expression of $\mathrm{SE}_k(\boldsymbol{\theta})$ is

$$\begin{aligned}\nabla \mathrm{SE}_k(\boldsymbol{\theta}) = \frac{\tilde{\tau}_d \log_2(e)}{1+\mathrm{SINR}_k(\boldsymbol{\theta})}\nabla \mathrm{SINR}_k(\boldsymbol{\theta}) - \\ \frac{[1+\mathrm{SINR}_k(\boldsymbol{\theta})]^{-2}}{\sqrt{[1+\mathrm{SINR}_k(\boldsymbol{\theta})]^2-1}}\frac{\log_2(e)Q^{-1}(\epsilon_k)}{\sqrt{T(\tilde{\tau}_d^{-1})}}\nabla \mathrm{SINR}_k(\boldsymbol{\theta}),\end{aligned} \tag{68}$$

according to (9), the calculation result of $\nabla \mathrm{SINR}_k(\boldsymbol{\theta})$ is

$$\nabla \mathrm{SINR}_k(\boldsymbol{\theta}) = \frac{\nabla \mathrm{CP}_k(\boldsymbol{\theta})}{1+\mathrm{NI}_k(\boldsymbol{\theta})} - \frac{\mathrm{CP}_k(\boldsymbol{\theta})\nabla \mathrm{NI}_k(\boldsymbol{\theta})}{(1+\mathrm{NI}_k(\boldsymbol{\theta}))^2}, \tag{69}$$

and

$$\nabla \mathrm{CP}_k(\boldsymbol{\theta}) = \left[\underbrace{\mathbf{0}_M,\ldots,\mathbf{0}_M}_{k-1}, \varrho_{kk}^1(\boldsymbol{\theta}), \underbrace{\mathbf{0}_M,\ldots,\mathbf{0}_M}_{K-k}\right], \tag{70}$$

$$\begin{aligned}\nabla \mathrm{NI}_k(\boldsymbol{\theta}) = \Bigg[\overbrace{\varrho_{k1}^1(\boldsymbol{\theta})+\varrho_{k1}^2(\boldsymbol{\theta}),\ldots,\varrho_{k(k-1)}^1(\boldsymbol{\theta})+\varrho_{k(k-1)}^2(\boldsymbol{\theta})}^{k-1}, \\ \underbrace{\varrho_{kk}^2(\boldsymbol{\theta}), \varrho_{k(k+1)}^1(\boldsymbol{\theta})+\varrho_{k(k+1)}^2(\boldsymbol{\theta}),\ldots,\varrho_{kK}^1(\boldsymbol{\theta})+\varrho_{kK}^2(\boldsymbol{\theta})}_{K-k}\Bigg],\end{aligned} \tag{71}$$

where $\varrho_{kk'}^1(\boldsymbol{\theta}) \triangleq 2\rho_d a_g \boldsymbol{\gamma}_{kk'}(\boldsymbol{\theta}_k^T\boldsymbol{\gamma}_{kk'}) \in \mathbb{R}^M$ and $\varrho_{kk'}^2(\boldsymbol{\theta}) \triangleq 2\rho_d\boldsymbol{\theta}_{k'} \odot [\beta_{1k}-\gamma_{1k};\ldots;\beta_{Mk}-\gamma_{Mk}] \in \mathbb{R}^M$. Thus, the calculation of $\nabla \mathrm{SE}_k(\boldsymbol{\theta})$ is completed.

2) The calculation of $\nabla P_{\text{alt}}(\boldsymbol{\theta})$: due to the complexity form of $P_{\text{alt}}(\boldsymbol{\theta})$, we divide it into the following four parts

$$P_{\text{alt}}(\boldsymbol{\theta}) = P_{\text{alt-1}}(\boldsymbol{\theta}) + P_{\text{alt-2}}(\boldsymbol{\theta}) + P_{\text{alt-3}}(\boldsymbol{\theta}) + P^{\text{Fix}}, \tag{72}$$

where $P_{\text{alt-1}}(\boldsymbol{\theta}) \triangleq \sum_{m=1}^{M}\frac{\rho_d N_0\|\boldsymbol{\theta}_m\|^2}{\alpha_m^{\text{A}}}$ represents the total AP transmit power, $P_{\text{alt-2}}(\boldsymbol{\theta}) \triangleq \sum_{m=1}^{M}\sigma_2(\boldsymbol{\theta}_m)NP_m^{\text{A,chain}}$ and $P_{\text{alt-3}}(\boldsymbol{\theta}) = \sum_{m=1}^{M}B\xi_m^{\text{F}}\left(\sum_{k=1}^{K}\sigma_1(\theta_{mk})\mathrm{SE}_k(\boldsymbol{\theta})\right)$. According to the gradient rule for multi-variable function, we can obtain the following expression

$$\nabla_{\boldsymbol{\theta}} P_{\text{alt-1}}(\boldsymbol{\theta}) = 2\rho_d N_0\left[\frac{\boldsymbol{\theta}_1^T}{\alpha_1^{\text{A}}},\ldots,\frac{\boldsymbol{\theta}_M^T}{\alpha_M^{\text{A}}}\right]^T, \tag{73}$$

$$\nabla_{\boldsymbol{\theta}} P_{\text{alt-2}}(\boldsymbol{\theta}) = N\left[\nabla\sigma_2(\boldsymbol{\theta_1})P_1^{\text{A,chain}},\ldots,\nabla\sigma_2(\boldsymbol{\theta_M})P_M^{\text{A,chain}}\right]^T, \tag{74}$$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} P_{\text{alt-3}}(\boldsymbol{\theta}) = B\begin{bmatrix}\xi_1^{\text{F}}\mathrm{SE}_1(\boldsymbol{\theta})\nabla\sigma_1(\theta_{11}) & \ldots & \xi_1^{\text{F}}\mathrm{SE}_K(\boldsymbol{\theta})\nabla\sigma_1(\theta_{1K}) \\ \vdots & \ddots & \vdots \\ \xi_M^{\text{F}}\mathrm{SE}_1(\boldsymbol{\theta})\nabla\sigma_1(\theta_{M1}) & \ldots & \xi_M^{\text{F}}\mathrm{SE}_K(\boldsymbol{\theta})\nabla\sigma_1(\theta_{MK})\end{bmatrix} \\ + \sum_{m=1}^{M}B\xi_m^{\text{F}}\left(\sum_{k=1}^{K}\sigma_1(\theta_{mk})\nabla\mathrm{SE}_k(\boldsymbol{\theta})\right),\end{aligned} \tag{75}$$

where $\nabla\sigma_1(\theta_{mk}) = \frac{2\delta_1\theta_{mk}}{(\theta_{mk}^2+\delta_1)^2}$ and $\nabla\sigma_2(\boldsymbol{\theta}_m) = \frac{2\delta_2\boldsymbol{\theta}_m^T}{(\|\boldsymbol{\theta}_m\|^2+\delta_2)^2}$.

3) The calculation of $\nabla Q_1(\boldsymbol{\theta})$ *and* $\nabla Q_2(\boldsymbol{\theta})$: based on the above gradient, it is straightforward to obtain the gradients of the two penalty functions as

$$\nabla Q_1(\boldsymbol{\theta}) = -2\sum_{k=1}^{K}\{[\max(0,\mathrm{SE}_{th}-\mathrm{SE}_k(\boldsymbol{\theta}))]\nabla\mathrm{SE}_k(\boldsymbol{\theta})\}, \tag{76}$$

$$\begin{aligned}\nabla Q_2(\boldsymbol{\theta}) = 2\sum_{m=1}^{M}\Bigg\{\left[\max\left(0,\sum\nolimits_{k=1}^{K}\sigma_1(\theta_{mk})\mathrm{SE}_k(\boldsymbol{\theta})-C_{\max}\right)\right] \\ \times\left(\nabla\varrho_m^3(\boldsymbol{\theta})+\sum_{k=1}^{K}\sigma_1(\theta_{mk})\nabla\mathrm{SE}_k(\boldsymbol{\theta})\right)\Bigg\},\end{aligned} \tag{77}$$

where

$$\begin{aligned}\nabla\varrho_m^3(\boldsymbol{\theta}) \triangleq \Bigg[\overbrace{\mathbf{0}_K,\ldots,\mathbf{0}_K}^{m-1}, [\nabla\sigma_1(\theta_{m1})\mathrm{SE}_1(\boldsymbol{\theta}),\ldots, \\ \ldots,\nabla\sigma_1(\theta_{mK})\mathrm{SE}_K(\boldsymbol{\theta})]^T, \underbrace{\mathbf{0}_K,\ldots,\mathbf{0}_K}_{M-m}\Bigg]^T.\end{aligned} \tag{78}$$

By summing the gradients of the above components, we can obtain the final gradient of $\nabla f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta})$. Next, we measure the computational complexity of the gradient calculation in terms of the number of real-valued floating point operations (flops). Specifically, In the $\nabla\mathrm{SE}_k(\boldsymbol{\theta})$ gradient computation, the dominant complexity comes from the calculation of $\nabla\mathrm{NI}_k(\boldsymbol{\theta})$ in the denominator of $\nabla\mathrm{SINR}_k(\boldsymbol{\theta})$, which requires $(M+1)(5K-3)$ flops. Therefore, computing $\nabla\mathrm{SE}_k(\boldsymbol{\theta})$ order is $\mathcal{O}(MK)$. In the total $\nabla f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta})$ gradient computation, $\nabla Q_2(\boldsymbol{\theta})$ and $P_{\text{alt-3}}(\boldsymbol{\theta})$ dominate. Since both $Q_2(\boldsymbol{\theta})$ and $P_{\text{alt-3}}(\boldsymbol{\theta})$ involves an outer loop over 1 to M and an inner loop over 1 to K, so the overall complexity order is $\mathcal{O}(M^2K^2)$, completing the proof.

Appendix C
Proof of Lemma 2

The Lipschitz smoothness of $\mathrm{SINR}(\cdot)$ under the conventional Shannon capacity theorem has been proven in [24, Appendix A]. Therefore, this paper focuses only on the Lipschitz smoothness and continuity[5] of the channel dispersion function

[5]Smoothness and continuity correspond to the first-order information and zero-order information (i.e. numerical value) of a function, respectively [31].

$V_k(\cdot)$ and the proposed characterization function $\sigma(\cdot)$ in the FBL rate analysis. Specifically, we first analyze the Lipschitz continuity of the channel dispersion function

$$\begin{aligned}\frac{|V_k(\mathbf{X})-V_k(\mathbf{Y})|}{\log_2^2(e)} &= \left|\frac{1}{\left(1+\mathrm{SINR}_k(\mathbf{Y})\right)^2}-\frac{1}{\left(1+\mathrm{SINR}_k(\mathbf{X})\right)^2}\right| \\ &= \frac{\left|\left(2+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y})\right)\left(\mathrm{SINR}_k(\mathbf{X})-\mathrm{SINR}_k(\mathbf{Y})\right)\right|}{\left(1+\mathrm{SINR}_k(\mathbf{X})\right)^2\left(1+\mathrm{SINR}_k(\mathbf{Y})\right)^2} \\ &\overset{(a)}{<} \frac{\left(2+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y})\right)\left|\left(\mathrm{SINR}_k(\mathbf{X})-\mathrm{SINR}_k(\mathbf{Y})\right)\right|}{\left(1+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y})+\mathrm{SINR}_k(\mathbf{X})\,\mathrm{SINR}_k(\mathbf{Y})\right)} \\ &\overset{(b)}{\leq} \frac{\left(2+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y})\right)\left|\left(\mathrm{SINR}_k(\mathbf{X})-\mathrm{SINR}_k(\mathbf{Y})\right)\right|}{\left(1+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y})\right)} \\ &\overset{(c)}{\leq} 2L_{\mathrm{SINR}_k}\|\mathbf{X}-\mathbf{Y}\|. \end{aligned} \tag{79}$$

Here (a) is obtained from the fact that $(1+\mathrm{SINR}_k(\cdot))^2>1$, and (b) is obtained by neglecting the non-negative terms in the denominator. Since $\frac{(2+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y}))}{(1+\mathrm{SINR}_k(\mathbf{X})+\mathrm{SINR}_k(\mathbf{Y}))}\leq 2$ and $\mathrm{SINR}_k(\cdot)$ is L_{SINR_k}-Lipschitz continuous [24], then we get the inference of (c). Therefore, $V_k(\cdot)$ possesses the $2\log_2^2(e)L_{\mathrm{SINR}_k}$-Lipschitz continuous property. Note that in the $\mathrm{SE}_k(\cdot)$, $V_k(\cdot)$ also undergoes a square root operation. Therefore, we further consider the following relationship

$$\begin{aligned}\left|\sqrt{V_k(\mathbf{X})}-\sqrt{V_k(\mathbf{Y})}\right| &\overset{(a)}{\leq} \frac{|V_k(\mathbf{X})-V_k(\mathbf{Y})|}{\sqrt{V_k(\mathbf{X})}+\sqrt{V_k(\mathbf{Y})}} \\ &\overset{(b)}{<} 2\log_2^2(e)L_{\mathrm{SINR}_k}C_v\|\mathbf{X}-\mathbf{Y}\|, \end{aligned} \tag{80}$$

where (a) is obtained from the absolute value difference inequality, and (b) is derived from the boundedness of $V_k(\cdot)$. Since $V_k(\mathbf{X})\in(0,\log_2^2(e)],\forall\mathbf{X}\in\mathcal{C}$, there exists a positive constant C_v such that $\frac{1}{\sqrt{V_k(\mathbf{X})}+\sqrt{V_k(\mathbf{Y})}}\leq C_v$. Based on the composition property of Lipschitz functions [24], the Lipschitz continuity of $\sqrt{V_k(\cdot)}$ is established. Therefore, according to (10), the SE function also satisfies Lipschitz continuity due to the linear combination property of Lipschitz functions.

Next, we aim to prove the Lipschitz smoothness property of SE. According to (68), it suffices to establish the Lipschitz continuity of $\frac{[1+\mathrm{SINR}_k(\cdot)]^{-2}}{\sqrt{[1+\mathrm{SINR}_k(\cdot)]^2-1}}$. To achieve this, we utilize the following theorem.

Quotient rule of Lipschitz Functions [24]: Let $f_1:\mathbb{R}^n\to\mathbb{R}^n$ and $f_2:\mathbb{R}^n\to\mathbb{R}$ be L_1-continuous and L_2-continuous, respectively. Furthermore, as $\|f_1(\mathbf{x})\|\leq C_1$ and $|f_2(\mathbf{x})|\geq C_2$, then $\frac{f_1}{f_2}$ is $C_1L_1+\frac{L_2}{C_2^2}$ Lipschitz continuous function.

According to (79) and (80), the Lipschitz continuity of both the numerator and denominator of $\frac{[1+\mathrm{SINR}_k(\cdot)]^{-2}}{\sqrt{[1+\mathrm{SINR}_k(\cdot)]^2-1}}$ can be readily proved. Furthermore, since $[1+\mathrm{SINR}_k(\cdot)]^{-2}<1$ and $\sqrt{[1+\mathrm{SINR}_k(\cdot)]^2-1}>\sqrt{(1+2^{\frac{\mathrm{SE}_{th}}{\tau_d}})^2-1}$, the Lipschitz smoothness property of SE can be derived by the above quotient rule of Lipschitz functions.

Next, we proceed to prove the Lipschitz continuity and smoothness of the characterization function. According to the proof process of Proposition 1, the gradient of the characterization function is $\nabla\sigma(\mathbf{x})=\frac{\varpi\delta\|\mathbf{x}\|^{\varpi-2}\mathbf{x}}{(\|\mathbf{x}\|^{\varpi}+\delta)^2}$. Since the denominator is always greater than zero and the power allocation coefficients are subject to finite constraints, it is straightforward to prove that the gradient is bounded. Therefore, according to the *Lagrange mean value theorem*, the Lipschitz continuity of $\delta(\cdot)$ always holds. Furthermore, the absolute value of gradient difference satisfies

$$\begin{aligned}\|\nabla\sigma(\mathbf{x})-\nabla\sigma(\mathbf{y})\| &= \left\|\frac{\varpi\delta\|\mathbf{x}\|^{\varpi-2}\mathbf{x}}{(\|\mathbf{x}\|^{\varpi}+\delta)^2}-\frac{\varpi\delta\|\mathbf{y}\|^{\varpi-2}\mathbf{y}}{(\|\mathbf{y}\|^{\varpi}+\delta)^2}\right\| \\ &\leq \left\|\frac{(\|\mathbf{x}\|^{\varpi-2}-\|\mathbf{y}\|^{\varpi-2})\mathbf{x}}{(\|\mathbf{x}\|^{\varpi}+\delta)^2}\right\|+\left\|\frac{\|\mathbf{y}\|^{\varpi-2}(\mathbf{x}-\mathbf{y})}{(\|\mathbf{y}\|^{\varpi}+\delta)^2}\right\|. \end{aligned} \tag{81}$$

The above result is obtained by the triangle inequality of absolute values for scaling. It is easy to prove that for any $\varpi\geq 2$, the above individual two parts are always Lipschitz continuous [40]. Due to space limitations, we omit the detailed process here.

By combining [24, Appendix A] with the above discussion, it can be readily verified that both the SE expression in URLLC communications and the approximate power consumption model adopted in this paper exhibit Lipschitz smoothness. This provides the theoretical basis for employing first-order fast algorithms as an alternative to conventional convex optimization techniques in the URLLC communications scenario, thus completing the proof.

Appendix D
Proof of Theorem 2

According to (57), the complete expression for the update projection of $\boldsymbol{\theta}^{(t)}$ is

$$\begin{aligned}\boldsymbol{\theta}^{(t)} &= \arg\min_{\boldsymbol{\theta}\in\mathcal{C}}\ \|\boldsymbol{\theta}-\mathbf{Z}^{(t)}-\alpha^{(t)}\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})\|^2 \\ &= \arg\max_{\boldsymbol{\theta}\in\mathcal{C}}\left\langle\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}),\boldsymbol{\theta}-\mathbf{Z}^{(t)}\right\rangle_F-\frac{1}{2\alpha^{(t)}}\|\boldsymbol{\theta}-\mathbf{Z}^{(t)}\|^2. \end{aligned} \tag{82}$$

Note that when $\boldsymbol{\theta}=\mathbf{Z}^{(t)}$, the optimal objective in (82) is 0, thus the following inequality holds

$$\left\langle\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}),\boldsymbol{\theta}^{(t)}-\mathbf{Z}^{(t)}\right\rangle_F-\frac{1}{2\alpha^{(t)}}\|\boldsymbol{\theta}^{(t)}-\mathbf{Z}^{(t)}\|^2\geq 0. \tag{83}$$

Since Lemma 2 has proven that $f_{\vartheta^{(i)},\chi}(\cdot)$ possesses the L_f-smooth property, its function values and gradient information satisfy the following condition [37]

$$f_{\vartheta^{(i)},\chi}(\mathbf{X})\geq f_{\vartheta^{(i)},\chi}(\mathbf{Y})+\left\langle\nabla f_{\vartheta^{(i)},\chi}(\mathbf{X}),\mathbf{Y}-\mathbf{X}\right\rangle_F-\frac{L_f}{2}\|\mathbf{Y}-\mathbf{X}\|^2. \tag{84}$$

Combing (83) and (84), we can obtain

$$\begin{aligned} f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) &\geq f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})-\frac{L_f}{2}\|\boldsymbol{\theta}^{(t)}-\mathbf{Z}^{(t)}\|^2 \\ &+\left\langle\nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}),\boldsymbol{\theta}^{(t)}-\mathbf{Z}^{(t)}\right\rangle_F \\ &\geq f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})+\left(\frac{1}{2\alpha^{(t)}}-\frac{L_f}{2}\right)\|\boldsymbol{\theta}^{(t)}-\mathbf{Z}^{(t)}\|^2. \end{aligned} \tag{85}$$

Clearly, if $\alpha^{(t)}\leq\frac{1}{L_f}$, then $f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)})\geq f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})$. Furthermore, according to (60), the value of $f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})$ has two different values depending on two conditions. When the accelerated extrapolated value is large, we have $f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})=f_{\vartheta^{(i)},\chi}(\mathbf{V}^{(t-1)})>f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)})$. Conversely, we have $f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)})=f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)})\geq f_{\vartheta^{(i)},\chi}(\mathbf{V}^{(t-1)})$.

In conclusion, the following monotonicity always holds

$$f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) \geq f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}) \geq f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)}). \tag{86}$$

From (85) and (86), we have

$$f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t)}) - f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(t-1)}) \geq \left(\frac{1}{2\alpha^{(t)}} - \frac{L_f}{2}\right) \|\boldsymbol{\theta}^{(t)} - \mathbf{Z}^{(t)}\|^2 \tag{87}$$

which results in the following summation inequality

$$\begin{aligned} \infty &> f^{\star}_{\vartheta^{(i)},\chi} - f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(0)}) \\ &\geq \sum_{t=0}^{\infty} \left(\frac{1}{2\alpha^{(t)}} - \frac{L_f}{2}\right) \|\boldsymbol{\theta}^{(t)} - \mathbf{Z}^{(t)}\|^2, \end{aligned} \tag{88}$$

where $f^{\star}_{\vartheta^{(i)},\chi}$ denotes the objective value of the accumulation points. Since the summation on the RHS of the above equation is bounded, we can conclude that as $\alpha^{(t)} < \frac{1}{L_f}$, $\boldsymbol{\theta}^{(\infty)} \to \mathbf{Z}^{(\infty)}$.

Since (82) is actually a convex problem and the global optimal solution can be obtained from (58), thus $\boldsymbol{\theta}^{(t)} \in \mathcal{C}$ satisfies the following optimality condition

$$\left\langle \nabla f_{\vartheta^{(i)},\chi}(\mathbf{Z}^{(t)}) - \frac{1}{\alpha^{(t)}} \left(\boldsymbol{\theta}^{(t)} - \mathbf{Z}^{(t)}\right), \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \right\rangle_F \leq 0. \tag{89}$$

Letting $t \to \infty$, we can simplify the above equation as

$$\left\langle \nabla f_{\vartheta^{(i)},\chi}(\boldsymbol{\theta}^{(\infty)}), \boldsymbol{\theta} - \boldsymbol{\theta}^{(\infty)} \right\rangle_F \leq 0. \tag{90}$$

The above equation has a clear physical interpretation: In the vicinity of the point $\boldsymbol{\theta}^{(\infty)}$, no better solution exists that can further increase the value of $f_{\vartheta^{(i)},\chi}(\cdot)$. This indicates that, for given $\vartheta^{(i)}$ and χ, the APG-AM algorithm converges to a stationary point of $\mathcal{P}_6$ [37], thus completing the proof.

REFERENCES

[1] A. A. Nasir, H. D. Tuan, H. Q. Ngo, T. Q. Duong and H. V. Poor, "Cell-Free Massive MIMO in the Short Blocklength Regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5861-5871, Sept. 2021.

[2] Q. Peng, H. Ren, C. Pan, N. Liu and M. Elkashlan, "Resource Allocation for Uplink Cell-Free Massive MIMO Enabled URLLC in a Smart Factory," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 553-568, Jan. 2023.

[3] Y. Huang, Y. Jiang, F. -C. Zheng, P. Zhu, D. Wang and X. You, "Energy-Efficient Optimization in User-Centric Cell-Free Massive MIMO Systems for URLLC With Finite Blocklength Communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12801-12814, Sept. 2024.

[4] Q. Peng, H. Ren, M. Dong, M. Elkashlan, K. -K. Wong and L. Hanzo, "Resource Allocation for Cell-Free Massive MIMO-Aided URLLC Systems Relying on Pilot Sharing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2193-2207, July 2023.

[5] B. Chong *et al.*, "Performance Optimization on Cell-Free Massive MIMO-Aided URLLC Systems With User Grouping," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 13977-13992, Oct. 2024.

[6] Q. Peng, H. Ren, C. Pan, N. Liu and M. Elkashlan, "Resource Allocation for Cell-Free Massive MIMO-Enabled URLLC Downlink Systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 7669-7684, June 2023.

[7] Y. Huang *et al.*, "Effective Energy Efficiency of Cell-Free mMIMO Systems for URLLC With Probabilistic Delay Bounds and Finite Blocklength Communications," *IEEE Trans. Wireless Commun.*, vol. 24, no. 3, pp. 2279-2296, March 2025.

[8] B. Chong and H. Lu, "Statistical Delay Performance Analysis for URLLC in Uplink Cell-Free Massive MIMO Systems: A Stochastic Network Calculus Perspective," *IEEE Trans. Wireless Commun.*, vol. 24, no. 2, pp. 1162-1177, Feb. 2025.

[9] H. Ren *et al.*, "Joint Pilot and Payload Power Allocation for Massive-MIMO-Enabled URLLC IIoT Networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816-830, May 2020.

[10] A. Lancho, G. Durisi and L. Sanguinetti, "Cell-Free Massive MIMO for URLLC: A Finite-Blocklength Analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8723-8735, Dec. 2023.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[12] G. Chen, Z. Wang, H. Lin, Y. Huang and L. Yang, "Computationally Efficient Unsupervised Deep Learning for Robust Joint AP Clustering and Beamforming Design in Cell-Free Systems," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 4250-4266, May 2025.

[13] X. Yan, Z. Wang, Y. Jia, Z. Zhang and Y. Huang, "Access Point Selection and Beamforming Design for Cell-Free Network: From Fractional Programming to GNN," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 9345-9360, Aug. 2024.

[14] Q. Chen, Z. Wang, C. Qi, Z. Gao, Y. Huang and D. Niyato, "Decentralized Likelihood Ascent Search-Aided Detection for Distributed Large-Scale MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 4160-4173, May 2025.

[15] J. Tian, Y. Han, S. Jin, J. Zhang and J. Wang, "Analytical Channel Modeling: From MIMO to Extra Large-Scale MIMO," *Chin. J. Electron.*, vol. 34, no. 1, pp. 1-15, January 2025.

[16] N. Li and P. Fan, "Distributed Cell-Free Massive MIMO Versus Cellular Massive MIMO Under UE Hardware Impairments," *Chin. J. Electron.*, vol. 33, no. 5, pp. 1274-1285, September 2024.

[17] R. Pinto Antonioli *et al.*, "On the Energy Efficiency of Cell-Free Systems With Limited Fronthauls: Is Coherent Transmission Always the Best Alternative?," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8729-8743, Oct. 2022.

[18] H. Q. Ngo, L. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the Total Energy Efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25-39, Mar. 2018.

[19] K. -G. Nguyen, Q. -D. Vu, M. Juntti and L. -N. Tran, "Energy Efficiency Maximization for C-RANs: Discrete Monotonic Optimization, Penalty, and ℓ_0-Approximation Methods," *IEEE Trans. Signal Proc.*, vol. 66, no. 17, pp. 4435-4449, Sept.1, 2018.

[20] O. Tervo, L. -N. Tran and M. Juntti, "Optimal Energy-Efficient Transmit Beamforming for Multi-User MISO Downlink," *IEEE Trans. Signal Proc.*, vol. 63, no. 20, pp. 5574-5588, Oct.15, 2015.

[21] T. T. Vu, D. T. Ngo, M. N. Dao, S. Durrani and R. H. Middleton, "Spectral and Energy Efficiency Maximization for Content-Centric C-RANs With Edge Caching," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6628-6642, Dec. 2018.

[22] V. -D. Nguyen, T. Q. Duong, H. D. Tuan, O. -S. Shin and H. V. Poor, "Spectral and Energy Efficiencies in Full-Duplex Wireless Information and Power Transfer," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2220-2233, May 2017.

[23] G. Dong *et al.* "Energy-Efficiency-Oriented Joint User Association and Power Allocation in Distributed Massive MIMO Systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5794-5808, June 2019.

[24] T. C. Mai *et al.* "Energy Efficiency Maximization in Large-Scale Cell-Free Massive MIMO: A Projected Gradient Approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6357-6371, Aug. 2022.

[25] B. Yan, Z. Wang, J. Zhang and Y. Huang, "Joint Antenna Activation and Power Allocation for Energy-Efficient Cell-Free Massive MIMO Systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 1, pp. 243-247, Jan. 2025.

[26] M. Farooq, H. Q. Ngo, E.-K. Hong and L.-N. Tran, "Utility Maximization for Large-Scale Cell-Free Massive MIMO Downlink," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 7050-7062, Oct. 2021.

[27] H. Li, C. Fang and Z. Lin, "Accelerated First-Order Optimization Algorithms for Machine Learning," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 2067-2082, Nov. 2020.

[28] P. Luong *et al.*, "Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RAN," *IEEE Trans. Signal Proc.*, vol. 65, no. 21, pp. 5605-5620, Nov.1, 2017.

[29] Y. -F. Liu et al., "A Survey of Recent Advances in Optimization Methods for Wireless Communications," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 11, pp. 2992-3031, Nov. 2024.

[30] A. Papazafeiropoulos, J. An, P. Kourtessis, T. Ratnarajah and S. Chatzinotas, "Achievable Rate Optimization for Stacked Intelligent Metasurface-Assisted Holographic MIMO Communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 13173-13186, Oct. 2024.

[31] L. Du *et al.*, "Cell-Free Massive MIMO: Joint Maximum-Ratio and Zero-Forcing Precoder With Power Control," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3741-3756, June 2021.

[32] C. Hao *et al.*, "Joint User Association and Power Control for Cell-Free Massive MIMO," *IEEE Int. Things J.*, vol. 11, no. 9, pp. 15823-15841, 1 May1, 2024.

[33] G. Wang *et al.*, "Green Cell-Free Massive MIMO: An Optimization Embedded Deep Reinforcement Learning Approach," *IEEE Trans. Signal Proc.*, vol. 72, pp. 2751-2766, 2024.
[34] T. Van Chien, E. Björnson and E. G. Larsson, "Joint Power Allocation and Load Balancing Optimization for Energy-Efficient Cell-Free Massive MIMO Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798-6812, Oct. 2020.
[35] G. Interdonato, M. Karlsson, E. Björnson and E. G. Larsson, "Local Partial Zero-Forcing Precoding for Cell-Free Massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758-4774, July 2020.
[36] H. H. M. Tam, H. D. Tuan, D. T. Ngo, T. Q. Duong, and H. V. Poor, "Joint load balancing and interference management for small-cell heterogeneous networks with limited backhaul capacity, *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 872-884, Feb. 2017.
[37] Q. Li, *et al*. "Convergence analysis of proximal gradient with momentum for nonconvex optimization." *International Conference on Machine Learning*. PMLR, 2017.
[38] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan and L. Hanzo, "The Non-Coherent Ultra-Dense C-RAN Is Capable of Outperforming Its Coherent Counterpart at a Limited Fronthaul Capacity," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2549-2560, Nov. 2018.
[39] Y. Sun, P. Babu and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Trans. Signal Proc.*, vol. 65, no. 3, pp. 794-816, 1 Feb.1, 2017.
[40] N. Weaver, *Lipschitz Algebras*, 2nd ed. Singapore: World Scientific, 2018.

BIOGRAPHIES

Bin Yan received the B.S. degree in information engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2023, where he is currently pursuing the Ph.D. degree in signal and information processing. His research interests include efficient communication resource allocation and precoding design in cell-free massive MIMO systems. (Email: bin_yan@seu.edu.cn)

Zheng Wang (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, and the M.S. degree in communications from University of Manchester, Manchester, U.K., in 2010. He received the Ph.D degree in communication engineering from Imperial College London, UK, in 2015.

Since 2021, he has been an Associate Professor in the School of Information and Engineering, Southeast University (SEU), Nanjing, China. From 2015 to 2016 he served as a Research Associate at Imperial College London, UK. From 2016 to 2017 he was an senior engineer with Radio Access Network R&D division, Huawei Technologies Co.. From 2017 to 2020 he was an Associate Professor at the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include massive MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications.

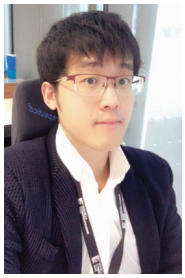
Amin Sakzad (Member, IEEE) received the PhD degree in applied mathematics, in 2011. Starting from 2016, he held a postdoctoral research fellowship with the Faculty of IT, Monash University. He became a lecturer, in 2017 and since 2020, he is a senior lecturer with the Department of Software,Systems,and Cybersecurity, Monash University. His research focuses on lattice based cryptography and wireless network coding.

Chuan Zhang (Senior Member, IEEE) received the B.E. degree in microelectronics and the M.E. degree in very-large scale integration (VLSI) design from Nanjing University, Nanjing, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities (UMN), USA, in 2012.

He is currently the Young Chair Professor with Southeast University. He is also with the LEADS, National Mobile Communications Research Laboratory, Frontiers Science Center for Mobile Information Communications and Security of MoE, Quantum Information Center of Southeast University, and the Purple Mountain Laboratories, Nanjing, China. His current research interests are algorithms and implementations for signal processing and communication systems.

Yongming Huang (Fellow, IEEE) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2007.

Since March 2007, he has been a Faculty Member of the School of Information Science and Engineering, Southeast University, where he is currently a Full Professor. He has also been the Director of the Pervasive Communication Research Center, Purple Mountain Laboratories, since 2019. From 2008 to 2009, he visited the Signal Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden. His current research interests include intelligent 5G/6G mobile communications and millimeter wave wireless communications. He has published over 200 peer-reviewed articles, hold over 80 invention patents. He submitted around 20 technical contributions to IEEE standards and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. He served as an Associate Editor for IEEE Transactions on Signal Processing and a Guest Editor for the IEEE Journal on Selected Areas in Communications. He is currently an Editor-at-Large of the IEEE Open Journal of the Communications Society and an Associate Editor of the IEEE Wireless Communications Letters.



Derrick Wing Kwan Ng (Fellow, IEEE) received his bachelor's degree (with first-class Honors) and the Master of Philosophy degree in electronic engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2006 and 2008, respectively, and his Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada, in November 2012. Following his Ph.D., he was a senior postdoctoral fellow at the Institute for Digital Communications, Friedrich-Alexander University Erlangen-N¨urnberg (FAU), Germany. He is currently a Scientia Associate Professor with the University of New South Wales, Sydney, NSW, Australia.

His research interests include global optimization, integrated sensing and communication (ISAC), physical layer security, IRS-assisted communication, UAV-assisted communication, wireless information and power transfer, and green (energy-efficient) wireless communications. He has been recognized as a Highly Cited Researcher by Clarivate Analytics (Web of Science) since 2018. He was the recipient of the Australian Research Council (ARC) Discovery Early Career Researcher Award 2017, IEEE Communications Society Leonard G. Abraham Prize 2023, IEEE Communications Society Stephen O. Rice Prize 2022, Best Paper Awards at the WCSP 2020, 2021, IEEE TCGCC Best Journal Paper Award 2018, INISCOM 2018, IEEE International Conference on Communications (ICC) 2018, 2021, 2023, 2024, IEEE International Conference on Computing, Networking and Communications (ICNC) 2016, IEEE Wireless Communications and Networking Conference (WCNC) 2012, IEEE Global Telecommunication Conference (Globecom) 2011, 2021, 2023 and IEEE Third International Conference on Communications and Networking in China 2008. From January 2012 to December 2019, he served as an Editorial Assistant to the Editor-in-Chief of the IEEE Transactions on Communications. He is also an Area Editor of the IEEE Transactions on Communications, an Associate Editor-in-Chief of the IEEE Open Journal of the Communications Society, and a member of the IEEE Transactions on Wireless Communications executive editorial committee.