Joint Antenna Activation and Power Allocation for Energy-Efficient Cell-Free Massive MIMO Systems

Bin Yan, Zheng Wang, *Senior Member, IEEE*, Jiayi Zhang, *Senior Member, IEEE*, and Yongming Huang, *Senior Member, IEEE*

***Abstract*—In this letter, a joint antenna activation and power allocation (JAAPA) scheme is proposed for the downlink energy efficiency (EE) optimization in cell-free massive multiple-input multiple-output (MIMO) systems. Different from the traditional optimization approaches that only focus on power allocation, the proposed JAAPA scheme introduces the activated antenna number into the consideration of EE optimization, thus leading to a further improvement of EE. Specifically, the JAAPA scheme is formulated as a non-convex fractional optimization problem. In order to solve it in an efficient approach, we adopt the monotone accelerated proximal gradient (APG) algorithm to obtain a sub-optimal solution. Finally, simulation results validate that the proposed JAAPA scheme achieves a better trade-off between EE performance and computational complexity compared to the existing resource allocation schemes.**

***Index Terms*—Energy efficiency, accelerated proximal gradient, antenna activation, power allocation, joint optimization.**

I. Introduction

NOWADAYS, energy efficiency (EE) has emerged as an important performance metric in cell-free massive MIMO systems [1]. As a fundamental method to optimize EE, power allocation not only manages the near-far effect but also balances the trade-off between spectral efficiency (SE) and power consumption [2]–[7]. In general, optimizing EE through power allocation usually establishes a non-convex optimization problem so that an efficient solution is highly desired in practice [2]. Specifically, the sequential convex approximation (SCA) is employed to transform the non-convex problem into a series of standard convex problems, which can be iteratively addressed by general solvers, i.e. CVX and YALMIP [3]–[6]. Meanwhile, the accelerated proximal gradient (APG) algorithm is widely applied by taking advantages of the first-order information, which significantly reduces the computational complexity with comparable performance to SCA methods [7]–[9].

However, previous works of EE joint optimization in cell-free systems only focus on the power allocation and access point (AP) selection, where the impact of the number of activated antennas on EE has not been throughly investigated yet [2], [3], [5]–[7]. In fact, as more antennas are deployed at APs to improve SE, more power consumption is also incurred with the growth of activated antennas, which may deteriorate EE at the same time [10], [11]. To this end, the number of activated antennas actually plays an important role in EE optimization, which should be taken into account. More specifically, such a joint optimization problem have been studied in distributed massive MIMO and multi-user MIMO systems, revealing that the detrimental effect of excessive antenna activation on EE is prevalent across different scenarios [4], [11].

In this letter, besides power allocation, the impact of antenna activation is also considered in the EE optimization problem. Specifically, the EE expression is extended under the fractional-exponent normalized conjugate beamforming (FENCB) precoding framework and therefore a joint antenna activation and power allocation (JAAPA) scheme is proposed to achieve reasonable resource allocation. Furthermore, the characteristics of SE and power consumption regarding to the number of activated antennas are explored, and a condition for adopting the minimum level of antenna activation is derived. Finally, necessary problem transformations are implemented to convert the joint non-convex optimization problem established by JAAPA into an accessible form, where the APG algorithm is used to efficiently obtain a sub-optimal solution.

II. System Model

Consider a canonical cell-free massive MIMO downlink system with M multi-antenna APs serving K single-antenna users simultaneously, while each AP has n activated antennas. Specifically, the propagation channel between the k-th user and the m-th AP is given by $\mathbf{g}_{mk} = \sqrt{\beta_{mk}}\mathbf{h}_{mk} \in \mathbb{C}^n$, where β_{mk} represents the large-scale fading coefficient (LSFC) , and $\mathbf{h}_{mk} \in \mathbb{C}^n$ is the small-scale fading vector. All APs are connected to the central processing unit (CPU) via backhaul links, and the system operates in time division duplex (TDD) mode. During the uplink training phase, all K users simultaneously transmit pilot sequences to the APs. Let τ_c and τ_u be the length of the coherence interval and the number of symbols used for uplink training, respectively.

Under the minimum mean square error (MMSE) criterion, the estimated channel vector $\widehat{\mathbf{g}}_{mk}$ is distributed according to $\mathcal{CN}(\mathbf{0}_n, \gamma_{mk}\mathbf{I}_n)$, and the mean-square is denoted by [3]

$$\gamma_{mk} \triangleq \frac{\rho_u \tau_u \beta_{mk}^2}{1 + \rho_u \tau_u \sum_{j=1}^{K} \beta_{mj} |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j|^2}, \tag{1}$$

Note that ρ_u represents the normalized uplink transmit signal-to-noise ratio (SNR), $\boldsymbol{\psi}_k \in \mathbb{C}^{\tau_u}$ denotes the pilot sequence transmitted by the k-th user.

During the downlink data transmission phase, local distributed precoding is widely used to reduce backhaul overhead and enhance system scalability [2]. Up to now, low-complexity EE optimization scheme has only been restricted within CB

Bin Yan, Zheng Wang and Yongming Huang are with School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mails: bin_yan@seu.edu.cn; wznuaa@gmail.com). Jiayi Zhang is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: jiayizhang@bjtu.edu.cn).

precoding [7], remaining the possibility of further enhancing performance through more advanced precoding methods [6], [10]. To explore the potential of other CB variants precoding methods, the distributed FENCB precoding framework is considered [12]

$$\mathbf{w}_{mk} = \frac{\widehat{\mathbf{g}}_{mk}^*}{\|\widehat{\mathbf{g}}_{mk}\|^{\alpha+1}}, \tag{2}$$

as a general solution, where $\alpha = -1, 0, 1$ corresponds to CB, NCB, and ECB respectively [3], [6], [7], [10]. In fact, α represents the *channel inversion rate* and could be any real value. Accordingly, when adopting the FENCB scheme, the non-negative power allocation coefficients $\boldsymbol{\eta} \in \mathbb{R}^{M\times K}$ must satisfy the per-AP transmit power constraint [12]

$$P_{\text{AP,m}}(\boldsymbol{\eta}) = \rho_d \frac{\Gamma(n-\alpha)}{\Gamma(n)} \sum_{k=1}^{K} \frac{\eta_{mk}}{\gamma_{mk}^{\alpha}} \leq \rho_d, \quad \forall m, \tag{3}$$

where ρ_d represents the maximum downlink normalized transmit power, $\Gamma(\cdot)$ is the *Gamma* function and $n > \alpha$. Based on the use and then forget principle [12], the effective signal-to-interference-plus-noise ratio (SINR) of k-th user is given as

$$\begin{aligned}\text{SINR}_k(\boldsymbol{\eta}) = \rho_d \left(\sum_{m=1}^{M} \sqrt{\eta_{mk}} a_{mkk}\right)^2 \Bigg\{1+ \\ \rho_d \sum_{j=1}^{K}\sum_{m=1}^{M} \eta_{mj} b_{mkj} + \rho_d \sum_{j\neq k}^{K} \left(\sum_{m=1}^{M} \sqrt{\eta_{mj}} a_{mkj}\right)^2 \Bigg\}^{-1},\end{aligned} \tag{4}$$

with

$$\begin{cases} a_{mkj} = \dfrac{\Gamma(n-\frac{\alpha-1}{2})}{\Gamma(n)} \dfrac{\gamma_{mk}^{1/2}}{\gamma_{mj}^{\alpha/2}} |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j|, \\ b_{mkj} = \dfrac{\Gamma(n-\alpha)}{\Gamma(n)} (n-\alpha-1) \dfrac{\gamma_{mk}}{\gamma_{mj}^{\alpha}} |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j|^2 \\ \qquad -a_{mkj}^2 + \dfrac{\Gamma(n-\alpha)}{\Gamma(n)} \dfrac{\beta_{mk}}{\gamma_{mj}^{\alpha}}. \end{cases}$$

The proportion of downlink data transmission within the entire coherence interval is defined as $\overline{\tau_d} = 1 - \frac{\tau_u}{\tau_c}$, then the achievable downlink SE of the k-th user can be obtained by *Shannon's theorem*, i.e.,

$$\text{SE}_k(\boldsymbol{\eta}) = \overline{\tau_d} \log_2 \left(1 + \text{SINR}_k(\boldsymbol{\eta})\right). \tag{5}$$

Based on (5), the global EE (bit/Joule) is defined as the ratio of total throughput to system power consumption in cell-free massive MIMO [3]

$$\text{EE}(\boldsymbol{\eta}) = \frac{B \sum_{k=1}^{K} \text{SE}_k(\boldsymbol{\eta})}{P_{\text{total}}(\boldsymbol{\eta})}, \tag{6}$$

where B represents the system bandwidth. Note that the above expression for EE is solely a function of the power coefficients, regardless of the activated antenna numbers.

Given $\text{EE}(\boldsymbol{\eta})$ in (6), the traditional EE optimization is achieved by executing power allocation scheme, which is mathematically represented as [3], [5]–[7]

$$\begin{aligned} \mathcal{P}_1: \quad \max_{\boldsymbol{\eta}} \quad & \text{EE}(\boldsymbol{\eta}) && \text{(7a)} \\ \text{s.t.} \quad & \text{SE}_k(\boldsymbol{\eta}) \geq \text{SE}_{th}, \forall k, && \text{(7b)} \\ & \{(3); \eta_{mk} \geq 0, \forall m, \forall k\}. && \text{(7c)} \end{aligned}$$

Here, in constraint (7b), SE_{th} serves as a minimum SE threshold, ensuring that all users can communicate reliably. Constraint (7c) represents the power limitation conditions. However, this scheme only focuses on power allocation optimization with all the available antennas being activated, which probably turns out to be inefficient for EE optimization.

III. Joint Optimization Scheme and Energy Efficiency Analysis

To achieve simultaneous optimization of power and activated antennas, we first need to rewrite the SINR expression (4) as (8) by expanding the independent variable from only $(\boldsymbol{\theta})$ to $(\boldsymbol{\theta}, n)$. Specifically, the variable substitution $\theta_{mk} = \sqrt{\eta_{mk}\gamma_{mk}^{-\alpha}}$ is performed for mathematical tractability, which is inspired by equation (3). Principly, this substitution confines the Lipschitz gradient of the SINR function within a suitable magnitude, preventing slow or infeasible convergence of first-order algorithms due to extreme gradient constants. Furthermore, this variable substitution is a generalization of [7] and [9], applicable to the transformation of any utility function within the FENCB precoding framework. The same variable substitution and expansion can also be applied to the power consumption function, where the details are omitted here.

Overall, the EE expression can be extended to optimize both power allocation and the number of activated antennas simultaneously, leading to the proposed JAAPA scheme[1]

$$\begin{aligned} \mathcal{P}_2: \quad \max_{\boldsymbol{\theta}\in\mathcal{C}_1, n\in\mathcal{C}_2} \quad & \overline{\text{EE}}(\boldsymbol{\theta}, n) \triangleq B \frac{\text{SE}(\boldsymbol{\theta}, n)}{P(\boldsymbol{\theta}, n)}, && \text{(11a)} \\ \text{s.t.} \quad & \text{SE}_k(\boldsymbol{\theta}, n) \geq \text{SE}_{th}, \forall k, && \text{(11b)} \end{aligned}$$

[1]Note that in (11a), the denominator's power consumption is different from $P_{\text{total}}(\boldsymbol{\eta})$ in (6), as it ignores the traffic-dependent power. This equivalent treatment aims to simplify the optimization of the objective function. Due to space limitations, please refer to [3, Appendix B] for the detailed procedure.

$$\text{SINR}_k(\boldsymbol{\theta}, n) = \frac{\rho_d \Gamma_n^2 \left(\sum_{m=1}^{M} \theta_{mk} \sqrt{\gamma_{mk}}\right)^2}{1 + \rho_d \sum_{j=1}^{K}\sum_{m=1}^{M} \theta_{mj}^2 (\Phi_{mkj} + \Psi_{mk}) + \rho_d \sum_{j\neq k}^{K} \left(\sum_{m=1}^{M} \theta_{mj} \Omega_{mkj}\right)^2} \tag{8}$$

$$\text{with} \quad \Gamma_n \triangleq \frac{\Gamma(n-\frac{\alpha-1}{2})}{\Gamma(n)}, \ \widetilde{\Gamma}_n \triangleq \frac{\Gamma(n-\alpha)}{\Gamma(n)}, \ \overline{\Gamma}_n \triangleq n-\alpha-1, \tag{9}$$

$$\text{and} \quad \Phi_{mkj} \triangleq (\widetilde{\Gamma}_n \overline{\Gamma}_n - \Gamma_n^2)\gamma_{mk} |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j|^2, \ \Psi_{mk} \triangleq \widetilde{\Gamma}_n \beta_{mk}, \ \Omega_{mkj} \triangleq \Gamma_n \sqrt{\gamma_{mk}} |\boldsymbol{\psi}_k^H \boldsymbol{\psi}_j|. \tag{10}$$

with

$$\mathcal{C}_1 \triangleq \{\boldsymbol{\theta} | \widetilde{\Gamma}_n \sum_{k=1}^{K} \theta_{mk}^2 \leq 1, \forall m; \boldsymbol{\theta} \in \mathbb{R}^{M\times K} \geq 0\}, \quad (12a)$$

$$\mathcal{C}_2 \triangleq \{n | \max(\lfloor \alpha \rfloor + 1, 1) \leq n \leq N_{\max}; n \in \mathbb{Z}_+\}, \quad (12b)$$

$$\mathrm{SE}(\boldsymbol{\theta}, n) \triangleq \sum_{k=1}^{K} \mathrm{SE}_k(\boldsymbol{\theta}, n) = \sum_{k=1}^{K} \overline{\tau_d} \log_2 \left[1 + \mathrm{SINR}_k(\boldsymbol{\theta}, n)\right], \quad (12c)$$

$$P(\boldsymbol{\theta}, n) \triangleq \rho_d N_0 \sum_{m=1}^{M} \frac{\widetilde{\Gamma}_n}{\alpha_m} \left(\sum_{k=1}^{K} \theta_{mk}^2\right) + \sum_{m=1}^{M} (n P_{\mathrm{tc},m} + P_{0,m}). \quad (12d)$$

Compared to $\mathcal{P}_1$, the optimization problem $\mathcal{P}_2$ introduces an additional constraint on the activated antenna number n. Typically, $\lfloor \alpha \rfloor$ represents the floor function of α, (12b) restricts n meets practical requirements, which is greater than the channel inversion rate and less than the maximum number of available antennas. Notably, each AP is assumed to have the same number of activated antennas for simplicity. In equation (12d), where N_0 denotes the noise power, $0 < \alpha_m \leq 1$ is the power amplifier efficiency, $P_{\mathrm{tc},m}$ is the internal power to run the circuit components, and P_0 is the traffic-independent power consumption [3].

Theorem 1. *When adopting the FENCB precoding framework, if the following condition is satisfied*

$$\sum_{m=1}^{M} P_{\mathrm{tc},m} \geq \frac{2K\overline{\tau_d}}{\ln 2}, \quad (13)$$

then the optimal number of activated antennas for $\mathcal{P}_2$ is

$$n^\star = \max(\lfloor \alpha \rfloor + 1, 1). \quad (14)$$

Proof. We elucidate this theorem by analyzing the growth rate for SE and system power consumption as n increases. Firstly, consider the following upper bound for $\mathrm{SE}(\boldsymbol{\theta}, n)$

$$\begin{aligned}
&\mathrm{SE}(\boldsymbol{\theta}, n) = \sum_{k=1}^{K} \overline{\tau_d} \log_2 \left[1 + \mathrm{SINR}_k(\boldsymbol{\theta}, n)\right] \\
&< \sum_{k=1}^{K} \overline{\tau_d} \log_2 \left[1 + \frac{\Gamma_n^2 \left(\sum_{m=1}^{M} \theta_{mk} \sqrt{\gamma_{mk}}\right)^2}{\widetilde{\Gamma}_n \sum_{j=1}^{K} \sum_{m=1}^{M} \theta_{mj}^2 \beta_{mk}}\right] \\
&= \sum_{k=1}^{K} \overline{\tau_d} \log_2 \left[\frac{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})}{\widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})}\right] \triangleq \overline{\mathrm{SE}}(\boldsymbol{\theta}, n), \quad (15)
\end{aligned}$$

where $S_{1k}(\boldsymbol{\theta}) = \left(\sum_{m=1}^{M} \theta_{mk} \sqrt{\gamma_{mk}}\right)^2$ and $S_{2k}(\boldsymbol{\theta}) = \sum_{j=1}^{K} \sum_{m=1}^{M} \theta_{mj}^2 \beta_{mk}$. Specifically, the upper bound $\overline{\mathrm{SE}}(\boldsymbol{\theta}, n)$ is derived by neglecting certain interference terms in $\mathrm{SE}(\boldsymbol{\theta}, n)$. As the impact of pilot contamination decreases, the approximation between the upper bound and the original value will improve. The partial derivative of the upper bound with respect to n can be expressed as

$$\begin{aligned}
\frac{\partial \overline{\mathrm{SE}}(\boldsymbol{\theta}, n)}{\partial n} &= \frac{\overline{\tau_d}}{\ln 2} \sum_{k=1}^{K} \left[\frac{\frac{\partial \Gamma_n^2}{\partial n} S_{1k}(\boldsymbol{\theta}) + \frac{\partial \widetilde{\Gamma}_n}{\partial n} S_{2k}(\boldsymbol{\theta})}{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_N S_{2k}(\boldsymbol{\theta})} - \frac{\partial \widetilde{\Gamma}_n}{\partial n} \frac{1}{\widetilde{\Gamma}_n}\right] \\
&\overset{(a)}{=} \frac{\overline{\tau_d}}{\ln 2} \sum_{k=1}^{K} \left[\frac{2\Gamma_n^2 (\Upsilon(n - \frac{\alpha-1}{2}) - \Upsilon(n)) S_{1k}(\boldsymbol{\theta})}{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})}\right] \\
&+ \frac{\overline{\tau_d}}{\ln 2} \sum_{k=1}^{K} \left[\frac{\widetilde{\Gamma}_n (\Upsilon(n - \alpha) - \Upsilon(n)) S_{2k}(\boldsymbol{\theta})}{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})} - \frac{\partial \widetilde{\Gamma}_n}{\partial n} \frac{1}{\widetilde{\Gamma}_n}\right] \\
&\overset{(b)}{<} \frac{\overline{\tau_d}}{\ln 2} \sum_{k=1}^{K} \left[\frac{2\Gamma_n^2 (\Upsilon(n - \frac{\alpha-1}{2}) - \Upsilon(n)) S_{1k}(\boldsymbol{\theta})}{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})}\right] \\
&+ \frac{\overline{\tau_d}}{\ln 2} \sum_{k=1}^{K} \left[\frac{2\widetilde{\Gamma}_n (\Upsilon(n - \frac{\alpha-1}{2}) - \Upsilon(n)) S_{2k}(\boldsymbol{\theta})}{\Gamma_n^2 S_{1k}(\boldsymbol{\theta}) + \widetilde{\Gamma}_n S_{2k}(\boldsymbol{\theta})} - \frac{\partial \widetilde{\Gamma}_n}{\partial n} \frac{1}{\widetilde{\Gamma}_n}\right] \\
&= \frac{K\overline{\tau_d}}{\ln 2} \left[2\Upsilon(n - \frac{\alpha - 1}{2}) - \Upsilon(n) - \Upsilon(n - \alpha)\right] \\
&\overset{(c)}{\approx} \frac{K\overline{\tau_d}}{\ln 2} \ln\left(1 + \frac{1}{n - \alpha} + \frac{(\alpha - 1)^2}{4n(n - \alpha)}\right) \\
&+ \frac{K\overline{\tau_d}}{\ln 2} \frac{2n + \alpha^2 - \alpha}{2n(n - \alpha)(2n - \alpha + 1))} < \frac{2K\overline{\tau_d}}{\ln 2}. \quad (16)
\end{aligned}$$

Here, equation (a) is derived by taking the derivative of the Gamma function $\frac{\partial \Gamma(n)}{\partial n} = \Upsilon(n)\Gamma(n)$, where $\Upsilon(n)$ represents the digamma function [13]. Using the inequality $\Upsilon(n - \alpha) \leq \Upsilon(n - \frac{\alpha-1}{2})$, the upper bound (b) holds when $\alpha \geq -1$. In practical scenarios, α is typically set to -1, 0, or 1, thus making this upper bound applicable. Equation (c) is derived by approximating $\Upsilon(n) \approx \ln(n) - \frac{1}{2n}$ [13].

From equation (16), the upper bound of $\frac{\partial \overline{\mathrm{SE}}(\boldsymbol{\theta}, n)}{\partial n}$ is always greater than 0 and asymptotically approaches 0 as n increases. Specifically, the upper bound of SE increases with n, but its growth rate is primarily dictated by the logarithmic term $\ln(1 + \frac{1}{n-\alpha})$, resulting in a slower increment as n increases. Moreover, its fastest growth rate does not exceed $\frac{2K\overline{\tau_d}}{\ln 2}$, which is solely determined by the system parameters. On the other hand, from equation (12d), we can see that even if the AP transmit power is ignored, the system power consumption still increases linearly with n, and the lower bound growth rate depends on the circuit internal power $\sum_{m=1}^{M} P_{\mathrm{tc},m}$.

Overall, the power consumption increases linearly, while the SE growth rate shows a logarithmic decay trend. As n reaches a certain upper bound, the power consumption will dominate, and the excessive antenna activation would deteriorate EE, indicating the existence of the optimal number of activated antennas $n^\star$. Furthermore, if $\sum_{m=1}^{M} P_{\mathrm{tc},m} \geq \frac{2K\overline{\tau_d}}{\ln 2}$, the SE growth rate is lower than the power consumption even with the minimum number of activated antennas. Therefore, antenna activation should be carried out at the minimum level, it means n ought to be the lower bound of the feasible set $\mathcal{C}_2$, i.e.,

$$n^\star = \max(\lfloor \alpha \rfloor + 1, 1),$$

completing the proof. □

IV. Problem Reformulation and APG Algorithm

In this section, we will employ the monotone APG algorithm to efficiently solve the joint non-convex fractional optimization problem $\mathcal{P}_2$ formulated by the JAAPA scheme.

Theoretically, the classic APG algorithm is designed to cope with the unconstrained optimization problems [8]. In order to transform the original problem $\mathcal{P}_2$ into an amenable form for

Algorithm 1 The JAAPA scheme with APG execution

Input: $\overline{\boldsymbol{\theta}}^{(0)}, N_{\max}, \xi = 0.1, \epsilon > 0, \zeta = 10, 0 < \alpha_y, \alpha_\theta < \frac{1}{L_f}$.
Output: Stationary point $\overline{\boldsymbol{\theta}}^{\star}$
1: **Initialization:** $\overline{\boldsymbol{\theta}}^{(1)} = \mathbf{z}^{(1)} = \overline{\boldsymbol{\theta}}^{(0)}$
2: **repeat**
3: Set: $i = 1, t^{(0)} = t^{(1)} = 1$
4: **repeat**
5: Update $\mathbf{y}^{(i)}$ via (18)
6: Update $\mathbf{z}^{(i)}$ and $\mathbf{v}^{(i)}$ via (19) and (20) respectively
7: Update $\overline{\boldsymbol{\theta}}^{(i)}$ via (21)
8: $t^{(i+1)} = \frac{\sqrt{4(t^{(i)})^2+1}+1}{2}$, $i = i+1$
9: **until** $|\frac{F_\xi(\overline{\boldsymbol{\theta}}^{(i)}) - F_\xi(\overline{\boldsymbol{\theta}}^{(i-1)})}{F_\xi(\overline{\boldsymbol{\theta}}^{(i)})}| \le \epsilon$
10: Updata the strating point for the next iteration: $\overline{\boldsymbol{\theta}}^{(0)} = \overline{\boldsymbol{\theta}}^{(1)} = \mathbf{z}^{(1)} = \overline{\boldsymbol{\theta}}^{(i)}$
11: Increase the penalty coefficient $\xi = \xi \times \zeta$
12: **until** $|\frac{Q(\overline{\boldsymbol{\theta}}^{(i)}) - Q(\overline{\boldsymbol{\theta}}^{(i-1)})}{Q(\overline{\boldsymbol{\theta}}^{(i)})}| \le \epsilon$
13: **return** $\overline{\boldsymbol{\theta}}^{\star} = \overline{\boldsymbol{\theta}}^{(i)}$

APG execution, the exterior penalty method is considered to handle the SE constraints, resulting in the following problem

$$\mathcal{P}_3: \quad \max_{\boldsymbol{\theta}\in\mathcal{C}_1, n\in\mathcal{C}_2} \; F_\xi(\boldsymbol{\theta}, n) \triangleq \overline{\mathrm{EE}}(\boldsymbol{\theta}, n) - \xi Q(\boldsymbol{\theta}, n). \quad (17)$$

Specifically, $Q(\boldsymbol{\theta}, n) = \sum_{k=1}^{K} \left[\max\left(0, \mathrm{SE}_{th} - \mathrm{SE}_k\left(\boldsymbol{\theta}, n\right)\right)\right]^2$ is the quadratic loss function corresponding to the constraint (11b). This loss imposes a penalty to force the solution move towards the feasible region as the constraint is not satisfied. $\xi > 0$ is the penalty coefficient, controling the magnitude of the penalty and increases progressively during the iterations.

To facilitate APG execution, we concatenate $\boldsymbol{\theta}$ and n into a vector, denoted as $\overline{\boldsymbol{\theta}} = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_K^T, n]^T \in \mathbb{R}^{MK+1}$. Let i represents the iteration index, APG uses momentum acceleration to compute the extrapolation point

$$\mathbf{y}^{(i)} = \overline{\boldsymbol{\theta}}^{(i)} + \frac{t^{(i-1)}}{t^{(i)}}(\mathbf{z}^{(i)} - \overline{\boldsymbol{\theta}}^{(i)}) + \frac{t^{(i-1)-1}}{t^{(i)}}(\overline{\boldsymbol{\theta}}^{(i)} - \overline{\boldsymbol{\theta}}^{(i-1)}), \quad (18)$$

here, $t^{(i)}$ is the extrapolation parameter and computed recursively. From (18), the extrapolation point undergoes gradient ascent within a given step size α_y to maximize the objective function. Meanwhile, the proximal operation is performed to ensure that the obtained point fall within the feasible set, i.e.,

$$\begin{aligned}\mathbf{z}^{(i+1)} &= \mathrm{prox}_{\mathcal{C}}\left(\mathbf{y}^{(i)} + \alpha_y \nabla F_\xi(\mathbf{y}^{(i)})\right) \\ &= \arg\min_{\overline{\boldsymbol{\theta}}^{(i+1)}\in\mathcal{C}} \|\overline{\boldsymbol{\theta}}^{(i+1)} - \left(\mathbf{y}^{(i)} + \alpha_y \nabla F_\xi(\mathbf{y}^{(i)})\right)\|^2, \end{aligned} \quad (19)$$

where $\mathcal{C} \triangleq \mathcal{C}_1 \bigcap \mathcal{C}_2$, and α_y can be set through backtracking line method to find the largest possible value [9]. The solution of (19) requires gradient and proximal projection calculations, which are similar to [7], and is omitted here for simplicity.

On the other hand, the point $\mathbf{y}^{(i)}$ may be a bad extrapolation, causing the solution to be trapped near saddle point [8]. Therefore, APG also employs the same operation as (19) to set up an additional monitor to correct unfavorable point

$$\mathbf{v}^{(i+1)} = \mathrm{prox}_{\mathcal{C}}\left(\overline{\boldsymbol{\theta}}^{(i)} + \alpha_\theta \nabla F_\xi(\overline{\boldsymbol{\theta}}^{(i)})\right). \quad (20)$$

Then, by comparing the accelerated point $\mathbf{z}^{(i+1)}$ and unaccelerated point $\mathbf{v}^{(i+1)}$, the point with a larger objective function value is selected as the next iteration point

$$\overline{\boldsymbol{\theta}}^{(i+1)} = \begin{cases} \mathbf{z}^{(i+1)}, & \text{if } F_\xi(\mathbf{z}^{(i+1)}) \ge F_\xi(\mathbf{v}^{(i+1)}), \\ \mathbf{v}^{(i+1)}, & \text{otherwise.} \end{cases} \quad (21)$$

Repeat the above update process until the objective function stabilizes. Based on [7, Proposition 1], $F_\xi(\boldsymbol{\theta}, n)$ possesses proper property with L_f-Lipschitz continuous gradient and bounded from above. Consequently, during the iterative solution process by using APG, $F_\xi(\boldsymbol{\theta}, n)$ will exhibit a monotonically non-decreasing behavior and ultimately converge to a approximate stationary point for $\mathcal{P}_2$. Besides, the complexity order for performing the first-order APG algorithm for JAAPA scheme is $\mathcal{O}(K^2M)$ per iteration. In contrast, the second-order cone (SOC) general solvers based on SCA have an iteration complexity of $\mathcal{O}(\sqrt{K+M}K^4M^3)$ [3], [5], [6], [10].

V. Numerical Results

In this section, we evaluate the performance of the proposed JAAPA scheme in terms of EE. To confirm the feasibility of the APG algorithm for the proposed scheme in large-scale systems, we set $M = 100, K = 40, \tau_u = 40, \tau_c = 200$. If not otherwise mentioned, the other parameters in our system follow the same settings as in [7]. The initial values of power allocation coefficients follows the maximal-ratio principle [10], i.e. $\theta_{mk}^{(0)} = \sqrt{\widetilde{\Gamma}_N^{-1}\gamma_{mk}(\sum_{k=1}^{K}\gamma_{mk})^{-1}}, \forall m, \forall k$. Additionally, the initial number of activated antennas for all APs is simply set to $\max(\lfloor\alpha\rfloor + 1, 1)$.

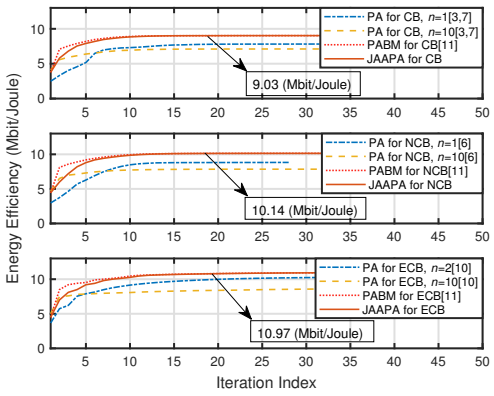

Fig. 1. Downlink EE versus the number of iteration index.

In Fig. 1, the proposed JAAPA scheme is applied to EE optimization problem by algorithm 1 for CB, NCB, and ECB precoding, respectively. The comparison benchmarks include approaches that utilize only power allocation (PA) [3], [6], [7], [10], as well as a joint optimization algorithm that employs the bisection method (BM) for both power allocation and antenna activation, namely PABM [11]. Obviously, the JAAPA scheme

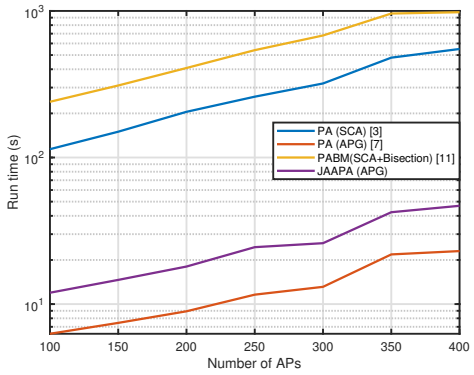

Fig. 2. Run time versus the number of APs (CB precoding).

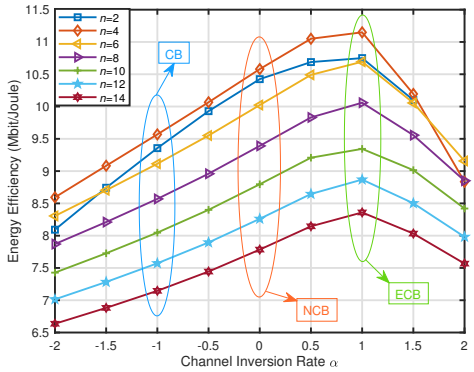

Fig. 3. Downlink EE versus channel inversion rate α.

can converge to PABM after a few iterations, achieving a significant performance improvement of approximately 20% compared to the approach that relies solely on PA[2]. Furthermore, as shown in Fig. 2, the proposed JAAPA (APG) can reduce runtime by approximately 20 times compared to the PABM, emerged as a pragmatic way for large-scale systems.

As seen in Fig. 3, the variations in EE under different channel inversion rates (corresponding to various CB variants) is compared. To provide an intuitive understanding of the impact of antenna activation on green communication, the number of activated antennas is varied uniformly. Based on this comparison, we can draw the following two conclusions:

- The highest EE is always achieved within ECB precoding among different CB variants. This is attributed to the near-deterministic nature of the effective channel, resulting from the normalization by the squared norm of the channel information [10].
- EE exhibits an almost linear degradation trend with excessive growth in antenna activation. Consequently, the proposed algorithm can identify $n^\star$, which holds significant importance in practical applications.

Fig. 4 illustrates the impact of $P_{\text{tc,m}}$ on the optimal number of activated antennas $n^\star$. Without loss of generality, the PA

[2]Since the performance achieved by SCA and APG for power allocation is equivalent [7], we only utilize APG for performance comparison in Fig. 1. In Fig. 2, we further distinguish between the two optimization algorithms during the complexity comparison.

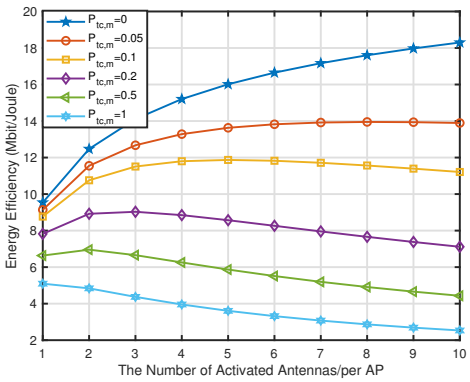

Fig. 4. Downlink EE versus the number of activated antennas under different internal power $P_{\text{tc,m}}$ (CB precoding).

(APG) scheme within CB precoding is adopted here [7]. Evidently, as $P_{\text{tc,m}}$ increases, $n^\star$ steadily approaches the lower bound of $\mathcal{C}_2$. Based on Theorem 1, we can conclude that if $P_{\text{tc,m}} \geq 0.92, \forall m$, then $n^\star = \max(\lfloor \alpha \rfloor + 1, 1) = 1$, which is consistent with Fig. 4. Additionally, note that when $P_{\text{tc,m}} = 0$, EE is able to increase indefinitely with n. However, it is evident that circuit operation incurs power consumption. Thus, finding ways to reduce this consumption to provide substantial improvements in EE is a viable direction for future research.

REFERENCES

[1] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective Multiple Antenna Technologies for Beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637-1660, Aug. 2020.

[2] Z. Liu, J. Zhang, Z. Liu, D. W. K. Ng and B. Ai, "Joint Cooperative Clustering and Power Control for Energy-Efficient Cell-Free XL-MIMO with Multi-Agent Reinforcement Learning," *IEEE Trans. Commun.*, to appear, 2024.

[3] 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.

[4] G. Dong, H. Zhang, S. Jin and D. Yuan, "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.

[5] L. D. Nguyen, T. Q. Duong, H. Q. Ngo and K. Tourki, "Energy Efficiency in Cell-Free Massive MIMO with Zero-Forcing Precoding Design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871-1874, Aug. 2017.

[6] B. Yan, N. Zhou and Z. Wang, "Energy Efficiency Optimization in Cell-Free Massive MIMO With Normalized Conjugate Beamforming," *In proc. IEEE Int. Conf. Commun. Technol.*, Wuxi, China, pp. 346-351, 2023.

[7] T. C. Mai, H. Q. Ngo and L. -N. Tran, "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.

[8] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," *Adv. Neural Inf. Process. Syst.*,,vol. 28, 2015, pp. 379-387

[9] 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.

[10] G. Interdonato, H. Q. Ngo and E. G. Larsson, "Enhanced Normalized Conjugate Beamforming for Cell-Free Massive MIMO," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 2863-2877, May 2021.

[11] Z. Chang et al., "Energy-Efficient Resource Allocation for Wireless Powered Massive MIMO System With Imperfect CSI," *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 2, pp. 121-130, June 2017.

[12] G. Interdonato and S. Buzzi, "Conjugate Beamforming with Fractional-Exponent Normalization and Scalable Power Control in Cell-Free Massive MIMO," *in Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, Lucca, Italy, pp.396-400, 2021.

[13] Mortici, Cristinel. "New approximations of the gamma function in terms of the digamma function." *Applied Mathem. Lett.* 23.1 (2010): 97-100.