Nonlinear Power Amplifier-Resilient Cell-Free Massive MIMO: A Joint Optimization Approach

Wei Jiang, *Senior Member, IEEE,* and Hans Dieter Schotten, *Member, IEEE*

Abstract—This letter proposes an analytical model to study analyzes the effects of power amplifiers (PAs) on the downlink of cell-free massive MIMO systems. We model signal transmission incorporating nonlinear PA distortion and derive a unified spectral efficiency (SE) expression applicable to arbitrary precoding schemes. To combat PA-induced performance loss, a tractable joint optimization approach, as well as its low-complexity alternative, for user-centric clustering and max-min power control is proposed based on a conservative approximation.

Index Terms—Cell-free massive MIMO, max-min power control, nonlinear power amplifier, user association, user fairness

I. Introduction

CELL-FREE (CF) massive multi-input multi-output (CF-mMIMO) has garnered significant attention recently. Nevertheless, power amplifiers (PAs)—critical components in wireless transmitters—introduce nonlinear signal distortion, particularly when operating near their saturation point. Ignoring this nonlinearity can lead to overestimated system performance and cause bias into algorithm design.

To date, only a few existing works provide initial insights. Mokhtari et al. [1] quantified sum-rate degradation under PA distortion using maximum ratio (MR) precoding with equal power allocation. Subsequent studies by Jadidi et al. [2] and Khoueini et al. [3] investigated uplink and downlink impacts, respectively, assuming single-antenna access points (APs) and MR precoding. While these studies offer valuable theoretical insights, several key dimensions remain unexplored, e.g.,

- *Scalability limits*: [1]–[3] use conventional CF configuration where all APs serve all users, neglecting the scalability gain of user-centric dynamic clustering [4];
- *Precoding generality constraint*: [1]–[3] focus solely on MR precoding, omitting state-of-the-art techniques such as zero-forcing (ZF) and minimum mean square error (MMSE) precoding [5].

To bridge these gaps, this letter presents a unified analytical framework and a novel design for PA-resilient CF-mMIMO systems. Specifically, we develop a downlink signal model that captures nonlinear PA distortion, and derive a spectral efficiency (SE) expression applicable to arbitrary linear precoding schemes. Our analysis reveals that the achievable SE under PA-induced distortion critically depends on both user-centric clustering and power control. Leveraging this insight, we propose a tractable approximation approach, that adaptively handles AP-user association and max–min fairness transmission power to effectively suppress the performance loss due to nonlinear PA distortion. The novelty of our work lies in moving beyond prior studies that focus either on PA-agnostic clustering [4], [6] and power optimization [5], [7] or on PA distortion analysis in non-clustering systems [1]–[3].

The remainder of this letter is structured as follows. Section II introduces the system model. Section III presents the SE analysis under PA distortion. Section IV formulates the joint optimization design. Section V analyzes computational complexity and offers a low-complexity algorithm.

II. System Model

We consider a CF-mMIMO system comprising L APs, each equipped with N_a co-located antennas, distributed throughout a coverage area to serve K users, where the total number of antennas $M = L \times N_a \gg K$. Each user equipment (UE) is generally equipped with a single antenna. The sets of indices for APs and users are denoted by $\mathbb{L} = \{1, \dots, L\}$ and $\mathbb{K} = \{1, \dots, K\}$, respectively. The wireless channel from AP l, $\forall l \in \mathbb{L}$ to UE k, $\forall k \in \mathbb{K}$ is denoted by $\mathbf{h}_{kl} \in \mathbb{C}^{N_a}$. Adopting the block fading assumption, each *coherent block*—a time-frequency span of τ_c symbols—maintains a quasi-static channel response. In practical deployments, closely spaced antennas at an AP exhibit spatial correlation due to their physical proximity. Hence, each coherence block applies an independent realization from *correlated* Rayleigh fading, defined as $\mathbf{h}_{kl} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{kl})$, where $\mathbf{R}_{kl} = \mathbb{E}[\mathbf{h}_{kl}\mathbf{h}_{kl}^H]$ stands for the spatial correlation matrix. Using linear MMSE estimation [5], the estimated channel $\hat{\mathbf{h}}_{kl}$ follows a complex Gaussian distribution $\hat{\mathbf{h}}_{kl} \sim \mathcal{CN}\left(\mathbf{0}, p_u\tau_p\mathbf{R}_{kl}\mathbf{\Gamma}_{kl}^{-1}\mathbf{R}_{kl}\right)$, where $\mathbf{\Gamma}_{kl} = p_u\tau_p \sum_{k' \in \mathcal{P}_k} \mathbf{R}_{k'l} + \sigma_z^2\mathbf{I}_{N_a}$, and p_u denotes the maximal transmit power of the UEs, τ_p is the pilot sequence length, $\mathcal{P}_k$ is the set of users sharing the same pilot as user k, and σ_z^2 indicates the variance of thermal noise. The estimation error, defined as $\tilde{\mathbf{h}}_{kl} = \mathbf{h}_{kl} - \hat{\mathbf{h}}_{kl}$, is attributed to both noise and pilot contamination, following $\mathcal{CN}(\mathbf{0}, \mathbf{\Theta}_{kl})$, where the associated error covariance matrix is given by $\mathbf{\Theta}_{kl} = \mathbb{E}\left[\tilde{\mathbf{h}}_{kl}\tilde{\mathbf{h}}_{kl}^H\right] = \mathbf{R}_{kl} - p_u\tau_p\mathbf{R}_{kl}\mathbf{\Gamma}_{kl}^{-1}\mathbf{R}_{kl}$.

III. Downlink Transmission

In the downlink, the data symbols intended for the K users are assumed to be independent, zero-mean random variables with unit variance. These symbols are jointly expressed as

$\mathbf{x} = [x_1, \ldots, x_K]^T$, satisfying $\mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{I}_K$. Under the user-centric dynamic clustering strategy, each AP maintains a set of associated users, denoted by $\mathbb{K}_l$ for AP l, where $\mathbb{K}_l \subseteq \mathbb{K}$. Each AP spatially multiplexes the symbols of its associated users $k \in \mathbb{K}_l$, producing

$$\mathbf{s}_l = \sqrt{p_a} \sum_{k \in \mathbb{K}} u_{kl} \sqrt{\eta_{kl}} \mathbf{w}_{kl} x_k, \tag{1}$$

where p_a denotes the AP power constraint, η_{kl} is the power coefficient for AP l to user k, u_{kl} is the association indicator for user-centric clustering, which equals 1 if user k is served by AP l (i.e., $k \in \mathbb{K}_l$), and 0 otherwise. The vector $\mathbf{w}_{kl} \in \mathbb{C}^{N_a}$ denotes the precoding vector employed by AP l for user k, with the normalization constraint $\mathbb{E}\left[|\mathbf{w}_{kl}|^2\right] = 1$.

A. Nonlinear Power Amplification

A common approach to model PA nonlinearity is the Bussgang decomposition [8]. For the feeding signal $\mathbf{s}_l$, the amplified transmit signal at AP l is given by

$$\mathbf{x}_l = \mathcal{G}(\mathbf{s}_l) = \alpha_l \sqrt{p_a} \sum_{k \in \mathbb{K}} u_{kl} \sqrt{\eta_{kl}} \mathbf{w}_{kl} x_k + \mathbf{d}_l, \tag{2}$$

where:

- $\mathcal{G}(\cdot)$: A general nonlinear PA function.
- α_l is a complex linear gain, and $\alpha_l = \mathbb{E}\left[\mathbf{s}_l^H \mathcal{G}(\mathbf{s}_l)\right] / \mathbb{E}[\|\mathbf{s}_l\|^2]$.
- $\mathbf{d}_l \sim \mathcal{CN}(0, \sigma_d^2 \mathbf{I}_{N_a})$ is complex Gaussian distortion, uncorrelated with $\mathbf{s}_l$ (i.e., $\mathbb{E}[\mathbf{s}_l^H \mathbf{d}_l] = \mathbf{0}$), where

$$\sigma_d^2 = \beta_l \mathbb{E}\left[\|\mathbf{s}_l\|^2\right] = \beta_l p_a \sum_{k \in \mathbb{K}} u_{kl} \eta_{kl} \tag{3}$$

stands for distortion variance, and β_l represents distortion-to-signal ratio.

Thus, the actual transmitted power is expressed by

$$\mathbb{E}[\|\mathbf{x}_l\|^2] = \mathbb{E}\left[\left\|\alpha_l \sqrt{p_a} \sum_{k \in \mathbb{K}} u_{kl} \sqrt{\eta_{kl}} \mathbf{w}_{kl} x_k + \mathbf{d}_l\right\|^2\right]$$
$$= |\alpha_l|^2 p_a \sum_{k \in \mathbb{K}} u_{kl} \eta_{kl} \mathbb{E}[\|\mathbf{w}_{kl}\|^2] + \mathbb{E}[\|\mathbf{d}_l\|^2], \tag{4}$$

applying the property of binary variables $u_{kl}^2 = u_{kl}$. The distortion power is $\mathbb{E}[\|\mathbf{d}_l\|^2] = \text{tr}\left(\mathbb{E}[\mathbf{d}_l \mathbf{d}_l^H]\right) = \text{tr}\left(\sigma_d^2 \mathbf{I}_{N_a}\right) = \beta_l p_a N_a \sum_{k \in \mathbb{K}} u_{kl} \eta_{kl}$, since $\mathbb{E}[\|\mathbf{w}_{kl}\|^2] = 1$. Now, (4) simplifies to $\mathbb{E}[\|\mathbf{x}_l\|^2] = \left(|\alpha_l|^2 p_a + \beta_l p_a N_a\right) \sum_{k \in \mathbb{K}} u_{kl} \eta_{kl}$. To ensure the AP respects its power budget, the transmitted power must satisfy $\mathbb{E}[\|\mathbf{x}_l\|^2] \leq p_a$, imposing the constraint:

$$\sum_{k \in \mathbb{K}} u_{kl} \eta_{kl} \leq \frac{1}{|\alpha_l|^2 + \beta_l N_a}, \quad \forall l \in \mathbb{L}. \tag{5}$$

B. Precoding Schemes

To ensure scalability, each AP l independently performs local signal processing using its own channel estimates, denoted by $\hat{\mathbf{H}}_l = \left[\hat{\mathbf{h}}_{1l}, \cdots, \hat{\mathbf{h}}_{Kl}\right] \in \mathbb{C}^{N_a \times K}$. The applicable precoding schemes identified include:

1) **MR**: This simple approach, a.k.a. conjugate beamforming, sets the precoding matrix at AP l as $\mathbf{V}_l^{mr} = \hat{\mathbf{H}}_l^*$ to maximize the desired signal strength.
2) **ZF**: Aiming to cancel multi-user interference, the ZF precoder is the pseudo-inverse of $\hat{\mathbf{H}}_l$, given by $\mathbf{V}_l^{zf} = \hat{\mathbf{H}}_l^* (\hat{\mathbf{H}}_l^T \hat{\mathbf{H}}_l^*)^{-1}$ so as to meet $\hat{\mathbf{H}}_l^T \mathbf{V}_l^{zf} = \mathbf{I}_K$. This condition holds exclusively when the channel matrix satisfies the

full-rank assumption, when $\hat{\mathbf{H}}_l$ has full column rank, requiring $N_a \geq K$.

3) **Regularized ZF (RZF)**: In practice, due to scalability requirement, APs may be equipped with fewer antennas than users ($N_a < K$), rendering ZF inapplicable. To handle rank-deficient cases, regularized ZF is used, i.e., $\mathbf{V}_l^{rzf} = \hat{\mathbf{H}}_l^* \left(\hat{\mathbf{H}}_l^T \hat{\mathbf{H}}_l^* + \sigma_z^2 \mathbf{I}_K\right)^{-1}$, where the regularization term $\sigma_z^2 \mathbf{I}_K$ ensures matrix invertibility.
4) **MMSE**: The balance between interference suppression and noise enhancement is achieved by minimizing the mean squared error between the transmitted symbol and the received signal. Its precoding matrix is given by $\mathbf{V}_l^{mmse} = \left(p_u \hat{\mathbf{H}}_l \mathbf{E}_l \hat{\mathbf{H}}_l^H + p_u \sum_{k \in \mathbb{K}} \eta_{kl} \mathbf{\Theta}_{kl} + \sigma_z^2 \mathbf{I}_{N_a}\right)^{-1} \hat{\mathbf{H}}_l$, where $\mathbf{E}_l = \text{diag}(\eta_{1l}, \ldots, \eta_{Kl})$ is a diagonal matrix of power coefficients.

Let $\mathbf{v}_{kl} \in \mathbb{C}^{N_a}$ denote the k-th column of $\mathbf{V}_l \in \mathbb{C}^{N_a \times K}$, i.e., $\mathbf{v}_{kl} = \mathbf{V}_l(:, k)$. The required precoding vector $\mathbf{w}_{kl}$ is obtained by normalizing $\mathbf{v}_{kl}$ as $\mathbf{w}_{kl} = \frac{\mathbf{v}_{kl}}{\|\mathbf{v}_{kl}\|}$, $\forall l \in \mathbb{L}, \forall k \in \mathbb{K}$, ensuring it satisfies the normalization condition $\mathbb{E}\left[\|\mathbf{w}_{kl}\|^2\right] = 1$.

C. Performance Analysis

The received signal at user k is $y_k = \sum_{l \in \mathbb{L}} \mathbf{h}_{kl}^T \mathbf{x}_l + n_k$, where noise $n_k \sim \mathcal{CN}(0, \sigma_z^2)$. The expression can be expanded as

$$y_k = \sqrt{p_a} \sum_{l \in \mathbb{L}} \alpha_l \sum_{k' \in \mathbb{K}} u_{k'l} \sqrt{\eta_{k'l}} \mathbf{h}_{kl}^T \mathbf{w}_{k'l} x_{k'} + \sum_{l \in \mathbb{L}} \mathbf{h}_{kl}^T \mathbf{d}_l + n_k. \tag{6}$$

Massive MIMO systems typically operate in time-division duplexing to avoid prohibitive downlink training overhead that scales with M. Therefore, users typically do not know channel estimates due to the absence of downlink pilots, making coherent detection impractical. Instead, signal detection is performed using large-scale fading decoding (LSFD), where the statistical mean $\mathbb{E}[\mathbf{w}_{kl}^H \mathbf{h}_{kl}]^1$, $\forall k, l$, serves as an approximation. This statistical mismatch leads to an additional performance degradation known as channel uncertainty error. To facilitate the derivation, (6) is decomposed into

$$y_k = \underbrace{\sqrt{p_a} \sum_{l \in \mathbb{L}} \alpha_l u_{kl} \sqrt{\eta_{kl}} \mathbb{E}[\mathbf{h}_{kl}^T \mathbf{w}_{kl}] x_k}_{\mathcal{S}_1: \text{ desired signal over channel statistics}}$$
$$+ \underbrace{\sqrt{p_a} \sum_{l \in \mathbb{L}} \alpha_l u_{kl} \sqrt{\eta_{kl}} \left(\mathbf{h}_{kl}^T \mathbf{w}_{kl} - \mathbb{E}[\mathbf{h}_{kl}^T \mathbf{w}_{kl}]\right) x_k}_{\mathcal{J}_1: \text{ channel uncertainty error}} \tag{7}$$
$$+ \underbrace{\sqrt{p_a} \sum_{l \in \mathbb{L}} \alpha_l \sum_{k' \in \mathbb{K} \setminus \{k\}} u_{k'l} \sqrt{\eta_{k'l}} \mathbf{h}_{kl}^T \mathbf{w}_{k'l} x_{k'}}_{\mathcal{J}_2: \text{ inter-user interference}} + \underbrace{\sum_{l \in \mathbb{L}} \mathbf{h}_{kl}^T \mathbf{d}_l}_{\mathcal{J}_3: \text{PA distortion}} + n_k.$$

Proposition 1: The achievable SE of user k is $R_k = \mathbb{E}\left[\log_2(1 + \gamma_k)\right]$, where the expectation is over channel real-

---

[1] Since the precoder $\mathbf{w}_{kl}$ is a function of the channel estimate $\hat{\mathbf{h}}_{kl}$ and is therefore uncorrelated with the channel estimation error $\tilde{\mathbf{h}}_{kl}$, it follows that $\mathbb{E}[\mathbf{w}_{kl}^H \mathbf{h}_{kl}] = \mathbb{E}[\mathbf{w}_{kl}^H (\hat{\mathbf{h}}_{kl} + \tilde{\mathbf{h}}_{kl})] = \mathbb{E}[\mathbf{w}_{kl}^H \hat{\mathbf{h}}_{kl}]$. This expectation can be derived from channel statistics, see, e.g., [6, Corol. 3].

izations and the instantaneous effective SINR is

$$\gamma_k = \frac{\left|\sum_{l\in\mathbb{L}} \alpha_l u_{kl} \sqrt{\eta_{kl}}\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2}{\left\{\begin{array}{l} \sigma_z^2/p_a + \sum_{k'\in\mathbb{K}}\left(\sum_{l\in\mathbb{L}} u_{k'l}\eta_{k'l}|\alpha_l|^2\mathbb{E}[|\mathbf{h}_{kl}^T\mathbf{w}_{k'l}|^2]\right) \\ - \sum_{l\in\mathbb{L}} u_{kl}\eta_{kl}|\alpha_l|^2 \left|\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2 + \sum_{l\in\mathbb{L}}\beta_l\left(\sum_{k'\in\mathbb{K}} u_{k'l}\eta_{k'l}\right)\mathrm{tr}(\mathbf{R}_{kl}) \end{array}\right\}} \tag{8}$$

Proof: Using the standard capacity lower bounds (cf. [6, Prop. 3]), the SINR is given by $\gamma_k = \frac{|\mathcal{S}_1|^2}{\mathbb{E}[|\mathcal{J}_1+\mathcal{J}_2+\mathcal{J}_3|^2]+\sigma_z^2}$, where the power gain of the desired signal is

$$|\mathcal{S}_1|^2 = p_a \left|\sum_{l\in\mathbb{L}} \alpha_l u_{kl} \sqrt{\eta_{kl}}\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2. \tag{9}$$

Due to the independence among data symbols and signal distortion, $\mathcal{J}_1$, $\mathcal{J}_2$, and $\mathcal{J}_3$ in (7) are uncorrelated. This implies that $\mathbb{E}\left[|\mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3|^2\right] = \mathbb{E}\left[|\mathcal{J}_1|^2\right] + \mathbb{E}\left[|\mathcal{J}_2|^2\right] + \mathbb{E}\left[|\mathcal{J}_3|^2\right]$. We now compute the variance of $\mathcal{J}_1$ (cf. [7, Th. 1]) as

$$\mathbb{E}\left[|\mathcal{J}_1|^2\right] = p_a \sum_{l\in\mathbb{L}} u_{kl}\eta_{kl}|\alpha_l|^2\mathbb{E}\left[\left|\mathbf{h}_{kl}^T\mathbf{w}_{kl} - \mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2\right] \tag{10}$$

$$= p_a \sum_{l\in\mathbb{L}} u_{kl}\eta_{kl}|\alpha_l|^2\left(\mathbb{E}\left[\left|\mathbf{h}_{kl}^T\mathbf{w}_{kl}\right|^2\right] - \left|\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2\right).$$

Next, the variance of $\mathcal{J}_2$ is computed as

$$\mathbb{E}\left[|\mathcal{J}_2|^2\right] = p_a \sum_{l\in\mathbb{L}} |\alpha_l|^2 \sum_{k'\in\mathbb{K}\setminus\{k\}} u_{k'l}\eta_{k'l}\mathbb{E}[|\mathbf{h}_{kl}^T\mathbf{w}_{k'l}|^2]. \tag{11}$$

Finally,

$$\mathbb{E}\left[|\mathcal{J}_3|^2\right] = \sum_{l\in\mathbb{L}} \sigma_d^2\mathbb{E}[\|\mathbf{h}_{kl}\|^2] = p_a \sum_{l\in\mathbb{L}}\beta_l\left(\sum_{k'\in\mathbb{K}} u_{k'l}\eta_{k'l}\right)\mathrm{tr}(\mathbf{R}_{kl}). \tag{12}$$

Applying $\mathcal{S}_1$, $\mathcal{J}_1$, $\mathcal{J}_2$, and $\mathcal{J}_3$ yields (8). ∎

IV. JOINT OPTIMIZATION OF USER ASSOCIATION AND MAX-MIN POWER ALLOCATION

Analyzing (3) and (8) reveal that user-centric AP association (u_{kl}) and power control (η_{kl}) decide the variance of PA distortion and the resultant SINR. To perserve uniformly high-quality service under PA-induced distortion, we propose jointly optimizing user association and max-min power allocation. The problem formulation is

$$\max_{\mathbf{U},\,\boldsymbol{\eta}} \quad \min_{k\in\mathbb{K}} \quad \gamma_k \tag{13}$$

$$\text{s.t.} \quad \begin{cases} \sum_{k\in\mathbb{K}} u_{kl}\eta_{kl} \leq \dfrac{1}{|\alpha_l|^2 + \beta_l N_a}, & \forall l\in\mathbb{L}, \\ \eta_{kl} \geq 0, \quad u_{kl}\in\{0,1\}, & \forall k\in\mathbb{K}, l\in\mathbb{L}, \end{cases}$$

where $\mathbf{U} = [u_{kl}]$ is the binary association matrix, and $\boldsymbol{\eta} = [\eta_{kl}]$ is the power allocation matrix. The numerator of γ_k is a square of a sum of square roots, which is not convex in $\boldsymbol{\eta}$. To convexify this term, we introduce auxiliary variables v_{kl}, collectively denoted as $\boldsymbol{\nu} = \{v_{kl}\}_{k\in\mathbb{K},l\in\mathbb{L}}$, and impose the following constraint:

$$v_{kl}^2 \leq \eta_{kl}, \quad \forall k\in\mathbb{K}, l\in\mathbb{L}. \tag{14}$$

It can be equivalently expressed as a convex rotated second-order cone (SOC) constraint, given by

$$\left\|[2v_{kl};\ 1 - \eta_{kl}]\right\| \leq 1 + \eta_{kl}, \quad \forall k\in\mathbb{K}, l\in\mathbb{L}. \tag{15}$$

This standard reformulation is efficient for embedding our problem within a convex SOC programming framework. Due to the objective's incentive to maximize power, (14) is a tight relaxation, i.e., $v_{kl} = \sqrt{\eta_{kl}}$ at optimum. Then, the effective received signal for user k can be denoted by

$$\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu}) = \sum_{l\in\mathbb{L}} \alpha_l u_{kl} v_{kl}\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}], \tag{16}$$

which becomes linear in v_{kl}. Define the additive terms in the denominator of (8) as

$$\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}) = \sum_{k'\in\mathbb{K}}\left(\sum_{l\in\mathbb{L}} u_{k'l}v_{k'l}^2|\alpha_l|^2\mathbb{E}\left[|\mathbf{h}_{kl}^T\mathbf{w}_{k'l}|^2\right]\right) \tag{17}$$
$$+ \sum_{l\in\mathbb{L}}\beta_l\left(\sum_{k'\in\mathbb{K}} u_{k'l}v_{k'l}^2\right)\mathrm{tr}(\mathbf{R}_{kl}) + \sigma_z^2/p_a,$$

and the subtracted term as $C_k(\mathbf{U},\boldsymbol{\nu}) = \sum_{l\in\mathbb{L}} u_{kl}v_{kl}^2|\alpha_l|^2 \left|\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2$. Thus, the denominator becomes $\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}) - C_k(\mathbf{U},\boldsymbol{\nu})$. Introducing a slack variable γ_t to represent the common SINR target, the constraint $\gamma_k \geq \gamma_t$ is rewritten to

$$|\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})|^2 \geq \gamma_t\left(\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}) - C_k(\mathbf{U},\boldsymbol{\nu})\right), \quad \forall k\in\mathbb{K}. \tag{18}$$

The constraint in (18) is still non-convex due to the quadratic term $|\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})|^2$. To convexify it, we build a stricter constraint:

$$(\Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\})^2 \geq \gamma_t\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}), \quad \forall k\in\mathbb{K}. \tag{19}$$

Remark 1: Since $|\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})| \geq \Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\}$ and $C_k(\mathbf{U},\boldsymbol{\nu}) \geq 0$, it follows that $(\Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\})^2 \geq \gamma_t\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}) \implies |\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})|^2 \geq \gamma_t(\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}) - C_k(\mathbf{U},\boldsymbol{\nu}))$, i.e., (19) is a sufficient (conservative) condition for (18). Thus feasibility under (19) guarantees feasibility under (18).

Lemma 1: The convexification above guarantees a close approximation because

$$|\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})| \approx \Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\}, \quad \text{and} \quad C_k(\mathbf{U},\boldsymbol{\nu}) \ll \mathcal{B}_k(\mathbf{U},\boldsymbol{\nu}).$$

Proof: The first approximation follows from typical precoder phase alignment (see Sec. III-B): precoder phases are chosen to align with their channels, so $\mathbf{h}_{kl}^T\mathbf{w}_{kl}$ are real — for example, with MR precoding one has $\mathbf{h}_{kl}^T\mathbf{w}_{kl} = \|\mathbf{h}_{kl}\|$. Consequently the sum in (16) that defines $\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})$ is composed mainly of real-valued contributions, making $|\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})| \approx \Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\}$ a reasonable approximation. For the second relation note that $\mathbb{E}[|\mathbf{h}_{kl}^T\mathbf{w}_{kl}|^2] - \left|\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2 = \mathrm{Var}(\mathbf{h}_{kl}^T\mathbf{w}_{kl}) \geq 0$, so $\mathbb{E}[|\mathbf{h}_{kl}^T\mathbf{w}_{kl}|^2] \geq \left|\mathbb{E}[\mathbf{h}_{kl}^T\mathbf{w}_{kl}]\right|^2$. In other words, terms collected in C_k are typically small compared with the squared-mean terms (collected in $\mathcal{B}_k$). This separation becomes more pronounced with extra noncoherent terms $k' \neq k$, especially when K is large. Hence, $C_k(\mathbf{U},\boldsymbol{\nu}) \ll \mathcal{B}_k(\mathbf{U},\boldsymbol{\nu})$ is generally satisfied. Numerical results presented later verify this claim. ∎

Since $\Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\}$ is linear in v_{kl}, we reformulate (19) into an SOC form:

$$\Re\{\mathcal{A}_k(\mathbf{U},\boldsymbol{\nu})\} \geq \sqrt{\gamma_t}\|\mathbf{b}_k(\mathbf{U},\boldsymbol{\nu})\|, \quad \forall k\in\mathbb{K}, \tag{20}$$

where $\mathbf{b}_k(\mathbf{U},\boldsymbol{\nu})$ is a vector that stacks the square roots of all $\mathcal{B}_k(\mathbf{U},\boldsymbol{\nu})$'s terms, namely

$$\mathbf{b}_k(\mathbf{U},\boldsymbol{\nu}) = \tag{21}$$
$$\left[\frac{\sigma_z}{\sqrt{p_a}};\ \left\{v_{k'l}|\alpha_l|\sqrt{u_{k'l}\mathbb{E}[|\mathbf{h}_{kl}^T\mathbf{w}_{k'l}|^2]};\ v_{k'l}\sqrt{u_{k'l}\beta_l\mathrm{tr}(\mathbf{R}_{kl})}\right\}_{k'\in\mathbb{K},\,l\in\mathbb{L}}\right].$$

Putting all these elements together, (13) is transformed into a tractable optimization problem:

$$\max_{\mathbf{U}, \, \boldsymbol{\nu}, \, \boldsymbol{\eta}, \, \gamma_t} \gamma_t \tag{22}$$

$$\text{s.t.} \begin{cases} \Re\{\mathcal{A}_k(\mathbf{U}, \boldsymbol{\nu})\} \geq \sqrt{\gamma_t}\|\mathbf{b}_k(\mathbf{U}, \boldsymbol{\nu})\|, & \forall k \in \mathbb{K}, \\ \|[\, 2\nu_{kl}; \, 1 - \eta_{kl}]\| \leq 1 + \eta_{kl}, & \forall k \in \mathbb{K}, l \in \mathbb{L}, \\ \sum_{k \in \mathbb{K}} u_{kl}\nu_{kl}^2 \leq \dfrac{1}{|\alpha_l|^2 + \beta_l N_a}, & \forall l \in \mathbb{L}, \\ \nu_{kl} \geq 0, \quad \eta_{kl} \geq 0, \quad u_{kl} \in \{0,1\}, & \forall k \in \mathbb{K}, \, l \in \mathbb{L}. \end{cases}$$

This problem can be solved by a bisection method on γ_t, as depicted in Algorithm 1.

Algorithm 1: Joint Optimization Approach

Initialization:

$t \leftarrow 0, \quad \gamma_{\text{low}}^{(0)} \leftarrow 0, \quad \gamma_{\text{high}}^{(0)} \leftarrow \max_{k \in \mathbb{K}} \left(\dfrac{p_a |\sum_{l \in \mathbb{L}} \alpha_l \mathbb{E}[\mathbf{h}_{kl}^T \mathbf{w}_{kl}]|^2}{\sigma_z^2} \right)$

while $\gamma_{high}^{(t)} - \gamma_{low}^{(t)} > \epsilon$ **do**

  $\gamma_t \leftarrow \frac{1}{2}(\gamma_{\text{low}}^{(t)} + \gamma_{\text{high}}^{(t)})$

  Convex Feasibility Check:

    Find $\{\mathbf{U}, \boldsymbol{\nu}\}$

$$\text{s.t.} \begin{cases} \Re\{\mathcal{A}_k(\mathbf{U}, \boldsymbol{\nu})\} \geq \sqrt{\gamma_t}\|\mathbf{b}_k(\mathbf{U}, \boldsymbol{\nu})\|, & \forall k \in \mathbb{K}, \\ \|[\, 2\nu_{kl}; \, 1 - \eta_{kl}]\| \leq 1 + \eta_{kl}, & \forall k \in \mathbb{K}, l \in \mathbb{L}, \\ \sum_{k \in \mathbb{K}} u_{kl}\nu_{kl}^2 \leq \dfrac{1}{|\alpha_l|^2 + \beta_l N_a}, & \forall l \in \mathbb{L}, \\ \nu_{kl} \geq 0, \quad u_{kl} \in \{0,1\}, & \forall k \in \mathbb{K}, \, l \in \mathbb{L}. \end{cases}$$

  if *feasible* **then**

    $\gamma_{\text{low}}^{(t+1)} \leftarrow \gamma_t, \quad \gamma_{\text{high}}^{(t+1)} \leftarrow \gamma_{\text{high}}^{(t)}, \quad \mathbf{U}^* \leftarrow \mathbf{U}, \boldsymbol{\nu}^* \leftarrow \boldsymbol{\nu}$

  end

  else

    $\gamma_{\text{low}}^{(t+1)} \leftarrow \gamma_{\text{low}}^{(t)}, \quad \gamma_{\text{high}}^{(t+1)} \leftarrow \gamma_t$

  end

  $t \leftarrow t + 1$

end

return $\{\mathbf{U}^*, \boldsymbol{\nu}^*\}$

V. COMPUTATIONAL COST AND LOW-COMPLEXITY ALTERNATIVE

The joint user association and power optimization (JUP) constitutes a mixed-integer SOC program. It involves $n = L \times K$ binary variables u_{kl}, indicating user-AP associations, and up to n continuous power variables η_{kl}, conditioned on the associated $u_{kl} = 1$. In the worst case, all 2^n combinations of $\mathbf{U}$ must be explored, resulting in an exponential complexity of $O(2^n)$. Such exponential scaling may limit the applicability of this approach to large-scale networks. To address this challenge, we propose a two-stage decoupled strategy:

1) *Fixed User Association:* Each user is associated with its geographically closest AP(s), thereby eliminating the binary variables u_{kl} from the optimization problem.

2) *Max–Min Power Control:* With $\mathbf{U}$ fixed, the problem reduces to optimizing $\{\eta_{kl}\}$ to maximize the minimum achievable SE. Power is allocated only to active associations (i.e., where $u_{kl} = 1$), so the effective number of power variables is smaller than $L \times K$.

TABLE I
KEY SIMULATION PARAMETERS

Parameter	Value
Coverage Radius	1000 m (3GPP Microcell)
Path Loss Model	$-30.5 - 36.7 \log_{10}(d)$ (dB)
Shadow Fading	$\mathcal{N}(0, 4^2)$ dB, Gaussian normal
Number of Active Users	$K = 4$
AP Configuration	32 APs, each with 2 antennas
Power Constraints	$p_u = 200$ mW and $p_a = 100$ mW
Noise Spectral Density	-174 dBm/Hz
Noise Figure	9 dB
System Bandwidth	5 MHz
Antenna Array Type	ULA with half-wavelength spacing
Spatial Correlation Model	Gaussian scattering
Angular Spread (Std. Dev.)	$30°$
Coherence Interval	$\tau_c = 200$ channel uses
Pilot Contamination	$\tau_p = 2$ (2 users per pilot)
PA Gain Coefficients	$0.8 + 0.1i \cdot \mathcal{N}(0, 1)$, standard normal
PA Nonlinearity Factors	$0.05 + 0.1 \cdot \mathcal{U}(0, 1)$, uniform distribution

For reference and upper-bound analysis, we compare against the conventional max-min power control [7], which optimizes all $L \times K$ power variables. This problem is a standard SOC program with polynomial-time solvability using interior-point methods, corresponding to the complexity on the order of $O(n^3)$. By decoupling user association and power control, this low-complexity alternative (named JUP-Lo) enables efficient optimization for larger network deployments.

TABLE II
COMPLEXITY COMPARISON OF DIFFERENT OPTIMIZATION SCHEMES

Algorithm	JUP	JUP-Lo	Max-Min Optimization
Complexity	$O\left(2^{L \times K}\right)$	$O\left((L \times K)^3\right)$	$O\left((L \times K)^3\right)$

VI. SIMULATION RESULTS AND DISCUSSIONS

We compare the proposed joint approach with four benchmark schemes: 1) CF configuration with equal power allocation (EPA); 2) CF with max-min power optimization; 3) User-centric (UC) method where each AP serves its nearest user; and 4) Ideal case without PA nonlinearity (i.e., $\alpha_l = 1$ and $\beta_l = 0$). Key simulation parameters are listed in Table I. The performance under different precoding strategies—MR, RZF, and MMSE—is illustrated in Figs. 1a, 1b, and 1c, respectively. Each figure shows the cumulative distribution function (CDF) of SE across users. Particular attention is given to the 95%-likely SE (i.e., the 5th percentile of each CDF), which serves as a key indicator of user fairness and cell-edge performance. The mixed-integer SOC problem in (22) is implemented using CVX [9] in conjunction with the Gurobi optimizer [10]. Prior to optimization, the expectations in (22) are precomputed by averaging over 100 channel realizations.

Key observations from the numerical results are as follows:

- *PA Nonlinearity Impact:* The results show that the 95%-likely SE under nonlinear PA is very poor, i.e., 0.0024, 0.0025, and 0.0020 bps/Hz for MR, RZF, and MMSE, respectively. In contrast, in the ideal case, the SEs reach 0.5635, 0.6542, and 0.6767 bps/Hz, respectively. This
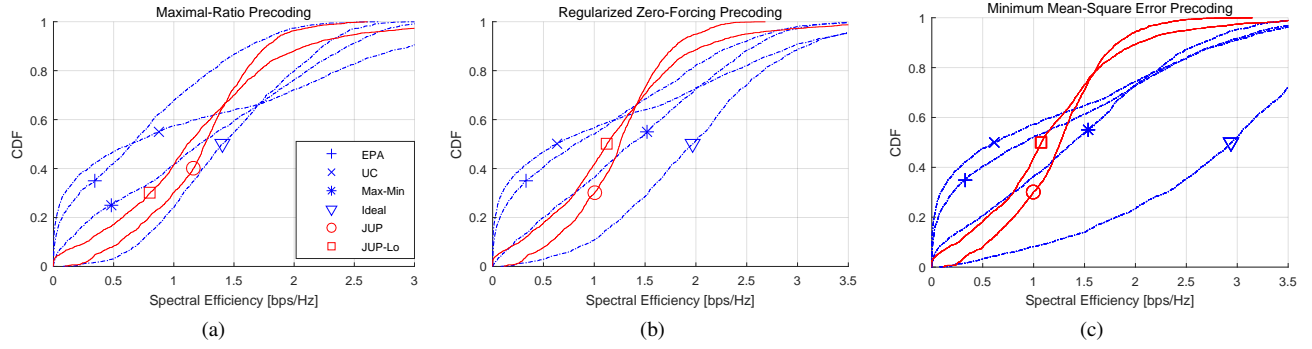
Fig. 1. The CDF of achievable spectral efficiency under MR, RZF, and MMSE precoding. User fairness (worst-case performance) is represented by the 5th percentile of each curve. All curves use the same marker styles as in Fig. 1a.

wide performance gap and near-zero SE values underscore the significant adverse impact of PA nonlinearity and highlight the need for PA-resilient design.

- *Baseline Limitations:* The applied UC approach yields inferior 95%-likely SE (MR: 0.0004, RZF: 0.0003, MMSE: 0.0003 bps/Hz), evident in the CDFs' early plateaus. While reducing fronthaul overhead, its exclusion of non-nearest users creates fairness gaps. The conventional max-min power optimization performs better, achieving 0.0289, 0.0956, and 0.0268 bps/Hz under MR, RZF, and MMSE precoding, respectively. However, using only optimal power optimization remains inadequate—its curves show limited rightward extension compared to ideal cases, indicating its sensitivity to PA nonlinearity.

- *Joint Optimization:* During the simulation, we observe that $\mathcal{A}_k$ is real for all three precoding schemes and that in most cases $\mathcal{C}_k/\mathcal{B}_k < 10\%$, thereby supporting Lemma 1. Our proposed approach significantly enhances performance (MR: 0.3514, RZF: 0.3602, MMSE: 0.3586 bps/Hz), visualized through *Right-shifted CDF curves* approaching that of ideal PA. It remarkably outperforms the max-min power optimization, achieving gains of approximately 1115 %, 276 %, and 1238 % in 95%-likely SE for MR, RZF, and MMSE precoding, respectively.

It is not feasible to evaluate complexity numerically as the number of optimization iterations is not fixed. Therefore, we use the average runtime as a metric to evaluate complexity. On a computer equipped with an Intel i7-4790 processor and 32GB of memory, the average time costs are approximately 17.07, 7.6, and 7.1 seconds for JUP, max-min, and JUP-Lo, respectively, as shown in Fig.2. While maintaining lower complexity, the simplified approach still outperforms conventional baselines. It achieves consistent 95%-likely performance across all precoding schemes (MR: 0.0531, RZF: 0.0548, MMSE: 0.0567 bps/Hz). These results represent a 2.1× improvement over max-min power control (0.0268 bps/Hz) and a 178× enhancement compared to UC (0.0003 bps/Hz) under MMSE precoding.

VII. Conclusion

This letter has addressed the challenge of investigated PA nonlinearity in cell-free massive MIMO downlink systems.
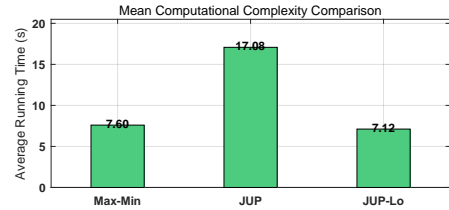

Fig. 2. Comparison of computational complexity.

We proposed a unified analytical framework that models PA-induced distortion across arbitrary linear precoding schemes. Then, a joint optimization approach of user association and power control was approximated to suppress alleviate the performance loss raised by PA nonlinearity. This joint approach, as well as its low-complexity variant, remarkably outperforms conventional hardware-agnostic baselines in terms of 95%-likely performance, providing a PA-resilient CF solution.

References

[1] Z. Mokhtari and R. Dinis, "Sum-rate of cell-free massive MIMO systems with power amplifier non-linearity," *IEEE Access*, vol. 9, p. 141927–141937, Oct. 2021.

[2] M. M. Jadidi *et al.*, "Performance analysis and power allocation for uplink cell-free massive MIMO system with nonlinear power amplifier," *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5473–5485, Sep. 2024.

[3] A. M. Khoueini *et al.*, "Performance analysis and power allocation for downlink cell-free massive MIMO system with nonlinear power amplifier," *IEEE Trans. Green Commun. Netw.*, Apr. 2025, Early Access.

[4] S. Buzzi *et al.*, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

[5] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77 – 90, Jan. 2020.

[6] ——, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[7] H. Q. Ngo *et al.*, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[8] O. T. Demir and E. Björnson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.

[9] M. Grant and S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming*, CVX Research, Inc., 2014.

[10] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2024. [Online]. Available: https://www.gurobi.com