IEEE Open Journal of the Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJCOMS.XXXX.XXXXXX

Joint RIS-RSMA Optimization for Asynchronous Multi-user Cell-Free MIMO

GHADIR DOGIM[1], AHMED EL-MAHDY[1], AND FALKO DRESSLER[2]

[1]German University in Cairo, Cairo, Egypt (e-mail: ghadir.mostafa@icloud.com; ahmed.elmahdy@guc.edu.eg)
[2]Technische Universität Berlin, Berlin, Germany (e-mail: dressler@ccs-labs.org)

CORRESPONDING AUTHOR: Falko Dressler (e-mail: dressler@ccs-labs.org).

ABSTRACT Cell-free multiple-input multiple-output (CF-MIMO) systems assisted by reconfigurable intelligent surfaces (RISs) and rate-splitting multiple access (RSMA) are promising for interference management in distributed sixth-generation (6G) networks. In practical CF-MIMO deployments, independent access-point (AP) oscillators and unequal propagation delays introduce AP-dependent phase variations that reduce coherent combining and degrade precoding performance. This paper develops a unified RIS-RSMA optimization framework for asynchronous downlink CF-MIMO.

The RSMA common precoder, private precoders, and passive RIS phase shifts are jointly optimized within a sum-rate maximization problem subject to total transmit-power, per-user quality-of-service, and RIS unit-modulus constraints. The resulting non-convex problem is addressed through a weighted minimum mean-square error (WMMSE)-based alternating optimization framework with Gauss–Seidel coordinate optimization for the RIS phase subproblem. An analytical common-stream power expression is derived to explain the limited benefit of decoupled common-precoder designs. The analysis shows that independent oscillator phase drifts at the APs attenuate coherent cross-AP combining terms and can restrict the common rate through the weakest-user decoding constraint.

In the default overloaded CF-MIMO configuration, the tested heuristic common-precoder schemes provide less than 0.2% gain over the corresponding optimized space-division multiple access (SDMA) baseline when combined with WMMSE-optimized private precoders. In contrast, the proposed joint RSMA design achieves approximately 12–15% sum-rate gain at moderate-to-high transmit powers and retains a similar high-power advantage in a larger-scale validation. Targeted comparisons using WMMSE and regularized zero-forcing (RZF) private precoders, together with the common-power allocation results, support the need for joint common/private precoder design. Additional results show that the gain persists across the tested RIS sizes, oscillator phase-noise variances, Rician factors, loading ratios, channel realizations, and multi-RIS deployments. Distributed multi-RIS deployment also improves spatial coverage uniformity under a fixed total element budget.

INDEX TERMS Alternating optimization, asynchronous reception, cell-free MIMO (CF-MIMO), common precoder design, passive beamforming, rate-splitting multiple access (RSMA), reconfigurable intelligent surface (RIS), weighted minimum mean-square error (WMMSE).

Notation: Upper and lower boldface represent matrices and vectors, respectively. $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^*$ denote transpose, conjugate transpose, and conjugate. $\|\cdot\|$ and $\|\cdot\|_F$ denote the Euclidean and Frobenius norms. $\mathrm{Tr}(\cdot)$ is the trace. $\mathrm{diag}(\boldsymbol{a})$ creates a diagonal matrix. $\mathcal{CN}(\mathbf{0},\mathbf{R})$ is circularly symmetric complex Gaussian. $\mathrm{Re}\{\cdot\}$ extracts the real part. $\mathbb{E}[\cdot]$ denotes expectation.

I. Introduction

CELL-FREE multiple-input multiple-output (CF-MIMO) has emerged as a transformative architecture for sixth-generation (6G) wireless networks, in which a large number of geographically distributed access points (APs) coherently and cooperatively serve all users on the same time-frequency resources without cell boundaries. By serving users through multiple geographically distributed

APs, CF-MIMO improves coverage uniformity, provides macro-diversity, and can outperform conventional small-cell architectures when appropriate centralized or cooperative processing is used [1], [2].

However, the practical deployment of CF-MIMO systems introduces a fundamental challenge that co-located massive MIMO does not face. Each AP is equipped with an independent local oscillator and is located at a different distance from each user, causing the signals arriving at user equipment (UEs) to experience heterogeneous propagation delays and independent oscillator phase drifts. These synchronization impairments degrade the phase coherence assumed by conventional precoding designs and increase residual multi-user interference, particularly when multiple users share limited spatial degrees of freedom [3]–[5].

Reconfigurable intelligent surfaces (RISs) have recently attracted significant attention as a cost-effective means of improving wireless propagation environments. Composed of a large number of passive reflecting elements with adjustable phase shifts, RISs can create virtual line-of-sight paths, enhance signal strength at target locations, and suppress interference without active radio-frequency chains [6], [7]. When deployed alongside CF-MIMO, RISs complement active AP beamforming by adding controllable reflected propagation paths, and prior RIS-aided cell-free studies have reported spectral- and energy-efficiency gains [8]–[10]. Nevertheless, joint optimization of active AP precoders and passive RIS phase shifts remains challenging due to the coupling between these variables and the unit-modulus constraint on the RIS elements.

Rate-splitting multiple access (RSMA) softly bridges the two extremes of fully decoding interference, as in non-orthogonal multiple access (NOMA), and treating interference entirely as noise, as in space-division multiple access (SDMA), by splitting messages into common and private parts [11]–[14]. This flexibility makes RSMA attractive under imperfect channel state information (CSI), overloaded transmission, and heterogeneous channel conditions, where it has been shown to outperform SDMA and NOMA in representative multi-antenna settings [13]–[16].

Despite the theoretical advantages of RSMA, a critical gap exists between theory and practice. Several existing RSMA implementations rely on *heuristic* common precoder constructions that decouple the design of the common beam from the private precoders. In these approaches, private precoders are first determined using conventional techniques like zero-forcing [17], and the common precoder is subsequently derived using simple rules such as dominant-eigenvector projection [13] or equal-gain combining [18]. A power-splitting parameter α is then searched over a grid to allocate power between the common and private streams. While such sequential approaches reduce computational complexity, their effectiveness depends critically on the array geometry. In co-located MIMO settings, heuristic common beams can exploit the shared array aperture and have been shown to provide useful RSMA gains under several channel and loading conditions [13], [14]. In CF-MIMO, however, the common precoder must coherently combine signals across geographically separated APs with fundamentally different spatial signatures to each user. To the best of our knowledge, the effectiveness of such decoupled common-precoder designs has not been systematically investigated in the considered overloaded asynchronous CF-MIMO setting. This paper addresses this question and shows, through analysis and simulation, that the tested heuristic common-precoder designs provide negligible rate-splitting gain in this setting when the private precoders are jointly optimized.

From the above survey, the existing literature does not jointly address three coupled aspects of the considered problem. First, RIS phase design and RSMA precoding are usually optimized without explicitly accounting for asynchronous transmission in distributed cell-free deployments. Second, existing cell-free RSMA designs often construct the common precoder separately from the private precoders, rather than jointly optimizing both sets of AP precoders. Third, the effect of decoupled common precoder design on the worst-user common rate has not been analyzed when the private precoders and RIS phase shifts are jointly optimized in an overloaded asynchronous cell-free deployment.

A. Contributions

This work develops a joint RIS-RSMA optimization framework for asynchronous distributed cell-free MIMO systems. The proposed design jointly optimizes the RSMA common stream, private streams, and RIS phase shifts. It also accounts for the phase variations introduced by asynchronous transmission. In distributed deployments, the RIS affects both common- and private-stream channels. The common-stream rate is limited by the weakest user. Therefore, decoupled common-precoder and RIS designs can be sensitive to phase misalignment across distributed APs.

The main contributions are summarized below.

- A joint RIS-RSMA formulation is developed for asynchronous downlink cell-free MIMO systems. The formulation accounts for passive RIS reflection, asynchronous phase variations, and instantaneous CSI at the network side. It jointly optimizes the common precoder, private precoders, and RIS phase shifts within one sum-rate maximization problem.
- A weighted minimum mean-square error (WMMSE)-based alternating optimization algorithm is proposed for the resulting non-convex problem. The AP precoders for the common and private streams are optimized jointly through the WMMSE reformulation. The RIS phase shifts are optimized under the unit-modulus constraint using Gauss–Seidel coordinate descent.
- An analytical common-stream power expression is derived to explain the limited benefit of decoupled common-precoder designs in the considered overloaded

asynchronous CF-MIMO setting. The analysis shows that AP-dependent phase variations attenuate coherent cross-AP combining and can restrict the common rate through the weakest-user decoding constraint. Targeted comparisons using WMMSE and RZF private precoders, together with the common-power allocation results, support the need for joint common/private precoder design.

This paper addresses the critical gap in the existing cell-free RSMA literature. Prior works have demonstrated RSMA gains using heuristic common-precoder designs in favourable underloaded, co-located, or simple-precoding settings. In contrast, this work shows that such decoupled designs can lose most of their RSMA benefit in asynchronous distributed cell-free architectures when the private precoders are already optimized. The proposed joint design restores these gains by optimizing the common precoder, private precoders, and RIS phase shifts within the same framework.

B. Paper Organization

Section III presents the system and channel models, including the RSMA transmission model and the synchronization impairments. Section IV formulates the sum-rate maximization problem and develops the proposed joint WMMSE–AO optimization framework. Section V describes the baseline schemes and examines the limited gain of the tested heuristic common-precoder designs in the considered overloaded asynchronous CF-MIMO setting. Section VI presents the performance evaluation. Section VII concludes the paper. Detailed derivations appear in the Appendices.

II. Related Work

A. Cell-Free MIMO with RIS

The integration of RIS into cell-free MIMO has been explored from several perspectives. Pan *et al.* [9] studied multicell MIMO with RIS using joint active and passive precoding to maximize the weighted sum rate; however, their framework is limited to multi-cell architectures and does not extend to the fully cooperative cell-free paradigm. Chien *et al.* [10] investigated RIS-assisted cell-free massive MIMO over spatially correlated channels, deriving achievable rate expressions and demonstrating coverage improvements. However, their analysis relies on maximum-ratio and local minimum mean-square error (MMSE) processing rather than centralized joint precoder optimization, and does not consider synchronization impairments or rate-splitting transmission.

More recently, Zhu *et al.* [19] developed a multi-agent reinforcement learning framework for joint precoding and phase shift optimization in RIS-aided CF-MIMO; while effective for distributed implementation, such data-driven approaches suffer from high training latency, lack of interpretability, and do not address the interference management limitations that RSMA is designed to overcome. Notably, none of these works considers rate-splitting transmission or the impact of asynchronous reception on joint RIS and precoder design.

B. RSMA in Multi-Antenna Systems

The theoretical foundations of RSMA for multi-antenna downlink systems were established through a series of influential works. Mao *et al.* [13] developed the WMMSE-based precoder design framework for RSMA in multiple-input single-output (MISO) broadcast channels, showing that RSMA bridges and generalizes both SDMA and NOMA; however, the analysis is restricted to co-located single-cell scenarios with perfect CSI. Joudeh and Clerckx [15] proposed a sum-rate maximization framework with partial channel state information at the transmitter (CSIT), demonstrating the robustness of RSMA to CSI imperfections, but did not consider distributed antenna architectures or passive beamforming through RIS.

For overloaded multi-antenna systems, Dizdar *et al.* [14] developed low-complexity RSMA designs that maintain performance advantages when $K > BM$, yet their heuristic common precoder constructions have not been validated in the cell-free setting where APs are geographically separated. In the context of RIS-aided systems, Bansal *et al.* [20] investigated RSMA for RIS-aided multi-user MISO but assumed a single co-located transmitter, and Soleymani *et al.* [21] studied RSMA optimization in beyond-diagonal RIS-assisted ultra-reliable low-latency communication (URLLC) systems, focusing on finite blocklength effects rather than distributed precoder design.

Soleymani *et al.* [22] further extended RIS-aided RSMA to full MIMO broadcast channels under URLLC constraints, confirming the synergy between RIS and RSMA under high user loads; however, this work also assumes a co-located base station (BS) and does not address the unique challenges of cell-free architectures. While these works comprehensively address RSMA in co-located settings, the extension to distributed cell-free architectures where the common precoder must span geographically separated APs rather than a single co-located array introduces fundamentally different design challenges that remain unaddressed.

C. Cell-Free RSMA

The application of rate-splitting to cell-free massive MIMO remains relatively emerging, and the studies reviewed here predominantly rely on heuristic common precoder designs. Mishra *et al.* [18] investigated an RSMA-assisted framework for massive machine-type communications in cell-free massive MIMO utilizing max-min power control. While their work demonstrated gains under pilot contamination using conjugate beamforming for private streams, their common precoder was constructed heuristically rather than jointly optimized. Furthermore, their system operates in the large-antenna regime ($B \gg K$) where channel hardening provides favorable conditions for simple precoding, leaving the effectiveness of such heuristic designs in the overloaded regime ($K > BM$) entirely unexamined.

Zheng *et al.* [23] provided the most relevant prior work by studying asynchronous cell-free massive MIMO with rate-splitting, deriving closed-form spectral efficiency expressions under oscillator phase-noise. Nevertheless, their framework employs conjugate beamforming for private precoders and a heuristic equal-ratio-sum common precoder, and does not include RIS or joint precoder optimization. Zheng *et al.* [24] derived closed-form spectral efficiency expressions for rate-splitting in cell-free massive MIMO with spatially correlated Rician fading and proposed a generative artificial intelligence (AI) approach for resource optimization, but again relied on predefined common precoder structures rather than joint design.

Zhang *et al.* [25] extended RSMA to CF-MIMO-URLLC systems with short-packet transmission under imperfect CSI; however, their work focused on power control rather than joint beamformer optimization. Most recently, Zhang *et al.* [26] investigated RSMA for uplink cell-free massive MIMO with low-resolution analog-to-digital converters (ADCs), demonstrating rate and energy efficiency gains under hardware constraints; however, the work considers the uplink direction and does not address joint downlink precoder-RIS optimization or synchronization impairments.

A common thread across all these cell-free RSMA works is their reliance on heuristic common precoder designs typically matched-filter or conjugate beamforming based directions combined with sequential power allocation. To the best of our knowledge, prior work has not investigated whether these heuristic approaches remain effective when private precoders are jointly optimized via WMMSE in the considered overloaded asynchronous CF-MIMO setting. This is precisely the gap addressed in this paper, and as shown in the results, the heuristic approaches provide negligible gain.

D. Asynchronous Cell-Free MIMO

The impact of hardware impairments and asynchronous reception on cell-free massive MIMO has been studied from multiple angles, but without considering joint RSMA–RIS optimization. Zheng et al. [3] analyzed cell-free massive MIMO systems with asynchronous reception, capturing both propagation delay phases and oscillator phase drifts via Wiener processes, and derived closed-form downlink spectral efficiency expressions; however, their analysis predates the integration of RSMA and RIS into cell-free systems. Fang et al. [4] analyzed cell-free massive MIMO with oscillator phase-noise, deriving performance bounds and proposing power control strategies, but did not consider rate-splitting or RIS-aided transmission. Li et al. [5] studied the impacts of asynchronous reception on cell-free distributed massive MIMO, quantifying the performance loss due to propagation delay offsets; however, no interference management technique beyond conventional precoding was considered. As noted above, Zheng et al. [23] provided the most comprehensive treatment of asynchronous cell-free MIMO with rate-splitting, but relied on conjugate beamforming and heuristic common precoding without RIS integration or joint precoder optimization.

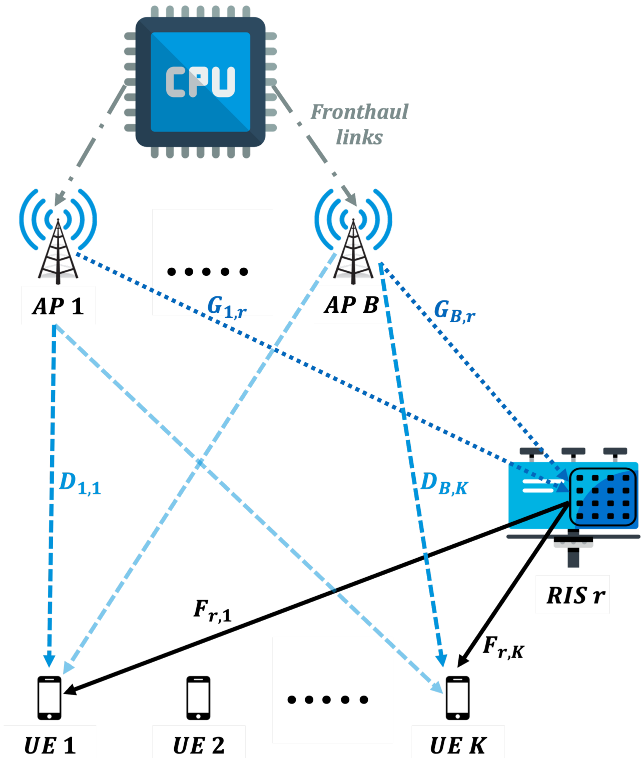


FIGURE 1. System model of the asynchronous RIS-aided CF-MIMO downlink. B distributed APs, each with M antennas and an independent local oscillator, cooperatively serve K single-antenna UEs via a CPU through fronthaul links. A passive RIS with N reflecting elements assists the transmission. Three channel types are shown: $\mathbf{D}_{b,k}$ AP–UE direct links (dashed), $\mathbf{G}_{b,r}$ AP–RIS links (dotted), and $\mathbf{F}_{r,k}$ RIS–UE links (solid).

III. System Model

A. Network Architecture

We consider an asynchronous RIS-aided CF-MIMO downlink system as illustrated in Fig. 1. The network consists of B distributed APs, each equipped with M antennas, one passive RIS with N reflecting elements, and K single-antenna UEs. A central processing unit (CPU) coordinates all APs through fronthaul links, enabling centralized joint precoding and resource allocation. [1] Each AP operates with an independent local oscillator, introducing AP-specific phase drifts that vary across the network. The APs are geographically distributed across the coverage area, and UEs are randomly clustered around a reference distance d from the origin.

[1]In the Open RAN (O-RAN) architecture [27], the APs correspond to Open Radio Units (O-RUs) and the CPU maps to the Open Distributed Unit (O-DU) where centralized baseband processing is performed. The AP/CPU terminology is retained for consistency with the cell-free massive MIMO literature [1]–[3].

The downlink transmission follows a coherence block model. During each coherence block, the CPU computes the precoding vectors for each AP and the RIS phase shift matrix based on channel estimates obtained from the uplink pilot phase. These are held fixed during the data transmission phase, even as the oscillator phases drift according to the Wiener process model described in Section III-D. In this work, the CPU is assumed to have error-free instantaneous CSI at the beginning of each coherence block. The precoders and RIS phases are designed using this block-start CSI and are then held fixed during data transmission, while the actual channels evolve according to the deterministic propagation-delay phases and the stochastic Wiener oscillator-drift model in Section III-D. This assumption allows the reported performance degradation to be attributed to the asynchronous channel mismatch rather than to channel-estimation error. Robust design under imperfect CSI, pilot contamination, and estimation uncertainty is left for future work.

As shown in Fig. 1, three types of channel links are present: the direct AP–UE channels $\mathbf{D}_{b,k}$ (dashed lines), the AP–RIS channels $\mathbf{G}_{b,r}$ (dotted lines), and the RIS–UE channels $\mathbf{F}_{r,k}$ (solid lines). The RIS passively reflects incident signals from the APs toward the UEs, creating additional propagation paths that enhance the effective channel quality. Each RIS element is modeled by a unit-modulus reflection coefficient, so the RIS adjusts the phase of the incident signal without active signal amplification [6], [7]. The RIS phase-shift vector is jointly optimized with the RSMA common and private transmit precoders at the APs, as detailed in Section IV. For notational consistency, the RIS is indexed by r, APs by $b \in \{1, \ldots, B\}$, and UEs by $k \in \{1, \ldots, K\}$.

B. Channel Model

The composite channel between each AP and UE comprises a direct link and a reflected link through the RIS. We model each link separately below, selecting channel models that reflect the distinct propagation characteristics of each link type. All numerical values used in the channel models are summarized in Table 2 in Section VI.

1) AP-UE Channel (Rayleigh)

The direct channel from AP b to UE k is modeled as

$$\mathbf{D}_{b,k} = \sqrt{\beta_{b,k}^{\mathrm{BU}}}\, \tilde{\mathbf{D}}_{b,k} \in \mathbb{C}^{M\times 1}, \tag{1}$$

where $\beta_{b,k}^{\mathrm{BU}} = C_0\, d_{b,k}^{-\kappa_{\mathrm{BU}}}$ is the distance-dependent large-scale path loss, $d_{b,k}$ is the distance between AP b and UE k, C_0 is the path-loss intercept at a reference distance of 1 m, and κ_{BU} is the path-loss exponent. The small-scale fading vector $\tilde{\mathbf{D}}_{b,k}$ has independent and identically distributed (i.i.d.) entries drawn from $\mathcal{CN}(0, 1)$, yielding a Rayleigh fading model. This choice is appropriate for the AP-UE link because the distributed APs and randomly located UEs typically lack a dominant line-of-sight (LoS) component, particularly in indoor hotspot and dense urban deployments where rich scattering is prevalent.

2) AP-RIS Channel (LoS-Dominant)

The channel from AP b to the RIS is modeled as

$$\mathbf{G}_{b,r} = \sqrt{\beta_{b,r}^{\mathrm{BR}}}\, e^{j2\pi\mathbf{U}} \in \mathbb{C}^{N\times M}, \tag{2}$$

where $\beta_{b,r}^{\mathrm{BR}} = 2C_0\, d_{b,r}^{-\kappa_{\mathrm{BR}}}$ is the large-scale path loss with $d_{b,r}$ denoting the distance between AP b and the RIS, κ_{BR} is the path-loss exponent, and $\mathbf{U}$ is a deterministic phase matrix capturing the LoS array response. The multiplicative factor of 2 reflects the stronger AP–RIS link assumed for an elevated or deliberately placed RIS deployment. A LoS-dominant AP–RIS model is therefore adopted, as commonly done in RIS-aided MIMO studies [6], [7]. The spatial asymmetry introduced by a single RIS deployment is examined later in Section VI-K.

3) RIS-UE Channel (Rayleigh / Rician)

The channel from the RIS to UE k is modeled as

$$\mathbf{F}_{r,k} = \sqrt{\beta_{r,k}^{\mathrm{RU}}}\, \tilde{\mathbf{F}}_{r,k} \in \mathbb{C}^{N\times 1}, \tag{3}$$

where $\beta_{r,k}^{\mathrm{RU}} = 2C_0\, d_{r,k}^{-\kappa_{\mathrm{RU}}}$ is the large-scale path loss, $d_{r,k}$ is the distance between the RIS and UE k, and κ_{RU} is the path-loss exponent. By default, the small-scale fading $\tilde{\mathbf{F}}_{r,k}$ has i.i.d. entries drawn from $\mathcal{CN}(0, 1)$, yielding a Rayleigh model, which serves as a conservative baseline to evaluate RSMA performance under challenging fading conditions. In practice, the elevated RIS deployment often provides a LoS component to ground-level UEs. To capture this, the Rician fading sensitivity analysis in Section VI-I generalizes the RIS-UE channel to a Rician fading model as follows.

$$\mathbf{F}_{r,k} = \sqrt{\beta_{r,k}^{\mathrm{RU}}} \left(\sqrt{\frac{K_R}{K_R+1}}\, \mathbf{F}_{r,k}^{\mathrm{LoS}} + \sqrt{\frac{1}{K_R+1}}\, \tilde{\mathbf{F}}_{r,k} \right), \tag{4}$$

where K_R is the Rician K-factor controlling the ratio of LoS power to scattered power, and $\mathbf{F}_{r,k}^{\mathrm{LoS}}$ is the deterministic LoS component based on the array steering vector. Setting $K_R = 0$ recovers the default Rayleigh model.

C. Composite Synchronous Channel

Before introducing the synchronization impairments, we first define the composite synchronous channel. In the absence of any synchronization or local oscillator mismatch, the effective channel from AP b to UE k through both the direct and RIS-reflected paths is given by

$$\mathbf{h}_{b,k}^{\mathrm{sync}} = \underbrace{\mathbf{D}_{b,k}}_{\text{Direct path}} + \underbrace{\mathbf{G}_{b,r}^{H}\boldsymbol{\Phi}\,\mathbf{F}_{r,k}}_{\text{Reflected path}}, \tag{5}$$

where $\mathbf{D}_{b,k} \in \mathbb{C}^{M\times 1}$ is the direct AP-UE channel defined in (1), $\mathbf{G}_{b,r} \in \mathbb{C}^{N\times M}$ is the AP-RIS channel defined in (2), $\mathbf{F}_{r,k} \in \mathbb{C}^{N\times 1}$ is the RIS-UE channel defined in (3), and $\boldsymbol{\Phi} = \mathrm{diag}(e^{j\phi_1}, \ldots, e^{j\phi_N})$ is the diagonal RIS phase shift matrix, where $\phi_n \in [0, 2\pi)$ is the phase shift applied by the n-th reflecting element and $|e^{j\phi_n}| = 1$ for all n. The first term represents the direct propagation path, while the second term represents the cascaded AP$\rightarrow$RIS$\rightarrow$UE reflected path, where the RIS coherently combines N reflected signal components. This synchronous composite channel is used to optimize the precoders and RIS phase shifts at the beginning of each coherence block. The resulting design is then held fixed while its performance is evaluated over the asynchronous channel evolution described in Section D.

D. Synchronization Impairments

In practical cell-free deployments, the signals received at each UE from different APs do not arrive phase-coherently. The asynchronous reception model contains two components: deterministic propagation-delay phase shifts due to unequal link distances and stochastic oscillator phase drift modeled as a discrete-time Wiener process [3]–[5], [23]. Unlike co-located massive MIMO, where the antennas usually share a common oscillator and an approximately common propagation geometry to each UE, CF-MIMO APs have independent oscillators and AP-dependent propagation delays. These impairments are therefore *AP-specific* and *time-varying*, making them particularly damaging to precoder performance.

1) Propagation Delay Phase

Due to the different physical positions of the distributed APs, the distance $d_{b,k}$ between AP b and UE k varies across all AP–UE pairs. For each UE k, the closest AP,

$$b_k^{\mathrm{ref}} = \arg \min_{b \in \{1,\ldots,B\}} d_{b,k}, \tag{6}$$

is selected as the propagation-delay reference. The relative distance between AP b and the reference AP for UE k is defined as

$$\Delta d_{b,k} = d_{b,k} - d_{b_k^{\mathrm{ref}},k}. \tag{7}$$

This distance variation introduces a deterministic phase offset on each direct AP–UE link, given by

$$\Theta_{b,k} = \exp\left(-j\,\frac{2\pi\,\Delta d_{b,k}}{c\,T_s}\right), \tag{8}$$

where c is the speed of light and T_s is the symbol duration. The ratio $\Delta d_{b,k}/(c\,T_s)$ represents the propagation delay relative to the closest AP, normalized by the symbol period. Since the AP positions are fixed and the UE locations change slowly relative to the coherence block duration, $\Theta_{b,k}$ is effectively constant within each block but differs across AP–UE pairs. By construction, $\Delta d_{b_k^{\mathrm{ref}},k} = 0$ and therefore $\Theta_{b_k^{\mathrm{ref}},k} = 1$. All other direct-link delay phases are measured relative to this reference [23].

2) Oscillator Phase-Noise (Wiener Process)

Each AP b is equipped with an independent free-running local oscillator that generates the carrier signal for up-conversion. Due to manufacturing imperfections and thermal noise in the oscillator circuitry, the carrier phase drifts randomly over time. Following the standard discrete-time Wiener process model widely adopted in the literature [3], [4], [23], the oscillator phase of AP b at the p-th time instant within a coherence block is given by

$$\psi_b^{\mathrm{AP}}[p] = \psi_b^{\mathrm{AP}}[p-1] + \Delta\psi_b[p], \quad \Delta\psi_b[p] \sim \mathcal{N}(0, \sigma_\phi^2), \tag{9}$$

where $\psi_b^{\mathrm{AP}}[p]$ is the cumulative oscillator phase of AP b at time instant p, $\Delta\psi_b[p]$ is the random phase increment between consecutive time instants, and

$$\sigma_\phi^2 = 4\pi^2 f_c^2 c_{\mathrm{osc}} T_s$$

is the per-step phase-noise variance. Here, f_c is the carrier frequency and c_{osc} is the oscillator quality coefficient, where smaller values indicate greater oscillator stability. The Wiener model captures the accumulation of oscillator phase errors over time.

Similarly, each UE k has an oscillator phase $\psi_k^{\mathrm{UE}}[p]$ with independent increments of variance σ_ϕ^2. The oscillator processes are independent across APs and UEs, and their initial phases are set to zero at $p = 1$.

For compactness, define the AP oscillator-phase vector at time instant p as

$$\boldsymbol{\psi}^{\mathrm{AP}}[p] = [\psi_1^{\mathrm{AP}}[p], \ldots, \psi_B^{\mathrm{AP}}[p]]^T.$$

Under the zero initial-phase convention, each $\psi_b^{\mathrm{AP}}[p]$ is the sum of the $p-1$ independent Wiener-process increments $\Delta\psi_b[\tau]$, $\tau = 2, \ldots, p$. Hence,

$$\boldsymbol{\psi}^{\mathrm{AP}}[p] \sim \mathcal{N}\big(\mathbf{0}, (p-1)\sigma_\phi^2 \mathbf{I}_B\big)\,.$$

3) Composite Asynchronous Channel

Combining the propagation delay phase, the oscillator phase drifts, and the synchronous channel, the effective asynchronous channel from AP b to UE k at time instant p is

$$\mathbf{h}_{b,k}^H[p] = \underbrace{\Theta_{b,k}\, e^{j(\psi_b^{\mathrm{AP}}[p] + \psi_k^{\mathrm{UE}}[p])}\, \mathbf{D}_{b,k}^H}_{\text{Direct path}} + \underbrace{\mathbf{F}_{r,k}^H[p]\, \boldsymbol{\Phi}^H\, \mathbf{G}_{b,r}[p]}_{\text{Reflected path}}\,. \tag{10}$$

Here, $\mathbf{G}_{b,r}[p] \in \mathbb{C}^{N \times M}$ and $\mathbf{F}_{r,k}[p] \in \mathbb{C}^{N \times 1}$ denote the time-varying AP–RIS and RIS–UE channels, respectively. They are obtained by applying the corresponding link-specific propagation-delay and oscillator phase factors to the channels defined in (2) and (3). Thus, the direct path is affected through $\Theta_{b,k}$ and the AP and UE oscillator phases, while the reflected path inherits the corresponding phase variations through both constituent links.

The precoders and RIS phase shifts are designed using the synchronous channel $\mathbf{h}_{b,k}^{\mathrm{sync}}$ at the beginning of each coherence block and are then held fixed during data transmission. The realized channel follows (10) for $p = 1, \ldots, P$. As the oscillator phase errors accumulate, the mismatch between the design and realized channels generally increases, reducing the coherent precoding gain and increasing residual multi-user interference. This mechanism explains how synchronization impairments degrade cell-free MIMO performance and motivates the use of RSMA to manage the resulting interference through the common stream.

E. RSMA Transmission

In RSMA, the message intended for each UE k is split at the CPU into a common part and a private part. The common parts from all K users are jointly encoded into a single common stream s_0, which is designed to be decoded by every UE via successive interference cancellation (SIC) [28]. The remaining private portion of each message is independently encoded into a private stream s_k, $k = 1, \ldots, K$, and decoded only by the intended UE. Residual inter-user interference is treated as noise. All streams satisfy unit power, i.e., $\mathbb{E}[|s_0|^2] = \mathbb{E}[|s_k|^2] = 1$.

At each AP b, the precoding matrix is $\mathbf{W}_b = [\mathbf{w}_{b,0}, \mathbf{w}_{b,1}, \ldots, \mathbf{w}_{b,K}] \in \mathbb{C}^{M\times(K+1)}$, where $\mathbf{w}_{b,0} \in \mathbb{C}^{M\times 1}$ is the common precoder and $\mathbf{w}_{b,k} \in \mathbb{C}^{M\times 1}$ is the private precoder for UE k at AP b. The transmitted signal from AP b is

$$\mathbf{x}_b = \underbrace{\mathbf{w}_{b,0}\, s_0}_{\text{Common stream}} + \underbrace{\sum_{k=1}^{K} \mathbf{w}_{b,k}\, s_k}_{\text{Private streams}} . \tag{11}$$

The signal received at UE k from AP b at time instant p is

$$\begin{aligned} y_{b,k}[p] = &\underbrace{\mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,0}\, s_0}_{\text{Common stream from AP } b} + \underbrace{\mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,k}\, s_k}_{\text{Desired private stream from AP } b} \\ &+ \underbrace{\sum_{j\neq k} \mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,j}\, s_j}_{\text{Inter-user interference from AP } b} . \end{aligned} \tag{12}$$

Since all B APs transmit simultaneously on the same time-frequency resource, the total received signal at UE k is the superposition across all APs plus noise.

$$\begin{aligned} y_k[p] &= \sum_{b=1}^{B} y_{b,k}[p] + n_k \\ &= \underbrace{\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,0}\, s_0}_{\text{Desired common stream}} + \underbrace{\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,k}\, s_k}_{\text{Desired private stream}} \\ &\quad + \underbrace{\sum_{\substack{j=1\\ j\neq k}}^{K} \sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\, \mathbf{w}_{b,j}\, s_j}_{\text{Multi-user interference}} + \underbrace{n_k}_{\text{Noise}}, \end{aligned} \tag{13}$$

where $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise at UE k. The key distinction from co-located MIMO is that each AP b contributes through a different asynchronous channel $\mathbf{h}_{b,k}^{H}[p]$, with AP-specific delay phases and oscillator drifts. This results in the degradation of the coherent combining across APs as the oscillator phases drift over time. This combining is essential for both the common and private precoding gains. Each UE first decodes the common stream s_0 by treating all private streams as noise. The signal-to-interference-plus-noise ratio (SINR) for decoding the common stream at UE k is

$$\gamma_{c,k}[p] = \frac{\left|\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\mathbf{w}_{b,0}\right|^2}{\sum_{j=1}^{K} \left|\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\mathbf{w}_{b,j}\right|^2 + \sigma^2}, \tag{14}$$

where the numerator is the coherently combined common signal power across all B APs and the denominator contains the total private stream interference from all K users plus noise. Note that since the common stream is decoded first, all K private streams appear as interference. After successful decoding, UE k reconstructs and subtracts the common stream from $y_k[p]$ via SIC, assuming perfect cancellation. The UE then decodes its own private stream s_k, yielding the private SINR given as

$$\gamma_{p,k}[p] = \frac{\left|\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\mathbf{w}_{b,k}\right|^2}{\sum_{\substack{j=1\\ j\neq k}}^{K} \left|\sum_{b=1}^{B} \mathbf{h}_{b,k}^{H}[p]\mathbf{w}_{b,j}\right|^2 + \sigma^2}, \tag{15}$$

where the common stream no longer appears in the denominator (removed by SIC) and the interference is now from the $K - 1$ private streams of other users only. Since every UE must successfully decode s_0, the achievable common rate is limited by the weakest user.

Because the precoders are designed based on the synchronous design channel at $p = 1$ while the actual channel evolves according to (10) over $p = 1, \ldots, P$, the achievable rates are evaluated by averaging the instantaneous rate expressions over the P asynchronous time instants. For the common stream, a conservative per-instant weakest-user metric is adopted, where at each time instant the common rate is limited by the UE with the smallest instantaneous common-stream SINR. The common and private rates are therefore

$$R_c = \frac{1}{P} \sum_{p=1}^{P} \min_k \log_2\left(1 + \gamma_{c,k}[p]\right), \tag{16}$$

$$R_{p,k} = \frac{1}{P} \sum_{p=1}^{P} \log_2\left(1 + \gamma_{p,k}[p]\right), \tag{17}$$

and the system sum rate is

$$R_{\text{sum}} = R_c + \sum_{k=1}^{K} R_{p,k}. \tag{18}$$

In the synchronous reference case ($P = 1$), these expressions reduce to the instantaneous rates evaluated at the design channel.

IV. Proposed Joint RSMA–RIS Optimization

Our objective is to maximize the sum rate R_{sum} by jointly optimizing the common and private precoders $\{\mathbf{w}_{b,j}\}$ at all APs and the RIS phase shift matrix $\mathbf{\Phi}$, subject to practical system constraints.

A. Problem Formulation

The common stream is decoded by all UEs at the common rate R_c. The nonnegative variable c_k denotes the portion of this common rate allocated to UE k, with $\sum_{k=1}^{K} c_k \leq R_c$. The total rate allocated to UE k is therefore $c_k + R_{p,k}$.

The optimization problem is formulated as follows.

$$\mathcal{P}_1: \max_{\{\mathbf{w}_{b,j}\}, \mathbf{\Phi}, R_c, \{c_k\}} \quad R_c + \sum_{k=1}^{K} R_{p,k} \tag{19a}$$

$$\text{s.t.} \quad R_c \leq \log_2(1+\gamma_{c,k}), \quad \forall k, \tag{19b}$$

$$\sum_{k=1}^{K} c_k \leq R_c, \quad c_k \geq 0, \quad \forall k, \tag{19c}$$

$$\sum_{b=1}^{B}\sum_{j=0}^{K} \|\mathbf{w}_{b,j}\|^2 \leq P_{\max}, \tag{19d}$$

$$c_k + R_{p,k} \geq R_{\min}, \quad \forall k, \tag{19e}$$

$$|\theta_n| = 1, \quad \forall n. \tag{19f}$$

The objective (19a) maximizes the system sum rate, which includes the common-stream rate R_c and the individual private rates $R_{p,k}$. Constraint (19b) ensures that every UE can successfully decode the common stream, requiring that R_c does not exceed the decodable rate at the weakest user. Constraint (19c) ensures that the total common-rate allocation across all users does not exceed the achievable common rate. Constraint (19d) imposes a network-wide total transmit-power budget $P_{\max}$ jointly across all APs and all $K+1$ transmitted streams. Constraint (19e) guarantees a minimum quality-of-service (QoS) rate $R_{\min}$ for each UE, where the total rate allocated to UE k comprises its common-rate allocation c_k and its private rate $R_{p,k}$. Constraint (19f) enforces the unit-modulus property on each RIS reflecting element, reflecting the hardware limitation that passive RIS elements can only adjust the phase of the incident signal without amplification.

Problem $\mathcal{P}_1$ is non-convex due to the coupling between the precoders $\{\mathbf{w}_{b,j}\}$ and the RIS phase shifts $\mathbf{\Phi}$ in the SINR expressions, as well as the unit-modulus constraint on the RIS reflection coefficients (19f). To handle this, an alternating optimization (AO) approach is employed that decomposes $\mathcal{P}_1$ into two tractable subproblems: (i) a convex WMMSE precoder optimization for fixed RIS phase shifts, and (ii) a quadratically constrained quadratic program (QCQP) RIS phase optimization for fixed precoders. The two subproblems are solved alternately until convergence.

B. WMMSE Reformulation for Precoder Optimization

For the purpose of optimization, we stack the per-AP channels into aggregate per-UE channels and the per-AP precoders into aggregate precoders as follows.

$$\mathbf{h}_k = [\mathbf{h}_{1,k}^T, \mathbf{h}_{2,k}^T, \ldots, \mathbf{h}_{B,k}^T]^T \in \mathbb{C}^{BM\times 1}, \tag{20}$$

$$\mathbf{w}_j = [\mathbf{w}_{1,j}^T, \mathbf{w}_{2,j}^T, \ldots, \mathbf{w}_{B,j}^T]^T \in \mathbb{C}^{BM\times 1}. \tag{21}$$

The full precoder matrix stacking all $K+1$ streams is denoted $\mathbf{W} = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{BM\times(K+1)}$. With this notation, $\sum_{b=1}^{B} \mathbf{h}_{b,k}^H \mathbf{w}_{b,j} = \mathbf{h}_k^H \mathbf{w}_j$.

Direct maximization of the sum rate R_{sum} over the precoders is intractable because the rate expressions involve logarithmic terms with coupled precoder variables in both the numerator and denominator. The key insight of the WMMSE framework [13], [15], [29] is that rate maximization can be equivalently reformulated as a weighted mean-square error (MSE) minimization, which is convex in the precoders when the receive equalizers and WMMSE weights are fixed. The detailed derivation of this equivalence and its extension to RSMA is provided in Appendix A.

For fixed RIS phase shifts $\mathbf{\Phi}$, the aggregate effective channel $\mathbf{h}_k$ defined in (20) is known. Consider a scalar receive equalizer $g_{i,k} \in \mathbb{C}$ applied at UE k for stream $i \in \{c, p\}$. For a given UE k, let $\mathbf{w}_i = \mathbf{w}_0$ when $i = c$ and $\mathbf{w}_i = \mathbf{w}_k$ when $i = p$. The MSE between the transmitted and decoded symbols measures the quality of the combined precoder-equalizer design. The MMSE-optimal equalizer and the corresponding optimal weight that links MSE minimization to rate maximization are given as follows.

$$g_{i,k}^{\star} = \frac{\mathbf{h}_k^H \mathbf{w}_i}{|\mathbf{h}_k^H \mathbf{w}_i|^2 + I_{i,k} + \sigma^2}, \tag{22}$$

$$\alpha_{i,k}^{\star} = (\varepsilon_{i,k}^{\star})^{-1} = 1 + \gamma_{i,k}, \tag{23}$$

where $g_{i,k}^{\star}$ is the MMSE receiver that minimizes the estimation error, $\varepsilon_{i,k}^{\star}$ is the resulting minimum MSE, and $\alpha_{i,k}^{\star}$ is the optimal WMMSE weight. The interference terms are defined as

$$I_{c,k} = \sum_{j=1}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2, \quad I_{p,k} = \sum_{j\neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2, \tag{24}$$

corresponding to the common-stream interference, where all K private streams contribute before SIC, and the private-stream interference, where only $K-1$ other private streams remain after SIC. Using the optimal equalizers $g_{i,k}^{\star}$ from (22) and the optimal weights $\alpha_{i,k}^{\star}$ from (23), we define the composite variable $\beta_{i,k} \triangleq \sqrt{\alpha_{i,k}}\, g_{i,k}^{*}$ and construct the WMMSE surrogate functions that serve as concave lower bounds on the individual rate contributions.

$$f_{p,k}(\mathbf{W}) = 2\sqrt{\alpha_{p,k}}\, \mathrm{Re}\{\beta_{p,k}^{*} \mathbf{h}_k^H \mathbf{w}_k\} - |\beta_{p,k}|^2 \left(\sum_{j=1}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \right), \tag{25}$$

$$f_{c,k}(\mathbf{W}) = 2\sqrt{\alpha_{c,k}}\, \mathrm{Re}\{\beta_{c,k}^{*} \mathbf{h}_k^H \mathbf{w}_0\} - |\beta_{c,k}|^2 \left(\sum_{j=0}^{K} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \right), \tag{26}$$

where $f_{p,k}(\mathbf{W})$ and $f_{c,k}(\mathbf{W})$ are the surrogates for the private rate $R_{p,k}$ and the common rate contribution at UE k, respectively.

Both surrogates are *concave quadratic* functions of $\mathbf{W}$, which makes the resulting precoder subproblem (27) a convex optimization problem solvable to global optimality via interior-point methods [30].

Substituting the surrogates (25)-(26) into $\mathcal{P}_1$ in (19) yields a convex subproblem in $\mathbf{W}$ via two reformulations. First, each private rate $R_{p,k}$ in (17) is replaced by its surrogate $f_{p,k}(\mathbf{W})$ from (25). Second, the common rate R_c in (16) is represented by the auxiliary variable $c_{\min}$, while c_k denotes the portion of this common rate allocated to UE k. The $\min_k$ operator is then handled through per-user constraints $c_{\min} \leq \log_2(1+\gamma_{c,k})$ for all k, which yield constraint (27c). The variables $\{c_k\}$ appear in the per-user QoS constraint (27d), contributing each user's portion of the common rate toward its total rate requirement. The detailed derivation is provided in Appendix A. The resulting subproblem is

$$\max_{\mathbf{W},c_{\min},\{c_k\}} \sum_{k=1}^{K} f_{p,k}(\mathbf{W}) + \ln 2 \cdot c_{\min} \tag{27a}$$

$$\text{s.t.} \;\; \|\mathbf{W}\|_F^2 \leq P_{\max} \tag{27b}$$

$$c_{\min} \leq r_{c,k}^{(0)} + \frac{f_{c,k}(\mathbf{W})}{\ln 2}, \; \forall k \tag{27c}$$

$$f_{p,k}(\mathbf{W}) + \ln 2 \cdot c_k \geq \eta_k, \; \forall k \tag{27d}$$

$$\sum_{k=1}^{K} c_k \leq c_{\min}, \;\; c_{\min}, c_k \geq 0, \tag{27e}$$

where $r_{c,k}^{(0)} = \log_2 \alpha_{c,k} + (1-\alpha_{c,k})/\ln 2$ is the constant term in the common-stream WMMSE lower bound and $\eta_k = R_{\min} \ln 2 - \ln \alpha_{p,k} + \alpha_{p,k} - 1$ is the QoS threshold.

Problem (27) is convex; the objective is concave quadratic in $\mathbf{W}$ and linear in $c_{\min}$, and all constraints are convex. It is then solved via interior-point methods using CVX [31]. The critical advantage over the heuristic approaches in Section V-D is that the common precoder $\mathbf{w}_0$ and the private precoders $\{\mathbf{w}_k\}$ are optimized *jointly within a single convex program*, allowing the solver to find a beam structure where the common stream constructively complements the private streams rather than interfering with them.

C. RIS Phase Optimization via Gauss-Seidel Coordinate Descent

With the precoders $\mathbf{W}$ fixed from the WMMSE subproblem, the remaining optimization is over the RIS phase shifts. We define the phase shift vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]^T$ with $\theta_n = e^{j\phi_n}$, so that $\mathbf{\Phi} = \mathrm{diag}(\boldsymbol{\theta})$. To expose the dependence on $\boldsymbol{\theta}$, we decompose the effective channel of UE k as

$$\mathbf{h}_k^H = \underbrace{\mathbf{d}_k^H}_{\text{Direct path}} + \underbrace{\boldsymbol{\theta}^H \mathbf{V}_k}_{\text{Reflected path}}, \tag{28}$$

where $\mathbf{d}_k^H = [\mathbf{D}_{1,k}^H, \ldots, \mathbf{D}_{B,k}^H] \in \mathbb{C}^{1\times BM}$ is the aggregate direct channel, $\mathbf{G} = [\mathbf{G}_{1,r}, \ldots, \mathbf{G}_{B,r}] \in \mathbb{C}^{N\times BM}$ is the concatenated AP-RIS channel, and $\mathbf{V}_k = \mathrm{diag}(\mathbf{F}_{r,k}^*)\,\mathbf{G} \in \mathbb{C}^{N\times BM}$ is the cascaded RIS channel matrix. The direct path $\mathbf{d}_k^H$ does not depend on $\boldsymbol{\theta}$, while the reflected path depends linearly on $\boldsymbol{\theta}$ through $\mathbf{V}_k$. Substituting (28) into the WMMSE objective and collecting quadratic and linear terms in $\boldsymbol{\theta}$, the RIS optimization reduces to the following quadratically constrained quadratic program (QCQP).

$$\min_{\boldsymbol{\theta}} \; \boldsymbol{\theta}^H \mathbf{Q}\,\boldsymbol{\theta} - 2\,\mathrm{Re}\{\boldsymbol{\theta}^H \mathbf{q}\} \quad \text{s.t. } |\theta_n| = 1, \; \forall n, \tag{29}$$

where $\mathbf{Q} \in \mathbb{C}^{N\times N}$ is a positive semidefinite Hermitian matrix encoding the quadratic interaction between RIS elements through all user channels, and $\mathbf{q} \in \mathbb{C}^{N\times 1}$ captures the cross-interaction between the direct and reflected paths. The derivation of $\mathbf{Q}$ and $\mathbf{q}$ is detailed in Appendix B. The unit-modulus constraints $|\theta_n| = 1$ make (29) non-convex despite the quadratic objective.

Problem (29) is solved via Gauss–Seidel coordinate descent, which updates one RIS element at a time while keeping the others fixed. When $\{\theta_m\}_{m\neq n}$ are held constant, the objective becomes a function of θ_n alone. Expanding the quadratic and linear terms and using $|\theta_n|^2 = 1$, the diagonal contribution $Q_{n,n}$ is constant, and the objective reduces to

$$J(\theta_n) = -2\,\mathrm{Re}\{c_n^* \theta_n\} + \text{const.}, \tag{30}$$

where

$$c_n = q_n - \sum_{m\neq n} Q_{n,m}\,\theta_m. \tag{31}$$

Here, c_n aggregates the linear coefficient and the inter-element coupling from all other elements. Minimizing $-2\,\mathrm{Re}\{c_n^*\theta_n\}$ subject to $|\theta_n| = 1$ has the closed-form solution

$$\theta_n \leftarrow e^{j\angle(c_n)}. \tag{32}$$

Each update has $\mathcal{O}(N)$ complexity and does not increase the RIS-subproblem objective. Appendix C establishes the monotonicity and convergence of the corresponding objective-value sequence.

D. Overall Algorithm

The complete algorithm is summarized in Algorithm 1. At each iteration, the algorithm first updates the MMSE receivers and weights via (22)-(23), then solves the convex precoder subproblem (27), and finally updates the RIS phases via (32). The iterations continue until the relative change in the achieved sum rate falls below ϵ or the maximum number of iterations $T_{\max}$ is reached.

E. Convergence and Complexity

For fixed RIS phases, the common and private precoders are updated by solving the convex WMMSE subproblem. For fixed precoders, each Gauss–Seidel coordinate update does not increase the objective of the RIS subproblem. As established in Appendix C, this objective is bounded below, and the corresponding objective-value sequence therefore converges to a finite value. In the implementation, the overall AO iterations are terminated when the relative change in the achieved sum rate falls below ϵ or when the maximum number of iterations $T_{\max}$ is reached.

Algorithm 1 Joint RSMA–RIS Optimization via WMMSE–AO

Input: Channels $\{\mathbf{D}, \mathbf{G}, \mathbf{F}\}$, $P_{\max}$, $R_{\min}$
Initialize precoders $\mathbf{W}^{(0)}$ and RIS phases $\boldsymbol{\theta}^{(0)}$
for $t = 1, \ldots, T_{\max}$ **do**
    Update effective channels $\mathbf{h}_k$ using $\boldsymbol{\theta}^{(t-1)}$
    Compute $g_{i,k}^{\star}$, $\alpha_{i,k}^{\star}$ via (22)-(23)
    Solve (27) $\rightarrow \mathbf{W}^{(t)}$
    Compute $\mathbf{Q}$, $\mathbf{q}$; update $\boldsymbol{\theta}^{(t)}$ via (32)
    if $|R_{\text{sum}}^{(t)} - R_{\text{sum}}^{(t-1)}|/R_{\text{sum}}^{(t-1)} < \epsilon$ **then**
        break
    end if
end for
Output: $\mathbf{W}^{\star}$, $\boldsymbol{\theta}^{\star}$, $R_{\text{sum}}^{\star}$

TABLE 1. Per-Iteration Computational Complexity Comparison

Scheme	Precoder Subproblem	RIS Subproblem
Proposed RSMA	$\mathcal{O}\big((BM)^3(K+1)\big)$	$\mathcal{O}(N^2 N_{\text{sweep}})$
SDMA (optimized)	$\mathcal{O}\big((BM)^3 K\big)$	$\mathcal{O}(N^2 N_{\text{sweep}})$
Heuristic RSMA	$\mathcal{O}\big((BM)^3 K + BMK\big)$	$\mathcal{O}(N^2 N_{\text{sweep}})$
NOMA	$\mathcal{O}(BMK)$	$\mathcal{O}(N^2 N_{\text{sweep}})$

Notes: The $(K+1)$ term accounts for the additional common precoder $\mathbf{w}_0$. For $K = 12$, the precoder-subproblem complexity term of the proposed RSMA design is approximately $(K+1)/K = 13/12 \approx 1.08$ times that of optimized SDMA. The RIS-subproblem complexity is identical for both schemes.

The per-iteration complexity is dominated by two components: the convex precoder subproblem (27), which requires $\mathcal{O}\big((BM)^3(K+1)\big)$ operations via interior-point methods, and the Gauss-Seidel RIS update, which requires $\mathcal{O}(N^2 N_{\text{sweep}})$ operations over N_{sweep} coordinate sweeps. A detailed comparison across all schemes is provided in Table 1.

F. Real-Time Feasibility

The proposed WMMSE–AO framework is intended primarily as a centralized reference design for quantifying the achievable gain of joint RSMA–RIS optimization. It is not intended as a fully optimized real-time physical layer (PHY) implementation. Each AO iteration consists of a convex WMMSE update that jointly optimizes the common and private precoders. This is followed by an element-wise Gauss-Seidel RIS phase update.

The main computational components are the convex WMMSE precoder update in (27) and the element-wise Gauss-Seidel RIS update in (32). Their per-iteration complexities are analyzed in Section IV-E and summarized in Table 1. Let

$$T_{\text{alg}} = I_{\text{AO}}\big(T_{\text{WMMSE}} + T_{\text{RIS}} + T_{\text{misc}}\big) \tag{33}$$

denote the total computation time required to obtain one solution. Here, I_{AO} is the number of AO iterations, T_{WMMSE} is the time required for the joint common/private precoder update, and T_{RIS} is the time required for the Gauss-Seidel RIS sweep. The term T_{misc} accounts for channel aggregation, MMSE receiver and weight updates, objective evaluation, and convergence checks.

The relevant latency budget is determined by the interval over which the design CSI remains valid. It is not determined by the asynchronous evaluation horizon, which comprises P time instants separated by T_s. In the adopted coherence-block model, the precoders and RIS phases are computed from the channel estimate available at the beginning of the block. They are then held fixed during data transmission.

Hence, an online implementation must complete the optimization within the channel-update interval. This interval is upper-bounded by the channel coherence time T_c. We do not claim that the CVX-based interior-point implementation satisfies this latency requirement on general-purpose hardware. Its role is to provide a centralized reference implementation for evaluating the performance of the proposed joint design. Real-time deployment would require implementation-specific profiling and lower-complexity solver realizations, as discussed below.

Several implementation choices can reduce the online computational burden. First, the AO procedure already uses an SDMA warm-start. The RSMA precoder is initialized from the converged SDMA solution. The same principle can be extended across coherence blocks by reusing the previous-block solution when consecutive channel estimates are correlated. Early stopping can further reduce the iteration count. Fig. 2 shows that most of the sum-rate improvement is obtained within the first few iterations.

Second, the framework jointly optimizes the active precoders and RIS phases using the instantaneous channel estimate available at the beginning of each coherence block. In practice, however, these variables can be updated on different timescales. Following the two-timescale principle in [32], a practical implementation can update the RIS phases less frequently using statistical CSI, while updating the active precoders more often using instantaneous effective CSI. This removes the RIS update from most per-block operations, leaving only the precoder update in most channel blocks.

Third, the generic CVX-based interior-point solver could be replaced by customized first-order or primal-dual methods [33], [34]. Such methods can provide more predictable per-iteration complexity and are better suited to parallelization and hardware acceleration than a general-purpose convex solver.

Alternatively, selected WMMSE update steps could be unfolded into a trainable architecture. Deep-unfolding approaches have been demonstrated for multi-user MIMO (MU-MIMO) beamforming [35], [36]. They have also been applied to RSMA digital precoding within learning-based RIS-aided systems [37]. These alternatives provide complementary paths for reducing the online computation time of the proposed framework.

From a deployment perspective, the centralized coordination assumed in this work is consistent with O-RAN architectures, where baseband processing can be hosted at the O-RAN Distributed Unit (O-DU) and slower control functions can be supported by the near-real-time RAN Intelligent Controller (near-RT RIC) [27], [38]. Accordingly, fast per-block precoder updates are better suited to the O-DU, whereas slower RIS updates can be coordinated by the near-RT RIC.

V. Baseline Schemes for Performance Benchmarking

To evaluate the proposed joint RSMA–RIS framework, four baseline categories are considered. The precoder design and RIS configuration used for each baseline are specified in the corresponding subsection.

A. Optimized SDMA

The optimized SDMA baseline uses the same WMMSE–AO framework and RIS design as the proposed RSMA scheme, but without rate splitting. The common precoder is set to $\mathbf{w}_{b,0} = \mathbf{0}$ for all APs, and the full transmit power $P_{\max}$ is allocated to the private streams. Subproblem (27) is simplified by removing the common rate variable $c_{\min}$ and its associated constraints (27c)-(27e). This isolates the contribution of rate splitting from other algorithmic improvements.

B. RZF-SDMA

The RZF-SDMA baseline isolates the effect of the private precoder design. The RIS phases are first obtained from the optimized SDMA AO procedure and are then held fixed. For the resulting aggregate effective channel matrix $\mathbf{H}$, the RZF precoder is given by

$$\mathbf{W}_{\mathrm{RZF}} = \sqrt{P_{\max}}\, \frac{\mathbf{H}\left(\mathbf{H}^H\mathbf{H} + \frac{K\sigma^2}{P_{\max}}\mathbf{I}_K\right)^{-1}}{\left\|\mathbf{H}\left(\mathbf{H}^H\mathbf{H} + \frac{K\sigma^2}{P_{\max}}\mathbf{I}_K\right)^{-1}\right\|_F}. \tag{34}$$

The regularization parameter $K\sigma^2/P_{\max}$ limits noise enhancement and improves numerical conditioning [39]. As the transmit power increases, the regularization parameter decreases and the precoder approaches an unregularized channel-inversion solution. However, exact cancellation of all multi-user interference is not possible in the considered overloaded regime $K > BM$. Like SDMA, this baseline contains no common stream and allocates the full transmit power to the private streams.

C. NOMA

The NOMA baseline uses conventional power-domain NOMA: maximum-ratio transmission (MRT) precoders $\mathbf{w}_{b,k} \propto \mathbf{h}_{b,k}^{\mathrm{sync}}$ with inverse-channel-gain power allocation, and receiver-side SIC with users ordered by ascending channel gain [11]. For this comparison, the RIS phases are fixed to those obtained from the optimized SDMA procedure. More sophisticated NOMA designs, such as WMMSE-based variants, are beyond the scope of this paper. The baseline is included to provide a conventional power-domain NOMA comparison with the proposed joint WMMSE–RSMA framework.

D. Heuristic RSMA

In existing RSMA implementations, the common precoder $\mathbf{w}_0 \in \mathbb{C}^{BM\times 1}$ is designed *independently* of the private precoders $\{\mathbf{w}_k\}$. Let $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{BM\times K}$ denote the aggregate channel matrix stacking all K user channels column-wise. Four heuristic constructions of the unnormalized common precoder $\hat{\mathbf{w}}_0$ are considered:

- *Maximum-Ratio (MR)* [13]: $\hat{\mathbf{w}}_0^{\mathrm{MR}}$ is the eigenvector of $\mathbf{H}\mathbf{H}^H$ associated with its largest eigenvalue, maximizing the total received common-signal energy $\sum_k |\mathbf{h}_k^H\mathbf{w}_0|^2$ across users.
- *Equal-Gain Combining (EGC)* [18]: $\hat{\mathbf{w}}_0^{\mathrm{EGC}} = \sum_{k=1}^{K}\mathbf{h}_k/\|\mathbf{h}_k\|$, which gives every user equal phase weight regardless of channel strength.
- *Signal-to-Leakage-plus-Noise Ratio (SLNR)* [40]: $\hat{\mathbf{w}}_0^{\mathrm{SLNR}}$ is obtained by iteratively maximizing $\min_k \gamma_{c,k}$ via a reweighted generalized eigenvalue problem.
- *Multicast semidefinite relaxation (SDR)* [41]: $\hat{\mathbf{w}}_0^{\mathrm{SDR}}$ is the rank-one solution of $\max_{\|\mathbf{w}\|\le 1}\min_k |\mathbf{h}_k^H\mathbf{w}|^2$, recovered from a semidefinite relaxation followed by Gaussian randomization.

After $\hat{\mathbf{w}}_0$ is computed by any of the four rules above, it is normalized and combined with the WMMSE-optimized private precoders, with the total power split between common and private streams via a one-dimensional grid search. All four schemes share the same WMMSE private precoder optimization and RIS design as the proposed scheme, differing only in the common precoder design, which is constructed independently rather than jointly optimized with the private precoders. A detailed comparison of the per-iteration computational cost across all schemes is provided in Table 1.

E. Limited Gain of Heuristic Common-Precoder Designs

The limited gain of the heuristic common-precoder designs in Section V-D is examined next. These designs have shown effectiveness in co-located MIMO systems [13], [14], but provide little gain in the considered overloaded asynchronous CF-MIMO setting. All schemes use the same SIC procedure, in which the common stream is decoded and removed before the private stream is decoded. They differ only in the common precoder design. In the heuristic schemes, $\hat{\mathbf{w}}_0$ is constructed using the channel matrix before $\{\mathbf{w}_k\}$ are optimized. The common precoder therefore remains fixed rather than being optimized jointly with the private precoders.

1) Failure of the Decoupled Common-Precoder Design

For a fixed heuristic common direction $\hat{\mathbf{w}}_0$, let $\alpha P_{\max}$ denote the power allocated to the common stream and $(1-\alpha)P_{\max}$ the power allocated to the private streams. Let $R_c^{\mathrm{heur}}(\alpha)$ and $R_{p,k}^{\mathrm{heur}}(\alpha)$ denote the common and private rates in (16) and (17), respectively, under heuristic RSMA with common-power fraction α. Let $R_{p,k}^{\mathrm{SDMA}}$ denote the private rate in (17) under optimized SDMA. Since SDMA contains no common stream, $R_{p,k}^{\mathrm{SDMA}}$ is independent of α.

The aggregate private-rate loss relative to optimized SDMA is

$$\Delta R_p(\alpha) = \sum_{k=1}^{K} \left(R_{p,k}^{\mathrm{SDMA}} - R_{p,k}^{\mathrm{heur}}(\alpha) \right). \tag{35}$$

The resulting sum-rate change is therefore

$$\Delta R_{\mathrm{sum}}(\alpha) = R_c^{\mathrm{heur}}(\alpha) - \Delta R_p(\alpha). \tag{36}$$

A gain over optimized SDMA is obtained only when

$$R_c^{\mathrm{heur}}(\alpha) > \Delta R_p(\alpha) \tag{37}$$

for some $\alpha > 0$. Thus, the weakest-user common rate must exceed the total private-rate reduction caused by allocating power to the common stream. This condition becomes difficult to satisfy when AP-specific oscillator drift attenuates coherent cross-AP combining, as quantified in Proposition 1.

In a distributed cell-free deployment, this condition is difficult to satisfy because the common stream must combine coherently across geographically separated APs at all users. Let

$$a_{b,k} = \mathbf{h}_{b,k}^H[1]\mathbf{w}_{b,0} \tag{38}$$

denote the contribution of AP b to the common stream at UE k at the initial time instant $p = 1$.

For compactness, let $\mathbf{a}_k = [a_{1,k}, \ldots, a_{B,k}]^T$ collect the fixed common-stream contributions at the design instant $p = 1$.

The channel $\mathbf{h}_{b,k}[1]$ includes the direct and RIS-assisted components, the fixed propagation-delay phases, and the RIS response. Under the adopted channel model, the direct and reflected components associated with AP b experience the same AP oscillator drift, while the UE oscillator phase is common to all AP contributions received by UE k. The common-stream power at UE k and time instant p can therefore be written as

$$S_{c,k}[p] = \left| \sum_{b=1}^{B} a_{b,k} e^{j\delta_{b,k}[p]} \right|^2, \tag{39}$$

where $\delta_{b,k}[p] = \left(\psi_b^{\mathrm{AP}}[p] - \psi_b^{\mathrm{AP}}[1]\right) + \left(\psi_k^{\mathrm{UE}}[p] - \psi_k^{\mathrm{UE}}[1]\right)$ denotes the residual oscillator-induced phase change of the AP-b contribution relative to $p = 1$. Using $|z|^2 = zz^*$ and grouping each pair of conjugate cross terms, (39) becomes

$$S_{c,k}[p] = \underbrace{\sum_{b=1}^{B} |a_{b,k}|^2}_{\text{Individual-AP power terms}} + \underbrace{\sum_{b=1}^{B} \sum_{\substack{b'=1 \\ b' \neq b}}^{B} a_{b,k} a_{b',k}^* e^{j\left(\delta_{b,k}[p] - \delta_{b',k}[p]\right)}}_{\text{Cross-AP combining terms}}. \tag{40}$$

The first term contains the individual AP power contributions. The second contains pairwise cross-AP combining terms, which occur in conjugate pairs and therefore sum to a real value. For UE k, define the relative phase error between APs b and b' as

$$\Delta\delta_{b,b',k}[p] \triangleq \delta_{b,k}[p] - \delta_{b',k}[p]. \tag{41}$$

The UE oscillator phase is common to both AP contributions and therefore cancels in this difference. Hence,

$$\Delta\delta_{b,b',k}[p] = \left(\psi_b^{\mathrm{AP}}[p] - \psi_b^{\mathrm{AP}}[1]\right) - \left(\psi_{b'}^{\mathrm{AP}}[p] - \psi_{b'}^{\mathrm{AP}}[1]\right). \tag{42}$$

Since the oscillator phases are initialized to zero at $p = 1$, this reduces to

$$\Delta\delta_{b,b',k}[p] = \psi_b^{\mathrm{AP}}[p] - \psi_{b'}^{\mathrm{AP}}[p]. \tag{43}$$

The expression in (40) therefore depends on the oscillator processes only through the AP phase differences. This justifies taking the expectation with respect to $\boldsymbol{\psi}^{\mathrm{AP}}[p]$ alone. Each AP oscillator accumulates $p-1$ independent Gaussian increments with per-step variance σ_ϕ^2. Since the AP oscillator processes are independent,

$$\Delta\delta_{b,b',k}[p] \sim \mathcal{N}\left(0,\, 2(p-1)\sigma_\phi^2\right). \tag{44}$$

For a zero-mean Gaussian variable X with variance v, $\mathbb{E}[e^{jX}] = e^{-v/2}$. Applying this identity gives

$$\mathbb{E}\left[e^{j\Delta\delta_{b,b',k}[p]}\right] = e^{-(p-1)\sigma_\phi^2}. \tag{45}$$

The resulting expected common-stream power is stated next.

Proposition 1 (Attenuation of coherent cross-AP combining)**.** *Conditioned on the fixed common-stream contribution vector $\mathbf{a}_k$ defined at $p = 1$, the expected received common-stream power at UE k over the random AP oscillator-phase vector $\boldsymbol{\psi}^{\mathrm{AP}}[p]$ is given by*

$$\mathbb{E}_{\boldsymbol{\psi}^{\mathrm{AP}}[p]}[S_{c,k}[p] \mid \mathbf{a}_k] = \sum_{b=1}^{B} |a_{b,k}|^2 + \sum_{b=1}^{B} \sum_{\substack{b'=1 \\ b' \neq b}}^{B} a_{b,k} a_{b',k}^* e^{-(p-1)\sigma_\phi^2}. \tag{46}$$

Thus, AP-specific oscillator drift attenuates the coherent cross-AP combining terms by $e^{-(p-1)\sigma_\phi^2}$, while leaving the individual AP power terms unchanged.

Proof:
The result follows by taking the expectation of (40) over $\psi^{\mathrm{AP}}[p]$, generated by the stochastic AP Wiener-process increments, while conditioning on the fixed common-stream contribution vector $\mathbf{a}_k$ defined at $p=1$. Substitution of (45) yields (46). ∎

Two limiting regimes follow from Proposition 1. When $(p-1)\sigma_\phi^2 \to 0$, the attenuation factor approaches one. Both the individual AP power terms and the cross-AP combining terms are then retained, giving $|\sum_b a_{b,k}|^2$. When $(p-1)\sigma_\phi^2 \to \infty$, the attenuation factor approaches zero and the cross-AP terms vanish in expectation. The individual AP power terms remain, giving $\sum_b |a_{b,k}|^2$. Thus, for a UE whose design-time AP contributions combine constructively, oscillator drift progressively reduces the additional power provided by coherent cross-AP combining. Since the common rate in (16) is limited by the weakest UE, this reduction at one UE can restrict the common rate available to all users.

This attenuation makes the gain condition in (37) more difficult to satisfy. A fixed heuristic direction $\hat{\mathbf{w}}_0$ is constructed independently of the private precoders and is therefore not jointly optimized to balance the weakest-user common rate against the aggregate private-rate loss. Moreover, allocating a fraction α of the transmit power to the common stream leaves less power for the K private streams, and the resulting rate reductions are summed in $\Delta R_p(\alpha)$.

For the tested heuristic precoders in the considered overloaded CF-MIMO setting, the numerical results show that the condition in (37) is not satisfied. Consequently, the power search selects α close to zero, causing heuristic RSMA to approach optimized SDMA. The analysis above identifies the cross-AP attenuation mechanism that contributes to the limited gain of the tested heuristic designs observed in Figs. 6–7.

2) What Joint Optimization Recovers

The proposed WMMSE framework in Section IV removes this decoupling by optimizing $\mathbf{w}_0$ and $\{\mathbf{w}_k\}$ jointly within the same convex precoder subproblem. The common precoder is therefore designed in coordination with the private precoders, enabling the common stream to exploit cross-AP combining while accounting for the interference generated by the private streams. The results in Section VI support this distinction. The heuristic common-precoder schemes remain close to optimized SDMA, whereas the jointly optimized RSMA design achieves a clear gain in overloaded cell-free regimes.

VI. Performance Evaluation

The system parameters are summarized in Table 2. Users are randomly clustered around $(d, 0)$ with radius 1 m. All reported sum rates are averaged over $N_{\mathrm{MC}} = 100$ independent Monte Carlo realizations of the channels, user positions, and oscillator phase-noise.

TABLE 2. Simulation Parameters

Parameter	Value
Number of APs (B)	4
Antennas per AP (M)	2
Number of users (K)	12
Network loading (K/BM)	1.5
RIS elements (N)	100
Carrier frequency (f_c)	2 GHz
Symbol duration (T_s)	10 μs
Noise power (σ^2)	10^{-11} W
Oscillator coefficient (c_{osc})	10^{-18}
Async. time instants (P)	50
Total transmit-power budget ($P_{\max}$)	-20 to 30 dBW
User distance (d)	80 m
Path-loss intercept (C_0)	10^{-3}
AP-UE path-loss exponent (κ_{BU})	3.5
AP-RIS path-loss exponent (κ_{BR})	2.0
RIS-UE path-loss exponent (κ_{RU})	2.1
Rician K-factor (K_R)	0 (default)
AO iterations ($T_{\max}$)	80
QoS fraction	0.20
Monte Carlo runs	100

The per-user QoS threshold is selected adaptively as $R_{\min} = 0.20 \min_k R_k^{\text{no-RIS}}$, where $R_k^{\text{no-RIS}}$ is the rate of user k under the no-RIS MMSE initialization. This provides a conservative, realization-dependent QoS requirement.

The implementation records the solver status of every QoS-constrained WMMSE precoder update. Across all reported Monte Carlo trials, the counter recorded zero infeasible or inaccurate QoS-constrained RSMA updates. Hence, the QoS-relaxed fallback was not invoked in the reported results. If needed, this fallback relaxes only the per-user QoS constraints. The remaining precoder constraints are retained, and the RIS unit-modulus constraints remain enforced separately in the Gauss-Seidel RIS update.

For each realization, the asynchronous performance is evaluated by computing R_{sum} in (18), which averages the instantaneous rate expressions across the P asynchronous time instants using precoders designed on the synchronous channel ($p = 1$). The optimized SDMA baseline uses the same WMMSE–AO RIS optimization framework as the proposed method but removes the common stream, as defined in Section V. Throughout the results, the RSMA gain refers to the relative sum-rate improvement of the proposed RSMA over optimized SDMA, expressed as a percentage, $(R_{\mathrm{RSMA}} - R_{\mathrm{SDMA}})/R_{\mathrm{SDMA}} \times 100\%$. When absolute differences are reported, the unit bps/Hz is stated explicitly.

The default configuration represents practical scenarios such as indoor hotspots, factory floors, or dense urban picocells where a limited number of multi-antenna APs cooperatively serve a user cluster. The default loading ratio

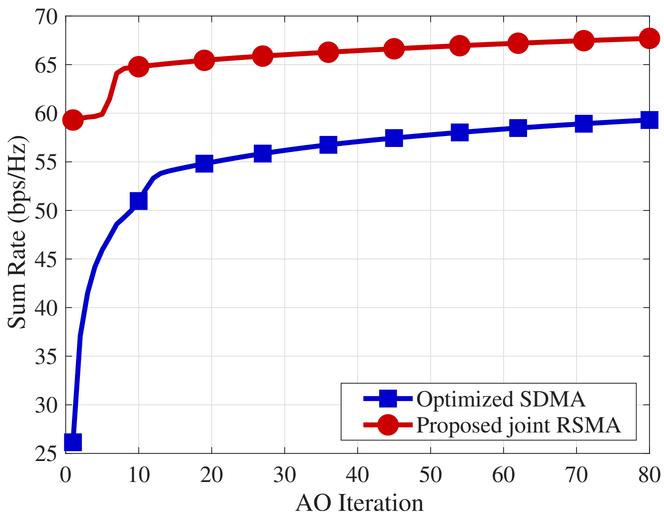


FIGURE 2. Convergence behaviour of the proposed AO algorithm at $P_{\max} = 10$ dBW. The proposed RSMA is warm-started from the converged SDMA solution. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

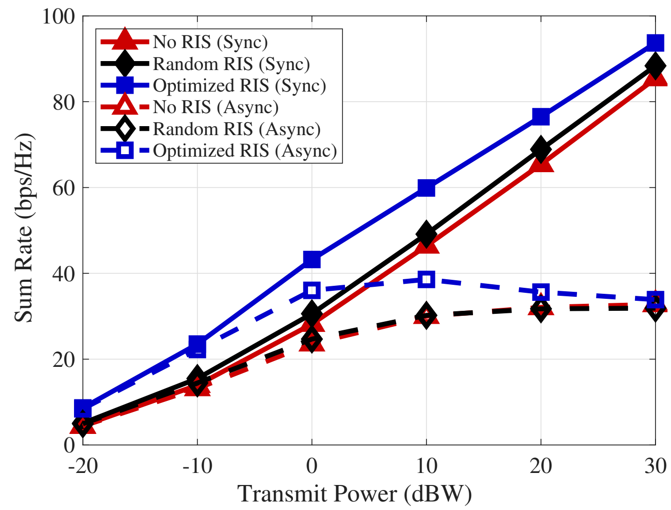


FIGURE 3. SDMA sum rate under different RIS configurations. Solid and dashed lines denote synchronous and asynchronous channels, respectively. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

TABLE 3. Sensitivity of the proposed RSMA gain to the SDMA warm-start. The cold-initialization variant uses random RIS phases and MMSE-initialized private precoders on the resulting effective channel, without any SDMA warm-start. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

$P_{\max}$ (dBW)	SDMA baseline	RSMA warm-start	RSMA cold-init	Warm gain over SDMA	Cold gain over SDMA
0	43.81	44.96	45.79	+2.6%	+4.5%
20	78.09	88.19	81.11	+12.9%	+3.9%

places the system in the overloaded regime where RSMA provides its primary benefit. In classical cell-free massive MIMO with $B \gg K$ and single-antenna APs, the system is underloaded and SDMA can already suppress all interference, leaving no role for rate splitting. The distributed AP geometry is central to these findings, as each geographically separated AP contributes a different spatial signature to each user, preventing an independently designed common precoder from simultaneously serving all users. Section VI-H confirms that this effect persists at larger system scale.

A. Convergence

The convergence behaviour discussed in Section IV-E is illustrated in Fig. 2. The SDMA baseline converges from random initialization, reaching approximately 55 bps/Hz within the first 15–20 iterations and converging to approximately 59 bps/Hz. The proposed RSMA, warm-started from the converged SDMA solution, benefits immediately from introducing the common stream and converges to approximately 68 bps/Hz within a few additional iterations. Both curves exhibit monotonic non-decrease, with the gap between the converged values reflecting the RSMA gain over optimized SDMA at this power level.

The sensitivity of the proposed RSMA design to initialization is examined in Table 3. The warm-started variant initializes the RSMA AO procedure from the converged SDMA solution. The cold-initialization variant instead draws the RIS phases uniformly at random and initializes the private precoders using an MMSE design on the resulting effective channel, without using the converged SDMA solution. Both variants use the same AO iteration budget and stopping criterion.

The cold-initialized variant still improves over the optimized SDMA baseline at both tested transmit powers, achieving a $+4.5\%$ gain at $P_{\max} = 0$ dBW and a $+3.9\%$ gain at $P_{\max} = 20$ dBW. Therefore, the SDMA warm-start is not essential for obtaining a positive RSMA gain.

The sensitivity to initialization becomes more pronounced at high transmit power. At $P_{\max} = 20$ dBW, the warm-started RSMA design achieves 88.19 bps/Hz, while the cold-initialized variant achieves 81.11 bps/Hz. The corresponding optimized SDMA baseline is 78.09 bps/Hz. This gives an RSMA gain of $+12.9\%$ with the warm-start and $+3.9\%$ with the cold initialization. The two initializations therefore do not achieve comparable performance in this regime, and the SDMA warm-start leads to a substantially higher converged sum rate.

At $P_{\max} = 0$ dBW, the cold-initialized variant gives a slightly higher sum rate than the warm-started variant. The difference is small, and both initializations achieve comparable performance at this lower transmit power. This indicates that the solution is less sensitive to initialization when multi-user interference is less dominant.

B. Sum Rate vs. Transmit Power

The role of the RIS is examined in Figs. 3 and 4, which show the SDMA and proposed RSMA sum rates under three RIS configurations: no RIS, random RIS phases, and jointly optimized RIS phases. In both figures, the no-RIS and random-RIS curves nearly overlap, confirming that un-

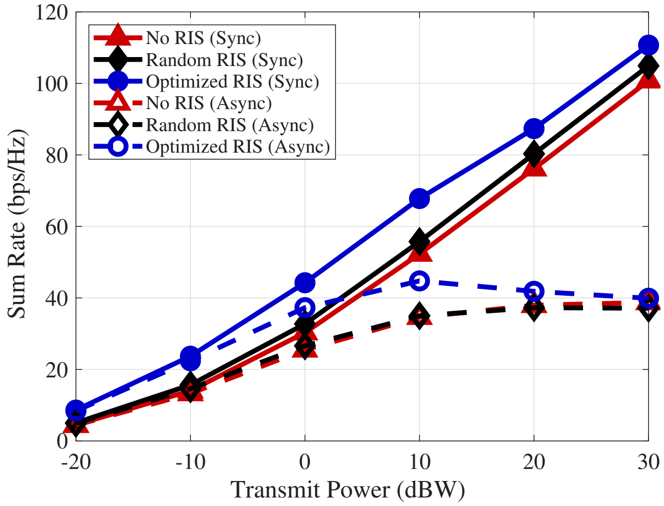


FIGURE 4. Proposed RSMA sum rate under different RIS configurations. Solid and dashed lines denote synchronous and asynchronous channels, respectively. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

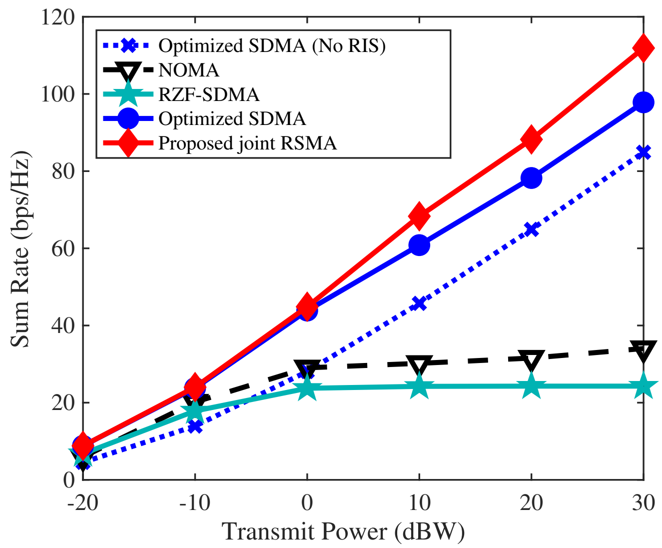


FIGURE 5. Sum rate versus total transmit power. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$, $K/BM = 1.5$.

optimized RIS phases provide negligible benefit. Joint AP-RIS phase optimization yields a clear gain over the no-RIS baseline across the tested power range. The proposed RSMA maintains higher absolute rates than SDMA under every RIS configuration, at every power level, and under both synchronous and asynchronous channels. Both figures also show that the asynchronous curves saturate above $P_{\max} = 10$ dBW, illustrating that oscillator drift caps the achievable sum rate in the high-power regime.

Fig. 5 compares the sum rate of the proposed RSMA framework with the Optimized SDMA, RZF-SDMA, NOMA, and No-RIS SDMA baselines as a function of the total transmit power under the default configuration ($B = 4$, $K = 12$, $M = 2$). Three key observations emerge.

1) RSMA gain grows with transmit power: At low transmit power ($P_{\max} = -20$ dBW), all schemes operate in the noise-limited regime and the proposed RSMA provides only a marginal gain over optimized SDMA. As the transmit power increases, the system becomes increasingly interference-limited and the benefit of rate splitting becomes more pronounced. At $P_{\max} = 30$ dBW, the proposed RSMA achieves 111.89 bps/Hz, compared with 97.83 bps/Hz for optimized SDMA, corresponding to a sum-rate gain of approximately 14.4%.

2) RZF-SDMA saturates in the overloaded regime: The RZF-SDMA baseline, in which the private precoders are computed in closed form without WMMSE optimization, achieves 24.23 bps/Hz at $P_{\max} = 10$ dBW and only 24.29 bps/Hz at $P_{\max} = 30$ dBW. This limited increase is consistent with the overloaded setting, where the available spatial dimension $BM = 8$ is smaller than the user load $K = 12$ and the RZF precoder cannot fully suppress multi-user interference. The result supports the use of WMMSE-optimized SDMA as the primary baseline throughout the paper.

3) Comparison with conventional NOMA: Power-domain NOMA with MRT precoding and SIC consistently underperforms the optimized SDMA baseline across the entire power range. At $P_{\max} = 30$ dBW, NOMA achieves only about 34 bps/Hz, compared with 97.83 bps/Hz for optimized SDMA, a deficit of approximately 65%. These results show that the considered power-domain NOMA baseline achieves lower sum rates than WMMSE-based spatial precoding in this overloaded cell-free MIMO configuration.

C. Common-Precoder Design Analysis

This subsection examines whether the RSMA gain is determined by the common-precoder design. Representative heuristic common precoders are first compared with the optimized SDMA baseline. The effect of the private precoder design is then isolated by repeating the comparison with RZF private precoders.

1) Limited Gain of the Tested Heuristic Common-Precoder Designs

Fig. 6 compares the proposed joint RSMA design with optimized SDMA, RZF-SDMA, and the MR- and EGC-based heuristic RSMA schemes defined in Section V-D. The RZF-SDMA curve is the same fixed-RIS baseline used in Fig. 5. Only MR and EGC are shown for visual clarity. The remaining heuristic designs are included in Table 4.

The heuristic RSMA curves remain close to the optimized SDMA baseline across the entire power range. At $P_{\max} = 20$ dBW, MR and EGC both achieve approximately 78.30 bps/Hz, compared with 78.22 bps/Hz for optimized SDMA. This corresponds to a gain of only about 0.1%. By comparison, the proposed joint RSMA design provides approximately 10.0 bps/Hz additional sum rate at the same

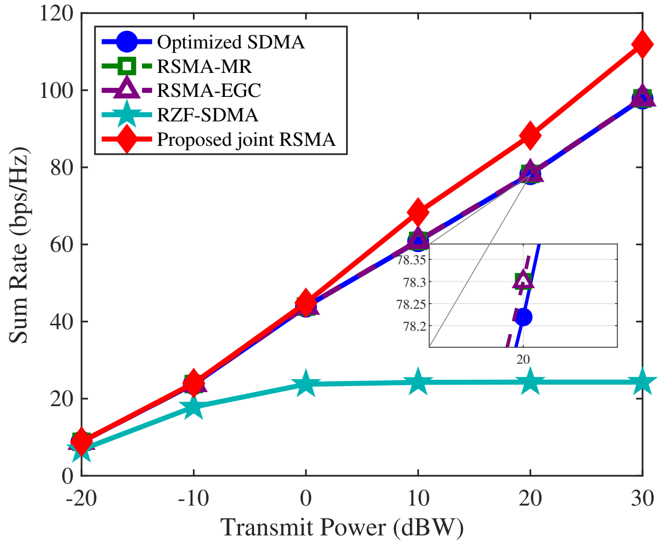


FIGURE 6. **Sum rate versus transmit power for optimized SDMA, RZF-SDMA, the MR- and EGC-based heuristic RSMA schemes, and the proposed joint RSMA design. The zoomed inset at $P_{\max} = 20$ dBW magnifies the overlap between the heuristic RSMA schemes and optimized SDMA. System parameters:** $B = 4$, $K = 12$, $M = 2$, $N = 100$.

operating point, corresponding to a gain of about 12.8%. The similar behaviour of MR and EGC indicates that the limited gain is not specific to either common-precoder construction.

The RZF-SDMA baseline saturates near 24 bps/Hz. This behaviour is consistent with the considered overloaded setting, where the available spatial dimension $BM = 8$ is smaller than the user load $K = 12$ and exact cancellation of all multi-user interference is not possible. The negligible gains of MR and EGC also contrast with [23], where heuristic common precoding provided gains of 0.1–2.2 bps/Hz when combined with conjugate private beamforming. This contrast motivates the private precoder isolation test presented next.

2) Isolating the Private-Precoder Contribution

To determine whether the limited heuristic gain is caused only by the WMMSE private precoder design, Table 4 evaluates each heuristic common precoder with both WMMSE-optimized and closed-form RZF private precoders. For the RZF-based comparisons, the RIS phases obtained from the optimized SDMA procedure are held fixed. The percentage gains are measured relative to optimized SDMA in the WMMSE column and RZF-SDMA in the RZF column.

With RZF private precoding, the tested heuristic common precoders provide at most a 1.86% gain, achieved by multicast SDR, compared with at most 0.15% when WMMSE-optimized private precoders are used. The larger gains with RZF suggest that a heuristic common stream has more opportunity to contribute when the private precoders leave more residual interference. However, the gain remains substantially below the 12.30% achieved by the proposed joint RSMA design over optimized SDMA.

TABLE 4. **Sum-rate comparison of the tested common-precoder designs with WMMSE-optimized and RZF private precoders at $P_{\max} = 10$ dBW. Values in parentheses denote the percentage gain over the corresponding no-common-stream baseline. System parameters:** $B = 4$, $K = 12$, $M = 2$, $N = 100$.

Scheme	WMMSE private (bps/Hz)	RZF private (bps/Hz)
No common stream	60.83	24.23
Heuristic MR	60.92 (+0.15%)	24.48 (+1.06%)
Heuristic EGC	60.92 (+0.15%)	24.37 (+0.59%)
Heuristic SLNR	60.92 (+0.15%)	24.23 (+0.02%)
Multicast SDR	60.92 (+0.15%)	24.68 (+1.86%)
Proposed joint RSMA	**68.31 (+12.30%)**	—

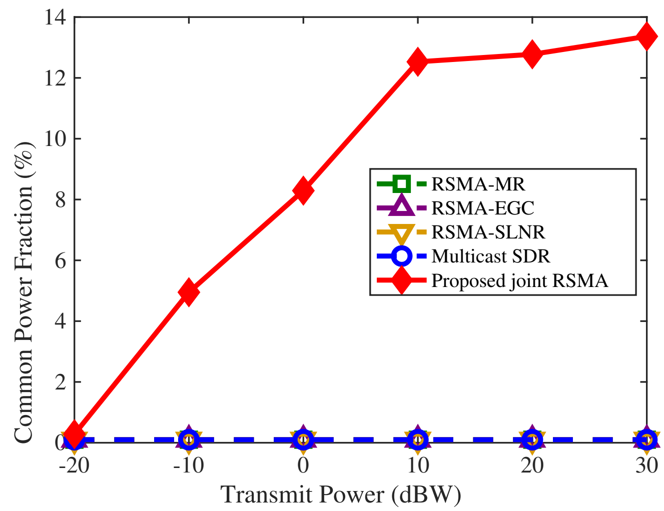


FIGURE 7. **Fraction of total transmit power allocated to the common stream versus transmit power, for the proposed joint RSMA and four heuristic RSMA designs (MR [13], EGC [18], SLNR [40], and Multicast SDR [41]). System parameters:** $B = 4$, $K = 12$, $M = 2$, $N = 100$.

The limited heuristic gain therefore persists with both private-precoder designs and cannot be explained solely by the use of WMMSE-optimized private precoders. This result is consistent with the gain condition in (37) and supports joint optimization of the common and private precoders in the considered overloaded CF-MIMO setting.

D. RSMA Gain Mechanism: Common-Power Allocation and Total Gain

Fig. 7 shows the fraction of the total transmit power allocated to the common stream. A value near zero indicates operation close to optimized SDMA, while a larger value indicates greater reliance on rate splitting.

The proposed joint RSMA design allocates little power to the common stream in the noise-limited regime ($P_{\max} \leq -10$ dBW). As the transmit power increases and multi-user interference becomes more significant, the common-power

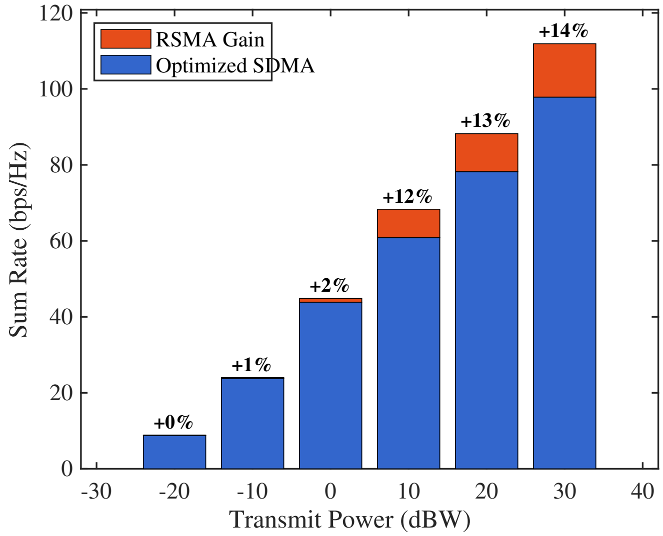


FIGURE 8. Sum rate at each transmit power level: optimized SDMA baseline (blue) and additional rate from the proposed joint RSMA (red), with relative gain annotated. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

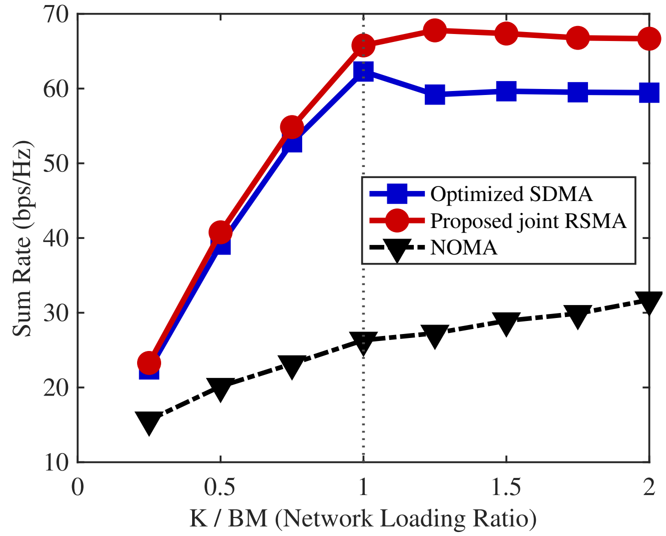


FIGURE 9. Sum rate versus the loading ratio K/BM **at** $P_{\max} = 10$ **dBW for the optimized SDMA baseline, NOMA, and the proposed joint RSMA scheme. The vertical dashed line marks the transition from underloaded to overloaded operation at** $K/BM = 1$**. System parameters:** $B = 4$, $M = 2$, $N = 100$.

fraction rises to approximately 12.5% at $P_{\max} = 10$ dBW and 13.4% at $P_{\max} = 30$ dBW. In contrast, all four heuristic schemes allocate only about 0.1% of the total power to the common stream across the tested power range and therefore remain close to optimized SDMA.

Fig. 8 shows the resulting sum-rate improvement of the proposed joint RSMA design over optimized SDMA. The absolute gain increases from approximately 7.5 bps/Hz at $P_{\max} = 10$ dBW to 14.1 bps/Hz at $P_{\max} = 30$ dBW, corresponding to relative gains of 12.3% and 14.4%, respectively. Together, the two figures show that the proposed scheme makes increasing use of the common stream as the system becomes interference-limited, whereas the decoupled heuristic designs retain a negligible common-power allocation and provide little additional sum rate.

E. Network Loading Ratio (K/BM) Analysis

The network loading ratio K/BM represents the ratio of users to available spatial dimensions and determines the extent to which multi-user interference can be managed through spatial separation. This subsection examines how the benefit of the proposed joint RSMA design changes as the system moves from underloaded to overloaded operation. The number of users is varied from $K = 2$ to $K = 16$ with $B = 4$ and $M = 2$, giving a loading ratio from $K/BM = 0.25$ to 2.0.

Fig. 9 shows that the optimized SDMA baseline and the proposed joint RSMA scheme achieve similar sum rates in the underloaded regime, where sufficient spatial degrees of freedom are available to separate the users. Once K/BM exceeds unity, the system becomes overloaded and the optimized SDMA baseline saturates around 60 bps/Hz. In the same region, the proposed joint RSMA scheme reaches approximately 67 bps/Hz and maintains a clear advantage over optimized SDMA. The common stream helps manage residual multi-user interference that private precoding alone cannot fully suppress. The considered NOMA baseline remains below both SDMA and RSMA across the loading range, indicating lower performance for this power-domain NOMA design in the considered cell-free MIMO setting.

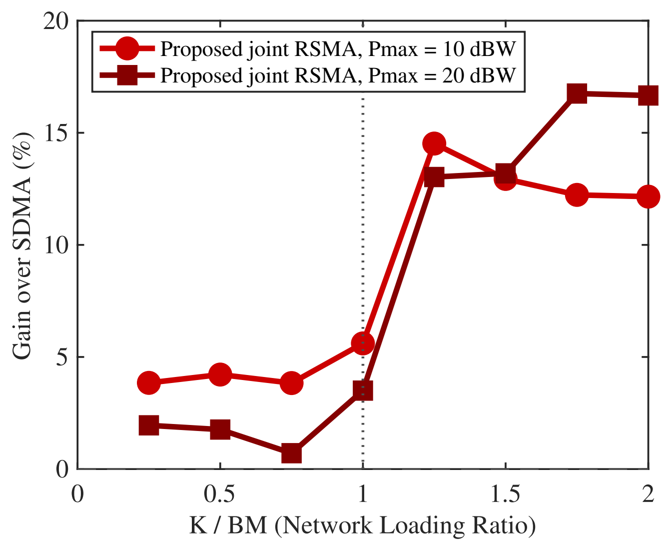


FIGURE 10. Percentage gain of the proposed joint RSMA scheme over the optimized SDMA baseline versus the loading ratio K/BM **at** $P_{\max} = 10$ **dBW and** 20 **dBW. The vertical dashed line marks the transition from underloaded to overloaded operation at** $K/BM = 1$**. System parameters:** $B = 4$, $M = 2$, $N = 100$.

Fig. 10 quantifies the relative gain of the proposed joint RSMA scheme over the optimized SDMA baseline. The gain exhibits a clear transition around $K/BM = 1$. In the underloaded regime ($K/BM \leq 1$), the gain remains

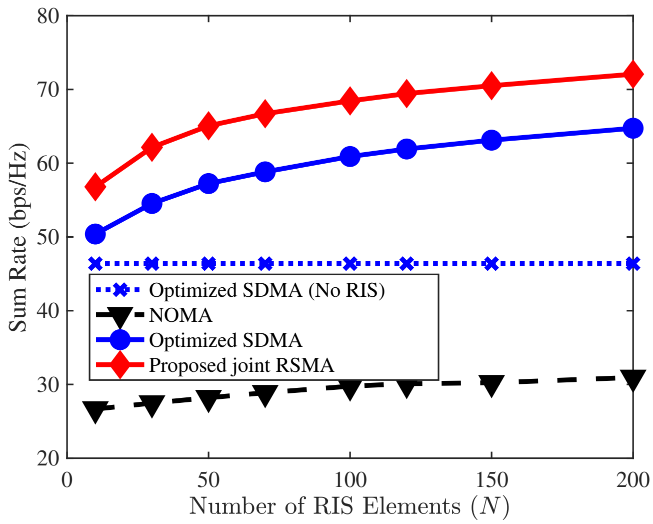


FIGURE 11. **Sum rate versus the number of RIS elements N at $P_{\max} = 10$ dBW for SDMA without RIS, NOMA, the optimized SDMA baseline, and the proposed joint RSMA. System parameters: $B = 4$, $K = 12$, $M = 2$.**

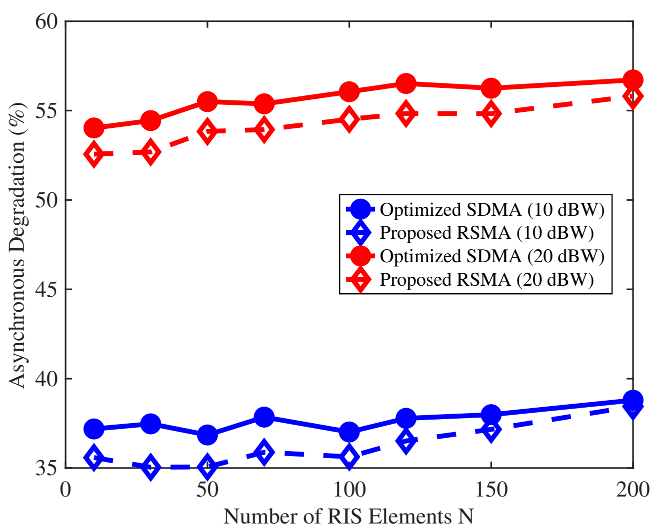


FIGURE 12. **Asynchronous degradation versus the number of RIS elements N for the optimized SDMA baseline and the proposed joint RSMA at $P_{\max} = 10$ dBW and 20 dBW. System parameters: $B = 4$, $K = 12$, $M = 2$.**

modest, ranging from approximately 0.7–5.6% depending on the transmit power. Once the system enters the overloaded regime ($K/BM > 1$), the gain rises to approximately 12–17% and remains at this higher level for both $P_{\max} = 10$ dBW and 20 dBW. These results confirm that the proposed joint RSMA design is most valuable in overloaded deployments, where spatial degrees of freedom are scarce and multi-user interference cannot be resolved by spatial separation alone.

F. Scaling With the Number of RIS Elements

The number of RIS elements is swept from $N = 10$ to 200, while all remaining parameters follow Table 2. Fig. 11 shows the resulting sum-rate performance at $P_{\max} = 10$ dBW. Both optimized SDMA and the proposed joint RSMA design benefit from the controllable reflected paths created by the optimized RIS phases, which improve the effective AP–UE channel gain. The optimized SDMA sum rate increases from approximately 50 bps/Hz at $N = 10$ to 65 bps/Hz at $N = 200$. Over the same range, the proposed joint RSMA sum rate increases from approximately 57 to 72 bps/Hz. The absolute advantage of the proposed joint RSMA design remains relatively stable at approximately 6–8 bps/Hz. This indicates that increasing the RIS size improves both schemes similarly. The RSMA advantage is therefore not primarily driven by the number of RIS elements, but by the joint optimization of the common and private precoders. NOMA remains below both optimized SDMA and joint RSMA throughout the tested range.

Fig. 12 shows that the asynchronous degradation increases gradually with N for both SDMA and RSMA. As the RIS size increases, the reflected AP–RIS–UE paths contribute more strongly to the effective channel. These paths are affected by AP oscillator phase drift and propagation-delay mismatch. Consequently, the loss of phase coherence produces a larger degradation when the reflected component becomes more prominent. The degradation is also greater at $P_{\max} = 20$ dBW than at $P_{\max} = 10$ dBW. At the higher transmit power, the system is more interference-limited and therefore more sensitive to synchronization errors. Together with the sum-rate improvement observed as N increases, these results reveal a trade-off between passive beamforming gain and robustness to asynchronous phase and delay mismatch.

G. Oscillator Phase-Noise Sensitivity

The oscillator phase-noise variance σ_ϕ^2 is swept from -50 dB to -20 dB to evaluate the sensitivity of the proposed scheme to synchronization mismatch in distributed AP deployments. Fig. 13 shows the sum rate at $P_{\max} = 10$ dBW under synchronous and asynchronous conditions. The corresponding synchronous reference rates are 60.5 and 67.9 bps/Hz for the optimized SDMA baseline and the proposed joint RSMA, respectively. These synchronous rates are independent of σ_ϕ^2, whereas the asynchronous rates decrease as the phase-noise variance increases.

Wiener phase drift accumulates within the coherence block and creates a mismatch between the channels used for precoder design and the realized channels. This mismatch reduces the coherent precoding gain and increases residual multi-user interference.

Fig. 14 quantifies the corresponding asynchronous degradation for optimized SDMA, the proposed joint RSMA design, and heuristic RSMA-MR. At $P_{\max} = 10$ dBW, the degradation increases from approximately 1–2% at $\sigma_\phi^2 = -50$ dB to approximately 62–63% at $\sigma_\phi^2 = -20$ dB. At $P_{\max} = 20$ dBW, it increases from approximately 7–8% to 74–75%. The greater degradation at the higher transmit

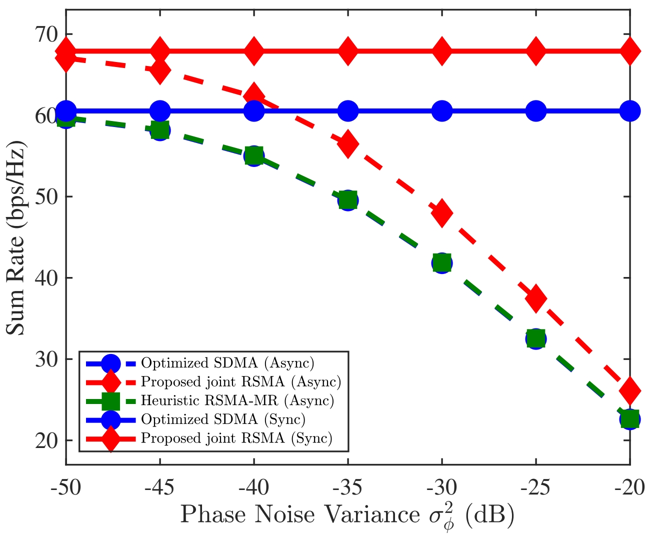


FIGURE 13. Sum rate versus oscillator phase-noise variance σ_ϕ^2 for the optimized SDMA baseline and the proposed joint RSMA under synchronous and asynchronous conditions at $P_{\max} = 10$ dBW. The heuristic RSMA-MR is shown under asynchronous operation. System parameters: $B = 4$, $K = 12$, $M = 2$, and $N = 100$.

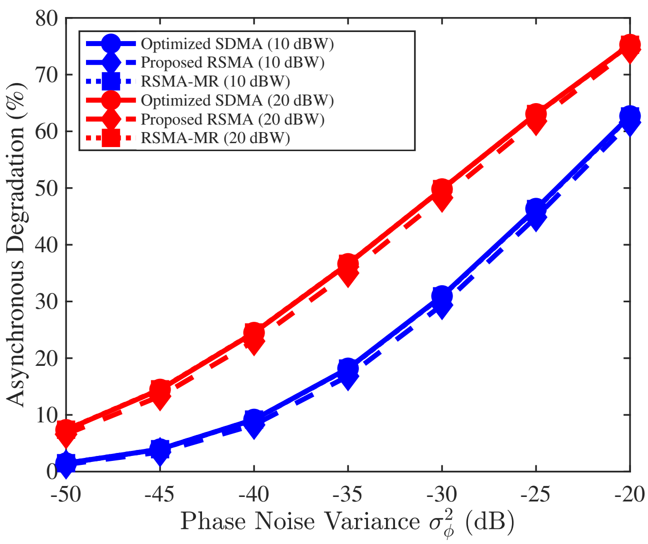


FIGURE 14. Asynchronous degradation versus σ_ϕ^2 for the optimized SDMA baseline, the proposed joint RSMA, and the heuristic RSMA-MR at $P_{\max} = 10$ dBW and 20 dBW. System parameters: $B = 4$, $K = 12$, $M = 2$, and $N = 100$.

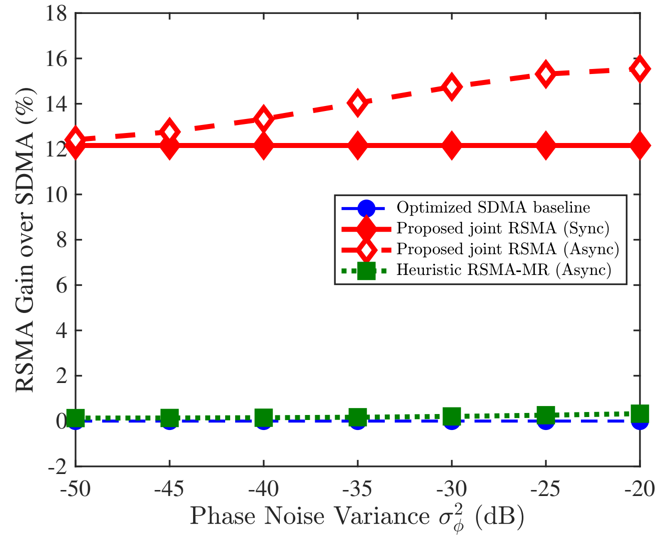


FIGURE 15. Gain of the proposed joint RSMA over the optimized SDMA baseline versus σ_ϕ^2 under synchronous and asynchronous conditions at $P_{\max} = 10$ dBW. The gain of the heuristic RSMA-MR over the optimized SDMA baseline under asynchronous operation is also shown for reference. System parameters: $B = 4$, $K = 12$, $M = 2$, and $N = 100$.

TABLE 5. Scalability validation at larger scale ($B = 8$, $M = 2$, $K = 24$, $BM = 16$, $K/BM = 1.5$, $N = 100$). The Best heuristic column reports the best result among MR, EGC, SLNR, and Multicast SDR common precoders. The Gain column reports the relative improvement of the proposed joint RSMA framework over the optimized SDMA baseline.

	Sum Rate (bps/Hz)					Gain
$P_{\max}$ (dBW)	NOMA	RZF	SDMA	Best heuristic	Proposed	(%)
-20	6.3	6.4	9.5	9.5	9.5	0.0
-10	23.3	17.9	26.0	26.0	26.0	0.1
$+0$	33.4	33.0	51.5	51.5	52.8	2.6
$+10$	36.5	40.6	84.5	84.6	89.6	6.0
$+20$	38.9	40.9	115.6	115.8	129.5	12.0
$+30$	42.2	40.9	152.3	152.4	171.6	12.6

power highlights the increased importance of accurate inter-AP synchronization in interference-limited operation.

Optimized SDMA and heuristic RSMA-MR exhibit similar degradation, whereas the proposed joint RSMA design is slightly less affected across the tested range. Joint common/private precoder optimization therefore provides a modest robustness benefit under phase-noise mismatch. However, it does not remove the need for accurate inter-AP synchronization.

Fig. 15 shows the relative gain over the optimized SDMA baseline at $P_{\max} = 10$ dBW. The synchronous RSMA gain remains constant at approximately 12%, since the synchronous rates are independent of the oscillator phase-noise variance. Under asynchronous operation, the relative gain increases gradually with σ_ϕ^2, reaching approximately 15–16% at $\sigma_\phi^2 = -20$ dB.

The absolute sum rates of both schemes decrease as the phase-noise variance increases. However, optimized SDMA loses a slightly larger fraction of its rate than the proposed joint RSMA design, causing the relative gain to increase. The heuristic RSMA-MR gain remains close to zero across the tested range, showing that the heuristic common precoder does not provide the same relative advantage.

H. Scalability Validation: Larger Cell-Free Configuration

To verify that the findings are not specific to the default configuration, the power sweep is repeated with $B = 8$ APs, $K = 24$ users, and $M = 2$ antennas per AP. The total number of transmit dimensions therefore increases from

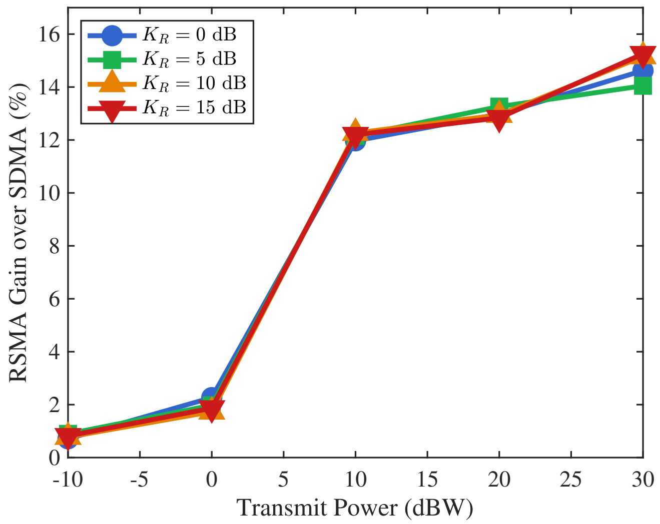


FIGURE 16. RSMA gain over the optimized SDMA baseline versus transmit power for different Rician factors K_R on the RIS–UE link. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

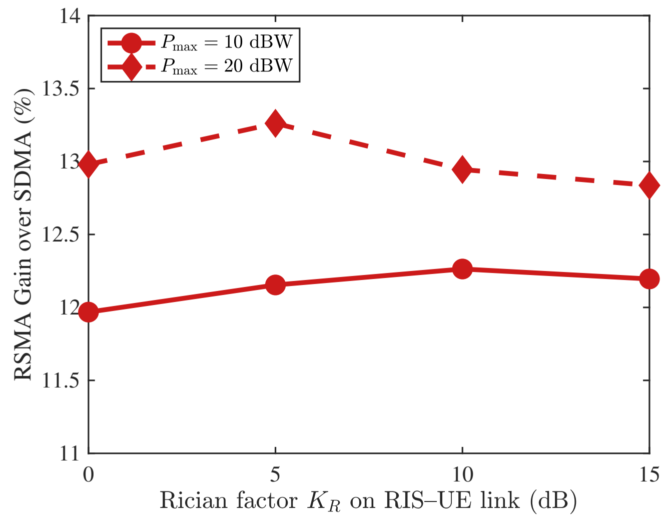


FIGURE 17. Gain of the proposed joint RSMA over the optimized SDMA baseline versus the Rician factor K_R on the RIS–UE link at $P_{\max} = 10$ dBW and 20 dBW. System parameters: $B = 4$, $K = 12$, $M = 2$, $N = 100$.

$BM = 8$ to $BM = 16$. The loading ratio is maintained at $K/BM = 1.5$, while the RIS size remains $N = 100$. Table 5 confirms three observations.

1) The heuristic common-precoder limitation persists at larger scale. The best heuristic RSMA result remains close to the optimized SDMA baseline across the entire power range. At $P_{\max} = 20$ dBW, the corresponding sum rates are 115.8 and 115.6 bps/Hz, respectively. The differences remain below 0.2%, showing that the decoupled common-precoder designs provide negligible gain after increasing the network scale.

2) The proposed joint RSMA gain is preserved at high transmit power. The proposed design achieves gains of 12.0% and 12.6% over optimized SDMA at $P_{\max} = 20$ and 30 dBW, respectively. The gain remains small in the noise-limited regime and increases as multi-user interference becomes more significant. This trend is consistent with the default configuration and confirms the benefit of joint interference management under spatial overload.

3) RZF and NOMA remain noncompetitive in the considered configuration. At high transmit power, RZF and NOMA saturate near 41 and 42 bps/Hz, respectively. Both remain substantially below optimized SDMA and the proposed joint RSMA design because the number of users exceeds the available spatial dimensions.

The larger-scale validation therefore confirms that the main findings persist when the numbers of APs and users are increased.

I. Rician Fading Sensitivity

To assess whether the proposed RSMA gain depends on the RIS–UE small-scale fading model, the Rician factor is swept using (4). Fig. 16 shows the gain of the proposed joint RSMA design over optimized SDMA versus transmit power for different K_R values. The curves overlap closely across the full power range. The gain is small in the noise-limited regime and increases as multi-user interference becomes more significant. At $P_{\max} = 30$ dBW, it remains approximately 14–15% across the tested K_R values.

Fig. 17 presents the same gain as a function of K_R at two fixed transmit-power levels. The proposed joint RSMA design maintains a gain of approximately 12–13% at both $P_{\max} = 10$ dBW and 20 dBW. The Rician model preserves the average RIS–UE channel power, $\mathbb{E}\{\|\mathbf{h}\|^2\}$, across all K_R values, although the instantaneous power $\|\mathbf{h}\|^2$ varies between channel realizations. Thus, varying K_R changes the relative weighting of the LoS and non-line-of-sight (NLoS) components without changing the average link strength. The limited variation across K_R shows that the relative RSMA advantage remains stable under the tested RIS–UE fading structures. This suggests that the gain is governed primarily by joint interference management in the overloaded setting rather than by a particular Rician factor.

J. Sum-Rate Distribution and Fairness

The average sum-rate analysis is complemented by examining the statistical distribution across channel realizations and the inter-user fairness via Jain's fairness index (JFI) [42].

1) Sum-Rate cumulative distribution function (CDF)

Fig. 18 presents the empirical sum-rate CDF at $P_{\max} = 10$ dBW, capturing the variation in network performance across random channel realizations and user placements. Under synchronous conditions, the proposed joint RSMA framework increases the median sum rate from 59.3 to 67.1 bps/Hz and the 10th-percentile sum rate from 57.7 to 64.9 bps/Hz. The near-uniform rightward shift indicates that

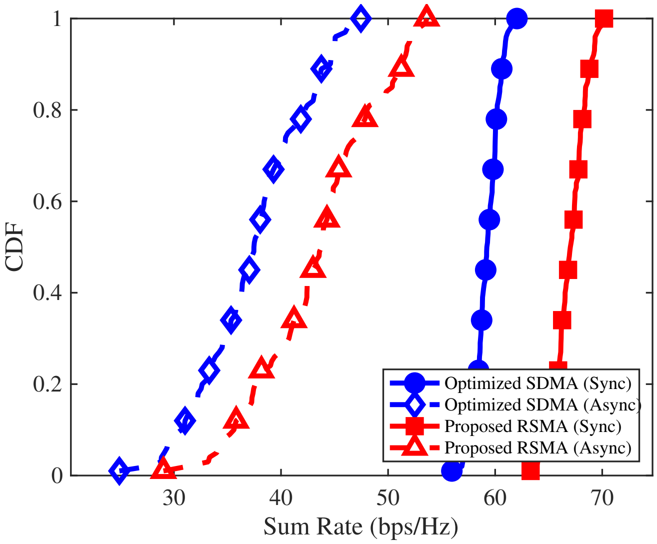


FIGURE 18. **Empirical CDF of the sum rate at $P_{\max} = 10$ dBW under synchronous and asynchronous conditions. System parameters: $B = 4$, $K = 12$, $M = 2$, and $N = 100$.**

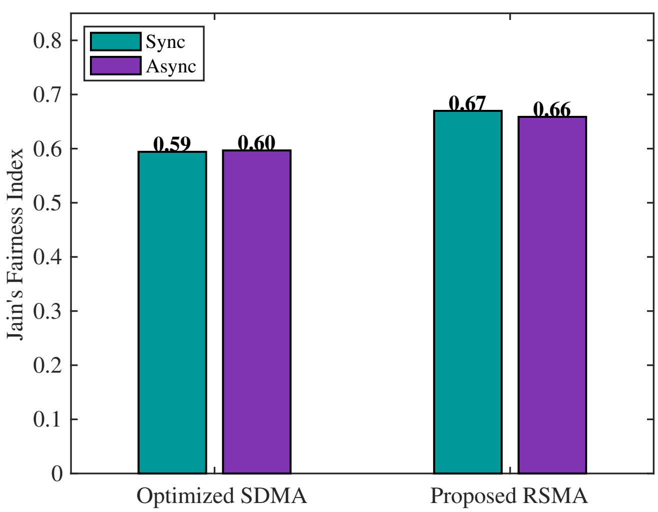


FIGURE 19. **Jain's fairness index at $P_{\max} = 20$ dBW for the optimized SDMA baseline and the proposed joint RSMA framework under synchronous and asynchronous conditions. Values are annotated above each bar. System parameters: $B = 4$, $K = 12$, $M = 2$, and $N = 100$.**

the gain persists across the distribution rather than being driven by a small number of favourable channel realizations.

Under asynchronous conditions, all curves shift toward lower rates, consistent with the phase-noise analysis in Section VI-G. Nevertheless, the proposed joint RSMA framework retains a clear advantage over optimized SDMA. The median sum rate increases from 37.5 to 43.8 bps/Hz, while the 10th-percentile value increases from 31.0 to 35.7 bps/Hz. The similar percentage gains at the median and the 10th percentile indicate that RSMA provides a consistent benefit under both typical and less favourable channel realizations.

2) Jain's Fairness Index

Although the objective is sum-rate maximization rather than direct fairness maximization, the proposed framework achieves a higher Jain's fairness index in the considered configuration. Jain's fairness index [42] is defined as

$$\mathcal{J} = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2} \in \left[\tfrac{1}{K}, 1\right], \tag{47}$$

where $\mathcal{J} = 1$ corresponds to identical rates across all users.

Fig. 19 shows that the proposed joint RSMA framework achieves $\mathcal{J} = 0.67$ under synchronous conditions and $\mathcal{J} = 0.66$ under asynchronous conditions, compared with 0.59 and 0.60, respectively, for optimized SDMA. The higher index indicates a more balanced distribution of user rates under both operating conditions.

The common stream must be decodable by all users, so its design is influenced by the weakest common-stream SINR. This property may contribute to the more balanced rate profile observed for the proposed joint RSMA design. The per-user QoS constraints also prevent individual rates from falling below $R_{\min}$ in both RSMA and SDMA. The observed JFI improvement therefore represents an additional benefit of the proposed design in the considered configuration.

K. Extension to Distributed Multi-RIS Deployments

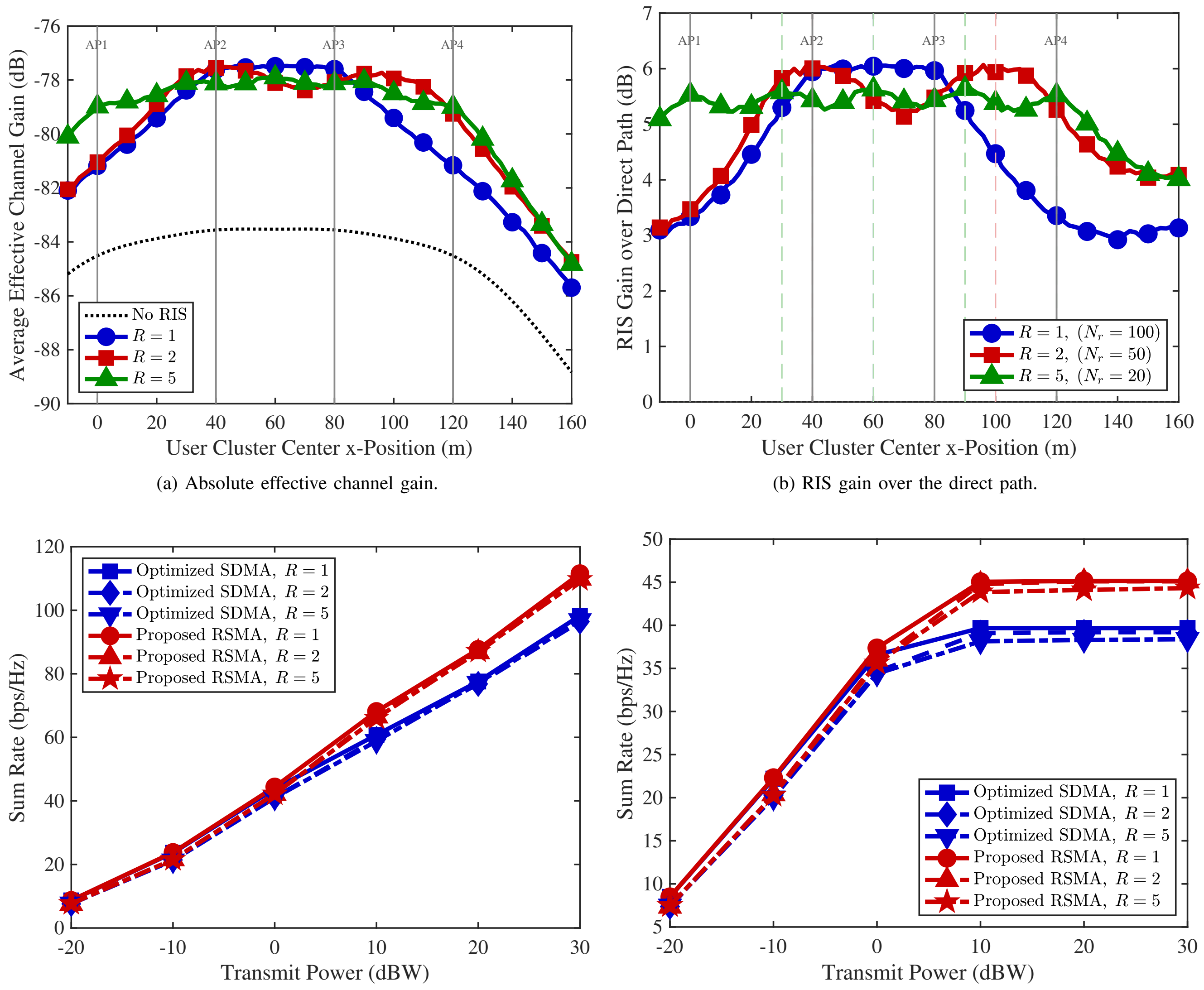
To examine the spatial coverage limitations of a single RIS and extend the proposed framework to distributed deployments, the fixed element budget $N_{\text{total}} = 100$ is divided across $R \in \{1, 2, 5\}$ surfaces. The RIS positions are $x \in \{60\}$ m for $R = 1$, $x \in \{40, 100\}$ m for $R = 2$, and $x \in \{0, 30, 60, 90, 120\}$ m for $R = 5$, with all surfaces placed at $y = 10$ m. Each surface contains $N_r = N_{\text{total}}/R$ elements, while the remaining parameters follow Table 2.

The phase-coefficient vectors of all surfaces are concatenated as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_R^T]^T \in \mathbb{C}^{N_{\text{total}} \times 1}$. Since N_{total} is fixed, the number of scalar phase updates per sweep remains unchanged as R increases.

Fig. 20a shows the absolute effective channel gain. The $R = 1$ and $R = 2$ configurations exhibit greater spatial variation, whereas $R = 5$ maintains a more consistent effective channel gain across the AP span.

Fig. 20b isolates the RIS contribution relative to the direct path. In the $R = 1$ reference deployment, the surface is placed near the centre of the AP coverage span to avoid favouring either boundary. For all configurations, user locations closer to an RIS benefit more strongly from the reflected path, producing local gain peaks near the RIS positions. As R increases, these peaks become more closely spaced. Their greater overlap for $R = 5$ produces a substantially flatter RIS-gain profile.

Across the AP span, the maximum-to-minimum variation in Fig. 20b decreases from 2.74 dB for $R = 1$ to 0.41 dB for $R = 5$. Compared with $R = 1$, the $R = 5$ configuration also increases the RIS gain at $x = 0$ and $x = 120$ m by approximately 2.2 dB. Distributing the RIS elements across



(a) Absolute effective channel gain.

(b) RIS gain over the direct path.

(c) Sum rate under synchronous channels.

(d) Sum rate under asynchronous channels.

FIGURE 20. Impact of distributing $N_{\text{total}} = 100$ RIS elements across $R \in \{1, 2, 5\}$ surfaces: (a) absolute effective channel gain versus user-cluster position; (b) RIS gain over the direct path versus user-cluster position; (c) sum rate versus total transmit power under synchronous channels; and (d) sum rate versus total transmit power under asynchronous channels.

multiple surfaces therefore improves coverage uniformity and the RIS gain near the coverage boundaries, with only a small reduction in the maximum gain.

Figs. 20c and 20d compare the sum-rate performance under synchronous and asynchronous conditions. Across the tested transmit-power range, the maximum-to-minimum variation among $R \in \{1, 2, 5\}$ remains below 3 bps/Hz for both SDMA and RSMA. For the proposed RSMA design, the largest variation is 2.24 bps/Hz under synchronous operation and 2.03 bps/Hz under asynchronous operation.

At $P_{\max} = 30$ dBW, the synchronous RSMA sum rate ranges from 109.63 to 111.52 bps/Hz across the three deployments, while the asynchronous values range from 44.31 to 45.14 bps/Hz. The corresponding asynchronous gains over optimized SDMA are 13.78%, 15.07%, and 15.45% for $R = 1$, $R = 2$, and $R = 5$, respectively.

The limited variation across R suggests that, for the considered deployment, aggregate throughput depends more strongly on the fixed total element budget than on its distribution among the surfaces. Redistributing the elements primarily changes the spatial locations of the reflected-path gains, providing substantially greater coverage uniformity with only minor changes in the network sum rate.

Although distributing the RIS elements improves coverage uniformity, the reflected-path gain still depends on the user positions relative to the surfaces. This motivates considering simultaneous transmitting and reflecting RIS (STAR-RIS) architectures for deployments containing users on both sides of a surface [43], [44]. By supporting simultaneous trans-

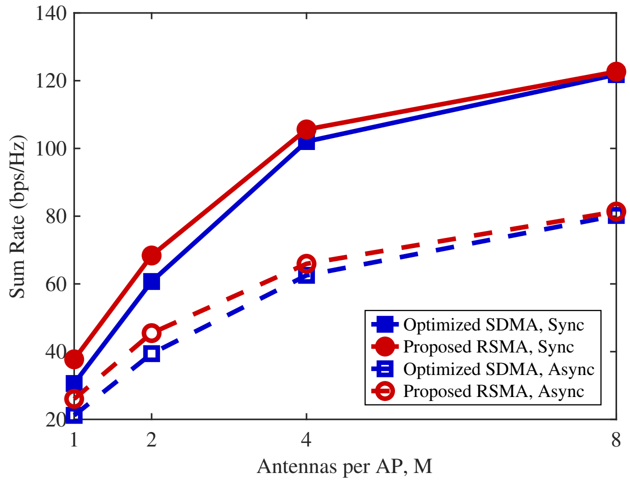


FIGURE 21. **Sum rate versus the number of antennas per AP, M, for synchronous and asynchronous channels at $B = 4$, $K = 12$, $N = 100$, and $P_{\max} = 10$ dBW. The RSMA gain over the optimized SDMA scheme is largest in the spatial-dimension-limited regime ($M = 1$) and narrows as M grows.**

mission and reflection, a STAR-RIS could extend coverage beyond the reflecting side and reduce spatial coverage asymmetry. Extending the proposed framework would require separate transmission and reflection coefficients, together with energy-splitting or mode-switching constraints. This extension is left for future work.

L. Impact of Antenna Configuration

The preceding results fix the number of antennas at each AP at $M = 2$. To assess how the proposed scheme scales with the number of antennas per AP, the sum rate is evaluated for $M \in \{1, 2, 4, 8\}$ at fixed $B = 4$, $K = 12$, $N = 100$, and $P_{\max} = 10$ dBW. The total number of transmit dimensions is BM, so this sweep varies the spatial loading ratio $K/(BM)$ from 3 at $M = 1$ to 0.375 at $M = 8$. Both synchronous and asynchronous operation are considered.

As shown in Fig. 21, the sum rates of both schemes increase with M because the additional antennas provide greater beamforming gain and more spatial dimensions for interference suppression. The relative advantage of RSMA is largest at small M and decreases as the loading ratio $K/(BM)$ falls. At $M = 1$, only $BM = 4$ transmit dimensions are available for $K = 12$ users. The system is therefore heavily overloaded, and the proposed RSMA design achieves its largest gain over optimized SDMA. As M increases, the additional spatial dimensions allow the private precoders to suppress multi-user interference more effectively, reducing the benefit of the common stream. At $M = 8$, where $BM = 32 > K$, the two schemes achieve nearly identical sum rates. The same qualitative trend is observed under asynchronous operation, although the achievable rates are lower. These results show that the RSMA advantage is most pronounced under spatial overload and diminishes as the available transmit dimensions exceed the user load.

The present formulation assumes single-antenna UEs, consistent with the scalar received-signal model, SINR expressions, MMSE equalizers, and SIC operations used throughout the paper. An extension to multi-antenna UEs with one private stream per user would require vector received-signal models and receive-combining vectors at each UE. The scalar SINR, MSE, and equalizer expressions would then be replaced by their vector and matrix counterparts, and the WMMSE reformulation would need to be modified accordingly. Supporting multiple streams per UE would require a further extension of the precoding and SIC structure. These extensions are not considered in the present single-antenna UE formulation and are left for future work.

VII. Conclusion

This paper studied RIS-assisted RSMA for asynchronous downlink cell-free MIMO systems. A WMMSE-based alternating optimization framework was developed to jointly optimize the RSMA common precoder, private precoders, and passive RIS phase shifts. The analysis showed that independent oscillator phase drifts at the APs attenuate the coherent cross-AP combining terms in the power of the common stream. Since the common rate is limited by the weakest user, this attenuation helps explain why decoupled heuristic designs for the common precoder provide little benefit in the considered overloaded CF-MIMO setting.

The numerical results showed that the tested heuristic common precoders provide less than 0.2% gain over optimized SDMA when combined with WMMSE-optimized private precoders. In contrast, the proposed joint RSMA design achieves approximately 12–15% sum-rate gain in the main $B = 4$, $K = 12$ configuration. A similar high-power advantage is retained in the larger $B = 8$, $K = 24$ validation. The results also show that the gain increases with spatial loading and remains relatively stable across the tested RIS sizes, oscillator phase-noise variances, and Rician factors. The empirical sum-rate CDF shows that the improvement persists across channel realizations, while the proposed design achieves a higher Jain's fairness index than optimized SDMA under both synchronous and asynchronous operation.

The multi-RIS results show that distributing a fixed RIS element budget across multiple RIS surfaces improves spatial coverage uniformity while retaining the aggregate phase-vector formulation and element-wise Gauss–Seidel update. The antenna-scaling results further show that the RSMA advantage is largest under spatial overload and diminishes as the number of antennas per AP increases.

From a practical perspective, the proposed WMMSE–AO framework provides a centralized reference design for evaluating the benefit of joint RIS and RSMA optimization. Future work will consider reducing the computational burden through deep unfolding and developing robust designs under imperfect CSI and imperfect SIC. Other directions include multi-antenna UEs, multilayer RSMA, deployments using STAR-RIS, and user-centric AP clustering.

Appendix A
WMMSE Reformulation for Joint RSMA Precoder Design

This appendix derives the WMMSE surrogates (25)-(26), the common-rate reformulation (27c), and the QoS constraint (27d) used in the convex subproblem (27). The classical rate-WMMSE identity [29] is adapted to the cell-free RSMA setting with aggregate channels $\mathbf{h}_k$ and aggregate precoders $\mathbf{w}_j$ defined in (20)-(21). The derivation in this appendix applies to the synchronous design channel used in the optimization, corresponding to $P = 1$. The general asynchronous rate expressions in (16)–(18) are used for performance evaluation after the precoders and RIS phases have been designed.

A. Rate-WMMSE Equivalence

Let $\hat{s}_{p,k} = g_{p,k} y_k$ denote the estimated private symbol using a scalar equalizer $g_{p,k} \in \mathbb{C}$. Assuming uncorrelated noise and unit-power data symbols ($\mathbb{E}[|s_k|^2] = 1$), the MSE $\varepsilon_{p,k} = \mathbb{E}[|\hat{s}_{p,k} - s_k|^2]$ expands to

$$\varepsilon_{p,k} = |1 - g_{p,k}\mathbf{h}_k^H \mathbf{w}_k|^2 + |g_{p,k}|^2 \left(\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \right). \tag{48}$$

Proposition 2 (Rate-WMMSE Equivalence)**.** *For any stream $i \in \{c, p\}$ with SINR $\gamma_{i,k}$, the achievable Shannon rate is mathematically equivalent to the maximum of a concave WMMSE surrogate function.*

$$\max_{g_{i,k},\, \alpha_{i,k} > 0} \left(\log_2 \alpha_{i,k} - \frac{\alpha_{i,k}\varepsilon_{i,k}}{\ln 2} + \frac{1}{\ln 2} \right) = \log_2(1 + \gamma_{i,k}). \tag{49}$$

This equivalence holds when the receiver equalizer $g_{i,k}$ and the penalty weight $\alpha_{i,k}$ are optimally chosen. For the private and common streams respectively, these optimal parameters are

$$g_{p,k}^\star = \frac{\mathbf{h}_k^H \mathbf{w}_k}{\sum_{j=1}^K |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2}, \quad g_{c,k}^\star = \frac{\mathbf{h}_k^H \mathbf{w}_0}{\sum_{j=0}^K |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2}, \tag{50}$$

$$\alpha_{p,k}^\star = (\varepsilon_{p,k}^\star)^{-1} = 1 + \gamma_{p,k}, \quad \alpha_{c,k}^\star = (\varepsilon_{c,k}^\star)^{-1} = 1 + \gamma_{c,k}. \tag{51}$$

Proof:
Taking the Wirtinger derivative of $\varepsilon_{i,k}$ with respect to $g_{i,k}^*$ and setting it to zero yields the MMSE equalizers in (50). These equalizers result in the minimum MSE $\varepsilon_{i,k}^\star = (1 + \gamma_{i,k})^{-1}$. Setting the derivative of the surrogate in (49) with respect to $\alpha_{i,k}$ to zero yields the optimal weights $\alpha_{i,k}^\star = 1/\varepsilon_{i,k}^\star$. Evaluating the surrogate at these optimal parameters recovers the Shannon rate. ■

Physically, $g_{c,k}^\star$ operates pre-SIC and faces all interference ($j \geq 0$), whereas $g_{p,k}^\star$ operates post-SIC with the common stream removed ($j \geq 1$).

B. Construction of the Concave Surrogates

Proposition 2 enables the construction of a tight global lower bound on the Shannon rate. For any equalizer $g_{i,k}$ and weight $\alpha_{i,k} > 0$, with $i \in \{c, p\}$, the surrogate in (49) satisfies

$$\log_2(1 + \gamma_{i,k}) \geq \log_2 \alpha_{i,k} + \frac{1 - \alpha_{i,k}\varepsilon_{i,k}}{\ln 2}, \tag{52}$$

with equality when $g_{i,k} = g_{i,k}^\star$ and $\alpha_{i,k} = \alpha_{i,k}^\star = (\varepsilon_{i,k}^\star)^{-1}$.

To reformulate the optimization into a concave quadratic form suitable for convex solvers, the composite variable $\beta_{i,k} \triangleq \sqrt{\alpha_{i,k}}\, g_{i,k}^*$ is introduced. Substituting $\beta_{i,k}$ into the MSE expression (48) and isolating the terms dependent on the precoding matrix $\mathbf{W}$ yields

$$\log_2(1 + \gamma_{i,k}) \geq r_{i,k}^{(0)} + \frac{f_{i,k}(\mathbf{W})}{\ln 2}. \tag{53}$$

Here, $f_{i,k}(\mathbf{W})$ contains the linear and quadratic precoder terms defined in (25)-(26), and $r_{i,k}^{(0)}$ collects the terms that are constant with respect to $\mathbf{W}$ in the WMMSE surrogate for fixed $g_{i,k}$ and $\alpha_{i,k}$.

$$r_{i,k}^{(0)} = \log_2 \alpha_{i,k} + \frac{1 - \alpha_{i,k}}{\ln 2}. \tag{54}$$

The factor $1/\ln 2$ in (53) converts the natural-log WMMSE identity to base-2 spectral efficiency, matching the rate definition in (18).

C. Common-Rate Reformulation

For the synchronous design-channel case ($P = 1$), (16) reduces to $R_c = \min_k \log_2(1 + \gamma_{c,k})$. This minimum-rate expression is nonsmooth and is equivalently reformulated by introducing the auxiliary variable $c_{\min}$ and imposing the following per-user constraints:

$$c_{\min} \leq \log_2(1 + \gamma_{c,k}), \quad \forall k. \tag{55}$$

Substituting the concave common-stream surrogate (53) into the right-hand side of (55) yields

$$c_{\min} \leq r_{c,k}^{(0)} + \frac{f_{c,k}(\mathbf{W})}{\ln 2}, \quad \forall k, \tag{56}$$

which corresponds to the optimization constraint (27c).

D. QoS Constraint Reformulation

To convexify the QoS requirement (19e), the common-rate shares c_k defined in (27e) are substituted for R_c, and the concave surrogate (53) is applied to $R_{p,k}$, yielding

$$c_k + r_{p,k}^{(0)} + \frac{f_{p,k}(\mathbf{W})}{\ln 2} \geq R_{\min}, \quad \forall k. \tag{57}$$

Substituting (54) into (57) and scaling by $\ln 2$ recovers the linear form in (27d).

$$f_{p,k}(\mathbf{W}) + (\ln 2)c_k \geq \eta_k, \quad \forall k, \tag{58}$$

where $\eta_k = (\ln 2)R_{\min} - \ln \alpha_{p,k} + \alpha_{p,k} - 1$.

Appendix B
RIS QCQP Parameter Derivation

This appendix derives the objective parameters $\mathbf{Q}$ and $\mathbf{q}$ for the RIS optimization subproblem (29). For fixed precoders, the effective channel from (28) decomposes as $\mathbf{h}_k^H = \mathbf{d}_k^H + \boldsymbol{\theta}^H \mathbf{V}_k$, where $\mathbf{V}_k = \mathrm{diag}(\mathbf{F}_{r,k}^*)\,\mathbf{G} \in \mathbb{C}^{N \times BM}$ is the cascaded RIS channel, with $\mathbf{G} = [\mathbf{G}_{1,r}, \ldots, \mathbf{G}_{B,r}]$ as defined in (28).

To isolate the phase shift vector $\boldsymbol{\theta}$, the precoders are projected onto the cascaded and direct channels.

$$\mathbf{a}_{j,k} = \mathbf{V}_k \mathbf{w}_j \in \mathbb{C}^N, \quad b_{j,k} = \mathbf{d}_k^H \mathbf{w}_j \in \mathbb{C}. \tag{59}$$

Substituting this decomposition into the WMMSE surrogate (25), the received power for stream j at user k expands to

$$|\mathbf{h}_k^H \mathbf{w}_j|^2 = \boldsymbol{\theta}^H \mathbf{a}_{j,k} \mathbf{a}_{j,k}^H \boldsymbol{\theta} + 2\,\mathrm{Re}\{\boldsymbol{\theta}^H \mathbf{a}_{j,k} b_{j,k}^*\} + |b_{j,k}|^2. \tag{60}$$

Collecting the quadratic and linear terms in $\boldsymbol{\theta}$ across all users and streams, and using the WMMSE composite variable $\beta_{p,k} \triangleq \sqrt{\alpha_{p,k}}\, g_{p,k}^*$ from Appendix A, the SDMA parameters are

$$\mathbf{Q} = \sum_{k=1}^{K} |\beta_{p,k}|^2 \left(\mathbf{a}_{k,k}\mathbf{a}_{k,k}^H + \sum_{j \neq k} \mathbf{a}_{j,k}\mathbf{a}_{j,k}^H \right), \tag{61}$$

$$\mathbf{q} = \sum_{k=1}^{K} \left[\sqrt{\alpha_{p,k}}\, \beta_{p,k}^*\, \mathbf{a}_{k,k} - |\beta_{p,k}|^2 \sum_{j=1}^{K} \mathbf{a}_{j,k} b_{j,k}^* \right]. \tag{62}$$

For RSMA, each user performs two-stage decoding (pre-SIC for the common stream and post-SIC for the private stream), requiring two distinct WMMSE composite variables per user: $\beta_{c,k} \triangleq \sqrt{\alpha_{c,k}}\, g_{c,k}^*$ and $\beta_{p,k} \triangleq \sqrt{\alpha_{p,k}}\, g_{p,k}^*$. Summing the common-stream and private-stream contributions yields

$$\mathbf{Q} = \sum_{k=1}^{K} \left[|\beta_{c,k}|^2 \sum_{j=0}^{K} \mathbf{a}_{j,k}\mathbf{a}_{j,k}^H + |\beta_{p,k}|^2 \sum_{j=1}^{K} \mathbf{a}_{j,k}\mathbf{a}_{j,k}^H \right], \tag{63}$$

$$\begin{aligned} \mathbf{q} = &\sum_{k=1}^{K} \left[\sqrt{\alpha_{c,k}}\, \beta_{c,k}^*\, \mathbf{a}_{0,k} - |\beta_{c,k}|^2 \sum_{j=0}^{K} \mathbf{a}_{j,k} b_{j,k}^* \right] \\ &+ \sum_{k=1}^{K} \left[\sqrt{\alpha_{p,k}}\, \beta_{p,k}^*\, \mathbf{a}_{k,k} - |\beta_{p,k}|^2 \sum_{j=1}^{K} \mathbf{a}_{j,k} b_{j,k}^* \right]. \end{aligned} \tag{64}$$

Appendix C
Monotonicity and Objective-Value Convergence of the Gauss–Seidel Update

This appendix establishes that the element-wise Gauss–Seidel update (32) monotonically non-increases the RIS-subproblem objective (29), which guarantees convergence of the corresponding objective-value sequence.

Proposition 3 (Monotonicity of Coordinate Descent)**.** *For the objective $J(\boldsymbol{\theta}) = \boldsymbol{\theta}^H \mathbf{Q}\,\boldsymbol{\theta} - 2\,\mathrm{Re}\{\boldsymbol{\theta}^H \mathbf{q}\}$, each coordinate update $\theta_n \leftarrow e^{j\angle(c_n)}$ guarantees $J(\boldsymbol{\theta}^{\text{new}}) \leq J(\boldsymbol{\theta}^{\text{old}})$.*

Proof:

1) Coordinate Isolation: Fixing all other phase shifts $\{\theta_m\}_{m \neq n}$, $J(\boldsymbol{\theta})$ reduces to a function of θ_n alone. Since $|\theta_n| = 1$, the diagonal term $Q_{n,n}|\theta_n|^2 = Q_{n,n}$ is constant, simplifying the objective to

$$J(\theta_n) = -2\,\mathrm{Re}\{c_n^* \theta_n\} + \text{const.}, \tag{65}$$

where $c_n = q_n - \sum_{m \neq n} Q_{n,m}\theta_m$.

2) Optimal Phase Update: Minimizing (65) subject to $|\theta_n| = 1$ requires aligning the phase of θ_n with c_n to maximize the real part, yielding the optimal update $\theta_n^{\text{new}} = e^{j\angle(c_n)}$.

3) Monotonic Decrease: Evaluating the objective at this newly aligned phase yields $J(\theta_n^{\text{new}}) = -2|c_n| + \text{const.}$ Since $\mathrm{Re}\{z\} \leq |z|$ for any complex number, $-|c_n| \leq -\mathrm{Re}\{c_n^* \theta_n^{\text{old}}\}$, which gives

$$J(\theta_n^{\text{new}}) \leq -2\,\mathrm{Re}\{c_n^* \theta_n^{\text{old}}\} + \text{const.} = J(\theta_n^{\text{old}}). \tag{66}$$

Sequentially updating all N elements therefore produces a monotonically non-increasing sequence of RIS-subproblem objective values. Since $J(\boldsymbol{\theta})$ is continuous on the compact unit-modulus feasible set, it is bounded below. Hence, the sequence of objective values converges to a finite limit. ■

REFERENCES

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.
[2] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2020.
[3] J. Zheng, Z. Zhao, J. Zhang, J. Cheng, and V. C. M. Leung, "Performance analysis of cell-free massive MIMO systems with asynchronous reception," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 190–195.
[4] Y. Fang, L. Qiu, X. Liang, and C. Ren, "Cell-free massive MIMO systems with oscillator phase noise: Performance analysis and power control," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 048–10 064, 2021.
[5] J. Li, M. Liu, P. Zhu, D. Wang, and X. You, "Impacts of asynchronous reception on cell-free distributed massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 11 106–11 110, 2021.
[6] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.
[7] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 990–1002, 2020.
[8] S.-N. Jin, D.-W. Yue, and H. H. Nguyen, "RIS-aided cell-free massive MIMO system: Joint design of transmit beamforming and phase shifts," *IEEE Systems Journal*, vol. 17, no. 2, pp. 3093–3104, 2023.
[9] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5218–5233, 2020.
[10] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, and B. Ottersten, "Reconfigurable intelligent surface-assisted cell-free massive MIMO systems over spatially correlated channels," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5106–5128, 2022.
[11] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

[12] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 98–105, 2016.

[13] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, 2018, art. no. 133.

[14] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access for downlink multi-antenna communications: Physical-layer design and link-level simulations," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2020, pp. 1–6.

[15] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4847–4861, 2016.

[16] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for 6G—Part I: Principles, applications, and future works," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2232–2236, 2022.

[17] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[18] A. Mishra, Y. Mao, L. Sanguinetti, and B. Clerckx, "Rate-splitting assisted massive machine-type communications in cell-free massive MIMO," *IEEE Communications Letters*, vol. 26, no. 6, pp. 1358–1362, 2022.

[19] Y. Zhu, E. Shi, Z. Liu, J. Zhang, and B. Ai, "Multi-agent reinforcement learning-based joint precoding and phase shift optimization for RIS-aided cell-free massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 9, pp. 14 015–14 020, 2024.

[20] A. Bansal, K. Singh, B. Clerckx, C.-P. Li, and M.-S. Alouini, "Rate-splitting multiple access for intelligent reflecting surface-aided multi-user communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9217–9229, 2021.

[21] M. Soleymani, I. Santamaria, E. A. Jorswieck, and B. Clerckx, "Optimization of rate-splitting multiple access in beyond diagonal RIS-assisted URLLC systems," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 5063–5078, 2024.

[22] M. Soleymani, I. Santamaria, E. A. Jorswieck, M. Di Renzo, R. Schober, and L. Hanzo, "Rate splitting multiple access for RIS-aided URLLC MIMO broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 25, no. 3, pp. 1529–1546, 2026.

[23] J. Zheng, J. Zhang, J. Cheng, V. C. M. Leung, D. W. K. Ng, and B. Ai, "Asynchronous cell-free massive MIMO with rate-splitting," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1366–1382, 2023.

[24] J. Zheng, J. Zhang, H. Du, R. Zhang, D. Niyato, O. A. Dobre, and B. Ai, "Rate-splitting for cell-free massive MIMO: Performance analysis and generative AI approach," *IEEE Transactions on Communications*, vol. 73, no. 4, pp. 2431–2447, 2025.

[25] Y. Zhang, H. Zhao, Y. Mao, W. Xia, W. Lu, and H. Zhu, "Rate-splitting multiple access in cell-free massive MIMO-URLLC systems: Achievable rate analysis and optimization," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 6752–6767, 2024.

[26] Y. Zhang, W. Xia, H. Zhao, Y. Mao, J. Zhang, and G. Zheng, "Enhancing uplink performance for cell-free massive MIMO with low-resolution ADCs by RSMA," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 3, pp. 720–735, 2025.

[27] M. Polese, M. Dohler, F. Dressler, M. Erol-Kantarci, R. Jana, R. Knopp, and T. Melodia, "Empowering the 6G cellular architecture with Open RAN," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 245–262, 2024.

[28] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 349–353, 2020.

[29] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[31] CVX Research, Inc., "CVX: MATLAB software for disciplined convex programming, version 2.2," [Online]. Available: https://cvxr.com/cvx, Jan. 2020.

[32] J. Dai, J. Ge, K. Zhi, C. Pan, Z. Zhang, J. Wang, and X. You, "Two-timescale transmission design for RIS-aided cell-free massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 23, no. 6, pp. 6498–6517, 2024.

[33] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[34] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[35] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1394–1410, 2021.

[36] A. Chowdhury, G. Verma, A. Swami, and S. Segarra, "Deep graph unfolding for beamforming in MU-MIMO interference networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4889–4903, 2024.

[37] M. Wu, Z. Gao, Y. Huang, Z. Xiao, D. W. K. Ng, and Z. Zhang, "Deep learning-based rate-splitting multiple access for reconfigurable intelligent surface-aided Tera-Hertz massive MIMO," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1431–1451, 2023.

[38] Ö. T. Demir, M. Masoudi, E. Björnson, and C. Cavdar, "Cell-free massive MIMO in O-RAN: Energy-aware joint orchestration of cloud, fronthaul, and radio resources," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 356–372, 2024.

[39] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part I: Channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, 2005.

[40] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1711–1721, 2007.

[41] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, 2006.

[42] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Digital Equipment Corporation, Hudson, MA, USA, DEC Research Report TR-301, Sep. 1984.

[43] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Communications Letters*, vol. 25, no. 9, pp. 3134–3138, 2021.

[44] W. Khalid, Z. Kaleem, R. Ullah, T. Van Chien, S. Noh, and H. Yu, "Simultaneous transmitting and reflecting-reconfigurable intelligent surface in 6G: Design guidelines and future perspectives," *IEEE Network*, vol. 37, no. 5, pp. 173–181, 2023.