Low-Complexity Beamforming and Rate Allocation for RSMA

Diluka Galappaththige, *Member, IEEE*, and Chintha Tellambura, *Fellow, IEEE*,

***Abstract*—This paper develops a low-complexity joint beamforming and rate allocation design for rate-splitting multiple access (RSMA) systems. Specifically, a multi-antenna base station (BS) serves multiple users to maximize the sum rate subject to common rate and transmit power constraints. The resulting problem is highly non-convex due to the coupling between beamforming and rate variables. Conventional approaches based on convex-concave procedure algorithm (CCPA) and fractional programming (FP) incur high computational complexity and scalability limitations. We propose an alternating optimization framework based on manifold optimization and the augmented Lagrangian method to solve the problem efficiently without semidefinite lifting, rank relaxation, or approximations. Numerical results show that, with 16 BS antennas, the proposed method is approximately $10\times$ and $8\times$ faster than CCPA and FP benchmarks, respectively, while achieving $5.2\,\%$ and $3.4\,\%$ sum-rate gains and higher constraint satisfaction probability.**

***Index Terms*—Rate-splitting multiple access, transmit beamforming, rate allocation, manifold algorithm.**

I. Introduction

Rate-splitting multiple access (RSMA) mitigates interference by partitioning each user's message into common and private components [1], [2]. Successive interference cancellation (SIC) decoding methods are applied, enabling partial interference decoding and partial noise treatment. This yields higher spectral and energy efficiency than space-division multiple access (SDMA) and non-orthogonal multiple access (NOMA) [1], [2]. However, these gains depend on the joint optimization of beamforming and common-rate allocation.

Existing solutions use successive convex approximation (SCA), semidefinite relaxation (SDR), and fractional programming (FP) [3]–[7]. For instance, [3] proposes an SDR-SCA-based beamforming design with dynamic common-stream decoding, while [4] employs FP and Karush-Kuhn-Tucker (KKT) conditions for joint beamforming and rate allocation. Reconfigurable intelligent surfaces (RIS)- and simultaneously transmitting and reflecting (STAR)-RIS-assisted RSMA designs in [5], [6] adopt SCA-based alternating optimization (AO), and [7] develops a convex-concave procedure algorithm (CCPA) via SDR-SCA for max-min fairness.

However, SDR lifts the problem into a high-dimensional semidefinite space, incurring substantial computational overhead, while the relaxed rank-one constraints require Gaussian randomization (GR), often leading to performance loss [8]. SCA methods rely on successive local approximations, yielding iterative, computationally intensive procedures that are sensitive to initialization. FP approaches introduce auxiliary variables and multi-block updates, thereby slowing convergence and limiting the effective enforcement of constraints.

As a result, these methods often incur high computational complexity, limited scalability with increasing numbers of antennas and users, and degraded feasibility due to relaxation and approximation errors, which can reduce the constraint satisfaction probability (CSP) (Section VI) [9], [10]. This motivates the development of more efficient and scalable solutions. In this work, we show that the beamforming variables lie on a complex sphere manifold (CSM), a structure that is not exploited in existing RSMA approaches [3]–[7]. Leveraging this property significantly reduces the search-space dimension while preserving the original problem structure. However, incorporating the CSM into RSMA optimization is nontrivial due to the presence of coupled non-convex constraints and the need to maintain feasibility throughout the iterative updates.

D. Galappaththige and C. Tellambura are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G 1H9, Canada (e-mail: {diluka.lg, ct4}@ualberta.ca).

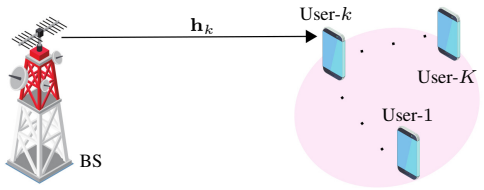

Fig. 1: An RSMA-assisted communication system.

Motivated by these limitations and building on our previous work [9], we propose a low-complexity joint RSMA beamforming and rate allocation problem for a multi-antenna base station (BS) serving multiple users, i.e., Fast-RSMA (FRSMA). Unlike existing SDR/SCA- and FP-based methods, FRSMA directly searches on the CSM while incorporating the constraints through the augmented Lagrangian method (ALM) [10], [11]. Therefore, this method avoids the limitations of the existing solutions. This problem is non-convex and is divided into a beamforming and a common-rate allocation subproblem. The former is solved using an ALM-based manifold optimization (MO) approach over a CSM. The proposed method iteratively updates the optimization variables, Lagrange multipliers, and penalty parameters to ensure constraint satisfaction. The common rate allocation subproblem is efficiently solved via convex optimization.

Simulation results demonstrate a favorable performance-complexity tradeoff. For example, with 16 BS antennas, runtime is reduced by $10\times$ and $8\times$ compared to CCPA and FP benchmarks, respectively, while improving the sum rate by $5.2\,\%$ and $3.4\,\%$, and consistently achieving higher CSP. Thus, improved RSMA performance is achieved with significantly lower computational cost, while maintaining scalability with respect to the number of antennas and users.

Notation: $\mathbf{I}_M$ is the $M \times M$ identity matrix. $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$ is a complex Gaussian vector with mean $\boldsymbol{\mu}$ and co-variance $\mathbf{R}$. $\mathbf{1}_{\{x\}}$ is 1 if $x > 0$ and 0 otherwise.

II. Preliminaries

A. System, Channel, and Signal Models

We consider an RSMA BS equipped with M antennas serving K single-antenna users (Fig. 1) [1], [12]. With time-division duplexing for both channel estimation and data trans-

mission, channel state information (CSI) can be estimated using orthogonal pilots [12], which is highly accurate. Thus, we assume perfect CSI availability.

During each fading block, the channel between the BS and the k-th user, i.e., $\mathbf{h}_k$, is defined as $\mathbf{h}_k = \zeta_k^{1/2}\tilde{\mathbf{h}}_k$, where ζ_k denotes the large-scale fading coefficient denoting pathloss and shadowing, while $\tilde{\mathbf{h}}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ represents the small-scale Rayleigh fading [12]. Small-scale fading varies across coherence intervals and requires frequent estimation, whereas large-scale fading changes slowly and remains approximately constant over many intervals [13], and thus does not require re-estimation at every coherence block.

The BS uses RSMA to send data to the users [1], [12]. In particular, the message q_k intended for the k-th user is split into a common part $q_{c,k}$ and a private part $q_{p,k}$, i.e., $q_k = \{q_{c,k}, q_{p,k}\}$ for $k \in \{1, \ldots, K\}$. The common parts of all users, i.e., $\{q_{c,1}, \ldots, q_{c,K}\}$ are jointly encoded into the common data stream q_c while the private parts, i.e., $\{q_{p,1}, \ldots, q_{p,K}\}$ are encoded into private data streams $\{q_1, \ldots, q_K\}$ [1], [12]. The BS then linearly precodes the common and private data streams using the precoders/beamforming $\mathbf{w}_c, \mathbf{w}_k \in \mathbb{C}^{M\times 1}$ for $k \in \{1, \ldots, K\}$, respectively [1], [12]. The BS transmitted signal can be given as $\mathbf{x} = \mathbf{w}_c q_c + \sum_{i=1}^{K} \mathbf{w}_i q_i$, where q_c and $\{q_i\}_{i=\{1,\ldots,K\}}$ are mutually independent. The received signal at the k-th user is given by $y_k = \mathbf{h}_k^{\mathrm{H}}\mathbf{w}_c q_c + \sum_{i=1}^{K} \mathbf{h}_k^{\mathrm{H}}\mathbf{w}_i q_i + n_k$, where $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at the k-th user, with 0 mean and σ^2 variance.

III. Communication Performance

1) Common Rate: The users first decode the common message by treating private signals as interference. To this end, the k-th user rate for decoding common data is given as $\mathcal{R}_{c,k} = \log_2(1 + \gamma_{c,k})$, where $\gamma_{c,k}$ is the received signal-to-interference-plus-noise ratio (SINR) at the k-th user for decoding q_c, and given by

$$\gamma_{c,k} = \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_c|^2}{\sum_{i=1}^{K} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_i|^2 + \sigma^2}. \tag{1}$$

To ensure that all users successfully decode the common data stream, the rate of decoding the common stream q_c should not exceed $\mathcal{R}_c = \min_{k\in\{1,\ldots,K\}} \mathcal{R}_{c,k}$ [2]. As $\mathcal{R}_c$ is shared by K users, it follows that $\sum_{k=1}^{K} C_k = \mathcal{R}_c$, where C_k is the portion of the common rate allocated for the k-th user message $q_{c,k}$.

2) Private Rate: Once q_c is decoded, the k-th user subtracts it from the received signal prior to decoding its intended private message q_k. We assume that perfect CSI is available at the user. Under this assumption, the common signal, $\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_c q_c$, can be perfectly removed from the received signal. We are adopting this simplification to facilitate analysis and to focus on the core optimization aspects of the proposed system. To this end, the private rate of the k-th user is given by $\mathcal{R}_{p,k} = \log_2(1+\gamma_{p,k})$, where $\gamma_{p,k}$ is the received SINR at the k-th user for decoding q_k and given by

$$\gamma_{p,k} = \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k|^2}{\sum_{i\neq k}^{K} |\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_i|^2 + \sigma^2}. \tag{2}$$

3) Total Rate: The total achievable rate of the k-th user, including the portion of common rate transmitting $q_{c,k}$ and private rate transmitting $q_{p,k}$, is given by $\mathcal{R}_k = C_k + \mathcal{R}_{p,k}$.

IV. Problem Formulation

Here, the joint beamforming and common-rate optimization problem is formulated. The goal is to maximize the users' communication sum rate, subject to constraints on the users' common rates and the maximum allowable BS transmit power. The optimization problem is:

$$\mathbf{P}: \max_{\mathbf{w}_c, \{\mathbf{w}_k\}_{k=1}^K, \{C_k\}_{k=1}^K} \sum_{k=1}^{K} C_k + \log_2(1+\gamma_{p,k}), \tag{3a}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} C_i \leq \mathcal{R}_{c,k}, \ \forall k, \tag{3b}$$

$$C_k \geq 0, \ \forall k, \tag{3c}$$

$$\|\mathbf{w}_c\|^2 + \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 \leq p_{\max}, \tag{3d}$$

where conditions (3b) and (3c) guarantee the successful decoding of the common stream, and (3d) sets the BS transmit power to be less than $p_{\max}$. Problem (3) captures the essential coupling between beamforming and rate allocation variables, which constitutes the core challenge in RSMA design [1].

While prior works consider different objectives and system models, such as weighted sum-rate maximization [4], dynamic decoding strategies [3], RIS/STAR-RIS-assisted systems [5], [6], and max-min fairness [7], they all involve solving joint beamforming and rate allocation problems under non-convex constraints. This underscores the importance of developing a low-complexity and scalable solution to the baseline problem in (3), which can serve as a unifying framework for more advanced settings. In particular, the proposed FRSMA framework is readily extendable to alternative objectives (e.g., weighted sum-rate or fairness), additional system components (e.g., RIS/STAR-RIS), and practical considerations (e.g., imperfect CSI) by suitably modifying the objective and constraint set while preserving the overall solution structure.

V. Proposed Solution

Problem $\mathbf{P}$ is non-convex due to the coupling among the variables. To address this, we adopt the AO technique [14]. Thus, $\mathbf{P}$ is decomposed into two sub-problems: (i) optimization of the BS beamformers $\mathbf{w}_c$ and $\{\mathbf{w}_k\}_{k=1}^K$ with fixed common rates $\{C_k\}_{k=1}^K$, and (ii) optimization of the common rates $\{C_k\}_{k=1}^K$ with fixed BS beamformers $\mathbf{w}_c$ and $\{\mathbf{w}_k\}_{k=1}^K$.

A. Sub-Problem 1: Optimization Over $\mathbf{w}_c$ and $\{\mathbf{w}_k\}_{k=1}^K$

With fixed $\{C_k\}_{k=1}^K$, $\mathbf{P}$ becomes a beamforming problem:

$$\mathbf{P}_w: \max_{\mathbf{w}_c, \{\mathbf{w}_k\}_{k=1}^K} \sum_{k=1}^{K} \log_2(1+\gamma_{p,k}), \tag{4a}$$

$$\text{s.t.} \quad \gamma_{c,k} \geq \Gamma_c^{\mathrm{th}}, \ \forall k, \tag{4b}$$

$$\|\mathbf{w}_c\|^2 + \sum_{i=1}^{K} \|\mathbf{w}_i\|^2 \leq p_{\max}, \tag{4c}$$

where $\Gamma_c^{\mathrm{th}} \triangleq 2^{\sum_{i=1}^{K} C_k} - 1$. Next, we introduce several mathematical notations. The beamforming vectors are organized into a single matrix $\mathbf{W} = [\mathbf{w}_c, \mathbf{w}_1 \ldots, \mathbf{w}_K] \in \mathbb{C}^{M\times(K+1)}$. An

index matrix, $\mathbf{E} = \mathbf{I}_{K+1} \in \mathbb{R}^{(K+1)\times(K+1)}$, is defined. The combination of $\mathbf{W}$ and $\mathbf{E}$ can represent each column individually from $\mathbf{W}$, i.e., $\mathbf{w}_c = \mathbf{W}\mathbf{E}_1$ and $\mathbf{w}_k = \mathbf{W}\mathbf{E}_{k+1}$, where $\mathbf{E}_k$ is the k-th column of $\mathbf{E}$. To this end, the beamforming vectors in the SINR are replaced with $\mathbf{W}$ and $\mathbf{E}$ [9].

Then, we use MO and FP to address $\mathbf{P}_w$. To handle the sum-log challenge in $\mathbf{P}_w$, the FP is utilized to substitute auxiliary variables, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$, for each SINR term in (4a) that ensures $\mu_k \leq \gamma_{p,k}$ [15]. Then, $\mathbf{P}_w$ is reformulated as [15]

$$\mathbf{P}_{w1}: \max_{\mathbf{W},\boldsymbol{\mu}} \; f(\mathbf{W},\boldsymbol{\mu}) = \frac{1}{\ln(2)}\sum\nolimits_{k=1}^{K}\ln(1+\mu_k) + \frac{1}{\ln(2)}\sum\nolimits_{k=1}^{K}\left(-\mu_k + \frac{(1+\mu_k)\gamma_{p,k}}{1+\gamma_{p,k}}\right), \quad (5a)$$

$$\text{s.t.} \quad (4b)-(4c). \quad (5b)$$

$\mathbf{P}_{w1}$ is a two-part optimization problem: (i) an outer optimization over $\mathbf{W}$ with fixed $\boldsymbol{\mu}$ and (ii) an inner optimization over $\boldsymbol{\mu}$ with fixed $\mathbf{W}$ [15]. Then, $\mathbf{P}_{w1}$ is solved by alternatively optimizing $\mathbf{W}$ and $\boldsymbol{\mu}$ until the objective function converges.

1) Optimizing $\boldsymbol{\mu}$: With fixed $\mathbf{W}$, $f(\mathbf{W},\boldsymbol{\mu})$ is a concave differentiable function over $\boldsymbol{\mu}$. Thus, the optimal $\boldsymbol{\mu}$ is obtained as $\frac{\partial f(\mathbf{W},\boldsymbol{\mu})}{\partial \mu_k} = 0$, i.e., the optimal $\mu_k^* = \gamma_{p,k}$ [9].

2) Optimizing $\mathbf{W}$: With fixed $\boldsymbol{\mu}$, the objective (5a) is simplified, eliminating the constant terms with respect to $\mathbf{W}$. Then, $\mathbf{P}_{w1}$ is reformulated as

$$\mathbf{P}_{w2}: \max_{\mathbf{W}} \sum_{k=1}^{K}\frac{\hat{\mu}_k|\mathbf{h}_k^{\mathrm{H}}\mathbf{W}\mathbf{E}_{k+1}|^2}{\sum_{i=1}^{K}|\mathbf{h}_k^{\mathrm{H}}\mathbf{W}\mathbf{E}_{i+1}|^2+\sigma^2}, \quad (6a)$$

$$\text{s.t.} \quad (4b)-(4c), \quad (6b)$$

where $\hat{\mu}_k = 1+\mu_k$. Note that $\mathbf{P}_{w2}$ maintains equivalence with the original problem $\mathbf{P}_w$, resulting in no performance loss [9].

An efficient approach is proposed to solve $\mathbf{P}_{w2}$ using MO, which constrains the search space to a manifold that locally resembles Euclidean space. By exploiting this geometric structure, MO enables efficient navigation toward optimal solutions [10]. For $\mathbf{P}_{w2}$, the corresponding manifold is a complex sphere of dimension $(M+1)(K+1)$, leading to a significant reduction in computational complexity [9]. Furthermore, operating directly on the manifold avoids the need for relaxations or approximations, common limitations of conventional methods, thereby improving both accuracy and efficiency.

To this end, we introduce a modified matrix $\mathbf{V} = [\mathbf{v}_c, \mathbf{v}_1, \cdots, \mathbf{v}_K]$ to handle the inequality and to normalize the power constraint (4c), resulting in $\mathrm{Tr}(\mathbf{V}\mathbf{V}^{\mathrm{H}}) = \mathrm{Tr}(\mathbf{W}\mathbf{W}^{\mathrm{H}}) + \|\mathbf{z}\|_2^2$, where $\mathbf{v}_c = [\mathbf{w}_c^{\mathrm{T}}, z_c]^{\mathrm{T}}$, $\mathbf{v}_k = [\mathbf{w}_k^{\mathrm{T}}, z_k]^{\mathrm{T}}$, and $\mathbf{z} = [z_c, z_1, \cdots, z_K]^{\mathrm{T}}$ is an auxiliary vector introduced to simplify power normalization while preserving the original constraint [9]. This results in a CSM $\mathcal{M} = \{\mathbf{V} \in \mathbb{C}^{(M+1)\times(K+1)} \mid \mathrm{Tr}(\mathbf{V}\mathbf{V}^{\mathrm{H}}) = 1\}$. Thus, $\mathbf{P}_{w2}$ is reformulated into a constrained optimization over $\mathcal{M}$ as

$$\mathbf{P}_{w3}: \min_{\mathbf{V}\in\mathcal{M}} \; f(\mathbf{V}) = -\sum_{k=1}^{K}\frac{\hat{\mu}_k|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{k+1}|^2}{\sum_{i=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{i+1}|^2+\sigma^2}, \quad (7a)$$

$$\text{s.t.} \quad u_k(\mathbf{V}) = \Gamma_c^{\mathrm{th}} - \frac{|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_1|^2}{\sum_{i=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{i+1}|^2+\sigma^2} \leq 0, \quad (7b)$$

where $\hat{\mathbf{h}}_k = \sqrt{p_{\max}}[\mathbf{h}_k, 0]$ is adjusted to match the problem's dimensionality. Here, $f(\mathbf{V})$ and $u_k(\mathbf{V})$ are continuous differentiable functions from $\mathcal{M}$ to $\mathbb{R}$. However, the constraint (7b) goes beyond the manifold constraint. Thus, we employ the ALM to incorporate (7b) into the objective with a penalty term [10]. The resulting Lagrangian cost function is given as

$$\mathcal{L}_\rho(\mathbf{V},\boldsymbol{\lambda}) = f(\mathbf{V}) + \frac{\rho}{2}\sum\nolimits_{k=1}^{K}\max\left\{0, \lambda_k/\rho + u_k(\mathbf{V})\right\}^2, \quad (8)$$

where $\rho > 0$ is the penalty parameter controlling the enforcement of the constraints and $\boldsymbol{\lambda} \geq 0 \in \mathbb{R}^N$ is the Lagrange multiplier vector. Following standard ALM approaches [10], [11], ρ is initialized with a moderate positive value and adaptively increased only when the reduction in constraint violation becomes insufficient, thereby balancing convergence stability and feasibility enforcement. The ALM optimizes $\mathbf{V}$ for a given $\boldsymbol{\lambda}$ using the MO and updates $\boldsymbol{\lambda}$ with a gradient-type rule [11]. The resulting optimization problem can be given as

$$\mathbf{P}_{w4}: \min_{\mathbf{V}\in\mathcal{M},\boldsymbol{\lambda}} \quad \mathcal{L}_\rho(\mathbf{V},\boldsymbol{\lambda}). \quad (9)$$

Optimizing $\mathbf{P}_{w4}$ over $\mathcal{M}$ involves the following steps [9], [10]: (i) Riemannian gradient computation, (ii) Search direction, (iii) Retraction (mapping), and (iv) Lagrange multiplier update. For more details, insights, and algorithmic details, interested readers are referred to [9], [10], and we omit them for brevity. Moreover, at iteration t, the Euclidean gradient $\nabla_{\mathbf{V}_t}\mathcal{L}_\rho(\mathbf{V},\boldsymbol{\lambda})$, where $\nabla_{\mathbf{V}_t}$ denotes the gradient with respect to the current iterate $\mathbf{V}_t$, required for the Riemannian gradient computation is given in (10).

B. Sub-Problem 2: Optimizing the common rate allocations

With fixed $\mathbf{w}_c$ and $\{\mathbf{w}_k\}_{k=1}^K$, $\mathbf{P}$ becomes a common rate allocation problem as follows:

$$\mathbf{P}_C: \max_{\{C_k\}_{k=1}^K} \quad \sum\nolimits_{k=1}^{K} C_k, \quad (11a)$$

$$\text{s.t.} \quad \sum\nolimits_{i=1}^{K} C_i \leq \log_2(1+\gamma_{c,k}), \; \forall k, \quad (11b)$$

$$C_k \geq 0, \; \forall k. \quad (11c)$$

Problem $\mathbf{P}_C$ is a convex problem with respect to $\{C_k\}_{k=1}^K$ and can be addressed using CVX Matlab [8].

C. Computational Complexity and Convergence

The computational burden of the FRSMA mainly arises from the MO iterations. Specifically, the computational cost per iteration is $\mathcal{O}(MK+MK^3)$. Assuming that the algorithm converges after R iterations, the overall complexity is therefore on the order of $\mathcal{O}\big(R(MK+MK^3)\big)$ [9], [10].

At iteration t, the update satisfies $\mathcal{L}_\rho(\mathbf{V}_{t+1},\boldsymbol{\lambda}_t) \leq \mathcal{L}_\rho(\mathbf{V}_t,\boldsymbol{\lambda}_t) + \epsilon_t$, with $\epsilon_t \to 0$ [11]. Since $\mathcal{L}_\rho(\cdot)$ is lower bounded over the compact CSM, the sequence converges. Moreover, each block is updated to a stationary point of the corresponding subproblem, and the surrogate functions satisfy first-order consistency. The ALM updates ensure asymptotic feasibility. Under standard constraint qualification and bounded multiplier assumptions, any accumulation point satisfies the KKT conditions [9]–[11]. The iterations proceed until $\|\mathbf{V}_{t+1} - \mathbf{V}_t\| \leq d_{\min}$.

$$\nabla_{\mathbf{V}_t}\mathcal{L}_\rho(\mathbf{V},\boldsymbol{\lambda}) = \sum_{k=1}^{K} -\hat{\mu}_k \left(\frac{2\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{k+1}\hat{\mathbf{h}}_k\mathbf{E}_{k+1}^{\mathrm{H}}}{\sum_{j=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{j+1}|^2+\sigma^2} - \sum_{i=1}^{K} \frac{2|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{k+1}|^2\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{i+1}\hat{\mathbf{h}}_k\mathbf{E}_{i+1}^{\mathrm{H}}}{\left(\sum_{j=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{j+1}|^2+\sigma^2\right)^2} \right)$$
$$- 2\rho\sum_{k=1}^{K}\mathbf{1}_{\left\{\frac{\lambda_k}{\rho}+u_k(\mathbf{V})\right\}}\left(\frac{\kappa_k}{\rho}+u_k(\mathbf{V})\right)\left(\frac{2\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_1\hat{\mathbf{h}}_k\mathbf{E}_1^{\mathrm{H}}}{\sum_{j=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{j+1}|^2+\sigma^2} - \sum_{i=1}^{K}\frac{2|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_1|^2\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{i+1}\hat{\mathbf{h}}_k\mathbf{E}_{i+1}^{\mathrm{H}}}{\left(\sum_{j=1}^{K}|\hat{\mathbf{h}}_k^{\mathrm{H}}\mathbf{V}\mathbf{E}_{j+1}|^2+\sigma^2\right)^2}\right) \quad (10)$$

VI. Simulation Results

We model the large-scale fading ζ_k using the 3GPP urban micro (UMi) model with $f_c = 3\,\mathrm{GHz}$ operating frequency [16, Table B.1.2.1]. The AWGN variance, σ^2, is given by $\sigma^2 = 10\log_{10}(N_0 B N_f)\,\mathrm{dBm}$, where $N_0 = -174\,\mathrm{dBm/Hz}$, $B = 10\,\mathrm{MHz}$ is the bandwidth, and $N_f = 10\,\mathrm{dB}$ is the noise figure. The BS is placed at $\{0,0\}$ while the users are randomly distributed in a circle centered at $\{80,0\}$ with a radius of $10\,\mathrm{m}$. Unless stated otherwise, all algorithmic parameters follow those in [9]. Simulations are evaluated for 10^3 iterations.

We compare the proposed algorithm for the RSMA-assisted communication systems against the following benchmarks:

1) CCPA Benchmarking: This uses SDR and SCA beamforming techniques, in line with [5]–[7]. Specifically, the beamforming vectors are lifted to semidefinite matrices as $\mathbf{W}_c \triangleq \mathbf{w}_c\mathbf{w}_c^{\mathrm{H}}$ and $\mathbf{W}_k \triangleq \mathbf{w}_k\mathbf{w}_k^{\mathrm{H}}$ for $k \in \{1,\cdots,K\}$. Here, $\mathbf{W}_c \in \mathbb{C}^{M\times M}$ and $\mathbf{W}_k \in \mathbb{C}^{M\times M}$ are positive semidefinite matrices, i.e., $\mathbf{W}_c \succeq 0$ and $\mathbf{W}_k \succeq 0$, with $\mathrm{Rank}(\mathbf{W}_c) = 1$ and $\mathrm{Rank}(\mathbf{W}_k) = 1$. The SCA technique is applied to convexify the logarithmic objective via first-order Taylor approximation [12], reformulating $\mathbf{P}$ as a standard SDP by relaxing the rank-one constraints [8]. Since the relaxed problem may yield higher-rank solutions, approximate rank-one beamformers are recovered through GR [8].

2) FP Benchmarking: This method employs FP to solve $\mathbf{P}_w$, following a similar approach to [4]. Specifically, a quadratic transform is first applied to the objective function, converting the fractional terms within the logarithm into a subtractive form by introducing auxiliary variables for each SINR term. The resulting problem is then solved via AO between the beamforming variables and the introduced auxiliary variables, which admit closed-form updates. For $\mathbf{P}_C$, the common rate allocation strategy described in Section V-B is adopted.

To ensure a fair comparison, all algorithms are evaluated under identical system configurations, including the same channel realizations, initialization strategy, transmit power constraints, and convergence criteria. Specifically, the beamforming variables are initialized using random complex Gaussian vectors with independent $\mathcal{CN}(0,1)$ entries and normalized to satisfy the transmit power constraint. The same initialization is used for all compared algorithms.

Fig. 2 compares the convergence of the FRSMA, CCPA, and FP schemes for different $M \in \{8,12,16\}$. Convergence is achieved when the normalized change in the objective function falls below 10^{-3}. As per Fig. 2, the proposed FRSMA algorithm converges significantly faster than the benchmark schemes. In particular, the sum rate rapidly increases and stabilizes within fewer than 3 iterations for all considered values of M. In contrast, the CCPA and FP methods require a larger number of iterations to converge, with at least 8 and 5 iterations, respectively. Moreover, this convergence behavior remains consistent across different M. These results demonstrate that the proposed FRSMA algorithm achieves faster convergence compared to existing methods.

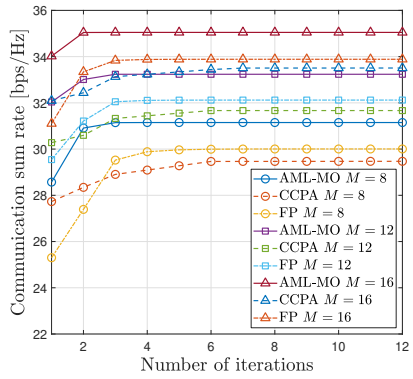

Fig. 2: Convergence rate of FRSMA, CCPA, and FP algorithms.

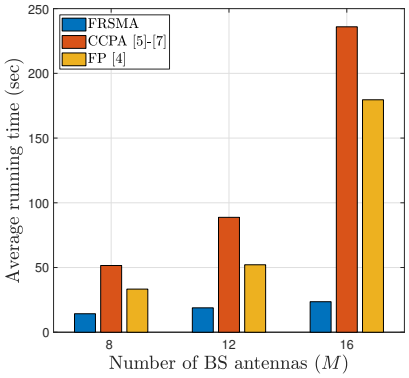

Fig. 3: Average runtimes of the three algorithms versus the number of BS antennas.

Fig. 3 presents the average runtimes of the FRSMA, CCPA, and FP schemes for varying M, based on MATLAB simulations conducted on an Intel® Core™ i7 @ 2.50 GHz. As expected, the runtime increases with M due to the higher problem dimensionality. Nevertheless, the proposed FRSMA algorithm consistently achieves lower runtime compared to both CCPA and FP across all considered values of M. For instance, at $M = 16$, FRSMA is approximately $10\times$ and $8\times$ faster than CCPA and FP, respectively.

Fig. 4 illustrates the communication sum rate achieved by the FRSMA, CCPA, and FP schemes as a function of M. As expected, the sum rate increases with M for all schemes due to the additional spatial degrees of freedom provided by larger antenna arrays. Moreover, the proposed FRSMA algorithm consistently achieves a higher sum rate compared to both CCPA and FP across all considered values of M. For example, at $M = 16$, it attains approximately $5.2\,\%$ and $3.4\,\%$ higher sum rate than CCPA and FP, respectively.

Table I presents the CSP of the FRSMA, CCPA, and FP

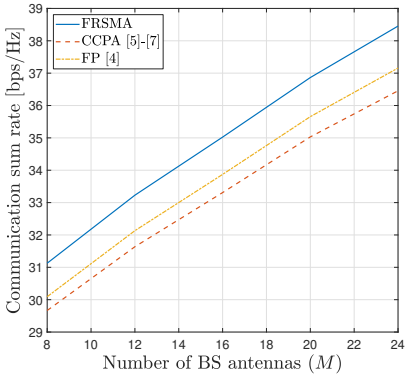

Fig. 4: Communication sum rate as a function of the number of BS antennas.

TABLE I: CSP comparison.

Setup		Algorithm		
M	K	**FRSMA**	**CCPA**	**FP**
8	2	95 %	48 %	88 %
8	4	93 %	47 %	85 %
16	2	98 %	52 %	91 %
16	4	97 %	51 %	90 %

schemes under different system configurations. The CSP is defined as the probability that the obtained beamforming vectors and rate allocation factors satisfy all constraints in (3). Let $\hat{\mathbf{w}}_c$, $\{\hat{\mathbf{w}}_k\}_{k=1}^K$, and $\{\hat{C}_k\}_{k=1}^K$ denote the beamforming vectors and rate allocation factors obtained from a given algorithm, and let $\mathcal{F}$ represent the feasible set. Then, the CSP is given by $\text{CSP} = \mathbb{P}((\hat{\mathbf{w}}_c, \hat{\mathbf{w}}_k, \hat{C}_k) \in \mathcal{F})$. In practice, CSP is estimated empirically over N_{sim} independent channel realizations as $\text{CSP} \approx 1/N_{\text{sim}} \sum_{i=1}^{N_{\text{sim}}} \mathbb{I}\big((\hat{\mathbf{w}}_c^{(i)}, \hat{\mathbf{w}}_k^{(i)}, \hat{C}_k^{(i)}) \in \mathcal{F}\big)$, where $\mathbb{I}(\cdot)$ is the indicator function.

We observe that FRSMA consistently achieves higher CSP compared to the benchmark schemes across all configurations. For instance, at $M = 8$ and $K = 2$, FRSMA attains $95\,\%$, whereas CCPA and FP achieve $48\,\%$ and $88\,\%$, respectively. Moreover, the CSP of FRSMA remains high as the number of users increases, indicating stable performance across different system dimensions. These results demonstrate the reliability of the proposed approach in satisfying the problem constraints.

The performance gains of FRSMA stem from its efficient handling of the original non-convex problem without relying on relaxations [9], [10]. In particular, the CCPA relies on SDR and SCA, which lift the problem into a higher-dimensional space (i.e., $M^2(K+1)$) and require GR for rank-one recovery, leading to increased computational complexity and potential performance loss. Similarly, the FP method introduces auxiliary variables and alternating updates across decoupled subproblems, which results in slower convergence and less effective constraint enforcement. In contrast, FRSMA exploits the intrinsic geometry of the beamforming variables via MO, reducing the search space to $(M+1)(K+1)$, and incorporates the constraints directly through an augmented Lagrangian framework. This enables efficient updates without relaxation or post-processing, thereby achieving faster convergence, lower runtime, improved sum rate, and high CSP.

VII. Conclusion

This paper develops a low-complexity framework for joint RSMA beamforming and rate allocation in a multi-antenna BS serving multiple users, maximizing sum rate under common rate and power constraints. The resulting non-convex problem is solved via an AO-based FRSMA approach that exploits the manifold structure to reduce the dimension of the search space to $(M+1)(K+1)$. The proposed method significantly lowers complexity while improving sum rate and CSP over CCPA and FP-based methods, demonstrating strong scalability with the number of antennas and users.

Limitations and Future Work: While FRSMA demonstrates clear performance gains under ideal conditions, its behavior in challenging regimes, such as low SNR, highly orthogonal user channels, severe CSI errors, imperfect SIC, tight common-rate constraints, and poor initialization, warrants further investigation. This work assumes centralized scheduling, perfect SIC, and ideal CSI. Practical aspects, including signaling latency, SIC processing delay, and imperfect SIC, are not modeled. Extending FRSMA to robust designs under those conditions is an important direction for future work.

References

[1] Y. Mao *et al.*, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2073–2126, 4th Quart. 2022.

[2] C. Xu, B. Clerckx, S. Chen, Y. Mao, and J. Zhang, "Rate-splitting multiple access for multi-antenna joint radar and communications," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1332–1347, Nov. 2021.

[3] Y. Wang, V. W. S. Wong, and J. Wang, "Flexible rate-splitting multiple access with finite blocklength," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1398–1412, May 2023.

[4] T. Fang and Y. Mao, "Rate splitting multiple access: Optimal beamforming structure and efficient optimization algorithms," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15 642–15 657, Oct. 2024.

[5] Y. Yang *et al.*, "Joint beamforming and rate splitting design for RIS assisted RSMA systems," in *Proc. Int. Conf. Future Commun. Netw. (FCN)*, Nov. 2024, pp. 1–6.

[6] R. Shahcheragh and K. Mohamed-pour, "Beamforming design for STAR-RIS assisted secure wireless communication system under hardware impairments," *EURASIP J. Wireless Commun. Netw.*, Jul. 2024.

[7] Z. Qiu, Y. Mao, S. Ma, and B. Clerckx, "Robust max–min fair beamforming design for rate splitting multiple access-aided visible light communications," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 10 043–10 057, Apr. 2025.

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

[9] S. Zargari, D. Galappaththige, C. Tellambura, and H. V. Poor, "A Riemannian manifold approach to constrained resource allocation in ISAC," *IEEE Trans. Commun.*, vol. 73, no. 5, pp. 3655–3670, May 2025.

[10] C. Liu and N. Boumal, "Simple algorithms for optimization on Riemannian manifolds with constraints," *Appl. Math. Optim.*, vol. 82, pp. 949–981, Mar. 2020.

[11] E. G. Birgin and J. M. Martínez, *Practical Augmented Lagrangian Methods for Constrained Optimization*. Philadelphia, PA: Soc. Ind. Appl. Math., 2014.

[12] D. Galappaththige and C. Tellambura, "Sum rate maximization for RSMA-assisted CF mMIMO networks with SWIPT users," *IEEE Wireless Commun. Lett.*, vol. 13, no. 5, pp. 1300–1304, May 2024.

[13] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[14] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel & Scientific Computations*, vol. 11, no. 4, pp. 351–368, Dec. 2003.

[15] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[16] "3GPP TR 36.814, further advancements for E-UTRA physical layer aspects, V.9.0.0 Rel. 9," Mar. 2010. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2493