Asynchronous 2-layer Full-duplex Cooperative RSMA with Imperfect Channel State Information and Imperfect Successive Interference Cancellation

Sama Wahb[a,*], Ahmed El-Mahdy[a], Falko Dressler[b]

[a] *Communications Engineering, German University in Cairo, Cairo, Egypt*
[b] *School of Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany*

Abstract

Sixth-generation (6G) wireless systems require advanced multiple access techniques capable of delivering high data rates, low latency, and robust performance under practical impairments. Rate-Splitting Multiple Access (RSMA) has emerged as a strong candidate due to its efficient interference management and robustness to imperfect Channel State Information (CSI). However, its performance is often limited by the weakest user, and existing cooperative approaches mainly rely on half-duplex relaying and ideal assumptions. In this paper, a downlink full-duplex multi-user 2-layer cooperative RSMA (C-RSMA) framework is proposed under asynchronous reception, imperfect CSI, and imperfect Successive Interference Cancellation (SIC). The 2-layer structure enhances interference mitigation and fairness, while full-duplex relaying improves spectral efficiency. An alternative optimization technique based on Weighted Minimum Mean Square Error (WMMSE) is used to jointly optimize precoding, rate allocation, and relay power to maximize the minimum user rate under latency constraints. Numerical results show that the proposed scheme enhances fairness and robustness over other schemes.

Keywords: Rate Splitting Multiple Access, Cooperative Technology, Imperfect CSI, Imperfect SIC, Asynchronous Reception

1. Introduction

Nowadays, sixth-generation mobile communications (6G) and beyond are attracting momentous attention from academia and industry due to their ability to enable the internet of everything (IoT) and provide services with higher throughput, ultra-reliability, heterogeneous quality of service, massive connectivity, and ultra-reliable low latency. 6G networks are expected to provide a data rate of 1 Tbps, a latency of 1 ms, and a connection density of 106 devices per square kilometer. Furthermore, there will be a significant increase in the number of connected devices.

One of the major challenges in meeting the aforementioned 6G requirements is effective interference management. Interference is expected to be both severe and unpredictable, and improper handling can significantly degrade system performance. Traditional techniques such as Space Division Multiple Access (SDMA) and Non-Orthogonal Multiple Access (NOMA) have shown limitations in managing interference efficiently. In SDMA, interference is treated purely as noise, which is only effective when the interference is weak. While, in NOMA, interference is fully decoded, which is only practical when the interference is strong. For optimal performance, the proportion of interference to be decoded and that to be treated as noise should be adaptively adjusted based on the actual interference level. One approach to achieve this adaptability is to dynamically switch between SDMA and NOMA schemes. Moreover, imperfect CSI can also be considered as one of the main sources of multi-user interference. The imperfect CSI makes the interference management more challenging, as the interference cannot be managed easily by the precoders at the transmitter because interference cannot be eliminated since the channel is not known accurately. Therefore, it becomes essential to find adequate multiple access (MA) techniques that can perform well in case of imperfect CSI, be flexible enough to adapt to the interference level rather than operate in the two extremes of fully treating interference as noise and fully decoding interference, and be general enough to outperform all existing MA techniques in order to enhance the system performance. One such MA technique that can achieve these requirements is rate splitting multiple access (RSMA).

RSMA uses different schemes: 1-layer RSMA and 2-layer RSMA. In 1-layer RSMA, the message of each user is split into two parts: common part and private part. The common parts of the users are combined together and encoded into one stream, while the private parts are encoded independently. While in the 2-layer RSMA, the message of each is split into three parts: inter-group common message, inner-group common message, and private message. The inter-group common messages from all users are combined together into one inter-group common message and is decoded by all users. The inner-group common messages of

*Corrsponding author
Email address: sama.ashraf@guc.edu.eg (Sama Wahb)

the users in the same group are combined into one inner-group common message and is decoded by the users in that group only. While the private messages are independently encoded into private streams that is decoded by its intended user only. This additional common layer provides higher degrees of freedom and greater flexibility in interference management and power allocation among streams, resulting in improved spectral and energy efficiency compared to 1-layer RSMA.

A main drawback of RSMA is that the system performance is often constrained by the user with the worst channel condition, limiting the achievable system rates. C-RSMA has been proposed to mitigate this bottleneck, where stronger users relay information to weaker users. However, many studies rely on half-duplex (HD) relaying, which suffers from wasted system resources. Despite the promising performance of 2-layer RSMA, several limitations remain in the existing literature. Most works assume perfect CSI and perfect SIC, which are difficult to achieve in practical systems due to channel estimation errors and hardware imperfections. Furthermore, due to the varying signal arrival times at the users, asynchronous reception naturally occurs in cooperative communications.

To address these gaps, this paper proposes a full-duplex (FD) multi-user 2-layer C-RSMA system under asynchronous reception, imperfect CSI, and imperfect SIC, with the goal of improving fairness of users. Full-duplex user relaying improves the spectral efficiency by allowing simultaneous transmission and reception, which is critical for 6G scenarios. Moreover, we incorporate a latency constraint to ensure real-time performance under practical system impairments.

Our main contributions can be summarized as follows:

- We integrate the 2-layer RSMA with full-duplex user relaying to improve the maximum minimum rate (Max-Min rate) of multi-user communications system under imperfect CSI, imperfect SIC, and asynchronous reception. By jointly considering 2-layer RSMA and full-duplex user cooperation under asynchronous reception and imperfect CSI/SIC, this work offers insights for the design of practical multi-user systems. The additional common layer in 2-layer RSMA provides enhanced flexibility in interference handling, allowing the system to better accommodate channel uncertainty and residual interference arising from non-ideal SIC. Full-duplex cooperative relaying reduces the fairness limitation of conventional RSMA by allowing strong users to continuously assist weaker users without losing spectral efficiency. Furthermore, the analysis demonstrates that asynchronous signal arrivals introduce structured phase distortions that have a tangible impact on achievable rates and cannot be neglected in realistic deployments. Overall, these findings highlight the importance of designing RSMA-based systems under realistic conditions rather than ideal assumptions.
- We formulate an optimization problem that maximizes the minimum rate of users by jointly optimizing the precoding matrix, the inter-group common rate allocation, the inner-group common rate allocation, the head relay power, and the group relay power while restricting the latency to be less than a certain threshold value.
- We solve the optimization problem using alternative optimization based on weighted minimum mean square error (WMMSE). The performance of the system is compared with other systems mentioned in literature to show the superiority of this system.

The rest of this paper is organized as follows: Section 2 reviews the related work and highlights the research gap addressed in this study. Section 3 presents the system model. Section 4 formulates the optimization problem for max–min rate performance under latency constraints. Section 5 details the proposed solution based on the WMMSE-driven alternative optimization algorithm. Section 6 provides the numerical results and performance comparisons with benchmark schemes. Finally, Section 7 concludes the paper.

2. Related Work

Traditional multiple access techniques are unable to meet the demands of future networks. RSMA has emerged as a promising technique, thanks to its ability to unify and generalize existing multiple access schemes. In particular, 2-layer RSMA has been shown to outperform 1-layer RSMA due to its ability in handling interference and distributing power efficiently. Many research works demonstrate the potential and the performance enhancements of 2-layer RSMA. Sadeghabadi et al. [1] study RSMA beamforming design for MIMO broadcast channels under a total power constraint. They derive a parameterization of optimal beamforming for different performance maximization problems. Their approach balances SNR maximization and inter-stream interference reduction, followed by power allocation to complete the design. Qi et al. [2] explore how the transmit power is reduced in the intelligent reflecting surface (IRS)-assisted 2-layer RSMA systems by jointly optimizing the user clustering, common rate distribution, transmit beamforming, and IRS passive beamforming using a low-complexity user clustering scheme. Simulation results demonstrate that the proposed scheme can significantly reduce power consumption compared with the benchmarks. Yang et al. [3] improve the sum rate by integrating RSMA with IRS. Numerical results indicate that the proposed scheme evidently surpasses 1-layer RSMA, OMA, SDMA, and no IRS schemes, especially in overloaded scenarios. Sharma et al. [4] study a simultaneous transmitting and reflecting (STAR) IRS assisted communication system, where a base station serves cell-edge users via a 2-layer RSMA protocol. The work analyzes the outage probability for energy splitting and mode switching modes, accounting

for spatial correlation caused by the STAR-RIS structure. Liang et al. [5] provide analytical expressions and numerical results for outage performance of an unmanned aerial vehicle terrestrial 2-layer RSMA communication system with uniformly distributed users and an interfering base station. It is concluded that the power allocation proportion between common and private signals and the fading parameters can significantly affect the outage performance. Niu et al. [6] integrate active IRS with RSMA for 1-layer and 2-layer RSMA. Their objective is to achieve a tradeoff between spectral efficiency (SE) and energy efficiency (EE). Simulation results show that through reconfiguring the wireless communication environment, the active IRS can achieve an adjustable tradeoff between the SE and EE, and the proposed active IRS assisted RSMA scheme has superiority over the other systems. Liu et al. [7] propose an active IRS aided 2-layer RSMA scheme for the mmWave hybrid antenna array, where a user grouping method and analog beamformer are designed to avoid the narrow beam width constraint. The digital precoder, the rate splitting vector, and the reflecting vector are jointly optimized in order to achieve the max-min fairness. RSMA can be integrated with satellite communication where a downlink non-orthogonal broadcast and uni-cast (NOBU) joint transmission scheme is considered by combining the 1-layer RSMA and multi-beam satellite system. Also, a general NOBU joint transmission using the 2-layer RSMA strategy is proposed [8].

All the above mentioned papers consider perfect channel state information (CSI), but other papers consider more practical scenarios where the channel state is imperfect. Park et al. [9] propose a new approach for linear precoding optimization in downlink MIMO with RSMA in case of imperfect CSI. The objective is to maximize the sum spectral efficiency of the 2-layer RSMA while complete analysis is missing. Pereira et al. [10] use a neural network to overcome one of the major challenges for 2-layer RSMA, which is the necessity to know the optimal clustering. The accuracy and performance metrics show that the neural network is able to learn and cluster the users based on the noisy channel response. Li et al. [11] provide an overview of integrating RSMA and IRS. It covers both 1-layer RSMA and 2-layer RSMA. It illustrates the rate region of IRS-aided RSMA. It discusses the major advantages, the research challenges, and open problems for IRS-aided RSMA systems. Jolly et al. [12] improve the weighted sum rate (WSR) for a multi-user downlink IRS assisted 2-layer RSMA. The users are divided into near users served by BS and cell-edge users served by IRS. The scenario of imperfect CSIT is introduced for comparison purposes. Simulations show that the proposed beamformer design for IRS-RSMA achieves better performance when compared to the IRS-NOMA transmission scheme. Bansal et al. [13] derive closed-form expressions of outage probability for both cell-edge users and near users for a downlink multi-user multiple IRS assisted 2-layer RSMA. Also, an On-Off scheme for controlling the IRSs with practical phase shifts has been proposed. It shows that the proposed system outperforms the other mentioned systems. Moreover, the impact of channel estimation errors, increasing the number of reflecting elements, and the number of cell-edge users on the outage probability performance has been described. Dai et al. [14] propose a general framework of 2-layer RSMA that is particularly suited to massive MIMO systems. A closed form expressions for power allocation is derived which provides insights into the effects of system parameters. Simulation results validate the significant sum rate gain of 2-layer RSMA over various baselines. Li et al. [15] propose a 2-layer RSMA framework with a balanced clustering design for cell-free massive MIMO (CF-mMIMO) systems, aiming to enhance spectral efficiency while maintaining low complexity and scalability. A two-stage optimization approach is introduced, incorporating spectral clustering to adaptively determine cluster numbers and jointly optimize inner and outer RSMA for improved interference management. Simulation results demonstrate notable SE gains and efficient scalability compared to conventional user-centric clustering schemes.

None of the aforementioned papers address the critical practical impairment of imperfect successive interference cancellation (SIC), which is unavoidable in real-world systems due to hardware limitations and channel estimation error [16]. Sadeghabadi et al. [17] propose low-complexity 2-layer RSMA schemes and introduce RSMA with Interference Nulling to reduce precoding complexity by partially nulling interference. They formulate a convex upper bound problem for the sum rate to compare existing algorithms, and conditions under which the upper bound problem is tight are determined. An upper-bound-aided method is also developed to enhance existing designs when the bound is tight. Simulation results indicate that the proposed approaches enhance performance over existing ones as well as approach their upper bounds in certain instances.

All the aforementioned papers share a common limitation of RSMA, namely that the system rate is constrained by the user with the worst channel conditions. This drawback can be mitigated by integrating RSMA with cooperative user relaying. Khisa et al. [18] propose a 2-layer half duplex C-RSMA scheme in a multi-user system where one cell-center-user (CCU) and one cell-edge-user (CEU) form a pair to maximize the overall sum rate. Each CCU relays the inner-group common message to its paired CEU to improve the signal quality of the CEU. The results show that the proposed system outperforms the other mentioned systems. However, this paper focuses on half-duplex relaying, which leads to inefficient resource utilization and increased latency. To address this limitation, full-duplex relaying should be adopted to fully exploit the available resources by enabling simultaneous transmission and reception and hence enhance spectral efficiency. Furthermore, papers [19], [20], and [21] investigate asynchronous reception, which is not focused on in the above papers, but none of these papers consider the RSMA technique. From the preceding investigations, it is evident that asynchronous reception of

Table 1: A summary of selected 2-layer RSMA-related works

Reference	Cooperative Technology	Full-duplex	Imperfect CSI	Imperfect SIC	Latency Constraint	Asynchronous Reception
[1]	✘	✘	✘	✘	✘	✘
[2]	✘	✘	✘	✘	✘	✘
[3]	✘	✘	✘	✘	✘	✘
[4]	✘	✘	✘	✘	✘	✘
[5]	✘	✘	✘	✘	✘	✘
[6]	✘	✘	✘	✘	✘	✘
[7]	✘	✘	✘	✘	✘	✘
[8]	✘	✘	✘	✘	✘	✘
[9]	✘	✘	✔	✘	✘	✘
[10]	✘	✘	✔	✘	✘	✘
[11]	✘	✘	✔	✘	✘	✘
[12]	✘	✘	✔	✘	✘	✘
[13]	✘	✘	✔	✘	✘	✘
[14]	✘	✘	✔	✘	✘	✘
[15]	✘	✘	✔	✘	✘	✘
[17]	✘	✘	✘	✔	✘	✘
[18]	✔	✘	✘	✘	✘	✘
Our Work	✔	✔	✔	✔	✔	✔

full-duplex multi-user 2-layer C-RSMA system operating under both imperfect CSI and imperfect SIC has not been previously explored. The current works are summarized in Table 1. In this work, we cover this limitation, and we maximize the minimum rate of multi-user in downlink full-duplex C-RSMA system with imperfect CSI and imperfect SIC under asynchronous reception and low latency. Although this work builds on RSMA and cooperative relaying, it introduces a fundamentally different design perspective. Specifically, the integration of 2-layer RSMA with cooperative relaying is shown to provide clear performance gains over conventional multiple-access schemes, non-cooperative RSMA, and 1-layer RSMA, particularly in terms of fairness and robustness. This provides new system-level insights and practical design guidelines for RSMA-based cooperative networks operating under realistic 6G conditions as shown in Table 1.

2.1. Innovation and Design Insights

This work significantly advances the cooperative RSMA framework in [18] by overcoming the fundamental limitations of half-duplex relaying and incorporating realistic system impairments. Unlike the half-duplex design, which suffers from an inherent 50% spectral efficiency loss due to orthogonal transmission phases and primarily targets sum-rate maximization under ideal assumptions, the proposed scheme introduces a full-duplex multi-user cooperative architecture that enables simultaneous transmission and reception, thereby enhancing spectral efficiency, reducing latency, and improving fairness. The head relay forwards the inter-group common stream while group relays forward inner-group common streams, forming a hierarchical two-layer structure. The key design insight is that RSMA and full-duplex relaying are complementary: although FD operation enhances the spectral efficiency, it also introduces additional interference sources such as residual self-interference, imperfect SIC, CSI uncertainty, and asynchronous reception. The two-layer RSMA structure provides enhanced interference management flexibility by partially decoding interference via common streams while treating the remaining interference as noise, thus stabilizing system performance under practical non-idealities. Furthermore, a latency constraint is explicitly integrated within a max–min fairness optimization framework. To the best of our knowledge, this is the first work to jointly incorporate full-duplex cooperative relaying, hierarchical two-layer RSMA, and multiple practical impairments into a unified fairness-oriented design suitable for realistic 6G deployments. A detailed comparison between the proposed work and [18] can be presented in Table 2.

While the works in [19]-[21] investigate asynchronous reception effects in multi-user systems. These studies model delay-induced phase shifts. However, they do not consider 2-layer RSMA transmission strategies, or cooperative rate-splitting. Moreover, these works typically do not integrate fairness-oriented optimization or practical impairments such as imperfect SIC within an RSMA framework. The proposed work extends beyond signal-level asynchronous analysis by embedding delay-dependent phase modeling into a full-duplex 2-layer cooperative RSMA system. The asynchronous phase shifts are explicitly incorporated into the channel model and SINR derivations, directly influencing common and private stream decoding. The presence

Table 2: Comparison between the proposed work and [18].

Aspect	**Proposed Work**	**[18] (Khisa et al., 2024)**
Relaying Mode	Full-Duplex (FD)	Half-Duplex (HD)
Relaying Scope	Dual-layer relaying: Head relay forwards inter-group and inner-group common streams; group relays forward inner-group common streams.	Single-layer relaying: Each relay forwards only the inner-group common stream to its paired user.
Channel State Information (CSI)	Imperfect CSI	Perfect CSI
Successive Interference Cancellation (SIC)	Imperfect SIC	Perfect SIC
Reception Type	Asynchronous reception	Synchronous reception
Performance Metrics	Max-min rate, fairness (Jain's index), and sum rate	Sum rate only
Key Novelty / Contribution	We consider 2-layer C-RSMA with full-duplex relaying under asynchronous reception, imperfect CSI, and imperfect SIC within a fairness-oriented framework.	2-layer C-RSMA with HD relaying under ideal assumptions (no asynchronous modeling or practical impairments).

of imperfect CSI and imperfect SIC further interacts with asynchronous reception, requiring robust joint WMMSE-based optimization.

To motivate the design of the proposed work, several practical design guidelines can be identified for 6G wireless systems. These guidelines highlight key architectural and optimization considerations required to ensure robustness, fairness, and latency-awareness in realistic multi-user deployments. First, full-duplex cooperative relaying should be preferred over half-duplex structures in fairness-critical and latency-sensitive scenarios, as eliminating orthogonal transmission phases significantly improves spectral efficiency while enabling continuous assistance to weak users. Second, asynchronous reception effects must be explicitly modeled in distributed and cooperative deployments, since delay-induced phase rotations reduce coherent combining gain and interact non-trivially with channel uncertainty; assuming perfect synchronization may lead to overly optimistic performance predictions. Third, imperfect successive interference cancellation (SIC) should be incorporated into the system design, particularly in multi-layer RSMA architectures, because residual interference propagates across decoding stages and can substantially degrade achievable rates under practical hardware and estimation constraints. Fourth, joint optimization of precoding and relay power allocation is essential in multi-impairment environments, as decoupled or heuristic approaches fail to capture the strong coupling between interference management, fairness, and latency. Finally, latency constraints should be embedded directly into the optimization formulation rather than

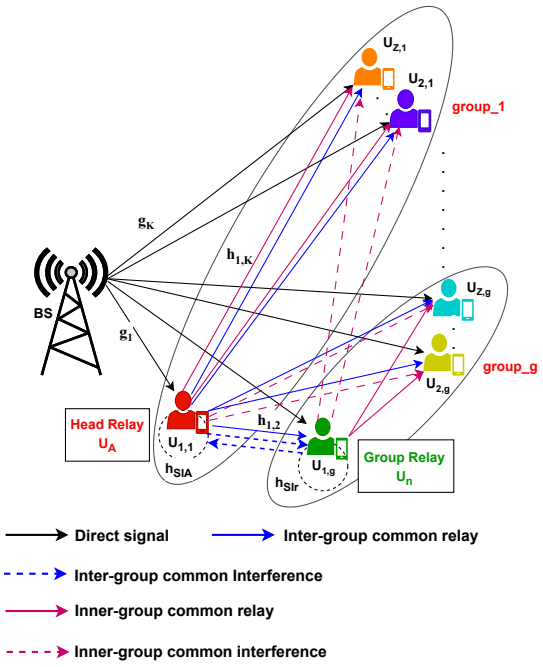

Figure 1: Full-duplex multi-user cooperative 2-layer RSMA system

treated as an external performance metric, ensuring that 6G systems simultaneously achieve reliability, fairness, and real-time service guarantees. Collectively, these guidelines highlight the necessity of impairment-aware, fairness-driven, and jointly optimized transmission strategies for realistic 6G multi-user networks.

3. System Model

We consider a downlink 2-layer full-duplex multi-user cooperative RSMA (FD C-RSMA) wireless communication system, as shown in Figure 1. The base station (BS) is equipped with N_t transmit antennas that serve K users with a single antenna, where the set of all users is denoted by $\mathcal{K} = \{1, 2, ..., K\}$. The K users are divided equally among G groups, and each group $g \in \mathcal{G} = \{1, \ldots, G\}$ contains $Z = \frac{K}{G}$ users. The user set $\mathcal{K}$ is divided into two non-overlapping subsets: denoted by $\mathcal{N}$ with $|\mathcal{N}| = N$, and the non-relay users, denoted by $\mathcal{M}$ with $|\mathcal{M}| = M$. These two subsets form a complete partition of the system, where $\mathcal{N} \cup \mathcal{M} = \mathcal{K}$. Each group-$g$ contains one full-duplex user relay (U_n); therefore, the number of relays is equal to the number of groups, but both have different notation to facilitate the analysis. To improve readability, the main symbols and their physical meanings are summarized in Table 3.

The channel between the BS and user-k is denoted by $\mathbf{g}_k \in \mathbb{C}^{N_t \times 1}$, and the channel between two different users is denoted by $h_{i,j}$ where $i \neq j$, and $i, j \in \mathcal{K}$. It is assumed that the nearest user to the BS has the best channel condition among the rest of the users as it has the highest estimated channel gain $\| \hat{\mathbf{g}}_k \|^2$, and acts as the head relay (U_A).

Table 3: List of Main Notations

Parameter	Description
A) System parameters	
N_t	Base station transmit antennas
$\mathcal{K}$	Set of all users
K	Total number of users
$\mathcal{G}$	Set of groups
G	Number of groups
g	Group index
Z	Number of users per group
$\mathcal{N}$	Set of full-duplex relay users
N	Number of relay users
n	Relay user index
$\mathcal{M}$	Set of non-relay users
M	Number of non-relay users
m	Non-relay user index
A	The head relay index
β	SIC imperfection factor
$\sigma^2_{e,k}$	CSI error variance for user k
n_0	AWGN noise
σ^2_{SI}	Residual self-interference power
B) Channel parameters	
$\mathbf{g}_k$	Channel between BS and user-k
$h_{i,j}$	Channel between users i and j
$\hat{\mathbf{g}}_k$	Estimated channel between BS and user-k
$\hat{h}_{i,j}$	Estimated inter-user channel
$\delta_{n,k}$	Asynchronous phase shift between relay n and user k
C) Optimization variables	
$\mathbf{x}$	Transmitted signal from BS
$\mathbf{p}_c$	Precoding vector for inter-group common stream
s_c	Inter-group common stream
$\mathbf{p}_{c_g}$	Precoding vector for inner-group common stream of group g
s_{c_g}	Inner-group common stream of group g
$\mathbf{p}_k$	Precoding vector for private stream of user-k
s_k	Private stream of user-k
$\mathbf{P}$	Overall precoding matrix
P_A	Head relay (U_A) transmit power
P_n	Group relay (U_n) transmit power
B) Constraints parameters	
P_t	Maximum transmit power of the BS
L_k	Normalized data size of user k
$t_{\max}$	Maximum latency threshold

User selection within each group is based on channel conditions. Specifically, the strongest user is selected as the head relay U_A, where $A \in \mathcal{N}$, and the next strongest $(N-1)$ users are selected as the group relays U_n, with one relay assigned to each group. After selecting the relay users, the remaining users are distributed across the groups according to their channel strengths. Specifically, each relay U_n is assigned $(Z-1)$ of the weaker users so that every group contains exactly Z users. The strongest relay is grouped with the weakest $(Z-1)$ users, the second strongest relay with the next weakest $(Z-1)$ users, and so on, until all groups are formed. Each group therefore consists of one relay user and $(Z-1)$ non-relay users, ensuring a balanced cooperative structure based on channel conditions. For example, if the channels between the BS and the users are arranged in descending order, where $\mathbf{g}_1 \geq \mathbf{g}_2 \geq \mathbf{g}_3 \geq .... \geq \mathbf{g}_{K-1} \geq \mathbf{g}_K$. Then U_1 acts as the head relay, and group-1 consists of Z users defined as:

$$\text{Group-1: } \{U_1, U_{K-(Z-2)}, \ldots, U_{K-1}, U_K\},$$

while group-2 consists also of Z users defined as:

$$\text{Group-2: } \{U_2, U_{K-2(Z-2)-1}, \ldots, U_{K-(Z-2)-1}\},$$

where U_2 is the group relay, and so on. Each user employs 2-layer RSMA, where its message is split into three parts: an inter-group common message, an inner-group common message, and a private message. The N full-duplex relay users are divided into one head relay (U_A), where $A \in \mathcal{N}$, and $(N-1)$ group relays (U_n). The head relay is responsible for forwarding the inter-group common message to the rest of the users in the system and forwarding the inner-group common message to the rest of the users in the same group. While, group relays forwards only the inner-group common message to the remaining users within the same group.

3.1. Imperfect CSI Channel Model

We consider both imperfect CSI and SIC in the system for a more realistic assumption. The imperfect CSI can be represented as $\mathbf{g}_k = \hat{\mathbf{g}}_k + \mathbf{e}_k$, where $\hat{\mathbf{g}}_k$ is the estimate of the channel $\mathbf{g}_k$, and $\mathbf{e}_k$ is the channel estimation error, which is considered as a random variable with zero mean and variance $\sigma^2_{e,k}$. Similarly, the imperfect channel between the users i and j is expressed as $h_{ij} = \hat{h}_{i,j} + e_{i,j}$.

3.2. Asynchronous Reception

In FD C-RSMA, the head relay and the group relay receive the signal from the BS and retransmit it to the intended users. Since the positions of the relays and their distances from the users are different, this causes unavoidable differences in the arrival time of the signal, resulting in asynchronous reception, which introduces a phase shift $\delta_{n,k}$ between the nth relay and the kth user and is given by [19]:

$$\delta_{n,k} = e^{-j2\pi \frac{t_{n,k}}{L_k}}, \tag{1}$$

where $t_{n,k} = d_{n,k}/c$ is the timing offset of the signal intended for the kth user and transmitted by the nth relay. Also, $d_{n,k}$, c, and L_k are the distance between the kth user and the nth relay, the speed of light, and the normalized data size, respectively. We assume that the first arrived

signal to the user k is from BS and its timing offset is zero, and hence this delay will affect the reception of the signal between users only. Considering the effect of the delay phase, the channel between two different users is given by

$$h_{n,k} = \delta_{n,k}\hat{h}_{n,k} + e_{n,k}. \tag{2}$$

by substituting (1) in (2), and expressing the estimated channel in amplitude-phase form as $|\hat{h}_{n,k}|e^{j\theta_{n,k}}$, the equivalent phase-rotated channel representation is given by

$$h_{n,k} = |\hat{h}_{n,k}|e^{j\left(\theta_{n,k} - 2\pi\frac{t_{n,k}}{L_k}\right)} + e_{n,k}. \tag{3}$$

where $\theta_{n,k}$ denotes the phase of the estimated channel between the kth user and the nth relay. It is observed that the asynchronous factor $\delta_{n,k}$ preserves the channel magnitude while introducing a phase rotation. Although this rotation does not change individual channel power terms, it modifies the relative phase among simultaneously received signal components. Furthermore, since the WMMSE-based optimization jointly updates the precoders, equalizers, and weights using the complex channel coefficients rather than their magnitudes only, the phase shift introduced by $\delta_{n,k}$ alters the equalizer and weight computations and consequently the resulting precoder solution. Therefore, asynchronous reception affects system performance through phase-sensitive signal interactions rather than through power scaling, leading to different optimized power allocation and achievable rate performance compared to the synchronous case.

4. Problem Formulation

In 2-layer RSMA, the message of each user W_k is split into three parts: an inter-group common message $W_{c,k}$, an inner-group common message W_{c,k_g}, and a private message $W_{p,k}$. All the inter-group common messages are combined together and encoded into a common stream s_c using a common codebook shared by all users. The inner-group common messages of the users in group-g are combined together into one inner-group common stream s_{c_g} using a common codebook shared by users in the same group-g. While all the private messages are encoded independently and transformed to private stream s_k. Then, the overall encoded streams can be written as $\mathbf{s} = [s_c, s_{c_1}, s_{c_2}, ..., s_{c_G}, s_1, s_2, ..., s_K]$, hence the signal transmitted by BS can be written as:

$$\mathbf{x} = \mathbf{p}_c s_c + \sum_{g=1}^{G} \mathbf{p}_{c_g} s_{c_g} + \sum_{k=1}^{K} \mathbf{p}_k s_k \tag{4}$$

where $\mathbf{P} = [\mathbf{p}_c, \mathbf{p}_{c_1}, ..., \mathbf{p}_{c_G}, \mathbf{p}_1, ..., \mathbf{p}_K]$ is the precoding matrix, and $\mathbf{P} \in \mathbb{C}^{N_t \times (K+G+1)}$. It is assumed that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$ and the transmit power is $tr(\mathbf{P}\mathbf{P}^H) \leq P_t$, where P_t is the maximum transmit power of the BS.

The achievable data rate analysis is divided into three parts: the achievable rate analysis for head relay (U_A), the achievable rate analysis for group relay (U_n) where $n \neq A$, and the achievable rate analysis for non-relay users (U_m).

4.1. Achievable Rate Analysis for Head Relay U_A

The received signal at the head relay user (U_A) received from BS is given by:

$$y_A = \underbrace{(\hat{\mathbf{g}}_A^H + \mathbf{e}_A^H)\mathbf{x}}_{\textit{desired signal from BS}} + \underbrace{h_{SI_A}\sqrt{P_A}(s_c + s_{c_g})}_{\textit{self-interference from inter and inner common message}} + \underbrace{\sum_{\substack{l=1 \\ l\neq A}}^{N}(\delta_{l,A}\hat{h}_{l,A} + e_{l,A})\sqrt{P_l}s_{c_l}}_{\textit{interference from the group relays in other groups}} + n_0, \tag{5}$$

where h_{SI_A} is the self-interference channel at the head relay (U_A) from relaying the inter-group common stream and inner-group common stream, which follows $h_{SI_A} \sim \mathcal{CN}(0, \sigma_{SI_A}^2)$. While n_0 is the complex additive white Gaussian noise (AWGN) at the users with power spectral density height denoted by N_0, and P_A is the head relay (U_A) transmit power. To model the residual self-interference after cancellation at the full-duplex relays, we adopt a linear residual SI model, which captures the dominant component of the remaining interference and effectively reflects its impact on system degradation. This model is widely used in the full-duplex literature, as it provides a clean baseline for understanding the core behavior of self-interference [22]. Moreover, it aligns with practical transceiver architectures, where the combination of analog and digital cancellation suppresses most nonlinear distortion components, leaving a dominant linear residual term proportional to the transmitted signal. Therefore, the uncancelled self-interference can be accurately represented using a linear attenuation factor or an equivalent residual interference power term. When U_A receives the transmitted signal from the BS, it first decodes the inter-group common stream s_c by treating the inner-group common stream s_{c_g} and private streams s_k as interference. In order to simplify the representation of equations, let: $W = \| \hat{\mathbf{g}}_A^H \mathbf{p}_i \|^2$ represents the estimated channel of the private streams of the users, $X = \sigma_{e,A}^2 \| \mathbf{p}_i \|^2$ represents the channel estimation error of the private streams of the users, $Y = \| \hat{\mathbf{g}}_A^H \mathbf{p}_{c_j} \|^2$ represents the estimated channel of the inner-group common streams of the groups, $Z = \sigma_{e,A}^2 \| \mathbf{p}_{c_j} \|^2$ represents the channel estimation error of the inner-group common streams of the groups, and $INT_A = 2P_A \| h_{SI_A} \|^2 + \sum_{\substack{l=1 \\ l\neq A}}^{N} P_l(\| \delta_{l,A}\hat{h}_{l,A} \|^2 + \sigma_{l,A}^2) + N_0$ represents the residual interference at the head relay and the interference from the group relays in other groups due to the transmission of the inner-group common message. Thus, the achievable rate of decoding the inter-group common stream at the head relay user (U_A) is $R_A^C = \log_2(1 + SINR_A^C)$. Then, using (4) and (5), the $SINR_A^C$ is given by (6).

After decoding the inter-group common stream, successive interference cancellation (SIC) is applied to remove it from (5). However, due to imperfect SIC, the inter-group common stream is not perfectly removed, leaving residual

$$SINR_A^C = \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2}{\sum_{i=1}^{K}(W_i + X_i) + \sum_{j=1}^{G}(Y_j + Z_j) + \sigma_{e,A}^2 \parallel \mathbf{p}_c \parallel^2 + INT_A}, \tag{6}$$

$$SINR_{g,A}^C = \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_g} \parallel^2}{\sum_{i=1}^{K}(W_i + X_i) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_j + Z_j) + INT_A + \sigma_{e,A}^2 \parallel \mathbf{p}_{c_g} \parallel^2 + B} \tag{7}$$

$$SINR_{g,A}^P = \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_A \parallel^2}{\sum_{\substack{i=1 \\ i\neq A}}^{K}(W_i + X_i) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_j + Z_j) + INT_A + B + C} \tag{8}$$

$$SINR_n^{BS,C} = \frac{\parallel \hat{\mathbf{g}}_n^H \mathbf{p}_c \parallel^2}{\sum_{i=1}^{K}(W_{n,i} + X_{n,i}) + \sum_{j=1}^{G}(Y_{n,j} + Z_{n,j}) + INT_n + \sigma_{e,n}^2 \parallel \mathbf{p}_c \parallel^2}, \tag{9}$$

$$SINR_{g,n}^C = \frac{\parallel \hat{\mathbf{g}}_n^H \mathbf{p}_{c_g} \parallel^2}{\sum_{i=1}^{K}(W_{n,i} + X_{n,i}) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_{n,j} + Z_{n,j}) + INT_n + \sigma_{e,n}^2 \parallel \mathbf{p}_{c_g} \parallel^2 + D} \tag{10}$$

interference. This residual interference from the inter-group common stream affects the decoding of the inner-group common stream and the private stream of each user. Then the rate of decoding the inner-group common stream at the head relay (U_A) is given by $R_{g,A}^C = \log_2(1 + SINR_{g,A}^C)$, where $SINR_{g,A}^C$ is given by (7), where B represents the residual interference due to the imperfect SIC of the inter-group common stream and is given by $B = \beta \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2 + \beta\sigma_{e,A}^2 \parallel \mathbf{p}_c \parallel^2$, where β is the SIC imperfection factor. In practical receivers, successive interference cancellation (SIC) is inherently imperfect due to channel estimation errors, hardware impairments, and synchronization mismatches. To capture this effect, we model imperfect SIC using a residual interference factor $\beta \in (0, 1)$, which represents the fraction of the canceled signal power that remains after SIC. While $\beta = 0$ corresponds to ideal SIC, practical systems typically achieve interference suppression levels of about $70 - 90\%$, leading to $\beta \in [0.1, 0.3]$. In the numerical results of this work, we adopt $\beta = 0.2$, corresponding to an 80% suppression level, as a conservative and realistic choice, particularly suitable for full-duplex cooperative systems operating under imperfect CSI and asynchronous reception.

After decoding the inner-group common stream, SIC is used to remove the inner-group common stream from the received signal given in (5); hence, the achievable rate to decode the private stream at head relay (U_A) is given by $R_{g,A}^P = \log_2(1 + SINR_{g,A}^P)$, where $SINR_{g,A}^P$ is represented as (8), where C represents the residual interference due to the imperfect SIC of the inner-group common stream and is given by $C = \sigma_{e,A}^2 \parallel \mathbf{p}_A \parallel^2 + \beta \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_g} \parallel^2 + \beta\sigma_{e,A}^2 \parallel \mathbf{p}_{c_g} \parallel^2$.

4.2. Achievable Rate Analysis for Group Relays U_n

In the proposed FD C-RSMA, the group relays receive signals from the BS and the head relay. Then, these full-duplex group relays forward the inner-group common stream only to the rest of the users within the same group after decoding it. Therefore, the received signal at the group relay user (U_n) transmitted by BS is given by:

$$y_{g,n}^{BS} = \underbrace{(\hat{\mathbf{g}}_n^H + \mathbf{e}_n^H)\mathbf{x}}_{\textbf{\textit{desired signal from BS}}} + \underbrace{h_{SI_n}\sqrt{P_n}s_{c_g}}_{\textbf{\textit{self-interference from inner-common message}}} + \underbrace{\sum_{\substack{l=1 \\ l\neq n}}^{N}(\delta_{l,n}\hat{h}_{l,n} + e_{l,n})\sqrt{P_l}s_{c_l}}_{\textbf{\textit{interference from the group relays in other groups}}} + n_0, \tag{11}$$

while the received signal at the group relay user (U_n) transmitted by the head relay (U_A) is given by:

$$y_{g,n}^{A} = \underbrace{(\delta_{A,n}\hat{h}_{A,n} + e_{A,n})\sqrt{P_A}s_c}_{\textbf{\textit{relayed signal by head relay}}} + \underbrace{h_{SI_n}\sqrt{P_n}s_{c_g}}_{\textbf{\textit{self-interference from inner-common message}}} + \underbrace{\sum_{\substack{l=1 \\ l\neq n}}^{N}(\delta_{l,n}\hat{h}_{l,n} + e_{l,n})\sqrt{P_l}s_{c_l}}_{\textbf{\textit{interference from the group relays in other groups}}} + n_0, \tag{12}$$

where P_n is the group relay (U_n) transmit power, and h_{SI_n} is the self-interference channel at the group relay (U_n) which follows $h_{SI_n} \sim \mathcal{CN}(0, \sigma_{SI_n}^2)$. We assume that the signals from the BS, the head user relay U_A, and the group user relay (U_n) can be fully resolved at users so that they can be appropriately co-phased and merged by maximal ratio combining (MRC) [23]. In practical cooperative systems, signals transmitted by the base station and relays are rarely perfectly time-aligned due to propagation differences, hardware processing delays, and independent oscillators. These timing offsets manifest as phase rotations in narrowband systems, reducing coherent signal combining and limiting the achievable maximal-ratio combining (MRC) gain, particularly under imperfect CSI. Although practical

$$SINR_{g,n}^{P} = \frac{\parallel \hat{\mathbf{g}}_n^H \mathbf{p}_n \parallel^2}{\sum_{\substack{i=1 \\ i\neq n}}^{K}(W_{n,i}+X_{n,i}) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_{n,j}+Z_{n,j}) + INT_n + D + E} \tag{13}$$

$$R_m^C = \log_2\Bigg(1+\frac{\parallel \delta_{A,m}\hat{h}_{A,m} \parallel^2 P_A}{\sigma_{A,m}^2 P_A + \sum_{l=1}^{N} P_l(\parallel \delta_{l,m}\hat{h}_{l,m} \parallel^2 +\sigma_{l,m}^2) + N_0} + \frac{\parallel \hat{\mathbf{g}}_m^H \mathbf{p}_c \parallel^2}{\sum_{i=1}^{K}(W_{m,i}+X_{m,i}) + \sum_{j=1}^{G}(Y_{m,j}+Z_{m,j}) + \sum_{l=1}^{N} P_l(\parallel \delta_{l,m}\hat{h}_{l,m} \parallel^2 +\sigma_{l,m}^2) + \sigma_{e,m}^2 \parallel \mathbf{p}_c \parallel^2 + N_0}\Bigg), \tag{14}$$

$$R_{g,m}^C = \log_2\left(1 + \frac{\parallel \delta_{n,m}\hat{h}_{n,m} \parallel^2 P_n}{\sigma_{n,m}^2 P_n + INT_m} + \frac{\parallel \hat{\mathbf{g}}_m^H \mathbf{p}_{c_g} \parallel^2}{\sum_{i=1}^{K}(W_{m,i}+X_{m,i}) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_{m,j}+Z_{m,j}) + INT_m + \sigma_{e,m}^2 \parallel \mathbf{p}_{c_g} \parallel^2 + F}\right), \tag{15}$$

$$SINR_{g,m}^{P} = \frac{\parallel \hat{\mathbf{g}}_m^H \mathbf{p}_m \parallel^2}{\sum_{\substack{i=1 \\ i\neq m}}^{K}(W_{m,i}+X_{m,i}) + \sum_{\substack{j=1 \\ j\neq g}}^{G}(Y_{m,j}+Z_{m,j}) + INT_m + F + G} \tag{16}$$

MRC implementation must consider synchronization errors, pilot contamination, hardware impairments, and processing latency, it remains attractive due to its low complexity and robustness. Importantly, the proposed 2-layer FD C-RSMA architecture reduces the performance degradation caused by asynchronous reception by leveraging additional common streams for enhanced interference management, thereby preserving robustness and fairness under realistic timing misalignment conditions. By adopting MRC under non-ideal CSI and asynchronous conditions, it is important to emphasize that, although perfect time alignment is not assumed, the asynchronous components can still be coherently combined based on their estimated effective channels, which inherently include the residual phase rotations induced by timing offsets. As a result, asynchronous reception does not prevent signal combining but rather manifests as a reduction in the effective combining gain. This modeling choice reflects practical receiver operation, where residual synchronization errors are absorbed into the channel estimates and treated as part of the composite channel response used for MRC.

When the group user relay (U_n) receives the transmitted signal, it first decodes the inter-group common stream s_c by treating the inner-group common stream and private streams as interference. In order to simplify the representation of equations, let: $W_n{=}\parallel \hat{\mathbf{g}}_n^H \mathbf{p}_i \parallel^2$, $X_n{=}\sigma_{e,n}^2 \parallel \mathbf{p}_i \parallel^2$, $Y_n{=}\parallel \hat{\mathbf{g}}_n^H \mathbf{p}_{c_j} \parallel^2$, $Z_n{=}\sigma_{e,n}^2 \parallel \mathbf{p}_{c_j} \parallel^2$, and $INT_n{=}P_n \parallel h_{SI_n} \parallel^2 + \sum_{\substack{l=1 \\ l\neq n}}^{N} P_l(\parallel \delta_{l,n}\hat{h}_{l,n} \parallel^2 +\sigma_{l,n}^2) + N_0$. Therefore, using (11) and (12), the achievable rate of decoding the inter-group common stream at the group relay user (U_n) is given by

$$R_n^C = \log_2(1 + \underbrace{\frac{\parallel \delta_{A,n}\hat{h}_{A,n} \parallel^2 P_A}{\sigma_{A,n}^2 P_A + INT_n}}_{SINR_n^{A,C}} + SINR_n^{BS,C}), \tag{17}$$

where $SINR_n^{A,C}$ is the signal to interference plus noise ratio for the transmitted signal from the head relay user to the n_{th} group relay user. While $SINR_n^{BS,C}$ is the signal to interference plus noise ratio for the transmitted signal from the base station to the n_{th} group relay user, where $SINR_n^{BS,C}$ is given by (9), for $n \in \mathcal{N}$, and $n \neq A$. As mentioned before, the inter-group common stream is transmitted by the BS and the head relay, and then received by the group relays and the non-relay users, then it is removed by SIC from (11). Then the achievable rate of decoding the inner-group common stream at the group relay (U_n) is $R_{g,n}^C = \log_2(1 + SINR_{g,n}^C)$, where $SINR_{g,n}^C$ is represented as (10), where $D = \beta \parallel \hat{\mathbf{g}}_n^H \mathbf{p}_c \parallel^2 + \beta\sigma_{e,n}^2 \parallel \mathbf{p}_c \parallel^2$.

Finally, the $SINR_{g,n}^P$ can be obtained from (11) after decoding the inter-group common stream and inner-group common stream and removing it using SIC. $SINR_{g,n}^P$ is given by (13), where $E = \sigma_{e,n}^2 \parallel \mathbf{p}_n \parallel^2 + \beta \parallel \hat{\mathbf{g}}_n^H \mathbf{p}_{c_g} \parallel^2 + \beta\sigma_{e,n}^2 \parallel \mathbf{p}_{c_g} \parallel^2$. Therefore, the achievable rate of the private stream at the group relay (U_n) is given by $R_{g,n}^P = \log_2(1 + SINR_{g,n}^P)$.

4.3. Achievable Rate Analysis for Non-Relay Users U_m

Since we consider full-duplex communications system, each non-relay user receives multiple signals at the same time. These non-relay users receive the transmitted signal from BS, the inter-group common stream from the head relay user (U_A), and the inner-group common stream from the group relay users. Then, the received signal at the non-relay user (U_m) in group-g transmitted by BS is given

by:

$$y_{g,m}^{BS} = \underbrace{(\hat{\mathbf{g}}_m^H + \mathbf{e}_m^H)\mathbf{x}}_{\textbf{desired signal from BS}} + \underbrace{(\delta_{n,m}\hat{h}_{n,m} + e_{n,m})\sqrt{P_n}s_{c_g}}_{\textbf{relayed signal by group relay}} + \underbrace{\sum_{\substack{l=1 \\ l\neq n}}^{N}(\delta_{l,m}\hat{h}_{l,m} + e_{l,m})\sqrt{P_l}s_{c_l}}_{\textbf{interference from the group relays in other groups}} + n_0, \quad (18)$$

The received signal at the non-relay user (U_m) in group-g transmitted by head relay (U_A) is given by:

$$y_{g,m}^{A} = \underbrace{(\delta_{A,m}\hat{h}_{A,m} + e_{A,m})\sqrt{P_A}s_c}_{\textbf{relayed signal by head relay}} + n_0 + \underbrace{(\delta_{n,m}\hat{h}_{n,m} + e_{n,m})\sqrt{P_n}s_{c_g}}_{\textbf{relayed signal by group relay}} + \underbrace{\sum_{\substack{l=1 \\ l\neq n}}^{N}(\delta_{l,m}\hat{h}_{l,m} + e_{l,m})\sqrt{P_l}s_{c_l}}_{\textbf{interference from the group relays in other groups}}, \quad (19)$$

for $m, n \in g$. To simplify the representation of equations, let: $W_m = \| \hat{\mathbf{g}}_m^H \mathbf{p}_i \|^2$, $X_m = \sigma_{e,m}^2 \| \mathbf{p}_i \|^2$, $Y_m = \| \hat{\mathbf{g}}_m^H \mathbf{p}_{c_j} \|^2$, $Z_m = \sigma_{e,m}^2 \| \mathbf{p}_{c_j} \|^2$, and $INT_m = \sum_{\substack{l=1 \\ l\neq n}}^{N} P_l(\| \delta_{l,m}\hat{h}_{l,m} \|^2 + \sigma_{l,m}^2) + N_0$. Thus, using (18) and (19), the achievable rate of decoding the inter-group common stream at the non-relay user (U_m) is given by (14).

After decoding the inter-group common stream, it is removed from the received signals. Then, the rate of decoding the inner-group common stream at the non-relay user (U_m) is represented as (15), where $F = \beta \| \hat{\mathbf{g}}_m^H \mathbf{p}_c \|^2 + \beta\sigma_{e,m}^2 \| \mathbf{p}_c \|^2$. Furthermore, after decoding the inter-group common stream and inner-group common stream and removing them from the received signals, the achievable rate to decode the private stream at non-relay user (U_m) is given by $R_{g,m}^P = \log_2(1 + SINR_{g,m}^P)$. Given that $SINR_{g,m}^P$ is represented as (16), where $G = \sigma_{e,m}^2 \| \mathbf{p}_m \|^2 + \beta \| \hat{\mathbf{g}}_m^H \mathbf{p}_{c_g} \|^2 + \beta\sigma_{e,m}^2 \| \mathbf{p}_{c_g} \|^2$.

After determining the achievable rate for each user in the system, we proceed to calculate the total achievable rate for each user $R_{k,tot}$. But, to guarantee that each user can decode the inter-group common stream successfully, the achievable rate of the inter-group common stream should be $R_c = \min(R_1^C, R_2^C, ..., R_K^C)$, R_c is shared by all users in the system. As the inter-group common stream contains information from each user, thus it should satisfy $\sum_{i=1}^{K} a_i \leq R_c$, where a_i is the portion of the inter-group common stream allocated to each user.

Furthermore, in order to guarantee successful decoding of the inner-group common stream at each user in group-g, the achievable rate of the inner-group common stream should be $R_{c,g} = \min(R_{g,1}^C, R_{g,2}^C, ..., R_{g,Z}^C)$, $R_{c,g}$ is shared by users in group-g, thus it should satisfy $\sum_{i=1}^{Z} b_{g,i} \leq R_{c,g}$, where $b_{g,i}$ is the portion of the inner-group common stream allocated to each user in group-g. Hence, the total achievable rate at user-k is given by

$$R_{k,tot} = a_k + b_{g,k} + R_{g,k}^P, \quad \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \quad (20)$$

4.4. Optimization Problem

The objective is to maximize the minimum rate (Max-Min rate) of all users in order to ensure fairness among users; this can be done by jointly optimizing the precoding matrix $\mathbf{P}$, the inter-group common rate allocation $\boldsymbol{a}$, the inner-group common rate allocation $\mathbf{b}_g$, the head relay power P_A, the group relay power P_n. The optimization problem can be formulated as the following:

$$\max_{\mathbf{P},\boldsymbol{a},\boldsymbol{b}_g,P_A,P_n} \quad \min\ R_{k,tot} \quad (21)$$

$$s.t: \sum_{i=1}^{K} a_i \leq R_c, \quad (21a)$$

$$\sum_{i=1}^{Z} b_{g,i} \leq R_{c,g}, \forall g \in \mathcal{G} \quad (21b)$$

$$a_k \geq 0, \forall k \in \mathcal{K}, \quad (21c)$$

$$b_{g,k} \geq 0, \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \quad (21d)$$

$$tr(\mathbf{P}\mathbf{P}^H) \leq P_t, \quad (21e)$$

$$\frac{L_k}{R_{k,tot}} \leq t_{max}, \quad (21f)$$

$$P_A \leq P_{A,max}, \quad (21g)$$

$$P_n \leq P_{n,max}, \forall n \in \mathcal{N}, n \neq A \quad (21h)$$

where constraint (21a) guarantees that each user is able to decode the inter-group common stream. Constraint (21b) guarantees that each user in group-g is able to decode its inner-group common stream. Constraints (21c) and (21d) show that the portion of the inter-group and inner-group common rate allocated to each user should be a positive value, while (21e), (21g), and (21h) refer to the power budget at the BS, the head relay, and the group relay, respectively, and (21f) guarantees the user's requirement of transmission latency, where L_k is the normalized data size, and t_{max} is the maximum acceptable latency.

The objective function, constraints (21a), (21b), and (21f) are non-convex. Accordingly, let v represent the minimum rate of all users; then the optimization problem (21) can be written as

$$\max_{\mathbf{P},\boldsymbol{a},\boldsymbol{b}_g,P_A,P_n,v} \quad v \quad (22)$$

$$s.t: R_{g,k}^P + b_{g,k} + a_k \geq v, \quad \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \quad (22a)$$

$$R_{g,k}^P + b_{g,k} + a_k \geq \frac{L_k}{t_{max}}, \quad \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \quad (22b)$$

$$(21a), (21b), (21c), (21d), (21e), (21g), (21h) \quad (22c)$$

Constraint (22a) states that the minimum rate (v) should be less than the total achievable rate ($R_{k,tot}$) for each user. Constraint (22b) is obtained after substituting (20) in ($21f$).

5. Solution of the Optimization Problem

In this section, we focus on optimizing the precoding matrix, the inter-group common rate allocation, the inner-group common rate allocation, the head relay power, and the group relay power to maximize the minimum rate (v). But, since (22) is non-convex problem, we propose an alternative optimization (AO) algorithm based on weighted minimum mean square error (WMMSE) to solve this optimization problem. The complete procedure is summarized in **Algorithm** 1, which iteratively optimizes the precoding matrix, rate and power allocation until convergence. The detailed analysis of this algorithm is provided in Appendix A for each type of users: the head relay (U_A), the group relay (U_n), and the non-relay users (U_m).

5.1. WMMSE Reformulation

The mean square error (MSE) of the inter-group common stream, the inner-group commen stream, and private stream at the head relay (U_A) is defined respectively as

$$\epsilon_A^C = \mathbb{E}[\| \ q_A^C y_A - s_c \ \|^2] \tag{23}$$

$$\epsilon_{g,A}^C = \mathbb{E}[\| \ (q_{g,A}^C (y_A - (1-\sqrt{\beta})(\hat{\mathbf{g}}_A^H + \mathbf{e}_A^H)\mathbf{p}_c s_c) - s_{c_g} \ \|^2] \tag{24}$$

$$\begin{aligned}\epsilon_{g,A}^P = \mathbb{E}[\| \ (q_{g,A}^P (y_A - (1-\sqrt{\beta})(\hat{\mathbf{g}}_A^H + \mathbf{e}_A^H)\mathbf{p}_c s_c \\ - (1-\sqrt{\beta})(\hat{\mathbf{g}}_A^H + \mathbf{e}_A^H)\mathbf{p}_{c_g} s_{c_g})) - s_A \ \|^2]\end{aligned} \tag{25}$$

where q_A^C, $q_{g,A}^C$, and $q_{g,A}^P$ are the scalar equalizer of the inter-group common stream, the inner-group common stream, and the private stream at the head relay (U_A), respectively. In order to achieve the minimum mean square error (MMSE), we should obtain the optimal equalizers. This can be done by differentiating the MSE equations with respect to the scaler equalizer and equate them to zero. The MMSE of the three streams of head relay user can be obtained by substituting the obtained optimal equalizers in the MSE equation. Consequently, the achievable rate of the inter-group common stream, the inner-group common stream, and the private stream at the head relay user (U_A) is given respectively by

$$R_A^C = log_2(1 + SINR_A^C) = -log_2(\tilde{\epsilon}_A^C) \tag{26}$$

$$R_{g,A}^C = -log_2(\tilde{\epsilon}_{g,A}^C) \tag{27}$$

$$R_{g,A}^P = -log_2(\tilde{\epsilon}_{g,A}^P) \tag{28}$$

where $\tilde{\epsilon}_A^C$, $\tilde{\epsilon}_{g,A}^C$, and $\tilde{\epsilon}_{g,A}^P$ are the MMSE of the inter-group common stream, the inner-group common stream, and the private stream respectively. Meanwhile, the equation of the augmented weighted mean square error (WMSE) is generally stated in [24], from which we can write that the augmented WMSE of the inter-group common stream, the inner-group common stream, and the private stream for the head relay can be expressed respectively as

$$Z_A^C = w_A^C \tilde{\epsilon}_A^C - log_2(w_A^C) \tag{29}$$

$$Z_{g,A}^C = w_{g,A}^C \tilde{\epsilon}_{g,A}^C - log_2(w_{g,A}^C) \tag{30}$$

$$Z_{g,A}^P = w_{g,A}^P \tilde{\epsilon}_{g,A}^P - log_2(w_{g,A}^P) \tag{31}$$

Consequently, the WMMSE of the head relay's inter-group common stream is given by $\tilde{Z}_A^C = 1 - R_A^C$, while the WMMSE of the inner-group common stream and private stream for the head relay user can be expressed, respectively, as $\tilde{Z}_{g,A}^C = 1 - R_{g,A}^C$, and $\tilde{Z}_{g,A}^P = 1 - R_{g,A}^P$.

While for the group relay, the WMMSE of the inter-group common stream, inner-group common stream, and private stream of the head relay can be written respectively as

$$\tilde{Z}_n^C \approx 2 - R_n^C \tag{32}$$

$$\tilde{Z}_{g,n}^C = 1 - log_2(\frac{1}{\tilde{\epsilon}_{g,n}^C}) = 1 - R_{g,n}^C \tag{33}$$

$$\tilde{Z}_{g,n}^P = 1 - log_2(\frac{1}{\tilde{\epsilon}_{g,n}^P}) = 1 - R_{g,n}^P \tag{34}$$

where R_n^C, $R_{g,n}^C$, and $R_{g,n}^P$ are the achievable rates of the inter-group common stream, the inner-group common stream, and the private stream of the group relay.

For the non-relay user, we follow the same methodology used for the head relay user and the group relay users. As mentioned before, each non-relay user receives three streams: the superimposed message from the BS, the relayed inter-group common message from the head relay, and the relayed inner-group common message from the group relay. Therefore, MRC is performed when calculating the achievable rate for inter-group common stream and inner-group common stream. Hence, the WMMSE of them and the private stream can be written as

$$\tilde{Z}_m^C = 2 - log_2(\frac{1}{\tilde{\epsilon}_m^{BS,C}\tilde{\epsilon}_m^{A,C}}) \approx 2 - R_m^C \tag{35}$$

$$\tilde{Z}_{g,m}^C = 2 - log_2(\frac{1}{\tilde{\epsilon}_{g,m}^{BS,C}\tilde{\epsilon}_{g,m}^{n,C}}) \approx 2 - R_{g,m}^C \tag{36}$$

$$\tilde{Z}_{g,m}^P = 1 - log_2(\frac{1}{\tilde{\epsilon}_{g,m}^P}) = 1 - R_{g,m}^P \tag{37}$$

where $\tilde{\epsilon}_m^{BS,C}$, $\tilde{\epsilon}_m^{A,C}$, $\tilde{\epsilon}_{g,m}^{BS,C}$, $\tilde{\epsilon}_{g,m}^{n,C}$, and $\tilde{\epsilon}_{g,m}^P$ are the MMSE of the inter-group common stream due to the BS, the inter-group common stream due to the head relay, the inner-group common stream due to the BS, the inner-group common stream due to the head relay, and the private stream at the non-relay users, respectively.

5.2. Reformulation of the Optimization Problem

Now, the objective is to minimize the the WMMSE instead of maximizing the minimum rate. Thus, the optimization problem (21) can be reformulated in term of

WMMSE as follows

$$\min_{\mathbf{P},P_A,P_n,\boldsymbol{x},\boldsymbol{y}_g,\tilde{\boldsymbol{q}},\boldsymbol{w}.\eta} \eta \tag{38}$$

$$s.t: \tilde{Z}^P_{g,k} + y_{g,k} + x_k \leq 1 - \eta, \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \tag{38a}$$

$$\tilde{Z}^P_{g,k} + y_{g,k} + x_k \leq 1 - \frac{L_k}{t_{max}}, \forall k \in \mathcal{K}, \forall g \in \mathcal{G} \tag{38b}$$

$$\sum_{i=1}^{K} x_i + 1 \geq \tilde{Z}^C_k, \text{for } \ k = A, \tag{38c}$$

$$\sum_{i=1}^{K} x_i + 2 \geq \tilde{Z}^C_k, \text{for } \ k \neq A, \tag{38d}$$

$$\sum_{i=1}^{Z} y_{g,i} + 1 \geq \tilde{Z}^C_{g,k}, \forall g \in \mathcal{G}, i \in g, k \in \mathcal{N}, \tag{38e}$$

$$\sum_{i=1}^{Z} y_{g,i} + 2 \geq \tilde{Z}^C_{g,k}, \text{for } \ k \notin \mathcal{N}, \tag{38f}$$

$$x_k \leq 0, \forall k \in \mathcal{K}, \tag{38g}$$

$$y_{g,k} \leq 0, \forall k \in \mathcal{K}, \forall g \in \mathcal{G}, \tag{38h}$$

$$tr(\mathbf{P}\mathbf{P}^H) \leq P_t, \tag{38i}$$

$$P_A \leq P_{A,max}, \tag{38j}$$

$$P_n \leq P_{n,max}, \forall n \in \mathcal{N}, n \neq A \tag{38k}$$

where η is the maximum WMMSE of all users, $\boldsymbol{x} = \boldsymbol{-a}$, and $\boldsymbol{y}_g = \boldsymbol{-b}_g$. Also, $\tilde{\boldsymbol{q}}$=[$\tilde{q}^C_A$, $\tilde{q}^C_{g,A}$, $\tilde{q}^P_{g,A}$, $\tilde{q}^{BS,C}_n$, $\tilde{q}^{A,C}_n$, $\tilde{q}^C_{g,n}$, $\tilde{q}^P_{g,n}$, $\tilde{q}^{BS,C}_m$,$\tilde{q}^{A,C}_m$, $\tilde{q}^{BS,C}_{g,m}$, $\tilde{q}^{n,C}_{g,m}$, $\tilde{q}^P_{g,n}$], $\forall n \in \mathcal{N}, \forall m \in \mathcal{M}$, and $\boldsymbol{w}$ = [$w^C_A, w^C_{g,A}$, $w^P_{g,A}, w^{BS,C}_n, w^{A,C}_n, w^C_{g,n}, w^P_{g,n}, w^{BS,C}_m w^{A,C}_m$, $w^{BS,C}_{g,m}, w^{n,C}_{g,m}, w^P_{g,n}$], $\forall n \in \mathcal{N}, \forall m \in \mathcal{M}$ represent the equalizers and weights associated with MMSE, respectively. Furthermore, $\tilde{Z}^C_k$, $\tilde{Z}^C_{g,k}$, and $\tilde{Z}^P_{g,k}$ are substituted by their derived expression shown in the previous analysis. The problem (38) is not jointly convex over all the optimization variables $(\mathbf{P}, P_A, P_n, \boldsymbol{x}, \boldsymbol{y}_g, \tilde{\boldsymbol{q}}, \boldsymbol{w})$. To address this, the problem is decomposed into two sub-problems. In the first part, the optimal equalizers $\tilde{\boldsymbol{q}}$ and the optimal weights $\boldsymbol{w}$ are computed using the derived closed forms while the values of $(\mathbf{P}, P_A, P_n, \boldsymbol{x}, \boldsymbol{y}_g)$ are given initially. In the second part, the calculated values of the equalizers $\tilde{\boldsymbol{q}}$ and the weights $\boldsymbol{w}$ are used to optimize $(\mathbf{P}, P_A, P_n, \boldsymbol{x}, \boldsymbol{y}_g)$ using (38). Now, the problem (38) is convex and can be solved iteratively using the proposed algorithm shown in **Algorithm** 1. The algorithm starts by initializing the precoding matrix $(\mathbf{P})$, then this initialization is used to calculate the optimal equalizers $(\tilde{\boldsymbol{q}})$ using the derived equations. After getting the optimal equalizers, the optimal MMSE $(\tilde{\epsilon})$ for all streams are also calculated using the derived equation, and then these MMSEs are substituted in the weights $(\boldsymbol{w})$ equations to get their optimal values that minimize the WMSE, so WMMSE is reached. Given $(\tilde{\boldsymbol{q}})$ and $(\boldsymbol{w})$, the problem (38) is solved to get the optimized values for $(\mathbf{P}, P_A, P_n, \boldsymbol{x}, \boldsymbol{y}_g)$, and the objective function. The algorithm continues iteratively updating these variables till convergence, this corresponds to **steps** 5 and 6 in **Algorithm** 1 and is repeated in **steps** 14 and 15. Then, we compare the new optimized Max-Min rate (objective function) with the previous value if it is greater than a certain tolerance value (κ), this process will be repeated until convergence. The overall optimization procedure is illustrated in the flowchart shown in Figure 2, which summarizes the main update steps of the proposed algorithm. The proposed algorithm builds directly on the traditional WMMSE framework but extends it significantly to address a more complex and realistic system model. The application context, system model, and optimization variables are novel and more comprehensive compared to traditional uses of WMMSE. A brief comparison between the proposed algorithm and the traditional one is shown in Table 5.

5.3. Convergence and Complexity

The proposed WMMSE-based alternating optimization algorithm is guaranteed to converge due to its block-wise monotonic structure. When the equalizers and weights are updated, the WMMSE of each stream is minimized, which is equivalent to increasing the corresponding achievable rates and therefore cannot degrade the global objective. When the precoding matrix and rate-allocation variables are optimized, the total WMMSE is further reduced, which directly corresponds to improving or maintaining the minimum user rate. This ensures that the objective value is non-increasing across iterations. Since the WMMSE metric is lower-bounded by zero and the max-min rate is upper-bounded due to finite transmit power constraints, the resulting sequence is both monotonic and bounded, and is therefore guaranteed to converge. To ensure both computational efficiency and solution accuracy, the algorithm iteratively updates the variables until the improvement in the Max-Min rate is negligible. Numerical results in Figure 7 confirm this behavior, showing fast and smooth convergence within a small number of iterations. While for the time complexity per AO iteration, it is dominated by the convex precoding matrix and rate-allocation update in (38). The decision dimension of this problem is dominated by the precoding matrix $\mathbf{P} \in \mathbb{C}^{N_t \times (K+G+1)}$ leading to $O(N_t(K+G))$ real optimization variables. Using second-order cone, the per-iteration complexity is therefore on the order of $O((N_t(K+G))^{3.5})$ The equalizer and weight updates only require vector–matrix operations with complexity $O(N_t(K+G)^2)$ which is negligible compared to the convex-solver cost. Hence, the overall complexity of the proposed algorithm is approximately $O(I_{AO}(N_t(K+G))^{3.5})$, where I_{AO} is the number of AO iterations required for convergence. This makes the algorithm computationally efficient and suitable for near-real-time operation. As for practical setups, convergence occurs within about 20 iterations, requiring less than 10^6 operations. On modern DSP or GPU hardware, this corresponds to under 1 ms processing time, well within 6G latency limits. Therefore, the algorithm can adapt dynamically to channel variations at the frame level while maintaining the latency constraint. In addition to the above time complexity analysis, the space complexity of the proposed algorithm is dominated

Table 4: Time and space complexity comparison of different multiple access schemes.

Aspect	Per-Iteration Time Complexity	Space Complexity
Proposed System	$O((N_t(K+G))^{3.5})$	$O(KN_t) + O(N_t(K+G)) + O(K(K+G))$
1-layer RSMA	$O((N_t(K+1))^{3.5})$	$O(KN_t) + O(N_t(K+1)) + O(K^2)$
Traditional WMMSE-SDMA	$O((N_tK)^{3.5})$	$O(KN_t) + O(N_tK) + O(K^2)$
NOMA	$O(K^{3.5}) + O(K^2)$	$O(K^2)$

by storing the BS–user channel estimates and the precoding matrix. Specifically, the channel estimates require $O(KN_t)$ memory, while the precoding matrix $\mathbf{P} \in \mathbb{C}^{N_t \times (K+G+1)}$ requires $O(N_t(K+G))$ storage, and the auxiliary WMMSE variables (equalizers and weights) scale as $O(K(K+G))$. Hence, the overall memory requirement grows polynomially with the system dimensions and remains practical for moderate-to-large values of K and N_t.

Compared with conventional WMMSE-based SDMA schemes, whose per-iteration complexity scales as $O((N_t K)^{3.5})$, and 1-layer RSMA with complexity $O((N_t(K+1))^{3.5})$, the proposed two-layer cooperative RSMA framework scales as $O((N_t(K+G))^{3.5})$, where the effective stream dimension increases from K to $K+G+1$ due to the additional group-wise common messages. Since the number of groups G is typically much smaller than the number of users K, the additional computational and memory overhead remains moderate and scales polynomially with system size. Although NOMA exhibits lower optimization dimensionality, it lacks the robustness and fairness guarantees of layered RSMA and suffers from strong SIC coupling complexity. A complexity comparison with SDMA, one-layer RSMA, and NOMA is summarized in Table 4, confirming that despite incorporating full-duplex relaying, asynchronous reception, imperfect CSI, and imperfect SIC, the proposed framework remains computationally scalable and suitable for real-time 6G networks.

5.4. Practical Impairments Sensitivity

The practical impairments have a significant impact on the overall system performance. The first impairment is the SIC imperfection factor (β). The Max-Min rate is inversely proportional to β; as β increases, residual interference from imperfectly decoded common streams (both inter-group and intra-group) increases, thereby reducing the SINR of the desired streams. The second impairment, the channel estimation error variance ($\sigma^2_{e,k}$), also exhibits an inverse relationship with the Max-Min rate. Imperfect CSI leads to inaccurate precoding and unmanaged multi-user interference, with the resulting errors propagating through all SINR computations. The results further reveal that imperfect CSI has the dominant effect on system performance. The final impairment, the phase shift due to asynchronous reception ($\delta_{n,k}$), causes the Max-Min rate to decrease, and its adverse effect becomes more pronounced for smaller normalized data sizes (L_k). A smaller L_k amplifies the effective phase shift, resulting in greater distortion and a more significant performance degradation. All these observations are validated by the simulation results in section 6, which clearly demonstrate the sensitivity of the proposed system to these practical impairments.

Algorithm 1: Proposed optimization algorithm based on WMMSE

Given : tolerance κ
Initialize : $\mathbf{P}^{(0)}, v^{(0)} = 0$;

$l = 0$;
calculate $\tilde{\boldsymbol{q}}^{(0)}$ using $\mathbf{P}^{(0)}$;
calculate $\tilde{\epsilon}^{(0)}$ using $\tilde{\boldsymbol{q}}^{(0)}$;
calculate $\boldsymbol{w}^{(0)}$ using $\tilde{\epsilon}^{(0)}$;
Solve (38) using $\mathbf{P}^{(0)}$,$\tilde{\boldsymbol{q}}^{(0)}$,$\boldsymbol{w}^{(0)}$;
Find the optimal value for the objective function η^* and optimal variables $\mathbf{P}^{(*)}, P_A^{(*)}, P_n^{(*)}, \boldsymbol{x}^{(*)}, \boldsymbol{y}_g^{(*)}$;
$l = l + 1$;
Update the old variable $\mathbf{P}^{(1)} \leftarrow \mathbf{P}^{(*)}$, $P_A^{(1)} \leftarrow P_A^{(*)}$, $P_n^{(1)} \leftarrow P_n^{(*)}$, $\boldsymbol{x}^{(1)} \leftarrow \boldsymbol{x}^{(*)}$, $\boldsymbol{y}_g^{(1)} \leftarrow \boldsymbol{y}_g^{(*)}$;
Let $\boldsymbol{a}^{(1)} = \text{-}\boldsymbol{x}^{(1)}$, $\boldsymbol{b}_g^{(1)} = \text{-}\boldsymbol{y}_g^{(1)}$;
Update $v^{(1)} = \min\{(1 - \tilde{Z}_{g,k}^{P^{(1)}}) + a_k^{(1)} + b_{g,k}^{(1)}\}$ $\forall k \in \mathcal{K}$;
while $\mid v^{(l)} - v^{(l-1)} \mid > \kappa$ **do**
  $l = l + 1$;
  calculate $\tilde{\boldsymbol{q}}^{(l)}$ using $\mathbf{P}^{(l-1)}$, $\tilde{\epsilon}^{(l)}$ using $\tilde{\boldsymbol{q}}^{(l)}$, and $\boldsymbol{w}^{(l)}$ using $\tilde{\epsilon}^{(l)}$;
  Solve (38) using $\mathbf{P}^{(l-1)}$,$\tilde{\boldsymbol{q}}^{(l)}$,$\boldsymbol{w}^{(l)}$;
  Find the optimal value for the objective function η^* and optimal variables $\mathbf{P}^{(*)}, P_A^{(*)}, P_n^{(*)}, \boldsymbol{x}^{(*)}, \boldsymbol{y}_g^{(*)}$;
  Update the old variables: $\mathbf{P}^{(l)} \leftarrow \mathbf{P}^{(*)}$, $P_A^{(L)} \leftarrow P_A^{(*)}$, $P_n^{(l)} \leftarrow P_n^{(*)}$, $\boldsymbol{x}^{(l)} \leftarrow \boldsymbol{x}^{(*)}$, $\boldsymbol{y}_g^{(l)} \leftarrow \boldsymbol{y}_g^{(*)}, v^{(l)} \leftarrow v^{(*)}$;
end

5.5. Practical Applications

The proposed system is inherently suited to meet key 6G requirements such as massive connectivity, high spectral efficiency, and reliable operation under practical impairments, making it a strong candidate for integration into emerging 6G technologies like collaborative vehicle links and 6G relay-assisted IoT. For collaborative vehicle links: V2X communication faces key challenges such as asynchronous reception, where signals from multiple relay vehicles and roadside units arrive at different times due to varying distances. In addition, imperfect CSI and SIC, resulting from rapidly changing channel conditions and hardware limitations. The proposed full-duplex multi-user

Table 5: Comparison Between Traditional WMMSE and Proposed WMMSE-Based Algorithm

Aspect	**Traditional WMMSE Algorithm**	**Proposed WMMSE-Based Algorithm**	**Similarity / Difference**
Objective Function	Minimizes the weighted sum mean square error (WMSE) to indirectly maximize sum-rate or fairness metrics.	Minimizes the maximum WMMSE (η) across all users to achieve Max–Min rate fairness under latency and power constraints.	Similar principle but different fairness-oriented objective.
Application Context	Often used for single-layer MIMO, MU-MISO.	Designed for Full-duplex 2-layer cooperative RSMA system with multiple relay and practical impairments.	Extended to multi-hop, multi-layer, imperfect-channel environment.
Convergence	Well-known to converge to a local optimum.	Shown to converge within a few iterations.	Same iterative convergence principle.
Optimization Variables	Typically optimizes precoding matrices, equalizers, weights and user rates.	Jointly optimizes: precoding matrix, equalizers, weights, inter/inner-group common rates, head relay power, and group relay power.	Extended variable set to account for relaying, layered RSMA, and practical impairments.
Fairness Consideration	Usually sum-rate focused; fairness not explicitly enforced.	Explicitly maximizes minimum user rate (Max-Min fairness) under latency constraints.	Strong fairness guarantee integrated into optimization.
Practicality	Limited to ideal or simplified scenarios.	Designed for realistic 6G conditions (asynchrony, CSI/SIC imperfections, full-duplex relaying).	Aligned with practical deployment challenges.

2-layer C-RSMA system is designed to effectively mitigate these challenges, maintaining reliable performance even under such non-ideal conditions. Furthermore, it enhances road safety by ensuring that all vehicles, including those with weaker signal reception, receive critical coordination and safety messages promptly and reliably. While for 6G relay-assisted IoT: The proposed full-duplex multi-user 2-layer C-RSMA system offers significant enhancements for IoT networks by improving spectral and energy efficiency, which are crucial for large-scale device connectivity in 6G environments. By enabling full-duplex relaying, spectral utilization increases effectively and the communication latency is reduced, which is vital for real-time IoT applications. The 2-layer cooperative framework further allows stronger devices or access points to assist weaker nodes through intelligent rate-splitting, ensuring fairness and reliable data delivery under challenging channel conditions. Additionally, by explicitly considering imperfect CSI, imperfect SIC, and asynchronous reception, the system maintains robust performance in dynamic and realistic IoT scenarios with mobile devices and rapidly varying channels. Consequently, the proposed approach is highly suitable for practical 6G IoT applications such as smart cities, industrial automation, and connected healthcare.

6. Numerical Results

In this section, numerical results are obtained to evaluate the performance of the proposed full-duplex 2-layer multi-user cooperative rate splitting system with imperfect CSI and imperfect SIC. The performance is measured for synchronous and asynchronous reception. In the simulation, the users are randomly distributed, where the one with the best channel condition acts as the head relay (U_A). We assume that the power of the head relay does not equal the power of the group relays $P_A \neq P_n$, as the head relay forwards the inter-group common stream to all users in the system, so it requires higher power than that of the group relay that relays the stream to users in its group only. The simulation parameters are given in Table 6, and the simulation is done using MATLAB software and YALMIP optimization tool. The simulation parameters in Table 6 are chosen to reflect realistic 6G deployment scenarios while maintaining consistency with established literature for fair comparison. The BS transmit power range from -10 dB to 10 dB captures both noise-limited and interference-limited regimes, enabling comprehensive evaluation of system behavior across operating conditions. The channel estimation error variance ($\sigma_{e,k} = -40$ dB) represents moderately accurate CSI achievable with practical

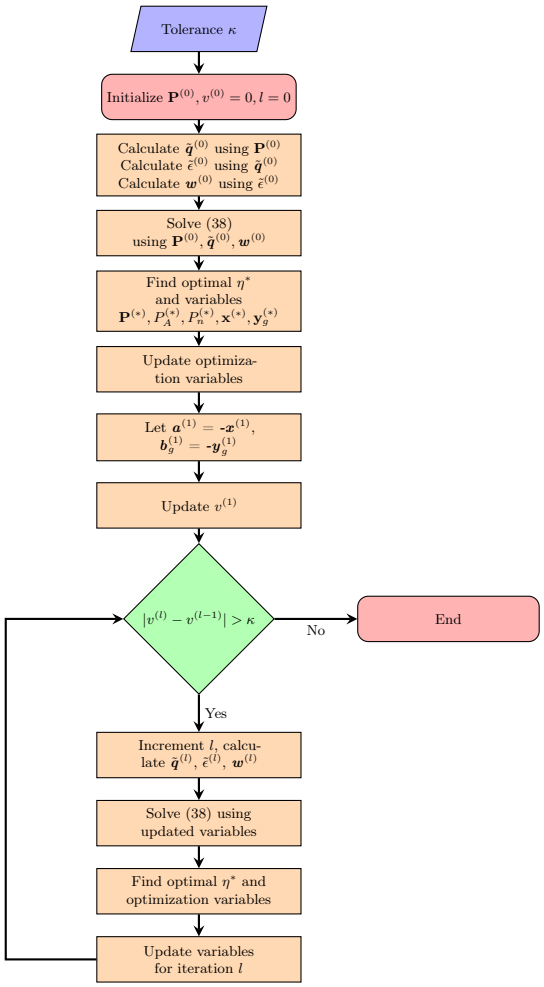

Figure 2: Flowchart of the proposed optimization algorithm

pilot-based estimation in urban macro-cell environments, while the SIC imperfection factor ($\beta = 0.2$) corresponds to 80% interference suppression which is a realistic target for modern transceivers given hardware limitations and estimation errors. Self-interference power ($\sigma_{SI} = -30$ dB) models the residual after analog and digital cancellation in full-duplex systems, consistent with state-of-the-art cancellation capabilities. The normalized data size ($L_k = -60$ dB) and latency threshold ($t_{max} = 1$ ms) align with 6G ultra-reliable low-latency communication requirements.

We compare the performance of the proposed system with the performance of different relaying systems that are found in literature [18]. The performance is measured in terms of the maximum minimum rate (Max-Min). The Max-Min rate is optimized using Algorithm 1 which optimizes the precoding matrix, inter-group common rate allocation, inner-group common allocation, and relays power. The performance of the proposed system is compared with the following systems:

- **2-layer without inter-group relaying:** 2-layer C-RSMA relays inner-group common stream only with both imperfect CSI and SIC [18].
- **1-layer without inter-group stream:** 1-layer C-RSMA relays inner-group common stream only with both imperfect CSI and SIC. In this 1-layer C-RSMA, the message of each user is divided into two parts: an inner-group common message and a private message. Each user decodes its inner-group common stream by treating the other inner-group common streams and private streams as interference, then each user decodes its own private stream by treating the other private streams and the other inner-group common streams except its own group as interference. There is no head relay, only group relays that relay the inner-group common stream only to the users in the same group [18].
- **1-layer:** 1-layer C-RSMA relays inter-group common stream only with both imperfect CSI and imperfect SIC. This is the default 1-layer C-RSMA system where the message of each user is divided into a common message and a private message. Each user decodes the common stream first by treating the private streams as interference, then each user decodes its own private stream by treating the other private streams as interference. There are no group relays, only a head relay that relays the common stream to all users [18].
- **Proposed 2-layer system Asynch:** 2-layer C-RSMA relays both inter-group common stream and inner-group common stream with both imperfect CSI and imperfect SIC with asynchronous reception.
- **2-layer without inter-group relaying Asynch:** 2-layer C-RSMA relays inner-group common stream only with both imperfect CSI and imperfect SIC with asynchronous reception.

The notation that appears in the legends of the figures is the following.

- **ICSI**: imperfect channel state information.
- **PCSI**: perfect channel state information.
- **ISIC**: imperfect successive interference cancellation.
- **PSIC**: perfect successive interference cancellation.

Figure 3 shows the Max-Min rate of the different synchronous systems using ICSI and ISIC versus BS transmitted power P_t. The proposed system demonstrates superior performance compared to all other schemes, achieving a 5.8% improvement when $P_t = 6dB$ over the 2-layer without inter-group relaying system. This gain is attributed to the relaying of both the inter-group and inner-group common streams, which jointly enhance interference management and the rate of the weakest users, thereby overcoming the limitation of the conventional RSMA systems. The figure also shows that the proposed system outperforms the other systems described in [18].

Table 6: List of Simulation Parameters

Parameter	Value
BS antenna, N_t	2
Power transmitted, P_t	6 dB
Head relay power, $P_{A,max}$	0 dB
Group relay power, $P_{n,max}$	-5 dB
AWGN power, N_0	-40 dB
Channel error variance, $\sigma^2_{e,k}$	-40 dB "unless stated"
Normalized data size, L_k	-60 dB
SIC imperfection factor, β	0.2 "unless stated"
SI power, σ^2_{SI}	-30 dB
Latency threshold, t_{min}	1ms
Path-loss exponent	2
Tolerance, κ	0.001
Number of users, K	6

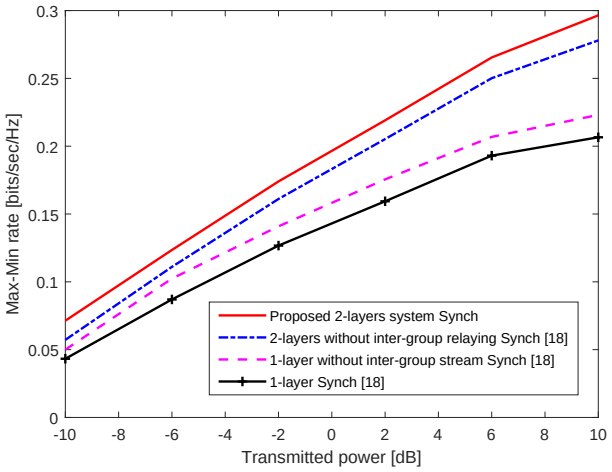

Figure 3: Max-Min Rate vs. BS transmitted power for synchronous cooperative schemes

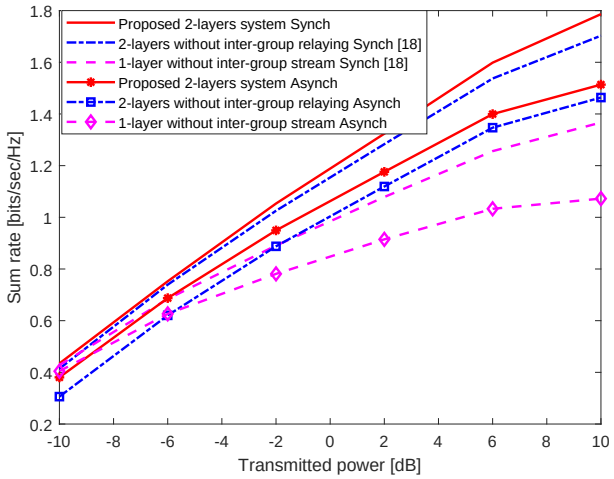

Figure 4: Sum Rate vs. BS transmitted power for synchronous and asynchronous cooperative schemes

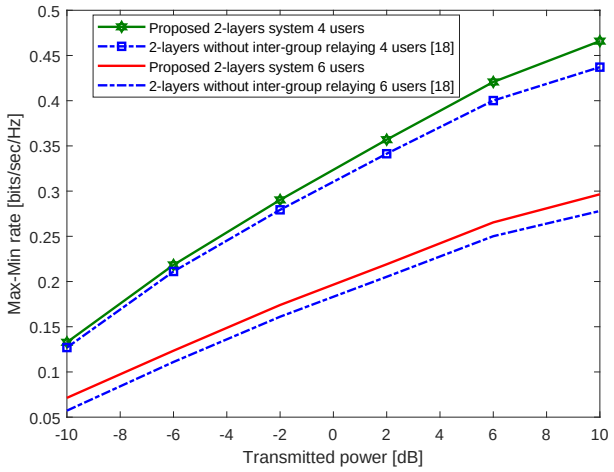

Figure 5: Performance comparison of 4-user and 6-user scenarios for different systems under synchronous reception.

We examine the performance of the proposed system in term of another parameter which is achievable sum rate as shown in Figure 4. Figure 4 shows the achievable sum rate of different systems under synchronous and asynchronous reception versus BS transmitted power P_t. It is shown that the proposed system outperforms all the other systems in both synchronization settings. This performance gain highlights the benefit of introducing an additional common layer combined with inter-group relaying, which enhances interference management and spectral efficiency. It is also noticeable that the systems with asynchronous reception have lower rate when compared with their corresponding ones with synchronous reception due to increasing the interference.

Figure 5 shows the Max-Min rate versus BS transmitted power under synchronous reception with ICSI and ISIC. It compares the performance of the proposed system and 2-layer without inter-group relaying system for $K = 4$ and $K = 6$ users. It is shown that the systems with 6 users have lower rate than their corresponding ones with 4 users due to increasing the multi-user interference.

Figure 6 shows the Max-Min rate versus BS transmitted power. It compares the performance of the proposed system and 2-layer without inter-group relaying system in case of synchronous and asynchronous reception. It can be seen that the systems with asynchronous reception have lower rate when compared with their corresponding ones with synchronous reception. This is because asynchronous reception causes inter-symbol interference, which increases the error in detecting the received signal, and hence degrades the system performance.

Figure 7 shows the convergence of the proposed algorithm for different systems in case of synchronous and asynchronous reception. It plots the Max-Min rate versus the number of iterations. It can be seen that all the mentioned systems converge within the first 6 iterations.

Figure 8 compares the performance of the proposed system with several benchmark schemes; namely non-orthogonal multiple access (NOMA) and space division multiple access

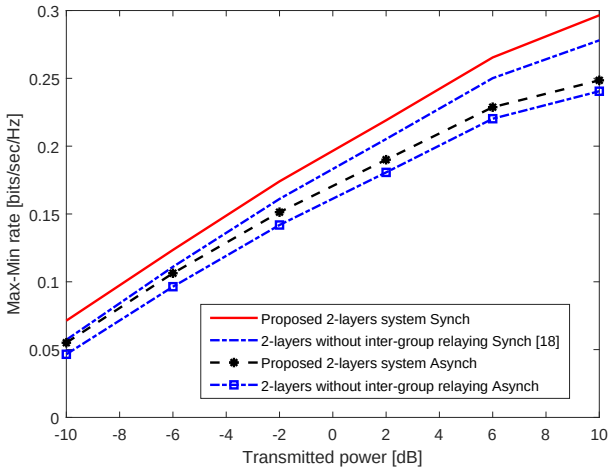

Figure 6: Impact of asynchronous reception on Max-Min rate performance

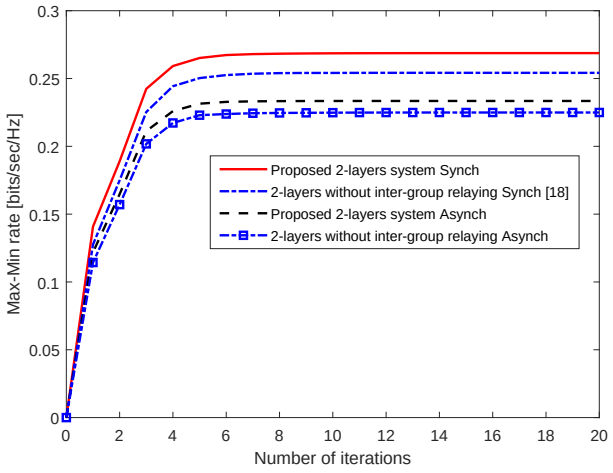

Figure 7: Convergence behavior of the proposed algorithm for different system configurations

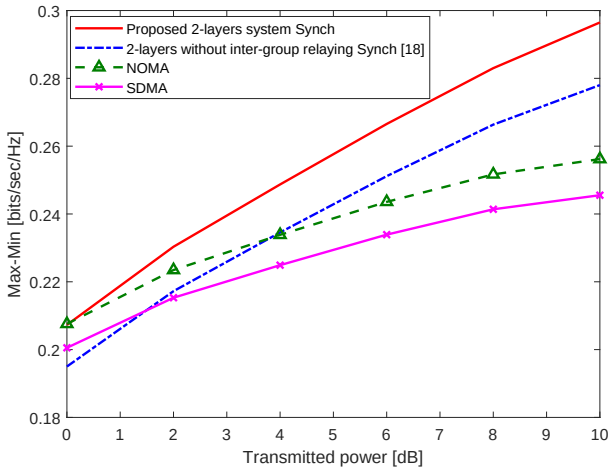

Figure 8: Max–Min rate comparison among the proposed 2-layer FD C-RSMA system, the reference scheme, SDMA, and NOMA under synchronous reception

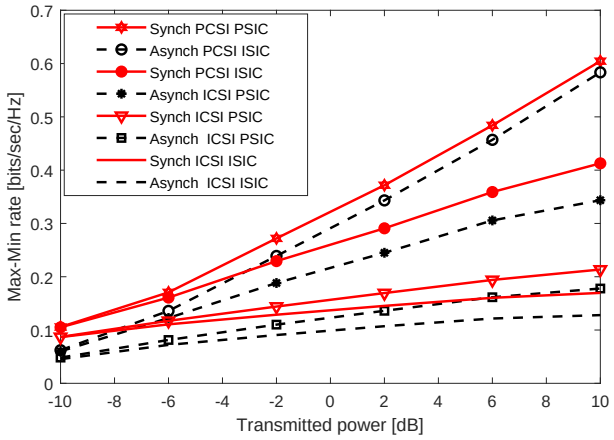

Figure 9: Max-Min rate vs. transmitted power with different impairments at $\sigma_e^2 = 5 * 10^{-4}$ in case of synchronous and asynchronous reception

(SDMA) schemes. It is clear that the proposed system outperforms all other schemes. RSMA-based strategies inherently achieve superior performance because they manage multi-user interference more efficiently, offering a flexible balance between partially decoding interference and partially treating it as noise. In addition, the proposed framework leverages full-duplex relays to forward both inter-group and inner-group common streams, which enhances link reliability and the rates of weaker users. The results also show that NOMA achieves higher Max–Min rates than SDMA since it fully decodes interference, whereas SDMA treats all interference as noise.

Figure 9 shows the Max-Min rate versus the BS transmitted power for the proposed 2-layer system with different impairments. It is worth noting that the channel imperfection has a significant effect on the system performance, as the figure shows that the system with imperfect CSI and perfect SIC has lower rate than the system with perfect CSI and imperfect SIC, and this implies that imperfect CSI has more of an effect on the performance of the system than imperfect SIC. In addition, the systems with asynchronous reception have worse performance than their corresponding ones with synchronous reception.

Figure 10 shows the Max-Min rate for different SIC imperfection β versus the BS transmitted power. This figure compares the performance of the proposed system in case of synchronous and asynchronous reception. It is shown that the systems with higher SIC imperfection factor have lower rate, this is because as β increases, the interference increases and the rate decreases.

Figure 11 shows the Max-Min rate versus the power transmitted of the BS for the proposed system in case of synchronous and asynchronous reception for different channel estimation error variance (σ_e^2). It is shown that the systems with higher estimation error variance have lower Max-Min rate compared with the other systems. This is because as estimation error increases, the interference and error increase, and thus the rate decreases.

Figure 12 shows the Max-Min rate versus the transmit-

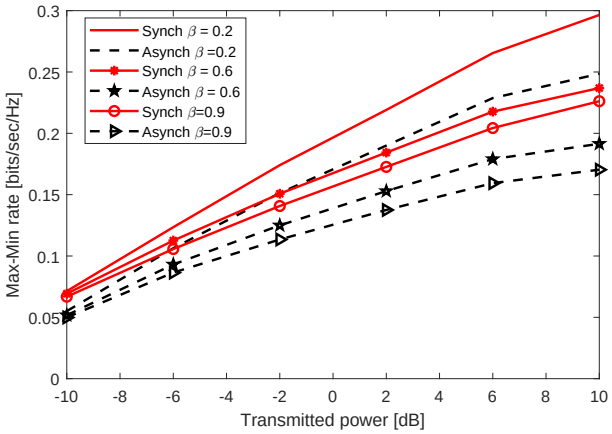

Figure 10: Sensitivity of Max-Min rate to the SIC imperfection factor (β) in case of synchronous and asynchronous reception

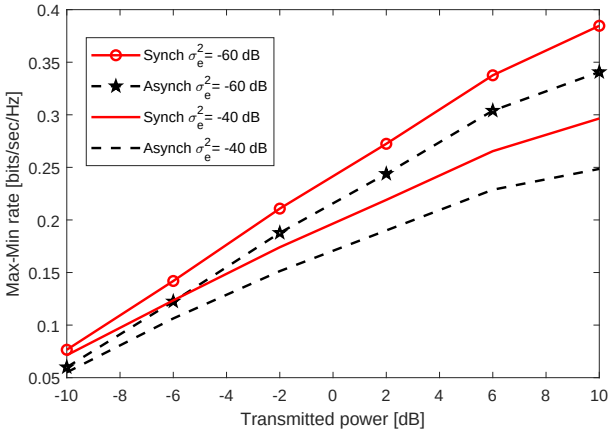

Figure 11: Sensitivity of Max-Min Rate to channel estimation error variance (σ_e^2) in case of synchronous and asynchronous reception

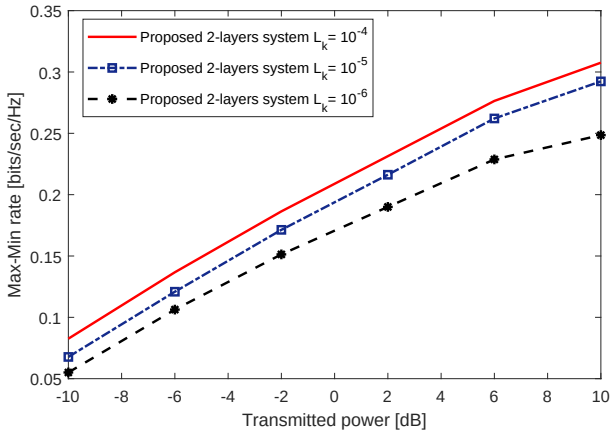

Figure 12: Impact of normalized data size (L) on asynchronous reception performance of the proposed system

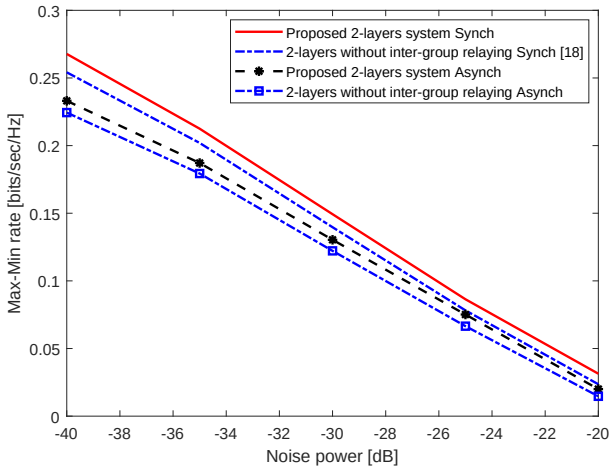

Figure 13: Max-Min rate vs. noise power for different systems in case of synchronous and asynchronous reception

ted power for the proposed system with different normalized data sizes (L_k). It is shown that the systems with lower data size have lower rate, this is because as data size decreases, the phase shift increases, causing more distortion and interference in the received signal.

Figure 13 shows the Max-Min rate versus the noise power. It compares the performance of the proposed system and 2-layer without inter-group relaying system in case of synchronous and asynchronous reception. It is shown that as the noise power increases, the Max-Min rate decreases due to increasing the effective interference and estimation errors experienced by the users.

Figure 14 shows the max–min rate versus the latency threshold for synchronous and asynchronous reception schemes. It can be observed that the proposed system consistently outperforms the 1-layer schemes, demonstrating the benefit of the additional common layer in managing interference and improving user fairness. It is also noticeable that the max–min rate decreases as the latency threshold increases, since latency is modeled as a constraint rather than a performance metric. Specifically, increasing the latency threshold relaxes the minimum rate requirement imposed on each user, allowing the optimizer to allocate fewer resources to the weakest user while still satisfying the latency constraint. Consequently, the optimization prioritizes interference management and power efficiency instead of maintaining a high minimum user rate, which leads to a reduction in the achieved max–min rate. This decreasing trend is therefore a direct consequence of the latency constraint formulation and reflects the inherent trade-off between fairness and delay tolerance. From a practical 6G perspective, this behavior indicates that enforcing stricter latency requirements (smaller t_{max}) implicitly pushes the system toward stronger fairness guarantees, as higher minimum rates are required to meet tight delay constraints. Conversely, when the latency constraint is relaxed, the system gains additional flexibility in resource allocation, which may slightly reduce fairness but improves overall interference management and power efficiency. Therefore, the figure highlights an important design trade-off where ultra-low latency operation inherently requires sustaining higher per-user rates, while delay-tolerant services allow more flexible and energy-efficient transmission strategies.

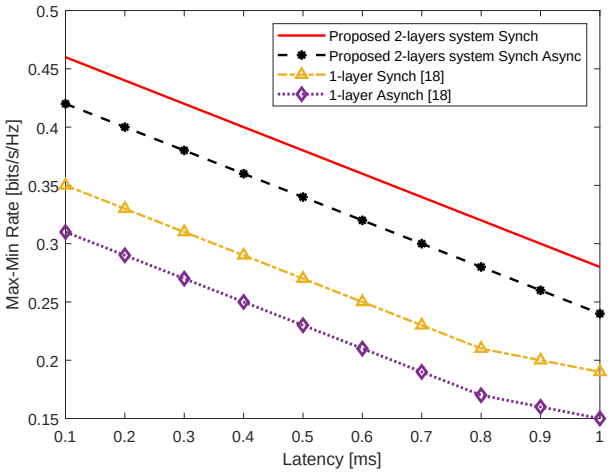

Figure 14: Latency–performance comparison of 1-layer and 2-layer RSMA schemes under synchronous and asynchronous reception.

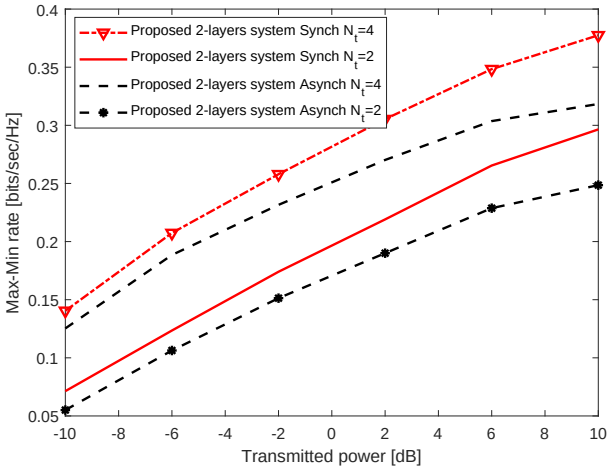

Figure 15: Max-Min rate vs. transmitted power for the proposed system with different number of antennas N_t

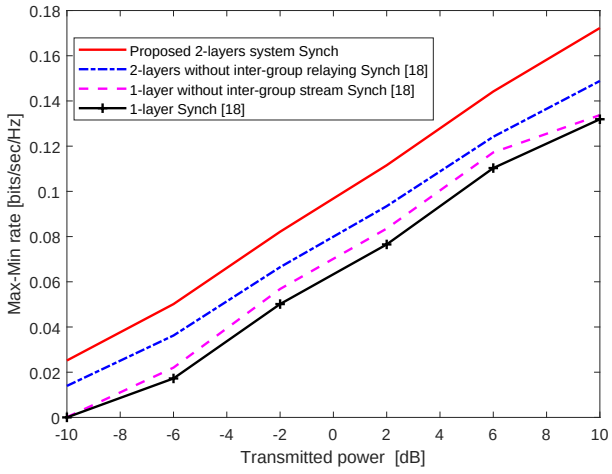

Figure 16: Max-Min rate vs. transmitted power synchronous cooperative schemes using COST2100 channel

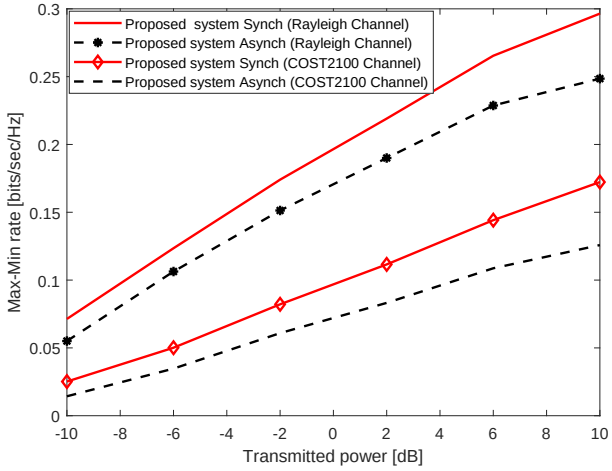

Figure 17: Max–min rate comparison of the proposed system under Rayleigh and COST2100 channel models

Figure 15 shows the Max-Min rate versus the transmitted power for the proposed system in case of synchronous and asynchronous reception for different number of antennas. The figure highlights the impact of increasing the number of BS antennas. As the number of antenna increases, the Max–Min rate improves due to the enhanced spatial diversity and antenna gain.

To enhance practical relevance beyond the considered Rayleigh channel model, we validate the proposed work using the COST2100 dataset [25], which captures realistic spatial correlation and propagation characteristics. This additional evaluation demonstrates that the proposed two-layer full-duplex cooperative RSMA scheme maintains its fairness and robustness performance under practical channel conditions as shown in Figure 16 and Figure 17. As shown in Figure 16 the proposed system outperforms all other considered schemes even under realistic COST2100 channel conditions, thereby validating its robustness and efficiency in practical propagation environments. While Figure 17 shows the Max-Min rate versus the transmitted power for the proposed system in case of synchronous and asynchronous reception for different channel models. It is noticeable that the systems using COST2100 channel have lower performance than their corresponding ones using Rayleigh channel, as COST2100 channel model captures realistic spatial correlation and cluster-based propagation, which reduces effective multiplexing and combining gains.

6.1. Jain's Index

To quantitatively measure how evenly the achievable rates are distributed among users, we adopt Jain's Fairness Index, a well-established metric for measuring fairness in multi-user communication systems. The index provides a normalized value between 0 and 1, where values closer to 1 indicate a more uniform and fair allocation of rates across all users, while values closer to 0 reflect strong imbalance. This makes Jain's Index an effective tool for evaluating the fairness performance of rate-splitting and cooperative

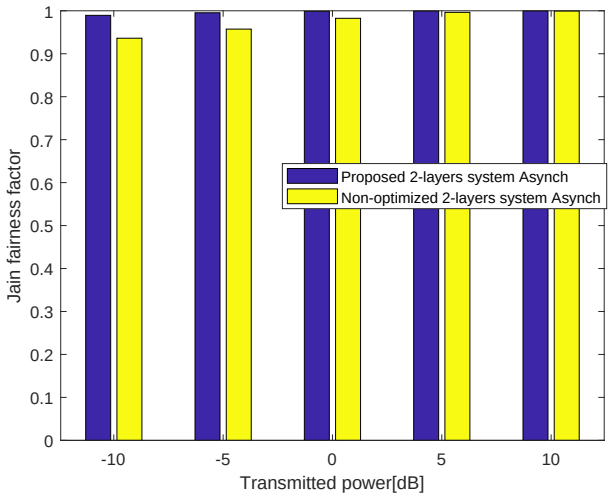

Figure 18: Jain's fairness index for the proposed optimized and non-optimized system

Table 7: Jain's Fairness Index

System	-10 dB	-5 dB	0 dB	5 dB	10 dB
Proposed	0.9894	0.9954	0.9989	0.9996	0.9998
Non-opt.	0.9362	0.9574	0.9825	0.9964	0.9995

relaying schemes, and it is given by

$$J = \frac{(\sum_{k=1}^{K} R_{k,tot})^2}{K \sum_{k=1}^{K} R_{k,tot}^2} \tag{39}$$

Figure 18 plots jain's index versus the transmitted power for the optimized and non-optimized proposed schemes. The figure demonstrates that the proposed system achieves a high level of fairness among users. It is shown that the optimized system attains a Jain's fairness index very close to 1, even under low transmit power conditions, confirming that users receive nearly equal rates. Moreover, the fairness of the optimized scheme is consistently higher than that of the non-optimized system across all power levels, highlighting the effectiveness of the optimization framework in balancing user performance. This graph can be summarized in the Table 7.

7. Conclusion

In this paper, we integrate 2-layer RSMA with full-duplex user relaying to address the limitations of 1-layer RSMA. The proposed system is analyzed under practical conditions, including imperfect CSI and imperfect SIC. In addition, we consider asynchronous reception, which is inevitable in wireless communications due to the arrival of signals from multiple sources located at different positions. The objective is to maximize the minimum rate of the users by jointly optimizing the precoding matrix, the inter-group common rate allocation, the inner-group common rate allocation, the head relay power, and the group relay power. The optimization problem is solved using an alternative optimization (AO) algorithm based on weighted minimum mean square error and solved iteratively. A key strength of the proposed framework is that it jointly incorporates 2-layer RSMA, full-duplex cooperation, and major practical impairments within a single unified model, addressing an evident gap in existing RSMA studies. Simulation results show that the proposed system has better performance when compared with other systems. Moreover, simulation results show that the asynchronous reception degrades the system performance. This integrated and realistic treatment enhances the relevance of the proposed design for practical 6G communication scenarios.

Appendix A. Derivation of the WMMSE for Head Relay (U_A) and Group Relay (U_n)

In this appendix we derive the WMMSE of head relay and group relay. The MSE of the inter-group common stream at the head relay (U_A) given in (23) can be expanded as

$$\epsilon_A^C = \mathbb{E}[\| q_A^C y_A \|^2 - 2\mathcal{R}\{q_A^C y_A s_c^*\} + \| s_c \|^2], \tag{A.1}$$

Using (5), the first and the second term of (A.1) can be written respectively as:

$$\begin{aligned}\mathbb{E}[\| q_A^C y_A \|^2] =& \| q_A^C \|^2 (\| \hat{\mathbf{g}}_A^H \mathbf{p}_c \|^2 + \sigma_{e,A}^2 \| \mathbf{p}_c \|^2) \\ &+ \sum_{i=1}^{K} (\| \hat{\mathbf{g}}_A^H \mathbf{p}_i \|^2 + \sigma_{e,A}^2 \| \mathbf{p}_i \|^2) + \sum_{j=1}^{G} (\| \hat{\mathbf{g}}_A^H \mathbf{p}_{c_j} \|^2 \\ &+ \sigma_{e,A}^2 \| \mathbf{p}_{c_j} \|^2) + \sum_{\substack{l=1 \\ l \neq A}}^{N} P_l (\| \delta_{l,A} \hat{h}_{l,A} \|^2 + \sigma_{l,A}^2) \\ &+ 2P_A \| h_{SI_A} \|^2 + N_0 \end{aligned} \tag{A.2}$$

and

$$\begin{aligned}\mathbb{E}[2\mathcal{R}\{q_A^C y_A s_c^*\}] =& 2\mathcal{R}\{q_A^C [(\hat{\mathbf{g}}_A^H + \mathbf{e}_A^H) \\ &(\mathbf{p}_c s_c + \sum_{g=1}^{G} \mathbf{p}_{c_g} \mathbf{s}_{c_g} + \sum_{k=1}^{K} \mathbf{p}_k \mathbf{s}_k) + h_{SI_A} \sqrt{P_A} (s_c + s_{c_g}) \\ &+ \sum_{\substack{l=1 \\ l \neq A}}^{N} (\delta_{l,A} \hat{h}_{l,A} + e_{l,A}) \sqrt{P_l} s_{c_l} + n_0] s_c^*\} \end{aligned} \tag{A.3}$$

But as mentioned before, $\mathbb{E}[\mathbf{e}_k] = 0$ and $\mathbb{E}[h_{SI_A}] = 0$. Also, $\mathbb{E}[\mathbf{s}_k \mathbf{s}_k^*] = \mathbf{I}$, this means that $\mathbb{E}[\mathbf{s}_k] = 0$, therefore $\mathbb{E}[\mathbf{s}_k \mathbf{s}_i^*] = 0$ where $k \neq i$ as the two streams are independent on each other. Therefore (A.3) can be rewritten as

$$\mathbb{E}[2\mathcal{R}\{q_A^C y_A s_c^*\}] = 2\mathcal{R}\{q_A^C \hat{\mathbf{g}}_A^H \mathbf{p}_c\} \tag{A.4}$$

while $\| s_c \|^2 = 1$. Therefore (23) can be written as

$$\begin{aligned}\epsilon_A^C =& \| q_A^C \|^2 T_A^C - 2\mathcal{R}\{q_A^C \hat{\mathbf{g}}_A^H \mathbf{p}_c\} + 1 \\ =& q_A^C q_A^{C^*} T_A^C - q_A^{C^*} (\hat{\mathbf{g}}_A^H \mathbf{p}_c) - q_A^C (\hat{\mathbf{g}}_A^H \mathbf{p}_c)^* + 1 \end{aligned} \tag{A.5}$$

where T_A^C is given by

$$T_A^C = (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_c \parallel^2) + \sum_{i=1}^{K} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_i \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_i \parallel^2) + \sum_{j=1}^{G} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_j} \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_{c_j} \parallel^2) + 2P_A \parallel h_{SI_A} \parallel^2 + \sum_{\substack{l=1 \\ l \neq A}}^{N} P_l (\parallel \delta_{l,A} \hat{h}_{l,A} \parallel^2 + \sigma_{l,A}^2) + N_0 \tag{A.6}$$

After decoding the inter-group common stream, the MSE of the inner-group common stream at the head relay (U_A) given in (24) can be rewritten as

$$\begin{aligned} \epsilon_{g,A}^C &= \parallel q_{g,A}^C \parallel^2 T_{g,A}^C - 2\mathcal{R}\{q_{g,A}^C \hat{\mathbf{g}}_A^H \mathbf{p}_{c_g}\} + 1 \\ &= q_{g,A}^C q_{g,A}^{C^*} T_{g,A}^C - q_{g,A}^{C^*} (\hat{\mathbf{g}}_A^H \mathbf{p}_{c_g}) - q_{g,A}^C (\hat{\mathbf{g}}_A^H \mathbf{p}_{c_g})^* + 1 \end{aligned} \tag{A.7}$$

where $T_{g,A}^C$ is given by

$$T_{g,A}^C = (\beta \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2 + \beta\sigma_{e,A}^2 \parallel \mathbf{p}_c \parallel^2) + \sum_{i=1}^{K} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_i \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_i \parallel^2) + \sum_{j=1}^{G} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_j} \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_{c_j} \parallel^2) + 2P_A \parallel h_{SI_A} \parallel^2 + \sum_{\substack{l=1 \\ l \neq A}}^{N} P_l (\parallel \delta_{l,A} \hat{h}_{l,A} \parallel^2 + \sigma_{l,A}^2) + N_0 \tag{A.8}$$

After decoding the inner-group common stream, the MSE of the private stream at the head relay (U_A) given in (25) can be rewritten as

$$\begin{aligned} \epsilon_{g,A}^P &= \parallel q_{g,A}^P \parallel^2 T_{g,A}^P - 2\mathcal{R}\{q_{g,A}^P \hat{\mathbf{g}}_A^H \mathbf{p}_A\} + 1 \\ &= q_{g,A}^P q_{g,A}^{P^*} T_{g,A}^P - q_{g,A}^{P^*} (\hat{\mathbf{g}}_A^H \mathbf{p}_A) - q_{g,A}^P (\hat{\mathbf{g}}_A^H \mathbf{p}_A)^* + 1 \end{aligned} \tag{A.9}$$

where $T_{g,A}^P$ is given by

$$T_{g,A}^P = (\beta \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2 + \beta\sigma_{e,A}^2 \parallel \mathbf{p}_c \parallel^2) + (\beta \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_g} \parallel^2 + \beta\sigma_{e,A}^2 \parallel \mathbf{p}_{c_g} \parallel^2) + \sum_{i=1}^{K} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_i \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_i \parallel^2) + \sum_{\substack{j=1 \\ j \neq g}}^{G} (\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c_j} \parallel^2 + \sigma_{e,A}^2 \parallel \mathbf{p}_{c_j} \parallel^2) + 2P_A \parallel h_{SI_A} \parallel^2 + \sum_{\substack{l=1 \\ l \neq A}}^{N} P_l (\parallel \delta_{l,A} \hat{h}_{l,A} \parallel^2 + \sigma_{l,A}^2) + N_0 \tag{A.10}$$

In order to achieve the minimum mean square error (MMSE), we should obtain the optimal equalizers. This can be done by solving $\frac{\partial \epsilon_A^C}{\partial q_A^{C*}} = 0$, $\frac{\partial \epsilon_{g,A}^C}{\partial q_{g,A}^{C*}} = 0$, and $\frac{\partial \epsilon_{g,A}^P}{\partial q_{g,A}^{P*}} = 0$, which result in $\tilde{q}_A^C = \frac{\hat{\mathbf{g}}_A^H \mathbf{p}_c}{T_A^C}$, $\tilde{q}_{g,A}^C = \frac{\hat{\mathbf{g}}_A^H \mathbf{p}_{c_g}}{T_{g,A}^C}$, and $\tilde{q}_{g,A}^P = \frac{\hat{\mathbf{g}}_A^H \mathbf{p}_A}{T_{g,A}^P}$. Where $\tilde{q}_A^C$, $\tilde{q}_{g,A}^C$, and $\tilde{q}_{g,A}^P$ are the optimal equalizers for the inter-group common stream, the inner-group common stream, and the private stream for the head relay, respectively. These optimal equalizers are computed in **Step** 2 and updated in **Step** 13 in **Algorithm** 1, and applied also for the other two types of users.

The MMSE of the inter-group common stream of head relay user can be obtained by substituting the obtained optimal equalizers in the MSE equation (A.5):

$$\begin{aligned} \tilde{\epsilon}_A^C &= \parallel \frac{\hat{\mathbf{g}}_A^H \mathbf{p}_c}{T_A^C} \parallel^2 T_A^C - 2\mathcal{R}\{\frac{\hat{\mathbf{g}}_A^H \mathbf{p}_c}{T_A^C} \hat{\mathbf{g}}_A^H \mathbf{p}_c\} + 1 \\ &= 1 - \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2}{T_A^C} = \frac{1}{1 + \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2}{I_A^C}} \end{aligned} \tag{A.11}$$

From (6), $SINR_A^C$ can be written as $SINR_A^C = \frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2}{I_A^C}$, where $I_A^C = T_A^C - \parallel \hat{\mathbf{g}}_A^H \mathbf{p}_c \parallel^2$. Then from (A.11), $\tilde{\epsilon}_A^C$ can be written as $\tilde{\epsilon}_A^C = \frac{1}{1+SINR_A^C}$. Consequently, the achievable rate of the inter-group common stream at the head relay user (U_A) is given by

$$R_A^C = log_2(1 + SINR_A^C) = -log_2(\tilde{\epsilon}_A^C) \tag{A.12}$$

Following the same procedures, the achievable rate of the inner-group common stream can be written as $R_{g,A}^C = -log_2(\tilde{\epsilon}_{g,A}^C)$, where $\tilde{\epsilon}_{g,A}^C$ is given by $\tilde{\epsilon}_{g,A}^C = \frac{1}{1+\frac{\parallel \hat{\mathbf{g}}_A^H \mathbf{p}_{c,g} \parallel^2}{I_{g,A}^C}} = \frac{1}{1+SINR_{g,A}^C}$. Similarly, the achievable rate of the private stream can be written as $R_{g,A}^P = -log_2(\tilde{\epsilon}_{g,A}^P)$, where $\tilde{\epsilon}_{g,A}^P$ is given by $\tilde{\epsilon}_{g,A}^P = \frac{1}{1+SINR_{g,A}^P}$. These MMSE are computed in **Step** 3 and updated in **Step** 13 in **Algorithm** 1, and applied also for the other two types of users. Meanwhile, the augmented WMSE of the inter-group common stream for the head relay can be expressed as

$$Z_A^C = w_A^C \tilde{\epsilon}_A^C - log_2(w_A^C) \tag{A.13}$$

where w_A^C is the weight associated with the head relay's inter-group common MSE. In order to get the weight (w_A^C) that achieves WMMSE, we solve $\frac{\partial Z_A^C}{\partial w_A^C} = 0$, which results in $w_A^C = \frac{1}{\tilde{\epsilon}_A^C}$, and then the WMMSE of the inter-group common stream for the head relay is given by

$$\tilde{Z_A}^C = 1 - log_2(\frac{1}{\tilde{\epsilon}_A^C}) \tag{A.14}$$

By substituting (A.12) in (A.14), the WMMSE of the head relay's inter-group common stream is given by $\tilde{Z}_A^C = 1 - R_A^C$.

Similarly, the augmented WMSE of the inner-group common stream and private stream for the head relay user can be expressed as $Z_{g,A}^C = w_{g,A}^C \tilde{\epsilon}_{g,A}^C - log_2(w_{g,A}^C)$, and

$Z_{g,A}^P = w_{g,A}^P \tilde{\epsilon}_{g,A}^P - log_2(w_{g,A}^P)$, respectively. Following the same procedure as in the inter-group common stream, we get the optimal weights, which are computed in **Step** 4 and updated in **Step** 13 in **Algorithm** 1; therefore, the WMMSE of the inner-group common stream and private stream for the head relay user can be expressed, respectively, as $\tilde{Z}_{g,A}^C = 1 - R_{g,A}^C$, and $\tilde{Z}_{g,A}^P = 1 - R_{g,A}^P$.

We apply the same techniques used for deriving the WMMSE for the head relay. Therefore, the mean square error (MSE) of the inter-group common stream at the group relay (U_n) is divided into two parts: the first part is due to the base station, and the second part is due to the head relay. The MSE of the inter-group common stream due to the base station is given by $\epsilon_n^{BS,C} = \mathbb{E}[\| q_n^{BS,C} y_{g,n}^{BS} - s_c \|^2]$, and can be written as

$$\begin{aligned}\epsilon_n^{BS,C} =& q_n^{BS,C} q_n^{BS,C^*} T_n^{BS,C} - q_n^{BS,C^*}(\hat{\mathbf{g}}_n^H \mathbf{p}_c) \\ & - q_n^{BS,C}(\hat{\mathbf{g}}_n^H \mathbf{p}_c)^* + 1 \end{aligned} \tag{A.15}$$

where $q_n^{BS,C}$ is the scalar equalizer of the inter-group common stream at the group relay (U_n) due to the BS, and $T_n^{BS,C}$ is given by

$$\begin{aligned}T_n^{BS,C} =& (\| \hat{\mathbf{g}}_n^H \mathbf{p}_c \|^2 + \sigma_{e,n}^2 \| \mathbf{p}_c \|^2) + \sum_{i=1}^{K} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_i \|^2 \\ & + \sigma_{e,n}^2 \| \mathbf{p}_i \|^2) + \sum_{j=1}^{G} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_{c_j} \|^2 + \sigma_{e,n}^2 \| \mathbf{p}_{c_j} \|^2) \\ & + P_n \| h_{SI_n} \|^2 + \sum_{\substack{l=1 \\ l \neq n}}^{N} P_l (\| \delta_{l,n} \hat{h}_{l,n} \|^2 + \sigma_{l,n}^2) + N_0 \end{aligned} \tag{A.16}$$

where the MSE of the inter-group common stream due to the head relay is given by $\epsilon_n^{A,C} = \mathbb{E}[\| q_n^{A,C} y_{g,n}^{A} - s_c \|^2]$, and can be written as

$$\begin{aligned}\epsilon_n^{A,C} =& q_n^{A,C} q_n^{A,C^*} T_n^{A,C} - q_n^{A,C^*}(\delta_{A,n} \hat{h}_{A,n} \sqrt{P_A}) \\ & - q_n^{A,C}(\delta_{A,n} \hat{h}_{A,n} \sqrt{P_A})^* + 1 \end{aligned} \tag{A.17}$$

where $q_n^{A,C}$ is the scalar equalizer of the inter-group common stream at the group relay (U_n) due to the head relay, and $T_n^{A,C}$ is given by

$$\begin{aligned}T_n^{A,C} =& (\| \delta_{A,n} \hat{h}_{A,n} \|^2 + \sigma_{A,n}^2) P_A + P_n \| h_{SI_n} \|^2 \\ & + \sum_{\substack{l=1 \\ l \neq n}}^{N} P_l (\| \delta_{l,n} \hat{h}_{l,n} \|^2 + \sigma_{l,n}^2) + N_0 \end{aligned} \tag{A.18}$$

After decoding the inter-group common stream, the MSE of the inner-group common stream at the group relay (U_n) is given by

$$\epsilon_{g,n}^C = q_{g,n}^C q_{g,n}^{C^*} T_{g,n}^C - q_{g,n}^{C^*}(\hat{\mathbf{g}}_n^H \mathbf{p}_{c_g}) - q_{g,n}^C (\hat{\mathbf{g}}_n^H \mathbf{p}_{c_g})^* + 1 \tag{A.19}$$

where $q_{g,n}^C$ is the scalar equalizer of the inner-group common stream at the group relay (U_n), and $T_{g,n}^C$ is given by

$$\begin{aligned}T_{g,n}^C =& (\beta \| \hat{\mathbf{g}}_n^H \mathbf{p}_c \|^2 + \beta \sigma_{e,n}^2 \| \mathbf{p}_c \|^2) + \sum_{i=1}^{K} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_i \|^2 \\ & + \sigma_{e,n}^2 \| \mathbf{p}_i \|^2) + \sum_{j=1}^{G} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_{c_j} \|^2 + \sigma_{e,n}^2 \| \mathbf{p}_{c_j} \|^2) \\ & + P_n \| h_{SI_n} \|^2 + \sum_{\substack{l=1 \\ l \neq n}}^{N} P_l (\| \delta_{l,n} \hat{h}_{l,n} \|^2 + \sigma_{l,n}^2) + N_0 \end{aligned} \tag{A.20}$$

After decoding the inner-group common stream, the MSE of the private stream at the group relay (U_n) is defined as

$$\epsilon_{g,n}^P = q_{g,n}^P q_{g,n}^{P^*} T_{g,n}^P - q_{g,n}^{P^*}(\hat{\mathbf{g}}_n^H \mathbf{p}_n) - q_{g,n}^P (\hat{\mathbf{g}}_n^H \mathbf{p}_n)^* + 1 \tag{A.21}$$

where $q_{g,n}^P$ is the scalar equalizer of the private stream at the group relay (U_n), and $T_{g,n}^P$ is given by

$$\begin{aligned}T_{g,n}^P =& (\beta \| \hat{\mathbf{g}}_n^H \mathbf{p}_c \|^2 + \beta \sigma_{e,n}^2 \| \mathbf{p}_c \|^2) + (\beta \| \hat{\mathbf{g}}_n^H \mathbf{p}_{c_g} \|^2 \\ & + \beta \sigma_{e,n}^2 \| \mathbf{p}_{c_g} \|^2) + \sum_{i=1}^{K} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_i \|^2 + \sigma_{e,n}^2 \| \mathbf{p}_i \|^2) \\ & + \sum_{\substack{j=1 \\ j \neq g}}^{G} (\| \hat{\mathbf{g}}_n^H \mathbf{p}_{c_j} \|^2 + \sigma_{e,n}^2 \| \mathbf{p}_{c_j} \|^2) + P_n \| h_{SI_n} \|^2 \\ & + \sum_{\substack{l=1 \\ l \neq n}}^{N} P_l (\| \delta_{l,n} \hat{h}_{l,n} \|^2 + \sigma_{l,n}^2) + N_0 \end{aligned} \tag{A.22}$$

As before, the optimal equalizers are obtained by solving $\frac{\partial \epsilon_n^{BS,C}}{\partial q_n^{BS,C^*}} = 0$, $\frac{\partial \epsilon_n^{A,C}}{\partial q_n^{A,C^*}} = 0$, $\frac{\partial \epsilon_{g,n}^C}{\partial q_{g,n}^{C^*}} = 0$, and $\frac{\partial \epsilon_{g,n}^P}{\partial q_{g,n}^{P^*}} = 0$, which results in $\tilde{q}_n^{BS,C} = \frac{\hat{\mathbf{g}}_n^H \mathbf{p}_c}{T_n^{BS,C}}$, $\tilde{q}_n^{A,C} = \frac{\delta_{A,n} \hat{h}_{A,n} \sqrt{P_A}}{T_n^{A,C}}$, $\tilde{q}_{g,n}^C = \frac{\hat{\mathbf{g}}_n^H \mathbf{p}_{c_g}}{T_{g,n}^C}$, and $\tilde{q}_{g,n}^P = \frac{\hat{\mathbf{g}}_n^H \mathbf{p}_n}{T_{g,n}^P}$. Where $\tilde{q}_n^{BS,C}$, $\tilde{q}_n^{A,C}$, $\tilde{q}_{g,n}^C$, and $\tilde{q}_{g,n}^P$ are the optimal equalizers for the inter-group common stream due to the BS and the head relay, the inner-group common stream, and the private stream for the group relay, respectively. The MMSE of the inter-group common stream, the inner-group common stream, and the private stream of the group relay can be obtained by substituting the optimal equalizers in the derived MSE equations (A.15), (A.17), (A.19), and (A.21), and the corresponding achievable rates are given respectively by

$$\begin{aligned}R_n^C &= log_2(1 + SINR_n^{A,C} + SINR_n^{BS,C}) \\ &= log_2(\frac{1}{\tilde{\epsilon}_n^{A,C}} + \frac{1}{\tilde{\epsilon}_n^{BS,C}} - 1) \end{aligned} \tag{A.23}$$

$$R_{g,n}^C = -log_2(\tilde{\epsilon}_{g,n}^C) \tag{A.24}$$

$$R_{g,n}^P = -log_2(\tilde{\epsilon}_{g,n}^P) \tag{A.25}$$

Following the same procedures in Section 5.1, the WMMSE of inter-group common stream is given by

$$\tilde{Z}_n^C = 2 - log_2(\frac{1}{\tilde{\epsilon}_n^{BS,C}\tilde{\epsilon}_n^{A,C}}) \tag{A.26}$$

Hence, after some mathematical manipulations, the WMMSE of inter-group common stream, inner-group common stream and private stream of the group relay can be given receptively by

$$\tilde{Z}_n^C \approx 2 - R_n^C \tag{A.27}$$

$$\tilde{Z}_{g,n}^C = 1 - log_2(\frac{1}{\tilde{\epsilon}_{g,n}^C}) = 1 - R_{g,n}^C \tag{A.28}$$

$$\tilde{Z}_{g,n}^P = 1 - log_2(\frac{1}{\tilde{\epsilon}_{g,n}^P}) = 1 - R_{g,n}^P \tag{A.29}$$

References

[1] Elaheh Sadeghabadi and Steven D. Blostein. Reduced complexity rate-splitting multiple access beamforming for generalized objectives. *IEEE Access*, 12:155958–155975, 2024. doi: 10.1109/ACCESS.2024.3483688.

[2] Ting Qi, Bin Lyu, and Changsheng You. Transmit power minimization for irs-assisted hierarchical rate-splitting multiple access systems. *IEEE Wireless Communications Letters*, 14(5): 1386–1390, 2025. doi: 10.1109/LWC.2025.3543172.

[3] Zhixiang Yang, Lei Feng, Fanqin Zhou, Kunyi Xie, Xuesong Qiu, Zehui Xiong, Chau Yuen, and Zhu Han. User grouping-based hierarchical rate-splitting multiple access for reconfigurable intelligent surface-assisted downlink communication. *IEEE Transactions on Cognitive Communications and Networking*, pages 1–1, 2025. doi: 10.1109/TCCN.2025.3549221.

[4] Prabhat Kumar Sharma and Urbashi Mitra. 2- layer hierarchical rsma for star ris assisted cell-edge communications. In *2025 National Conference on Communications (NCC)*, pages 1–6, 2025. doi: 10.1109/NCC63735.2025.10983309.

[5] Jiarui Liang, Hang Deng, Yuanyuan Peng, Jinpeng Song, Shuai Wang, and Gaofeng Pan. Uav-terrestrial rsma communication suffering terrestrial interference. In *2024 International Conference on Ubiquitous Communication (Ucom)*, pages 53–56, 2024. doi: 10.1109/Ucom62433.2024.10695926.

[6] Hehao Niu, Zhi Lin, Kang An, Jiangzhou Wang, Gan Zheng, Naofal Al-Dhahir, and Kai-Kit Wong. Active ris assisted rate-splitting multiple access network: Spectral and energy efficiency tradeoff. *IEEE Journal on Selected Areas in Communications*, 41(5):1452–1467, 2023. doi: 10.1109/JSAC.2023.3240718.

[7] Penglu Liu, Yong Li, Wei Cheng, Xiaodai Dong, and Limeng Dong. Active intelligent reflecting surface aided rsma for millimeter-wave hybrid antenna array. *IEEE Transactions on Communications*, 71(9):5287–5302, 2023. doi: 10.1109/TCOMM.2023.3285290.

[8] Zhiqiang Li, Shiji Wang, Shuai Han, and Cheng Li. Non-orthogonal broadcast and unicast joint transmission for multi-beam satellite system. *IEEE Transactions on Broadcasting*, 69 (3):647–660, 2023. doi: 10.1109/TBC.2023.3267239.

[9] Jeonghun Park, Jinseok Choi, Namyoon Lee, Wonjae Shin, and H. Vincent Poor. Rate-splitting multiple access for downlink mimo: A generalized power iteration approach. *IEEE Transactions on Wireless Communications*, 22(3):1588–1603, 2023. doi: 10.1109/TWC.2022.3205480.

[10] Roberto Pereira, Anay Ajit Deshpande, Cristian J. Vaca-Rubio, Xavier Mestre, Andrea Zanella, David Gregoratti, Elisabeth de Carvalho, and Petar Popovski. User clustering for rate splitting using machine learning. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 722–726, 2022. doi: 10.23919/EUSIPCO55093.2022.9909639.

[11] Hongyu Li, Yijie Mao, Onur Dizdar, and Bruno Clerckx. Rate-splitting multiple access for 6g—part iii: Interplay with reconfigurable intelligent surfaces. *IEEE Communications Letters*, 26 (10):2242–2246, 2022. doi: 10.1109/LCOMM.2022.3192041.

[12] Aditya Jolly, Keshav Singh, and Sudip Biswas. Rsma for irs aided 6g communication systems: Joint active and passive beamforming design. In *2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pages 7–12, 2021. doi: 10.1109/ANTS52808.2021.9936974.

[13] Ankur Bansal, Keshav Singh, and Chih-Peng Li. Analysis of hierarchical rate splitting for intelligent reflecting surfaces-aided downlink multiuser miso communications. *IEEE Open Journal of the Communications Society*, 2:785–798, 2021. doi: 10.1109/OJCOMS.2021.3070340.

[14] Mingbo Dai, Bruno Clerckx, David Gesbert, and Giuseppe Caire. A rate splitting strategy for massive mimo with imperfect csit. *IEEE Transactions on Wireless Communications*, 15(7):4611–4624, 2016. doi: 10.1109/TWC.2016.2543212.

[15] Ziheng Li, Tong Wang, Lin Gao, Yufei Jiang, and Zhihua Yang. A two-layer rsma framework with balanced clustering design for cell-free massive mimo systems. *IEEE Internet of Things Journal*, pages 1–1, 2025. doi: 10.1109/JIOT.2025.3584822.

[16] Sultangali Arzykulov, Galymzhan Nauryzbayev, Mohammad S. Hashmi, Ahmed M. Eltawil, Khaled M. Rabie, and Shakhmaran Seilov. Hardware- and interference-limited cognitive iot relaying noma networks with imperfect sic over generalized non-homogeneous fading channels. *IEEE Access*, 8:72942–72956, 2020. doi: 10.1109/ACCESS.2020.2987873.

[17] Elaheh Sadeghabadi and Steven Blostein. Precoding design and sum rate upper bound of rsma using interference nulling. In *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 663–668, 2022. doi: 10.1109/ICCWorkshops53468.2022.9814632.

[18] Shreya Khisa, Mohamed Elhattab, Chadi Assi, and Sanaa Sharafeddine. Joint user pairing and resource allocation optimization in downlink 2-layer cooperative rsma networks. In *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2024. doi: 10.1109/WCNC57260.2024.10570959.

[19] Jiakang Zheng, Jiayi Zhang, Julian Cheng, Victor C. M. Leung, Derrick Wing Kwan Ng, and Bo Ai. Asynchronous cell-free massive mimo with rate-splitting. *IEEE Journal on Selected Areas in Communications*, 41(5):1366–1382, 2023. doi: 10.1109/JSAC.2023.3240709.

[20] Guoyu Li, Shaochuan Wu, Changsheng You, Wenbin Zhang, Guanyu Shang, and Xiaokang Zhou. Cell-free massive mimo-ofdm: Asynchronous reception and performance analysis. *IEEE Internet of Things Journal*, 11(7):11894–11906, 2024. doi: 10.1109/JIOT.2023.3333532.

[21] Guoyu Li, Shaochuan Wu, Changsheng You, Wenbin Zhang, and Guanyu Shang. Asynchronous cell-free massive mimo-ofdm: Mixed coherent and non-coherent transmissions. *IEEE Communications Letters*, 29(2):363–367, 2025. doi: 10.1109/LCOMM.2024.3518485.

[22] Jinyuan Gu, Mingxing Wang, Wei Duan, Lei Zhang, and Huaiping Zhang. Ris-noma communications over nakagamim fading with imperfect successive interference cancellation. *Computer Communications*, 226-227:107926, 2024. ISSN 0140-3664. doi: https://doi.org/10.1016/j.comcom.2024.107926. URL `https://www.sciencedirect.com/science/article/pii/S0140366424002718`.

[23] Shreya Khisa, Mohammed Almekhlafi, Mohamed Elhattab, and Chadi Assi. Full duplex cooperative rate splitting multiple access for a miso broadcast channel with two users. *IEEE Communications Letters*, 26(8):1913–1917, 2022. doi: 10.1109/LCOMM.2022.3173894.

[24] Jian Zhang, Bruno Clerckx, Jianhua Ge, and Yijie Mao. Cooperative rate splitting for miso broadcast channel with user relaying, and performance benefits over cooperative noma. *IEEE Signal Processing Letters*, 26(11):1678–1682, 2019. doi: 10.1109/LSP.2019.2942994.

[25] L. Liu, J. Poutanen, F. Quitin, K. Haneda, F. Tufvesson, P. De Doncker, P. Vainikainen, and C. Oestges. The cost 2100 mimo channel model. *IEEE Wireless Communications*, 19(6):92–99, December 2012.

Sama Wahb received the B.Sc. degree in 2023 and the M.Sc. degree in digital communications in 2024. She is currently pursuing the Ph.D. degree and works as a Teaching Assistant at the German University in Cairo. Her research interests include digital communication systems, and emerging 6G technologies.

Ahmed El-Mahdy was born in Port Said, Egypt, in 1963. He received the B.S., M.S., and Ph.D. degrees in electrical engineering in 1986, 1992, and 1999, respectively. He is currently a Professor in the Faculty of Information Engineering and Technology at the German University in Cairo. His research interests include wireless communications and signal processing.

Falko Dressler is a Chair for Telecommunication Networks at TU Berlin. He received the M.Sc. and Ph.D. degrees from the University of Erlangen in 1998 and 2003, respectively. His research interests include adaptive wireless networking across multiple bands and wireless-based sensing, with applications in ad hoc and sensor networks, the Internet of Things, and cyber-physical systems.