Malicious RIS Meets RSMA: Unveiling the Robustness of Rate Splitting to RIS-Induced Attacks

Arthur S. de Sena[†], André Gomes[‡], Jacek Kibiłda[‡], Nurul H. Mahmood[†],
Luiz A. DaSilva[‡], Matti Latva-aho[†]
[†] University of Oulu, 6G Flagship, Finland [‡] Virginia Tech, CCI, USA

Emails: {arthur.sena, nurulhuda.mahmood, matti.latva-aho}@oulu.fi, {gomesa, jkibilda, ldasilva}@vt.edu

***Abstract*—While the robustness of rate-splitting multiple access (RSMA) to imperfect channel state information (CSI) is well-documented, its susceptibility to attacks launched with malicious reconfigurable intelligent surfaces (RISs) remains unexplored. This paper fills this gap by investigating three potential RIS-induced attacks against RSMA in a multi-user multiple-input multiple-output (MIMO) network: random interference, aligned interference, and mitigation attack. The random interference attack employs random RIS coefficients to disrupt RSMA. The other two attacks are triggered by optimizing the RIS through weighted-sum strategies based on the projected gradient method. Simulation results reveal significant degradation caused by all the attacks under perfect CSI conditions. Remarkably, when imperfect CSI is considered, RSMA, owing to its flexible power allocation strategy designed to counter CSI-related interference, can be robust to the attacks even when the base station is blind to them. It is also shown that RSMA can significantly outperform conventional space-division multiple access (SDMA).**

***Index Terms*—Reconfigurable intelligent surface, rate-splitting multiple access, physical-layer security, multi-user MIMO.**

I. Introduction

Next-generation multiple-input multiple-output (MIMO) systems, employing an ever-increasing number of antennas, are a central pillar for future sixth-generation (6G) wireless networks, as they allow for simultaneous transmissions to multiple users under the same frequency and time slot. One of the major challenges lies in maintaining low inter-user interference while serving multiple users. Conventional approaches employ linear precoding strategies to implement space-division multiple access (SDMA), which requires accurate channel state information (CSI). In practical systems, only imperfect CSI is available, which inevitably leads to residual inter-user interference at the receiver [1].

To overcome the limitations of SDMA, rate-splitting multiple access (RSMA) was proposed [1]–[5]. RSMA implements a flexible split transmission scheme aided by successive interference cancellation (SIC) at the receiver side. At the base station (BS), one part of the users' data is encoded into private messages and transmitted via private precoders, while the other part is encoded into a common message and broadcast through a common precoder [1], [2]. The private precoders are designed similarly as in SDMA, making them sensitive to imperfect CSI. In contrast, the common precoder is constructed as a multicast precoder, which increases the system's robustness to inter-user interference.

The reliance of multi-user MIMO systems on CSI makes them susceptible to attacks that alter the propagation environment [6]–[9]. These attacks may employ a reconfigurable intelligent surface (RIS), a low-power planar array of nearly passive reconfigurable elements that can be dynamically controlled to adjust the propagation environment. While RIS is primarily seen as a performance-enhancing technology [2], it may also trigger powerful attacks against wireless links [6]–[13]. In the SDMA case, among other threats, RISs can disrupt the channel estimation process and boost CSI inaccuracy, making linear precoders incapable of tackling inter-user interference. Different malicious attack schemes have been identified, including attacks with random RIS coefficients [6], [7], optimized attacks in which the RIS-associated channels are aligned to boost interference [8], and cancellation attacks, where the RIS is optimized to add up the direct and reflected signals destructively at the receiver [9]. While the robustness of RSMA to interference originating from CSI imperfections has been demonstrated [1], [3], [5], its susceptibility to RIS-induced attacks remains an open question.

This paper delves for the first time into potential adversarial attacks that can be launched with the help of a malicious RIS against RSMA. We investigate a downlink multi-user MIMO network, in which a nearby attacker controls an adversarial RIS, as can be seen in Fig. 1. For this scenario, the attacker exploits the training protocol employed at the BS to execute three different RIS attack options not yet explored in the context of RSMA: *random interference*, *aligned interference*, and *mitigation attack*. The *random interference attack* attempts to degrade the transmission of common and private data messages by configuring the RIS with random phase shifts. The *aligned interference attack* tries to maximize the reflected power by exploiting the channels associated with the RIS to further diminish the effectiveness of RSMA precoders. Finally, the *mitigation attack* attempts to minimize the signal power at the users by destructively adding the reflected RIS channels to the legitimate direct user channels. For the latter two attacks, we consider weighted-sum strategies based on the projected gradient method to optimize the adversarial RIS phase shifts.

Our numerical results reveal a remarkable property of RSMA that manifests itself in scenarios with imperfect CSI.

André is now with Rowan University, USA. Email: gomesa@rowan.edu.

By flexibly allocating power to the common message, RSMA can (unintentionally) mitigate the impact of the attacks, considerably outperforming SDMA in all the considered threat scenarios under imperfect CSI. Counterintuitively, the proposed attacks have the strongest impact under perfect CSI conditions. The results show that the attacks can potentially cause significant impact in all scenarios, with the severest performance degradation observed for the mitigation attack.

Notation: Boldface lower-case letters denote vectors and upper-case represent matrices. The ith column of a matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{:,i}$, the transpose and Hermitian transpose of $\mathbf{A}$ are represented by $\mathbf{A}^T$ and $\mathbf{A}^H$, respectively, $\mathbf{1}_{M,N}$ is the $M \times N$ all-ones matrix, and $\diamond$ represents the Khatri-Rao product. The operator $\text{vec}\{\cdot\}$ transforms an $M \times N$ matrix into a column vector, $\text{vecd}\{\cdot\}$ converts the diagonal elements of an $M \times M$ matrix into a column vector, $\text{diag}\{\cdot\}$ transforms a vector of length M into an $M \times M$ diagonal matrix, and $\angle(z)$ returns the phase of the complex scalar z.

II. System Model

This paper studies the downlink MIMO system illustrated in Fig. 1, where one BS employing M antennas performs downlink data transmission to K single-antenna users, represented as $\mathcal{K} = \{1, 2, \cdots, K\}$, utilizing RSMA. We assume an attacker deploys one malicious RIS with L reflecting elements. For the attacks to be effective, the RIS is set to an absorption mode during the channel estimation phase and turned on only when the data transmission starts, as considered in [6]–[8].

We adopt in this work the single-layer RSMA, where the BS transmits a single common message and users are required to carry out a single-layer SIC [1]. Under the single-layer RSMA protocol, the message for each user is first split into common and private parts at the BS. All the users' common parts are then encoded into a single common super symbol x^c, while the private parts are individually mapped into private symbols x_k^p. Next, the common and private symbols x^c and x_k^p are linearly precoded and superimposed in the power domain for transmission, resulting in the following data vector

$$\mathbf{x} = \mathbf{p}^c\sqrt{P\alpha^c}x^c + \sum_{k=1}^{K}\mathbf{p}_k^p\sqrt{P\alpha_k^p}x_k^p \in \mathbb{C}^M, \tag{1}$$

where P is the total transmit power, α^c and α_k^p are the power allocation coefficients for the common and private symbols, and $\mathbf{p}^c \in \mathbb{C}^M$ and $\mathbf{p}_k^p \in \mathbb{C}^M$ are the linear precoding vectors responsible for transmitting the corresponding symbols.

A. Signal reception and performance metrics

After transmission, the superimposed RSMA data streams propagate through the direct link and the reflected one via the malicious RIS. As a result, the kth user will receive

$$y_k = \left(\mathbf{f}_k^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_k^H\right)\left(\mathbf{p}^c\sqrt{P\alpha^c}x^c + \sum_{i=1}^{K}\mathbf{p}_i^p\sqrt{P\alpha_i^p}x_i^p\right) + n_k, \tag{2}$$

where $\mathbf{h}_k \in \mathbb{C}^M$, $\mathbf{G} \in \mathbb{C}^{L\times M}$, and $\mathbf{f}_k \in \mathbb{C}^L$ model the wireless channels between the BS and the kth user (link BS-U), the BS and the RIS (link BS-RIS), and the RIS and the kth user (link RIS-U), respectively. $\mathbf{\Theta} = \text{diag}\{\mu_1 e^{-j\theta_1}, \cdots, \mu_L e^{-j\theta_L}\} \in \mathbb{C}^{L\times L}$ is the diagonal matrix modeling the reflections induced by the malicious RIS, satisfying $|\mu_l|^2 = 1$ and $\theta_l \in [0, 2\pi], \forall l = 1, \cdots, L$, and $n_k \in \mathbb{C}$ is the noise coefficient for the kth user, following the complex Gaussian distribution with zero mean and variance σ^2.

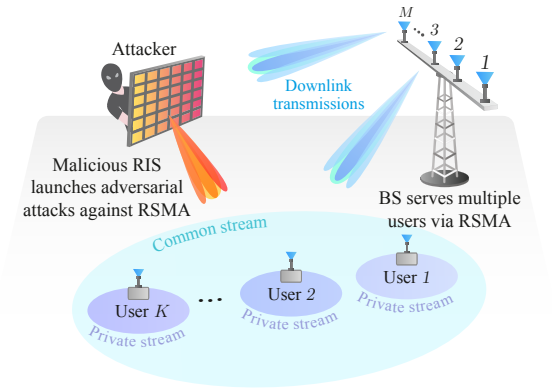

Fig. 1: An attacker deploys an RIS to perform adversarial attacks against RSMA in a downlink multi-user MIMO system.

The common message is recovered directly from (2) while the private messages are treated as noise. As a result, the common message will be decoded by the kth user with the following signal-to-interference-plus-noise ratio (SINR)

$$\gamma_k^c = \frac{|\left(\mathbf{f}_k^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_k^H\right)\mathbf{p}^c|^2 P\alpha^c}{\sum_{i=1}^{K}|\left(\mathbf{f}_k^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_k^H\right)\mathbf{p}_i^p|^2 P\alpha_i^p + \sigma^2}, \tag{3}$$

where the term in the denominator accounts for the interference generated by both the kth intended private message and unintended private messages, resulting from imperfect CSI used in the precoder design and the malicious RIS attack. The instantaneous common rate experienced at the kth user is then given by $R_k^c = \log_2(1 + \gamma_k^c)$. Since all users need to decode the common message, the actual allocated rate will be $R^c = \min_{\forall k\in\mathcal{K}}\{R_k^c\}$.

Once the common message is retrieved, SIC is executed to subtract it from (2). We assume that SIC can successfully remove the interference associated with the common message. However, the kth user still experiences residual interference of private messages intended for other users $k' \neq k \in \mathcal{K}$, i.e., due to CSI error and RIS interference. Thus, the private SINR for the kth user will be

$$\gamma_k^p = \frac{|\left(\mathbf{f}_k^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_k^H\right)\mathbf{p}_k^p|^2 P\alpha_k^p}{\sum_{i=1,i\neq k}^{K}|\left(\mathbf{f}_k^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_k^H\right)\mathbf{p}_i^p|^2 P\alpha_i^p + \sigma^2}, \tag{4}$$

resulting in the following achievable private rate $R_k^p = \log_2(1+\gamma_k^p)$. Consequently, the sum rate, in bits/s/Hz, experienced in the system can be expressed as $R = R^c + \sum_{k=1}^{K} R_k^p$.

B. CSI acquisition and precoder design at the BS

During the channel estimation phase, the attacker sets its RIS to an absorption mode so that the BS does not account for the reflected RIS channels in the estimation process. Nevertheless, we assume that the estimate of the legitimate fast-fading channels, $\mathbf{h}_k, \forall k \in \mathcal{K}$, acquired at the BS is

imperfect. This implies that the BS precoders are designed based on a corrupted version of $\mathbf{h}_k$, modeled by [3]

$$\hat{\mathbf{h}}_k = \sqrt{1-(\tau^{\text{BS-U}})^2}\mathbf{h}_k + \tau^{\text{BS-U}}\mathbf{z}_k, \tag{5}$$

where $\mathbf{z}_k$ is the error vector independent of $\mathbf{h}_k$, whose entries follow the complex Gaussian distribution with zero mean and unit variance, and the coefficient $\tau^{\text{BS-U}} \in [0,1]$ models the quality of the CSI estimation.

Given the above considerations, the private precoding vector $\mathbf{p}_k^p \in \mathbb{C}^M$ should be designed to ensure the (ideally) interference-free delivery of the private messages. More specifically, we wish to achieve $\left[\mathbf{h}_{k'}^H\right]\mathbf{p}_k^p \approx 0$, $\forall k' \neq k \in \mathcal{K}$. Such a goal can be accomplished by designing $\mathbf{p}_k^p$ as a zero-forcing precoder based on the acquired estimate of $\mathbf{h}_k$ in (5), as follows. First, let us define $\hat{\mathbf{H}} = \left[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \cdots, \hat{\mathbf{h}}_K\right] \in \mathbb{C}^{M\times K}$. Then, the private precoder for the kth user can be given by

$$\mathbf{p}_k^p = \left[\hat{\mathbf{H}}\left(\hat{\mathbf{H}}^H\hat{\mathbf{H}}\right)^{-1}\right]_{:,k} \in \mathbb{C}^M, \tag{6}$$

where it must be ensured that $M \geq K$.

As for the common precoder, $\mathbf{p}^c \in \mathbb{C}^M$, it should offer a satisfactory reception of the common message to all users. However, the design of such precoders generally leads to NP-hard problems that can only be solved sub-optimally through, conventionally, iterative methods with high computational complexity , e.g., semidefinite relaxation and successive convex approximation-based methods [14]. Alternatively, as in [3], we adopted a lower complexity weighted matched filter (MF) strategy, which can be computed in closed form by

$$\mathbf{p}^c = \sum_{i=1}^{K} \mu_k \hat{\mathbf{h}}_i, \tag{7}$$

where μ_k is the weight for the kth user. In particular, the weights are adjusted as $\mu_1 = \cdots = \mu_K = \frac{1}{\sqrt{\mathbf{s}^H\mathbf{s}}}$, with $\mathbf{s} = \sum_{i=1}^{K}\hat{\mathbf{h}}_i$, such that $\|\mathbf{p}^c\|_2^2 = 1$. The precoding method in (7) becomes asymptotically optimal as the number of transmit antennas grows large [3].

C. Power allocation

This paper adopts a simple yet effective adaptive power allocation strategy inspired by [1], where the power is allocated in such a way that the interference resulting from imperfect CSI in the SINR for the private messages in (4), corresponding to the legitimate BS-U channels, reaches approximately the same level as the noise power, i.e., $\sum_{i=1,i\neq k}^{K}|\mathbf{h}_k^H\mathbf{p}_i^p|^2 P\alpha_i^p \approx \sigma^2$. Specifically, the power allocation coefficient for the common message is determined as a function of the power allocated to the private messages as $\alpha^c = 1-\sum_{i=1}^{K}\alpha_i^p$, in which a uniform power allocation is employed across the private coefficients, such that $\alpha^p = \alpha_1^p = \cdots = \alpha_K^p$, based on the criteria:

$$\alpha^p = \min\left\{\frac{1}{K}, \frac{\sigma^2}{\min\limits_{\forall k}\sum_{i=1,i\neq k}^{K}|\mathbf{h}_k^H\mathbf{p}_i^p|^2 P}\right\}. \tag{8}$$

As can be noticed, the greater the interference levels generated by imperfect CSI, the more power is allocated to the common message. As explained in [1], making the interference power similar to the noise power ensures that the experienced data rates do not saturate in the high transmit power regime, i.e., $R \to \infty$ as $P \to \infty$ even under imperfect CSI scenarios. Note that for the above approach to work, the noise and interference powers (or their ratio) must be reported by each user to the BS in the training phase. More advanced allocation methods, such as in [4], [5], are left for future work.

III. Potential RIS-Induced Attacks against RSMA

The malicious goal of the attacker is to compute a reflection matrix $\boldsymbol{\Theta}$ that induces a performance degradation of the employed RSMA scheme, i.e., degrade the system sum rate. In the following subsections, we investigate three approaches to accomplish the goal and analyze the necessary CSI knowledge for their implementation.

A. Random interference attack

The most straightforward strategy that can cause performance degradation in RSMA consists of configuring randomly the RIS reflecting coefficients to launch passive jamming attacks, similar to the attack against SDMA proposed in [6]. Random RIS interference attacks should reduce the effectiveness of the precoders in (6) and (7), consequently leading to degradation in the rates of both common and private messages, without the need for any CSI knowledge. The strongest impact should be observed against the private messages, given that the associated rates will become interference-limited due to the inability of the private precoders to cancel out the inter-user interference propagating through the malicious RIS. It is noteworthy that, even though CSI is not needed to optimize the RIS, the attacker must know when the channel estimation and power allocation training phases happen. With this information, the attacker can configure the RIS to absorb impinging signals and avoid being detected by the BS.

B. Aligned interference attack

In our recent work [8], we demonstrated that if the attacker manages to acquire at least the illegitimate BS-RIS and RIS-U channels, it becomes possible to optimize the RIS to cause a powerful performance degradation. In this subsection, we show how such an optimized interference attack can be extended to RSMA.

For this attack to be effective in RSMA, as in the previous subsection, the RIS is set to absorption mode for both the channel estimation and power allocation. Moreover, we assume that the attacker has enough computational power to estimate the channels of the links BS-RIS and RIS-U, which are modeled as $\hat{\mathbf{G}} = \sqrt{1-(\tau^{\text{BS-RIS}})^2}\mathbf{G} + \tau^{\text{BS-RIS}}\tilde{\mathbf{Z}}$ and $\hat{\mathbf{f}}_k = \sqrt{1-(\tau^{\text{RIS-U}})^2}\mathbf{f}_k + \tau^{\text{RIS-U}}\tilde{\mathbf{z}}_k$, respectively, where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{z}}_k$ are the associated error matrix and vector with entries following the complex standard Gaussian distribution, and $\tau^{\text{BS-RIS}}$ and $\tau^{\text{RIS-U}}$ are the error coefficients of the corresponding links, similarly as in (5). Observe that the interference propagating through the cascade RIS channels, visible in the SINR of the common message in (3), also impacts the private messages. This implies that aligning $\hat{\mathbf{G}}$ with the channel vector $\hat{\mathbf{f}}_k$ of users $k \in \mathcal{K}$ should cause degradation to the rates of both common and private messages. With access only to

Algorithm 1: RIS-induced interference attack against RSMA

Input: $I, \delta \in (0,1), \{\omega_1, \cdots, \omega_K\}, \{\hat{\mathbf{K}}_1, \cdots, \hat{\mathbf{K}}_K\}$;

Initialize: $\boldsymbol{\theta}_{(1)} = \mathbf{1}_{L,1}$, $\Delta = \delta / \lambda_{\max}\left(\bar{\mathbf{K}}^H \bar{\mathbf{K}}\right)$, $\bar{\mathbf{K}} = \left[\sqrt{\omega_1}\hat{\mathbf{K}}_1^H \quad \cdots \quad \sqrt{\omega_K}\hat{\mathbf{K}}_K^H\right]^H$;

for $i = 1, 2, \cdots, I-1$ **do**

  Update in the direction of the gradient of (11a): $\boldsymbol{\vartheta} = \boldsymbol{\theta}_{(i)} + \Delta \bar{\mathbf{K}}^H \bar{\mathbf{K}} \boldsymbol{\theta}_{(i)}$;

  Compute the projection onto the unit 1-sphere: $\boldsymbol{\theta}_{(i+1)} = e^{j\angle(\boldsymbol{\vartheta})}$;

end

Output: $\boldsymbol{\Theta} = \mathrm{diag}\{\boldsymbol{\theta}_{(I)}\}$.

RIS-associated CSI, the attacker can launch an attack against RSMA through the following weighted-sum maximization

$$\arg\max_{\boldsymbol{\Theta}} \quad \sum_{k=1}^{K} \omega_k \|\hat{\mathbf{f}}_k^H \boldsymbol{\Theta} \hat{\mathbf{G}}\|_2^2, \tag{9a}$$

$$\text{s.t.} \quad \boldsymbol{\Theta} = \mathrm{diag}\{\mu_1 e^{-j\theta_1}, \cdots, \mu_L e^{-j\theta_L}\}, \tag{9b}$$

$$|\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}. \tag{9c}$$

where ω_k are weights that can be exploited to set the intensity of the attacks against each user. The current matrix structure of (9) is challenging to tackle. To achieve a tractable version of the problem, the attacker invokes the following Khatri-Rao identity: $\left(\mathbf{Z}^T \diamond \mathbf{X}\right) \mathrm{vecd}\{\mathbf{Y}\} = \mathrm{vec}\{\mathbf{XYZ}\}$. By relying on this property, we can define $\boldsymbol{\theta} \triangleq \mathrm{vecd}\{\boldsymbol{\Theta}\} \in \mathbb{C}^L$ and $\hat{\mathbf{K}}_k \triangleq \hat{\mathbf{G}}^T \diamond \hat{\mathbf{f}}_k^H \in \mathbb{C}^{M \times L}$. These transformations are then applied to (9), resulting in the following simpler problem

$$\arg\max_{\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \omega_k \left\| \hat{\mathbf{K}}_k \boldsymbol{\theta} \right\|_2^2, \tag{10a}$$

$$\text{s.t.} \quad |\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}. \tag{10b}$$

Lastly, the matrices in the objective function in (10a) are stacked to obtain the following equivalent problem

$$\arg\max_{\boldsymbol{\theta}} \quad \left\| \left[\sqrt{\omega_1}\hat{\mathbf{K}}_1^H \quad \cdots \quad \sqrt{\omega_K}\hat{\mathbf{K}}_K^H\right]^H \boldsymbol{\theta} \right\|_2^2, \tag{11a}$$

$$\text{s.t.} \quad |\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}. \tag{11b}$$

Despite the non-convex element-wise modulus constraint in (11b), the desired solution can be efficiently approximated through a projected gradient method, as in [8], [15], which is presented in Algorithm 1, in which $\lambda_{\max}(\cdot)$ computes the largest eigenvalue of a given matrix, Δ is the step size in the direction of the gradient, δ is a coefficient that controls the step size, and I denotes the number iterations.

C. Mitigation attack

By inspecting the expressions (3) and (4), it can be noticed that in the extreme case with no power allocated to the private messages, no interference will impact the SINR of the common message. In such a scenario, the worst effect that the attacks from the previous subsections can cause is to make the wireless channels mismatched with the precoder in (7), i.e., due to the unexpected contribution of the reflected RIS channels. Even though this might lead to performance degradation, the data rates on the common message are not interference-limited. In this case, the attacker must opt for an attack capable of mitigating the common signal to create a stronger impact. To this end, the RIS needs to be optimized

Algorithm 2: RIS-induced mitigation attack against RSMA

Input: $I, \delta \in (0,1), \{\omega_1, \cdots, \omega_K\}, \{\hat{\mathbf{K}}_1, \cdots, \hat{\mathbf{K}}_K\}, \{\hat{\mathbf{h}}_1, \cdots, \hat{\mathbf{h}}_K\}$;

Initialize: $\bar{\mathbf{K}} = \left[\sqrt{\omega_1}\hat{\mathbf{K}}_1 \quad \cdots \quad \sqrt{\omega_K}\hat{\mathbf{K}}_K\right]^H$, $\bar{\mathbf{h}} = \left[\sqrt{\omega_1}\hat{\mathbf{h}}_1^H \quad \cdots \quad \sqrt{\omega_K}\hat{\mathbf{h}}_K^H\right]^T$, $\boldsymbol{\theta}_{(1)} = e^{j\angle\left([(\bar{\mathbf{K}}^H\bar{\mathbf{K}})^{-1}\bar{\mathbf{K}}^H]\bar{\mathbf{h}}\right)}$, $\Delta = \delta / \lambda_{\max}\left(\bar{\mathbf{K}}^H \bar{\mathbf{K}}\right)$;

for $i = 1, 2, \cdots, I-1$ **do**

  Update in the opposite direction of the gradient of (14a): $\boldsymbol{\vartheta} = \boldsymbol{\theta}_{(i)} - \Delta \bar{\mathbf{K}}^H (\bar{\mathbf{K}} \boldsymbol{\theta}_{(i)} + \bar{\mathbf{h}})$;

  Compute the projection onto the unit 1-sphere: $\boldsymbol{\theta}_{(i+1)} = e^{j\angle(\boldsymbol{\vartheta})}$;

end

Output: $\boldsymbol{\Theta} = \mathrm{diag}\{\boldsymbol{\theta}_{(I)}\}$.

such that reflected channels add destructively with the direct BS-U channels. This can be accomplished with the following minimization problem

$$\arg\min_{\boldsymbol{\Theta}} \quad \sum_{k=1}^{K} \omega_k \|\hat{\mathbf{f}}_k^H \boldsymbol{\Theta} \hat{\mathbf{G}} + \hat{\mathbf{h}}_k^H\|_2^2, \tag{12a}$$

$$\text{s.t.} \quad \boldsymbol{\Theta} = \mathrm{diag}\{\mu_1 e^{-j\theta_1}, \cdots, \mu_L e^{-j\theta_L}\}, \tag{12b}$$

$$|\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}. \tag{12c}$$

By relying on the Khatri-Rao property introduced in the last subsection, problem (12) can be reformulated as follows

$$\arg\min_{\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \omega_k \|\hat{\mathbf{K}}_k \boldsymbol{\theta} + (\hat{\mathbf{h}}_k^H)^T\|_2^2, \tag{13a}$$

$$\text{s.t.} \quad |\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}, \tag{13b}$$

where $\boldsymbol{\theta}$ is the vector of reflecting coefficients, and $\hat{\mathbf{K}}_k$ is the matrix from the Khatri-Rao product, as in subsection III-B.

For solving (13), the terms of the sum in its objective function are stacked vertically so that the following is achieved

$$\arg\min_{\boldsymbol{\theta}} \quad \left\| \begin{bmatrix} \sqrt{\omega_1}\hat{\mathbf{K}}_1^H \\ \vdots \\ \sqrt{\omega_K}\hat{\mathbf{K}}_K^H \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \sqrt{\omega_1}(\hat{\mathbf{h}}_1^H)^T \\ \vdots \\ \sqrt{\omega_K}(\hat{\mathbf{h}}_K^H)^T \end{bmatrix} \right\|_2^2, \tag{14a}$$

$$\text{s.t.} \quad |\mu_l|^2 = 1, \ \forall l \in \{1, \cdots, L\}. \tag{14b}$$

The above problem is non-convex due to the element-wise unity modulus constraint. Still, like (11), this class of problems can be solved sub-optimally by exploiting a projected gradient strategy. The implemented approach is provided in Algorithm 2. Note that since all signals transmitted by the BS propagate through the same channels, this attack should impact the detection of both common and private messages. Therefore, users will experience performance degradation independently of the amount of power allocated to either of the messages. Consequently, in contrast to the other considered attacks, the mitigation scheme does not require the attacker to know when the power allocation training is carried out. This is important since for this adversarial scheme to be effective, the attacker needs the knowledge of estimates of the channels of the BS-RIS link, $\hat{\mathbf{G}}$, RIS-U link, $\hat{\mathbf{f}}_k$, and the legitimate BS-U link, $\hat{\mathbf{h}}_k^H$, similar to the attack proposed in [9]. Depending on the attacker's capabilities, obtaining the latter estimates may be costly or impractical. Thus, in our numerical results, we will evaluate a scenario where the attacker can only access imperfect channel estimates.

IV. Numerical Results

In this section, we investigate the severity of the proposed attacks against RSMA and the impact of different system parameters. We compare the performance of RSMA and the conventional SDMA, i.e., when $\alpha^c = 0$, to assess their robustness against the presented threats.

We implement the communication scenario where $K = 3$ single-antenna users are connected to a BS with $M = 10$ transmit antennas. The coordinates of users $1, 2$, and 3 are fixed at $(30, 15)$ m, $(50, 15)$ m, and $(55, 10)$ m, respectively, and the BS at $(0, 0)$ m. Moreover, the attacker's RIS comprises $L = 200$ reflecting elements and is deployed at the coordinate $(40, 5)$ m. With this setup, the path-loss coefficients are calculated as $(d^{\text{BS-RIS}})^{-\eta}$, $(d_k^{\text{RIS-U}})^{-\eta}$, and $(d_k^{\text{BS-U}})^{-\eta}$, for $k \in \{1, 2, 3\}$, where $d^{\text{BS-RIS}}$, $d_k^{\text{RIS-U}}$, and $d_k^{\text{BS-U}}$ represent the distances of the links BS-RIS, RIS-U, and BS-U, respectively, with η denoting the path-loss exponent, set to 2.5. As for Algorithms 1 and 2, the step size parameter is configured as $\delta = 0.99$, the number of iterations as $I = 3 \times 10^3$, and the weights are set to $\omega_1 = \omega_2 = \omega_3 = 1/3$. Furthermore, in the SDMA scheme, the total transmit power of the BS is allocated uniformly among the users, i.e., the fraction P/K is allocated to each user, and the noise variance is set to $\sigma^2 = -50$ dBm.

We start with Fig. 2, which presents sum rate curves for the case where both the BS and the attacker can estimate the channels perfectly. Because the BS has access to perfect CSI, the private precoders can completely remove inter-user interference. As a result, the power allocation strategy in subsection II-C, which cannot detect the RIS interference, will assign power primarily to the private messages, making RSMA perform identically as SDMA in all tested cases. As can be seen, in this ideal scenario with perfect CSI, all three kinds of RIS-induced attacks are able to severely deteriorate the performance of both RSMA and SDMA, with the random interference rendering the mildest impact, the aligned interference the second strongest, and the mitigation attack the strongest impact.

In Fig. 3, we can visualize the behavior of the considered multiple access schemes for the scenario with imperfect CSI at both the BS and the attacker, considering an error factor of $\tau^{\text{BS-U}} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = 0.3$. In this scenario, we see that the curves for the safe system are lower than those observed under perfect CSI in Fig. 2. Moreover, the sum rate experienced with SDMA saturates as a consequence of the dominant inter-user interference in high transmit power values, either due to imperfect CSI or the RIS attacks, or both. On the other hand, the flexible power allocation policy of RSMA shifts power to the common message as the detected interference from imperfect CSI starts to grow. Note that, although the BS is blind to the RIS-induced interference, assigning power to the common message as a way to overcome degradation from imperfect CSI can also significantly alleviate (unintentionally) the impact of the attacks. Also, even though the different attacks can still cause performance degradation to RSMA, the sum rate curves are no longer interference-limited, i.e., the

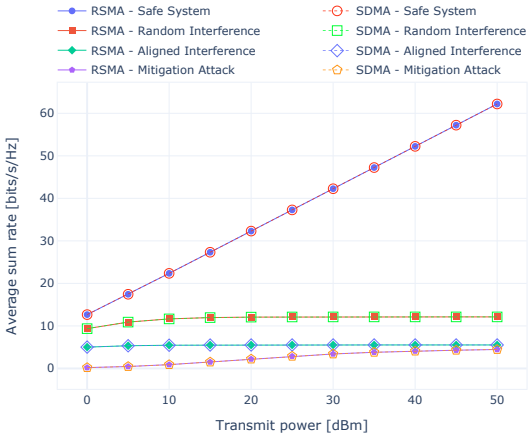

Fig. 2: Average sum rate for different attack strategies with perfect CSI in all links, i.e., $\tau^{\text{BS-U}} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = 0$ at both the BS and the attacker.

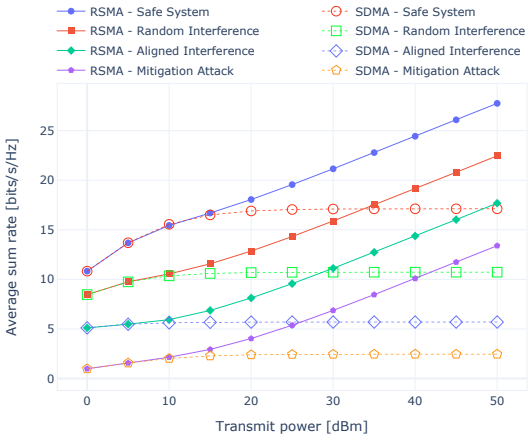

Fig. 3: Average sum rate for different attack strategies with imperfect CSI in all links, with $\tau^{\text{BS-U}} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = 0.3$ at both the BS and the attacker.

RSMA curves do not saturate as in the SDMA schemes. This confirms that RSMA is remarkably more robust than SDMA, even under such threat scenarios that may be difficult to detect.

Lastly, Figs. 4 and 5 investigate the impact of the CSI quality on the attacks' severity. In Fig. 4, specifically, we test different channel error values at the attacker, such that $\tilde{\tau} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = \tau^{\text{BS-U}}$, while considering perfect CSI at the BS. The figure shows that the impact of both the aligned interference and the mitigation attack diminishes with the increase of the channel error factor, approaching the performance observed under the random interference attack as $\tilde{\tau}$ gets large. Moreover, it is noteworthy that even with a high channel error, the mitigation attack remains the most impactful one, still slightly outperforming its aligned interference counterpart when $\tilde{\tau} = 0.9$. It can also be observed that, because the BS operates under perfect CSI, independently of the channel error at the attacker all attacks make the sum rate curves saturate in the high-power regime. This interference-limited behavior changes in Fig. 5, which investigates different error levels at the attacker but considers imperfect CSI at the BS, with $\tau^{\text{BS-U}} = 0.3$. Again, the dominance of the mitigation attack persists even under high levels of error. These results indicate that if the attacker acquires at least imperfect estimates of both BS-U and RIS-induced channels, the mitigation strategy is the

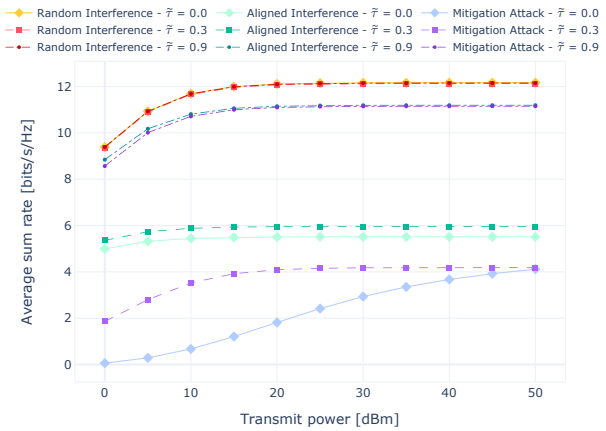

Fig. 4: Average sum rate for RSMA under different attack strategies with perfect CSI at the BS, i.e., $\tau^{\text{BS-U}} = 0$ for the BS, and various CSI error levels at the attacker, considering $\tilde{\tau} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = \tau^{\text{BS-U}}$ for the attacker.

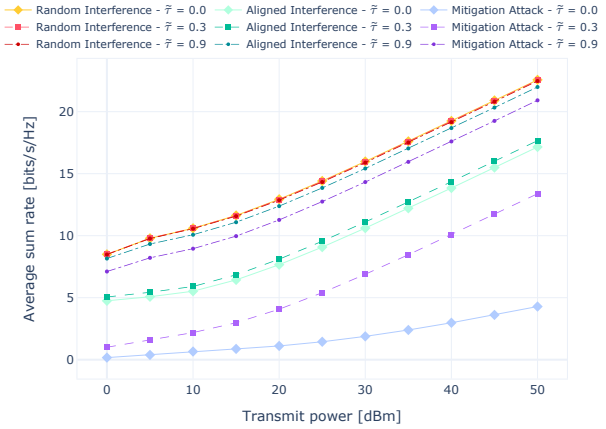

Fig. 5: Average sum rate for RSMA under different attack strategies with $\tau^{\text{BS-U}} = 0.3$ for the BS, and various CSI error levels at the attacker, considering $\tilde{\tau} = \tau^{\text{BS-RIS}} = \tau^{\text{RIS-U}} = \tau^{\text{BS-U}}$ for the attacker.

attack that poses the highest risk of performance degradation. Nevertheless, the flexible interference management of RSMA can reverse the impacts of the attacks to some extent, as long as $\tau^{\text{BS-U}} > 0$ at the BS. Counterintuitively, operating under imperfect CSI, instead of being detrimental, makes RSMA more robust to such adversarial RIS attacks.

V. Conclusions

This paper covered three potential RIS-induced attacks that can harm the performance of RSMA, namely random interference, aligned interference, and mitigation attack. For the two latter attacks, we presented algorithms based on projected gradient methods that can efficiently find RIS coefficients that lead to a strong degradation of the data rates of all connected users. Comprehensive simulation results demonstrated the severity of the different malicious schemes and revealed that RSMA can be robust even when the BS is blind to the attacks. We demonstrated that by smartly allocating power to the common message to avoid interference from imperfect CSI, RSMA can deliver data rates that considerably outperform SDMA under the presented security threats. In future work, we will explore strategies for further improving the robustness of RSMA and methods for countering such attacks.

Acknowledgments

This work was supported by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme within Hexa-X-II project (grant no. 101095759), by Business Finland via the 6GBridge - Local 6G project (grant no. 8002/31/2022), and by the Research Council of Finland through the 6G Flagship (grant no. 346208) and the 6G-ConCoRSe project (grant no. 24300065). This work also received support from the Commonwealth Cyber Initiative (CCI) in Virginia, US, an investment in the advancement of cyber R&D, innovation, and workforce development, and by the National Science Foundation under grants no. 2318798 and 2326599.

References

[1] B. Clerckx, Y. Mao, E. A. Jorswieck, J. Yuan, D. J. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, 2023.

[2] A. S. de Sena, P. H. J. Nardelli, D. B. da Costa, P. Popovski, and C. B. Papadias, "Rate-splitting multiple access and its interplay with intelligent reflecting surfaces," *IEEE Commun. Mag.*, vol. 60, no. 7, pp. 52–57, 2022.

[3] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, 2016.

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

[5] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging generalizing and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Networking*, no. 133, pp. 1–54, May 2018.

[6] H. Huang, Y. Zhang, H. Zhang, Y. Cai, A. L. Swindlehurst, and Z. Han, "Disco intelligent reflecting surfaces: Active channel aging for fully-passive jamming attacks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 806–819, 2024.

[7] H. Huang, Y. Zhang, H. Zhang, C. Zhang, and Z. Han, "Illegal intelligent reflecting surface based active channel aging: When jammer can attack without power and CSI," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 11 018–11 022, 2023.

[8] A. S. de Sena, J. Kibilda, N. H. Mahmood, A. Gomes, and M. Latva-aho, "Malicious RIS versus massive MIMO: Securing multiple access against RIS-based jamming attacks," *IEEE Wireless Commun. Lett.*, vol. 13, no. 4, pp. 989–993, 2024.

[9] B. Lyu, D. T. Hoang, S. Gong, D. Niyato, and D. I. Kim, "IRS-based wireless jamming attacks: When jammers can attack without power," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1663–1667, 2020.

[10] P. Staat, H. Elders-Boll, M. Heinrichs, C. Zenger, and C. Paar, "Mirror, mirror on the wall: Wireless environment reconfiguration attacks based on fast software-controlled surfaces," in *Proc. Asia Conf. on Computer and Communications Security*, 2022, p. 208–221.

[11] H. Chen and Y. Ghasempour, "Malicious mmWave reconfigurable surface: Eavesdropping through harmonic steering," in *Proc. Int. Workshop on Mobile Computing Systems and Applications*, 2022, pp. 54–60.

[12] Y. Wang, H. Lu, D. Zhao, Y. Deng, and A. Nallanathan, "Wireless communication in the presence of illegal reconfigurable intelligent surface: Signal leakage and interference attack," *IEEE Wireless Commun.*, vol. 29, no. 3, pp. 131–138, 2022.

[13] H. Wang, Z. Han, and A. L. Swindlehurst, "Channel reciprocity attacks using intelligent surfaces with non-diagonal phase shifts," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1469–1485, 2024.

[14] A. Konar and N. D. Sidiropoulos, "Fast approximation algorithms for a class of non-convex QCQP problems using first-order methods," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3494–3509, 2017.

[15] J. Tranter, N. D. Sidiropoulos, X. Fu, and A. Swami, "Fast unit-modulus least squares with applications in beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2875–2887, 2017.