Malicious RIS versus Massive MIMO: Securing Multiple Access against RIS-based Jamming Attacks

Arthur Sousa de Sena, *Member*, *IEEE*, Jacek Kibiłda, *Senior Member*, *IEEE*,
Nurul Huda Mahmood, André Gomes, *Member*, *IEEE*, Matti Latva-aho, *Fellow*, *IEEE*

***Abstract*—In this letter, we study an attack that leverages a reconfigurable intelligent surface (RIS) to induce harmful interference toward multiple users in massive multiple-input multiple-output (mMIMO) systems during the data transmission phase. We propose an efficient and flexible weighted-sum projected gradient-based algorithm for the attacker to optimize the RIS reflection coefficients without knowing legitimate user channels. Simulation results demonstrate that our malicious algorithm outperforms baseline strategies while offering adaptability for targeting specific users. To counter such a threat, we propose two reception strategies, which we show to be effective even if only an imperfect estimate of the cascade RIS channel is available.**

***Index Terms*—Reconfigurable intelligent surface, massive MIMO, passive jamming, physical-layer security**

I. Introduction

Reconfigurable intelligent surface (RIS) is a low-cost and low-power alternative to RF repeaters that is meant to enhance service availability and resilience of next-generation wireless networks by dynamically controlling the reflection of impinging electromagnetic waves. Coupled with a passive mode of operation, it is also an attractive technology in adversarial applications. A malicious actor may deploy a RIS or hijack an existing one to easily trigger attacks that utilize impinging signals to attack wireless communication links.

Ample recent works have proposed a variety of attacks, including creating destructive multipath [1], assisting an active jammer [2], or an eavesdropper [3]. However, a potentially more effective way to attack legitimate systems with a RIS hinges on exploiting vulnerabilities in wireless system design. For instance, active RIS architectures can create amplified destructive millimeter-wave beamforming [4], or alter RIS reflective coefficients between the transmission of data and control symbols, resulting in a substantial degradation in symbol error rate [5]. Similarly, by introducing malicious reflections during the channel estimation process, a malicious RIS can reduce the effectiveness of channel equalization [6], physical layer key generation rate [7], or corrupt the beamforming vectors [8]. Vulnerabilities in channel estimation are particularly critical for multiple access methods in massive MIMO (mMIMO) like space-division multiple access (SDMA).

SDMA-aided mMIMO involves two stages: the channel state information (CSI) acquisition, based on transmitted pilot sequences, and data transmission using a precoder design based on the acquired CSI. The dependency on CSI accuracy makes mMIMO systems vulnerable to pilot contamination attacks triggered by active transmitters [9]. However, this attack can also be conducted without additional power by altering the RIS phase shifts between pilot and data transmission. More worryingly, it was shown in [10], [11] that an attack against legitimate users can be made effective even with random RIS phase shifts, named "Disco Ball" attack.

In this letter, we propose a new and more potent version of the attack against downlink data transmission to multi-antenna users in a mMIMO system employing SDMA, whereby the malicious RIS does not randomly select the phase shifts, as in the Disco Ball strategy [10], [11], but efficiently optimizes them to increase the damage, all without the knowledge of legitimate channels. To accomplish this, we develop an efficient and flexible weighted-sum projected gradient-based algorithm to optimize the RIS coefficients. Subsequently, we propose two reception strategies that require only an estimate of the effective cascade RIS channel to mitigate the adverse effects of the attack. Simulation results demonstrate that our malicious algorithm outperforms baseline strategies in unleashing more powerful attacks and exhibits adaptability for targeting specific users. Our results also show that the proposed mitigation strategies are effective even when only an imperfect estimate of the cascade RIS channel is available.

Notation: The ith element of a vector $\mathbf{a}$ is denoted by $[\mathbf{a}]_i$, the (ij) entry of a matrix $\mathbf{A}$ by $[\mathbf{A}]_{ij}$, the submatrix of $\mathbf{A}$ formed by its rows (columns) from i to j by $[\mathbf{A}]_{i:j,:}$ ($[\mathbf{A}]_{:,i:j}$). The transpose and Hermitian transpose of $\mathbf{A}$ are represented by $\mathbf{A}^T$ and $\mathbf{A}^H$, respectively, $\mathbf{I}_M$ is the $M \times M$ identity matrix, $\mathbf{0}_{M,N}$ is the $M \times N$ zero matrix, and $\diamond$ represents the Khatri-Rao product. The operator $\mathrm{vec}\{\cdot\}$ transforms an $M \times N$ matrix into a column vector, $\mathrm{vecd}\{\cdot\}$ converts the diagonal elements of an $M \times M$ matrix into a column vector, $\mathrm{diagm}\{\cdot\}$ creates an $M \times M$ diagonal matrix from the diagonal elements of an arbitrary $M \times M$ matrix, $\mathrm{diag}\{\cdot\}$ transforms a vector of length M into an $M \times M$ diagonal matrix, $\angle(z)$ returns the phase of the complex number z, and $\mathrm{E}\{\cdot\}$ denotes expectation.

Arthur S. de Sena, Nurul H. Mahmood, and Matti Latva-aho are with the University of Oulu, Oulu, Finland (email: arthur.sena@oulu.fi, nurulhuda.mahmood@oulu.fi, matti.latva-aho@oulu.fi).

Jacek Kibiłda and André Gomes are with the Commonwealth Cyber Initiative, Virginia Tech, USA (email: jkibilda@vt.edu, gomesa@vt.edu).

The research leading to this paper received support from 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 Agreement No 101095759), the Academy of Finland under the 6G Flagship program (Grant No 346208) and the academy project ReWIN-6G (Grant No 357120). This material is also based upon work supported by the National Science Foundation, under Grants No. 2326599 and 2318798, and the Commonwealth Cyber Initiative.

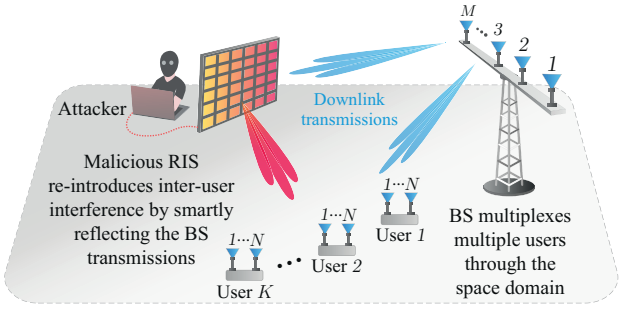

Fig. 1: An attacker uses a RIS to perform an interference attack against data transmission in downlink multi-user mMIMO.

II. System Model

We study a downlink multi-user mMIMO system, consisting of one base station (BS) employing M antennas and K spatially distributed users, represented by the index set $\mathcal{K} = \{1, 2, \cdots, K\}$, where each user is equipped with N antennas, such that $M \gg N$. The BS spatially multiplexes the users using an SDMA strategy, in which linear precoders are employed. In the presence of accurate CSI, such precoders can efficiently tackle mMIMO inter-user interference. However, as illustrated in Fig. 1, the system is jammed by an attacker that controls a malicious RIS comprising L reflecting elements. It is assumed the attacker has enough computational power to acquire the CSI of the cascade RIS channel, but not the direct channels between the users and the BS. With this information, the attacker can recycle impinging signals from the BS and passively re-introduce inter-user interference that conventional precoders cannot eliminate. Under the described scenario, the kth user receives the following signal

$$\mathbf{y}_k = \left(\mathbf{F}_k^H \mathbf{\Theta} \mathbf{G} + \mathbf{H}_k^H\right) \sum_{i=1}^{K} \mathbf{P}_i \sqrt{P} \mathbf{x}_i + \mathbf{n}_k \in \mathbb{C}^N, \quad (1)$$

where $\mathbf{x}_i = [\sqrt{\alpha_{i,1}} x_{i,1}, \cdots, \sqrt{\alpha_{i,S}} x_{i,S}]^T \in \mathbb{C}^S$ is the data vector comprising S symbols intended for the ith user, satisfying $\mathrm{E}\{|x_{i,s}|^2\} = 1$, with $\alpha_{i,s}$ denoting the power allocation coefficient for the sth symbol, P represents the transmit power budget, $\mathbf{P}_i \in \mathbb{C}^{M \times S}$ is the precoding matrix responsible for multiplexing users in space, $\mathbf{n}_k \in \mathbb{C}^N$ is the noise vector[1], whose entries follow the complex Gaussian distribution with zero mean and variance σ^2, and $\mathbf{\Theta}$ is the diagonal reflection matrix of the malicious RIS, whose amplitude and phase-shift coefficients associated with the lth reflecting element satisfy the constraints $|[\mathbf{\Theta}]_{ll}| = 1$ and $\angle([\mathbf{\Theta}]_{ll}) \in [0, 2\pi], \forall l = 1, \cdots, L$. The matrices $\mathbf{H}_k \in \mathbb{C}^{M \times N}$, $\mathbf{G} \in \mathbb{C}^{L \times M}$, and $\mathbf{F}_k \in \mathbb{C}^{L \times N}$ represent the 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, and are assumed to have correlated entries.

A. Precoding for User Multiplexing

To ensure the effectiveness of the passive RIS jamming attack, the attacker maliciously stays idle during the BS channel estimation phase. This means the BS can rely only on the jamming-free channels observed on each BS-U link to construct the precoding matrix $\mathbf{P}_k$. The BS then employs a two-layered spatial multiplexing strategy in which the desired precoder is structured as $\mathbf{P}_k \triangleq \mathbf{K}_k \mathbf{D}_k$, where $\mathbf{K}_k$ addresses inter-user interference and $\mathbf{D}_k$ maximizes the reception of intended signals at the kth user. By recalling the truncated eigendecomposition, the channel covariance matrix of the direct link BS-U for the kth user can be expressed as $\mathbf{\Sigma}_k \triangleq \mathrm{E}\{\mathbf{H}_k \mathbf{H}_k^H\} = \bar{\mathbf{U}}_k \bar{\mathbf{\Delta}}_k \bar{\mathbf{U}}_k^H$, where $\bar{\mathbf{\Delta}}_k \in \mathbb{R}^{r_k \times r_k}$ is a diagonal matrix comprising r_k nonzero eigenvalues of $\mathbf{\Sigma}_k$, and $\bar{\mathbf{U}}_k \in \mathbb{C}^{M \times r_k}$ is a tall matrix containing the associated eigenvectors. The users' channel matrices can then be factorized as $\mathbf{H}_k = \bar{\mathbf{U}}_k \bar{\mathbf{\Delta}}_k^{\frac{1}{2}} \bar{\mathbf{H}}_k$, where $\bar{\mathbf{H}}_k \in \mathbb{C}^{r_s \times N}$ is the reduced-dimension fast-fading channel matrix, whose entries follow the complex Gaussian distribution with zero mean and unity variance. Thus, the inter-user interference can be tackled if $[\bar{\mathbf{U}}_1 \ \cdots \ \bar{\mathbf{U}}_{k-1} \ \bar{\mathbf{U}}_{k+1} \ \cdots \ \bar{\mathbf{U}}_K]^H \mathbf{K}_k = \mathbf{\Lambda}_k^H \mathbf{K}_k = \mathbf{0}, \forall k \in \mathcal{K}$, where $\mathbf{\Lambda}_k \in \mathbb{C}^{M \times \sum_{k' \neq k} r_{k'}}$ is a block matrix with the eigenvectors corresponding to the non-zero eigenvalues of non-intended users. This implies that $\mathbf{K}_k$ must be designed from the orthogonal complement of the column space of $\mathbf{\Lambda}_k$, which can be achieved with the aid of its full singular value decomposition (SVD). More specifically, let $\mathbf{U}_{\mathbf{\Lambda}_k}$ denote the left eigenvector matrix of $\mathbf{\Lambda}_k$ computed via SVD. Then, since the last $M - \sum_{k' \neq k} r_{k'}$ columns of $\mathbf{U}_{\mathbf{\Lambda}_k}$ provide the orthonormal basis for the desired orthogonal complement, the precoder for the kth user can be computed as

$$\mathbf{K}_k = \frac{1}{\sqrt{S}} [\mathbf{U}_{\mathbf{\Lambda}_k}]_{:,(M-S+1):M} \in \mathbb{C}^{M \times S}, \quad (2)$$

where the parameter S sets the number of parallel symbols to be transmitted to each user, which must satisfy $S \leq \left(M - \sum_{k' \neq k} r_{k'}\right)$ and $S \leq \min\{r_k\}, \forall k \in \mathcal{K}$.

On the other hand, we design the precoder $\mathbf{D}_k$ to maximize the received power, i.e., $\max_{\mathbf{D}_k} \|\mathbf{H}_k^H \mathbf{K}_k \mathbf{D}_k\|_F^2$, at the kth user. This goal can be achieved through the full SVD of the projected channel $\mathbf{H}_k^H \mathbf{K}_k \in \mathbb{C}^{N \times S}$. Specifically, by recalling the SVD, we can decompose $\mathbf{H}_k^H \mathbf{K}_k = \hat{\mathbf{U}}_k \hat{\mathbf{\Delta}}_k \hat{\mathbf{V}}_k^H$. Then, the inner precoder can be obtained as $\mathbf{D}_k = \hat{\mathbf{V}}_k \in \mathbb{C}^{S \times S}$.

III. RIS-Assisted Attack Design

With the precoder in (2), the BS can effectively suppress inter-user interference experienced in the link BS-U. However, because the BS cannot detect the malicious RIS during channel estimation, inter-user interference propagating through the reflected BS-RIS-U link will inevitably be re-introduced to the users. Thus, the kth user will observe the following signal

$$\mathbf{y}_k = \mathbf{H}_k^H \mathbf{P}_k \sqrt{P} \mathbf{x}_k + \mathbf{F}_k^H \mathbf{\Theta} \mathbf{G} \sum_{i=1}^{K} \mathbf{P}_i \sqrt{P} \mathbf{x}_i + \mathbf{n}_k. \quad (3)$$

The attacker aims to re-introduce as much inter-user interference as possible in the system to degrade the detection of intended symbols at each targeted user $k \in \mathcal{K}$. To this end, the reflecting coefficients of the deployed RIS are tuned to steer all transmissions coming from the BS to jam multiple access.

The challenge is that the attacker is unlikely to have access to the legitimate BS-U channels or the BS precoders. Despite this, it can still exploit the RIS channels to boost inter-user interference. In this case, the attacker will attempt to match the channels of the BS-RIS link with those of the RIS-U

[1]Since the attacker controls only a passive RIS with no active amplifiers, any thermal noise introduced by the RIS reflecting elements is negligible.

link by maximizing $\left\|\mathbf{F}_k^H \boldsymbol{\Theta} \mathbf{G}\right\|_F^2$. However, since a single RIS has been deployed, the attacker cannot find a global set of reflecting coefficients to achieve its objective optimally for all users. Alternatively, the attacker may apply a flexible jamming framework to tune the intensity of the attack for each user. This framework is implemented by computing the Pareto optimal points for the following weighted-sum problem:

$$\max_{\boldsymbol{\Theta}} \sum_{k=1}^{K} \nu_k \left\|\mathbf{F}_k^H \boldsymbol{\Theta} \mathbf{G}\right\|_F^2, \tag{4a}$$

$$\text{s.t. } |[\boldsymbol{\Theta}]_{mn}| = 1,\ \forall m,n \in \{1,\cdots,L\} \mid m = n, \tag{4b}$$

$$|[\boldsymbol{\Theta}]_{mn}| = 0,\ \forall m,n \in \{1,\cdots,L\} \mid m \neq n, \tag{4c}$$

where the scalars ν_k are the optimization weights that the attacker can exploit to adjust the levels of interference to the different users. Note that constraint (4b) sets the amplitudes of the reflection coefficients to one, modeling the passive operation of the RIS, while (4c) ensures that $\boldsymbol{\Theta}$ is a diagonal matrix. These constraints, plus the matrix form of the objective function in (4a), make the problem difficult to solve. To tackle this complication, the attacker transforms the original formulation by exploiting the Khatri-Rao factorization:

$$\left(\mathbf{B}^T \diamond \mathbf{A}\right) \operatorname{vecd}\{\mathbf{X}\} = \operatorname{vec}\{\mathbf{AXB}\}, \tag{5}$$

where $\mathbf{X}$ is diagonal and $\mathbf{A}$ and $\mathbf{B}$ are arbitrary matrices of compatible dimensions. Specifically, by invoking (5), the following transformations can be applied:

$$\boldsymbol{\theta} \triangleq \operatorname{vecd}\{\boldsymbol{\Theta}\} \in \mathbb{C}^L, \qquad \mathbf{S}_k \triangleq \mathbf{G}^T \diamond \mathbf{F}_k^H \in \mathbb{C}^{MN \times L}.$$

Then, by plugging the above definitions into (4), the attacker achieves the following equivalent problem

$$\max_{\boldsymbol{\theta}} \sum_{k=1}^{K} \nu_k \left\|\mathbf{S}_k \boldsymbol{\theta}\right\|_2^2, \tag{6a}$$

$$\text{s.t. } |[\boldsymbol{\theta}]_n| = 1,\ \forall n \in \{1,\cdots,L\}. \tag{6b}$$

To simplify further the above optimization, the weighted sum in (6a) is transformed into a matrix equivalent form by vertically concatenating each term of the sum, as follows

$$\max_{\boldsymbol{\theta}} \left\| \begin{bmatrix} \sqrt{\nu_1}\mathbf{S}_1 \\ \vdots \\ \sqrt{\nu_K}\mathbf{S}_K \end{bmatrix} \boldsymbol{\theta} \right\|_2^2, \tag{7a}$$

$$\text{s.t. } |[\boldsymbol{\theta}]_n| = 1,\ \forall n \in \{1,\cdots,L\}. \tag{7b}$$

The problem in (7) is non-convex due to the element-wise modulus constraint and is, in general, NP-hard [12]. To overcome the challenge, we propose a suboptimal but practical strategy utilizing a projected gradient-based algorithm. The proposed strategy is presented in Algorithm 1, where T is the number of iterations, β is a parameter used to tune the step size α, and $\lambda_{\max}(\cdot)$ returns the largest eigenvalue of its argument. As demonstrated in [12, Theorem 1], this algorithm converges to a Karush-Kuhn-Tucker (KKT) point of the original problem as long as $\alpha < 1/\lambda^*$, with λ^* denoting the largest eigenvalue of $\bar{\mathbf{S}}^H \bar{\mathbf{S}}$, defined in line 1 of Algorithm 1.

It should be noted that the solution to (7) represents a feasible attack that may cause significant performance degradation to the users, as demonstrated in our simulation results. However, it does not represent a system-wide optimal RIS attack, for it does not assume that the attacker has access to important parameters, such as BS precoders and BS-U channel matrices, which may be difficult, if not impossible, to obtain by the attacker.

Algorithm 1: Malicious RIS optimization for multi-user jamming attacks in mMIMO systems

Input: $T, \beta \in (0,1), \{\nu_1,\cdots,\nu_K\}, \{\mathbf{S}_1,\cdots,\mathbf{S}_K\}$;

1 Initialize: $\bar{\mathbf{S}} = \begin{bmatrix} \sqrt{\nu_1}\mathbf{S}_1 \\ \vdots \\ \sqrt{\nu_K}\mathbf{S}_K \end{bmatrix}$, $\boldsymbol{\theta}_{(1)} = \mathbf{I}_{L,1}$, $\lambda^\star = \lambda_{\max}\left(\bar{\mathbf{S}}^H \bar{\mathbf{S}}\right)$, $\alpha = \frac{\beta}{\lambda^\star}$;

2 **for** $t = 1, 2, \cdots, T-1$ **do**

3 Compute the gradient of (7a): $\boldsymbol{\delta}_{(t)} = \bar{\mathbf{S}}^H\left(\bar{\mathbf{S}}\boldsymbol{\theta}_{(t)}\right)$;

4 Perform the gradient step: $\boldsymbol{\phi}_{(t+1)} = \boldsymbol{\theta}_{(t)} + \alpha\boldsymbol{\delta}_{(t)}$;

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

6 **end**

Output: $\boldsymbol{\Theta} = \operatorname{diag}\{\boldsymbol{\theta}_{(T)}\}$.

IV. Reception and Attack Mitigation

The BS precoders alone cannot cancel the attack by the RIS. This section addresses this issue by proposing two detection strategies to be deployed on the user side. In particular, we propose a two-layer reception strategy of the form $\mathbf{Q}_k\mathbf{W}_k$, where $\mathbf{W}_k$ is responsible for tackling the malicious RIS reflections and $\mathbf{Q}_k$ is a reception matrix responsible for removing inter-symbol interference. Moreover, while the CSI of the link BS-U is available to the network, we assume that only an estimate of the cascade BS-RIS-U channel $\mathbf{Z}_k^H \triangleq \mathbf{F}_k^H \boldsymbol{\Theta} \mathbf{G}$ can be made available to the users[2], which we model as follows

$$\hat{\mathbf{Z}}_k \triangleq \sqrt{1-\tau^2}\mathbf{Z}_k + \tau\mathbf{E}_k, \tag{8}$$

where $\mathbf{E}_k$ is the error matrix, which is independent of $\mathbf{Z}_k$ and whose entries follow the standard complex Gaussian distribution, and $\tau \in [0,1]$ models the level of error between the estimate $\hat{\mathbf{Z}}_k$ and the true channel $\mathbf{Z}_k$.

A. Full MITigation (F-MIT) of RIS attacks

By considering all RIS-reflected signals as interference, this first strategy aims to fully mitigate the attack. To this end, let us define $\tilde{\mathbf{P}} \triangleq [\mathbf{P}_1 \ \cdots \ \mathbf{P}_K] \in \mathbb{C}^{M \times KS}$ and $\tilde{\mathbf{x}} \triangleq [\mathbf{x}_1^T \ \cdots \ \mathbf{x}_K^T]^T \in \mathbb{C}^{KS}$. Then, (3) can be rewritten as

$$\mathbf{y}_k = \mathbf{H}_k^H \mathbf{P}_k \sqrt{P}\mathbf{x}_k + \mathbf{Z}_k^H \tilde{\mathbf{P}} \sqrt{P}\tilde{\mathbf{x}} + \mathbf{n}_k. \tag{9}$$

Now, to mitigate the attack, the kth user needs an inner receptor $\mathbf{W}_k$ that can achieve $\mathbf{W}_k\mathbf{Z}_k^H\tilde{\mathbf{P}} \approx \mathbf{0}_{S,S}$. Thus, $\mathbf{W}_k$ should lie in the left null space of the transformed channel matrix $\underline{\mathbf{Z}}_k \triangleq \mathbf{Z}_k^H\tilde{\mathbf{P}} \in \mathbb{C}^{N \times KS}$. Since the perfect knowledge of $\mathbf{Z}_k$ is not possible, users approximate the desired left null space from the left eigenvector matrix $\mathbf{U}_{\hat{\underline{\mathbf{Z}}}_k} \in \mathbb{C}^{N \times N}$ of the estimated channel $\hat{\underline{\mathbf{Z}}}_k \triangleq \hat{\mathbf{Z}}_k^H\tilde{\mathbf{P}} \in \mathbb{C}^{N \times KS}$. Note that, in addition to canceling interference, the matrix $\mathbf{W}_k$, with dimension $W \times N$, should satisfy $N \geq W \geq S$ for $\mathbf{Q}_k$ to remove the inter-symbol interference. Provided that $N > KS$ is satisfied, $\hat{\underline{\mathbf{Z}}}_k$ will have full column rank and a left null space of dimension $N - KS$. As a result, the perfect orthogonality with the estimated channel, i.e., $\mathbf{W}_k\hat{\underline{\mathbf{Z}}}_k = \mathbf{0}_{W,S}$, is achieved if and only if $W \leq N - KS$, otherwise residual interference

[2]Advanced interference estimation techniques can be employed in the users' devices to estimate the cascade RIS channels. However, this goes beyond the objectives of this work and arises as an interesting research direction.

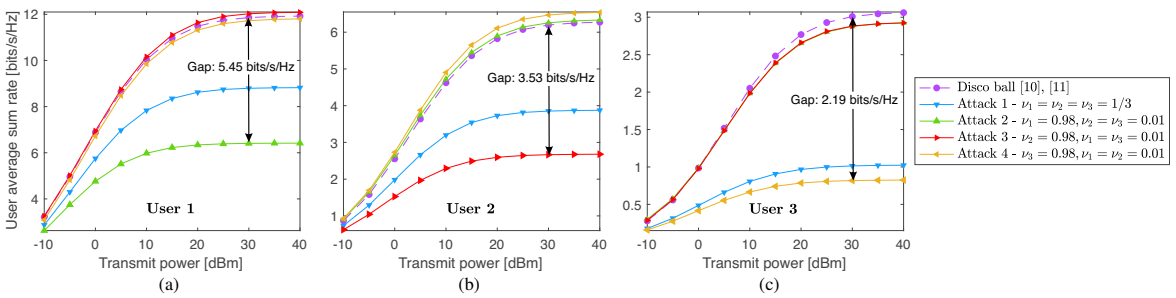

Fig. 2: User average sum rate under different attack scenarios: (a) for user 1, (b) for user 2, and (c) for user 3, considering $N = 4$ and $L = 200$.

is left. With these observations, $\mathbf{W}_k$ can be constructed from the last W rows of $\mathbf{U}^H_{\underline{\hat{\mathbf{Z}}}_k}$, as follows

$$\mathbf{W}_k = \left[\mathbf{U}^H_{\underline{\hat{\mathbf{Z}}}_k}\right]_{(N-W+1):N,:} \in \mathbb{C}^{W\times N}, \tag{10}$$

where we should satisfy $N > KS$ for the existence of a non-trivial left null space. Moreover, we adjust $W = \max\{N - KS, S\}$, which satisfies the requirement $W \geq S$.

Now, since the resulting effective channel $\mathbf{W}_k\mathbf{H}_k^H\mathbf{P}_k$ is an $W \times S$ matrix, with $S \leq W$, we can remove the inter-symbol interference by computing its left inverse, as follows

$$\mathbf{Q}_k = [(\mathbf{W}_k\mathbf{H}_k^H\mathbf{P}_k)^H\mathbf{W}_k\mathbf{H}_k^H\mathbf{P}_k]^{-1}(\mathbf{W}_k\mathbf{H}_k^H\mathbf{P}_k)^H. \tag{11}$$

Then, after filtering the signal in (9) with (10) and (11), the kth user detects the sth symbol with the following SINR

$$\gamma_k^s = \frac{P|[\mathbf{Q}_k\mathbf{W}_k\mathbf{H}_k^H\mathbf{P}_k\mathbf{x}_k]_s|^2}{P|[\mathbf{Q}_k\mathbf{W}_k\mathbf{Z}_k^H\tilde{\mathbf{P}}\tilde{\mathbf{x}}]_s|^2 + \sigma^2[\mathbf{Q}_k\mathbf{Q}_k^H]_{ss}}, \tag{12}$$

where the rightmost term in the denominator follows from $|[\mathbf{Q}_k\mathbf{W}_k\mathbf{n}_k]_s|^2 = \sigma^2[\mathbf{Q}_k\mathbf{W}_k\mathbf{W}_k^H\mathbf{Q}_k^H]_{ss} = \sigma^2[\mathbf{Q}_k\mathbf{Q}_k^H]_{ss}$, given that $\mathbf{W}_k$ is a semi-unitary matrix, i.e., $\mathbf{W}_k\mathbf{W}_k^H = \mathbf{I}_S$.

B. Harnessing and MITigation (H-MIT) of RIS attacks

In this subsection, we propose a second approach that relaxes the orthogonality constraint $W \leq N - SK$ demanded by the F-MIT strategy. In this method, only unintended reflected streams are selected for mitigation. As an added advantage, users can harness the RIS reflections to extract their own symbols. To attain this objective, we define $\tilde{\mathbf{P}}^\star \triangleq [\mathbf{P}_1 \cdots \mathbf{P}_{k-1}\, \mathbf{P}_{k+1} \cdots \mathbf{P}_K] \in \mathbb{C}^{M\times(K-1)S}$, and $\tilde{\mathbf{x}}^\star \triangleq [\mathbf{x}_1^T \cdots \mathbf{x}_{k-1}^T\, \mathbf{x}_{k+1}^T \cdots \mathbf{x}_K^T]^T \in \mathbb{C}^{(K-1)S}$, which allow us to rewrite the signal in (9) as

$$\mathbf{y}_k = (\mathbf{H}_k^H + \mathbf{Z}_k^H)\mathbf{P}_k\sqrt{P}\mathbf{x}_k + \mathbf{Z}_k^H\tilde{\mathbf{P}}^\star\sqrt{P}\tilde{\mathbf{x}}^\star + \mathbf{n}_k. \tag{13}$$

Thus, we need to achieve $\mathbf{W}_k\mathbf{Z}_k^H\tilde{\mathbf{P}}^\star \approx \mathbf{0}_{W,S}, \forall k \in \mathcal{K}$, which can be accomplished using a similar approach of that employed for the F-MIT strategy. Specifically, let $\underline{\hat{\mathbf{Z}}}_k^\star \triangleq \hat{\mathbf{Z}}_k^H\tilde{\mathbf{P}}^\star \in \mathbb{C}^{N\times(K-1)S}$ be the estimated transformed interference channel, and denote by $\mathbf{U}_{\underline{\hat{\mathbf{Z}}}_k^\star}$ the associated left eigenvector matrix obtained from the SVD of $\underline{\hat{\mathbf{Z}}}_k^\star$. Then, $\mathbf{W}_k$ can be obtained from the last W rows of $\mathbf{U}^H_{\underline{\hat{\mathbf{Z}}}_k^\star}$, as follows

$$\mathbf{W}_k = \left[\mathbf{U}^H_{\underline{\hat{\mathbf{Z}}}_k^\star}\right]_{(N-W+1):N,:} \in \mathbb{C}^{S\times N}, \tag{14}$$

with $W = \max\{N - (K-1)S, S\}$, where now we need $N > (K-1)S$ for achieving the desired non-trivial left null space, and only $W \leq N - (K-1)S$ for the orthogonality with the estimated selected interference channels.

By relying on the effectiveness of $\mathbf{W}_k$ to tackle $\mathbf{Z}_k^H\tilde{\mathbf{P}}^\star\sqrt{P}\tilde{\mathbf{x}}$ in (13), users address the remaining inter-symbol interference with the left inverse of the resulting estimated projected channel $\mathbf{W}_k(\mathbf{H}_k^H + \hat{\mathbf{Z}}_k^H)\mathbf{P}_k \in \mathbb{C}^{W\times S}$, which is given by

$$\mathbf{Q}_k = [(\mathbf{W}_k(\mathbf{H}_k^H + \hat{\mathbf{Z}}_k^H)\mathbf{P}_k)^H\mathbf{W}_k(\mathbf{H}_k^H + \hat{\mathbf{Z}}_k^H)\mathbf{P}_k]^{-1} \times (\mathbf{W}_k(\mathbf{H}_k^H + \hat{\mathbf{Z}}_k^H)\mathbf{P}_k)^H. \tag{15}$$

Note in (15) that the design of $\mathbf{Q}_k$ now depends on $\hat{\mathbf{Z}}_k$. This implies that users might also experience inter-symbol interference. By applying (15) to the signal in (13), the kth user recovers the sth symbol with the following SINR

$$\gamma_k^s = \frac{P|[\boldsymbol{\Xi}_k\mathbf{x}_k]_s|^2}{P|[\boldsymbol{\Upsilon}_k\mathbf{x}_k]_s|^2 + P|[\mathbf{Q}_k\mathbf{W}_k\mathbf{Z}_k^H\tilde{\mathbf{P}}^\star\tilde{\mathbf{x}}^\star]_s|^2 + \sigma^2[\mathbf{Q}_k\mathbf{Q}_k^H]_{ss}}. \tag{16}$$

where $\boldsymbol{\Xi}_k \triangleq \text{diagm}\{\mathbf{Q}_k\mathbf{W}_k(\mathbf{H}_k^H + \mathbf{Z}_k^H)\mathbf{P}_k\}$ is a diagonal matrix with the effective channel coefficients for intended symbols, and $\boldsymbol{\Upsilon}_k \triangleq \mathbf{Q}_k\mathbf{W}_k(\mathbf{H}_k^H + \mathbf{Z}_k^H)\mathbf{P}_k - \boldsymbol{\Xi}_k$ is a hollow matrix with the coefficients for inter-symbol interference.

V. Simulation Results

In this section, we report on the outcomes of Monte Carlo simulations. First, we compare the user sum rates of a mMIMO system exposed to RIS attacks leveraging random phase shifts, a.k.a. Disco Ball attacks [10], [11], and a system under attacks launched by our malicious RIS scheme considering four different threat scenarios. Subsequently, we compare the performance of a safe mMIMO (free from threats) and a system under our optimized attack with and without employing the proposed mitigation strategies, F-MIT and H-MIT.

In each case, the BS is equipped with a uniform linear array of $M = 60$ antennas, and it communicates with $K = 3$ multi-antenna users, such that $S = 2$ parallel data symbols are transmitted for each user. Moreover, we assume that the BS is located at the origin, whereas users $1, 2$, and 3 are located at the coordinates $(20, 0)$ m, $(20, 40)$ m, and $(50, 20)$ m, respectively. As for the attacker, unless otherwise stated, its malicious RIS is strategically deployed at the coordinate $(30, 20)$ m. With this geometrical scenario, the path-loss coefficients for the links BS-RIS, RIS-U, and BS-U are, respectively, computed 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 $d_k^{\text{BS-U}}$ are the distances, and η is the path-loss exponent set to 2.5 for all links. Regarding Algorithm 1, we set the step parameter to $\beta = 0.99$ and the number of iterations to $T = 3 \times 10^3$. Furthermore, we adjust the noise variance to $\sigma^2 = -40$ dBm and employ a uniform power allocation among users and symbols, such that $\alpha_{k,1} = \cdots = \alpha_{k,S} = 1$, $\forall k \in \{1, 2, 3\}$.

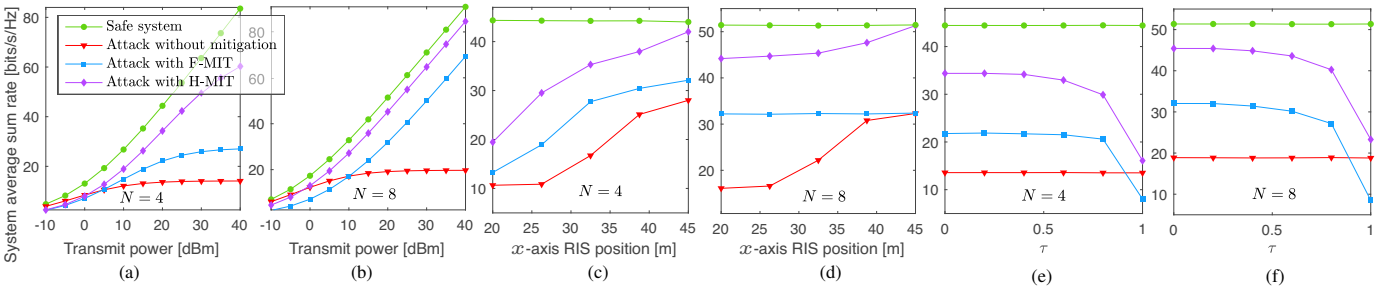

Fig. 3: System average sum rate (with $L = 200$ and $\nu_1 = \nu_2 = \nu_3 = 1/3$): (a) for $N = 4$ and (b) for $N = 8$ when $\tau = 0$; (c) for $N = 4$ and (d) for $N = 8$, considering different x-axis RIS positions and a fixed y-axis coordinate of 20 m, with $P = 20$ dBm and $\tau = 0$; (e) for $N = 4$ and (f) for $N = 8$ with different values of τ and $P = 20$ dBm.

Fig. 2 presents the user average sum rate curves, computed as $R_k = \mathrm{E}\{\sum_{s=1}^{S} \log_2(1 + \gamma_k^s)\}$, which informs the sum spectral efficiency for each user $k \in \{1, 2, 3\}$ under different attack scenarios. These results validate the proposed malicious RIS algorithm and demonstrate its efficiency and adaptability in executing targeted attacks against a specific user. As can be seen, this capability cannot be achieved with the Disco Ball attack. For instance, in Fig. 2(a), user 1 can achieve a sum rate of 8.79 bits/s/Hz when the transmit power is 30 dBm under the proposed RIS attack with equal weights (Attack 1). When the attacker targets this user, i.e., $\nu_1 = 0.98$ in Attack 2, its sum rate drops to only 6.41 bits/s/Hz, a 2.38 bits/s/Hz degradation compared to the equal-weighted case and impressive 5.45 bits/s/Hz compared to the Disco Ball attack. Similar behavior can also be verified for the other users in Figs. 2(b) and 2(c), where the targeted attacks launched by our optimized scheme are remarkably more powerful than the baseline counterpart.

Fig. 3 evaluates the system average sum rate, calculated as $\bar{R} = \mathrm{E}\{\sum_{k=1}^{K} \sum_{s=1}^{S} \log_2(1 + \gamma_k^s)\}$. Specifically, Figs. 3(a), 3(c), and 3(e) show that even when the constraints of (10) and (14) cannot be met, i.e., when $N = 4$, both F-MIT and H-MIT can still provide significant performance gains over the case without mitigation, as long as the transmit power is high enough. When $N = 8$, the perfect orthogonality with the estimated RIS channel can be achieved by both F-MIT and H-MIT, removing the saturation on the sum rate curves, as shown in Fig. 3(b). However, since these receptors also amplify the noise, attaining the same performance as the safe system is impossible. Nonetheless, we can see that recycling intended signals from the RIS-reflected signals is advantageous as it enables the H-MIT receptor to significantly outperform the F-MIT counterpart for $N = 4$ and $N = 8$. These performance gains are also reproduced in the subsequent figures. In Figs. 3(c) and 3(d), the RIS is moved to different positions along the x-axis direction, while keeping its y-axis coordinate fixed at 20 m. As can be seen, the closer the RIS is to the BS, the stronger the performance degradation caused by the proposed attack. Lastly, Figs. 3(e) and 3(f) present the system sum rates for the case when the estimation of the RIS channels is imperfect, revealing that the H-MIT receptor is beneficial even when the estimation error is high.

VI. Conclusions

In this letter, we have shed light on a powerful adversarial attack that can be launched with an optimized RIS against multiple users without requiring their legitimate channels. Simulation results demonstrated the adaptability and efficiency of the proposed RIS algorithm in significantly degrading the sum rates of users, highlighting its superiority over the state-of-the-art Disco Ball attack. This stresses the significance of developing countermeasures to safeguard next-generation wireless systems from emerging physical layer threats. In this regard, we proposed two defense mechanisms, F-MIT and H-MIT, for countering such malicious RIS attacks. Both receptors provided significant performance enhancements over unmitigated attack scenarios, even when their null space constraints could not be satisfied. Still, it is worth noting that users in low transmit power regimes may still be affected by the attack, and additional research and appropriate mitigation strategies are required to safeguard them. Our future work shall delve deeper into estimation strategies for enabling this and related RIS attacks and developing new robust mitigation strategies to counteract threats under realistic CSI conditions.

References

[1] 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.

[2] 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.

[3] 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.

[4] Z. Lin, H. Niu *et al.*, "Pain without gain: Destructive beamforming from a malicious RIS perspective in IoT networks," *IEEE IoT J.*, pp. 1–1, 2023.

[5] H. Alakoca, M. Namdar *et al.*, "Metasurface manipulation attacks: Potential security threats of RIS-aided 6G communications," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 24–30, 2022.

[6] 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.

[7] G. Li, L. Hu *et al.*, "Reconfigurable intelligent surface for physical layer key generation: Constructive or destructive?" *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 146–153, 2022.

[8] J. Yang, X. Ji, F. Wang, K. Huang, and L. Guo, "A novel pilot spoofing scheme via intelligent reflecting surface based on statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12 847–12 857, 2021.

[9] N. Wang, L. Jiao, A. Alipour-Fanid, M. Dabaghchian, and K. Zeng, "Pilot contamination attack detection for NOMA in 5G mm-wave massive MIMO networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1363–1378, 2019.

[10] 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.
[11] H. Huang, Y. Zhang *et al.*, "Disco intelligent reflecting surfaces: Active channel aging for fully-passive jamming attacks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 806–819, 2024.
[12] 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.