On the spectral efficiency of hybrid relay/RIS-assisted massive MIMO systems

On the Spectral Efficiency of Hybrid Relay/RIS-Assisted Massive MIMO Systems

Shih-Kai Chou, Hien Quoc Ngo, and Michail Matthaiou
Centre for Wireless Innovation (CWI), Queen's University Belfast, Belfast, BT3 9DT, U.K
Email: schou01@qub.ac.uk, hien.ngo@qub.ac.uk, m.matthaiou@qub.ac.uk

Abstract—**Reconfigurable intelligent surfaces (RISs) play an important role in extending the connectivity and improving the data rate of future wireless communication systems. However, the conventional (passive) RISs have their limitations and, thus, a hybrid-relay RIS (HR-RIS) architecture is proposed to reap the benefits of relaying systems with high power consumption but higher throughput, and passive RIS systems with cascaded fading effects but low complexity. In this paper, we investigate the performance of HR-RISs in a massive multiple-input multiple-output (M-MIMO) system with zero-forcing (ZF) processing, where channel state information (CSI) is unavailable. We first model the uplink/downlink channels and derive the linear minimum mean square error (LMMSE) estimate of the effective channels. We, then, derive a closed-form expression for the signal to interference and noise ratio (SINR) and spectral efficiency (SE). Finally, we provide some useful engineering insights with our asymptotic analysis and numerical results.**

I. Introduction

M-MIMO is one of the key technologies in fifth-generation (5G) cellular networks, due to its ability to provide high spectral and energy efficiencies with simple linear processing [1], [2]. However, the performance of M-MIMO is still limited by some users (UEs) with weak channels to the base station (BS) due to the long distance or/and heavy blockage. This phenomenon undermines the practical coverage of M-MIMO systems. To this end, new technologies have been proposed to extend the connectivity and improve the network coverage, and one particular technology which has attracted significant attention is RISs. A RIS consists of an array of reconfigurable, passive, and low cost reflective elements. These elements form a metasurface which can be controlled either locally or at network level in order to customize the propagation of the radio waves, and further to improve the network coverage through low-energy, low-complexity sensing, and basic operations [3], [4]. RISs have the ability to increase the spectral and energy efficiency [5], [6] and even provide better position accuracy [7] at the same time. Capitalizing on these benefits, RISs are considered to be one of the most promising technologies for the sixth generation (6G) communication systems [2].

Despite the advantages of RISs, the RIS-aided systems suffer from the cascaded fading effect and they can only offer passive gain. Such gains are generally limited compared to those provided by conventional relays. As a result, a hybrid-RIS-aided M-MIMO architecture was recently proposed in [8]. In HR-RIS-aided systems, a few active elements will be deployed along with passive ones. Compared to the conventional relay systems, they offer lower power consumption. Compared to conventional passive RIS-aided systems, HR-RIS-aided systems offer better performance. Therefore, these architectures have attracted increasing interest recently [8]–[11]. In [8], the authors considered a HR-RIS-aided single-input-single-output (SISO) system with compressed sensing and deep learning techniques. The work of [9] proposed fixed and dynamic HR-RIS architectures and showed that both architectures yield significant improvement in the SE and energy efficiency (EE), compared to conventional passive RIS-aided systems. In [10], the authors proposed an alternating optimization method to optimize the SE in a downlink HR-RIS-aided MIMO system, whereas, [12] investigated the performance of HR-RISs in a covert communication system, in which the transmit power and relay/reflection coefficients were jointly optimized to maximize the covert rate, and it was shown that the HR-RIS can significantly improve the covert rate compared to passive RIS. Similarly to HR-RISs, active RISs [13]–[15] have been investigated recently. The major difference between HR-RIS and active RIS comes from the portion of active elements. More precisely, a HR-RIS deploys only one or several active elements along with passive elements. These active elements are activated by low-powered RF chain and power amplifiers (PAs). On the other hand, active RISs have no passive elements. Importantly, the authors in [15] have confirmed that even with a few active elements, the HR-RISs can improve the SE noticeably.

Most of previous works make the idealistic assumption of perfect CSI, which is unrealistic in practice. Moreover, they either consider only one passive RIS/HR-RIS in the system or in SISO/MIMO scenarios. Furthermore, there is no precoding scheme utilized at the BS. In this paper, we bridge this gap by considering a HR-RIS-aided M-MIMO system, where several HR-RISs are deployed to assist the BS which equipped with a large number of antennas to communicate with single-antenna UEs. Our main contributions are as follows:

- We first model the uplink training and downlink payload data transmission of a HR-RIS-aided M-MIMO system

The work of H. Q. Ngo was supported by the U. K. Research and Innovation Future Leaders Fellowships under Grant MR /S017666/1. The work of M. Matthaiou was supported by a research grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant No. 101001331).

considering time-division duplex (TDD) operation. The linear minimum-mean-square-error (LMMSE) estimates of the effective channels are derived.
- We derive an approximating closed-form expression for the SE with ZF precoder used at the BS. Our closed-form result is very tight and enables us to obtain important insights into our considered system.
- We pursue an asymptotic analysis for a large number of antennas at the BS, number of HR-RISs and the number of elements.

Notation: Boldface upper and lower case letters denote matrices and column vectors, respectively. The subscript $(\cdot)^H$ and $(\cdot)^T$ stand for Hermitian and transpose operation, respectively, whereas, $\mathrm{diag}\{a_1,\ldots,a_N\}$ represents a diagonal matrix with diagonal entries $a_1,\ldots,a_N$, and $\mathrm{tr}\{\cdot\}$ is the trace operation. Furthermore, the notation $\mathbb{E}\{\cdot\}$ denotes the expectation operator, while $\mathbb{C}\{\cdot\}$ denotes the covariance operator. Finally, the symbols $|\cdot|$ and $\|\cdot\|$ indicate the absolute and norm value, respectively.

II. System Model

We consider a BS equipped with M antennas serving K single-antenna UEs, with the help of L HR-RISs, where the l-th RIS has total N elements (including N_{A} active elements and $N-N_{\mathrm{A}}$ passive ones) to reflect the signal from the BS to the UEs. Denote by $\mathbf{h}_{\mathrm{d},k}$, and $\mathbf{h}_{kl}^{\mathrm{UR}}$, $\mathbf{H}_l^{\mathrm{RB}}$, the direct channel from the k-th UE to the BS, the channel from the k-th UE to l-th RIS, and the channel from the l-th RIS to the BS, respectively. Then, the direct channel $\mathbf{h}_{\mathrm{d},k}$, the n-th element of $\mathbf{h}_{kl}^{\mathrm{UR}}$ (denoted by h_{nlk}^{UR}), and the n-th column vector of $\mathbf{H}_l^{\mathrm{RB}}$ (denoted by $\mathbf{h}_{nl}^{\mathrm{RB}}$) are, respectively, modeled as

$$\mathbf{h}_{\mathrm{d},k}=\sqrt{\beta_{\mathrm{d},k}}\tilde{\mathbf{h}}_{\mathrm{d},k}, \tag{1}$$

$$h_{nlk}^{\mathrm{UR}}=\sqrt{\beta_{kl}^{\mathrm{UR}}}\tilde{h}_{nlk}^{\mathrm{UR}}, \tag{2}$$

$$\mathbf{h}_{nl}^{\mathrm{RB}}=\sqrt{\beta_{l}^{\mathrm{RB}}}\tilde{\mathbf{h}}_{nl}^{\mathrm{RB}}, \tag{3}$$

where $\beta_{\mathrm{d},k}$, β_{kl}^{UR} and β_l^{RB} represent the corresponding large-scale fading coefficients. We assume that the elements of $\tilde{\mathbf{h}}_{\mathrm{d},k}$, $\tilde{h}_{nlk}^{\mathrm{UR}}$, $\tilde{\mathbf{h}}_{nl}^{\mathrm{RB}}$ are i.i.d. $\mathcal{CN}(0,1)$. In this paper, we consider TDD operation, i.e., the downlink includes two phases for each coherence interval: *uplink training* and *downlink payload data transmission* phases.

A. Uplink Training Phase

In the uplink training phase, all K UEs send pilot sequences simultaneously to the BS, and the BS estimates the channels to all UEs.

Let τ_{c} be the length of the coherence interval (in samples), and let τ_{p} be the length of uplink training duration, where $\tau_{\mathrm{p}}<\tau_{\mathrm{c}}$. Let $\mathbf{s}_k\in\mathbb{C}^{\tau_{\mathrm{p}}\times 1}$, where $\|\mathbf{s}_k\|^2=1$, be the pilot sequence sent by the k-th UE. We assume that all the pilot sequences sent by all K UEs are mutually orthogonal. In addition, let p_{u} be the power of each UE to transmit the pilot signal. Then, the signal received at the BS can be expressed as the superposition of the signal from the UEs to the BS through HR-RISs, and direct links from the UEs to the BS as

$$\begin{aligned}\mathbf{Y}_{\mathrm{BS}}^{\mathrm{U}}=&\sqrt{\tau_{\mathrm{p}}p_{\mathrm{u}}}\sum_{l=1}^{L}\sum_{k=1}^{K}\mathbf{H}_l^{\mathrm{RB}}\mathbf{\Xi}_{\mathrm{ALL},l}\mathbf{h}_{kl}^{\mathrm{UR}}\mathbf{s}_k^H\\&+\sqrt{\tau_{\mathrm{p}}p_{\mathrm{u}}}\sum_{k=1}^{K}\mathbf{h}_{\mathrm{d},k}\mathbf{s}_k^H\\&+\sum_{l=1}^{L}\mathbf{H}_l^{\mathrm{RB}}\mathbf{\Xi}_{\mathrm{A},l}\left(\mathbf{Z}_{\mathrm{A},l}+\mathbf{Z}_{\mathrm{SI},l}\right)+\mathbf{Z}_{\mathrm{BS}},\end{aligned} \tag{4}$$

where the symbols in (4) are defined as follows:

- $\mathbf{\Xi}_{\mathrm{ALL},l}$ is the processing matrix for the l-th RIS, which can be represented as
$$\mathbf{\Xi}_{\mathrm{ALL},l}=\mathrm{diag}\{\alpha_{1l},\ldots,\alpha_{Nl}\}, \tag{5}$$
where $\alpha_{nl}=|\alpha_{nl}|e^{j\theta_{nl}}$, θ_{nl} represents the phase shift and $|\alpha_{nl}|=1$ if $n\notin\mathcal{A}_l$, where $\mathcal{A}_l$ denotes the index set of the active elements in the l-th HR-RIS.
- $\mathbf{\Xi}_{\mathrm{A},l}=\mathrm{diag}\left\{\phi_{1,l},\ldots,\phi_{N,l}\right\}$, where
$$\phi_{n,l}=\begin{cases}|\alpha_{nl}|e^{j\theta_{nl}}, \text{ if } n\in\mathcal{A}_l\\ 0, \text{ otherwise.}\end{cases} \tag{6}$$
- $\mathbf{Z}_{\mathrm{A},l}$, $\mathbf{Z}_{\mathrm{SI},l}$, and $\mathbf{Z}_{\mathrm{BS}}$ represent the noise, self-interference (SI) associated with the active elements of the l-th RIS, and the noise at the BS, respectively. We assume that $\mathbf{Z}_{\mathrm{A},l}$, $\mathbf{Z}_{\mathrm{SI},l}$, and $\mathbf{Z}_{\mathrm{BS}}$ are mutually independent, and their elements are distributed as $\mathcal{CN}\left(0,\sigma_{\mathrm{A},l}^2\right)$, $\mathcal{CN}\left(0,\sigma_{\mathrm{SI},l}^2\right)$ and $\mathcal{CN}\left(0,\sigma_{\mathrm{BS}}^2\right)$, respectively.

By defining the effective channel from the k-th UE to the BS as

$$\begin{aligned}\mathbf{h}_k^{\mathrm{U}}&\triangleq\mathbf{h}_{\mathrm{d},k}+\sum_{l=1}^{L}\mathbf{H}_l^{\mathrm{RB}}\mathbf{\Xi}_{\mathrm{ALL},l}\mathbf{h}_{kl}^{\mathrm{UR}}\\&=\mathbf{h}_{\mathrm{d},k}+\sum_{l=1}^{L}\sum_{n=1}^{N}\alpha_{nl}h_{nlk}^{\mathrm{UR}}\mathbf{h}_{nl}^{\mathrm{RB}},\end{aligned} \tag{7}$$

equation (4) can be rewritten as

$$\begin{aligned}\mathbf{Y}_{\mathrm{BS}}^{\mathrm{U}}=&\sqrt{\tau_{\mathrm{p}}p_{\mathrm{u}}}\sum_{k=1}^{K}\mathbf{h}_k^{\mathrm{U}}\mathbf{s}_k^H+\sum_{l=1}^{L}\mathbf{H}_l^{\mathrm{RB}}\mathbf{\Xi}_{\mathrm{A},l}\left(\mathbf{Z}_{\mathrm{A},l}+\mathbf{Z}_{\mathrm{SI},l}\right)\\&+\mathbf{Z}_{\mathrm{BS}}.\end{aligned} \tag{8}$$

To estimate $\mathbf{h}_k^{\mathrm{U}}$, $\mathbf{Y}_{\mathrm{BS}}^{\mathrm{U}}$ is first projected onto $\mathbf{s}_k$ to obtain

$$\tilde{\mathbf{y}}_k=\mathbf{Y}_{\mathrm{BS}}^{\mathrm{U}}\mathbf{s}_k=\sqrt{\tau_{\mathrm{p}}p_{\mathrm{u}}}\mathbf{h}_k^{\mathrm{U}}+\tilde{\mathbf{z}}_k, \tag{9}$$

where $\tilde{\mathbf{z}}_k=\left(\sum_{l=1}^{L}\mathbf{H}_l^{\mathrm{RB}}\mathbf{\Xi}_{\mathrm{A},l}\left(\mathbf{Z}_{\mathrm{A},l}+\mathbf{Z}_{\mathrm{SI},l}\right)+\mathbf{Z}_{\mathrm{BS}}\right)\mathbf{s}_k$. Then, given $\tilde{\mathbf{y}}_k$, the LMMSE estimate of the effective channel $\mathbf{h}_k^{\mathrm{U}}$ is $\hat{\mathbf{h}}_k^{\mathrm{U}}$ expressed as (11), shown at the top of the next page. By denoting

$$\mathbf{C}_k=\beta_{\mathrm{d},k}\mathbf{I}_{\mathrm{M}}+\sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\mathbf{I}_{\mathrm{M}}, \tag{12}$$

$$\hat{\mathbf{h}}_k^{\mathrm{U}} = \mathbb{E}\left\{\mathbf{h}_k^{\mathrm{U}}\right\} + \mathbb{C}\left\{\mathbf{h}_k^{\mathrm{U}}, \tilde{\mathbf{y}}_k\right\} \mathbb{C}\left\{\tilde{\mathbf{y}}_k\right\}^{-1} \left(\tilde{\mathbf{y}} - \mathbb{E}\left\{\tilde{\mathbf{y}}_k^{\mathrm{U}}\right\}\right) \tag{10}$$

$$= \sqrt{\tau_{\mathrm{p}} p_{\mathrm{u}}} \left(\beta_{\mathrm{d},k}\mathbf{I}_{\mathrm{M}} + \sum_{l=1}^{L}\sum_{n=1}^{N} |\alpha_{nl}|^2 \beta_l^{\mathrm{RB}} \beta_{kl}^{\mathrm{UR}} \mathbf{I}_{\mathrm{M}}\right)$$

$$\times \left[\tau_{\mathrm{p}} p_{\mathrm{u}} \left(\beta_{\mathrm{d},k}\mathbf{I}_{\mathrm{M}} + \sum_{l=1}^{L}\sum_{n=1}^{N} |\alpha_{nl}|^2 \beta_l^{\mathrm{RB}} \beta_{kl}^{\mathrm{UR}} \mathbf{I}_{\mathrm{M}}\right) + \left(\sigma_{\mathrm{BS}}^2 \mathbf{I}_{\mathrm{M}} + \sum_{l=1}^{L}\sum_{n\in\mathcal{A}_l} |\alpha_{nl}|^2 \beta_l^{\mathrm{RB}} \left(\sigma_{\mathrm{SI},l}^2 + \sigma_{\mathrm{A},l}^2\right) \mathbf{I}_{\mathrm{M}}\right)\right]^{-1} \tilde{\mathbf{y}}_k. \tag{11}$$

and

$$\mathbf{C}_{\mathrm{z},k} = \sigma_{\mathrm{BS}}^2 \mathbf{I}_{\mathrm{M}} + \sum_{l=1}^{L}\sum_{n\in\mathcal{A}_l} |\alpha_l|^2 \beta_l^{\mathrm{RB}} \left(\sigma_{\mathrm{SI},l}^2 + \sigma_{\mathrm{A},l}^2\right) \mathbf{I}_{\mathrm{M}}, \tag{13}$$

we can rewrite the effective channel estimate (11) as

$$\hat{\mathbf{h}}_k^{\mathrm{U}} = \sqrt{\tau_{\mathrm{p}} p_{\mathrm{u}}} \mathbf{C}_k \left(\tau_{\mathrm{p}} p_{\mathrm{u}} \mathbf{C}_k + \mathbf{C}_{\mathrm{z},k}\right)^{-1} \tilde{\mathbf{y}}_k. \tag{14}$$

We define $\mathbf{H}^{\mathrm{U}} = \left[\mathbf{h}_1^{\mathrm{U}}, \ldots, \mathbf{h}_K^{\mathrm{U}}\right]$, which collects all the effective channel vectors from all K UEs and $\hat{\mathbf{H}}^{\mathrm{U}}$ corresponds to its LMMSE estimate.

Finally, the channel estimation error vector associated with the k-th UE is denoted by $\tilde{\mathbf{h}}_k^{\mathrm{U}} = \mathbf{h}_k^{\mathrm{U}} - \hat{\mathbf{h}}_k^{\mathrm{U}}$, and the corresponding channel estimation matrix is $\tilde{\mathbf{H}}^{\mathrm{U}} = \left[\tilde{\mathbf{h}}_1^{\mathrm{U}}, \ldots, \tilde{\mathbf{h}}_K^{\mathrm{U}}\right]$. From the LMMSE estimation property, $\tilde{\mathbf{h}}_k^{\mathrm{U}}$ is uncorrelated with $\hat{\mathbf{h}}_k^{\mathrm{U}}$. Since $\mathbb{E}\left\{\tilde{\mathbf{h}}_k^{\mathrm{U}}\right\} = \mathbf{0}$, we can have the covariance matrix $\mathbb{C}\left\{\tilde{\mathbf{h}}_k^{\mathrm{U}}\right\} = \mathbf{C}_k - \hat{\mathbf{C}}_k$, where $\hat{\mathbf{C}}_k = \mathbb{C}\left\{\hat{\mathbf{h}}_k^{\mathrm{U}}\right\}$.

B. Downlink Payload Data Transmission

In this phase, the BS intends to send signals to all K UEs. Let $\mathbf{W}$ be the precoding matrix, and $\mathbf{q} = [q_1, \ldots, q_K]^T$ be the symbol vector, where q_k, with $\mathbb{E}\left\{|q_k|\right\}^2 = 1$, is the symbol intended for the k-th UE. The downlink signals propagate on the direct channels and are also reflected through the HR-RISs. Thus, the signal received at the k-th UE can be expressed as

$$r_k = \sqrt{P_{\mathrm{d}}} \mathbf{h}_k^{\mathrm{U}^T} \mathbf{W}\mathbf{q} + \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{UR}^T} \mathbf{\Xi}_{\mathrm{A},l} \left(\mathbf{z}_{\mathrm{A},l}^D + \mathbf{z}_{\mathrm{SI},l}^D\right) + n_{\mathrm{UE},k}, \tag{15}$$

where P_{d} is the transmit power, $\mathbf{z}_{\mathrm{A},l}^D$, $\mathbf{z}_{\mathrm{SI},l}^D$ and $n_{\mathrm{UE},k}$ represent the noise caused by the active elements, residual SI and the noise at the k-th UE, respectively, whose elements have the distribution of $\mathcal{CN}\left(0, \sigma_{\mathrm{A},l}^{D^2}\right)$, $\mathcal{CN}\left(0, \sigma_{\mathrm{SI},l}^{D^2}\right)$ and $\mathcal{CN}\left(0, \sigma_{\mathrm{UE},k}^2\right)$, respectively.

In this paper, the ZF precoder is considered. With ZF, the precoding matrix is given

$$\mathbf{W} = \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}, \tag{16}$$

where $\hat{\mathbf{H}} = \left[\hat{\mathbf{h}}_1^{\mathrm{U}}, \ldots, \hat{\mathbf{h}}_K^{\mathrm{U}}\right]$, and $\mathbf{P} = \mathrm{diag}\left\{\sqrt{\eta_1}, \ldots, \sqrt{\eta_K}\right\}$ is the power control matrix chosen to satisfy the power constraint $\mathbb{E}\left\{||\mathbf{W}\mathbf{q}||^2\right\} \leq 1$ at the BS.

III. Performance Analysis

In this section, we first derive closed-form expression for the SE. Then, based on the closed-form results, we provide some insights into the system performance when the number of base station antennas and the number of RIS elements are very large.

A. Spectral Efficiency

The received signal at the k-th UE (15) can be rewritten as

$$r_k = \sqrt{P_{\mathrm{d}}}(\hat{\mathbf{h}}_k^{\mathrm{U}} + \tilde{\mathbf{h}}_k^{\mathrm{U}})^T \mathbf{W}\mathbf{q} + \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{UR}^T} \mathbf{\Xi}_{\mathrm{A},l}(\mathbf{z}_{\mathrm{A},l}^D + \mathbf{z}_{\mathrm{SI},l}^D) + n_{\mathrm{UE},k}. \tag{17}$$

By substituting (16) into (17), we obtain

$$\begin{aligned} r_k &= \sqrt{P_{\mathrm{d}}} \hat{\mathbf{h}}_k^{\mathrm{U}^T} \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q} + \tilde{\mathbf{h}}_k^{\mathrm{U}^T} \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q} \\ &\quad + \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{UR}^T} \Xi_{\mathrm{A},l}(\mathbf{z}_{\mathrm{A},l}^D + \mathbf{z}_{\mathrm{SI},l}^D) + n_{\mathrm{UE},k} \\ &= \sqrt{P_{\mathrm{d}}\eta_k} q_k + \tilde{\mathbf{h}}_k^{\mathrm{U}^T} \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q} \\ &\quad + \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{UR}^T} \Xi_{\mathrm{A},l}(\mathbf{z}_{\mathrm{A},l}^D + \mathbf{z}_{\mathrm{SI},l}^D) + n_{\mathrm{UE},k} \\ &= \sqrt{P_{\mathrm{d}}\eta_k} q_k + \tilde{\mathbf{h}}_k^{\mathrm{U}^T} \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q} + n_{\mathrm{ALL},k}, \end{aligned} \tag{18}$$

where $n_{\mathrm{ALL},k}$ is the effective noise at the k-th UE defined as $n_{\mathrm{ALL},k} = \sum_{l=1}^{L} \mathbf{h}_{l,k}^{\mathrm{UR}^T} \mathbf{\Xi}_{\mathrm{A},l}(\mathbf{z}_{\mathrm{A},l}^D + \mathbf{z}_{\mathrm{SI},l}^D) + n_{\mathrm{UE},k}$. In (18), the first term is the desired signal, the second and third terms represent the channel estimation error and noise. Since the last two terms are uncorrelated with the desired signal, we can obtain the following SE of the k-th UE:

$$\mathrm{SE}_k = \left(1 - \frac{\tau_{\mathrm{p}}}{\tau_c}\right) \log_2\left(1 + \mathrm{SINR}_k\right), \tag{19}$$

where SINR_k is given as

$$\mathrm{SINR}_k = \frac{P_{\mathrm{d}}\eta_k}{P_{\mathrm{d}}\mathbb{E}\left\{|\tilde{\mathbf{h}}_k^{\mathrm{U}^T} \hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q}|^2\right\} + \sigma_{\mathrm{ALL},k}^2}, \tag{20}$$

where

$$\sigma_{\mathrm{ALL},k}^2 = \sigma_{\mathrm{UE},k}^2 + \sum_{l=1}^{L}\sum_{n\in\mathcal{A}_l} |\alpha_{nl}|^2 \beta_l^{\mathrm{UR}} \left(\sigma_{\mathrm{SI},l}^{D^2} + \sigma_{\mathrm{A},l}^{D^2}\right).$$

The only undetermined term in (20) is

$$\mathbb{E}\left\{|\tilde{\mathbf{h}}_k^{\mathrm{U}^T}\hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\mathbf{P}\mathbf{q}|^2\right\}$$
$$= \mathrm{tr}\left\{\mathbf{P}^2\mathbb{E}\left\{(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\hat{\mathbf{H}}^T\tilde{\mathbf{h}}_k^{\mathrm{U}^*}\tilde{\mathbf{h}}_k^{\mathrm{U}^T}\hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\right\}\right\}. \tag{21}$$

Since the effective channels are cascaded by the direct and indirect channels between the BS and UEs, their channel estimates and channel estimation errors have complicated forms. Thus, it is very difficult (if not impossible) to obtain an exact closed-form result of (21). To render this problem more tractable, we propose to use a tight approximation. More precisely, by using the fact that the number of RISs elements is large, and by the central limit theorem, the effective channel $\mathbf{h}_k^{\mathrm{U}}$ given in (7) can be considered approximately Gaussian distributed, i.e. $\mathbf{h}_k^{\mathrm{U}} \sim \mathcal{CN}(0, \mathbf{C}_k)$. Thus, the channel estimate can be also approximately Gaussian distributed, i.e., $\hat{\mathbf{h}}_k^{\mathrm{U}} \sim \mathcal{CN}\left(0, \hat{\mathbf{C}}_k\right)$. In addition, since $\tilde{\mathbf{h}}_k^{\mathrm{U}} = \mathbf{h}_k^{\mathrm{U}} - \hat{\mathbf{h}}_k^{\mathrm{U}}$, we can then express the covariance matrix of $\tilde{\mathbf{h}}_k^{\mathrm{U}}$ as

$$\begin{aligned}\tilde{\mathbf{C}}_k &= \mathbf{C}_k - \hat{\mathbf{C}}_k \\ &= \left(\beta_{\mathrm{d},k} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\right)\mathbf{I}_{\mathrm{M}} \\ &\quad - \tau_{\mathrm{p}}p_{\mathrm{u}}\mathbf{C}_k\mathbf{E}_k^{-H}\mathbf{C}_k^H,\end{aligned} \tag{22}$$

where $\mathbf{E}_k = \mathbb{C}\{\tilde{\mathbf{y}}_k\}$. By using the above approximation, we can obtain the approximating closed-form expression for the spectral efficiency as in the following theorem.

Theorem 1: The spectral efficiency of the k-th UE can be approximated as

$$\begin{aligned}\mathrm{SE}_k \approx &\left(1 - \frac{\tau_{\mathrm{p}}}{\tau_c}\right) \times \\ &\log_2\left(1 + \frac{P_{\mathrm{d}}\eta_k}{P_{\mathrm{d}}\mathcal{B}_k\sum_{k'=1}^{K}\frac{\eta_{k'}}{(M-K)\hat{\mathcal{B}}_{k'}} + \sigma_{\mathrm{ALL},k}^2}\right),\end{aligned} \tag{23}$$

where $\mathcal{B}_k$ is denoted as

$$\begin{aligned}\mathcal{B}_k = &\beta_{\mathrm{d},k} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}} \\ &- \tau_{\mathrm{p}}p_{\mathrm{u}}\left(\beta_{\mathrm{d},k} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\right)^2 \\ &\times\left(\tau_{\mathrm{p}}p_{\mathrm{u}}\left(\beta_{\mathrm{d},k} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\right) + \right. \\ &\left. + \sigma_{\mathrm{BS}}^2 + \sum_{l=1}^{L}\sum_{n\in\mathcal{A}_l}|\alpha_{nl}|^2\beta_l^{\mathrm{UR}}\left(\sigma_{\mathrm{SI},l}^2 + \sigma_{\mathrm{A},l}^2\right)\right)^{-1},\end{aligned} \tag{24}$$

while $\hat{\mathcal{B}}_{k'}$ is defined as

$$\begin{aligned}\hat{\mathcal{B}}_{k'} = &\tau_{\mathrm{p}}p_{\mathrm{u}}\left(\beta_{\mathrm{d},k'} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\right)^2 \\ &\times\left(\tau_{\mathrm{p}}p_{\mathrm{u}}\left(\beta_{\mathrm{d},k'} + \sum_{l=1}^{L}\sum_{n=1}^{N}|\alpha_{nl}|^2\beta_l^{\mathrm{RB}}\beta_{kl}^{\mathrm{UR}}\right) + \right. \\ &\left. + \sigma_{\mathrm{BS}}^2 + \sum_{l=1}^{L}\sum_{n\in\mathcal{A}_l}|\alpha_{nl}|^2\beta_l^{\mathrm{UR}}\left(\sigma_{\mathrm{SI},l}^2 + \sigma_{\mathrm{A},l}^2\right)\right)^{-1}.\end{aligned} \tag{25}$$

Proof: See Appendix. ■

B. Asymptotic Analysis

In this section, we consider two scenarios: 1) $M \to \infty$ and 2) $NL \to \infty$. In addition, we consider full power control $\eta_k = \frac{(M-K)\hat{\mathcal{B}}_k}{K}$ which satisfies the power constraint at the BS: $\mathbb{E}\left\{||\mathbf{W}\mathbf{q}||^2\right\} \le 1$. For simplicity, we consider a simplified case where the large-scale fading coefficients of each transmission link are set to the same, i.e. $\beta_{\mathrm{d},k} = \beta_{\mathrm{d}}$, $\forall k$, $\beta_l^{\mathrm{RB}} = \beta^{\mathrm{RB}}$, $\forall l$, and $\beta_{kl}^{\mathrm{UR}} = \beta^{\mathrm{UR}}$, $\forall l$ and k, whereas $|\alpha_{nl}|^2$ is set as $|\alpha_{nl}|^2 = 1, \forall n \notin \mathcal{A}$, $\forall l$ and $|\alpha_{nl}|^2 = |\alpha|^2, \forall n \in \mathcal{A}$ and l. With this assumption, the SINR in (23) can be simplified as

$$\mathrm{SINR}_k = \frac{\frac{P_{\mathrm{d}}(M-K)\hat{\mathcal{B}}}{K}}{P_{\mathrm{d}}\mathcal{B} + \sigma_{\mathrm{ALL}}^2}, \tag{26}$$

where $\hat{\mathcal{B}}$, $\mathcal{B}$, and σ_{ALL}^2 are denoted as

$$\hat{\mathcal{B}} = \tau_{\mathrm{p}}p_{\mathrm{u}}\mu^2, \tag{27}$$

$$\mathcal{B} = \mu\left(\sigma_{\mathrm{BS}}^2 + LN_{\mathrm{A}}|\alpha|^2\beta^{\mathrm{UR}}\left(\sigma_{\mathrm{SI}}^2 + \sigma_{\mathrm{A}}^2\right)\right), \tag{28}$$

and

$$\begin{aligned}\sigma_{\mathrm{ALL}}^2 = &LN_{\mathrm{A}}|\alpha|^2\beta^{\mathrm{UR}}\left(\sigma_{\mathrm{SI}}^2 + \sigma_{\mathrm{A}}^2\right) \times \\ &\left(\tau_{\mathrm{p}}p_{\mathrm{u}}\mu\sigma_{\mathrm{BS}}^2 + LN_{\mathrm{A}}|\alpha|^2\beta^{\mathrm{UR}}\left(\sigma_{\mathrm{SI}}^2 + \sigma_{\mathrm{A}}^2\right)\right),\end{aligned} \tag{29}$$

where $\mu = \beta_d + \left(L\left(N - N_{\mathrm{A}} + N_{\mathrm{A}}|\alpha|^2\right)\right)\beta^{\mathrm{RB}}\beta^{\mathrm{UR}}$.

1) When $M \to \infty$: From (26), we obtain

$$\lim_{M\to\infty}\mathrm{SINR}_k = \infty. \tag{30}$$

We can see that when M goes to infinity, the effect of noise and self-interference of the active elements disappears, and hence, the SINR can increase without bound.

We next assume the transmit power at the BS as $P_{\mathrm{d}} = E_{\mathrm{d}}/M$, where E_{d} is fixed. From (26), we can have

$$\lim_{M\to\infty}\mathrm{SINR}_k = \frac{E_{\mathrm{d}}\hat{\mathcal{B}}}{K\sigma_{\mathrm{ALL}}^2}. \tag{31}$$

This implies that by using a very large M, due to the array gain, we can scale down the transmit power by $1/M$ without any performance degradation. In addition, the asymptotic SINR in this case depends mainly on the variances of the effective channel gain and the noise at the HR-RIS side.

2) When $N \to \infty$, and L is fixed: Here, we consider the case of $N \to \infty$, while the number of surfaces is fixed. In this case, the asymptotic SINR is

$$\lim_{N\to\infty} \mathrm{SINR}_k = \lim_{N\to\infty} \frac{\frac{N P_\mathrm{d} L (M-K) \tau_\mathrm{p} p_\mathrm{u} \beta^{\mathrm{RB}}}{K}}{P_\mathrm{d} \tau_\mathrm{p} p_\mathrm{u} \beta^{\mathrm{RB}} \beta^{\mathrm{UR}} + \sigma_{\mathrm{SI}}^2 + \sigma_{\mathrm{A}}^2} = \infty. \tag{32}$$

As we can see from (32), when N grows, the SINR can also increase without bound.

3) When $L \to \infty$, and N is fixed: As for the case of N being fixed and L going to infinity, we can deduce the asymptotic SINR as

$$\lim_{L\to\infty} \mathrm{SINR}_k = \frac{P_\mathrm{d} (M-K) \tau_\mathrm{p} p_\mathrm{u} \beta^{\mathrm{RB}} \left(N + N_\mathrm{A} \left(|\alpha|^2 - 1\right)\right)}{K N_\mathrm{A} \left(\sigma_{\mathrm{SI}}^2 + \sigma_{\mathrm{A}}^2\right) \left(P_\mathrm{d} |\alpha|^2 + \sigma_{\mathrm{BS}}^2\right)}. \tag{33}$$

As we can see from (33), as $L \to \infty$, the SINR grows proportionally to $(M-K)/K$. Although the HR-RISs bring the interference and cascaded noise into the system, they can offer an additional gain from the active elements. Moreover, (33) depends on only on β^{RB}. This implies that we should carefully design the distances between the HR-RISs and the BS to optimize the system performance.

IV. Numerical Results

In this section, we provide numerical results to verify our analysis. We consider the following setup: all UEs and RISs are randomly distributed over the coverage area of 1×1 km^2, and the BS is located at the center of the area. Let $d \in \{d_k^{\mathrm{BU}}, d_{kl}^{\mathrm{UR}}, d_l^{\mathrm{RB}}\}$ be the distance between the BS and the k-th UE, between the k-th UE and the l-th HR-RIS, and between the l-th HR-RIS and the BS, respectively. The large-scale fading coefficients are modeled as three-slope model [16], which can be expressed as

$$\beta = \begin{cases} \beta_0 - 35\log_{10}(d), & \text{if } d > d_1 \\ \beta_0 - 15\log_{10}(d_1) - 20\log_{10}(d), & \text{if } d_0 < d < d_1, \\ \beta_0 - 15\log_{10}(d_1) - 20\log_{10}(d_0), & \text{if } d \le d_0 \end{cases} \tag{34}$$

where $\beta_0 = -132.7$ dB, $d_0 = 10$ m, and $d_1 = 50$ m. We choose $\sigma_{\mathrm{BS}}^2 = \sigma_{\mathrm{A},nl}^2 = \sigma_{\mathrm{SI},nl}^2 = -170 + 10\log_{10} B_0 +$ NF dBm, where $B_0 = 20$ MHz and NF $= 9$ dB. In addition, $|\alpha_{nl}|$ is set to satisfy $\sum_{l=1}^{L}\sum_{n=1}^{N} |\alpha_{nl}|^2 \le P_{\mathrm{R,max}} = 5$ dBm with exhaustive search, while the phases are randomly generated [9], where it can be seen as the worst case.

Firstly, we validate the tightness of our approximating closed-form expression in Theorem 1, as well as the asymptotic results. In Fig. 1, we consider two cases: 1) $P_\mathrm{d} = 10$ dB, $p_\mathrm{u} = 0$ dB, where they are normalized by the noise power, and 2) $P_\mathrm{d} = E_\mathrm{d}/M$, where $E_\mathrm{d} = 20$ dB. Furthermore, we choose $\tau_\mathrm{p} = K$, $\tau_\mathrm{c} = 200$, $L = 40$, $N = 20$, $N_\mathrm{A} = 4$, $K = 5$. We also set the large-scale coefficients as $\beta_{\mathrm{d},k} = \beta_\mathrm{d}$, $\forall k$, $\beta_l^{\mathrm{RB}} = \beta^{\mathrm{RB}}$, $\forall l$, and $\beta_{kl}^{\mathrm{UR}} = \beta^{\mathrm{UR}}$,$\forall l$ and k. In Fig. 2, we use the same setup as in Fig. 1, except that M is fixed, and L changes. As shown in both Fig. 1 and Fig. 2, our theoretical analysis results align well with Monte-Carlo simulations as well as the asymptotic results. Next, in Fig. 3, we use the same setup as in Fig. 1

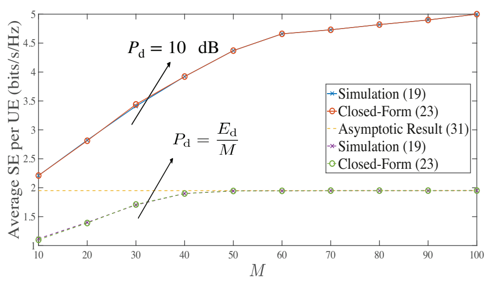

Fig. 1: SE versus M.

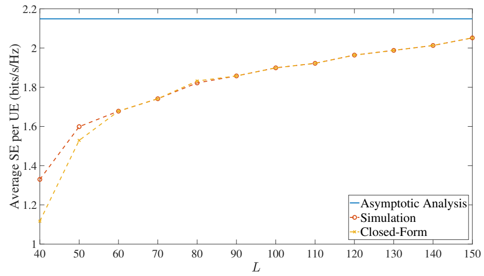

Fig. 2: SE versus L.

with the case of $K = 10$, $M = 50$, $P_\mathrm{d} = 10$ dB to compare the performance of HR-RIS and passive RIS-aided M-MIMO system. The figure shows the cumulative distribution of the SE per user. Compared to a passive RIS-aided system, a HR-RIS-aided system offers higher gain and the improvement is significant. More specifically, under this setup, the 95%-likely per-user downlink spectral efficiency of the conventional passive-aided system is about 0.15 bits/s/Hz, whereas, the HR-RIS-aided system yields 0.48 bits/s/Hz, that showcases a 3.2 times improvement than the conventional fully passive system. Finally, in Fig. 4, we demonstrate the impact of the number

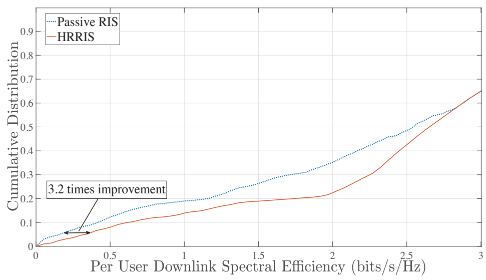

Fig. 3: The cumulative distribution of the SE per user for passive RIS and HR-RIS-aided systems.

of active elements and self-interference level of HR-RIS on the SE performance. In particular, we set $M = 40$, $L = 10$, $N = 200$, $N_\mathrm{A} \in [0, 200]$, $K = 10$, $P_\mathrm{R} = 10$ dB, and $\sigma_{\mathrm{SI}}^2 = \sigma_{\mathrm{A}}^2 = \{5, 30\}$ dB. The results showcase that a HR-RIS architecture, even with a small number of active elements, can produce a significant improvement in the SE. However, as the number of active elements increases, self-interference/noise at the HR-RISs has more effect, and hence, the SE decreases. A careful selection of the number of active elements is essential for improving the system performance.

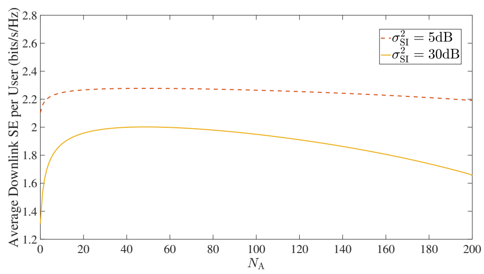

Fig. 4: Average per-user downlink SE versus N_{A}.

V. Conclusion

This work considered HR-RIS aided M-MIMO systems, in which we modeled the uplink training and downlink payload data transmission under TDD operation. We derived the LMMSE estimate of the effective channels, and the approximating closed-formed expression for the downlink SE under ZF processing. This approximation was based on the central limit theorem and shown to be very tight. In addition, we analyzed the asymptotic results when the numbers of BS antennas and RIS elements go to infinity, which helped us to obtain important insights. The power scaling law was also provided, i.e., we showed that when M goes to infinity, we can reduce the transmit power proportionally to $1/M$, while maintaining a give quality of service. The analytical derivations have been numerically verified by simulations, showing that by deploying several active elements in the RIS, the SE can improve significantly, compared to conventional RIS systems where all elements are passive.

Appendix

With the Gaussian approximation, (21) can be expressed as

$$\begin{aligned}(21) &\approx \mathrm{tr}\left\{\mathbf{P}^2\mathbb{E}\left\{(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\hat{\mathbf{H}}^T\tilde{\mathbf{C}}_k\hat{\mathbf{H}}^*(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\right\}\right\} \\ &= \mathcal{B}_k\mathrm{tr}\left\{\mathbf{P}^2\mathbb{E}\left\{(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\right\}\right\} \\ &= \mathcal{B}_k\sum_{k'=1}^{K}\eta_{k'}\mathbb{E}\left\{\left[(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\right]_{k'k'}\right\}. \end{aligned} \tag{35}$$

Since $\hat{\mathbf{H}}$ is constructed by complex Gaussian vectors, we can express it as

$$\hat{\mathbf{H}} = \hat{\mathbf{H}}_{\mathcal{W}}\mathbf{D}^{1/2}, \tag{36}$$

where $\hat{\mathbf{H}}_{\mathcal{W}}$ is an $M \times K$ matrix whose entries are i.i.d. $\mathcal{CN}(0,1)$, and $\mathbf{D}$ is a diagonal matrix whose k-th diagonal element is $\hat{\mathcal{B}}_k$. Then, we can express (35) as

$$\begin{aligned}&\mathcal{B}_k\sum_{k'=1}^{K}\eta_{k'}\mathbb{E}\left\{\left[(\hat{\mathbf{H}}^T\hat{\mathbf{H}}^*)^{-1}\right]_{k'k'}\right\} \\ &= \mathcal{B}_k\sum_{k'=1}^{K}\frac{\eta_{k'}}{K\hat{\mathcal{B}}_{k'}}\mathbb{E}\left\{\mathrm{tr}\left\{\left[(\hat{\mathbf{H}}_{\mathcal{W}}^T\hat{\mathbf{H}}_{\mathcal{W}}^*)^{-1}\right]\right\}\right\} \\ &= \mathcal{B}_k\sum_{k'=1}^{K}\frac{\eta_{k'}}{(M-K)\hat{\mathcal{B}}_{k'}}, \quad \text{for } M \geq K+1, \end{aligned} \tag{37}$$

where the last equality follows [17, Lemma 2.10]. Thus, we can arrive at the desired result in Theorem 1.

References

[1] J. Zhang, E. Björnson, D. W. K. Ng, M. Matthaiou, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[2] M. Matthaiou, *et al.*, "The road to 6G: Ten physical layer challenges for communications engineers," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 64–69, Jan. 2021.

[3] T. Nakanishi, T. Otani, Y. Tamayama, and M. Kitano, "Storage of electromagnetic waves in a metamaterial that mimics electromagnetically induced transparency," *Physical Review B*, vol. 87, no. 161110, Apr. 2013.

[4] M. Noor-A-Rahim *et al.*, "Towards industry 5.0: Intelligent reflecting surface (IRS) in smart manufacturing," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 72–78, Oct. 2022.

[5] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[6] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Jun. 2019.

[7] J. He, H. Wymeersch, T. Sanguanpuak, O. Silven, and M. Juntti, "Adaptive beamforming design for mmwave RIS-aided joint localization and communication," in *Proc. IEEE WCNC*, Apr. 2020.

[8] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE Access*, vol. 9, pp. 44 304–44 321, Mar. 2021.

[9] N. T. Nguyen, Q.-D. Vu, K. Lee, and M. Juntti, "Hybrid relay-reflecting intelligent surface-assisted wireless communication," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6228–6244, Jun. 2022.

[10] N. T. Nguyen, Q. D. Vu, K. Lee, and M. Juntti, "Spectral efficiency optimization for hybrid relay-reflecting intelligent surface," in *Proc. IEEE ICC*, Jun. 2021, pp. 1–6.

[11] B. Al-Nahhas, M. Obeed, A. Chaaban, and M. J. Hossain, "RIS-aided cell-free massive MIMO: Performance analysis and competitiveness," in *Proc. IEEE ICC*, Jun. 2021, pp. 1–6.

[12] J. Hu, X. Shi, S. Yan, Y. Chen, T. Zhao, and F. Shu, "Hybrid relay-reflecting intelligent surface-aided covert communications," 2022. [Online]. Available: https://arxiv.org/abs/2203.12223

[13] M. H. Khoshafa, T. M. N. Ngatched, M. H. Ahmed, and A. R. Ndjiongue, "Active reconfigurable intelligent surfaces-aided wireless communication system," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3699–3703, Nov. 2021.

[14] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active RIS vs. passive RIS: Which will prevail in 6G?" 2021. [Online]. Available: https://arxiv.org/abs/2103.15154

[15] R. Long, Y.-C. Liang, Y. Pei, and E. G. Larsson, "Active reconfigurable intelligent surface-aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4962–4975, Aug. 2021.

[16] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[17] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004. [Online]. Available: http://dx.doi.org/10.1561/0100000001