MIMO Full-duplex Networks with Limited Knowledge of the Relay State

MIMO Full-duplex Networks with Limited Knowledge of the Relay State

Alessandro Nordio, *Member, IEEE,* Carla Fabiana Chiasserini, *Fellow, IEEE*

Abstract

Full-duplex (FD)-enabled relay networks represent a relevant solution to two critical needs of next-generation networks, namely, radio coverage extension and high spectral efficiency of wireless communications. Under practical conditions, however, the FD mode may not be the best operational setting for the relay; rather, operating in half-duplex may be more convenient when harsh channel conditions add up to self-interference. One of the fundamental challenges in the design of FD relay networks is thus how to determine the relay operational mode and the value of transmit power at both the relay and the data source, so that the achievable data rate is maximized as time varies. We address this problem in a two-hop, MIMO network, accounting for practical operational conditions in which the source is unaware of the symbols that the relay is transmitting. In light of the problem complexity, we also derive a lower-bound to the maximum achievable rate, which proves to be tight, especially for low-medium SNR values. We then tackle massive MIMO networks, and exploit our asymptotic analysis in the number of antennas to derive a low-complexity, yet highly efficient, operational mode and transmit power allocation scheme for a finite-size scenario.

I. Introduction

The rapid increase in the mobile data demand and the need for an extended coverage of wireless connectivity have made multiple-input-multiple-output (MIMO) relay networks a highly attractive technology [1], [2]. In such systems, all involved nodes, namely, source, relay, and destination, may be equipped with multiple antennas for data transmission and reception. Then, spectral efficiency can be further increased by leveraging the full-duplex (FD) operation, so as to enable simultaneous transmission and reception over the same frequency band at the relay [3]. Given the promise of such high performance, MIMO-FD relay networks have been identified as one of the enabling technologies for 5G and beyond, efficiently catering data transfers in case of coverage holes or users at the cell edge [4]–[6].

Importantly, the increasing practical relevance of FD relay networks is also due to the recent advances in techniques for self-interference suppression at the relay. Thanks to such suppression

mechanisms, the signal leakage from the transmitter to the co-located receiver at the relay can be reduced, thus allowing for higher data rates. Examples of techniques for reducing self-interference in FD relay networks can be found, e.g., in [7]–[11]. Other works have instead focused on the characterization of such self-interference [12] and modelled it as a Gaussian noise affecting the relay receiver, with energy proportional to the relay transmit power [13], [14]. A third body of works have evaluated the impact of residual self-interference on the performance of FD relay networks, or have designed and analyzed solutions to cope with that. In particular, [15] focuses on the capacity of the Gaussian two-hop FD relay channel, assuming that the relay adopts the decoding-and-forward (DF) scheme, the nodes have a single antenna, and the source has perfect knowledge of the symbols transmitted by the relay at any time instant. Importantly, under the above conditions, [15] derives the optimal probability distribution of the source input and of the relay input. MIMO FD communications are instead addressed, e.g., in [16]–[23]. In the absence of direct link and in the presence of self-interference, [16] derives the capacity of the relay channel for the amplify-and-forward (AF) relaying scheme. AF relays and multiple transmit antennas are also considered in [17] where the SINR at the receiver is maximized as the relay transmit power varies. The studies in [18]–[22], instead, consider the design of linear source and relay precoders for MIMO FD AF relay communications. In particular, [18] derives a closed-form expression for precoding at the relay with rank-1 zero-forcing self-interference suppression, while [19] focuses on the effect of the residual self-interference due to imperfect cancellation. The work in [20] aims at maximizing the sum rate for a two-way relay channel, by jointly optimizing the source transmit power and the beamforming matrix at the relay. With a similar goal in mind, in [21] beamforming at the source is optimized while accounting for relay processing. At last, [22] presents a hybrid beamforming scheme at the source and relay for multiuser mmwave systems. Such a scheme maximizes the worst-case sum rate in the case of imperfect channel state information (CSI), while meeting the user quality of service requirements. Considering instead DF, [23] presents an optimal relay beamforming that minimizes the outage probability in MIMO FD DF relaying, when only partial CSI is available at the receiver.

It is worth noting that relays with FD capabilities can operate in either FD mode or half-duplex (HD) mode (the two of them also jointly referred to as X-duplex mode). Therefore, depending on the channel conditions and on the constraints on the system parameters, optimal transmission policies can be based on hybrid FD/HD relaying techniques [24]–[26]. Specifically, the authors in [24] consider the instantaneous and average spectral efficiency of a dual-hop

network with direct link between source and destination. They propose hybrid FD/HD relaying policies that optimally switch between the two operational modes, depending on the channel conditions and on the relay transmit power. The study in [25] considers a DF relaying technique and evaluates the block error rate for FD and HD, ultra reliable short-packet communications, and finite block-length codes. Finally, [26] derives the optimal transmission mode and power allocation in *single-antenna* FD relay networks when the relay implements the DF technique and only the relay transmit power distribution is available at the source.

In this paper, we consider both a MIMO and a massive MIMO (mMIMO) two-hop network where no direct source-destination link exists and the relay adopts the DF scheme. The relay operates in X-duplex mode (i.e., FD or HD) and the source has knowledge only of the transmit power distribution adopted by the relay over a given time horizon. More specifically, the source is not aware of the symbols that the relay is transmitting towards the destination. Such a scenario reflects the practical case in which the relay performs link-layer or physical-layer encryption, or it inserts in-band control information [27], [28]. Indeed, in all such situations the relay may send to the destination a different sequence of symbols with respect to the one received from the source. Furthermore, we consider that constraints on the average and maximum values of transmit power at the source and relay nodes may be in place. The input distribution at the source and the relay is assumed to be Gaussian, with variance not exceeding a given maximum value. With regard to the residual self-interference, this is modeled as an additive Gaussian noise with variance proportional to the relay transmit power, as often done in the related literature.

Given the above scenario, our objective is to maximize the data rate from the source to the destination, by optimally setting over time the transmit power to be used at the source node and at the relay node. It is worth noting that the latter also implies optimally setting the time fraction during which the relay should operate in FD and HD mode, as a null source or relay power correspond to the HD mode and positive values at both the source and the relay correspond to the FD mode. We remark that, to the best of our knowledge, none of the existing works has tackled the above aspects in a MIMO FD system using DF where the source has no full knowledge of the symbols transmitted by the relay. More in detail, our main contributions are as follows:

(i) We formulate an optimization problem to maximize the achievable rate in a MIMO, FD, 2-hop network. Importantly, we take the distribution of the transmit power at the relay as the main decision variable. Then we maximize the rate over the source-relay link, conditioned to

the value of the relay transmit power, so as to obtain a closed-form expression for the optimal transmit power at the source. Based on this result, we can identify different operational regions as the transmit power at the source and the relay, and the corresponding rate value over the first hop, vary.

(ii) The closed-form expression of the conditional achievable rate on the first hop is then plugged into the original problem formulation. In light of the problem complexity, we present an approximate version of the problem that can be efficiently solved numerically. Furthermore, we derive an analytical expression for the lower bound to the source-destination achievable rate and characterize the performance in the absence of residual self-interference.

(iii) We study an mMIMO scenario where the large number of antennas at the transmitting and receiving nodes allows for an asymptotic analysis, leading to an accurate, yet tractable, semi-analytical expression of the maximum achievable rate. Importantly, such an expression does not depend on the instantaneous channel conditions, rather on the asymptotic cumulative density function of the channel matrix eigenvalues. We exploit these results to derive a *hybrid* approach combining the optimal asymptotic distribution of the transmit power at the relay with the achievable rate maximization in the case of a finite number of antennas.

(iv) Through extensive numerical results, we show the behavior of the maximum achievable rate as the average and maximum transmit power values vary. Furthermore, and very relevantly, we demonstrate that our hybrid solution holds tight also in scenarios with a small number of antennas. This suggests that the source and relay transmit power do not need to be recomputed at every new channel instance, rather they can be computed once over a relatively long time horizon and still provide performance close to the optimum.

We remark that our analytical derivations optimizing the efficiency of MIMO-FD and mMIMO-FD relay systems can play an important role in the design of 5G-and-beyond wireless networks. Indeed, by identifying the optimal transmission power and operational mode of source and relay substantially contributes to improving the system performance in terms of data rate and power consumption.

The remainder of the paper is organized as follows. Sec. II provides some background on precoding for MIMO communications, reporting the main results that we leverage in our analysis. Sec. III introduces the system model and the formulation of the optimization problem, accounting for the system constraints. Then Sec. IV focuses on the maximization of the rate on the source-relay link, conditioned to the transmit power used at the relay. The complexity of the resulting

problem is discussed in Sec. V, where bounds and an approximate solution are also presented. The mMIMO case, along with an efficient semi-asymptotic approximation for the finite case, is analysed in Sec. VI, while Sec. VII shows some performance results. Finally, Sec. VIII concludes the paper.

II. Preliminaries on SVD precoding for MIMO

In this section, we recall the expression of the achievable rate in a MIMO system with n transmit and $m \geq n$ receive antennas, when an optimized singular value decomposition (SVD) precoder [29] is used at the transmitter. The received signal is given by $\mathbf{y} = \sqrt{p\alpha}\mathbf{H}\mathbf{\Pi}\mathbf{x} + \boldsymbol{\eta}$ where p is the transmit power, α is the path loss coefficient, $\mathbf{H}$ is the channel matrix, and $\mathbf{x}$ is the Gaussian complex channel input with $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{H}}] = \frac{1}{n}\mathbf{I}$. Also, $\mathbf{\Pi}$ is the precoding matrix and $\boldsymbol{\eta}$ is the Gaussian complex noise with covariance $\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^{\mathsf{H}}] = I\mathbf{I}$. Since the precoding matrix $\mathbf{\Pi}$ should not affect the average power of the transmitted signal, we must impose $\mathbb{E}[|\mathbf{\Pi}\mathbf{x}|^2] = 1$, which, by defining $\mathbf{Q} = \mathbf{\Pi}\mathbf{\Pi}^{\mathsf{H}}$, implies $\mathsf{Tr}\{\mathbf{Q}\} = n$. For a generic precoder applied at the transmitter, and for Gaussian input, the mutual information between $\mathbf{x}$ and $\mathbf{y}$ is given by $\rho = \log\left|\mathbf{I} + \frac{\gamma}{n}\mathbf{H}\mathbf{Q}\mathbf{H}^{\mathsf{H}}\right|$ with $\gamma = \frac{p\alpha}{I}$ being the signal to noise and interference ratio (SINR). Let $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$ be the singular value decomposition of $\mathbf{H}$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$ is the corresponding matrix of singular values. Also, let $n' \leq n$ be the rank of $\mathbf{H}$. If the matrix $\mathbf{H}$ is perfectly known at the transmitter, the precoder $\mathbf{\Pi}^\star$ maximizing ρ is such that [29] $\mathbf{\Pi}^\star\mathbf{\Pi}^{\star\mathsf{H}} = \mathbf{Q}^\star = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathsf{H}}$ where $\mathbf{D}$ is a diagonal matrix whose elements are $[\mathbf{D}]_{i,i} = \left[\phi - \frac{\zeta}{\gamma\lambda_i}\right]^+$, for $i = 1, \ldots, n'$, and $[\mathbf{D}]_{i,i} = 0$ for $n' < i \leq n$. Moreover, $[\cdot]^+ = \max\{0, \cdot\}$, $\lambda_i = \frac{1}{m}\left[\mathbf{\Sigma}\mathbf{\Sigma}^{\mathsf{H}}\right]_{i,i}$, $\zeta = \frac{n}{m}$, and the parameter ϕ satisfies the constraint:

$$\mathsf{Tr}\{\mathbf{Q}^\star\} = \mathsf{Tr}\{\mathbf{D}\} = \sum_{i=1}^{n'}\left[\phi - \frac{\zeta}{\gamma\lambda_i}\right]^+ = n\,. \tag{1}$$

By replacing the expression for $\mathbf{Q}^\star$ in the mutual information, we get:

$$\rho = \sum_{i=1}^{n'}\left[\log\frac{\phi\gamma\lambda_i}{\zeta}\right]^+. \tag{2}$$

Assume the eigenvalues λ_i to be ordered in decreasing order, i.e, $\lambda_i \geq \lambda_j$ for $i < j$, then, given γ and ϕ, the terms $\phi - \frac{\zeta}{\gamma\lambda_i}$ $(1 \leq i \leq n')$ in (1) decrease as i increases. Let $k \leq n'$ be the largest integer such that $\phi - \frac{\zeta}{\gamma\lambda_k} > 0$. Then we can solve (1) for ϕ and obtain $\phi = \frac{n}{k}\left(1 + \frac{\zeta}{\gamma}\ell_k\right)$

with $\ell_k = \frac{1}{n}\sum_{i=1}^{k}\frac{1}{\lambda_i}$. By substituting such value of ϕ in (2) and defining $c_k = \frac{\zeta k}{n\lambda_k} - \zeta\ell_k$ $(k = 1, \ldots, n')$, we can rewrite the mutual information as:

$$\rho = \sum_{i=1}^{k} \log\left[\frac{n\lambda_i}{k}\left(\frac{\gamma}{\zeta} + \ell_k\right)\right] \tag{3}$$

for $k = 1, \ldots, n' - 1$ and $c_k \leq \gamma \leq c_{k+1}$, and for $k = n'$ and $\gamma \geq c_{n'}$.

III. System model and problem formulation

In this section, we first introduce the network system under study and the communication model. Then, we formulate the optimization problem, maximizing the achievable data rate and accounting for the system constraints.

A. Communication network

We consider a two-hop, FD relay network where the source $\mathcal{S}$ and the relay $\mathcal{R}$ are equipped with n_1 and n_2 transmit antennas, respectively, while the relay and the destination $\mathcal{D}$ are equipped with $m_1 \geq n_1$ and $m_2 \geq n_2$ receive antennas, respectively. In the following, we will use the subscripts 1 and 2 to differentiate the parameters and variables related, respectively, to the first and second hop of the relay network.

The relay decodes the information received from the source, then it re-encodes the information and forwards it toward $\mathcal{D}$ (DF mode). To derive the maximum achievable rate, we assume that the source is saturated, i.e., it has always data to transmit; also, we consider that source and relay perform data transmissions according to a time division strategy. Importantly, we account for the practical case where source and relay are not synchronized on a per-symbol basis, and that the relay can modify link-layer or physical-layer control information as well as perform link-layer data encryption. This implies that, unlike most of the existing work, the source may not be aware of the symbols that the relay is transmitting towards the destination.

As for the communication channel, we assume that it remains constant during a time period T. Within T, source and relay can operate in one of the following modes: (i) $\mathcal{S}$ transmits while $\mathcal{R}$ receives only (HD-RX mode); (ii) $\mathcal{S}$ is silent while $\mathcal{R}$ transmits only (HD-TX mode); (iii) both $\mathcal{S}$ and $\mathcal{R}$ transmit at the same time and $\mathcal{R}$ operates in FD (FD mode). Notice that, within T, there may be multiple *phases*, each corresponding to one of the above operational modes and to specific values of transmit power at the source and at the relay. A pictorial representation is provided in Fig. 1.

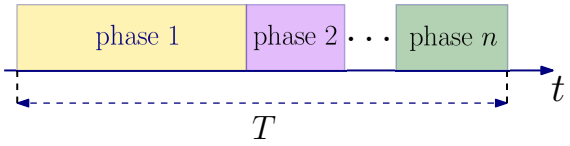

Fig. 1. Structure of the time period.

Let p_1 and p_2 be the average transmit power used by, respectively, $\mathcal{S}$ and $\mathcal{R}$, in a given phase, such that $p_1 \leq p_1^{\max}$ and $p_2 \leq p_2^{\max}$ so as to reflect the fact that in real-world systems the transmit power is limited to a maximum value. Additionally, as commonly assumed in communication systems, we impose that the average value of the transmit power over a period T is equal to $\mathbb{E}[p_1] = \bar{p}_1$ at the source and $\mathbb{E}[p_2] = \bar{p}_2$ at the relay.

In a given phase, the signal received at the relay can be expressed as

$$\mathbf{y}_1 = \sqrt{p_1 \alpha_1}\mathbf{H}_1\mathbf{\Pi}_1\mathbf{x}_1 + \boldsymbol{\nu} + \boldsymbol{\eta}_1 \qquad \text{where} \tag{4}$$

- $\mathbf{x}_1$ is the vector of symbols transmitted by $\mathcal{S}$, which are assumed to be Gaussian, complex, multivariate distributed with zero mean and covariance $\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^{\mathsf{H}}] = \frac{1}{n_1}\mathbf{I}$;
- $\mathbf{\Pi}_1$ is the optimal precoding matrix at $\mathcal{S}$ such that $\mathsf{Tr}\{\mathbf{\Pi}_1\mathbf{\Pi}_1^{\mathsf{H}}\} = n_1$;
- α_1 is the path loss coefficient;
- $\boldsymbol{\eta}_1$ is a random vector, independent of $\mathbf{x}_1$, accounting for thermal noise and interference: it is modeled as a complex Gaussian, multivariate random variable with independent and identically distributed (iid) entries, zero mean, and covariance $\mathbb{E}[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\mathsf{H}}] = I_1\mathbf{I}$;
- $\mathbf{H}_1$ is the channel matrix corresponding to the $\mathcal{S}$-$\mathcal{R}$ link, with rank $n_1' \leq n_1$; we assume that $\mathbf{H}_1$ (i.e., perfect CSI) is known at both $\mathcal{S}$ and $\mathcal{R}$;
- $\boldsymbol{\nu}$ is the residual self-interference at $\mathcal{R}$, modeled as a complex Gaussian, multivariate random variable independent of $\boldsymbol{\eta}_1$ and $\mathbf{x}_1$, with zero mean and covariance $\mathbb{E}[\boldsymbol{\nu}\boldsymbol{\nu}^{\mathsf{H}}] = \beta p_2 \mathbf{I}$ [14]. This implies that, as also done in [13], [14], the self-interference covariance matrix is proportional to the relay average transmit power p_2 through the coefficient β, representing the self-interference attenuation factor. We remark that β can take on any arbitrary value, and that $\mathbb{E}[\boldsymbol{\nu}\boldsymbol{\nu}^{\mathsf{H}}] = 0$ whenever the relay is silent (i.e., $p_2 = 0$ and there is no self-interference).

The signal received at the destination is instead given by

$$\mathbf{y}_2 = \sqrt{p_2 \alpha_2}\mathbf{H}_2\mathbf{\Pi}_2\mathbf{x}_2 + \boldsymbol{\eta}_2 \qquad \text{where} \tag{5}$$

- $\mathbf{x}_2$ is the vector of random symbols transmitted by $\mathcal{R}$, assumed to be Gaussian, complex multivariate with zero mean and covariance $\mathbb{E}[\mathbf{x}_2\mathbf{x}_2^{\mathsf{H}}] = \frac{1}{n_2}\mathbf{I}$;
- $\mathbf{\Pi}_2$ is the optimal precoding matrix at $\mathcal{R}$ such that $\mathsf{Tr}\{\mathbf{\Pi}_2\mathbf{\Pi}_2^{\mathsf{H}}\} = n_2$;
- α_2 is the path loss coefficient on the $\mathcal{R}$-$\mathcal{D}$ link;
- $\mathbf{H}_2$ is the $\mathcal{R}$-$\mathcal{D}$ channel matrix, with rank $n_2' \leq n_2$; perfect CSI knowledge is assumed at $\mathcal{R}$;
- $\boldsymbol{\eta}_2$ represents the noise plus interference at $\mathcal{D}$ and is modeled as a Gaussian, multivariate complex random variable independent of $\mathbf{x}_2$, with covariance $\mathbb{E}[\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\mathsf{H}}] = I_2\mathbf{I}$.

We underline that the precoding matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ are optimized as described in Sec. II.

Next, let $\lambda_{j,i}$ be the i-th ordered eigenvalue of $\frac{1}{m_j}\mathbf{H}_j^{\mathsf{H}}\mathbf{H}_j$, and let us define

$$\zeta_j = \frac{n_j}{m_j}; \quad \ell_{j,k} = \frac{1}{n_j}\sum_{i=1}^{k}\frac{1}{\lambda_{j,i}}; \quad c_{j,k} = \frac{\zeta_j k}{n_j\lambda_{j,k}} - \zeta_j\ell_{j,k}; \quad \gamma_1(p_1,p_2) = \frac{p_1\alpha_1}{I_1+\beta p_2}; \quad \gamma_2(p_2) = \frac{p_2\alpha_2}{I_2}\,.$$

Based on the above model and definitions, we can write the expression of the achievable rate on the $\mathcal{S}$-$\mathcal{R}$ and $\mathcal{R}$-$\mathcal{D}$ links, as

$$\rho_1(p_1,p_2) = \sum_{i=1}^{k}\log\left[\frac{n_1\lambda_{1,i}}{k}\left(\frac{\gamma_1(p_1,p_2)}{\zeta_1}+\ell_{1,k}\right)\right] \tag{6}$$

for $k = 1,\ldots,n_1'-1$ and $c_{1,k} \leq \gamma_1(p_1,p_2) < c_{1,k+1}$, and $k = n_1'$ and $\gamma_1(p_1,p_2) \geq c_{1,n_1'}$, and

$$\rho_2(p_2) = \sum_{i=1}^{k}\log\left[\frac{n_2\lambda_{2,i}}{k}\left(\frac{\gamma_2(p_2)}{\zeta_2}+\ell_{2,k}\right)\right] \tag{7}$$

for $k = 1,\ldots,n_2'-1$ and $c_{2,k} \leq \gamma_2(p_2) < c_{2,k+1}$, and for $k = n_2'$ and $\gamma_2(p_2) \geq c_{2,n_2'}$. Note that both ρ_1 and ρ_2 are piecewise functions composed of n_1' and n_2' pieces, respectively. While a more elaborated analysis is needed to further characterize the behavior of ρ_1, for ρ_2 we can already identify the points of transition between adjacent pieces of the rate function by solving $\gamma_2(p_2) = c_{2,k}$ for p_2. By doing that, we get:

$$\tau_k \triangleq \frac{\zeta_2 I_2}{\alpha_2}\left(\frac{k}{n_2\lambda_{2,k}} - \ell_{2,k}\right) \tag{8}$$

where we recall that $\lambda_{2,k}$ is the k-th eigenvalue of $\frac{1}{m_2}\mathbf{H}_2^{\mathsf{H}}\mathbf{H}_2$ and $\ell_{2,k} = \frac{1}{n_2}\sum_{i=1}^{k}\frac{1}{\lambda_{2,i}}$. Importantly, we remark that $\rho_2(p_2)$ is an increasing and concave function.

Finally, let $g(p_2)$, $p_2 \in [0, p_2^{\max}]$, be the distribution of the transmit power p_2 over period T at $\mathcal{R}$; thus

$$\int_0^{p_2^{\max}} g(p_2)\,\mathrm{d}p_2 = 1; \quad g(p_2) \geq 0\,. \tag{9}$$

Then, the average rates achieved over T on the $\mathcal{S}$-$\mathcal{R}$ and $\mathcal{R}$-$\mathcal{D}$ links are given by, respectively,

$$R_1(g,p_1)=\int_0^{p_2^{\max}} g(p_2)\rho_1(p_1,p_2)\,\mathrm{d}p_2; \quad R_2(g)=\int_0^{p_2^{\max}} g(p_2)\rho_2(p_2)\,\mathrm{d}p_2 \tag{10}$$

where the used notations highlight that the average rates depend on the specific choice of the distribution $g(\cdot)$. It is then easy to see that maximizing the achievable rate over a period T implies to optimally characterize the different phases in T, i.e., to determine: (i) the optimal values of transmit power to be used in each phase at the source, p_1, and at the relay, p_2 (or, equivalently, $g(p_2)$), and (ii) the optimal duration of each phase.

B. Maximizing the achievable rate

Let R be the achievable rate at which data are transferred from $\mathcal{S}$ to $\mathcal{D}$ in period T. We first observe that the achievable rate is given by the minimum between the rate over the $\mathcal{S}$-$\mathcal{R}$ link, and the one over the $\mathcal{R}$-$\mathcal{D}$ link. Therefore, the maximum achievable rate supported by the system can be found by solving

$$\textbf{P0:} \quad R=\max_{g(\cdot),p_1}\min\left\{R_1(g,p_1),R_2(g)\right\}=\max_{g(\cdot)}\min\left\{\max_{p_1}R_1(g,p_1),R_2(g)\right\} \tag{11}$$

subject to the constraints in (9), $\mathbb{E}[p_1]=\bar{p}_1$, $\mathbb{E}[p_2]=\bar{p}_2$, $p_2\in[0,p_2^{\max}]$, and $p_1\in[0,p_1^{\max}]$.

Looking at the expression of $R_1(g,p_1)$ in (10) and $\rho_1(p_1,p_2)$, we notice that the SINR, hence the rate, depends on both p_1 and p_2, as the residual self-interference creates a dependency between the performance of the first and the second link. Thus, to maximize the rate over the first hop, source and relay should coordinate their power allocation strategies. We then write the transmit power at $\mathcal{S}$ as a function of the transmit power at $\mathcal{R}$ ($p_1=p_1(p_2)$); as a consequence, the constraints on the average power over period T for the $\mathcal{S}$-$\mathcal{R}$ and the $\mathcal{R}$-$\mathcal{D}$ link are given by

$$\mathbb{E}[p_2]=\int_0^{p_2^{\max}} p_2 g(p_2)\,\mathrm{d}p_2=\bar{p}_2; \quad \mathbb{E}[p_1]=\int_0^{p_2^{\max}} p_1(p_2)g(p_2)\,\mathrm{d}p_2=\bar{p}_1\,. \tag{12}$$

Given the above expressions and as detailed in the following sections, we can consider *the distribution of the power at the relay, $g(\cdot)$, as the only decision variable* in our optimization problem **P0**. Indeed, the optimal transmit power at the source depends on p_2, and the optimal values of p_1 and p_2 determine the optimal source and relay operational modes. As an example, the case $p_1>0$, $p_2>0$ corresponds to the relay operating in FD mode, while $p_1=0$ and $p_2>0$ correspond to the HD-TX mode. Additionally, as will become clear in Sec. IV-B, given the optimal $g(\cdot)$, we can identify the optimal duration of the temporal phases within T, each corresponding to a different operational mode.

With the aim to solve **P0**, in the following we first maximize the rate on the $\mathcal{S}$-$\mathcal{R}$ link, thus finding the optimal expression for the transmit power at the source as a function of p_2. Then we optimize the minimum between R_1 and R_2 with respect to the distribution $g(\cdot)$.

IV. Rate optimization on the source-relay link

Here, we first derive the expression of the optimal transmit power p_1 (Sec. IV-A), which allows us to write the SINR at $\mathcal{R}$ and, hence, the rate achievable on the $\mathcal{S}$-$\mathcal{R}$ link (Sec. IV-B). We also provide numerical examples illustrating the behavior of such quantities, as the transmit power at the relay, p_2, varies.

A. Optimizing the transmit power at the source

The maximization over p_1 of $R_1(g,p_1)$ (see (11)) can be written as

$$\textbf{P1:}\quad R_1^\star(g)=\max_{p_1(\cdot)}\int_0^{p_2^{\max}} g(p_2)\rho_1(p_1,p_2)\,\mathrm{d}p_2 \qquad \text{s.t.}$$

$$\text{(a)}\quad 0\le p_1\le p_1^{\max};\quad \text{(b)}\quad \int_0^{p_2^{\max}} g(p_2)p_1(p_2)\,\mathrm{d}p_2=\bar{p}_1$$

where (a) is the constraint on the maximum transmit power at the source and (b) is the constraint on the average value of p_1 appearing in (12).

Proposition 4.1: The maximizer of the argument in **P1** is given by:

$$p_1^\star(p_2)=\begin{cases} p_1^{\max}, & \text{if } \ p_2\le\widehat{\omega} \\ \frac{\psi k}{n_1}-\frac{I_1+\beta p}{\alpha_1}\zeta_1\ell_{1,k}, & \text{if } \ k=1,\ldots,\widehat{k}-1 \text{ and } \omega_{k+1}\le p_2<\omega_k;\ k=\widehat{k} \text{ and } \widehat{\omega}\le p_2<\omega_{\widehat{k}} \\ 0 & \text{if } \ p_2\ge\omega_1 \end{cases} \tag{13}$$

where ψ is a parameter satisfying $\int_0^{p_2^{\max}} g(p_2)p_1^\star(p_2)\,\mathrm{d}p_2=\bar{p}_1$, $\widehat{\omega}=\alpha_1\frac{\widehat{k}\psi/n_1-p_1^{\max}}{\zeta_1\ell_{1,\widehat{k}}\beta}-\frac{I_1}{\beta}$, $\omega_k\triangleq\frac{\psi\lambda_{1,k}\alpha_1}{\zeta_1\beta}-\frac{I_1}{\beta}$, and $\widehat{k}=\arg\max_{k=1,\ldots,n_1'}\frac{k\psi/n_1-p_1^{\max}}{\ell_{1,k}}$.

Proof: The proof is provided in App. A. ∎

An example reporting the behavior of $p_1^\star(p_2)$ is depicted in Fig. 2 where $\widehat{k}=2$ and the shaded regions show the corresponding relay modes. Looking at Fig. 2, we can give the following interpretation of (13):

(a) $p_1^\star(p_2)$ decreases as the relay transmit power, p_2, increases. This is due to the fact that, as p_2 increases, the amount of residual self-interference at $\mathcal{R}$ grows (i.e., the $\mathcal{S}$-$\mathcal{R}$ channel becomes noisier). Thus, due to the constraint on the average transmit power, beyond a

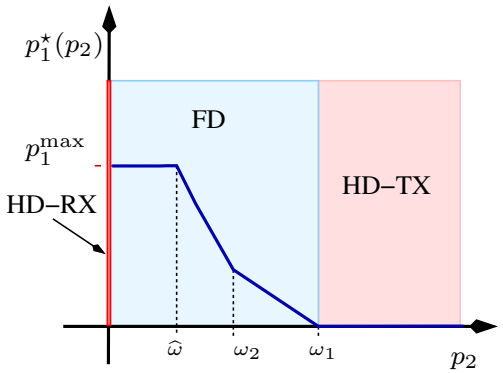

Fig. 2. Behavior of the function $p_1^\star(p_2)$. The shaded regions denote the different operational modes at the relay.

certain value of p_2 it is not convenient for the source to transfer data towards the relay. In other words, given its limited power budget, the optimal choice for the source becomes to save power and transmit when the level of self-interference is lower;

(b) $p_1^\star(p_2)$ is made of $\widehat{k}+2$ linear pieces, each corresponding to a specific operational region at $\mathcal{S}$;

(c) for $\widehat{\omega} \leq p_2 < \omega_1$, $p_1^\star(p_2)$ is made of $\widehat{k}$ pieces; the k-th piece corresponds to a situation where the source transmit power is allocated to the first k (best) parallel channels defined by the eigenmodes of the matrix $\mathbf{H}_1$;

(d) when $p_2 \geq \omega_1$, the $\mathcal{S}$-$\mathcal{R}$ link experiences such high levels of self-interference that the best strategy for $\mathcal{S}$ becomes to be silent ($p_1^\star = 0$) and for the relay to work in HD-TX mode;

(e) in the region $0 < p_2 < \omega_1$, the transmit power at both source and relay is strictly positive, hence the relay works in FD mode;

(f) when $p_2 = 0$, the transmit power at $\mathcal{S}$ is positive, thus the relay works in HD-RX mode;

(g) for $p_2 \leq \widehat{\omega}$, $p_1^\star$ takes its maximum value, $p_1^{\max}$.

Remark 4.1: Depending on the value of the system parameters, some of the threshold points, ω_k, of transition between different operational regions, as well as $\widehat{\omega}$, can assume negative values. However, being p_2 the transmit power at the relay, $p_1^\star(p_2)$ can only be evaluated for $p_2 \geq 0$, thus only the subset of positive values within each operational region should be considered.

B. Rate characterization on the source-relay link

When the optimized transmit power $p_1^\star(p_2)$ is adopted at the source, the SINR at $\mathcal{R}$ is given by $\gamma_1^\star(p_2) = \frac{\alpha_1 p_1^\star(p_2)}{I_1+\beta p_2}$, which, according to (13) and for $p_2 \in [0, p_2^{\max}]$, can be rewritten as

$$\gamma_1^\star(p_2)=\begin{cases} \frac{p_1^{\max}\alpha_1}{I_1+\beta p_2} & \text{if } p_2<\widehat{\omega} \\ \frac{\psi k \alpha_1}{n_1(I_1+\beta p_2)}-\zeta_1 \ell_{1,k} & \text{if } \omega_{k+1}\leq p_2<\omega_k, k=1,\ldots,\widehat{k}-1; \widehat{\omega}\leq p_2<\omega_{\widehat{k}}, k=\widehat{k} \\ 0 & \text{if } \omega_1 \leq p_2 \end{cases} . \tag{14}$$

We then substitute such an expression for $\gamma_1^\star(p_2)$ in (6) to obtain the achievable rate on the $\mathcal{S}$-$\mathcal{R}$ link: $\rho_1^\star(p_2) \triangleq \rho_1(p_1^\star(p_2), p_2)$. Since both (6) and (14) are piecewise defined functions, so is $\rho_1^\star(p_2)$. The transition points (between pieces, or, equivalently, regions) correspond to those values of p_2 located on the boundary between two adjacent pieces. An important transition point is $p_2 = \widehat{\omega}$, which separates the region where the source power reaches the maximum possible value, $p_1^\star = p_1^{\max}$, from the region where $p_1^\star < p_1^{\max}$. An example of $\rho_1^\star(p_2)$ is illustrated in Figures 3(left) and 3(center) where the transition point $\widehat{\omega}$ corresponds to the regions boundary highlighted by the red dashed line. In the region $p_2 > \widehat{\omega}$, the rate $\rho_1^\star(p_2)$ is piecewise defined and transitions between adjacent pieces occur when the SINR $\gamma_1^\star(p_2)$ crosses one of the thresholds $c_{1,k}$, $k = 1, \ldots, \widehat{k}$. Such transition points, whose expressions are reported after (13), are denoted with ω_k. For $p_2 < \widehat{\omega}$, the rate $\rho_1^\star(p_2)$ is, again, piecewise defined and the corresponding transition points are denoted with θ_k, $k = \widehat{k}+1, \ldots, n_1'$. Their expression can be obtained by solving for p_2 the equation $\gamma_1^\star(p_2) = \frac{p_1^{\max}\alpha_1}{I_1+\beta p_2} = c_{1,k}$, which yields (see the SINR expression for $p_2 < \widehat{\omega}$ in (14) and that for $c_{1,k}$) $\theta_k = \frac{p_1^{\max}\alpha_1}{\beta\zeta_1}\left(\frac{k}{n_1\lambda_{1,k}} - \ell_{1,k}\right)^{-1} - \frac{I_1}{\beta}$.

Summarizing, using (6) and (14), for $p_2 \in [0, p_2^{\max}]$ and $\widehat{k} < n_1'$, we have:

$$\rho_1^\star(p_2) = \begin{cases} \sum_{i=1}^{k} \log\left(\frac{n_1\lambda_{1,i}}{k}\left[\frac{p_1^{\max}\alpha_1/\zeta_1}{I_1+\beta p_2}+\ell_{1,k}\right]\right) & \text{if } \theta_{k+1} \leq p_2 < \theta_k, k = \widehat{k}+1, \ldots, n_1'-1; \\ & \theta_{\widehat{k}+1} \leq p_2 < \widehat{\omega}, k = \widehat{k}; \quad p_2 < \theta_{n_1}, k = n_1' \\ \sum_{i=1}^{k} \log\left(\frac{\psi\lambda_{1,i}\alpha_1}{\zeta_1(I_1+\beta p_2)}\right) & \text{if } \omega_{k+1} \leq p_2 < \omega_k, k = 1, \ldots, \widehat{k}-1 \\ & \widehat{\omega} \leq p_2 < \omega_{\widehat{k}}, k = \widehat{k} \\ 0 & \text{if } p_2 \geq \omega_1 \end{cases} \tag{15}$$

while, for $\widehat{k} = n_1'$, we get:

$$\rho_1^\star(p_2)=\begin{cases} \sum_{i=1}^{n_1'} \log\left(\lambda_{1,i}\left[\frac{p_1^{\max}\alpha_1/\zeta_1}{I_1+\beta p_2}+\ell_{1,n_1'}\right]\right) & \text{if } p_2 < \widehat{\omega} \\ \sum_{i=1}^{k} \log\left(\frac{\psi\lambda_{1,i}\alpha_1}{\zeta_1(I_1+\beta p_2)}\right) & \text{if } \omega_{k+1} \leq p_2 < \omega_k, k = 1, \ldots, n_1'-1 \\ & \widehat{\omega} \leq p_2 < \omega_{n_1}, k = n_1' \\ 0 & \text{if } p_2 \geq \omega_1 . \end{cases} \tag{16}$$

As can be seen from the above expression, for $p_2 \geq \widehat{\omega}$, the rate $\rho_1^\star(p_2)$ is continuous and convex, although not differentiable in $p_2 = \omega_k$, $k = 1, \ldots, \widehat{k}$. For $p_2 < \widehat{\omega}$, the rate $\rho_1^\star(p_2)$ is convex, and differentiable everywhere. However, although made of convex pieces, $\rho_1^\star(p_2)$ is not globally convex, since in $p_2 = \widehat{\omega}$ we have: $\lim_{p_2=\widehat{\omega}^-} \frac{\mathrm{d}\rho_1^\star(p_2)}{\mathrm{d}p_2} > \lim_{p_2=\widehat{\omega}^+} \frac{\mathrm{d}\rho_1^\star(p_2)}{\mathrm{d}p_2}$. Such considerations on the convexity of $\rho_1^\star(p_2)$ will be helpful for deriving the lower-bound in Sec. V-B.

The explicit expression of the rate $\rho_1^\star(p_2)$ in (15) and (16) allows us to write the solution to problem **P1** as $R_1^\star(g) = \int_0^{p_2^{\max}} g(p_2)\rho_1^\star(p_2)\,\mathrm{d}p_2$. Equipped with this result, we can now rewrite problem **P0** as

$$\textbf{P2:} \quad R = \max_{g(\cdot)} \min \left\{ \int_0^{p_2^{\max}} g(p_2)\rho_1^\star(p)\,\mathrm{d}p_2, \int_0^{p_2^{\max}} g(p_2)\rho_2(p_2)\,\mathrm{d}p_2 \right\} \text{s.t. (9) and (12).}$$

By solving **P2**, we can find the network data rate R and the optimal transmit power distribution at the relay, $g(\cdot)$.

Example 1: An example showing the behavior of $p_1^\star(p_2)$ and $\rho_1^\star(p_2)$ is depicted in Fig. 3(left) where $n_1 = m_1 = 3$, $\lambda_{1,i} = 7 - i$, $i = 1, \ldots, 6$, $\beta = \psi = \alpha_1 = I_1 = 1$, and $p_1^{\max} = 0.3$. Under this setting, we have $n_1 = n_1'$, $\zeta_1 = 1$, $\widehat{\omega} = 0.32$, $\widehat{k} = 2$, $\omega_1 = 2$, $\omega_2 = 1$, and $\theta_3 = -0.2286$. The latter is not shown in the figure since, being negative, it must be discarded.

Example 2: Fig. 3(center) still shows the behavior of $p_1^\star(p_2)$ and $\rho_1^\star(p_2)$, but under different settings. We have: $n_1 = n_1' = m_1 = 6$, $\lambda_{1,i} = 7 - i$, $i = 1, \ldots, 6$, $\beta = \psi = \alpha_1 = I_1 = 1$, and $p_1^{\max} = 0.15$. In this case, we get $\zeta_1 = 1$, $\widehat{\omega} = 2.4054$, $\widehat{k} = 3$, $\omega_k = 6 - k$, $k = 1, \ldots, \widehat{k}$, $\theta_4 = 1.3478$, $\theta_5 = -0.1429$, and $\theta_6 = -0.7465$. The latter two values are negative, hence they are not shown. Now assume that the solution of problem **P2** is obtained by the optimal distribution: $g^\star(p_2) = 0.2\delta(p_2) + 0.3\delta(p_2 - 2) + 0.5\delta(p_2 - 6)$ where $\delta(\cdot)$ is the Dirac delta function. Then, the optimal communication strategy will consist of 3 phases, one for each delta function. Specifically (see Fig. 3(center)),

- in phase 1, lasting $0.2T$, the relay transmit power is $p_2 = 0$, and the source transmit power is $p_1^\star(0) = p_1^{\max}$ (HD-RX mode);
- in phase 2, lasting $0.3T$, the source and relay transmit powers are both positive, namely, $p_2 = 2$ and $p_1^\star(2) = p_1^{\max}$ (FD mode);
- in phase 3, lasting $0.5T$, the source and relay transmit powers are $p_2 = 6$ and $p_1^\star(6) = 0$ (HD-TX mode).

Example 3: We now look at the behavior of $\rho_1^\star(p_2)$ and $\rho_2(p_2)$ (see Fig. 3(right)) when $n_1 = m_1 = n_2 = m_2 = 3$, $\lambda_{1,i} = 4 - i$, $i = 1, \ldots, 3$, $\lambda_{2,1} = 4$, $\lambda_{2,2} = 1/4$ $\lambda_{2,3} = 2/11$, $\beta = \psi = \alpha_1 =$

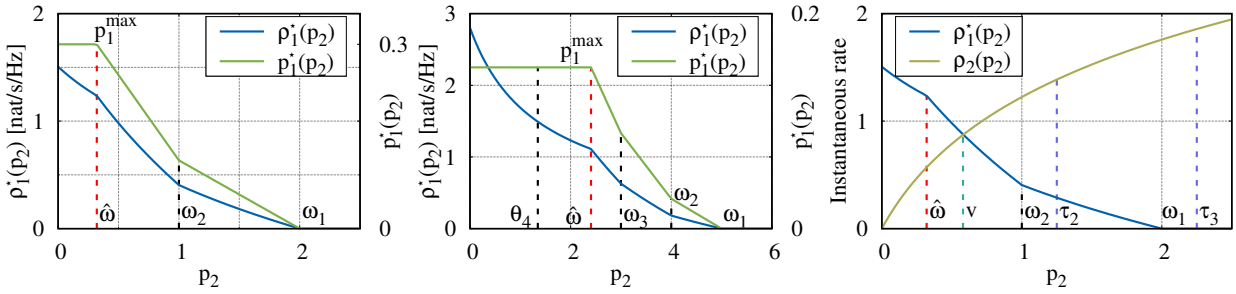

Fig. 3. Examples of the behavior of $p_1^\star$, $\rho_1^\star$, and ρ_2 as p_2 varies. The transition points of the piecewise functions are highlighted as well.

$I_1 = I_2 = 1$, $\alpha_2 = 0.2$, and $p_1^{\max} = 0.3$. We have: $n_1 = n_1'$, $n_2' = n_2$, $\zeta_1 = \zeta_2 = 1$, $\widehat{\omega} = 0.32$, $\widehat{k} = 2$, $\omega_k = 3 - k$, $k = 1, \ldots, \widehat{k}$, $\theta_3 = -0.2286$, $\tau_1 = 0$, $\tau_2 = 5/4$, and $\tau_3 = 9/4$ (the latter was defined in (8)). The figure also shows the value v for which $\rho_1^\star(v) = \rho_2(v)$.

V. Approximate solution and lower bound

We now look at the problem in **P2** and discuss its complexity. We first observe that the rate optimization has to be performed with respect to a probability function. Even in the simplest case where $n_1 = n_2 = 1$ and $p_1^{\max} \to \infty$, characterizing the solution space is extremely hard, as shown in [26]. It follows that, in the general case where $\rho_1^\star$ and ρ_2 are piecewise functions and $\rho_1^\star$ is not convex, solving **P2** is analytically unfeasible. We therefore present first an efficient numerical approximation obtained by discretizing the distribution $g(\cdot)$, then we derive a lower bound to R, which can be obtained without resorting to discretization and proves to be very tight, especially for low-medium SNR values. Finally, for the sake of completeness, we obtain the maximum achievable rate in absence of residual self-interference.

A. Approximate solution to Problem **P2**

We now present an approximate solution to Problem **P2**, obtained by discretizing the distribution $g(p_2)$ as follows:

$$g(p_2) = \sum_{n=1}^{N_p} g_n \delta(p_2 - \pi_n)$$

where N_p is the number of discretization points and $0 \leq \pi_1 < \pi_2 < \cdots < \pi_{N_p} \leq p_2^{\max}$. Let $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_{N_p}]^\mathsf{T}$, $\mathbf{g} = [g_1, \ldots, g_{N_p}]^\mathsf{T}$, $\mathbf{p}_1^\star = [p_1^\star(\pi_1), \ldots, p_1^\star(\pi_{N_p})]^\mathsf{T}$, $\boldsymbol{\rho}_1^\star = [\rho_1^\star(\pi_1), \ldots, \rho_1^\star(\pi_{N_p})]^\mathsf{T}$,

and $\boldsymbol{\rho}_2 = [\rho_2(\pi_1), \ldots, \rho_2(\pi_{N_p})]^{\mathsf{T}}$. We recall that rates $\rho_1^\star(p_2)$ and $\rho_2(p_2)$ have been defined in (15), (16), and (7), respectively. Then we can turn **P2** into the following discrete problem:

$$\textbf{P3:} \quad \widetilde{R} = \max_{\boldsymbol{\pi}} \max_{\mathbf{g}} \min \left\{ \mathbf{g}^{\mathsf{T}} \boldsymbol{\rho}_1^\star, \mathbf{g}^{\mathsf{T}} \boldsymbol{\rho}_2 \right\} \qquad \text{s.t.}$$

$$(a)\ \mathbf{g}^{\mathsf{T}} \mathbf{p}_1^\star = \bar{p}_1\,, \quad (b)\ \mathbf{g}^{\mathsf{T}} \boldsymbol{\pi} = \bar{p}_2\,, \quad (c)\ \mathbf{g} \geq \mathbf{0}\,, \quad (d)\ \mathbf{g}^{\mathsf{T}} \mathbf{1} = 1\,. \tag{17}$$

Observe that the problem in (17) requires to consider all possible choices of the vector $\boldsymbol{\pi}$, given a value of N_p. Clearly, the larger the N_p, the more cumbersome the computation of $\widetilde{R}$, but also the higher the values of rate that we obtain, since the solution space can be searched within a finer grained grid. As shown in Sec. VII, a value of $N_p = 50$ is sufficient to closely approximate R. Furthermore, a convenient, yet effective, way to solve **P3** is to fix $\boldsymbol{\pi}$ by taking N_p values logarithmically spaced in the range $[\epsilon p_2^{\max}, p_2^{\max}]$, with $\epsilon \ll 1$, including the values $\{0, \bar{p}_2\}$. In practice, the elements of $\boldsymbol{\pi}$ are given by $\pi_i = \epsilon^{1 - \frac{i}{N_p - 3}} p_2^{\max}$, for $i = 0, \ldots, N_p - 3$, $\pi_{N_p - 2} = 0$, and $\pi_{N_p - 1} = \bar{p}_2$.

The rationale behind this choice is that communication phases in a time period T may be characterized by values of the relay transmit power differing from each other by several orders of magnitude. Thus, selecting logarithmically spaced elements for $\boldsymbol{\pi}$ allows us to achieve a better approximation of R. Additionally, solutions to (11) may include phases where the relay is silent (i.e., it works in HD mode, $p_2 = 0$), transmits at average power ($p_2 = \bar{p}_2$), or maximum power ($p_2 = p_2^{\max}$); thus, we include such values of p_2 in $\boldsymbol{\pi}$ as well. By doing this, we can rewrite our problem as:

$$\textbf{P4}: \qquad \widehat{R} = \max_{\mathbf{g}} \min \left\{ \mathbf{g}^{\mathsf{T}} \boldsymbol{\rho}_1^\star, \mathbf{g}^{\mathsf{T}} \boldsymbol{\rho}_2 \right\} \qquad \text{s.t. (a), (b), (c), (d) in (17)}\,. \tag{18}$$

Importantly, problem **P4** can be solved through standard linear programming algorithms available in many dedicated software packages (e.g., Matlab or Octave, just to name a few). In a practical system, the routine solving **P4** can be executed at the relay (which has knowledge of both $\mathbf{H}_1$ and $\mathbf{H}_2$), once per channel coherence time T. The resulting optimal distribution of the relay transmit power is sent to $\mathcal{S}$, which can simply use it in the expression of $p_1^\star$ and compute its optimal transmit power. Note that, although vector $\mathbf{g}^\star$ has size N_p, only few of its elements are non-zero (in our analysis we found that at most 3 elements of $\mathbf{g}^\star$ are non-zero). It follows that, for each channel coherence time T, only such elements need to be transmitted to $\mathcal{S}$, thus entailing a very limited communication overhead.

A discussion on the accuracy of the proposed approximation method is reported in Sec. VII.

Remark 5.1: For a given input vector $\boldsymbol{\pi}$, problem **P4** provides, as output, the rate $\widehat{R}$ as well as the vector of coefficients $\mathbf{g}^\star$ maximizing the rate. In principle, within T there may be up to N_p phases, with the generic n-th phase lasting $g_n^\star T$ seconds and such that the relay transmits at average power π_n. In the practice, however, the resulting number of phases is small, since only few coefficients $g_n^\star$ are positive, as shown by our numerical results in Sec. VII.

B. Lower bound

We now aim at providing a lower bound to the maximum achievable rate, which can be derived without resorting to discretization. To this end, we consider (11) and plug in such expression of the rate the definitions of $R_1(g)$ and $R_2(g)$. By doing so, we can write:

$$\begin{aligned} R &= \max_{g(\cdot)} \min\{R_1(g), R_2(g)\} \\ &= \max_{g(\cdot)} \min\left\{ \int_0^{p_2^{\max}} g(p_2)\rho_1^\star(p_2)\,\mathrm{d}p_2, \int_0^{p_2^{\max}} g(p_2)\rho_2(p_2)\,\mathrm{d}p_2 \right\} \\ &\geq \max_{g(\cdot)} \int_0^{p_2^{\max}} g(p_2) \min\left\{\rho_1^\star(p_2), \rho_2(p_2)\right\}\,\mathrm{d}p_2 \\ &\triangleq R_{\mathrm{LB}}\,. \end{aligned} \tag{19}$$

Importantly, as shown below, one can prove that the optimal distribution of p_2 maximizing the above expression is given by the sum of two Dirac $\delta(\cdot)$ functions.

Proposition 5.1: The expression of the maximixer $g_{\mathrm{LB}}^\star(\mathrm{p}_2)$ providing the lower bound R_{LB} in (19) is given by $g_{\mathrm{LB}}^\star = u\delta(p_2-\pi_1) + (1-u)\delta(p_2-\pi_2)$, where u, π_1, π_2 are reported in Table I as the system parameters vary.

Proof: We first recall the result in [26, Lemma 4.1], which states that, given a continuous concave function, $\phi(p)$, and a probability density function (pdf), $g(p)$, with support in $[a,b]$ and average $\int_a^b pg(p)\,\mathrm{d}p = m$, then $\max_{g(\cdot)} \int_a^b \phi(p)g(p)\,\mathrm{d}p = \phi(m)$ and $\min_{g(\cdot)} \int_a^b \phi(p)g(p)\,\mathrm{d}p = \frac{b-m}{b-a}\phi(a) + \left(1-\frac{b-m}{b-a}\right)\phi(b)$ are obtained, respectively, for $g(p)=\delta(p-m)$ and $g(p)=\frac{b-m}{b-a}\delta(p-a)+\left(1-\frac{b-m}{b-a}\right)\delta(p-b)$. When instead $\phi(p)$ is convex, the expressions reported above have to be swapped.

Now, let v be the solution for p_2 of the equation $\rho_1^\star(p_2)=\rho_2(p_2)$. This solution always exists and is unique since $\rho_1^\star(p_2)$ and $\rho_2(p_2)$ are, respectively, a decreasing and an increasing function of p_2, $\rho_1^\star(0) > \rho_2(0)$, and $\rho_2(\omega_1) > \rho_1^\star(\omega_1) = 0$. We then proceed by identifying several cases

defined by sets of condones on the system parameters and we derive the expression of the bound for each of them. We first observe that, if $p_2^{\max} \leq v$, the bound in (19) can be rewritten as

$$R^{\mathrm{LB}} = \max_{g(\cdot)} \int_0^{p_2^{\max}} g(p_2)\rho_2(p_2)\,\mathrm{d}p_2 = \rho_2(\bar{p}_2)$$

by virtue of [26, Lemma 4.1]. In such a case, the maximizer is $g^\star_{\mathrm{LB}}(p_2) = \delta(p_2 - \bar{p}_2)$. Instead, when $p_2^{\max} > v$, the integral in (19) can be decomposed in a sum of two terms:

$$R^{\mathrm{LB}} = \max_{g(\cdot)} \left[\int_0^{v} g(p_2)\rho_2(p_2)\,\mathrm{d}p_2 + \int_v^{p_2^{\max}} g(p_2)\rho_1^\star(p_2)\,\mathrm{d}p_2 \right]. \tag{20}$$

Let $g_1(p_2)$ and $g_2(p_2)$ be two distributions with support, respectively, in $[v, p_2^{\max}]$ and $[0, v]$ and such that $g(p_2) = qg_1(p_2) + (1-q)g_2(p_2)$, with $q \in [0,1]$. By substituting such expression for $g(p_2)$ in the left constraint in (12), we get $(1-q)\bar{p}_{2,2} + q\bar{p}_{2,1} = \bar{p}_2$, thus $\bar{p}_{2,2} = \frac{\bar{p}_2 - q\bar{p}_{2,1}}{1-q}$ where $\bar{p}_{2,1} = \int_v^{p_2^{\max}} p_2 g_1(p_2)\,\mathrm{d}p_2$ and $\bar{p}_{2,2} = \int_0^{v} p_2 g_2(p_2)\,\mathrm{d}p_2$ are, respectively, the average value of $g_1(\cdot)$ and $g_2(\cdot)$. The maximization over $g(\cdot)$ in (20) can now be rewritten as the joint maximization over the distributions $g_1(\cdot)$ and $g_2(\cdot)$, and over the variables q and $\bar{p}_{2,1}$. Note that, since the average value of a convex distribution falls in its support, we have to impose:

$$(a) \qquad v \leq \bar{p}_{2,1} \leq p_2^{\max}; \qquad (b) \qquad 0 \leq \bar{p}_{2,2} \leq v\,. \tag{21}$$

By recalling that $\bar{p}_{2,2} = \frac{\bar{p}_2 - q\bar{p}_{2,1}}{1-q}$, the inequality (21)-(b) can be rewritten as $v + \frac{\bar{p}_2 - v}{q} \leq \bar{p}_{2,1} \leq \frac{\bar{p}_2}{q}$ which, merged with (21)-(a), provides $v + \left[\frac{\bar{p}_2 - v}{q}\right]^+ \leq \bar{p}_{2,1} \leq \min\left\{p_2^{\max}, \frac{\bar{p}_2}{q}\right\}$. The above set of values for $\bar{p}_{2,1}$ is non-empty if $v + \frac{\bar{p}_2 - v}{q} \leq p_2^{\max}$ and $v \leq \bar{p}_2/q$, i.e., if $\frac{\bar{p}_2 - v}{p_2^{\max} - v} \leq q \leq \frac{\bar{p}_2}{v}$. Finally, by recalling that $0 \leq q \leq 1$, the set of valid pairs $(q, \bar{p}_{2,1})$ is given by

$$\mathcal{Z} = \left\{ (q, \bar{p}_{2,1}) \,\middle|\, \left[\frac{\bar{p}_2 - v}{p_2^{\max} - v}\right]^+ \leq q \leq \min\left\{1, \frac{\bar{p}_2}{v}\right\}, v + \left[\frac{\bar{p}_2 - v}{q}\right]^+ \leq \bar{p}_{2,1} \leq \min\left\{p_2^{\max}, \frac{\bar{p}_2}{q}\right\} \right\}.$$

Equipped with this result, we can rewrite the bound in (20) as follows

$$R^{\mathrm{LB}} = \max_{(q,\bar{p}_{2,1}) \in \mathcal{Z}} \left[(1-q)\rho_2\left(\frac{\bar{p}_2 - q\bar{p}_{2,1}}{1-q}\right) + qW \right] \tag{22}$$

where we applied [26, Lemma 4.1], defined $W \triangleq \max_{g_1(\cdot)} \int_v^{p_2^{\max}} g_1(p_2)\rho_1^\star(p_2)\,\mathrm{d}p_2$, and exploited the relation $\bar{p}_{2,2} = \frac{\bar{p}_2 - q\bar{p}_{2,1}}{1-q}$. We now focus on the term W in (22); as shown in App. C, we have:

1) if $\widehat{\omega} \leq v$ or $\widehat{\omega} \geq p_2^{\max}$, then

$$W = \rho_1^\star(p_2^{\max}) + \frac{p_2^{\max} - \bar{p}_{2,1}}{p_2^{\max} - v}\left[\rho_1^\star(v) - \rho_1^\star(p_2^{\max})\right]; \tag{23}$$

2) if $v < \widehat{\omega} < p_2^{\max}$, then, by defining $z = \frac{\rho_1^\star(\widehat{\omega})-\rho_1^\star(p_2^{\max})}{p_2^{\max}-\widehat{\omega}} - \frac{\rho_1^\star(v)-\rho_1^\star(\widehat{\omega})}{\widehat{\omega}-v}$ we have

$$W = \begin{cases} -\frac{\rho_1^\star(\widehat{\omega})-\rho_1^\star(p_2^{\max})}{p_2^{\max}-\widehat{\omega}}\bar{p}_{2,1} + \frac{\rho_1^\star(\widehat{\omega})p_2^{\max}-\rho_1^\star(p_2^{\max})\widehat{\omega}}{p_2^{\max}-\widehat{\omega}} & \text{if } z \geq 0, \bar{p}_{2,1} \geq \widehat{\omega} \\ -\frac{\rho_1^\star(v)-\rho_1^\star(\widehat{\omega})}{\widehat{\omega}-v}\bar{p}_{2,1} + \frac{\rho_1^\star(v)\widehat{\omega}-\rho_1^\star(\widehat{\omega})v}{\widehat{\omega}-v} & \text{if } z \geq 0, \bar{p}_{2,1} < \widehat{\omega} \\ -\frac{\rho_1^\star(v)-\rho_1^\star(p_2^{\max})}{p_2^{\max}-v}\bar{p}_{2,1} + \frac{\rho_1^\star(v)p_2^{\max}-\rho_1^\star(p_2^{\max})v}{p_2^{\max}-v} & \text{if } z < 0 \end{cases} \tag{24}$$

Next, we observe that the first term in (22) decreases with $\bar{p}_{2,1}$, since $\rho_2(\cdot)$ is an increasing function. Moreover, also the expressions for W in (23) and (24) decrease as $\bar{p}_{2,1}$ increases. It follows that both terms in (22) decrease with $\bar{p}_{2,1}$; hence, from the definition of $\mathcal{Z}$, the maximizer turns out to be $\bar{p}_{2,1} = v + \left[\frac{\bar{p}_2-v}{q}\right]^+$. The above expression for $\bar{p}_{2,1}$ leads to two cases: for $\bar{p}_{2,1}$:

- when $\bar{p}_2 < v$, we have $\bar{p}_{2,1} = v$. In such a case, the expressions for W in (23) and (24) do not depend on q and, as detailed in App. B, we obtain $R^{\mathrm{LB}} = \rho_2(\bar{p}_2)$ and $g_{\mathrm{LB}}^\star = \delta(p_2 - \bar{p}_2)$.
- the expression $\bar{p}_{2,1} = v + \frac{\bar{p}_2-v}{q}$ holds when $\bar{p}_2 \geq v$. According to the definition of $\mathcal{Z}$, this constraint implies that the range of valid values for q is $\frac{\bar{p}_2-v}{p_2^{\max}-v} \leq q \leq 1$. Substituting $\bar{p}_{2,1} = v + \frac{\bar{p}_2-v}{q}$ into (22), the expression of the lower bound reduces to

$$R^{\mathrm{LB}} = \max_{\frac{\bar{p}_2-v}{p_2^{\max}-v} \leq q \leq 1} (1-q)\rho_2(v) + qW|_{\bar{p}_{2,1}=v+\frac{\bar{p}_2-v}{q}} . \tag{25}$$

Specifically, for $\widehat{\omega} \leq v$ or $\widehat{\omega} \geq p_2^{\max}$, we use the expression of W in (23) and obtain:

$$R^{\mathrm{LB}} = \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}\rho_1^\star(v) + \left(1 - \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}\right)\rho_1^\star(p_2^{\max}) .$$

Instead, for $v < \widehat{\omega} < p_2^{\max}$, we use the expression of W in (24). As for the first term in (24), we observe that $\bar{p}_{2,1} \geq \widehat{\omega}$ and $\bar{p}_{2,1} = v + \frac{\bar{p}_2-v}{q}$ imply $q \leq \frac{\bar{p}_2-v}{\widehat{\omega}-v}$. Therefore, we have

$$R^{\mathrm{LB}} = \begin{cases} \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-\widehat{\omega}}\rho_1^\star(\widehat{\omega}) + \left(1 - \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-\widehat{\omega}}\right)\rho_1^\star(p_2^{\max}) & \bar{p}_2 \geq \widehat{\omega} \\ \frac{\widehat{\omega}-\bar{p}_2}{\widehat{\omega}-v}\rho_1^\star(v) + \left(1 - \frac{\widehat{\omega}-\bar{p}_2}{\widehat{\omega}-v}\right)\rho_1^\star(\widehat{\omega}) & \bar{p}_2 < \widehat{\omega} \end{cases} \tag{26}$$

being $z \geq 0$. As for the second term in (24), we observe that $\bar{p}_{2,1} < \widehat{\omega}$ and $\bar{p}_{2,1} = v + \frac{\bar{p}_2-v}{q}$ imply $q > \frac{\bar{p}_2-v}{\widehat{\omega}-v}$. Thus,

$$\begin{aligned} R^{\mathrm{LB}} &= \max_{\max\left\{\frac{\bar{p}_2-v}{p_2^{\max}-v}, \frac{\bar{p}_2-v}{\widehat{\omega}-v}\right\} \leq q \leq 1} \left[\frac{\rho_1^\star(v)\widehat{\omega} - \rho_1^\star(v)\bar{p}_2 + \rho_1^\star(\widehat{\omega})\bar{p}_2 - \rho_1^\star(\widehat{\omega})v}{\widehat{\omega}-v}\right] \\ &= \frac{\widehat{\omega}-\bar{p}_2}{\widehat{\omega}-v}\rho_1^\star(v) + \left(1 - \frac{\widehat{\omega}-\bar{p}_2}{\widehat{\omega}-v}\right)\rho_1^\star(\widehat{\omega}) \end{aligned} \tag{27}$$

since $\bar{p}_2 \leq \bar{p}_{2,1} < \widehat{\omega}$ implies $\bar{p}_2 < \widehat{\omega}$. As for the third term in (24), we have

$$\begin{aligned} R^{\mathrm{LB}} &= \max_{\frac{\bar{p}_2-v}{p_2^{\max}-v} \leq q \leq 1} \left[\frac{\rho_1^\star(v)(p_2^{\max}-\bar{p}_2) + \rho_1^\star(p_2^{\max})(\bar{p}_2-v)}{p_2^{\max}-v}\right] \\ &= \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}\rho_1^\star(v) + \left(1 - \frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}\right)\rho_1^\star(p_2^{\max}) . \end{aligned} \tag{28}$$

TABLE I

PARAMETERS OF THE MAXIMIZER $g^{\star}_{\mathrm{LB}}(p_2)$ PROVIDING THE LOWER BOUND IN (19). $p_2 = v$ IS SUCH THAT $\rho_1^{\star}(v) = \rho_2(v)$

Condition	u	π_1	π_2
$p_2^{\max} \leq v$	1	$\bar{p}_2$	-
$p_2^{\max} > v$; $\bar{p}_2 < v$	1	$\bar{p}_2$	-
$p_2^{\max} > v$; $\bar{p}_2 \geq v$; $\widehat{\omega} \geq p_2^{\max}$ or $\widehat{\omega} \leq v$	$\frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}$	v	$p_2^{\max}$
$p_2^{\max} > v$; $\bar{p}_2 \geq v$; $v < \widehat{\omega} < p_2^{\max}$; $z < 0$	$\frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-v}$	v	$p_2^{\max}$
$p_2^{\max} > v$; $\bar{p}_2 \geq v$; $v < \widehat{\omega} < p_2^{\max}$; $z \geq 0$; $\bar{p}_2 < \widehat{\omega}$	$\frac{\widehat{\omega}-\bar{p}_2}{\widehat{\omega}-v}$	v	$\widehat{\omega}$
$p_2^{\max} > v$; $\bar{p}_2 \geq v$; $v < \widehat{\omega} < p_2^{\max}$; $z \geq 0$; $\bar{p}_2 \geq \widehat{\omega}$	$\frac{p_2^{\max}-\bar{p}_2}{p_2^{\max}-\widehat{\omega}}$	$\widehat{\omega}$	$p_2^{\max}$
$z = \frac{\rho_1^{\star}(\widehat{\omega})-\rho_1^{\star}(p_2^{\max})}{p_2^{\max}-\widehat{\omega}} - \frac{\rho_1^{\star}(v)-\rho_1^{\star}(\widehat{\omega})}{\widehat{\omega}-v}$			

The above results are summarized in Table I. ■

We remark that, as shown in Sec. VII. the lower bound to the maximum achievable rate that can be attained proves to be very tight, especially for low-medium SNR values.

C. Maximum achievable rate in absence of residual self-interference

The achievable rate in absence of residual self-interference can be obtained by setting $\beta = 0$. In this case, the SINR at $\mathcal{R}$ is a function of the source transmit power only, and is given by $\gamma_1(p_1) = \frac{p_1\alpha_1}{I_1}$. Then, in analogy to (7), the instantaneous rate achieved on the $\mathcal{S}$-$\mathcal{R}$ is given by

$$\rho_1^0(p_1) = \sum_{i=1}^{k} \log\left(\frac{n_1\lambda_{1,i}}{k}(\gamma_1(p_1) + \ell_{1,k})\right) \tag{29}$$

for $k = 1, \ldots, n_1' - 1$ and $c_{1,k} \leq \gamma_1(p_1) < c_{1,k+1}$, and for $k = n_1'$ and $\gamma_1(p_1) \geq c_{1,n_1'}$. Since the expressions of the rates $\rho_1^0(p_1)$ and $\rho_2(p_2)$ in, respectively, (29) and (7) are both concave functions, the maximization of the corresponding rates averaged over a time period can be performed by using [26, Lemma 4.1]. Such a maximization is therefore obtained for a transmit power at $\mathcal{S}$ and $\mathcal{R}$ equal to $\bar{p}_1$ and $\bar{p}_2$, respectively. Then the upper bound to R is given by

$$R^0 = \min\left\{\rho_1^0(\bar{p}_1), \rho_2(\bar{p}_2)\right\}.$$

VI. MASSIVE MIMO SYSTEMS AND HYBRID ASYMPTOTIC RATE APPROXIMATION

In this section, we first tackle mMIMO communications and present an approximated analysis considering that the number of antennas grows large (Sec. VI-A). We then leverage our asymp-

totic analysis to derive a low-complexity rate approximation that holds tight for systems with a limited number of antennas (Sec. VI-B).

A. *Asymptotic analysis of mMIMO systems*

Here we address the case where the number of antennas at $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{D}$ grows large – a relevant scenario in 5G-and-beyond communications. In particular, we consider an asymptotic regime where, as done in the previous sections, we assume a block fading channel with coherence time equal to T. Also, we have that matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are full rank (i.e., $n_1' = n_1$ and $n_2' = n_2$) and n_1, n_2, m_1, m_2 tend to infinity, while the ratios $\zeta_1 = n_1/m_1$, $\zeta_2 = n_2/m_2$, and $\zeta_r = n_2/m_1$ remain constant. Clearly, under such conditions, we need to take the normalized achievable rate as performance metric, i.e., $\frac{R}{n_1}$, as R would diverge. We can write:

$$\begin{aligned} R^{\infty} &\triangleq \lim_{n_1,n_2,m_1,m_2\to\infty} \frac{R}{n_1} \\ &= \lim_{n_1,n_2,m_1,m_2\to\infty} \max_{g(\cdot)} \min\left\{ \int_0^{p_2^{\max}} g(p_2) \frac{\rho_1^{\star}(p_2)}{n_1} \,\mathrm{d}p_2, \int_0^{p_2^{\max}} g(p_2) \frac{\rho_2(p_2)}{n_1} \,\mathrm{d}p_2 \right\} \\ &= \max_{g(\cdot)} \min\left\{ \int_0^{p_2^{\max}} g(p_2) \rho_1^{\star,\infty}(p_2) \,\mathrm{d}p_2, \frac{\zeta_r}{\zeta_1} \int_0^{p_2^{\max}} g(p_2) \rho_2^{\infty}(p_2) \,\mathrm{d}p_2 \right\} \end{aligned} \tag{30}$$

where we defined $\rho_1^{\star,\infty}(p_2) \triangleq \lim_{n_1,m_1\to\infty} \frac{\rho_1^{\star}(p_2)}{n_1}$ and $\rho_2^{\infty}(p_2) \triangleq \lim_{n_2,m_2\to\infty} \frac{\rho_2(p_2)}{n_2}$. In the following, we derive the expressions for $\rho_1^{\star,\infty}(p_2)$ and $\rho_2^{\infty}(p_2)$.

We denote the $\mathcal{S}$-$\mathcal{R}$ and $\mathcal{R}$-$\mathcal{D}$ links with the index $j = 1, 2$. Then consider link j and let $F_j^{(n_j)}(\lambda)$ be the empirical cumulative distribution function (cdf) of the eigenvalues $\lambda_{j,i}$ of $\frac{1}{m_j}\mathbf{H}_j^{\mathsf{H}}\mathbf{H}_j$, $i = 1, \ldots, n_j$. As often assumed in the literature [30], under some conditions on $\mathbf{H}_j$, the empirical cdf $F_j^{(n_j)}(\cdot)$ tends to the asymptotic cdfs $F_j(\lambda)$ as the number of transmit and receive antennas tends to infinity, with constant ratio $\zeta_j = \frac{n_j}{m_j} \leq 1$. Considering $\lambda_{j,i}$ to be ordered in decreasing order, by definition of empirical cdf, we have

$$F_j^{(n_j)}(\lambda_{j,k}) = 1 - \frac{k}{n_j}\,. \tag{31}$$

In such asymptotic regime, and for $k \to \infty$ with constant ratio $y = k/n_j$, the empirical cdf in (31) tends to

$$\lim_{k,n_j\to\infty} F_j^{(n_j)}(\lambda_{j,k}) = F_j(\lambda) = 1 - y\,. \tag{32}$$

Next, we denote with $f_j(\cdot)$ the corresponding pdf, with support $[a_j, b_j]$. In the above asymptotic regime, we can also write

$$\lim_{n_j,m_j,k\to\infty} \ell_{j,k} = \lim_{n_j,m_j,k\to\infty} \frac{1}{n_j} \sum_{i=1}^{k} \frac{1}{\lambda_{j,k}} = \int_{\lambda}^{b_j} \frac{1}{z} f_j(z) \,\mathrm{d}z \triangleq \ell_j^{\infty}(\lambda) \tag{33}$$

where, $\ell_{j,k}$ is as defined in (6) and (7), for $j = 1, 2$, and according to (32), $\lambda = F_j^{-1}(1-y)$.

Let us now focus on the $\mathcal{S}$-$\mathcal{R}$ link. The asymptotic expression of the optimized transmit power in (13) can be obtained by observing that, as the number of transmit antennas n_1 grows, the difference $\lambda_{1,k} - \lambda_{1,k+1}$ between any two adjacent ordered eigenvalues tends to 0, and so does the difference $\omega_k - \omega_{k+1}$, since it is proportional to $\lambda_{1,k} - \lambda_{1,k+1}$. Let λ_1 be the generic eigenvalue related to the $\mathcal{S}$-$\mathcal{R}$ link and $\omega = \frac{\psi\lambda_1\alpha_1}{\zeta_1\beta}$. According to (13), when the source power is smaller than $p_1^{\max}$, we have $\lim_{k,n_1\to\infty} \frac{\psi k}{n_1} - \frac{I_1+\beta p_2}{\alpha_1}\zeta_1\ell_{1,k} = \psi y - \frac{I_1+\beta p_2}{\alpha_1}\zeta_1\ell_1^{\infty}(\lambda_1)$ whenever $p_2 = \omega$, i.e., $\lambda_1 = \frac{I_1+\beta p_2}{\psi\alpha_1}\zeta_1$. Note that, to obtain the above expression, we exploited (31), (32), and (33). By recalling that $y = 1 - F_j(\lambda)$, we get

$$\psi y - \frac{I_1+\beta p_2}{\alpha_1}\zeta_1\ell_1^{\infty}(\lambda) \;\; = \;\; \psi\left[1 - F_1(\lambda_1) - \lambda_1\ell_1^{\infty}(\lambda_1)\right]. \tag{34}$$

In conclusion, the optimized source transmit power $p_1^{\star}(p_2)$ in (13) tends to the asymptotic limit

$$p_1^{\star,\infty}(p_2) \triangleq \min\left\{p_1^{\max}, \psi\left[1 - F_1(\lambda_1) - \lambda_1\ell_1^{\infty}(\lambda_1)\right]\right\} \tag{35}$$

where we recall that ψ is such that $\int_0^{p_2^{\max}} p_1^{\star,\infty}(p_2)\,\mathrm{d}p_2 = \bar{p}_1$. Similarly, the SINR on the $\mathcal{S}$-$\mathcal{R}$ link in (14) tends asymptotically to $\gamma_1^{\star,\infty}(p_2) = \min\left\{\frac{\zeta_1 p_1^{\max}}{\lambda_1\psi}, \zeta_1\left[\frac{1-F_1(\lambda_1)}{\lambda_1} - \ell_1^{\infty}(\lambda_1)\right]\right\}$. When $k < n_1$ and $c_{1,k} \le \gamma_1^{\star}(p_2) < c_{1,k+1}$, we replace the above expression in (6) and obtain

$$\begin{aligned}
\rho_1^{\star,\infty}(p_2) &= \lim_{n_1,m_1\to\infty}\frac{\rho_1^{\star}(p_2)}{n_1} \\
&= \lim_{n_1,m_1,k\to\infty}\frac{1}{n_1}\sum_{i=1}^{k}\log\left[\frac{n_1\lambda_{1,i}}{k}\left(\frac{\gamma_1^{\star}(p_2)}{\zeta_1} + \ell_{1,k}\right)\right] \\
&= \int_{\lambda_1}^{b_1}\log\left[\frac{z}{y}\left(\frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1} + \ell_1^{\infty}(\lambda_1)\right)\right] f_1(z)\,\mathrm{d}z\,.
\end{aligned} \tag{36}$$

Recall that $y = k/n_1$ and the eigenvalues are ordered in decreasing order, thus the k-th eigenvalue in a finite system corresponds to $z = \lambda_1$ in the asymptotic regime. Since asymptotically $c_{1,k} - c_{1,k+1} \to 0$, given $c_{1,k+1} \le \gamma_1^{\star}(p_2) \le c_{1,k}$, the above result holds if $\lim_{n_1,m_1,k\to\infty}\gamma_1^{\star}(p_2) = \lim_{n_1,m_1,k\to\infty} c_{1,k}$, i.e.,

$$\gamma_1^{\star,\infty}(p_2) = \zeta_1\left(\frac{y}{\lambda_1} - \ell_1^{\infty}(\lambda_1)\right). \tag{37}$$

By substituting (37) in (36), we obtain $\rho_1^{\star,\infty}(p_2) = J_1(\lambda_1) \quad \text{for} \quad \gamma_1^{\star,\infty}(p_2) = \zeta_1\left(\frac{1-F_1(\lambda_1)}{\lambda_1} - \ell_1^{\infty}(\lambda_1)\right)$ where we defined $J_1(\lambda_1) \triangleq \int_{\lambda_1}^{b_1} f_1(z)\log\frac{z}{\lambda_1}\,\mathrm{d}z$. When $k = n_1$, we observe that asymptoti-

cally $\lim_{n_1,m_1\to\infty}\lambda_{1,n_1}=a_1$. Therefore, the condition $\gamma_1^\star(p_2)\geq c_{1,n_1}$ becomes $\gamma_1^{\star,\infty}(p_2)>\zeta_1\left(\frac{1}{a_1}-\ell_1^\infty(a_1)\right)$. Hence, we have

$$\begin{aligned}\rho_1^{\star,\infty}(p_2) &= \lim_{n_1,m_1\to\infty}\frac{1}{n_1}\sum_{i=1}^{n_1}\log\left[\lambda_{1,i}\left(\frac{\gamma_1^\star(p_2)}{\zeta_1}+\ell_{1,n_1}\right)\right]\\ &= \int_{a_1}^{b_1}\log\left[z\left(\frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1}+\ell_1^\infty(a_1)\right)\right]f_1(z)\,\mathrm{d}z\\ &= J_1(a_1)+\log\left[a_1\left(\frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1}+\ell_1^\infty(a_1)\right)\right]. \end{aligned}\tag{38}$$

In conclusion, by defining $Q_j(s)\triangleq\int_s^{b_j}f_j(z)\log z\,\mathrm{d}z$, we have

$$\rho_1^{\star,\infty}(p_2)=\begin{cases} Q_1(\lambda_1)-\log\lambda_1 & \text{for } \frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1}=\frac{1-F_1(\lambda_1)}{\lambda_1}-\ell_1^\infty(\lambda_1)\\ Q_1(a_1)+\log\left(\frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1}+\ell_1^\infty(a_1)\right) & \text{for } \frac{\gamma_1^{\star,\infty}(p_2)}{\zeta_1}>\frac{1}{a_1}-\ell_1^\infty(a_1)\,.\end{cases}\tag{39}$$

Similarly, if λ_2 is the generic eigenvalue of $\frac{1}{m_2}\mathbf{H}_2^{\mathsf{H}}\mathbf{H}_2$, the asymptotic normalized rate $\rho_2^\infty(p_2)\triangleq\lim_{n_2,m_2\to\infty}\frac{\rho_2(p_2)}{n_2}$ for the $\mathcal{R}$-$\mathcal{D}$ link can be written as

$$\rho_2^{\infty}(p_2)=\begin{cases} Q_2(\lambda_2)-\log\lambda_2 & \text{for } \frac{\gamma_2(p_2)}{\zeta_2}=\frac{1-F_2(\lambda_2)}{\lambda_2}-\ell_2^\infty(\lambda_2)\\ Q_2(a_2)+\log\left(\frac{\gamma_2(p_2)}{\zeta_2}+\ell_2^\infty(a_2)\right) & \text{for } \frac{\gamma_2(p)}{\zeta_2}>\frac{1}{a_2}-\ell_2^\infty(a_2)\,.\end{cases}$$

Next, we consider the case of high practical relevance where the channel is affected by Rayleigh fading. In the latter case, we also consider that the rank of the matrix $\mathbb{E}[\mathbf{H}_j^{\mathsf{H}}]\mathbb{E}[\mathbf{H}_j]$ is constant and independent of n_j. Under these conditions, the asymptotic distribution, $f_j(\lambda)$, of the eigenvalues of the channel matrices $\frac{1}{m_j}\mathbf{H}_j^{\mathsf{H}}\mathbf{H}_j$ follows the Marčenko-Pastur law [30] $f_j(z)=\frac{1}{2\pi\zeta_j z}\sqrt{(z-a_j)(b_j-z)}$ with support $[a_j,b_j]$, and where $a_j=(1-\sqrt{\zeta_j})^2$, $b_j=(1+\sqrt{\zeta_j})^2$, and $0<\zeta_j\leq 1$. The corresponding cdf is

$$F_j(z)=\frac{1}{2}+zf_j(z)+\frac{a_j+b_j}{4\pi\zeta_j}\arcsin\left(\frac{2z-a_j-b_j}{b_j-a_j}\right)-\frac{\sqrt{a_jb_j}}{2\pi\zeta_j}\arcsin\left(\frac{(a_j+b_j)z-2a_jb_j}{z(b_j-a_j)}\right).$$

Moreover,

$$\ell_j^\infty(z)=\begin{cases}\frac{1}{2(1-\zeta_j)}+f_j(z)-\frac{1+\zeta_j}{2\pi\zeta_j(1-\zeta_j)}\arcsin\left(\frac{(a_j+b_j)z-2a_jb_j}{z(b_j-a_j)}\right)+\frac{1}{2\pi\zeta_j}\arcsin\left(\frac{2z-a_j-b_j}{b_j-a_j}\right) & \text{if } \zeta_j<1\\ \frac{1}{\pi}\sqrt{\frac{4-z}{z}}+\frac{1}{2\pi}\arcsin\left(\frac{z}{2}-1\right)-\frac{1}{4} & \text{if } \zeta_j=1\,.\end{cases}$$

Equipped with the above expressions, we can approximate the rate R^∞ by discretizing the distribution $g(\cdot)$, as done in Sec. V-A. From (30), we obtain:

$$\textbf{P5:}\quad \widehat{R}^\infty=\max_{\mathbf{g}}\min\left\{\mathbf{g}^{\mathsf{T}}\boldsymbol{\rho}_1^{\star,\infty},\frac{\zeta_r}{\zeta_1}\mathbf{g}^{\mathsf{T}}\boldsymbol{\rho}_2^{\infty}\right\}\qquad\text{s.t.}$$

$$(a)\ \mathbf{g}^{\mathsf{T}}\mathbf{p}_1^{\star,\infty}=\bar{p}_1\,,\quad (b)\ \mathbf{g}^{\mathsf{T}}\boldsymbol{\pi}=\bar{p}_2\,,\quad (c)\ \mathbf{g}\geq\mathbf{0}\,,\quad (d)\ \mathbf{g}^{\mathsf{T}}\mathbf{1}=1\tag{40}$$

where we recall that $g(p_2) = \sum_{n=1}^{N_p} g_n \delta(p_2 - \pi_n)$, $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_{N_p}]^{\mathsf{T}}$, $\mathbf{g} = [g_1, \ldots, g_{N_p}]^{\mathsf{T}}$, and we defined $\mathbf{p}_1^{\star,\infty} = [p_1^{\star,\infty}(\pi_1), \ldots, p_1^{\star,\infty}(\pi_{N_p})]^{\mathsf{T}}$, $\boldsymbol{\rho}_1^{\star,\infty} = [\rho_1^{\star,\infty}(\pi_1), \ldots, \rho_1^{\star,\infty}(\pi_{N_p})]^{\mathsf{T}}$ and $\boldsymbol{\rho}_2^{\infty} = [\rho_2^{\infty}(\pi_1), \ldots, \rho_2^{\infty}(\pi_{N_p})]^{\mathsf{T}}$. Again, when the vector $\boldsymbol{\pi}$ is suitably chosen, the normalized rate $\widehat{R}^{\infty}$ is an excellent approximation of R^{∞}, as also discussed in Sec. VII.

B. *Hybrid asymptotic approximation for MIMO systems*

We now present a hybrid methodology to approximate the rate R of a MIMO system with an arbitrary (limited) number of antennas, leveraging $g^{\infty}(p_2)$ and $p_1^{\star,\infty}(p_2)$ as sub-optimal choices for computing the rate in (11). The advantage of such an approach is that $g^{\infty}(p_2)$ does not depend on a specific channel realization, rather on the distribution of the channel gains. Thus, given the system parameters and the channel distribution, $g^{\infty}(p_2)$ has to be computed only once, as opposed to (11) that needs to be recomputed as frequently as once per channel coherence time. Since the relay does not need to send control information to the source any longer, our hybrid methodology eliminates the communication overhead; furthermore, it is robust to imperfect CSI, provided that the estimated channel matrices are still Rayleigh distributed.

Let $g^{\infty}(p_2) = \sum_{n=1}^{N_p} g_n^{\infty} \delta(p_2 - \pi_n)$ and $p_1^{\star,\infty}(p_2) = \sum_{n=1}^{N_p} p_1^{\star,\infty}(\pi_n) \delta(p_2 - \pi_n)$ be, respectively, the distribution of the relay transmit power and the source transmit power maximizing the asymptotic rate in (40), when the vector of relay powers, $\boldsymbol{\pi}$, is provided. By replacing in (11) $g(p_2)$ and $p_1(p_2)$ with $g^{\infty}(p_2)$ and $p_1^{\star,\infty}(p_2)$, respectively, we have:

$$\begin{aligned} \textbf{P6:} \quad R^{\mathrm{H}} &= \min\left\{R_1(g^{\infty}, p_1^{\star,\infty}), R_2(g^{\infty})\right\} \\ &= \min\left\{\boldsymbol{\rho}_1^{\mathsf{T}}\mathbf{g}^{\infty}, \boldsymbol{\rho}_2^{\mathsf{T}}\mathbf{g}^{\infty}\right\} \end{aligned} \tag{41}$$

where $\boldsymbol{\rho}_1 = [\rho_1(p_1^{\star,\infty}(\pi_1), \pi_1), \ldots, \rho_1(p_1^{\star,\infty}(\pi_{N_p}), \pi_{N_p})]^{\mathsf{T}}$ and $\boldsymbol{\rho}_2 = [\rho_2(\pi_1), \ldots, \rho_2(\pi_{N_p})]^{\mathsf{T}}$, and the functions $\rho_1(p_1, p_2)$ and $\rho_2(p_2)$ are defined as in (6) and (7), respectively. We remark that by replacing $g(p_2)$ and $p_1(p_2)$ with, respectively, $g^{\infty}(p_2)$ and $p_1^{\star,\infty}(p_2)$, the constraints in (9) and (12) are automatically satisfied.

Importantly, in Sec. VII we show that the adoption in **P2** of the above suboptimal choices for $g(p_2)$ and $p_1(p_2)$ (e.g., $g^{\infty}(p_2)$ and $p_1^{\star,\infty}(p_2)$) yields a very limited degradation of the achievable rate.

VII. Numerical results

To evaluate the performance of our proposed solution, we consider a scenario similar to that employed in [15], [26] where the distance between the source and the relay and between the

relay and the destination is set to $d = 300\,\text{m}$. We assume that communications take place using carrier frequency $f_c = 2.6\,\text{GHz}$ and signal bandwidth $B = 200\,\text{kHz}$. The path loss for both links is given by $\alpha_j = \left(\frac{c}{4\pi f_c}\right)^2 d^{-p_e}$ $j = 1, 2$, where $p_e = 3$ is the path loss exponent. The additive noise at both relay and destination has power spectral density $N_0 = -174$ dBm/Hz so that the noise terms are given by $I_1 = I_2 = N_0 B = -121\,\text{dBm}$. Furthermore, we assume that $\mathcal{S}$-$\mathcal{R}$ and $\mathcal{R}$-$\mathcal{D}$ links are affected by Rayleigh fading and that the entries of the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are random independent variables, having circularly-symmetric complex Gaussian distribution with zero-mean and unit-variance. Since such matrices have full rank with probability 1, in the following we assume $n_1' = n_1$ and $n_2' = n_2$.

We first evaluate the performance obtained through the approximate solution in (18) and we discuss its accuracy. To this end, in Fig. 4(left) we show the rate $\widehat{R}$, computed as in (18) and obtained for a single realization of the channel matrices $\mathbf{H}_1$ and $\mathbf{H}_2$, for $n_1 = m_1 = n_2 = m_2 = 4$, $p_2^{\max} = 23\,\text{dBm}$, $\bar{p}_2 = 20\,\text{dBm}$, $p_1^{\max} = \bar{p}_1 + 3\,\text{dB}$, and $\beta = -125\,\text{dB}$. Each curve refers to a different number of discretization points, N_p, which is also the size of vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_{N_p}]^{\mathsf{T}}$. The elements of $\boldsymbol{\pi}$ are selected as described in Sec. V-A. Looking at Fig. 4(left), we observe that:

- the rate $\widehat{R}$ increases with N_p. Values of N_p larger than 200 do not provide further gain;
- the solution of problem **P4** in (18) provides up to three positive weights, g_n, regardless the value of N_p;
- for $\bar{p}_1 < 20\,\text{dB}$ and $\bar{p}_1 > 38\,\text{dB}$, the curves do not depend on N_p, while they slightly differ in the range $20\,\text{dB} < \bar{p}_1 < 38\,\text{dB}$. Indeed, we have observed that, for low values of $\bar{p}_1$, (18) provides two positive weights g_n, namely for $p_2 = 0$ and $p_2 = p_2^{\max}$ while, for $\bar{p}_1 > 38\,\text{dB}$, the only positive weight is for $p_2 = \bar{p}_2$. As for $20\,\text{dB} < \bar{p}_1 < 38\,\text{dB}$, the curves converge very fast as N_p grows. We then consider the rate obtained for $N_p = 200$ as an almost perfect approximation of R.

Let $\widehat{R}_{N_p}$ be the rate obtained by solving (18) using as input a vector $\boldsymbol{\pi}$ composed of N_p elements, i.e., $N_p - 2$ logarithmically spaced elements plus the elements $\{0, \bar{p}_2\}$, as described in Sec. V-A where $\epsilon = 10^{-10}$. In Fig. 4(right), to quantify the effect of N_p on $\widehat{R}$, we show the relative rate loss $\widehat{R}^{\text{loss}} = |1 - \frac{\widehat{R}_{N_p}}{\widehat{R}_{200}}|$, averaged over 100 realizations of the channel matrices. As can be observed, $N_p = 5$ provides less than 5% rate loss, while for $N_p = 100$ the loss becomes negligible (lower than 10^{-4}).

In the following, we will use $N_p = 50$ as a fair trade-off between accuracy of the rate $\widehat{R}$ and

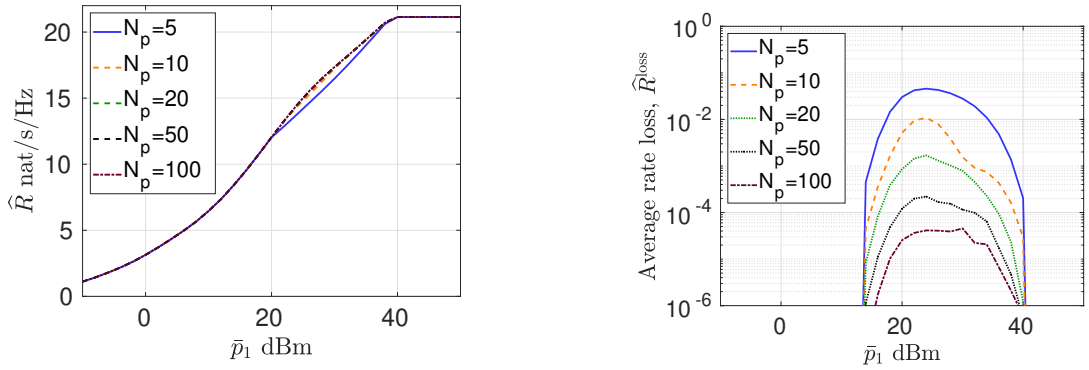

Fig. 4. Left: achievable rate as the number of discretization points N_p varies, for $n_1 = m_1 = n_2 = m_2 = 4$, $p_2^{\max} = 23$ dBm, $\bar{p}_2 = 20$ dBm, and $p_1^{\max} = \bar{p}_1 + 3$ dB. Right: rate loss as the number of discretization points N_p varies, for a given channel realization, $n_1 = m_1 = n_2 = m_2 = 4$, $p_2^{\max} = 23$ dBm, $\bar{p}_2 = 20$ dBm, and $p_1^{\max} = \bar{p}_1 + 3$ dB.

the complexity of problem **P4**, as well as for solving **P5** and computing the asymptotic rate. In particular, for the same system parameters considered above, Fig. 5 depicts the distribution of the achievable rate, $\widehat{R}$, versus the average transmit power at the source, $\bar{p}_1$, when the number of transmitting or receiving antennas at all devices is set to n (i.e., $n_1 = m_1 = n_2 = m_2 = n$). More specifically, we consider $n = 2$ (left plot), $n = 4$ (center plot), and $n = 16$ (right plot). For each value of $\bar{p}_1$, the lines limiting the shaded regions represent the percentiles $\{10, 20, \ldots, 90\}$ of the distribution. The solid line represents the median while the dashed line denotes the asymptotic rate $n\widehat{R}^{\infty}$. It can be observed that, as n increases, the distribution of $\widehat{R}$ rapidly converges to the asymptotic rate. Although the convergence is faster for lower values of $\bar{p}_1$, for $n = 16$ the 10-th and 90-th percentiles tightly match the asymptotic rate for all considered values of $\bar{p}_1$.

We remark that, since the discretization points we used are distributed over the entire support of the objective function in problem **P2** and the involved functions are $\mathcal{C}^{\infty}$ (except for a discrete set of points), rate $\widehat{R}$ converges to R as N_p grows. For the same reason, rate $\widehat{R}^{\infty}$ converges to R^{∞}. Moreover, as the number of antennas tends to infinity, the normalized rate $\frac{R}{n} \to R^{\infty}$ because of the convergence of the empirical eigenvalue distribution $F_j^{(n_j)}(\lambda_{j,k})$ to the asymptotic distribution $F_j(\lambda)$. We then conclude that, for sufficiently large values of N_p and n, rate $n\widehat{R}^{\infty}$ is an excellent approximation for R.

We now evaluate the performance of the hybrid technique outlined in Sec. VI-B. Fig. 6(left) compares the rate $\widehat{R}$ obtained by solving **P4**, against the rate R^{H}, obtained by solving **P5**. Both

rates are averaged over 100 realizations of the channel matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. We observe that $R^{\mathrm{H}} \leq \widehat{R}$ as expected, however the rate loss is very limited, especially when the number of antennas, n, is greater or equal to 4. Also, the figure shows that, for small and large values of $\bar{p}_2$, the two rates coincide. This is due to the fact that in such cases **P4** and **P5** provide, as output, the same maximizing distribution $g(\cdot)$.

In Fig. 6(right), we compare the rate achieved by the proposed techniques for $n_1 = m_1 = n_2 = m_2 = 4$. Specifically, *(i)* the solid line represents the rate $\widehat{R}$, obtained by solving **P4**; *(ii)* the dashed line denotes the lower-bound R^{LB} defined in (19); *(iii)* the dotted line refers to the rate, R^{H}, achieved by the hybrid technique proposed in Sec. VI-B; *(iv)* the dash-dotted line indicates the rate, R^{HD}, achieved by the system when the relay is constrained to work in the conventional HD mode, for which the rate is given by

$$R^{\mathrm{HD}} = \max_{\tau} \min \left\{ (1-\sigma)\rho_1 \left(\frac{\bar{p}_2}{1-\sigma}, 0 \right), \sigma \rho_2 \left(\frac{\bar{p}_1}{\tau} \right) \right\}$$

subject to the average power constraints in (12) and where $\rho_1(p_1, p_2)$ and $\rho_2(p_2)$ are defined in (6) and (7), respectively (this scheme implies that the time period is divided in two phases only, of duration σT and $(1-\sigma)T$, respectively); (v) the line with markers represents the rate, R^{FD} obtained when the relay always works in FD mode, and $p_2 = \bar{p}_2$ and $p_1 = \bar{p}_1$.

Fig. 6(top right) presents the achievable rate versus $\bar{p}_1$ for $\bar{p}_2 = \bar{p}_1$ and $p_1^{\max} = p_2^{\max} = \bar{p}_1 + 10\,\mathrm{dB}$. We observe that, again, the hybrid technique closely matches $\widehat{R}$, showing a performance loss lower than 1 dB. The bound R^{LB} is very tight for low values of transmit powers, whereas a relay exclusively working in HD mode, reduces the system performance of about 4 dB. The system working exclusively in FD mode shows instead a horizontal asymptote for high values of $\bar{p}_1$ and $\bar{p}_2$. Indeed, given the expressions of the SINR on the two links (see Sec. III-A), it is clear that, given $\bar{p}_1 = \bar{p}_2$, the SINR on the $\mathcal{S}$-$\mathcal{R}$ link becomes the network bottleneck as $\bar{p}_2$ increases.

Finally, in Fig. 6(bottom right) we set $\bar{p}_1 = 20\,\mathrm{dBm}$ and $p_1^{\max} = 23\,\mathrm{dBm}$, and we show the achievable rates versus the average SNR on the $\mathcal{R}$-$\mathcal{D}$ link, $\bar{\gamma}_2 = \bar{p}_2 \alpha_2 / I_2$. Also in this case, the hybrid technique exhibits a limited performance loss in the medium-high SNR region, whereas the lower bound is extremely tight for all low-medium values of SNR. Importantly, the results in Fig. 6 demonstrate that the hybrid technique is very appealing in relay communications even when the transceivers are equipped with a small number of antennas. Indeed, it requires essentially no communication overhead, at the price of a very small degradation in the achievable

communication rate. When instead the system works in FD mode all the time, we observe a severe drop in performance for high values of $\bar{\gamma}_2$: as $\bar{\gamma}_2$ grows while $\bar{p}_1$ remains constant, the SINR, hence the throughput, on the $\mathcal{S}$-$\mathcal{R}$ link tends to zero.

Finally, let us discuss the resulting operational modes adopted at the relay. Let t_{HD} be the fraction of time the relay works in HD mode. Referring to Fig. 6(bottom right) and to the curve obtained by solving problem **P4** (i.e., the rate $\widehat{R}$), we observed that $t_{\mathrm{HD}} = 0$ for $\bar{\gamma}_2 \leq 20\,\mathrm{dB}$, $t_{\mathrm{HD}} = 0.39$ for $\bar{\gamma}_2 = 30\,\mathrm{dB}$, and $t_{\mathrm{HD}} = 0.5$ for $\bar{\gamma}_2 \geq 40\,\mathrm{dB}$. As expected, t_{HD} increases with the relay average transmit power $\bar{p}_2$, which is proportional to $\bar{\gamma}_2$. Indeed, as $\bar{p}_2$ increases, the best option for the source is to remain silent for a larger fraction of time (HD-TX mode), according to the optimal power profile in (13). The capability of the relay to operate in FD mode is instead fully exploited in the low-medium SNR regime on the $\mathcal{R}$-$\mathcal{D}$ link.

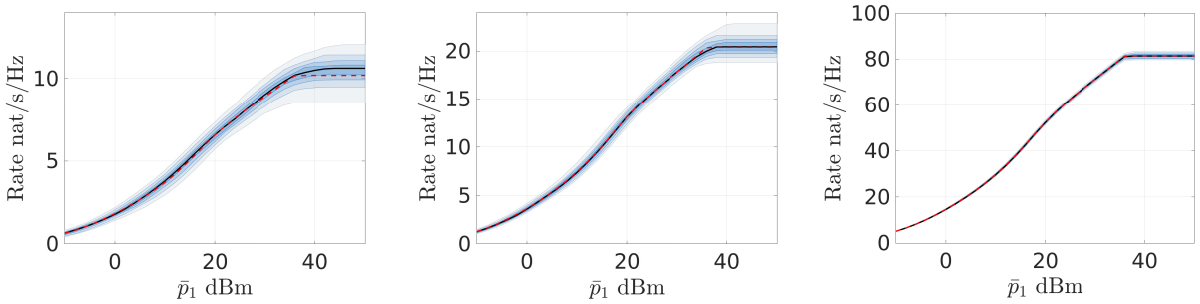

Fig. 5. Distribution of the achievable rate $\widehat{R}$, for $n_1 = m_1 = n_2 = m_2 = 2$ (left), $n_1 = m_1 = n_2 = m_2 = 4$ (center), and $n_1 = m_1 = n_2 = m_2 = 16$ (right) $p_2^{\max} = 23\,\mathrm{dBm}$, $\bar{p}_2 = 20\,\mathrm{dBm}$, and $p_1^{\max} = \bar{p}_1 + 3\,\mathrm{dB}$. The percentiles $\{10, 20, \ldots, 90\}$ are denoted by the lines limiting the shaded regions, while the asymptotic rate is represented by the red dashed line.

VIII. Conclusions

We modeled a two-hop relay network where the relay can work in FD or HD mode and all nodes are equipped with multiple antennas. Considering practical operational conditions, we analysed the achievable rate performance and determined the optimal operational mode at the relay as well as the transmit power level at both the relay and the source nodes. To do this, we first selected the transmit power at the relay as the main decision variable and, conditioned to the value of such variable, we derived a closed-form expression for the optimal transmit power at the source. We then leveraged the latter expression to get an analytical formulation of the overall problem. Given its complexity, we presented an approximate solution as well as an analytical

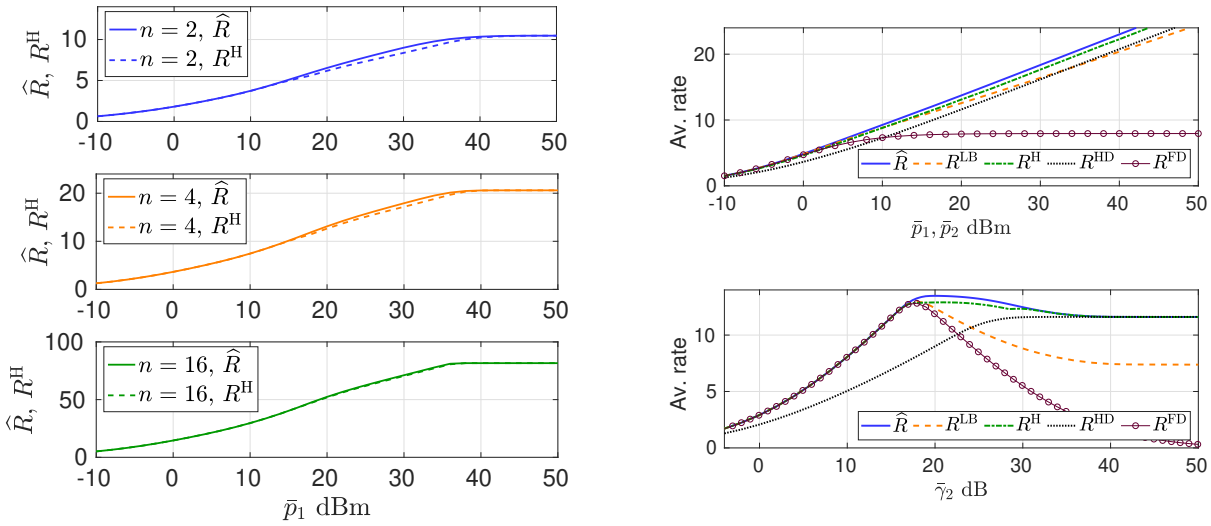

Fig. 6. Left: average achievable rate $\widehat{R}$ and R^{H} for different values of the the number of antennas, as $\bar{p}_1$ varies and for $p_2^{\max} = 23\,\mathrm{dBm}$, $\bar{p}_2 = 20\,\mathrm{dBm}$, and $p_1^{\max} = \bar{p}_1 + 3\,\mathrm{dB}$. Right-top: average achievable rate plotted versus $\bar{p}_1 = \bar{p}_2$, for $n_1 = m_1 = n_2 = m_2 = 4$, $p_1^{\max} = p_2^{\max} = \bar{p}_2 + 10\,\mathrm{dB}$. Right-bottom: average achievable rate plotted versus the average SNR $\bar{\gamma}_2$, for $n_1 = m_1 = n_2 = m_2 = 4$, $\bar{p}_1 = 20\,\mathrm{dB}$, $p_1^{\max} = 23\,\mathrm{dB}$, and $p_2^{\max} = \bar{p}_2 + 3\,\mathrm{dB}$.

expression for the lower bound to the source-destination achievable rate and for the performance in the absence of residual self-interference.

We then addressed an mMIMO scenario and, through an asymptotic analysis, we obtained a semi-analytical, low-complexity method to determine the optimal transmit power at both the source and the relay, and the optimal operational mode at the latter node. Remarkably, this expression is independent of the instantaneous channel conditions and can be conveniently used to obtain the optimal system settings, even when the nodes are equipped with a small number of antennas.

APPENDIX A
PROOF OF PROPOSITION 4.1

We are interested in finding the maximizer, $p_1^\star(p_2)$, of the objective function in **P1**. To this end, we write the Lagrangian of the problem $\mathcal{L}(p_1, p_2) = -g(p_2)\rho_1(p_1, p_2) + \psi' g(p_2) p_1$ where $\psi' \geq 0$ is the Lagrange multiplier. Then, when $P(p) \neq p_1^{\max}$ and $p_1(p_2) \neq 0$, by imposing $\frac{\partial \mathcal{L}(p_1,p_2)}{\partial p_1} = 0$, we obtain $\psi' = \frac{\partial \rho_1(p_1,p_2)}{\partial p_1}$. We recall that the SINR, $\gamma_1(p_1, p_2)$, on the source-relay link, takes the form $\gamma_1(p_2) = \frac{p_1 \alpha_1}{I_1 + \beta p_2}$ and the rate $\rho_1(p_1, p_2)$ in (6) can be rewritten as a piecewise function,

continuous in p_1

$$\rho_1(p_1,p_2)=\sum_{i=1}^{k}\log\left[\frac{\lambda_{1,i}n_1}{k}\left(\frac{P\alpha_1/\zeta_1}{I_1+\beta p}+\ell_{1,k}\right)\right] \tag{42}$$

for $k=1,\ldots,n_1'-1$ and $L_k(p_2)\leq p_1\leq L_{k+1}(p_2)$, and for $k=n_1'$ and $p_1\geq L_{n_1}(p)$, where $L_k(p_2)=\frac{c_{1,k}}{\alpha_1}(I_1+\beta p_2)$. Since $\frac{\partial\rho_1(p_1,p_2)}{\partial p_1}$ is continuous in p_1 and monotonically decreasing for $p_1>0$, the equation $\psi'=\frac{\partial\rho_1(p_1,p_2)}{\partial p_1}$ has a unique solution. As ψ' can be arbitrarily defined, let us set $\psi'=\frac{n_1}{\psi}$ where ψ is a new Lagrange multiplier. Then, by using (42), we first obtain $\psi=\frac{p_1+\frac{I_1+\beta p_2}{\alpha_1}\zeta_1\ell_{1,k}}{k/n_1}$ for $k=1,\ldots,n_1'-1$ and $L_k(p_2)\leq p_1<L_{k+1}(p_2)$, and for $k=n_1'$ and $p_1\geq L_{n_1'}(p_2)$. The solution of the above equation is $\widetilde{p_1}(p_2)=\frac{\psi k}{n_1}-\frac{I_1+\beta p_2}{\alpha_1}\zeta_1\ell_{1,k}$ for $k=1,\ldots,n_1'-1$ and $L_k(p_2)\leq\widetilde{p_1}(p_2)<L_{k+1}(p_2)$, and for $k=n_1'$ and $\widetilde{p_1}(p_2)\geq L_{n_1'}(p_2)$. The above expression can be simplified by observing that the intersection of $\widetilde{p_1}(p_2)$ with $L_k(p_2)$ is unique for every k. By solving $\widetilde{p_1}(p_2)=L_k(p_2)$, we get $p_2=\omega_k\triangleq\frac{\psi\lambda_{1,k}\alpha_1}{\zeta_1\beta}-\frac{I_1}{\beta}$. Thus,

$$\widetilde{p_1}(p_2)=\begin{cases}\frac{\psi k}{n_1}-\frac{I_1+\beta p}{\alpha_1}\zeta_1\ell_{1,k} & \text{if } \omega_{k+1}\leq p_2<\omega_k,\, k=1,\ldots,n_1'-1 \text{ and } p_2<\omega_{n_1'},\, k=n_1'\\ 0 & \text{if } p_2\geq\omega_1\,.\end{cases} \tag{43}$$

Note that the last line in (43) can be derived considering that $\widetilde{p_1}(p_2)$ is a decreasing function of p_2 with $p_2\in[0,p_2^{\max}]$. Since $\widetilde{p_1}(p_2)=0$ for $p_2=\omega_1$, the optimal source transmit power is zero also for any $p_2>\omega_1$. Then, we consider the constraint on the maximum transmit power at $\mathcal{S}$ and write $p_1^\star(p_2)=\min\{\widetilde{p_1}(p_2),p_1^{\max}\}$. Finally, we solve the equation $\widetilde{p_1}(p_2)=p_1^{\max}$ and obtain $p_2=\widehat{\omega}=\alpha_1\frac{\widehat{k}\psi/n_1-p_1^{\max}}{\zeta_1\ell_{1,\widehat{k}}\beta}-\frac{I_1}{\beta}$ where $\widehat{k}=\arg\max_{k=1,\ldots,n_1'}\frac{k\psi/n_1-p_1^{\max}}{\ell_{1,k}}$ represents the index of the piece of $\widetilde{p_1}(p_2)$ in which $\widehat{\omega}$ lies. We can then write the optimal transmit power at $\mathcal{S}$, $p_1^\star(p_2)$, as in (13).

APPENDIX B
DERIVATION OF THE LOWER BOUND FOR $\bar{p}_{2,1}=v$

The solution $\bar{p}_{2,1}=v$ holds when $\bar{p}_2<v$. However, since $\bar{p}_{2,1}$ is the average of $g_1(p_2)$, we can write $\bar{p}_{2,1}=\int_v^{p_2^{\max}}p_2g_1(p_2)\,\mathrm{d}p_2=v$, which implies $g_1(p_2)=\delta(p_2-v)$. Moreover, according to (22), the condition $\bar{p}_2<v$ implies that the range of valid values for q reduces to $0\leq q\leq\bar{p}_2/v$, since in this case $\frac{\bar{p}_2}{v}<1$ and $\frac{\bar{p}_2-v}{p_2^{\max}-v}<0$. Summarizing, under the conditions $p_2^{\max}>v$ and $\bar{p}_2<v$, the lower-bound in (22) becomes

$$R^{\mathrm{LB}}=\max_{0\leq q\leq\bar{p}_2/v}(1-q)\rho_2\left(\frac{\bar{p}_2-qv}{1-q}\right)+q\rho_1^\star(v)=\rho_2(\bar{p}_2)$$

In (44), we first used $g_1(p_2) = \delta(p_2 - v)$, then we exploited the relation $\rho_1^\star(v) = \rho_2(v)$, and, finally we observed that $h(q) = (1-q)\rho_2\left(\frac{\bar{p}_2 - qv}{1-q}\right) + q\rho_2(v)$ decreases with q since, $h(0) \geq h(\bar{p}_2/v)$, $h''(q) < 0$, and $h'(0) \leq 0$. Then, looking at (19), we note that the expression coincides with (44) for $g_{\mathrm{LB}}^\star(p_2) = \delta(p - \bar{p}_2)$, which thus results to be the maximizer.

APPENDIX C
PROOF OF (24)

We are interested in computing W. If $\widehat{\omega} \leq v$ or $\widehat{\omega} \geq p_2^{\max}$, then the function $\rho_1^\star(p_2)$ is convex for $p_2 \in [v, p_2^{\max}]$. Thanks to [26, Lemma 4.1], we have

$$W = \rho_1^\star(p_2^{\max}) + \frac{p_2^{\max} - \bar{p}_{2,1}}{p_2^{\max} - v}\left[\rho_1^\star(v) - \rho_1^\star(p_2^{\max})\right] . \tag{44}$$

If, instead, $v < \widehat{\omega} < p_2^{\max}$, the function $\rho_1^\star(p_2)$ is not convex for $p_2 \in [v, p_2^{\max}]$, but it is made of two convex pieces, as discussed in Sec. IV-B. Therefore, we can think of $g_1(\cdot)$ as a combination of two distributions, say $g_{1,1}(\cdot)$ and $g_{1,2}(\cdot)$, with support in $[v, \widehat{\omega}]$ and $[\widehat{\omega}, p_2^{\max}]$, respectively, and such that $g_1(x) = u g_{1,1}(x) + (1-u) g_{1,2}(x)$, with $0 \leq u \leq 1$. For sake of simplicity, let us make a change of notation and simply refer to $[v, \widehat{\omega}]$ as $[a, b]$, and $[\widehat{\omega}, p_2^{\max}]$ as $[b, c]$.

We also define the average of the distributions $g_{1,1}(\cdot)$ and $g_{1,2}(\cdot)$ as μ_1 and μ_2, respectively, with $a \leq \mu_1 \leq b$, and $b \leq \mu_2 \leq c$. Then, the constraint on the average of $g_1(\cdot)$ provides $\int_a^c x g_1(x)\,\mathrm{d}x = \mu = u\mu_1 + (1-u)\mu_2$, which leads to $\mu_2 = \frac{\mu - p\mu_1}{1-p}$. Thus, we can rewrite W as

$$\begin{aligned} W &= \max_{g_{1,1}(\cdot), g_{1,2}(\cdot), u} \left[u \int_a^b g_{1,1}(x)\rho(x)\,\mathrm{d}x + (1-u)\int_b^c g_{1,2}(x)\rho(x)\,\mathrm{d}x \right] \\ &\overset{(a)}{=} \max_{u, \mu_1} u[d_1\rho(a) + (1-d_1)\rho(b)] + (1-u)[d_2\rho(b) + (1-d_2)\rho(c)] \end{aligned} \tag{45}$$

where $d_1 = \frac{b-\mu_1}{b-a}$, $d_2 = \frac{c-\mu_2}{c-b}$, and in the last equality we applied twice the result in [26, Lemma 4.1], thanks to the convexity of $\rho(x)$ for $x \in [a, b)$ and $x \in (b, c]$. By merging this result with the inequalities $b \leq \mu_2 \leq c$, we obtain $c - \frac{c-\mu}{u} \leq \mu_1 \leq b - \frac{b-\mu}{u}$. Furthermore, by introducing the constraint $a \leq \mu_1 \leq b$, we get $\max\left\{a, c - \frac{c-\mu}{u}\right\} \leq \mu_1 \leq \min\left\{b, b - \frac{b-\mu}{u}\right\}$. After some algebra, such result can be rewritten as $[b-\mu]^+ \leq t \leq \min\left\{u(b-a), u(b-c) + c - \mu\right\}$ where we defined $t \triangleq u(b - \mu_1)$. Clearly, the set of possible values for t is non-empty if the constraints $0 \leq u(b-a), 0 \leq u(b-c) + c - \mu, b - \mu \leq u(b-a), b - \mu \leq u(b-c) + c - \mu, 0 \leq u \leq 1$ are met, which means $\left[\frac{b-\mu}{b-a}\right]^+ \leq u \leq \min\left\{1, \frac{c-\mu}{c-b}\right\}$. We now observe that W in (45) can be rewritten as $W = \max_{u,\mu_1} w$ where $w = -zt + y$, with $z \triangleq \frac{\rho(b) - \rho(c)}{c-b} - \frac{\rho(a) - \rho(b)}{b-a}$ and $y \triangleq \frac{\rho(b)(c-\mu) - \rho(c)(b-\mu)}{c-b}$. The expression for w varies depending on z and μ, as follows:

- $\mu \leq b$ and $z \geq 0$: when $z \geq 0$, w is maximized if t is minimized, i.e., we impose $t = [b-\mu]^+$. Since $\mu \leq b$, this reduces to $t = b - \mu$. Then, $W = -\frac{\rho(a)-\rho(b)}{b-a}\mu + \frac{b\rho(a)-a\rho(b)}{b-a}$.
- $\mu > b$ and $z \geq 0$: by following the same considerations as before, the maximum is achieved when $t = 0$ and we obtain $W = -\frac{\rho(b)-\rho(c)}{c-b}\mu + \frac{c\rho(b)-b\rho(c)}{c-b}$.
- $\mu \leq b$ and $z < 0$: when $z < 0$, w is maximized if t is maximized. Then we set $t = \min\{u(b-a), u(b-c)+c-\mu\}$. Moreover, when $\mu \leq b$, the range of possible values for u reduces to $\frac{b-\mu}{b-a} \leq u \leq 1$. The optimum is achieved when $u = \frac{c-\mu}{c-a}$, i.e., $t = \frac{c-\mu}{c-a}(b-a)$, and we get $W = -\frac{\rho(a)-\rho(c)}{c-a}\mu + \frac{c\rho(a)-a\rho(c)}{c-a}$.
- $\mu > b$ and $z < 0$: we proceed as above and choose $t = \min\{u(b-a), u(b-c)+c-\mu\}$. In this case, since $\mu > b$, the range of possible values for u reduces to $0 \leq u \leq \frac{c-\mu}{c-b}$. Again, the optimum is achieved when $u = \frac{c-\mu}{c-a}$ and we obtain $W = -\frac{\rho(a)-\rho(c)}{c-a}\mu + \frac{c\rho(a)-a\rho(c)}{c-a}$.

REFERENCES

[1] D. Kudathanthirige and G. A. Aruma Baduge, "Massive mimo configurations for multi-cell multi-user relay networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1849–1868, March 2018.

[2] F. Shu, Y. Zhou, R. Chen, J. Wang, J. Li, and B. Vucetic, "High-performance beamformer and low-complexity detector for df-based full-duplex MIMO relaying networks," *China Communications*, vol. 14, no. 2, pp. 173–182, February 2017.

[3] Z. Zhang, X. Chai, K. Long, A. V. Vasilakos, and L. Hanzo, "Full duplex techniques for 5G networks: self-interference cancellation, protocol design, and relay selection," *IEEE Comm. Mag.*, vol. 53, no. 5, pp. 128–137, May 2015.

[4] H. Alves, R. D. Souza, and M. E. Pellenz, "Brief survey on full-duplex relaying and its applications on 5G," in *IEEE Workshop on Computer Aided Modelling and Design of Communication Links and Networks*, Sep. 2015, pp. 17–21.

[5] X. Xia, K. Xu, Y. Wang, and Y. Xu, "A 5G-enabling technology: Benefits, feasibility, and limitations of in-band full-duplex mMIMO," *IEEE Veh. Tech. Mag.*, vol. 13, no. 3, pp. 81–90, Sep. 2018.

[6] X. Chen, G. Liu, Z. Ma, X. Zhang, W. Xu, and P. Fan, "Optimal power allocations for non-orthogonal multiple access over 5G full/half-duplex relaying mobile wireless networks," *IEEE Trans. on Wir. Comm.*, vol. 18, no. 1, pp. 77–92, Jan 2019.

[7] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of loopback self-interference in full-duplex MIMO relays," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 12, pp. 5983–5993, Dec 2011.

[8] Y. Sun, Y. Yang, and Y. Z. P. Si, R. Yang, "Novel self-interference suppression schemes based on Dempster-Shafer theory with network coding in two-way full-duplex MIMO relay," *J. of Wir. Comm. Netw.*, vol. 109, Apr 2016.

[9] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of loopback self-interference in full-duplex MIMO relays," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 12, pp. 5983–5993, Dec 2011.

[10] E. Everett, C. Shepard, L. Zhong, and A. Sabharwal, "SoftNull: Many-antenna full-duplex wireless via digital beamforming," *IEEE Trans. on Wir. Comm.*, vol. 15, no. 12, pp. 8077–8092, 2016.

[11] G. C. Alexandropoulos and M. Duarte, "Joint design of multi-tap analog cancellation and digital beamforming for reduced complexity full duplex MIMO systems," in *IEEE ICC*, 2017, pp. 1–7.

[12] A. Shojaeifard, K. Wong, M. Di Renzo, G. Zheng, K. A. Hamdi, and J. Tang, "Self-interference in full-duplex multi-user MIMO channels," *IEEE Comm. Lett.*, vol. 21, no. 4, pp. 841–844, April 2017.

[13] S. Li, M. Zhou, J. Wu, L. Song, Y. Li, and H. Li, "On the performance of X-duplex relaying," *IEEE Trans. on Wir. Comm.*, vol. 16, no. 3, pp. 1868–1880, March 2017.

[14] A. Behboodi, A. Chaaban, R. Mathar, and M. Alouini, "On full duplex gaussian relay channels with self-interference," in *IEEE ISIT*, July 2016, pp. 1864–1868.

[15] N. Zlatanov, E. Sippel, V. Jamali, and R. Schober, "Capacity of the gaussian two-hop full-duplex relay channel with residual self-interference," *IEEE Trans. on Comm.*, vol. 65, no. 3, pp. 1005–1021, March 2017.

[16] Y. Y. Kang and J. H. Cho, "Capacity of MIMO wireless channel with full-duplex amplify-and-forward relay," in *IEEE PIMRC*, Sep. 2009, pp. 117–121.

[17] Y. Y. Kang, B. Kwak, and J. H. Cho, "An optimal full-duplex AF relay for joint analog and digital domain self-interference cancellation," *IEEE Trans. on Comm.*, vol. 62, no. 8, pp. 2758–2772, Aug 2014.

[18] H. A. Suraweera, I. Krikidis, G. Zheng, C. Yuen, and P. J. Smith, "Low-complexity end-to-end performance optimization in mimo full-duplex relay systems," *IEEE Trans. on Wir. Comm.*, vol. 13, no. 2, pp. 913–927, 2014.

[19] Y. Shao and T. A. Gulliver, "Precoding design for two-way MIMO full-duplex amplify-and-forward relay communication systems," *IEEE Access*, vol. 7, pp. 76 458–76 469, 2019.

[20] G. Zheng, "Joint beamforming optimization and power control for full-duplex MIMO two-way relay channel," *IEEE Trans. on Sig. Proc.*, vol. 63, no. 3, pp. 555–566, 2015.

[21] Q. Shi, M. Hong, X. Gao, E. Song, Y. Cai, and W. Xu, "Joint source-relay design for full-duplex MIMO AF relay systems," *IEEE Trans. on Sig. Proc.*, vol. 64, no. 23, pp. 6118–6131, 2016.

[22] Y. Cai, Y. Xu, Q. Shi, B. Champagne, and L. Hanzo, "Robust joint hybrid transceiver design for millimeter wave full-duplex MIMO relay systems," *IEEE Trans. on Wir. Comm.*, vol. 18, no. 2, pp. 1199–1215, 2019.

[23] H. Shen, W. Xu, and C. Zhao, "Outage minimized full-duplex multiantenna df relaying with csi uncertainty," *IEEE Trans. on Veh. Tech.*, vol. 67, no. 9, pp. 9000–9005, Sep. 2018.

[24] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Trans. on Wir. Comm.*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[25] Y. Gu, H. Chen, Y. Li, and B. Vucetic, "Ultra-reliable short-packet communications: Half-duplex or full-duplex relaying?" *IEEE Wir. Comm. Lett.*, vol. 7, no. 3, pp. 348–351, June 2018.

[26] A. Nordio, C. Chiasserini, and E. Viterbo, "Optimal power allocation strategies in two-hop x-duplex relay channel," *IEEE Trans. on Comm.*, vol. 66, no. 7, pp. 2888–2903, July 2018.

[27] N. Aldaghri and H. Mahdavifar, "Physical layer secret key generation in static environments," *IEEE Trans. on Inf. Forens. and Sec.*, vol. 15, pp. 2692–2705, 2020.

[28] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, and X. Li, "In-band full-duplex relaying: A survey, research issues and challenges," *IEEE Commu. Surv. Tut.*, vol. 17, no. 2, pp. 500–524, 2015.

[29] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. on Telecom.*, vol. 10, no. 6, pp. 585–595, 1999.

[30] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Found. and Trends in Comm. and Inf. Th.*, vol. 1, no. 1, pp. 1–182, 2004.