# Rate-Splitting Unifying SDMA, OMA, NOMA, and Multicasting in MISO Broadcast Channel: A Simple Two-User Rate Analysis

Bruno Clerckx, *Senior Member, IEEE*, Yijie Mao, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Considering a two-user multi-antenna Broadcast Channel, this letter shows that linearly precoded Rate-Splitting (RS) with Successive Interference Cancellation (SIC) receivers is a flexible framework for non-orthogonal transmission that generalizes, and subsumes as special cases, four seemingly different strategies, namely Space Division Multiple Access (SDMA) based on linear precoding, Orthogonal Multiple Access (OMA), Non-Orthogonal Multiple Access (NOMA) based on linearly precoded superposition coding with SIC, and physical-layer multicasting. This letter studies the sum-rate and shows analytically how RS unifies, outperforms, and specializes to SDMA, OMA, NOMA, and multicasting as a function of the disparity of the channel strengths and the angle between the user channel directions.**

***Index Terms*—Rate-splitting, multi-antenna broadcast channel, rate analysis, SDMA, OMA, NOMA, multicasting.**

## I. Introduction

LINEARLY precoded Rate-Splitting (RS) with Successive Interference Cancellation (SIC) receivers has recently appeared as a powerful non-orthogonal transmission and robust interference management strategy for multi-antenna wireless networks [1]. Though originally introduced for the two-user Single-Input Single-Output Interference Channel (IC) in [2], RS has become an underpinning communication-theoretic strategy to tackle modern interference-related problems and has recently been successfully investigated in several Multiple-Input Single-Output (MISO) Broadcast Channel (BC) settings, namely, unicast-only transmission with perfect Channel State Information at the Transmitter (CSIT) [3], [4] and imperfect CSIT [5]–[13], (multigroup) multicast-only transmission [14], as well as superimposed unicast and multicast transmission [15]. Results highlight that RS provides significant benefits in terms of spectral efficiency [3], [6], [7], [9], [13]–[15], energy efficiency [4], robustness [8], and CSI feedback overhead reduction [6], [12] over conventional strategies used in LTE-A/5G that rely on fully treating interference as noise (e.g., conventional multi-user linear precoding and Space Division Multiple Access - SDMA) or fully decoding interference (e.g., power-domain Non-Orthogonal Multiple Access - NOMA [16]). The key behind realizing those benefits is the ability of RS, through splitting messages into common and private parts, to partially decode interference and partially treat interference as noise. Additionally, RS is an enabler for powerful multiple access designs that subsumes SDMA and NOMA as special cases and outperforms them both for a wide range of network loads (underloaded/overloaded regimes) and user deployments (for diverse channel directions/strengths and CSIT qualities) [3]. *In this letter*, we build upon this last observation and show considering a simple two-user MISO BC with perfect CSIT that RS is a flexible framework for non-orthogonal transmission that generalizes, and subsumes as special cases, four seemingly completely different strategies, namely SDMA based on linear precoding, Orthogonal Multiple Access (OMA) where a resource is fully taken up by a single user, power-domain NOMA based on linearly precoded superposition coding with SIC, and physical-layer multicasting. This is the first paper to show analytically how RS unifies, outperforms, and specializes to SDMA, OMA, NOMA, and multicasting as a function of the disparity of the user channel strengths and the angle between the user channel directions. To that end, this letter differs from, and nicely complements, past works that analytically studied the rate performance of RS with imperfect CSIT [6], [9], [12] or looked at RS from an optimization perspective [3], [7], [8].

*Notation:* $|.|$ and $\|.\|$ refer to the absolute value of a scalar and the $l_2$-norm of a vector. $\mathbf{I}$ is the identity matrix. $\mathbf{a}^H$ denotes the Hermitian transpose of vector $\mathbf{a}$. I.i.d. stands for independent and identically distributed. $\mathcal{CN}(0,\sigma^2)$ denotes the Circularly Symmetric Complex Gaussian distribution with zero mean and variance $\sigma^2$. $\sim$ stands for "distributed as".

## II. System Model: Rate-Splitting Architecture

We consider a MISO BC consisting of one transmitter with $n_t$ antennas and two single-antenna users. As per Fig. 1, the architecture relies on rate-splitting of two messages $W_1$ and $W_2$ intended for user-1 and user-2, respectively. To that end, the message $W_k$ of user-$k$ is split into a common part $W_{\mathrm{c},k}$ and a private part $W_{\mathrm{p},k}$. The common parts $W_{\mathrm{c},1}$, $W_{\mathrm{c},2}$ of both users are combined into the common message $W_{\mathrm{c}}$, which is encoded into the common stream $s_{\mathrm{c}}$ using a codebook shared by both users. Hence, $s_{\mathrm{c}}$ is a common stream required to be decoded by both users, and contains parts of the messages $W_1$ and $W_2$ intended for user-1 and user-2, respectively. The private parts $W_{\mathrm{p},1}$ and $W_{\mathrm{p},2}$, respectively containing the remaining parts of the messages $W_1$ and $W_2$, are independently encoded into the private stream $s_1$ for user-1 and $s_2$ for user-2. Out of the two messages $W_1$ and $W_2$, three streams $s_{\mathrm{c}}$, $s_1$, and $s_2$ are therefore created. The streams are linearly

Manuscript received September 20, 2019; accepted November 16, 2019. Date of publication November 20, 2019; date of current version March 9, 2020. This work was supported in part by Engineering and Physical Sciences Research Council (EPSRC) of the U.K. under Grant EP/N015312/1 and Grant EP/R511547/1, and in part by the U.S. National Science Foundation under Grant CCF-1908308. The associate editor coordinating the review of this article and approving it for publication was Y. Huang. *(Corresponding author: Bruno Clerckx.)*

B. Clerckx and Y. Mao are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: b.clerckx@imperial.ac.uk; y.mao16@imperial.ac.uk).

R. Schober is with the Institute of Digital Communications, University of Erlangen-Nuremberg, 91058 Erlangen, Germany (e-mail: robert.schober@fau.de).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/LWC.2019.2954518

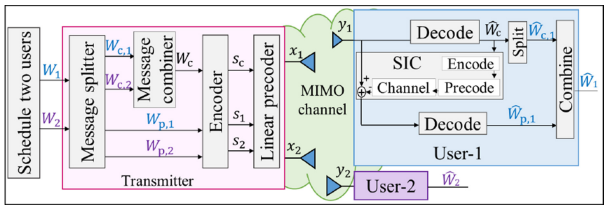



Fig. 1. Two-user system architecture with rate-splitting.

| | $s_1$ | $s_2$ | $s_c$ |
|---|---|---|---|
| SDMA | $W_1$ | $W_2$ | - |
| NOMA | $W_1$ | - | $W_2$ |
| OMA | $W_1$ | - | - |
| Multicasting | - | - | $W_1, W_2$ |
| RS | $W_{p,1}$ | $W_{p,2}$ | $W_{c,1}, W_{c,2}$ |

decoded by its intended user and treated as noise by the other user

decoded by both users

Fig. 2. Mapping of messages to streams.

precoded such that the transmit signal is given by

$$\mathbf{x} = \mathbf{p}_c s_c + \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2. \quad (1)$$

Defining $\mathbf{s} = [s_c, s_1, s_2]^T$ and assuming that $\mathbb{E}[\mathbf{s}\mathbf{s}^H]=\mathbf{I}$, the average transmit power constraint is written as $P_c+P_1+P_2 \leq P$ where $P_c = \|\mathbf{p}_c\|^2$ and $P_k = \|\mathbf{p}_k\|^2$ with $k = 1, 2$. We refer to $\mathbf{h}_k$ as the channel vector of user-$k$, such that the signal received at user-$k$ can be written as $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$, $k = 1, 2$, where $n_k \sim \mathcal{CN}(0,1)$ is Additive White Gaussian Noise (AWGN). We further write the channel vectors as the product of their norm and direction as $\mathbf{h}_k = \|\mathbf{h}_k\|\bar{\mathbf{h}}_k$, and assume without loss of generality $\|\mathbf{h}_1\| \geq \|\mathbf{h}_2\|$. We also assume perfect CSI at the transmitter and the receivers.

At each user-$k$, the common stream $s_c$ is first decoded into $\widehat{W}_c$ by treating the interference from the private streams as noise. Using SIC, $\widehat{W}_c$ is re-encoded, precoded, and subtracted from the received signal, such that user-$k$ can decode its private stream $s_k$ into $\widehat{W}_{p,k}$ by treating the remaining interference from the other private stream as noise. User-$k$ reconstructs the original message by extracting $\widehat{W}_{c,k}$ from $\widehat{W}_c$, and combining $\widehat{W}_{c,k}$ with $\widehat{W}_{p,k}$ into $\widehat{W}_k$. Assuming Gaussian signalling and ideal SIC, the rate of the common stream is given by

$$R_c = \min\left(\log_2\left(1 + \frac{\left|\mathbf{h}_1^H \mathbf{p}_c\right|^2}{1 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2 + \left|\mathbf{h}_1^H \mathbf{p}_2\right|^2}\right), \log_2\left(1 + \frac{\left|\mathbf{h}_2^H \mathbf{p}_c\right|^2}{1 + \left|\mathbf{h}_2^H \mathbf{p}_1\right|^2 + \left|\mathbf{h}_2^H \mathbf{p}_2\right|^2}\right)\right), \quad (2)$$

and the rates of the two private streams are obtained as

$$R_k = \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{p}_k\right|^2}{1 + \left|\mathbf{h}_k^H \mathbf{p}_j\right|^2}\right), k \neq j. \quad (3)$$

The rate of user-$k$ is given by $R_k + R_{c,k}$ where $R_{c,k}$ is the rate of the common part of the $k$th user's message, i.e., $W_{c,k}$, and it satisfies $R_{c,1} + R_{c,2} = R_c$. The sum-rate is therefore simply written as $R_s = \sum_{k=1,2} R_k + R_{c,k} = R_c + R_1 + R_2$.

By adjusting the message split and the power allocation to the common stream and the private streams, RS enables the decoding of part of the interference (thanks to the presence of the common stream) and treating the remaining part (the private stream of the other user) as noise. As a consequence, RS allows the exploration of a wide range of strategies. Among those strategies, there are four extreme cases, namely, SDMA, NOMA, OMA, and physical-layer multicasting.[1] Indeed, SDMA is obtained by allocating no power to the common stream ($P_c = 0$) such that $W_k$ is encoded directly into $s_k$. No interference is decoded at the receiver using the common message, and the interference between $s_1$ and $s_2$ is fully treated as noise. NOMA is obtained by encoding $W_2$ entirely into $s_c$ (i.e., $W_c = W_2$) and $W_1$ into $s_1$, and turning off $s_2$ ($P_2 = 0$). In this way, user-1 fully decodes the interference created by the message of user-2. OMA is a sub-strategy of SDMA and NOMA and is obtained when only user-1 (with the stronger channel gain) is scheduled ($P_c = 0, P_2 = 0$). Multicasting is obtained by combining and encoding both $W_1$ and $W_2$ into $s_c$, and turning off $s_1$ and $s_2$ ($P_1 = 0, P_2 = 0$). Fig. 2 illustrates the mapping of the messages to the streams.

*Remark 1:* The maximum number of interference-free streams (also called Degrees-of-Freedom DoF) in a two-user MISO BC is equal to 2. From the above system model, both SDMA and RS achieve such a DoF by precoding $s_1$ and $s_2$ using zero-forcing (ZF). On the other hand, OMA, NOMA, and multicasting achieve at most a DoF of 1 (irrespectively of how the precoders and power allocation are optimized), which leads to a rate loss at high Signal-to-Noise Ratio (SNR) in general multi-antenna settings, as highlighted in [3], [14].

## III. Sum-Rate Analysis

Our objective is to derive tractable and insightful sum-rate expressions in a two-user MISO BC to illustrate the flexibility of RS in unifying SDMA, OMA, NOMA, and multicasting.[2] To that end, we do not optimize the precoding directions jointly with the power allocation as in [3], [7] but rather fix the precoding directions using ZF for the private streams, and adjust the power allocation among all the streams.[3] This leads to $|\mathbf{h}_2^H \mathbf{p}_1| = 0$, $|\mathbf{h}_1^H \mathbf{p}_2| = 0$, and $|\mathbf{h}_k^H \mathbf{p}_k|^2 = \|\mathbf{h}_k\|^2 \rho P_k$, $k = 1, 2$, where $\rho = 1 - |\bar{\mathbf{h}}_1^H \bar{\mathbf{h}}_2|^2$ ($\rho = 0$ corresponds to aligned channels and $\rho = 1$ to orthogonal channels). The precoder of the common stream is then designed such that

$$\max_{\mathbf{p}_c} \min\left(\frac{|\mathbf{h}_1^H \mathbf{p}_c|^2}{1+|\mathbf{h}_1^H \mathbf{p}_1|^2}, \frac{|\mathbf{h}_2^H \mathbf{p}_c|^2}{1+|\mathbf{h}_2^H \mathbf{p}_2|^2}\right), \text{s.t. } \|\mathbf{p}_c\|^2 = P_c. \quad (4)$$

Defining $\gamma_k^2 = 1 + |\mathbf{h}_k^H \mathbf{p}_k|^2 = 1 + \|\mathbf{h}_k\|^2 \rho P_k$, $k = 1, 2$, and $\tilde{\mathbf{h}}_k = \mathbf{h}_k / \gamma_k$, the problem is re-written as

$$\max_{\mathbf{p}_c} \min\left(\left|\tilde{\mathbf{h}}_1^H \mathbf{p}_c\right|^2, \left|\tilde{\mathbf{h}}_2^H \mathbf{p}_c\right|^2\right), \text{ s.t. } \|\mathbf{p}_c\|^2 = P_c. \quad (5)$$

Following [17], the solution of (5) is $\mathbf{p}_c = \sqrt{P_c}\mathbf{f}_c$ with the precoder direction $\mathbf{f}_c$ ($\|\mathbf{f}_c\|^2 = 1$) given by

$$\mathbf{f}_c = \frac{1}{\sqrt{\lambda}}\left(\mu_1 \tilde{\mathbf{h}}_1 + \mu_2 \tilde{\mathbf{h}}_2 e^{-j\angle\alpha_{12}}\right), \quad (6)$$

where

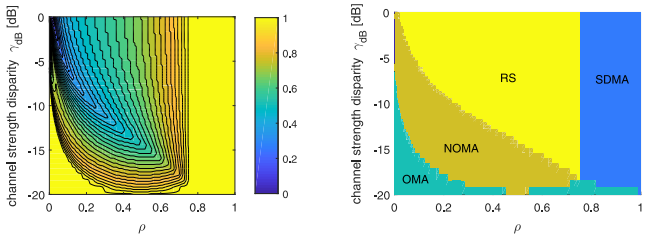
$$\lambda = \frac{\alpha_{11}\alpha_{22} - |\alpha_{12}|^2}{\alpha_{11} + \alpha_{22} - 2|\alpha_{12}|}, \quad (7)$$

$$\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} = \frac{1}{\alpha_{11} + \alpha_{22} - 2|\alpha_{12}|}\begin{bmatrix}\alpha_{22} - |\alpha_{12}| \\ \alpha_{11} - |\alpha_{12}|\end{bmatrix}, \quad (8)$$

[1]The complexity of RS versus those strategies is discussed in [3].

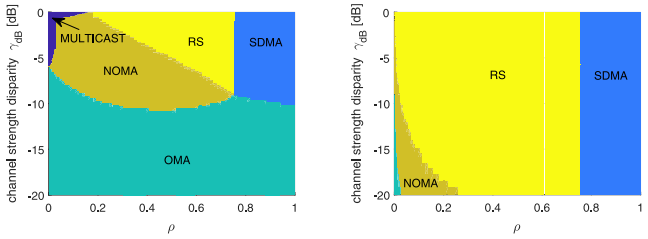
[2]Other metrics/setups accounting for fairness and $K$ users have been studied using optimization in [3], [8], [14]. Conclusions drawn here on the superiority and versatility of RS also hold for $K$ users and under fairness constraints.

[3]Simulations in Section IV show that the conclusions drawn with the simple precoders also hold with the numerically optimized precoders of [3], [7].




(a) Optimum $t$ (b) Regions of operation

Fig. 3. Optimum $t$ in (a) and regions of operation for RS, SDMA, NOMA, and OMA in (b). Precoders from Section III with $P = 100$ (SNR = 20dB).




(a) $P = 10$ (SNR=10dB) (b) $P = 1000$ (SNR=30dB)

Fig. 4. Regions of operation for RS, SDMA, NOMA, OMA and Multicast with precoders from Section III for $P = 10$, 1000 (SNR = 10dB, 30dB).

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12}^* & \alpha_{22} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{h}}_1^H \\ \tilde{\mathbf{h}}_2^H \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{h}}_1 & \tilde{\mathbf{h}}_2 \end{bmatrix}. \tag{9}$$

### A. Sum-Rate at Finite SNR

The sum-rate with the above precoder designs can be written as $R_{\rm s} = R_{\rm c} + \log_2(\gamma_1^2) + \log_2(\gamma_2^2)$, where $R_{\rm c} = \min(\log_2(1+|\tilde{\mathbf{h}}_1^H \mathbf{p}_{\rm c}|^2), \log_2(1+|\tilde{\mathbf{h}}_2^H \mathbf{p}_{\rm c}|^2))$. With $\mathbf{p}_{\rm c}$ as per (6), following [17], $|\tilde{\mathbf{h}}_1^H \mathbf{p}_{\rm c}| = |\tilde{\mathbf{h}}_2^H \mathbf{p}_{\rm c}|$, and we can write $R_{\rm c} = \log_2(1 + |\tilde{\mathbf{h}}_2^H \mathbf{p}_{\rm c}|^2)$, and the sum-rate simply as

$$R_{\rm s} = \log_2\left(\gamma_1^2\right) + \log_2\left(\gamma_2^2 + \left|\mathbf{h}_2^H \mathbf{p}_{\rm c}\right|^2\right). \tag{10}$$

Consider a fraction $t$ of the total transmit power $P$ is allocated to the private streams such that $P_1 + P_2 = tP$ and the remaining power $P_{\rm c} = (1-t)P$ is allocated to the common stream. For a given $t$, the optimal values of $P_1$ and $P_2$, maximizing the sum-rate of the private streams, are given by the Water-Filling (WF) solution

$$P_k = \max\left(\mu - \frac{1}{\|\mathbf{h}_k\|^2 \rho}, 0\right), k = 1, 2, \tag{11}$$

with the water level $\mu$ chosen such that $P_1 + P_2 = tP$, and set as $\mu = \frac{tP}{2} + \frac{1}{2\rho}[\frac{1}{\|\mathbf{h}_1\|^2} + \frac{1}{\|\mathbf{h}_2\|^2}]$ in the sequel. Let us also introduce $\Gamma = \frac{1}{\rho}[\frac{1}{\|\mathbf{h}_2\|^2} - \frac{1}{\|\mathbf{h}_1\|^2}]$, which is a function of two main parameters: $\rho$ reflecting the angle between the user channel directions, and $\frac{1}{\|\mathbf{h}_2\|^2} - \frac{1}{\|\mathbf{h}_1\|^2}$ reflecting the disparity of the channel strengths. We can then identify two main regimes.

*1) OMA/NOMA/Multicasting Regime:* If $\mu \leq \frac{1}{\|\mathbf{h}_2\|^2 \rho}$, i.e., $tP \leq \Gamma$, we set $P_2 = 0$ and $P_1 = tP$ according to (11), and RS specializes to multicasting for $t = 0$, NOMA for $0 < t < 1$, and OMA for $t = 1$. In this regime, $t$ needs to be adjusted so as to identify the best strategy among OMA, NOMA, and multicasting, and therefore efficiently allocate power across the common stream $s_{\rm c}$ and the private stream $s_1$.

*2) RS/SDMA Regime:* If $\mu > \frac{1}{\|\mathbf{h}_2\|^2 \rho}$, i.e., $tP > \Gamma$, the WF solution (11) leads to $P_1 = \mu - \frac{1}{\|\mathbf{h}_1\|^2 \rho} = \frac{tP}{2} + \frac{\Gamma}{2} > 0$ and $P_2 = \mu - \frac{1}{\|\mathbf{h}_2\|^2 \rho} = \frac{tP}{2} - \frac{\Gamma}{2} > 0$. RS specializes to SDMA whenever $t$ is set to 1, but does not specialize to any other known scheme for $0 < t < 1$. In this regime, $t$ needs to be adjusted, as explained in the sequel, so as to allocate the power efficiently across the common stream and the two private streams. Substituting the expressions of $P_k$ and $\gamma_k^2$, $k = 1, 2$, into (10), we can write

$$R_{\rm s} = \log_2\Big(ac + (ad + bc)t + bdt^2\Big), \tag{12}$$

where $b = \frac{\|\mathbf{h}_1\|^2 \rho P}{2}$, $a = 1 + \frac{\Gamma}{P} b$, $d = \frac{\|\mathbf{h}_2\|^2 \rho P}{2} - |\mathbf{h}_2^H \mathbf{f}_c|^2 P$, and $c = 1 - \frac{\Gamma}{P} d + |\mathbf{h}_2^H \mathbf{f}_c|^2 (P - \Gamma)$. The value of $t$ that maximizes $R_{\rm s}$ is the solution of $\frac{\partial R_s}{\partial t} = 0$, which is written as $t = -\frac{a}{2b} - \frac{c}{2d}$. Since $t \leq 1$, the optimal value $t^\star$ is given in closed form by (13) at the bottom of the next page. For $t^\star < 1$, RS yields a non-zero sum-rate enhancement over SDMA.

*Remark 2:* It is important to note that the solution $t = -\frac{a}{2b} - \frac{c}{2d}$ holds because the coefficients $a$, $b$, $c$, $d$ are not functions of $t$. This could appear surprising since $c$ and $d$ are functions of $\mathbf{f}_c$, which, according to (5), is a function of $P_1$ and $P_2$ and therefore of $t$. However, interestingly, in the regime where $P_1 > 0$ and $P_2 > 0$, we can show that $\mathbf{f}_c$ is not a function of $t$. Making use of $P_1 = \frac{tP}{2} + \frac{\Gamma}{2}$ and $P_2 = \frac{tP}{2} - \frac{\Gamma}{2}$, we can write $\gamma_k^2 = 1 + \|\mathbf{h}_k\|^2 \rho P_k = \frac{f(t)}{\|\mathbf{h}_j\|^2}$, $k, j = 1, 2$ and $k \neq j$, with $f(t) = \frac{\|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2 + \|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2 \rho P t}{2}$. We then obtain

$$\begin{aligned} &\max_{\mathbf{f}_{\rm c}} \min\Big(\left|\tilde{\mathbf{h}}_1^H \mathbf{f}_{\rm c}\right|^2, \left|\tilde{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2\Big) \\ &\Leftrightarrow \max_{\mathbf{f}_{\rm c}} \min\Big(\gamma_2^2 \left|\mathbf{h}_1^H \mathbf{f}_{\rm c}\right|^2, \gamma_1^2 \left|\mathbf{h}_2^H \mathbf{f}_{\rm c}\right|^2\Big) \\ &\Leftrightarrow \max_{\mathbf{f}_{\rm c}} \min\Big(f(t) \left|\bar{\mathbf{h}}_1^H \mathbf{f}_{\rm c}\right|^2, f(t) \left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2\Big) \\ &\Leftrightarrow \max_{\mathbf{f}_{\rm c}} \min\Big(\left|\bar{\mathbf{h}}_1^H \mathbf{f}_{\rm c}\right|^2, \left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2\Big), \end{aligned} \tag{14}$$

which reveals that $\mathbf{f}_{\rm c}$ is not a function of $t$ and the channel strength disparity, but only of the channel directions.

### B. Sum-Rate at High SNR

At high SNR, considering $0 < t \leq 1$ and $\rho > 0$, the solution in (11) allocates power uniformly across the two private streams as $P_1 = P_2 = \frac{tP}{2} > 0$. Hence, only RS and SDMA are suitable strategies at high SNR. The sum-rate in (10) can then be written as

$$R_{\rm s} \overset{P \nearrow}{=} \log_2\Big(\|\mathbf{h}_1\|^2 \rho\Big) + 2\log_2(P) + \log_2\Big(et^2 + ft\Big) \tag{15}$$

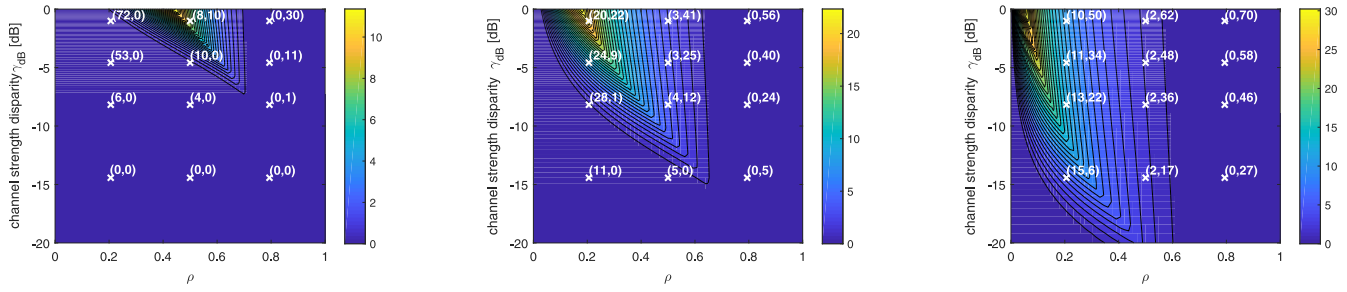
with $e = \frac{\|\mathbf{h}_2\|^2 \rho}{4} - \frac{|\mathbf{h}_2^H \mathbf{f}_{\rm c}|^2}{2}$, $f = \frac{|\mathbf{h}_2^H \mathbf{f}_{\rm c}|^2}{2}$. Not surprisingly, a DoF of 2 is achieved in (15). More interesting is the fact that RS brings a constant sum-rate enhancement over SDMA. Indeed, the value of $t$ that maximizes (15) is given by

$$t^\star = \min\left(\frac{-f}{2e}, 1\right) = \min\left(\frac{\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2}{2\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2 - \rho}, 1\right), \tag{16}$$

which coincides with (13) when $P \to \infty$, and leads to a high SNR non-zero (whenever $0 < t^\star < 1$) sum-rate gap between RS and SDMA ($t = 1$) given by

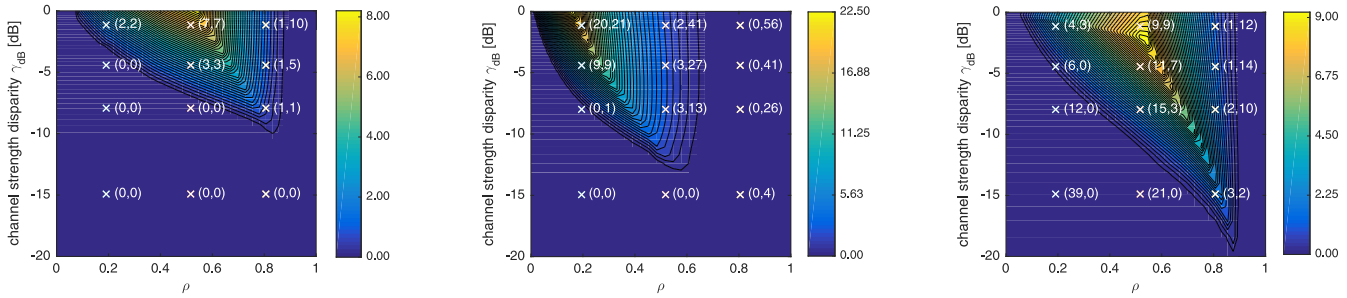
$$\Delta R_{\rm s} = R_{\rm s}|_{t^\star} - R_{\rm s}|_{t=1} = \log_2\left(\frac{\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^4}{\rho\Big(2\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\rm c}\right|^2 - \rho\Big)}\right). \tag{17}$$

$t^\star$ increases and $\Delta R_{\rm s}$ decreases as $\rho$ increases, and both are not a function of the channel strengths. The sum-rate gap between RS and NOMA/OMA/multicasting grows unbounded as $P \to \infty$ due to the difference in DoF (Remark 1).




(a) $P = 10$ (SNR=10dB) (b) $P = 100$ (SNR=20dB) (c) $P = 1000$ (SNR=30dB)

Fig. 5. Relative sum-rate gain [%] of RS over dynamic switching between SDMA and NOMA, with $n_t = 2$ and precoders from Section III. The values in brackets indicate sum-rate gains [%] over SDMA and NOMA, respectively.




(a) $u_1 = 10^{0.5}, u_2 = 1$ (b) $u_1 = 1, u_2 = 1$ (c) $u_1 = 1, u_2 = 10^{0.5}$

Fig. 6. Relative weighted sum-rate gain [%] of RS over dynamic switching between SDMA and NOMA for different values of weights $u_1, u_2$, with precoders based on WMMSE optimization, $n_t = 2$, and $P = 100$. The values in brackets indicate weighted sum-rate gains [%] over SDMA and NOMA, respectively.

### C. Discussions

We can draw several insights from the above analysis. *First*, for given $t$, $\rho$, $\|\mathbf{h}_1\|^2$, and $\|\mathbf{h}_2\|^2$, as $P$ increases, the SNRs of the private streams increase, while the Signal-to-Interference-plus-Noise Ratio (SINR) of the common stream ultimately saturates (interference limited regime). This suggests that the common message can only provide a constant rate improvement at high SNR, while the two private streams provide the DoF of 2. *Second*, the quantity $\rho$ is present in the SNRs of both private streams and has the effect of increasing/decreasing the SNRs of those two streams. A lower $\rho$ indicates that both private streams effectively operate at a lower SNR. According to (11), for a given $t$, a low $\rho$ favors power allocation to a single private stream (NOMA/OMA/Multicasting regime) over a wider range of $P$, and also leads to a smaller interference power (and therefore a higher rate) for the common stream. A higher $\rho$ leads to a higher effective SNR and therefore a better capability to support two private streams (RS/SDMA regime). *Third*, as the disparity of channel strengths increases, the WF solution allocates a larger amount of power to the stronger user (user-1) over a wider range of $P$ (for a given $t$). Beyond a certain disparity, for given $t$, $P$, and $\rho$, $P_2$ is turned off and RS specializes to NOMA/OMA.

## IV. Evaluations

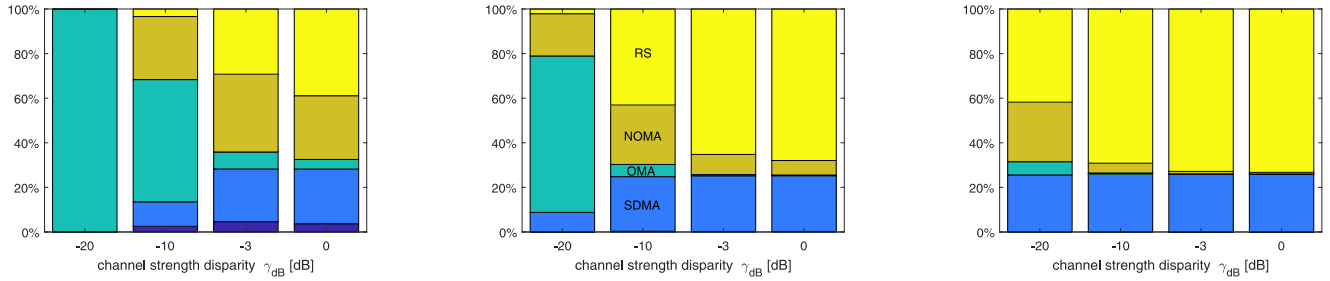
In this section, we first illustrate the above analysis and the preferred regions for the operation of NOMA, OMA, SDMA, and RS. We assume $n_t = 2$, and channel vectors given by $\mathbf{h}_1 = 1/\sqrt{2}\,[1, 1]^H$ and $\mathbf{h}_2 = \gamma/\sqrt{2}\,[1, e^{j\theta}]^H$.

Assuming the precoding strategies in Section III and the WF power allocation (11), the colors in Fig. 3(a) and (b) illustrate the optimum value (obtained from exhaustive search whenever not available in closed form) of $t$ that maximizes the sum-rate and the corresponding preferred communication strategy (RS, SDMA, NOMA, OMA) as a function of $\rho = 1 - |\bar{\mathbf{h}}_1^H \bar{\mathbf{h}}_2|^2$ (ranging from 0 to 1) and $\gamma_{\text{dB}} = 20\log_{10}(\gamma)$ (ranging from 0 to $-20$dB). For $P = 100$, given the noise variance and channel gain normalization, user-1 and user-2 have a long-term SNR of 20dB and 0 dB $\leq$ 20 dB $+\gamma_{\text{dB}} \leq$ 20 dB, respectively. Recall that SDMA is characterized by $t = 1, P_1 > 0, P_2 > 0$, NOMA by $0 < t < 1, P_1 > 0, P_2 = 0$, OMA by $t = 1, P_1 = P, P_2 = 0$, and multicast by $t = 0, P_1 = 0, P_2 = 0$. For all other regimes, RS does not specialize to any other well-established scheme and is simply referred to as RS. We observe that NOMA is preferred for deployments with small $\rho$, i.e., closely aligned users, and small $\gamma$, SDMA is preferred whenever $\rho$ is sufficiently large, i.e., semi-orthogonal users, and RS bridges those two extremes. OMA is preferred whenever $\gamma$ is very small.

Fig. 3 is obtained for $P = 100$. In Fig. 4, we illustrate the regions for $P = 10$ and $P = 1000$. The long term SNR of user-1 is therefore 10 dB and 30 dB, respectively, and that of user-2 is up to 20dB lower. As $P$ increases, RS becomes the dominant strategy for most deployment conditions.

Fig. 5 shows the relative sum-rate gain [%] of RS over dynamic switching between SDMA and NOMA, defined as $\frac{R_{\text{s}}^{\text{RS}} - \max(R_{\text{s}}^{\text{SDMA}}, R_{\text{s}}^{\text{NOMA}})}{\max(R_{\text{s}}^{\text{SDMA}}, R_{\text{s}}^{\text{NOMA}})} \times 100$, for $P = 10$, 100, 1000 and the precoders from Section III. RS provides explicit

$$t^\star = \min\Big(-\frac{a}{2b} - \frac{c}{2d}, 1\Big) = \min\left(\frac{\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\text{c}}\right|^2}{2\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\text{c}}\right|^2 - \rho} + \frac{1}{2\rho}\left(\frac{1}{\|\mathbf{h}_1\|^2} + \frac{1}{\|\mathbf{h}_2\|^2}\right)\left(\frac{2\rho - 2\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\text{c}}\right|^2}{2\left|\bar{\mathbf{h}}_2^H \mathbf{f}_{\text{c}}\right|^2 - \rho}\right)\frac{1}{P}, 1\right) \tag{13}$$




(a) $P = 10$ (SNR=10dB) (b) $P = 100$ (SNR=20dB) (c) $P = 1000$ (SNR=30dB)

Fig. 7. Percentage of operation of RS, SDMA, NOMA, OMA, and Multicast with precoders from Section III for $P = 10$, 100, 1000, with $n_t = 2$.

gains over dynamic switching for medium values of $\rho$. The values in brackets indicate the relative sum-rate gains [%] over SDMA and NOMA, respectively, i.e., $(\frac{R_{\rm s}^{\rm RS}-R_{\rm s}^{\rm SDMA}}{R_{\rm s}^{\rm SDMA}} \times 100, \frac{R_{\rm s}^{\rm RS}-R_{\rm s}^{\rm NOMA}}{R_{\rm s}^{\rm NOMA}} \times 100)$. Large gains over SDMA are observed for low to medium values of $\rho$, and over NOMA for medium to large values of $\rho$ at low SNR and for all values of $\rho$ and $\gamma_{\rm dB}$ at higher SNR. Values (0, 0) indicate that OMA is the preferred strategy, and that RS, SDMA, and NOMA all specialize to OMA.

Fig. 6 is similar to Fig. 5 but now the Weighted Minimum Mean Square Error (WMMSE) precoding optimization framework for RS developed in [3], [7] is adopted. Such framework optimizes all precoders $(\mathbf{p}_{\rm c}, \mathbf{p}_1, \mathbf{p}_2)$ jointly with the power allocations so as to maximize the weighted sum-rate $\sum_{k=1,2} u_k(R_k + R_{{\rm c},k})$. In those evaluations, the convergence tolerance of the WMMSE algorithm is set to $\epsilon = 10^{-3}$ [3]. When allocating equal weights or higher weights to the user with the stronger channel (namely user-1), NOMA has no benefit over SDMA. When a higher weight is given to the weaker user (user-2), NOMA is able to outperform SDMA. RS on the other hand always provides the same or better performance than both SDMA and NOMA for all weights, $\rho$, and $\gamma_{\rm dB}$. Though the precoders of Section III are simple and not optimal, the insights obtained from the analysis and Fig. 5 are inline with those obtained from Fig. 6. Hence, irrespectively of the precoding strategies, i.e., simple or optimized, RS unifies and outperforms SDMA, OMA, NOMA, and multicasting.

We now change the channel model and assume i.i.d. Rayleigh fading, i.e., the entries of $\mathbf{h}_1$ and $\mathbf{h}_2$ are $\mathcal{CN}(0, 1/n_t)$ and $\mathcal{CN}(0, \gamma^2/n_t)$. We generate 10000 channel realizations. Making use of the precoders in Section III, we identify the preferred (i.e., sum-rate maximizing) strategy for each channel realization. Fig. 7 displays the percentage a given strategy is the preferred option as a function of $P$ and $\gamma_{\rm dB}$ for $n_t = 2$. OMA is preferred for low $P$ and low $\gamma_{\rm dB}$, and RS becomes the preferred option as $P$ and/or $\gamma_{\rm dB}$ increase. At high SNR, RS is the preferred option for about 75% of the channel realizations and SDMA for the remaining 25%. Results with $n_t = 4$ (not reproduced here due to the space constraint) show that NOMA almost disappears from the set of preferred strategies, and SDMA becomes more dominant (for about 60% of the channel realizations and RS for the remaining 40%). This is natural since, as $n_t$ increases, the likelihood of experiencing large $\rho$ increases, and $t^\star$ has a higher chance of being equal to 1.

## V. Conclusion

RS unifies SDMA, OMA, NOMA, and multicasting under a single approach and provides a powerful framework for the design and optimization of non-orthogonal transmission, multiple access, and interference management strategies. Thanks to its versatility, RS has the potential to tackle challenges of modern communication systems and is a rich source of research problems for academia and industry, spanning fundamental limits, optimization, PHY/MAC layers, and standardization.

## References

[1] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[2] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, Jan. 1981.

[3] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, p. 133, May 2018.

[4] Y. Mao, B. Clerckx, and V. O. K. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2018, pp. 1–5.

[5] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan. 2013.

[6] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, Sep. 2015.

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

[8] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, Vol. 64, No. 23, pp. 6227–6242, Dec. 2016.

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

[10] A. G. Davoodi and S. A. Jafar, "GDoF of the MISO BC: Bridging the gap between finite precision CSIT and perfect CSIT," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1297–1301.

[11] E. Piovano and B. Clerckx, "Optimal DoF region of the $K$-user MISO BC with partial CSIT," *IEEE Commun. Lett.*, vol. 21, no. 11, pp. 2368–2371, Nov. 2017.

[12] M. Dai and B. Clerckx, "Multiuser millimeter wave beamforming strategies with quantized and statistical CSIT," *IEEE Trans. Wireless Commun.*, vol 16, no 11, pp7025–7038, Nov. 2017.

[13] G. Lu, L. Li, H. Tian, and F. Qian, "MMSE-based precoding for rate splitting systems with finite feedback," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 642–645, Mar. 2018.

[14] H. Joudeh and B. Clerckx, "Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7276–7289, Nov. 2017.

[15] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2018, pp. 1–5.

[16] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[17] C.-L. Hsiao, j.-C. Guey, W.-H. Sheen, and R.-J. Chen, "A two-user approximation-based transmit beamforming for physical-layer multicasting in mobile cellular downlink systems," *J. Chin. Inst. Eng.*, vol. 38, no. 6, pp. 742–750, 2015.