On the Optimality of NOMA in Two-User Downlink Multiple Antenna Channels

Eduard Jorswieck and Sepehr Rezvani
Institute for Communications Technology
Department of Information Theory and Communication Systems
TU Braunschweig, Lower Saxony, Germany
{jorswieck,rezvani}@ifn.ing.tu-bs.de

***Abstract*—Non-orthogonal multiple access (NOMA) is proposed as downlink precoding scheme in cellular wireless networks. Applying multiple antennas at the base station allows for spatial precoding and spatial multiplexing. In particular, the capacity region for multiple antenna downlink channels is achieved by dirty paper precoding (DPC) and time sharing. It cannot be achieved by linear superposition coding and successive interference cancellation (SIC). In this work, we investigate the performance of NOMA in multiple antenna downlink channels. We show that NOMA cannot achieve the sum capacity except for parallel channels. Furthermore, we propose an efficient algorithm to compute the minimum power for DPC with time sharing under rate constraint. The results are compared with NOMA and zero-forcing (ZF) precoding. The numerical experiments show that NOMA requires significantly more transmit power than DPC. Moreover, the performance gap between ZF precoding and NOMA highly depends on the SNR region, number of antennas, and minimum rate demands.**

***Index Terms*—Non-orthogonal multiple access, multiple-input single-output, broadcast channel, dirty paper precoding, zero-forcing**

I. Introduction

As the fifth-generation (5G) wireless networks have been commercialized, it is natural for the researchers to exploit the potential multiple access techniques in beyond 5G (B5G) and the sixth-generation (6G) wireless networks. Since spectrum is scarce resource and the number of mobile devices is steadily increasing, multiple access techniques for up- and downlink transmission are most important.

Non-orthogonal multiple access (NOMA) has been recently proposed for the 3rd generation partnership projects long-term evolution advanced (3GPP-LTE-A). It constitutes a promising technology of enhancing the spectral efficiency and achieving massive connectivity challenges by accommodating several users within the same orthogonal resource block, via multiplexing at different power levels. The information theoretic properties of NOMA are reviewed in [1] while a short description of the extensions to multiple-antenna cases is included. One important property for single-antenna single-cell NOMA systems is that the optimal decoding order for SIC depends only on the channel gain but not on the transmit strategies, i.e., transmit power. This follows from the degradedness of the SISO broadcast channel (BC) [2].

In [3], the application of multiple-input multiple-output (MIMO) techniques to NOMA is proposed. Precoding and detection matrices are proposed to improve the performance gap between MIMO-NOMA and conventional orthogonal multiple access schemes. The concept of signal alignment is applied to MIMO-NOMA up- and downlink systems in [4]. The optimal precoding for quality-of-service (QoS) optimization in two-user multiple antenna NOMA is solved in [5], [6]. Later, in [7], a comprehensive solution for the design, analysis, and optimization of a multiple-antenna non-orthogonal multiple access (NOMA) system for multiuser downlink communication with both time duplex division (TDD) and frequency duplex division (FDD) modes is provided. In [8] a review on multiple antenna techniques for NOMA is presented. Finally, beamforming design for multiple antenna NOMA systems by considering a power minimization problem under rate constraints [9]. The general MIMO BC is not degraded. It is shown that under specific channel condition, called quasi-degradation, NOMA has the same performance as DPC. Under this condition, the optimal decoding order of NOMA and coding order of DPC are independent from beamforming and only depend on the channel gains of the multiplexed users. Therefore, the question whether NOMA can achieve the capacity region is still open.

In a different approach, the maximization of the achievable sum-rate with NOMA is considered in [10] where the NOMA order is prefixed again according to the channel gains. In [11], clustering and linear beamforming is proposed to cancel the inter-cluster interference while the users are sorted based on their channel gains. Another approach for cell-free massive MIMO with NOMA is introduced in **Yikai18**.

Recognizing that NOMA does not achieve the capacity in multiple antenna downlink, a more general approach based on rate-splitting is proposed in [12]. The rate-splitting approach is more flexible and can cover spatial division multiple access (SDMA), NOMA and orthogonal multiple access (OMA) as special cases. Finally, the efficiency of NOMA in multiple antenna downlink transmission is studied in [13]. First, it is shown that NOMA is suboptimal in most operating regions. Second, alternative schemes, including rate-splitting are suggested to improve significantly the performance.

Motivated by recent studies on the optimality of NOMA, we consider a simple idealized two-user multiple antenna downlink transmission. First, we review the capacity region achieved by dirty paper precoding and time sharing as well as the achievable rate regions by NOMA and treating interference as noise (TIN). Next, we analytically show that NOMA cannot achieve the sum capacity in general, only if the channels are parallel. Finally, we derive an algorithm to find minimum transmit power for given rate requirements. This completes recent results [9] on the optimality of NOMA for the power minimization problem to include the complete convex hull (achieved with time sharing).

A. System Model

We consider the simple two-user multiple antenna downlink channel including a single BS with n transmit antennas and single antenna receivers. The transmitter applies a beamforming for each receiver. The beamforming vectors are denoted by $\boldsymbol{w}_1, \boldsymbol{w}_2$. They can be chosen under a transmit power constraint $||\boldsymbol{w}_1||^2 + ||\boldsymbol{w}_2||^2 \leq P$. The receiver noise is additive white Gaussian noise with variance σ^2. The received signal at the i-th receiver is given by

$$y_i = \boldsymbol{h}_i^H \boldsymbol{w}_1 x_1 + \boldsymbol{h}_i^H \boldsymbol{w}_2 x_2 + n_i, \tag{1}$$

with noise n_i, transmit signals x_1 and x_2 for receiver 1 and 2, respectively. The channels are denoted by vectors $\boldsymbol{h}_1, \boldsymbol{h}_2$. $\boldsymbol{x}^H$ denotes the Hermitian of the vector $\boldsymbol{x}$.

II. Preliminaries

Here, we review the available results and transceiver schemes for the multiple-antenna downlink channel. We start with the capacity region achieved by dirty paper precoding (DPC), followed by NOMA[1] and TIN. Finally, we briefly mention the duality theory between multi-antenna BC and multiple access channel (MAC) which helps us to prove our main results.

A. Dirty Paper Precoding

The capacity region of the multiple antenna BC is derived in [14]. It is shown that DPC can achieve the capacity region. The DPC achievable rate region is introduced in [14, Definition 5] as

$$\mathcal{R}^{\mathrm{DPC}} = \mathrm{cv}\Big\{ \bigcup_{\pi \in \Pi} \mathcal{R}^{\mathrm{DPC}}(\pi, P, \boldsymbol{h}_1, \boldsymbol{h}_2) \Big\}, \tag{2}$$

with Π as the two permutations for the two precoding orders π_1, where user 2 cancels user 1, and π_2, where user 1 cancels user 2. $\mathrm{cv}\{\cdot\}$ denotes the convex closure operator and P as the average transmit power constraints and $\mathcal{R}^{DPC}(\pi, P, \boldsymbol{h}_1, \boldsymbol{h}_2)$ corresponds to the achievable rate region of DPC with precoding order π. We introduce the abbreviation $C(x) = \log(1+x)$. The corresponding achievable rates are for $\pi_1 = [1, 2]$

$$R_1^{\mathrm{DPC}}(\pi_1, \rho, \boldsymbol{h}_1, \boldsymbol{h}_2) = C\left(\frac{\rho |\boldsymbol{h}_1^H \boldsymbol{w}_1|^2}{1 + \rho |\boldsymbol{h}_1^H \boldsymbol{w}_2|^2} \right), \tag{3}$$

$$R_2^{\mathrm{DPC}}(\pi_1, \rho, \boldsymbol{h}_1, \boldsymbol{h}_2) = C\left(\rho |\boldsymbol{h}_2^H \boldsymbol{w}_2|^2 \right), \tag{4}$$

[1]We use the term NOMA to refer to power-domain NOMA.

with tranmit SNR $\rho = \frac{P}{\sigma_n^2}$. The achievable rates for $\pi_2 = [2, 1]$ are given by

$$R_1^{DPC}(\pi_2, \rho, \boldsymbol{h}_1, \boldsymbol{h}_2) = C\left(\rho |\boldsymbol{h}_1^H \boldsymbol{w}_1|^2 \right), \tag{5}$$

$$R_2^{DPC}(\pi_2, \rho, \boldsymbol{h}_1, \boldsymbol{h}_2) = C\left(\frac{\rho |\boldsymbol{h}_2^H \boldsymbol{w}_2|^2}{1 + \rho |\boldsymbol{h}_2^H \boldsymbol{w}_1|^2} \right), \tag{6}$$

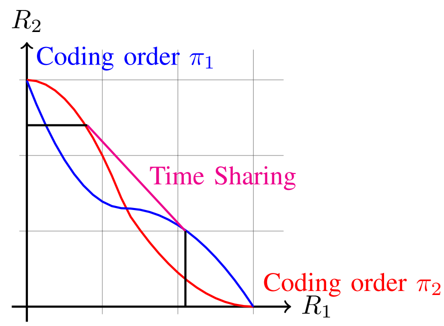

Fig. 1. Capacity region of MISO downlink channel. Coding order π_2 in red colour, coding order π_1 in blue color, and the convex hull, i.e., the maximum sum rate line in magenta. Furthermore, the region in which all rate points are achieved by maximizing the sum rate are indicated with black lines.

These two achievable rate regions in (3), (4) and (5),(6) correspond to the red and the blue curves in Figure 1. The convex closure operator corresponds to the magenta line in Figure 1. The complete capacity region are all points under the union of the red, magenta, and blue curves. Note that all points below the boundary are indicated by the black lines and the sum rate line and they can be achieved by distributing the sum rate to the users.

B. Non-Orthogonal Multiple Access

In the considered two-user NOMA, one user tries to fully decode the message of the other user, subtract it from its received signal and then decode its own message without interference. The requirement for this user to be able to decode the other users' message results in an additional rate constraint for the other users' data rate.

In single-antenna NOMA, the optimal decoding order simply depends on the received channel-to-noise ratio (CNR). The user with the better channel gain is able to decode the other users' signal independent of the power allocation and the corresponding rate upper-bound constraint is automatically fulfilled. The single-antenna broadcast channel is degraded and this strategy corresponds to the well-known superposition and successive interference cancellation (SIC) which achieves the capacity of the degraded broadcast channel [2, Theorem 5.3].

However, multiple-antenna BC are non-degraded and NOMA does neither achieve the capacity region nor is the optimal decoding order as simple as in the single-antenna case. Therefore, it is not sufficient to consider only the order based on the channel gains $||\boldsymbol{h}_1|| \geq ||\boldsymbol{h}_2||$. Instead, both NOMA orders need to be considered and the following achievable rates:

$$\begin{aligned} R_1^{N,\pi_2} &= C(\rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2) \quad (7)\\ R_2^{N,\pi_2} &= \min\left\{C\left(\frac{\rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2}{1+\rho|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2}\right), C\left(\frac{\rho|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2}{1+\rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2}\right)\right\}\\ R_1^{N,\pi_1} &= \min\left\{C\left(\frac{\rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2}{1+\rho|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2}\right), C\left(\frac{\rho|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2}{1+\rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2}\right)\right\}\\ R_2^{N,\pi_1} &= C(\rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2) \quad (8)\end{aligned}$$

The complete achievable rate region using downlink NOMA is given by

$$\mathcal{R}^N = \text{cv}\Big\{\bigcup_{\substack{i\in\{1,2\}\\ ||\boldsymbol{w}_1||^2+||\boldsymbol{w}_2||^2\le P}} (R_1^{N,\pi_i}, R_2^{N,\pi_i}\Big\}. \quad (9)$$

Therefore, the achievable maximum sum rate R^N_{sum} of NOMA can be written as

$$\begin{aligned}\max \quad \Bigg\{ &\min\Big\{C(\rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2) - C(\rho|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2) + \\ &C(\rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2 + \rho|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2),\\ &C(\rho|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2 + \rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2)\Big\},\\ &\min\Big\{C(\rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2) - C(\rho|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2) + \\ &C(\rho|\boldsymbol{h}_2^H\boldsymbol{w}_2|^2 + \rho|\boldsymbol{h}_2^H\boldsymbol{w}_1|^2),\\ &C(\rho|\boldsymbol{h}_1^H\boldsymbol{w}_2|^2 + \rho|\boldsymbol{h}_1^H\boldsymbol{w}_1|^2)\Big\}\Bigg\}. \quad (10)\end{aligned}$$

C. Treating Interference as Noise

TIN is a simple strategy where point-to-point codes are reused for smaller SINR values where interference is treated as noise [2, Section 6.4.3]. For small SNR it is sum capacity achieving in the interference channel. The achievable rates are easily computed by

$$R_i^{\text{TIN}} = C\left(\frac{\rho|\boldsymbol{h}_i^H\boldsymbol{w}_i|^2}{1+\rho|\boldsymbol{h}_i^H\boldsymbol{w}_j|^2}\right), \quad 1\le i\ne j\le 2, \quad (11)$$

with corresponding sum rate $R^{\text{TIN}}_{\text{sum}} = R^{\text{TIN}}_1 + R^{\text{TIN}}_2$. Within TIN, the zero-forcing beamforming solution is a special case with

$$\boldsymbol{w}_i^{\text{ZF}} = \frac{\Pi^{\perp}_{\boldsymbol{h}_j}\boldsymbol{h}_i}{||\Pi^{\perp}_{\boldsymbol{h}_j}||}, \qquad 1\le i\ne j\le 2.$$

The projector onto the orthogonal complement of a space spanned by $\boldsymbol{x}$ is denoted by $\Pi^{\perp}_{\boldsymbol{x}}$.

D. Uplink-Downlink Duality

The sum rate maximization problem in the multiple antenna BC can be solved by considering the corresponding uplink problem [15, Lemma 2]

$$R^{\text{DPC}}_{sum} = \max_{\substack{p_1,p_2\ge 0\\ p_1+p_2\le 1}} \log\det\left(\boldsymbol{I} + \rho p_1\boldsymbol{H}_1 + \rho p_2\boldsymbol{H}_2\right), \quad (12)$$

with channel matrices $\boldsymbol{H}_1 = \boldsymbol{h}_1\boldsymbol{h}_1^H$ and $\boldsymbol{H}_2 = \boldsymbol{h}_2\boldsymbol{h}_2^H$. Problem (12) is a convex programming problem which can be easily solved with CVXPY [16], [17]. The corresponding maximum sum rate points for the two decoding orders in the dual MAC can be explicitly computed with the optimal power allocation $\boldsymbol{p}^*$ as

$$\begin{aligned} R_1^{sc}(\pi_2) &= \log\det\left(\boldsymbol{I} + \rho p_1^*\boldsymbol{H}_1\right) \quad (13)\\ R_2^{sc}(\pi_2) &= \log\det\left(\boldsymbol{I} + \rho p_2^*\boldsymbol{H}_2\left[\boldsymbol{I} + \rho p_1\boldsymbol{H}_1\right]^{-1}\right) \quad (14)\end{aligned}$$

and vice versa for the decoding order π_1.

Furthermore in [18], it is shown that each point in the capacity region of the dual multiple antenna MAC under a sum power constraint can be achieved by the dual multiple antenna BC. Interestingly, the SIC decoding order is the reverse DPC coding order. And there exists a one-to-one mapping between the strategies that achieve the rate points.

III. Optimality of NOMA and Power Minimization

In this section, we present our main results regarding the optimality of the NOMA and TIN schemes compared to the capacity achieving DPC with time-sharing. We start with the sum rate maximization under power constraints and then present some conclusions for the power minimization under rate constraints.

A. Sum-Capacity Sub-Optimality of NOMA

Theorem 1. *In the two-user multiple antenna downlink transmission, NOMA cannot achieve the sum capacity in general. Only if both the channels are parallel, i.e., $\boldsymbol{h}_1 = \alpha\boldsymbol{h}_2$, $\alpha\in\mathbb{C}$, then NOMA is able to achieve the sum capacity.*

Proof. The proof is based on some auxiliary results collected in the appendix. First, we note that the sum capacity of the MISO BC can be written by applying the up- and downlink duality as in (12). While the optimization of the NOMA sum-rate in (10) with respect to the beamforming vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$ is difficult, an upper bound on the sum rate can be written as

$$R^N_{sum} \le C(\rho p_1||\boldsymbol{h}_1||^2 + \rho p_2||\boldsymbol{h}_2||^2), \quad (15)$$

where the bound is achieved with equality with $\boldsymbol{w}_1 = \frac{p_1}{||\boldsymbol{h}_1||^2}\boldsymbol{h}_1$ and $\boldsymbol{w}_2 = \frac{p_2}{||\boldsymbol{h}_2||^2}$ if $\boldsymbol{h}_1 = \alpha\boldsymbol{h}_2$ for some $\alpha\in\mathbb{C}$. The final step is to show that for $p_1, p_2 \ge 0$ (we set $\rho = 1$ for convenience) and arbitrary $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{C}^n$ holds

$$\det\left(\boldsymbol{I} + p_1\boldsymbol{H}_1 + p_2\boldsymbol{H}_2\right) \ge (1 + p_1||\boldsymbol{h}_1||^2 + p_2||\boldsymbol{h}_2||^2), \quad (16)$$

with equality if and only if $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are parallel. We start with the left-hand-side of the inequality and identify $\boldsymbol{A} = \boldsymbol{I} + p_1\boldsymbol{h}_1\boldsymbol{h}_1^H$ and apply the equality (18). In order to compute $\boldsymbol{A}^{-1}$ we apply the equality (19):

$$\boldsymbol{A}^{-1} = \left[\boldsymbol{I} + p_1\boldsymbol{h}_1\boldsymbol{h}_1^H\right]^{-1} = \boldsymbol{I} - \frac{p_1\boldsymbol{h}_1\boldsymbol{h}_1^H}{1 + p_1||\boldsymbol{h}_1||^2}.$$

Using this $\boldsymbol{A}^{-1}$ in (18), we get

$$\begin{aligned}
& \det(\boldsymbol{A} + p_2 \boldsymbol{h}_2 \boldsymbol{h}_2^H) \\
= & \left(1 + p_2 \boldsymbol{h}_2^H \left[\boldsymbol{I} - \frac{p_1 \boldsymbol{h}_1 \boldsymbol{h}_1^H}{1 + p_1 ||\boldsymbol{h}_1||^2}\right] \boldsymbol{h}_2\right)(1 + p_1||\boldsymbol{h}_1||^2) \\
= & \left(1 + p_2 \boldsymbol{h}_2 \boldsymbol{h}_2^H - \frac{p_1 p_2 |\boldsymbol{h}_1^H \boldsymbol{h}_2|^2}{1 + p_1||\boldsymbol{h}_1||^2}\right)(1 + p_1||\boldsymbol{h}_1||^2) \\
= & \Big(1 + p_2||\boldsymbol{h}_2||^2 + p_1||\boldsymbol{h}_1||^2 \\
& + p_1 p_2 (||\boldsymbol{h}_1||^2 ||\boldsymbol{h}_2||^2 - |\boldsymbol{h}_1^H \boldsymbol{h}_2|^2\Big) \\
\geq & \ (1 + p_2||\boldsymbol{h}_2||^2 + p_1||\boldsymbol{h}_1||^2), \qquad (17)
\end{aligned}$$

since $||\boldsymbol{h}_1||^2||\boldsymbol{h}_2||^2 \geq |\boldsymbol{h}_1^H \boldsymbol{h}_2|^2$ with equality if and only if $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are parallel. □

Note that the maximum of the RHS of (15) is achieved for allocating all power to the better channel, i.e., $C(\rho P \max(||\boldsymbol{h}_1||^2, ||\boldsymbol{h}_2||^2))$.

B. Achievable Rate Regions

In this section, we compare the achievable rate regions by DPC, NOMA, and ZF, including their convex hulls operation (time sharing), and by TIN. In order to obtain an efficient algorithm, we exploit the uplink-downlink duality again. While ZF has a simple closed-form solution

$$p_i^{ZF} = \frac{2^{R_i} - 1}{|\boldsymbol{h}_i \boldsymbol{w}_i^{ZF}|^2}, \qquad i = \{1, 2\},$$

and for NOMA a semidefinite program (SDP) can be found in [9], we propose the following algorithm to find the minimum power for DPC.

Data: Rate requirements R_1, R_2, channels $\boldsymbol{H}_1, \boldsymbol{H}_2$
Result: Minimum power required by DPC
Solve $\min_{p_1,p_2} p_1 + p_2$ s.t.
$\log\det(\boldsymbol{I} + p_1 \boldsymbol{H}_1 + p_2 \boldsymbol{H}_2) \geq R_1 + R_2$;
Denote found solution by p_1^*, p_2^* ;
Define $\bar{R}_1 = \log\det(\boldsymbol{I} + p_1^* \boldsymbol{H}_1)$ and $\bar{R}_2$ analogue;
if $\bar{R}_1 < R_1$ **then**
  Compute $p_1^* = \frac{2^{R_1}-1}{||\boldsymbol{h}_1||^2}$;
  Compute p_2^* s.t. $\log\det(\boldsymbol{I} + p_1^* \boldsymbol{H}_1 + p_2^* \boldsymbol{H}_2) \geq R_2 + \log\det(\boldsymbol{I} + p_1^* \boldsymbol{H}_1)$;
end
if $\bar{R}_2 < R_2$ **then**
  Compute $p_2^* = \frac{2^{R_2}-1}{||\boldsymbol{h}_2||^2}$;
  Compute p_1^* s.t. $\log\det(\boldsymbol{I} + p_1^* \boldsymbol{H}_1 + p_2^* \boldsymbol{H}_2) \geq R_1 + \log\det(\boldsymbol{I} + p_2^* \boldsymbol{H}_2)$;
end
return $p_1^* + p_2^*$;

Algorithm 1: Find minimum power with DPC.

Algorithm 1 first assumes that the rate requirements lie below the sum rate curve within the marked area in Figure (1) and computes the minimum sum power such that the sum rate is larger than $R_1 + R_2$. Next, it checks the two black horizontal and vertical lines. If the rate requirement is outside, the corresponding Pareto boundary is searched for in the two if clauses.

Note that Algorithm 1 is not iterative but consists of solving a simple convex programming problem and two if clauses, followed eventually by a closed form power allocation.

IV. Numerical Illustrations

A. Sum Rate Comparison

An example of the sum rate comparison between the capacity achieving DPC, the NOMA upper bound from (15), and TIN is shown in Figure 2.

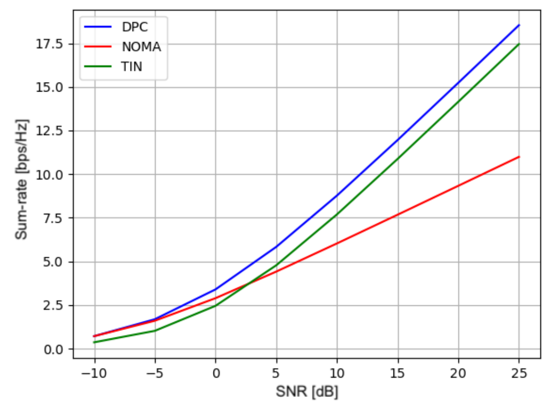

Fig. 2. Sum rate comparison between DPC, NOMA, and TIN for downlink transmission with three transmit antennas.

We can observe that the high-SNR slope of the sum rate of TIN and DPC are equal to 2 with a high-SNR power offset of about 2 dB, while the high-SNR slope of NOMA is 1. The reason for this behavior is shown in the inequality (16). The spatial multiplexing gain is visible on the LHS of (16) while it is lost on the RHS of (16).

B. Achievable Rate Regions

Next, we perform numerical experiments with the power minimization problem and compare the minimum power achieved with DPC, NOMA and TIN using ZF. We assume the same rate requirements for both users and the value on the $x-axis$ correspond to the rate requirement. 1000 channel realizations are randomly generated according to a Rayleigh distribution to compute the average minimum power.

The results in Figure 3 show that DPC significantly outperforms the two other schemes NOMA and TIN. Depending on the number of antennas, ZF outperforms NOMA for more transmit antennas, because of the spatial degrees of freedom. For higher rate requirements, ZF will also outperform NOMA. This illustrates that depending on the number of antennas, the rate requirements and the channel realizations, NOMA is outperformed by TIN with ZF or not.

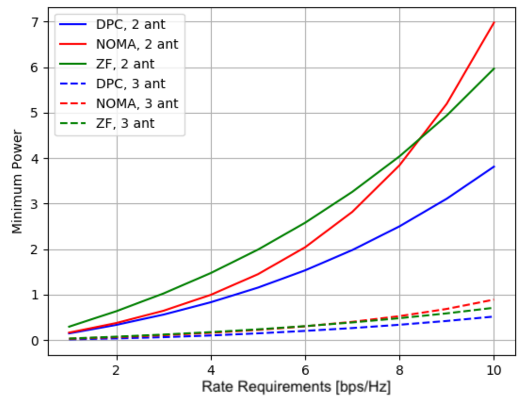

Fig. 3. Minimum power comparison between DPC, NOMA, and TIN for downlink transmission with two and three transmit antennas.

V. Conclusions and Future Work

In order to choose multiple access schemes for the multiple antenna downlink transmission, we have considered the capacity achieving scheme DPC including time sharing. The first important observation is that time sharing is required to achieve the complete capacity region for the multiple-antenna downlink channel. This is in stark contrast to single-antenna downlink transmission where linear superposition coding combined with simple SIC according to the order of channel gains, can achieve the capacity. This scheme corresponds to NOMA. In the multiple antenna case with two users, we show that NOMA cannot achieve the sum capacity expect for parallel channel realizations. Furthermore, we propose a fast algorithm to compute the minimum power to support rate requirements for the capacity region, and compare the results to NOMA and TIN with ZF. The results indicate that in multiple antenna channels, NOMA usually does not achieve the minimum power and the loss compared to the capacity achieving scheme grows with number of antennas and with higher rate requirements. The extension of the presented results to more than two users is straightforward and our current work.

Appendix

Here, we review two lemmas useful for the proof of our main results. The following lemma can be found e.g. in [19, Lemma 1.1]

Lemma 1. *If $\boldsymbol{A}$ is an invertible $n \times n$ matrix, and $\boldsymbol{v}$ is an n-dimensional column vector, then*

$$\det\left(\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^H\right) = \left(1 + \boldsymbol{v}^H \boldsymbol{A}^{-1}\boldsymbol{v}\right)\det(\boldsymbol{A}). \tag{18}$$

The other results that we apply is a simple variant of the the Sherman-Morrison-Woodbury-formula [20, Section 0.7.4]:

Lemma 2. *For any column vector $\boldsymbol{v}$*

$$\left(\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}^H\right)^{-1} = \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^H}{1 + ||\boldsymbol{v}||^2}. \tag{19}$$

References

[1] M. Vaezi and H. Vincent Poor, “NOMA: an information-theoretic perspective,” in *Multiple Access Techniques for 5G Wireless Networks and Beyond*, M. Vaezi, Z. Ding, and H. V. Poor, Eds. Cham: Springer International Publishing, 2019, pp. 167–193.

[2] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[3] Z. Ding, F. Adachi, and H. V. Poor, “The application of MIMO to non-orthogonal multiple access,” *IEEE Transactions on Wireless Communications*, vol. 15, pp. 537–552, 1 2016.

[4] Z. Ding, R. Schober, and H. V. Poor, “A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment,” *IEEE Transactions on Wireless Communications*, vol. 15, pp. 4438–4454, 6 2016.

[5] Z. Chen, Z. Ding, P. Xu, and X. Dai, “Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink,” *IEEE Communications Letters*, vol. 20, pp. 1263–1266, 6 2016.

[6] Z. Chen, Z. Ding, P. Xu, X. Dai, J. Xu, and D. W. K. Ng, “Comment on “optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink”,” *IEEE Communications Letters*, vol. 21, pp. 2109–2111, 9 2017.

[7] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, “On the design of multiple-antenna non-orthogonal multiple access,” in *Multiple Access Techniques for 5G Wireless Networks and Beyond*, M. Vaezi, Z. Ding, and H. V. Poor, Eds. Cham: Springer International Publishing, 2019, pp. 229–256.

[8] F.-Y. Tian and X.-M. Chen, “Multiple-antenna techniques in nonorthogonal multiple access: A review,” *Frontiers in Information Technology & Electronic Engineering*, vol. 20, pp. 1665–1697, 2019.

[9] J. Zhu, J. Wang, Y. Huang, K. Navaie, Z. Ding, and L. Yang, “On optimal beamforming design for downlink MISO NOMA systems,” *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 3008–3020, 3 2020.

[10] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, “A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems,” *IEEE Transactions on Signal Processing*, vol. 64, pp. 76–88, 1 2016.

[11] S. Ali, E. Hossain, and D. I. Kim, “Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation,” *IEEE Access*, vol. 5, pp. 565–577, 2017.

[12] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, “Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis,” *IEEE Wireless Communications Letters*, vol. 9, pp. 349–353, 3 2020.

[13] B. Clerckx, Y. Mao, R. Schober, *et al.*, “Is NOMA efficient in multi-antenna networks? a critical look at next generation multiple access techniques,” Jan. 2021. arXiv: 2101.04802 `[cs.IT]`.

[14] H. Weingarten, Y. Steinberg, and S. Shamai, “The capacity region of the Gaussian multiple-input multiple-output broadcast channel,” *IEEE Transactions on Information Theory*, vol. 52, pp. 3936–3964, 9 2006.

[15] P. Viswanath and D. N. C. Tse, “Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality,” *IEEE Trans. on Information Theory*, vol. 49, pp. 1912–1921, Aug. 2003.

[16] S. Diamond and S. Boyd, “CVXPY: A Python-embedded modeling language for convex optimization,” *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[17] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, “A rewriting system for convex optimization problems,” *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[18] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[19] J. Ding and A. Zhou, “Eigenvalues of rank-one updated matrices with some applications,” *Applied Mathematics Letters*, vol. 20, no. 12, pp. 1223–1226, 2007.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd. Cambridge University Press, 2013.