NOMA - A Paradigm Shift for Multiple Access for 5G and Beyond

Zhiguo Ding and Wei Liang

A Tutorial at International School on Toward Intelligent World

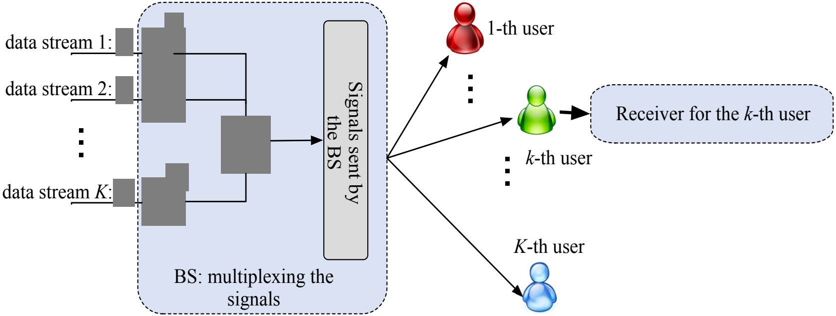
28 Aug 2018

Outline

Outline

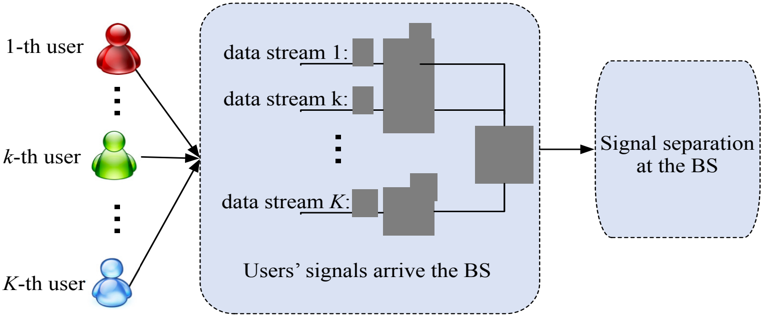
Non-orthogonal Multiple Access (NOMA)

- What is multiple access (MA)?
  - Techniques to serve multiple users with limited bandwidth.
  - An example for downlink multiple access



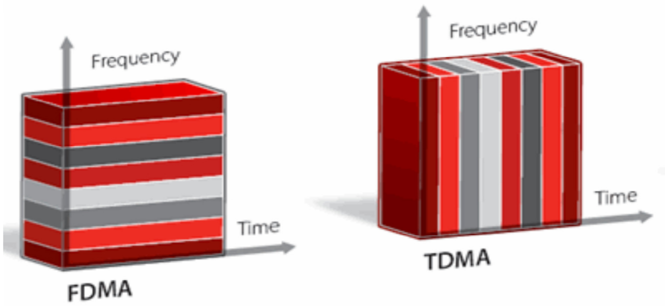
Non-orthogonal Multiple Access (NOMA)

- What is multiple access (MA)?
  - Techniques to serve multiple users with limited bandwidth.
  - An example for uplink multiple access



Non-orthogonal Multiple Access (NOMA)

- What kind of multiple access techniques have been used?
  - We have been using orthogonal multiple access (OMA).
  - TDMA: Orthogonal (non-overlapping) time slots are allocated to users.
  - FDMA: Orthogonal (non-overlapping) frequency channels are allocated to users.
  - ...



Non-orthogonal Multiple Access (NOMA) - (1/2)

Disadvantages of OMA

- Dilemma to realize a better trade-off between throughput and user fairness, illustrated in the following example:
    - A user with a poor connection to the base station (BS) is served by using OMA.
    - Spectral efficiency is low since this user cannot utilize the allocated bandwidth efficiently.
    - Since OMA is used, the bandwidth resources occupied by this user cannot be shared by the others.
- Difficult to support massive connectivity
    - Recall that the three key requirements for 5G are to support high throughput, low latency and massive connectivity

Non-orthogonal Multiple Access (NOMA) - (2/2)

- A promising solution is to break orthogonality $\rightarrow$ NOMA
  - The key idea of NOMA is to encourage spectrum sharing
  - Details for the advantages of NOMA are to be given in the remaining of this tutorial.
- NOMA is gaining ground on the competition of multiple access techniques for the next generation wireless networks
  - Adopted by many 5G MA concepts, including power-domain (PD) NOMA, sparse code multiple access (SCMA), multi-user sharing access (MUSA), pattern division multiple access (PDMA), lattice partition multiple access (LPMA), etc.
  - Used by 4G LTE-A, termed multi-user superposition transmission (MUST - to be discussed later).
  - Included in the forthcoming digital TV standard (ATSC 3.0).

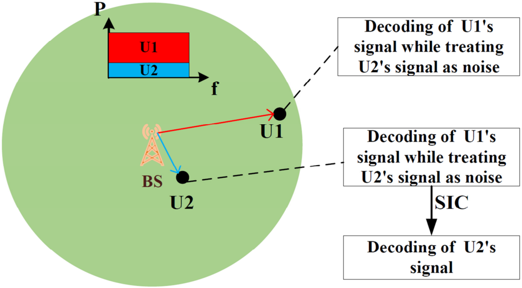
Outline

Two Forms of Single-Carrier NOMA

- In the following, we will focus on the case when the NOMA principle is applied to a single orthogonal resource block
    - This resource block may represent a single OFDMA subcarrier, a time slot, etc.
    - The use of a single carrier will be used as an example, given the popularity of OFDMA.
- With only a single carrier, the principle of NOMA can be implemented in the following two versions:
    - Power-domain NOMA.
    - Cognitive radio inspired NOMA.

Outline

Power Domain NOMA (1/5)



- All the users are served at the same time, frequency and code, but with different power levels.
- Users with better channel conditions get less power.
- Successive interference cancellation (SIC) is used.

[1]Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System level performance evaluation of downlink non-orthogonal multiple access (NOMA)", in PIMRC 2013.
[2]Z. Ding, Z. Yang, P. Fan and H. V. Poor, "On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users", IEEE SPL, 2014.

Power Domain NOMA (2/5)

For the example shown in the previous figure

- Denote the message to user i by s_i, its channel by h_i and its power allocation coefficient by α_i.
- Assume $|h_1| \leq |h_2|$, which means $\alpha_1 \geq \alpha_2$
- The base station sends a superimposed message, $\alpha_1 s_1 + \alpha_2 s_2$
- User i observes $y_i = h_i(\alpha_1 s_1 + \alpha_2 s_2) + n_i$, where n_i is noise.
- User 1 decodes its message directly with the following rate, $\log_2\left(1 + \frac{|h_1|^2\alpha_1^2}{|h_1|^2\alpha_2^2 + \frac{1}{\rho}}\right)$, where ρ is the transmit SNR.
- After SIC, user 2's rate is $\log_2\left(1 + \rho|h_2|^2\alpha_2^2\right)$, since $|h_1| \leq |h_2|$ and hence $\log_2\left(1 + \frac{|h_1|^2\alpha_1^2}{|h_1|^2\alpha_2^2 + \frac{1}{\rho}}\right) \leq \log_2\left(1 + \frac{|h_2|^2\alpha_1^2}{|h_2|^2\alpha_2^2 + \frac{1}{\rho}}\right)$

Power Domain NOMA (3/5)

Recall, with NOMA, user 1's achievable rate is $\log_2\left(1+\frac{|h_1|^2\alpha_1^2}{|h_1|^2\alpha_2^2+\frac{1}{\rho}}\right)$, and user 2's rate is $\log_2\left(1+\rho|h_2|^2\alpha_2^2\right)$.

- Example 1:
  - Consider $\rho \to \infty$ which implies a high SNR scenario, and $\rho|h_1|^2 \to 0$ which implies that user 1's channel experiences a deep fade.
  - The sum rate of NOMA becomes

$$\log_2\left(1+\frac{\alpha_1^2}{\alpha_2^2}\right)+\log_2\left(\rho|h_2|^2\alpha_2^2\right)=\log_2(\rho|h_2|^2).$$

  - The sum rate of OMA is

$$\frac{1}{2}\log_2\left(1+\rho|h_1|^2\right)+\frac{1}{2}\log_2\left(1+\rho|h_2|^2\right)\approx\frac{1}{2}\log_2\left(\rho|h_2|^2\right).$$

  - The performance gain of NOMA over OMA is obvious.

Power Domain NOMA (4/5)

- Example 2:
  - Assume that user 1 is an IoT device requiring only a low data rate and user 2 is a user demanding a high data rate.
  - When OFDMA is used, which is a typical example of OMA, each user is allocated a separate subcarrier.
  - In this example, the spectral efficiency of OMA is poor since the IoT device is served with more bandwidth than what it actually needs, while the broadband user is not assigned enough bandwidth.
  - On the other hand, the use of NOMA encourages spectrum sharing, i.e., the broadband user can also have access to the subcarrier occupied by the IoT device.
  - As a result, the use of NOMA efficiently supports massive connectivity and meets the users' diverse QoS requirements.

[3] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things,", IEEE Access, 2016.

Power Domain NOMA (5/5)

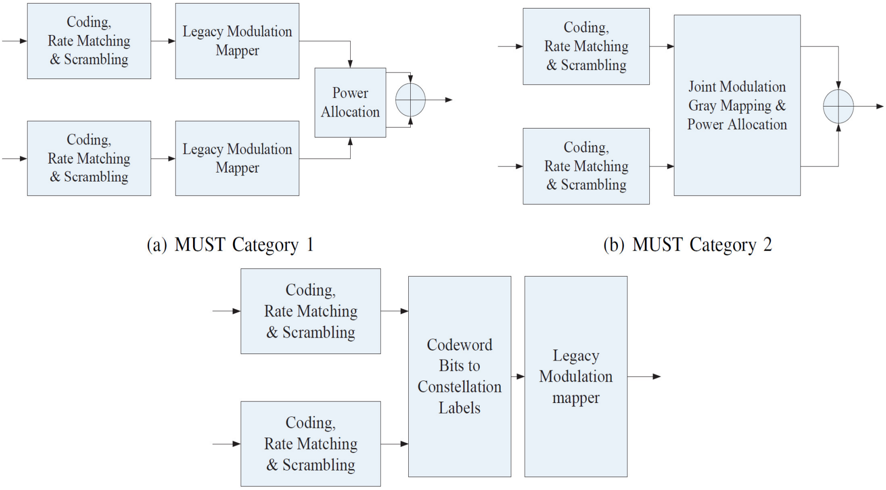
- Can the use of optimal resource allocation for OMA overcome the above described disadvantage?
- Consider the following OMA schemes
  - OMA-TYPE-I: optimal power allocation among the frequency channels, but the width of these channels is fixed.
  - OMA-TYPE-II: optimal power allocation and the width of the frequency channels can be optimally changed.
- Using tools from optimization theory, it can be rigorously proved that NOMA always outperforms OMA or at least achieves the same performance as OMA.
- Note that adaptive resource allocation for OMA introduces dynamic changes to the properties of the resource blocks
  - Frequency channels with very small widths might be needed, which might not be realistic in practice.

[4] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "A mathematical proof of the superiority of NOMA compared to conventional OMA", arXiv:1612.01069

Application of Power-Doman NOMA in 4G and 5G (1/2)

- Included in various whitepapers for 5G (DOCOMO, METIS, NGMN, ZTE, SK Telecom, etc.)
- Recently proposed to 3GPP-LTE (MUST)
  - At the 3GPP meeting in May 2015, it was decided to include MUST into LTE Advanced
  - At the 3GPP meeting in August 2015, 15 forms of MUST have been proposed by Huawei, Qualcomm, NTT DOCOMO, Nokia, Intel, LG Electronics, Samsung, ZTE, Alcatel Lucent, etc
  - For example, Huawei proposed three forms of NOMA:
    - Non-Orthogonal Multiple Access (NOMA)
    - Semi-Orthogonal Multiple Access (SOMA)
    - Rate-adaptive constellation Expansion Multiple Access (REMA)
  - At the 3GPP meeting in December 2015, NOMA has been included into LTE Release 13.

Application of Power-Doman NOMA in 4G and 5G (2/2)



(a) MUST Category 1

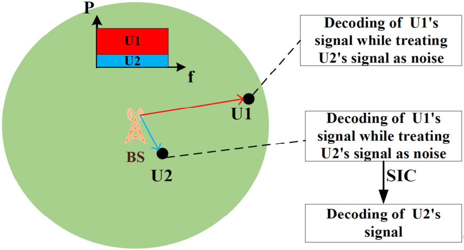
(b) MUST Category 2

(c) MUST Category 3

Outline

Synergy Between Cognitive Radio and NOMA (1/2)

- Conventional power-domain NOMA allocates more power to the user with poor channel conditions to ensures user fairness.
- But power domain NOMA cannot strictly guarantee the users' QoS targets.
- CR-NOMA can strictly guarantee the users' QoS requirements by using the fact that NOMA can be viewed as a special case of CR networks.
- For example, consider the following two-user scenario:



Synergy Between Cognitive Radio and NOMA (2/2)

- User 1 can be viewed as a primary user in a CR network:
  - If OMA is used, the orthogonal bandwidth allocated to user 1 cannot be accessed by other users.
  - Spectral efficiency is low since user 1 has a poor connection to the BS.
- The use of NOMA is equivalent to the application of the cognitive radio concept:
  - Specifically user 2, a user with better channel conditions, is admitted to the channel occupied by user 1.
  - Although user 1 causes extra interference at user 2 and hence reduces user 2's rate, the overall system throughput will be increased significantly since user 1 has a stronger connection to the BS.

[5] Z. Ding, P. Fan and H. V. Poor, "Impact of User Pairing on 5G Non-Orthogonal Multiple Access Downlink Transmissions", IEEE TVT, 2016.

Cognitive Radio Inspired NOMA - System Model

- Consider a downlink communication scenario with one BS and M single-antenna users.
- All the users share the same bandwidth resources, such as time slots, spreading codes, and subcarrier channels.
- Without loss of generality, assume that the users' channels have been ordered as $|h_1|^2 \leq \cdots \leq |h_M|^2$.
- Consider the situation in which the m-th user and the n-th user, $m < n$, are paired to perform NOMA.
  - Note that NOMA is implemented in LTE for user pairs.
- The insights obtained for the case with two selected users can be used for the design of dynamic user pairing approaches
  - A game theoretic approach for user pairing can be designed.

Cognitive Radio Inspired NOMA - Form I (1/3)

- Consider that user m is the primary user and user n is the secondary user.
- The rates achievable by the users are given by

$$R_m = \log\left(1 + \frac{|h_m|^2 a_m^2}{|h_m|^2 a_n^2 + \frac{1}{\rho}}\right), \tag{1}$$

and

$$R_n = \log\left(1 + \rho a_n^2 |h_n|^2\right), \tag{2}$$

respectively.
- Note user n can decode user m's signal if R_m is used.
- The rates of OMA, $\bar{R}_i$ for $i \in \{m, n\}$, are

$$\bar{R}_i = \frac{1}{2}\log\left(1 + \rho|h_i|^2\right). \tag{3}$$

Cognitive Radio Inspired NOMA - Form I (2/3)

- Consider that the targeted SINR at the m-th user is I, i.e., its target rate is $\log(1+I)$
- The choices of the power allocation coefficients, a_m and a_n, need to satisfy the following constraint:

$$\frac{|h_m|^2 a_m^2}{|h_m|^2 a_n^2 + \frac{1}{\rho}} \geq I. \tag{4}$$

- This implies that the maximal transmit power that can be allocated to the n-th user is given by

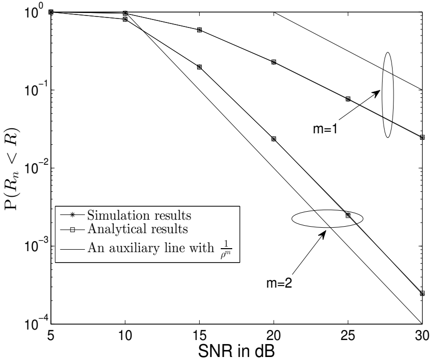
$$a_n^2 = \max\left\{0, \frac{|h_m|^2 - \frac{I}{\rho}}{|h_m|^2(1+I)}\right\}. \tag{5}$$

- Note that the choice of a_n in (5) is a function of the channel coefficient h_m, unlike the constant choice of a_n used by PD-NOMA in the previous section.

Cognitive Radio Inspired NOMA - Form I (3/3)

- With this cognitive radio inspired power allocation policy, the primary user (user m) experiences the same outage performance as with OMA
  - If user m's targeted rate cannot be supported, outage occurs in both the OMA and NOMA modes.
  - If user m's targeted rate can be supported, no outage occurs in the OMA mode, and in NOMA, sufficient power is given to user m to avoid any outage, before serving user n.
- What happens to the diversity gain of user n?
  - The diversity gain achieved by user n is m.
  - For example, if the best user, i.e., user M, is paired with the worst user, i.e., user 1, the diversity gain at the best user is 1, instead of M.
- This is expected since the chance of user n for being served is decided by user m's channel conditions.

Cognitive Radio NOMA - Form I: Simulations



The outage probability of the cognitive user. $M = 5$ and $n = M$.

- There are 5 users, where the user with the strongest channel is selected as the cognitive user.
- $m = 1$ ($m = 2$) corresponds to choosing the user with the weakest (second weakest) channel as the primary user.
- The curves with $\frac{1}{\rho^m}$ illustrate the diversity gain of m.
- As expected, the primary user's channel affects the cognitive user's diversity.

Cognitive Radio Inspired NOMA - Form II (1/3)

- Recall that we focus on a downlink NOMA system with M single-antenna users and one single-antenna base station.
- Without loss of generality, assume the users are ordered as

$$|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2,$$

where h_i is the Rayleigh fading channel gain.
- The m-th user and the n-th user are selected to perform NOMA, $1 \leq m < n \leq M$.
- The key idea of this form is to meet both of the users' QoS requirements.

[6] Z. Yang, Z. Ding, P. Fan and N. Al-Dhahir, "A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA", IEEE TWC, 2016.

Cognitive Radio Inspired NOMA - Form II (2/3)

- The rates of user m and user n in downlink NOMA are given by

$$R_{m,D}^{N} = \log_2\left(1 + \frac{\alpha_m |h_m|^2}{\alpha_n |h_m|^2 + 1/\rho}\right), \tag{6}$$

and

$$R_{n,D}^{N} = \log_2\left(1 + \alpha_n \rho |h_n|^2\right). \tag{7}$$

- The rate of user i in OMA, is given by

$$R_i^T = \frac{1}{2}\log_2\left(1 + \rho |h_i|^2\right), \; i \in \{m, n\}. \tag{8}$$

Cognitive Radio Inspired NOMA - Form II (3/3)

- Assume $R_{m,D}^N \geq R_m^T$, which results in the following constraint:

$$\log_2\left(1+\alpha_m^2\rho|h_m|^2\right) \geq \frac{1}{2}\log_2\left(1+\rho|h_m|^2\right) \Rightarrow \alpha_m^2 \geq \frac{1}{\sqrt{1+\rho|h_m|^2}+1}.$$

- Assume $R_{n,D}^N \geq R_n^T$, then

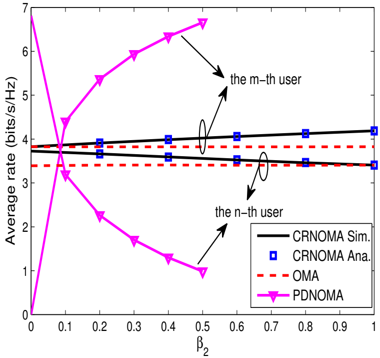
$$\log_2\left(1+\frac{\alpha_n^2|h_n|^2}{\alpha_m^2|h_n|^2+1/\rho}\right) \geq \frac{1}{2}\log_2\left(1+\rho|h_n|^2\right) \Rightarrow \alpha_m^2 \leq \frac{1}{\sqrt{1+\rho|h_n|^2}+1}$$

- Combining the two constraints, α_m^2 can be expressed as follows:

$$\alpha_m^2 = \frac{\beta_1}{W+1} + \frac{\beta_2}{V+1}, \tag{9}$$

where $W = \sqrt{1+\rho|h_m|^2}$, $V = \sqrt{1+\rho|h_n|^2}$, $0 \leq \beta_i \leq 1$, $i = 1, 2$, and $\beta_1 + \beta_2 = 1$.

Cognitive Radio NOMA - Form II: Simulations



The average rate for user n and user m in downlink CR-NOMA systems with $SNR = 20$ dB, $M = 5$, $m = 5$, and $n = 4$.

- As expected, the rate of user m increases with increasing β_2, and the rate of user n decreases with increasing β_2.
- Compare to PD-NOMA, CRNOMA offers better user fairness, since the ranges of the rates become smaller.
- Further optimization of the power allocation coefficients can increase the gap between OMA and NOMA.

Outline

Motivation and Introduction

- Why adopt hybrid NOMA?
  - Asking all users to participate in NOMA can cause significant complexity
  - This motivates the hybrid NOMA scheme
    - Users in a cell are divided into small groups
    - OMA is used to avoid inter-group interference
    - NOMA is implemented for the users within a single group
  - Who is to be grouped with whom?
    - A worst choice is to pair two users who have the same channel gains
    - Consider $h_1 = h_2$.
    - The sum rate of NOMA becomes $\log\left(1+\rho\alpha_1|h_1|^2\right)+\log\left(1+\frac{\rho\alpha_2|h_2|^2}{\rho\alpha_1|h_2|^2+1}\right)=\log(1+\rho|h_2|^2)$
    - The sum rate of OMA is $\sum_{i=1}^{2}\frac{1}{2}\log\left(1+\rho|h_i|^2\right)=\log\left(1+\rho|h_2|^2\right)$

[5] Z. Ding, P. Fan and H. V. Poor, "Impact of User Pairing on 5G Non-Orthogonal Multiple Access Downlink Transmissions", IEEE TVT, 2016.

Hybrid NOMA - Impact of User Pairing

- Consider a downlink communication scenario with one BS and M mobile users.
- All the users share the same bandwidth resources, such as time slots, spreading codes, and subcarrier channels.
- Without loss of generality, assume that the users' channels have been ordered as $|h_1|^2 \leq \cdots \leq |h_M|^2$.
- Consider the situation in which the m-th user and the n-th user, $m < n$, are paired to perform NOMA.
    - Note that in LTE a two-user version of NOMA is implemented.

Hybrid NOMA - Impact of User Pairing on the Sum Rate

Theorem 4.1

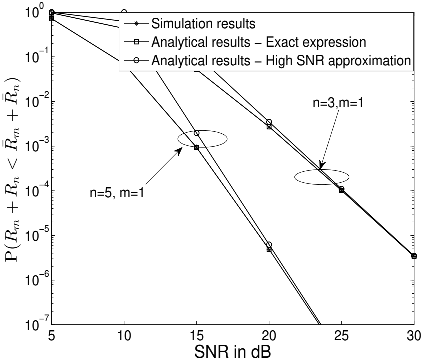
Suppose that the m-th and n-th ordered users are paired to perform NOMA. At high SNR, the probability that PD-NOMA achieves a lower sum rate than conventional MA is approximated as

$$\mathrm{P}(R_m + R_n < \bar{R}_m + \bar{R}_n) \approx \frac{1}{\rho^n}\left(\frac{\varpi_3 \varpi_2^n}{n} - \varpi_1 \varpi\right), \qquad (10)$$

where ϖ_i is a constant and not a function of ρ.

- This theorem shows that the probability $\mathrm{P}(R_m + R_n < \bar{R}_m + \bar{R}_n)$ decays at a rate of $\frac{1}{\rho^n}$.
- Therefore, to ensure a larger sum rate, n should be as large as possible, i.e., user n's channel conditions are critical to reduce $\mathrm{P}(R_m + R_n < \bar{R}_m + \bar{R}_n)$.

Impact of User Pairing on the Sum Rate - Simulations



The probability that NOMA realizes a lower sum rate than conventional MA. $M = 5$.

- There are $M = 5$ user and the user with the worst channel is always selected.
- As observed, scheduling the user with the best channel ($n = 5$) reduces the probability.
- Hence increasing the channel difference is helpful to ensure that NOMA outperforms OMA, in terms of sum rate.

Hybrid NOMA - Impact of User Pairing on Sum Rate Gap (1/3)

- In addition to the probability $\mathrm{P}(R_m + R_n < \bar{R}_m + \bar{R}_n)$, it is also of interest to study the sum rate outage probability

$$\mathrm{P}(R_m + R_n - \bar{R}_m - \bar{R}_n < R),$$

where R is a targeted performance gain.

- The probability studied previously can be viewed as a special case by setting $R = 0$.
- An interesting observation for the cases with $R > 0$ is that there will be an error floor for $\mathrm{P}(R_m + R_n - \bar{R}_m - \bar{R}_n < R)$, regardless of how large the SNR is, as shown in the following slide.

Hybrid NOMA - Impact of User Pairing on Sum Rate Gap (2/3)

- This can be shown by studying the following asymptotic expression of the sum rate gap:

$$
\begin{aligned}
& R_m + R_n - \bar{R}_m - \bar{R}_n \qquad (11)\\
& = \log\left(\frac{|h_m|^2 + \frac{1}{\rho}}{|h_m|^2 a_n^2 + \frac{1}{\rho}}\right) + \log\left(1 + \rho a_n^2 |h_n|^2\right) \\
& - \frac{1}{2}\log\left(1 + \rho|h_n|^2\right) - \frac{1}{2}\log\left(1 + \rho|h_m|^2\right) \\
& \underset{\rho\to\infty}{\to} \log\left(\frac{1}{a_n^2}\right) + \log\left(\rho a_n^2 |h_n|^2\right) - \log\left(\rho|h_m||h_n|\right) \\
& = \log|h_n| - \log|h_m|,
\end{aligned}
$$

which is not a function of SNR.

Hybrid NOMA - Impact of User Pairing on Sum Rate Gap (3/3)

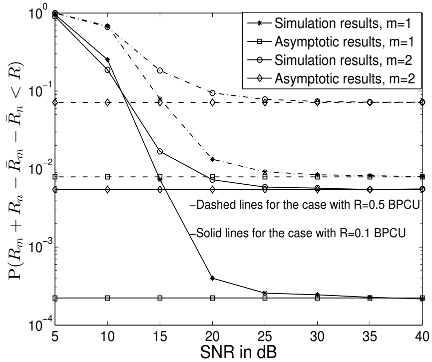
- Hence the probability can be expressed asymptotically as follows:

$$\mathrm{P}\left(R_m + R_n - \bar{R}_m - \bar{R}_n < R\right) \underset{\rho\to\infty}{\to} \mathrm{P}\left(\log|h_n| - \log|h_m| < R\right). \tag{12}$$

- When $R = 0$, $\mathrm{P}\left(R_m + R_n - \bar{R}_m - \bar{R}_n < R\right) \to 0$, which is consistent with the previous discussions.
- When $R \neq 0$, (12) implies that the probability can be expressed asymptotically as follows:

$$\mathrm{P}\left(R_m + R_n - \bar{R}_m - \bar{R}_n < R\right) \to \mathrm{P}\left(\frac{|h_n|^2}{|h_m|^2} < 2^{2R}\right). \tag{13}$$

Impact of User Pairing on the Sum Rate Gap - Simulations



The probability that the sum rate gap between PD-NOMA and conventional MA is larger than R. $M = 5$ and $n = M$.

- There are $M = 5$ users and the user with the best channel is always slected.
- There are error floors for the probabilities, since the difference between the two sum rates is a constant at high SNR.
- Increasing the channel difference is helpful to ensure that NOMA outperforms OMA, i.e., the probability for $m = 1$ is smaller than that for $m = 2$.

Hybrid NOMA - Impact of User Pairing on Individual Rate Outage (1/2)

- We focus on the probability that PD-NOMA can achieve a larger rate than orthogonal MA for the m-th user

$$\mathrm{P}(R_m > \bar{R}_m) = \mathrm{P}\left(\left(1 + \frac{|h_m|^2 a_m^2}{|h_m|^2 a_n^2 + \frac{1}{\rho}}\right)^2 > (1 + \rho|h_m|^2)\right).$$

- On the other hand, the probability that the n-th user can experience better performance in NOMA than OMA is

$$\mathrm{P}(R_n > \bar{R}_n) = \mathrm{P}\left(\log\left(1 + \rho a_n^2 |h_n|^2\right) > \frac{1}{2}\log(1 + \rho|h_n|^2\right).$$

- Their high SNR approximations are given by

$$\mathrm{P}(R_m > \bar{R}_m) \approx \frac{\varpi_1}{\rho^m}, \quad \mathrm{P}(R_n > \bar{R}_n) \approx 1 - \frac{\varpi_2}{\rho^n}, \tag{14}$$

where ϖ_i is a constant and not a function of the SNR.

Hybrid NOMA - Impact of User Pairing on Individual Rate Outage (2/2)

- Therefore, the two users will have totally different experiences in NOMA systems.
- A user with strong channel conditions is more willing to perform NOMA since $\mathrm{P}(R_n > \bar{R}_n) \to 1 - \frac{1}{\rho^n} \to 1$.
- A user with poor channel conditions is not willing to perform NOMA since $\mathrm{P}(R_m > \bar{R}_m) \to \frac{1}{\rho^m} \to 0$
- Therefore, it is preferable to pair two users whose channel conditions are significantly different, since the above results imply that m should be as small as possible and n should be as large as possible.
- These conclusions are consistent with the previous ones
  - A larger n decreases $\mathrm{P}(R_m + R_n < \bar{R}_m + \bar{R}_n)$
  - In order to enlarge the sum rate gap, it is ideal to schedule two users whose channel conditions are different.

Hybrid NOMA - Resource Allocation (1/2)

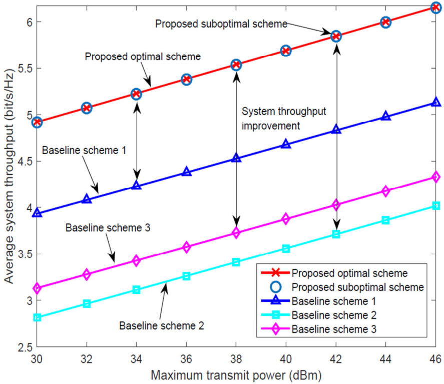
- Take resource allocation for MC-NOMA as an example.
- Resource allocation for MC-NOMA in general very difficult

$$\begin{aligned} \max_{power, subcarrier\ \ allocation} \quad & \text{(Weighted) sum rate} \\ \text{s.t.} \quad & \text{Power constraints} \\ & \text{Subcarrier allocation constraints} \end{aligned} \tag{15}$$

  - Containing two types of power allocation: within one NOMA group and between NOMA groups
  - It is also a mixed integer non-convex optimization problem
    - Monotonic optimization can be applied to solve such a non-convex optimization problem to obtain an optimal solution
    - A low-complexity suboptimal solution based on successive convex approximation can be obtained with a performance close to the optimal.

[7] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," IEEE TWC, 2017.

Hybrid NOMA - Resource Allocation (2/2)



Average system throughput (bits/s/Hz) vs. the maximum transmit power at the BS (dBm), P_{max}, for different resource allocation schemes and 6 users.

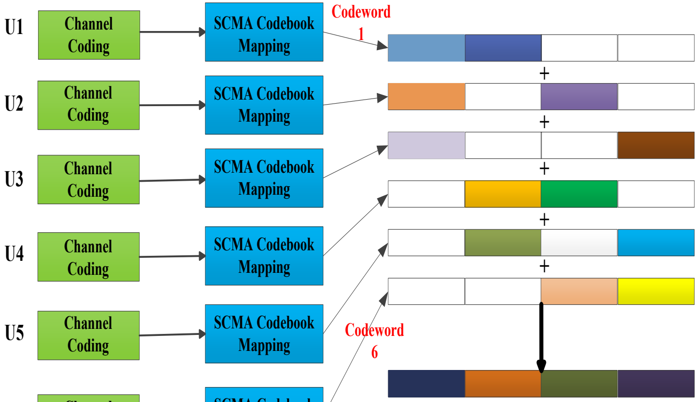
- For baseline scheme 1, a straightforward suboptimal joint power and subcarrier allocation for MC-NOMA is used.
- For baseline scheme 2, the user pair on each subcarrier is randomly selected and only power allocation is optimized.
- Baseline scheme 3 is a conventional MC-OMA scheme.

Outline

5G MC-NOMA

- Various practical forms of multi-carrier NOMA have been proposed for the 5G standard.
  - Multi-carrier NOMA achieves a favourable tradeoff between system performance and complexity.
- Both low density spreading (LDS) and sparse code multiple access (SCMA) are based on the idea that one user's information is spread over multiple subcarriers.
- However, the number of subcarriers assigned to each user is smaller than the total number of subcarriers
  - This is referred to as the low spreading (sparse) feature of these two versions of NOMA.
  - This feature ensures that the number of users utilizing the same subcarrier is not too large, such that the system complexity remains manageable.

- The key step of SCMA is how to map users to subcarriers.



$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

[8] H. Nikopour and H. Baligh, "Sparse code multiple access," IEEE PIMRC, 2013.

SCMA: An Example of MC-NOMA (1/2)

- Consider an SCMA system with 6 users and 4 subcarriers.
- The key step to implement SCMA is to design the factor graph matrix, which specifies which user's encoded messages are allocated to which subcarriers.
- A typical factor graph matrix for SCMA with 6 users and 4 subcarriers is the following

$$\mathbf{F} = \begin{array}{c} \begin{array}{cccccc} \text{U1} & \text{U2} & \text{U3} & \text{U4} & \text{U5} & \text{U6} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \leftarrow \text{subcarrier 1} \\ \leftarrow \text{subcarrier 2} \\ \leftarrow \text{subcarrier 3} \\ \leftarrow \text{subcarrier 4} \end{array}$$

where $[\mathbf{F}]_{i,j} = 1$ means that the j-th user can use the i-th subcarrier, and $[\mathbf{F}]_{i,j} = 0$ means that this user cannot use the subcarrier.

SCMA: An Example of MC-NOMA (2/2)

- The sparse feature of SCMA is reflected by the fact that there are only two non-zero entries in each column of $\mathbf{F}$, i.e., each user employs only two subcarriers.
- Since one user can use multiple subcarriers, SCMA employs multi-dimensional coding in order to ensure that the user's information is effectively spread over the subcarriers.
- Because one user's messages at different subcarriers are jointly encoded, SCMA requires joint decoding at the receiver, where the message passing algorithm (MPA) is used to ensure low complexity.
    - Joint decoding is an important feature of SCMA, which distinguishes it from power-domain NOMA, where SIC is employed.

Pattern Division Multiple Access (1/2)

- PDMA is another type of multi-carrier NOMA, but the low density spreading (sparse) feature is no longer strictly present
  - The number of subcarriers occupied by one user is not necessarily much smaller than the total number of subcarriers.
- Similar to the factor graph matrix for SCMA, the performance of PDMA is largely determined by the design of the subcarrier allocation matrix, referred to as the PDMA pattern matrix.
- Consider a case with five users and three subcarriers, and PDMA pattern matrix

$$\mathbf{Q} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}, \tag{16}$$

where the entries of this matrix indicate how the subcarriers are allocated to the users.

[9] S. Chen, et. al., "PDMA - a novel non-orthogonal multiple access for 5G radio networks", IEEE TVT, 2016

Pattern Division Multiple Access (2/2)

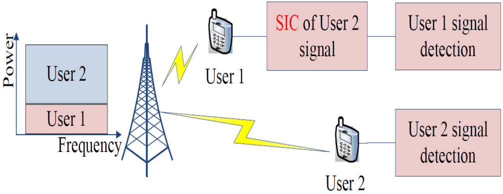
Consider a case with five users and three subcarriers, and PDMA pattern matrix

$$\mathbf{Q} = \begin{array}{cl} \begin{array}{ccccc} \text{U1} & \text{U2} & \text{U3} & \text{U4} & \text{U5} \end{array} & \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} & \begin{array}{l} \text{subcarrier 1} \\ \text{subcarrier 2} \\ \text{subcarrier 3} \end{array} \end{array}$$

- User 1 is able to transmit or receive on all subcarriers.
- User 5 uses the first subcarrier only.
- Therefore, different from LDS and SCMA, some users might be able to use all the subcarriers.
    - SCMA requires that each user occupies the same number of subcarriers.
    - This constraint is not required by PDMA and hence makes PDMA more flexible.

Outline

Motivation and Introduction



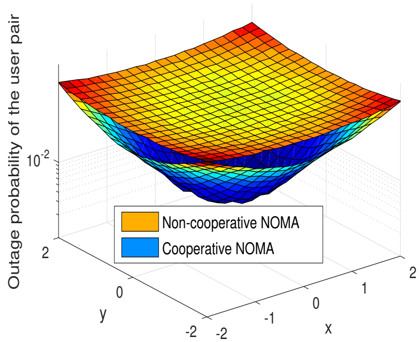
- User 2 suffers some performance loss in NOMA
- There is redundant information inherent in NOMA systems
  - Users with better channel conditions know the information sent to the other users.
- Cooperative NOMA exploits this feature
  - User 1 is a natural relay and helps user 2.
  - 3 time slots are needed for cooperative OMA, but cooperative NOMA only needs 2.

[10] Z. Ding, M. Peng and H. V. Poor, "Cooperative Non-Orthogonal Multiple Access in 5G Systems", IEEE CL, 2015.

A Simple Example (1/3)

- Consider a NOMA downlink with two users.
- Time slot I: BS sends the superimposed messages to both users
- Time slot II: The user with strong channel conditions is to help its partner by acting as a relay
- Simulation parameters are set as follows:
  - The BS is located at $(0, 0)$.
  - User 2 is located at $(5m, 0)$.
  - The x-y plane denotes the location of user 1.
  - A bounded path loss model is used to ensure all distances are greater than one. The path loss exponent is 3.
  - The transmit signal-to-noise ratio (SNR) is 30 dB.
  - The power allocation coefficient for user 2 and user 1 are $(a_A, a_B) = \left(\frac{4}{5}, \frac{1}{5}\right)$.
  - The targeted data rate is 0.5 bits per channel use (BPCU).

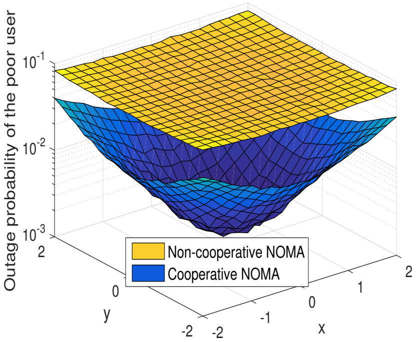
A Simple Example (2/3)



Overall outage probabilities achieved by cooperative NOMA and OMA.

- Overall outage means that outage happens if any of the two users is in outage.
- Cooperative NOMA achieves a lower outage probability than non-cooperative OMA.
- Both schemes are affected by the location of the strong user.
- A careful choice for user 1's location can significantly enlarge the performance gap between the two schemes.

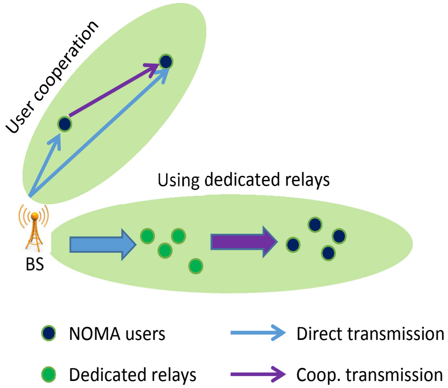
A Simple Example (3/3)



The poor user's outage probabilities achieved by cooperative and no-cooperative NOMA.

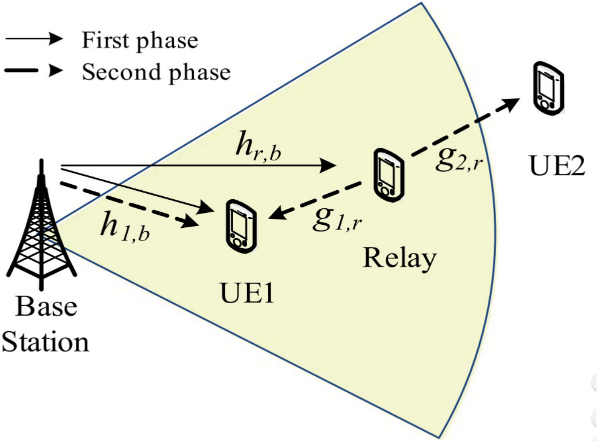
- This figure shows the outage performance for the user with poor channel conditions.
- The use of cooperation significantly helps the user with poor channel conditions.
- For non-cooperative NOMA, one user's location has no impact on the other's performance.

Outline



- We provide an example to show the benefits of this type of cooperative NOMA
  - Assume that there is a dedicated relay which is used to help two users located close to the cell edge.
  - There is no direct link between the base station and the users.
  - With cooperative OMA, four time slots are required.
  - With cooperative NOMA, only two time slots are needed.

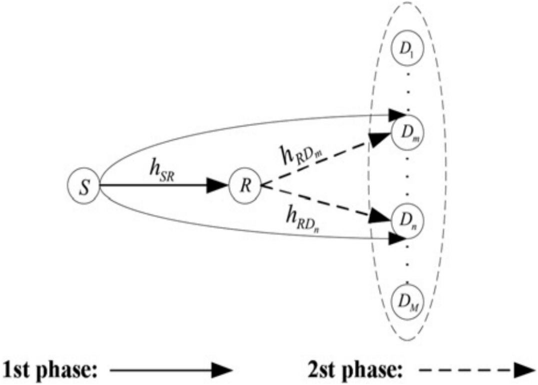
Variations of Cooperative NOMA (1/2)



Using a dedicated relay to help a user close to the cell edge.

[11] J. Kim and I. Lee, "Non-Orthogonal Multiple Access in Coordinated Direct and Relay Trans.", IEEE CL, 2015.

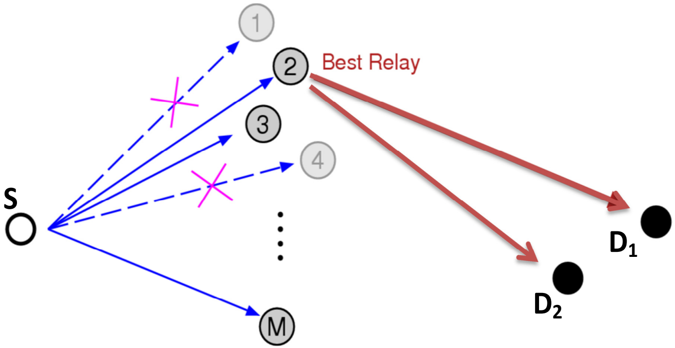
Variations of Cooperative NOMA (2/2)



Using a dedicated relay to help users close to the cell edge.

[12] J. Men and J. Ge, "Non-Orthogonal Multiple Access for Multiple-Antenna Relaying Networks"", IEEE CL, 2015.

Relay Selection for Cooperative NOMA (1/2)



Relay selection studies which relay to use when multiple relays are available.

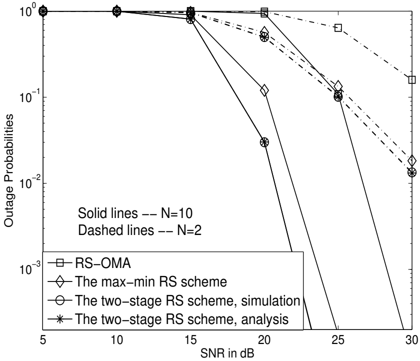
[13] Z. Ding, H. Dai and H. V. Poor, "Relay Selection for Cooperative NOMA", IEEE WCL, 2016.

Relay Selection for Cooperative NOMA (2/2)

Two types of relay selection can be used

- The max-min relay selection criterion
  - Select a relay whose incoming and outgoing channels are balanced.
  - This is the optimal strategy in conventional cooperative networks.
- The two-stage relay selection strategy
  - Stage one: Relays which can guarantee the performance of the user with strict QoS requirements are identified and grouped into a subset.
  - Stage two: Select from the qualified relay subset the relay that yields the largest rate for the other user which only needs to be served opportunistically.
- The two-stage relay selection strategy is optimal to minimize the overall outage probability.

Relay Selection for Cooperative NOMA - Simulations



Comparison between cooperative OMA and NOMA with different relay selection (RS) strategies. $R_1 = 0.5$ bit per channel use (BPCU), $R_2 = 2$ BPCU, and $\alpha_2 = \frac{1}{4}$.

- Both NOMA schemes with different relay selection strategies outperform OMA.
- The two-stage relay selection strategy outperforms the max-min scheme.
- All the curves have the same slope, which means that the same diversity gain is archived by the corresponding schemes.

Outline

Motivations

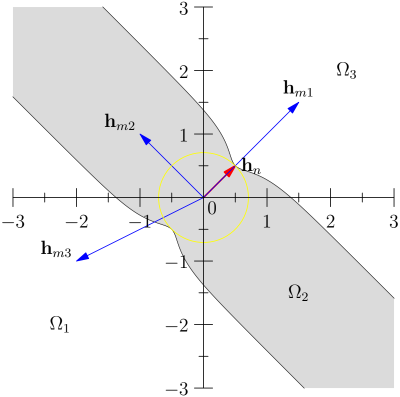
- Why to design MIMO-NOMA?
  - MIMO offers degrees of freedom to further improve the system throughput of NOMA.
- What are the challenges?
  - The key feature of NOMA is to exploit the difference between users' channel conditions.
  - In scenarios with single-antenna nodes, channels are scalar and it is easy to order the users based their channel conditions.
  - In MIMO, channels are in form of matrices/vectors, which makes it difficult to order users.
- The general idea behind MIMO-NOMA designs is introduced first.

Quasi-Degradation Criterion (1/2)

- In general, it is difficult to tell how optimal a MIMO-NOMA transmission scheme is
  - The capacity for a general MIMO broadcast channel is still unknown, and dirty paper coding (DPC) is commonly used as a benchmark.
  - Analyzing the gap between MIMO-NOMA and DPC in general is challenging.
- In the special case of MISO, the use of quasi-degradation tells us how optimal MISO-NOMA is:
  - The quasi-degradation criterion is to describe the condition under which the use of NOMA can realize the same performance as DPC.
- Consider the following example:
  - The BS has 2 antennas and there are 2 single-antenna users.
  - Here, $\mathbf{h}_n$ ($\mathbf{h}_m$) denotes the 2×1 channel vector of user n (m).

[14] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink", IEEE TSP, 2016.

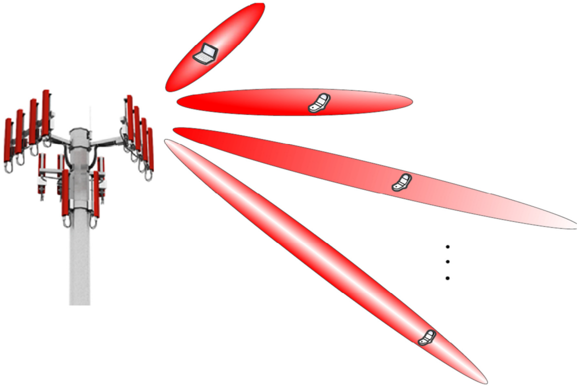
Quasi-Degradation Criterion (2/2)



An illustration of quasi-degradation when a BS equipped with two antennas communicates with two single-antenna users.

- An extreme example for quasi-degradation is that one user's channel vector is a scaled version of the other user's channel vector, i.e., $\mathbf{h}_n$ and $\mathbf{h}_{m1}$.
- An extreme example for non-quasi-degradation is that the users' channel vectors are mutually orthogonal, i.e., $\mathbf{h}_n$ and $\mathbf{h}_{m2}$.

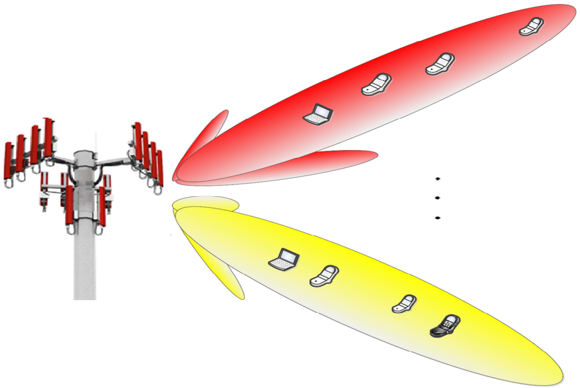
Practical Designs of MIMO-NOMA



One type of MIMO-NOMA approaches is to assign users to beams individually and order users according to their path losses.

[15] M. F. Hanif, Z. Ding, T. Ratnarajah and G. K. Karagiannidis, "A Minorization-Maximization Method for Optimizing Sum Rate in Non-Orthogonal Multiple Access Systems", IEEE TSP, 2016.

Outline



- In the part, we focus on approaches that decompose MIMO-NOMA into SISO-NOMA.
  - Applicable to both uplink and downlink transmission
  - Applicable to other types of 5G multiple access

Approach I: Without CSIT (1/6)

- Consider a downlink communication scenario with one BS equipped with M antennas and multiple users equipped with N antennas each.
- To make NOMA applicable, the users are randomly grouped into M clusters with K users in each cluster.
- It is assumed that $N \geq M$
    - Examples include ultra-densely deployed small cells in which we have low-cost and low-power small-cell base stations
    - Another example are cloud radio access networks (C-RANs)
    - Also applicable to Internet of Things, such as smart homes, in which the capability of a home base station is similar to that of laptops and other digital devices
- Scenarios in which a BS has more antennas than users will be discussed later.

---

[17] Z. Ding, F. Adachi and H. V. Poor, "The Application of MIMO to Non-Orthogonal Multiple Access", IEEE

Approach I: Without CSIT (2/6)

The signals transmitted by the BS are given by

$$\mathbf{x} = \mathbf{P}\tilde{\mathbf{s}}, \tag{17}$$

- the $M \times 1$ vector $\tilde{\mathbf{s}}$ is given by

$$\tilde{\mathbf{s}} = \begin{bmatrix} \alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K} \\ \vdots \\ \alpha_{M,1}s_{M,1} + \cdots + \alpha_{M,K}s_{M,K} \end{bmatrix} \triangleq \begin{bmatrix} \tilde{s}_1 \\ \vdots \\ \tilde{s}_M \end{bmatrix} \tag{18}$$

- $s_{m,k}$ denotes the information bearing signal to be transmitted to the k-th user in the m-th cluster
- $\alpha_{i,j}$ denotes the NOMA power allocation coefficient
- the design of the $M \times M$ precoding matrix $\mathbf{P}$ will be discussed later.

Approach I: Without CSIT (3/6)

- The observation at the k-th user in the first cluster is given by

$$\mathbf{y}_{1,k} = \mathbf{H}_{1,k}\mathbf{P}\tilde{\mathbf{s}} + \mathbf{n}_{1,k}, \tag{19}$$

  - $\mathbf{H}_{1,k}$ is the $N \times M$ Rayleigh fading channel matrix from the BS to the k-th user in the first cluster,
  - $\mathbf{n}_{1,k}$ is an additive Gaussian noise vector.
- After applying this detection vector $\mathbf{v}_{1,k}$, the signal model can be rewritten as follows:

$$\mathbf{v}_{1,k}^H\mathbf{y}_{1,k} = \mathbf{v}_{1,k}^H\mathbf{H}_{1,k}\mathbf{P}\tilde{\mathbf{s}} + \mathbf{v}_{1,k}^H\mathbf{n}_{1,k}. \tag{20}$$

- Denote the i-th column of $\mathbf{P}$ by $\mathbf{p}_i$. The above signal model can be rewritten as follows:

$$\mathbf{v}_{1,k}^H\mathbf{y}_{1,k} = \mathbf{v}_{1,k}^H\mathbf{H}_{1,k}\mathbf{p}_1\left(\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K}\right) \tag{21}$$
$$+ \sum_{m=2}^{M}\mathbf{v}_{1,k}^H\mathbf{H}_{1,k}\mathbf{p}_m\tilde{s}_m + \mathbf{v}_{1,k}^H\mathbf{n}_{1,k}.$$

Approach I: Without CSIT (4/6)

- The effective channel gains are ordered as follows:

$$|\mathbf{v}_{1,1}^H \mathbf{H}_{1,1} \mathbf{p}_1|^2 \geq \cdots \geq |\mathbf{v}_{1,K}^H \mathbf{H}_{1,K} \mathbf{p}_1|^2, \tag{22}$$

and we assume fixed power allocation as follows:

$$\alpha_{1,1} \leq \cdots \leq \alpha_{1,K}.$$

- The messages $s_{1,j}$, $K \geq j \geq (k+1)$, will be detected at the k-th user in the first cluster with the following SINR:

$$SINR_{1,k}^{j} = \tag{23}$$

$$\frac{|\mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{p}_1|^2 \alpha_{1,j}^2}{\sum_{l=1}^{j-1} |\mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{p}_1|^2 \alpha_{1,l}^2 + \sum_{m=2}^{M} |\mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{p}_m|^2 + |\mathbf{v}_{1,k}|^2 \frac{1}{\rho}}.$$

- If $\log(1 + SINR_{1,k}^{j}) > R_{1,j}$, SIC can be carried out.

Approach I: Without CSIT (5/6)

- To completely remove inter-cluster interference, the following constraints should be satisified

$$\mathbf{v}_{i,k}^H \mathbf{H}_{i,k} \mathbf{p}_m = 0, \quad \forall m \neq i. \tag{24}$$

- Without CSIT, we let $\mathbf{P} = \mathbf{I}_M$
  - Avoid asking the users to feedback all their CSI to the BS.
- With this $\mathbf{P}$, the constraints on the detection matrices become

$$\mathbf{v}_{i,k}^H \mathbf{h}_{m,ik} = 0, \tag{25}$$

  where $\mathbf{h}_{m,ik}$ is the m-th column of $\mathbf{H}_{i,k}$.
- Therefore at the k-th user in the i-th cluser, the constraints can be rewritten as follows:

$$\mathbf{v}_{i,k}^H \underbrace{\begin{bmatrix} \mathbf{h}_{1,ik} & \cdots & \mathbf{h}_{i-1,ik} & \mathbf{h}_{i+1,ik} & \cdots & \mathbf{h}_{M,ik} \end{bmatrix}}_{\tilde{\mathbf{H}}_{i,k}} = 0. \tag{26}$$

Approach I: Without CSIT (6/6)

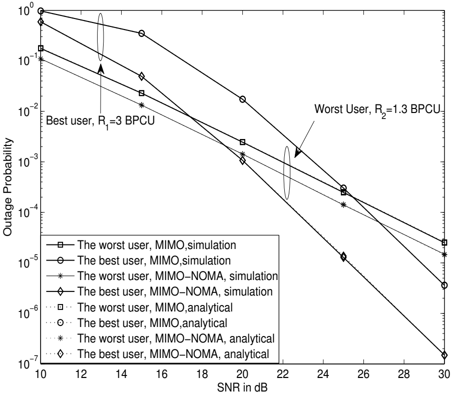
- Note that the dimension of $\tilde{\mathbf{H}}_{i,k}$ is $N \times (M-1)$ since it is a submatrix of $\mathbf{H}_{i,k}$ formed by removing one column.
- As a result, $\mathbf{v}_{i,k}$ can be obtained from the null space of $\tilde{\mathbf{H}}_{i,k}$,

$$\mathbf{v}_{i,k} = \mathbf{U}_{i,k}\mathbf{z}_{i,k}, \tag{27}$$

  - $\mathbf{U}_{i,k}$ contains all the left singular vectors of $\tilde{\mathbf{H}}_{i,k}$ corresponding to zero singular values,
  - $\mathbf{z}_{i,k}$ is a $(N-M+1)\times 1$ normalized vector and can be chosen following the maximal ratio combining (MRC) principle.
  - In order to ensure the existence of $\mathbf{v}_{i,k}$, $N \geq M$ is assumed.
- Note that with such $\mathbf{v}_{i,k}$, we have decomposed MIMO-NOMA into SISO-NOMA since

$$\mathbf{v}_{1,k}^H\mathbf{y}_{1,k} = \mathbf{v}_{1,k}^H\mathbf{H}_{1,k}\mathbf{p}_1\left(\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K}\right) + \mathbf{v}_{1,k}^H\mathbf{n}_{1,k}.$$

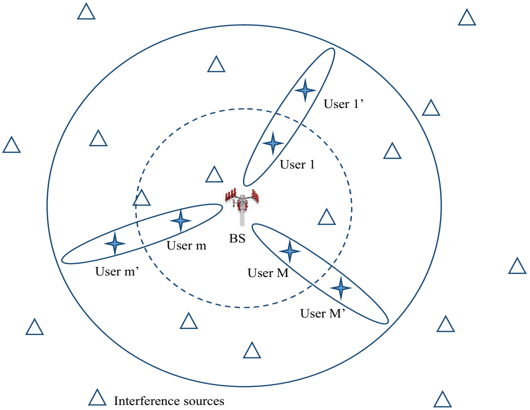
Approach I: Numerical Studies



Outage probabilities achieved by MIMO-OMA and MIMO-NOMA. $M = 2$, $K = 2$ and $N = 3$.

- For both of the two users, the use of NOMA can reduce their outage probabilities.
- The diversity gains achieved by the two users are related to their channel conditions.
- NOMA achieves the same diversity gain as OMA.

Approach II: With CSIT (1/9)



[18] Z. Ding, R. Schober, and H. V. Poor, "A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment", IEEE TWC, 2016.

Approach II: With CSIT (2/9)

- Consider a MIMO-NOMA downlink communication scenario with a base station equipped with M antennas and users equipped with N antennas.
- Assume $N > \frac{M}{2}$ in order to implement the concept of signal alignment.
- The users are assumed to be uniformly deployed in a disk, denoted by $\mathcal{D}$, i.e., the cell controlled by the base station.
    - The radius of the disk is r, and the BS is at the center of $\mathcal{D}$.
- Assume that the disk is divided into two regions.
    - The first region is a smaller disk, denoted by $\mathcal{D}_1$, with radius r_1 ($r_1 < r$) and the base station located at its origin.
    - The second region is a ring, denoted by $\mathcal{D}_2$, constructed from $\mathcal{D}$ by removing $\mathcal{D}_1$.
    - Assume that M pairs of users are selected, where user m, randomly located in $\mathcal{D}_1$, is paired with user m', randomly located in $\mathcal{D}_2$.

Approach II: with CSIT (3/9)

- Interference sources are distributed in $\mathcal{R}^2$ according to a homogeneous Poisson point process (PPP) Ψ_I of density λ_I.
  - The interference sources are equipped with a single antenna and use identical transmission powers, denoted by ρ_I.
  - Examples include cognitive radio networks with single-antenna secondary users, or single-antenna sensors in 5G powered IoT.
- Consider a composite channel model with Rayleigh fading and path loss, i.e., the channel matrix of user m is $\mathbf{H}_m = \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}$
  - $\mathbf{G}_m$ denotes an $N \times M$ matrix whose elements represent Rayleigh fading channel gains,
  - d_m denotes the distance from the base station to the user,
  - The path loss is modelled as follows $L(d_m) = \begin{cases} d_m^{\alpha}, & \text{if} \quad d_m > r_0 \\ r_0^{\alpha}, & \text{otherwise} \end{cases}$, where α denotes the path loss exponent and parameter r_0 avoids a singularity when the distance is small. Note that $r_1 \geq r_0$.

- The base station sends the following $M \times 1$ vector

$$\mathbf{s} = \begin{bmatrix} \alpha_1 s_1 + \alpha_{1'} s_{1'} \\ \vdots \\ \alpha_M s_M + \alpha_{M'} s_{M'} \end{bmatrix}, \tag{28}$$

  where s_m is the signal intended for the m-th user, α_m is the power allocation coefficient, and $\alpha_m^2 + \alpha_{m'}^2 = 1$.

- User m's observation is given by

$$\mathbf{y}_m = \frac{\mathbf{G}_m}{\sqrt{L(d_m)}} \mathbf{P}\mathbf{s} + \mathbf{w}_{I_m} + \mathbf{n}_m, \tag{29}$$

  - $\mathbf{P}$ is an $M \times M$ precoding matrix to be defined later
  - $\mathbf{w}_{I_m}$ is the overall co-channel interference received by user m
  - $\mathbf{n}_m$ denotes the noise vector.

- User m applies a detection vector $\mathbf{v}_m$ to its observation, and therefore the user's observation can be re-written as follows:

$$\mathbf{v}_m^H \mathbf{y}_m = \mathbf{v}_m^H \frac{\mathbf{G}_m}{\sqrt{L(d_m)}} \mathbf{p}_m (\alpha_m s_m + \alpha_{m'} s_{m'}) \\ + \underbrace{\sum_{i \neq m} \mathbf{v}_m^H \frac{\mathbf{G}_m}{\sqrt{L(d_m)}} \mathbf{p}_i (\alpha_i s_i + \alpha_{i'} s_{i'}) + \mathbf{v}_m^H (\mathbf{w}_{I_m} + \mathbf{n}_m)}_{interference\ \ (including\ \ inter-pair\ \ interference)\ +\ \ noise}, \tag{30}$$

where $\mathbf{p}_m$ denotes the m-th column of $\mathbf{P}$.

- In order to remove inter-pair interference, the following constraint has to be met:

$$\begin{bmatrix} \mathbf{v}_m^H \mathbf{G}_m \\ \mathbf{v}_{m'}^H \mathbf{G}_{m'} \end{bmatrix} \mathbf{p}_i = \mathbf{0}_{2\times 1}, \ \forall i \neq m. \tag{31}$$

- Why to use signal alignment
  - Without loss of generality, we focus on $\mathbf{p}_1$ which needs to satisfy the following constraint:

$$\begin{bmatrix} \mathbf{G}_2^H\mathbf{v}_2 & \mathbf{G}_{2'}^H\mathbf{v}_{2'} & \cdots & \mathbf{G}_M^H\mathbf{v}_M & \mathbf{G}_{M'}^H\mathbf{v}_{M'} \end{bmatrix}^H \mathbf{p}_1 = \mathbf{0}_{2(M-1)\times 1}.$$

  - Note that the dimension of the matrix above is $2(M-1)\times M$.
  - Therefore, in general, a non-zero vector $\mathbf{p}_i$ satisfying the above constraint does not exist.
  - The motivation to use signal alignment is to ensure the existence of $\mathbf{p}_i$
    - One straightforward approach is to serve fewer user pairs, i.e., reducing the number of user pairs to $\left(\frac{M}{2}+1\right)$.
    - However, this approach will reduce the overall system throughput.

- With signal alignment, the detection vectors should satisfy

$$\mathbf{v}_m^H \mathbf{G}_m = \mathbf{v}_{m'}^H \mathbf{G}_{m'}, \tag{32}$$

or equivalently $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix} \begin{bmatrix} \mathbf{v}_m^H & \mathbf{v}_{m'}^H \end{bmatrix}^H = \mathbf{0}_{M\times 1}$.

- Define $\mathbf{U}_m$ as the $2N \times (2N - M)$ matrix containing the $(2N - M)$ right singular vectors of $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix}$ corresponding to its zero singular values.
- Therefore, the detection vectors are designed as follows:

$$\begin{bmatrix} \mathbf{v}_m \\ \mathbf{v}_{m'} \end{bmatrix} = \mathbf{U}_m \mathbf{x}_m, \tag{33}$$

where $\mathbf{x}_m$ is a $(2N - M) \times 1$ vector and can be determined by using MRC. We normalize $\mathbf{x}_m$ to 2, i.e., $|\mathbf{x}|^2 = 2$.

- It is clear $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix} \mathbf{U}_m \mathbf{x}_m = \mathbf{0}_{M\times 1}$.

- The effect of the signal alignment is to project the channels of the two users in the same pair into the same direction.
- Define $\mathbf{g}_m \triangleq \mathbf{G}_m^H \mathbf{v}_m$ as the effective channel vector shared by the two users.
- The constraint for $\mathbf{p}_i$ can be rewritten as follows:

$$\begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_{i-1} & \mathbf{g}_{i+1} & \cdots & \mathbf{g}_M \end{bmatrix}^H \mathbf{p}_i = \mathbf{0}_{(M-1)\times 1}. \tag{34}$$

- Note that $\begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_{i-1} & \mathbf{g}_{i+1} & \cdots & \mathbf{g}_M \end{bmatrix}^H$ is an $(M-1) \times M$ matrix, so a $\mathbf{p}_i$ satisfying (34) exists.
- Define $\mathbf{G} \triangleq \begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_M \end{bmatrix}^H$. Precoding can be designed as

$$\mathbf{P} = \mathbf{G}^{-H}\mathbf{D}, \tag{35}$$

where $\mathbf{D}^2 = \mathrm{diag}\{\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{1,1}}, \cdots, \frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{M,M}}\}$.

- The signal model for user m can now be written as follows:

$$\mathbf{v}_m^H \mathbf{y}_m = \frac{(\alpha_m s_m + \alpha_{m'} s_{m'})}{\sqrt{L(d_m)(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}} + \mathbf{v}_m^H (\mathbf{w}_{I_m} + \mathbf{n}_m). \tag{36}$$

- Define $y_m = \mathbf{v}_m^H \mathbf{y}_m$, $h_m = \frac{1}{\sqrt{L(d_m)(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}$, $h_{m'} = \frac{1}{\sqrt{L(d_{m'})(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}$, $w_{I_m} = \mathbf{v}_m^H \mathbf{w}_{I_m}$, and $n_m = \mathbf{v}_m^H \mathbf{n}_m$.
- The two users receive the following scalar observations

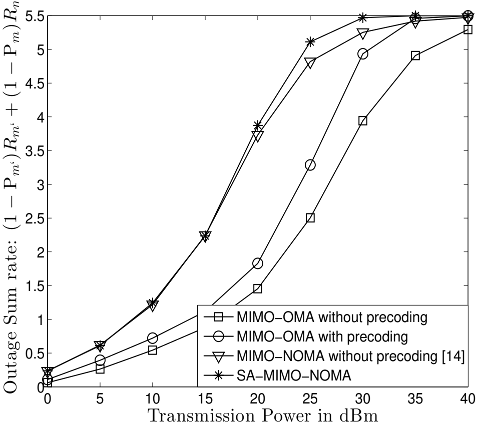
$$y_m = h_m(\alpha_m s_m + \alpha_{m'} s_{m'}) + w_{I_m} + n_m, \tag{37}$$

and

$$y_{m'} = h_{m'}(\alpha_m s_m + \alpha_{m'} s_{m'}) + w_{I_{m'}} + n_{m'}. \tag{38}$$

- We have decomposed MIMO-NOMA into SISO-NOMA.

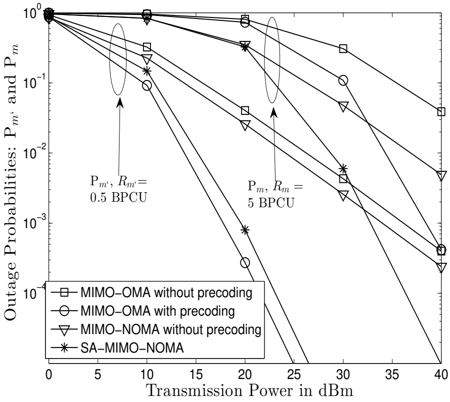
Approach II: Numerical Studies (1/2)



Outage sum rates achieved by the considered MIMO schemes. $M = N = 3$

- The proposed signal alignment based scheme is termed "SA-MIMO-NOMA".
- Both the MIMO-NOMA schemes outperform the MIMO-OMA schemes.
- In terms of outage rates, the difference between the two MIMO-NOMA schemes with and without precoding is small.

Approach II: Numerical Studies (2/2)



Outage probabilities achieved by the considered MIMO schemes. $M = N = 3$

- MIMO-NOMA with precoding outperforms MIMO-NOMA without precoding.
- The diversity gain achieved by MIMO-NOMA with precoding is larger than that of MIMO-NOMA without precoding.
- MIMO-NOMA and MIMO-OMA achieve the same diversity gain.

Approach III: Massive-MIMO-NOMA

- The previous approaches more or less require some assumptions about the number of antennas at the users.
- Can we apply NOMA to massive MIMO, where the BS has M antennas, each user has N antennas, and $M >> N$?
- Following the geometrical one-ring scattering model, we divide the users into K spatial clusters
    - There are L users in each cluster sharing the same spatial correlation matrix, denoted by $\mathbf{R}_k$
- By utilizing such spatial correlation, the two MIMO-NOMA approaches introduced before can be extended to massive MIMO.

[19] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA with Limited Feedback", IEEE SPL, 2016.

Outline

System Model (1/5)

- Consider a MIMO-NOMA system with one BS and two users
- The BS has M antennas and each user has N antennas.
- The two users' channels are statistically the same.
- The base station will transmit the following vector:

$$\mathbf{x} = \mathbf{P}\mathbf{s}. \tag{39}$$

- The vector $\mathbf{s}$ is constructed by using the NOMA approach

$$\mathbf{s} = \begin{bmatrix} \alpha_1 s_1 + \beta_1 w_1 & \cdots & \alpha_N s_N + \beta_N w_N \end{bmatrix}^T, \tag{40}$$

where s_i is the i-th stream sent to user 1, α_i is the power allocation coefficient for s_i, w_i and β_i are defined similarly.

[20] Z. Ding, L. Dai and H. V. Poor, "MIMO-NOMA Design for Small Packet Transmission in the Internet of Things", IEEE Access, 2016.

System Model (2/5)

- Users are ordered according to their QoS requirements, instead of their channels
  - Take intelligent transportation as an example
    - User 1 can be a vehicle receiving the incident warning information which is contained within a few bytes only.
    - User 2 can be another vehicle which is to perform some background tasks, such as downloading multimedia files.
- There are two sets of parameters to be designed, the precoding matrix **P** and the power allocation coefficients α_i.
- The aim of the proposed design is to realize two goals simultaneously.
  - One is to meet the QoS requirement at user 1 strictly, with a low data rate - CR power allocation
  - The other is to artificially create channel difference and serve user 2 opportunistically - Precoding

System Model (3/5)

- Assume that the QR decomposition of user 2's channel matrix, $\mathbf{H}_2$, is given by

$$\mathbf{H}_2^H = \mathbf{Q}_2\tilde{\mathbf{R}}_2, \tag{41}$$

  - $\mathbf{Q}_2$ is an $M \times M$ unitary matrix,
  - $\tilde{\mathbf{R}}_2$ is an $M \times N$ matrix obtained from the QR decomposition.
- Define $\mathbf{V}_2$ as an $M \times N$ matrix collecting the N left columns of $\mathbf{Q}_2$, and $\mathbf{R}_2$ is an $N \times N$ upper submatrix of $\tilde{\mathbf{R}}_2$. From the QR decomposition, we know that $\mathbf{H}_2^H = \mathbf{V}_2\mathbf{R}_2$.
- The precoding matrix $\mathbf{P}$ is set as $\mathbf{P} = \mathbf{V}_2$, which is to improve the signal strength at user 2.
- As can be seen in the following, this choice of the precoding matrix also degrades the channel conditions at user 1, which makes user 1 analogous to a cell edge user in a conventional NOMA setup.

System Model (4/5)

- User 2's observation can be expressed as follows:

$$\mathbf{y}_2 = \mathbf{R}_2^H \mathbf{s} + \mathbf{n}_2, \tag{42}$$

  where $\mathbf{n}_2$ is the noise vector.
- Since $\mathbf{R}_2^H$ is a lower triangular matrix, SIC can be carried out to cancel inter-layer interference (between w_i and w_j, $i \neq j$) and intra-layer interference (between s_i and w_i).
- Particularly, suppose that s_j and w_j from the previous layers, $j < i$, are decoded successfully.
  - User 2 can decode the message intended for user 1 at the i-th layer, s_i, with the following SINR: $\text{SINR}_{2,i'} = \frac{\alpha_i^2 [\mathbf{R}_2^H]_{i,i}^2}{\beta_i^2 [\mathbf{R}_2^H]_{i,i}^2 + \frac{1}{\rho}}$.
  - Provided that $\log(1 + \text{SINR}_{2,i'}) > R_{1,i}$, user 2 can successfully remove user 1's message, s_i, from its i-th layer, and its own message can be decoded with the following SNR:

$$\text{SNR}_{2,i} = \rho \beta_i^2 [\mathbf{R}_2^H]_{i,i}^2. \tag{43}$$

System Model (5/5)

- User 1's observation is given by

$$\mathbf{y}_1 = \mathbf{H}_1\mathbf{P}\mathbf{s} + \mathbf{n}_1. \tag{44}$$

- Analogously to the cell edge user in a conventional NOMA network, user 1 is not to decode w_i
  - The use of the QR based detection will result in a significant performance loss
  - Therefore, zero forcing is applied at user 1.
- Particularly, the system model at user 1 can be written as:

$$(\mathbf{H}_1\mathbf{V}_2)^{\dagger}\,\mathbf{y}_1 = \mathbf{s} + (\mathbf{H}_1\mathbf{V}_2)^{\dagger}\,\mathbf{n}_1, \tag{45}$$

where $(\mathbf{H}_1\mathbf{V}_2)^{\dagger} = \left(\mathbf{V}_2^H\mathbf{H}_1^H\mathbf{H}_1\mathbf{V}_2\right)^{-1}\mathbf{V}_2^H\mathbf{H}_1^H$.

- As a result, user 1 can decode its message with $\mathrm{SINR}_{1,i} = \frac{\alpha_i^2 z_i}{\beta_i^2 z_i + \frac{1}{\rho}}$, where $z_i = \frac{1}{\left[\left(\mathbf{V}_2^H\mathbf{H}_1^H\mathbf{H}_1\mathbf{V}_2\right)^{-1}\right]_{i,i}}$.

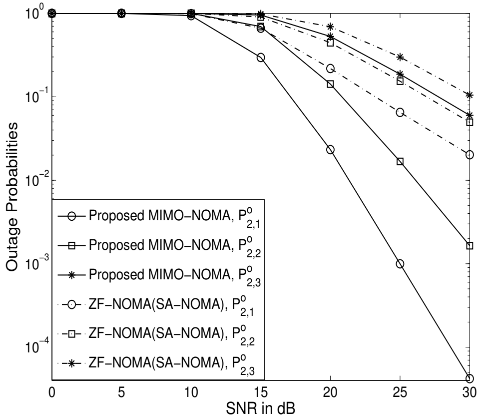
Impact of the Proposed Precoding Scheme (1/2)

- The two users' experiences with the proposed precoding scheme are different.
- The reception reliability at user 2 is determined by parameter x_i, where $x_i \triangleq [\mathbf{R}_2^H]_{i,i}^2$.
  - It follows a chi-square distribution with $2(M-i+1)$ degrees of freedom.
  - Therefore, more antennas at the base station can improve the receive signal strength at user 2 which is a function of $[\mathbf{R}_2^H]_{i,i}^2$.
- The reception reliability at user 1 is degraded due to the use of the precoding matrix, $\mathbf{P}$.
  - $\mathbf{H}_1\mathbf{P}$ is still an $N \times N$ complex Gaussian matrix
  - The use of the proposed precoding matrix <u>shrinks</u> user 1's channel matrix from an $N \times M$ complex Gaussian matrix to another complex Gaussian matrix with smaller size.
  - The effective channel gain is exponentially distributed, and its pdf is no longer a function of M.

Impact of the Proposed Precoding Scheme (2/2)

- The impact of the proposed precoding scheme can be illustrated by using the following extreme example.
- Consider the special case with $N = 1$, where the channel matrices become $1 \times M$ vectors, denoted by $\mathbf{h}_1$ and $\mathbf{h}_2$, respectively.
- After applying $\mathbf{P}$, the effective channel gain at user 2 is $|\mathbf{h}_2|^2$ which becomes stronger by increasing M.
- On the other hand, the effective channel gain at user 1 is always exponentially distributed, and the use of more antennas at the base station does not improve the transmission reliability at user 1.
- The benefits of this design are
  - Users' effective channel gains become very different
  - Users' effective channel conditions reflect their QoS targets

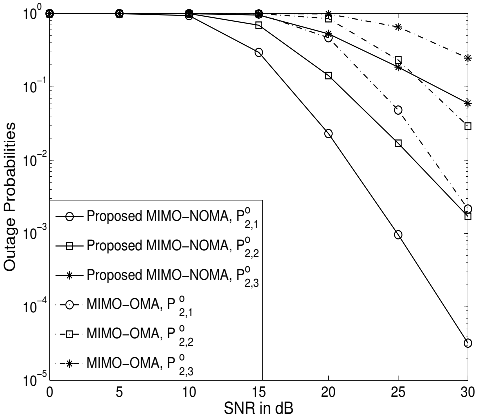
Numerical Studies (1/2)



Outage probabilities achieved by the considered MIMO-NOMA schemes. $M = N = 3$

- When users have the same path loss and $M = N$, ZF-NOMA and SA-NOMA are exactly the same.
- The new type of MIMO-NOMA can reduce the outage probability significantly.
- Over the three layers, ZF-NOMA can only achieve a diversity gain of 1, smaller than the new type of MIMO-NOMA.

Numerical Studies (2/2)



Outage probabilities achieved by the MIMO-NOMA and MIMO-OMA schemes. $M = N = 3$

- The new type of MIMO-NOMA can reduce the outage probability significantly, compared to MIMO-OMA.
- Over the three layers, the MIMO-NOMA and MIMO-OMA schemes achieve the same diversity gain.

Outline

Motivation for Combining MmWave and NOMA

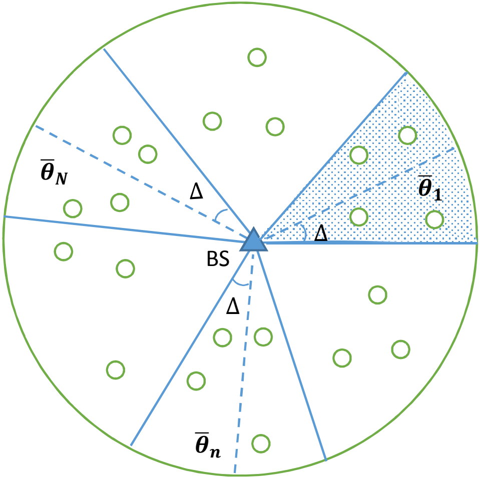
- NOMA is compatible to massive MIMO, PHY-security, CR networks, and other types of 5G techniques.
- In this part, we focus on the coexistence between NOMA and mmWave communications
  - Similar to NOMA, the motivation for using mmWave communications is motivated by the spectrum crunch.
  - The solution provided by mmWave communications is to use less occupied mmWave bands.
- Since there is more spectrum available, do we still need NOMA?
- The huge demand for bandwidth resources due to the exponential growth of broadband traffic can be met only by
  - Acquiring more radio spectrum
  - Efficiently using the acquired spectrum

System Model - Network Topology (1/2)

- Consider a mmWave-NOMA downlink transmission scenario with one base station communicating with multiple users.
- The base station is equipped with M antennas and each user has a single antenna.
- Denote the disk which is covered by the base station by $\mathcal{D}$.
  - Assume that the base station is located at the origin of $\mathcal{D}$
  - The radius of the disk is denoted by $R_{\mathcal{D}}$.
  - Assume that users are randomly deployed in the disk following the homogeneous Poisson point process, with density λ.
- We will show that the propagation characteristics of mmWave transmission is ideal for the implementation of NOMA.

[21] Z. Ding, P. Fan and H. V. Poor, "Random Beamforming in Millimeter-Wave NOMA Networks", IEEE Access, 2017.

System Model - Network Topology (2/2)



System Model - mmWave Channel Model (1/2)

- The mmWave-based channel vector from the base station to user k can be expressed as follows:

$$\mathbf{h}_k = \sqrt{M}\frac{a_{k,0}\mathbf{a}(\theta_k^0)}{\sqrt{1+d_k^{\alpha_{LOS}}}} + \sqrt{M}\sum_{l=1}^{L}\frac{a_{k,l}\mathbf{a}(\theta_k^l)}{\sqrt{1+d_k^{\alpha_{NLOS}}}}, \tag{46}$$

  - L is the number of multi-paths,
  - θ_k^l denotes the normalized direction of the l-th path,
  - d_k denotes the distance between the transceivers,
  - α_{NLOS} and α_{LOS} denote the path loss exponents,
  - $a_{k,l}$ denotes the complex gain and $a_{k,l} \sim CN(0,1)$.
  - and the normalized channel vector $\mathbf{a}(\theta)$ is given by

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}}\begin{bmatrix}1 & e^{-j\pi\theta} & \cdots & e^{-j\pi(M-1)\theta}\end{bmatrix}^T. \tag{47}$$

System Model - mmWave Channel Model (2/2)

- As discussed in the literature, the path loss of the NLOS components will be much more significant than that of the LOS component
- Therefore, the first term on the right-hand side of (46) is dominant, which yields the following simplified channel model

$$\mathbf{h}_k = \sqrt{M}\frac{a_k \mathbf{a}(\theta_k)}{\sqrt{1+d_k^{\alpha}}}, \tag{48}$$

where the subscripts of 0 and LOS have been omitted to simplify the notation.

- If two users have similar θ_k, their channels are highly correlated.

Random Beamforming for mmWave-NOMA (1/3)

- Many existing precoding and beamforming schemes for NOMA require global CSI at the base station.
- In order to reduce the system overhead, we consider the application of random beamforming to mmWave-NOMA
- For now, we focus on the case that a single beam, denoted by $\mathbf{p}$, is generated at the base station.
- Since analog precoding is preferable for mmWave systems, we use the following beamforming vector

$$\mathbf{p} = \mathbf{a}(\bar{\theta}), \tag{49}$$

  where $\bar{\theta}$ is uniformly distributed between -1 and 1.
- One straightforward solution for user scheduling is
  - Ask each user to feed its effective channel gain $|\mathbf{h}_j^H \mathbf{p}|^2$ back to the base station.
  - The base station schedules the user which has the strongest channel condition.

Random Beamforming for wwWave-NOMA (2/3)

- However, such an approach will still introduce a considerable system overhead, particularly if there are a lot of users in the cell.
- For mmWave, many users do not have to participate in the access competition, as explained in the following.
- Without loss of generality, user m is randomly chosen to be served on beam $\mathbf{p}$.
- The effective channel gain of this user can be written as follows:

$$|\mathbf{h}_j^H \mathbf{p}|^2 = M \frac{|a_j|^2 |\mathbf{p}^H \mathbf{a}(\theta_j)|^2}{1 + d_j^\alpha} \tag{50}$$

$$= \frac{|a_j|^2 \left| \sum_{l=0}^{M-1} e^{-j\pi l(\bar{\theta} - \theta_m)} \right|^2}{M(1 + d_j^\alpha)}.$$

Random Beamforming for wwWave-NOMA (3/3)

- This effective channel gain can be rewritten as follows:

$$|\mathbf{h}_j^H \mathbf{p}|^2 = \frac{|a_j|^2 \sin^2\left(\frac{\pi M(\bar{\theta}-\theta_j)}{2}\right)}{M(1+d_j^\alpha)\sin^2\left(\frac{\pi(\bar{\theta}-\theta_j)}{2}\right)} = \frac{|a_j|^2}{(1+d_j^\alpha)} F_M\left(\pi[\bar{\theta}-\theta_j]\right), \tag{51}$$

  where $F_M(x)$ denotes the Fejér kernel.
- Note that a Fejér kernel goes to zero quickly for increasing argument, i.e., $F_M(x) \to 0$ for $x \to \infty$.
- This means that a user can have a large channel gain if this user's channel vector is aligned with the direction of the beam.
- Therefore, we will schedule only the users who are located in the circular sector shown in the previous figure.

Implementation of NOMA (1/3)

- Consider that there are K users in the sector, $\mathcal{D}_\theta$,
- These users are ordered according to their effective channel gains as follows:

$$|\mathbf{h}_1^H\mathbf{p}|^2 \leq \cdots \leq |\mathbf{h}_K^H\mathbf{p}|^2. \tag{52}$$

- We assume that user i and user j, $1 \leq i < j \leq K$, are paired together for NOMA transmission on the randomly generated beam,
- The signal sent by the base station is given by

$$\mathbf{p}\left(\beta_i s_i + \beta_j s_j\right), \tag{53}$$

where β_i denotes the power allocation coefficient.
- Since $|\mathbf{h}_i^H\mathbf{p}|^2 < |\mathbf{h}_j^H\mathbf{p}|^2$, the application of NOMA means $\beta_i \geq \beta_j$, where $\beta_i^2 + \beta_j^2 = 1$.

Implementation of NOMA (2/3)

- Therefore, user i will receive the following observation

$$y_i = \mathbf{h}_i^H \mathbf{p} \left(\beta_i s_i + \beta_j s_j \right) + n_i. \tag{54}$$

- User i will treat its partner's message as noise and directly decode its information with the following SINR:

$$\text{SINR}_i = \frac{|\mathbf{h}_i^H \mathbf{p}|^2 \beta_i^2}{|\mathbf{h}_i^H \mathbf{p}|^2 \beta_j^2 + \frac{1}{\rho}}, \tag{55}$$

where ρ denotes the transmit SNR.

- As a result, the outage probability for user i to decode its information is given by

$$\mathrm{P}_{i|K}^{o} = \mathrm{P}\left(\log(1 + \text{SINR}_i) < R_i | K \right) = \mathrm{P}\left(\text{SINR}_i < \epsilon_i | K \right),$$

which is conditioned on the number of users in $\mathcal{D}_\theta$, where $\epsilon_i = 2^{R_i} - 1$.

Implementation of NOMA (3/3)

- User j first tries to decode its partner's message with the following SINR, $\text{SINR}_{i\to j} = \frac{|\mathbf{h}_j^H\mathbf{p}|^2\beta_i^2}{|\mathbf{h}_j^H\mathbf{p}|^2\beta_j^2+\frac{1}{\rho}}$.
- If $\text{SINR}_{i\to j} \geq \epsilon_i$, the user can decode its own message with the following SNR

$$\text{SINR}_j = \rho|\mathbf{h}_j^H\mathbf{p}|^2\beta_j^2. \tag{56}$$

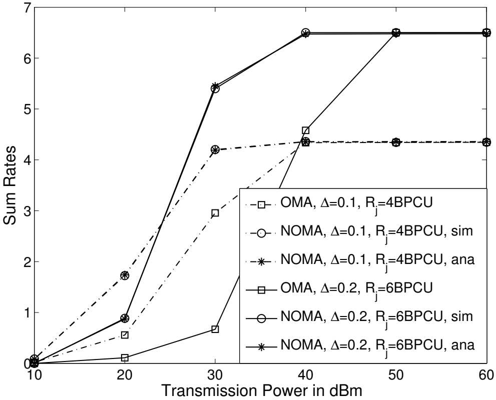
- Therefore, the conditional outage probability at user j is

$$\mathrm{P}^o_{j|K} = 1 - \mathrm{P}\left(\text{SINR}_{i\to j} > \epsilon_i, \text{SINR}_j > \epsilon_j|K\right). \tag{57}$$

- As a result, the outage sum rate is given by

$$R_{sum}^{NOMA} = \mathrm{P}(K=1)(1-\mathrm{P}^{1|K}_{OMA})R_1 + \sum_{k=2}^{\infty}\mathrm{P}(K=k) \times \left((1-\mathrm{P}^o_{i|K})R_i + (1-\mathrm{P}^o_{j|K})R_j\right). \tag{58}$$

Numerical Results



Sum Rates achieved by the OMA and NOMA schemes. $M = 4$, $\lambda = 1$, $\Delta = 0.1$, $R_i = 0.5$ BPCU, $i = 4$ and $j = 1$

- In general, mmWave-NOMA outperforms mmWave-OMA,
- When $R_j = 6$ BPCU and the transmission power is 30 dBm, the performance gain of the NOMA scheme over OMA is 5 BPCU.
- When the transmission power is very high, no outage will happen and the two schemes achieve the same outage sum rate.

Outline

Finite Resolution Analog Beamforming (1/3)

- A recent development in massive MIMO and mmWave networks is the use of finite resolution analog beamforming, which reduces hardware costs.
- Analog beamforming has the property that it does not alter the amplitude of a signal, but modifies its phase only, which is different from digital beamforming.
- The finite resolution constraint on analog beamforming is due to the fact that the number of phase shifts supported by a practical circuit is finite.

[22] Z. Ding, L. Dai, R. Schober and H. V. Poor, "NOMA meets finite resolution analog beamforming", IEEE CL, 2017.

Finite Resolution Analog Beamforming (2/3)

- Consider an example in which the users in a cell are divided into two groups, denoted by $\mathcal{S}_1$ and $\mathcal{S}_2$.
- The users in $\mathcal{S}_1$ have strict QoS requirements to meet, and the users in $\mathcal{S}_2$ can be served opportunistically.
- Take user 1 from $\mathcal{S}_1$ as an example.
- Depending on the values of this user's complex-valued channel coefficients, 1 or -1 will be chosen as the beamformer elements, as shown in the table.

Table 1: An example for finite resolution analog beamforming

	user 1 in $\mathcal{S}_1$	user 2 in $\mathcal{S}_1$	user 1 in $\mathcal{S}_2$	user 2 in $\mathcal{S}_2$
channel vectors	-0.19 + 0.66j -0.06 - 0.53j 0.34 - 0.03j 0.31 - 0.18j	-0.49 + 0.16j -0.35 + 0.22j -0.10 - 0.62j -0.06 + 0.41j	-0.27 - 0.11j -0.06 + 0.58j 0.31 - 0.05j 0.34 - 0.60j	-0.33 + 0.25j -0.45 + 0.10j -0.20 + 0.59j -0.45 - 0.15j
FRAB beam-formers	-1 -1 1 1	-1 -1 - 1 - 1	-1 -1 1 1	-1 -1 -1 -1

Finite Resolution Analog Beamforming (3/3)

- The use of finite resolution analog beamforming provides an opportunity for the implementation of NOMA.
- Again consider the example shown in the table. The base station forms two beams according to the channel state information (CSI) of the two users in $\mathcal{S}_1$.
- Because of the finite resolution of analog beamforming, these formed beams are also preferred by the users in $\mathcal{S}_2$, even though the users in the two groups have different CSI.

Table 2: An example for finite resolution analog beamforming

	user 1 in $\mathcal{S}_1$	user 2 in $\mathcal{S}_1$	user 1 in $\mathcal{S}_2$	user 2 in $\mathcal{S}_2$
channel vectors	-0.19 + 0.66j -0.06 - 0.53j 0.34 - 0.03j 0.31 - 0.18j	-0.49 + 0.16j -0.35 + 0.22j -0.10 - 0.62j -0.06 + 0.41j	-0.27 - 0.11j -0.06 + 0.58j 0.31 - 0.05j 0.34 - 0.60j	-0.33 + 0.25j -0.45 + 0.10j -0.20 + 0.59j -0.45 - 0.15j
FRAB beam-formers	-1 -1 1 1	-1 -1 - 1 - 1	-1 -1 1 1	-1 -1 -1 -1

System Model (1/2)

- Consider a NOMA downlink scenario, where the base station is equipped with M antennas.
- Assume that there are two groups of single-antenna users in the network.
- Denote by $\mathcal{S}_1$ a group containing users with strict quality of service (QoS) requirements, whose distances to the base station are denoted by d_{yk} and are assumed to be fixed.
- Denote by $\mathcal{S}_2$ a group of users to be served opportunistically, and these users are uniformly located in a disk-shaped area with radius r_1, where the base station is at its center.
- Denote the distances of users in $\mathcal{S}_2$ by d_{xi}.
- The $M \times 1$ channel vector of a user in $\mathcal{S}_1$ ($\mathcal{S}_2$) is denoted by $\mathbf{h}_k$ ($\mathbf{g}_i$).

System Model (2/2)

- Two types of channel models are considered here.
- The first one is based on the Rayleigh fading model.
- The second one is based on the mmWave model, i.e., a channel vector can be expressed as follows:

$$\mathbf{h}_k = \frac{a_k}{1 + d_{yk}^{\alpha}} \begin{bmatrix} 1 & e^{-j\pi\theta_k} & \cdots e^{-j\pi(M-1)\theta_k} \end{bmatrix}^T, \tag{59}$$

where α denotes the path loss exponent, θ_k is the normalized direction, and a_k denotes the fading attenuation coefficient. Note that for the purpose of illustration, only the line-of-sight path is considered for the mmWave model.

Implementation of Finite Resolution Analog Beamforming

- Suppose that the users in $\mathcal{S}_1$ are served with finite resolution analog beamforming.
- Denote by $\mathbf{f}_k$ the $M \times 1$ beamforming vector for user k, where each element of the beamforming vector is drawn from the following vector:

$$\bar{\mathbf{f}} = \begin{bmatrix} 1 & e^{j\frac{2\pi}{N_q}} & \dots & e^{j\frac{(N_q-1)2\pi}{N_q}} \end{bmatrix}, \tag{60}$$

  where N_q denotes the number of supported phase shifts.
- Therefore, the i-th element of $\bar{\mathbf{f}}$ is chosen as the m-th element of $\mathbf{f}_k$ based on the following criterion:

$$i_{k,m}^{*} = \underset{i \in \{1,\cdots,N_q\}}{\arg\min} \left| \bar{f}_i - \frac{h_{k,m}}{|h_{k,m}|} \right|^2, \tag{61}$$

  where $\bar{f}_i$ denotes the i-th element of $\bar{\mathbf{f}}$, and $h_{k,m}$ denotes the m-th element of user k's channel vector.

Implementation of NOMA (1/2)

- Suppose that only one user from $\mathcal{S}_2$ will be chosen to be paired with user k from $\mathcal{S}_1$ and denote this user by user i_k^*.
- The base station broadcasts a superposition of two users' messages on each beam.
- User k in $\mathcal{S}_1$ treats its partner's message as noise and decodes its own message with the following SINR:

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{f}_k|^2 \alpha_{0,k}^2}{|\mathbf{h}_k^H \mathbf{f}_k|^2 \alpha_{1,k}^2 + \sum_{l \in \mathcal{S}_1 \setminus k} |\mathbf{h}_k^H \mathbf{f}_l|^2 + \frac{M}{\rho}}, \tag{62}$$

where the factor M is shown in the denominator in order to ensure the transmission power normalization, and the power allocation coefficients are denoted by $\alpha_{n,k}$.

- Note that $\sum_{n=0}^{1} \alpha_{n,k}^2 = 1$ and $\alpha_{0,k} \geq \alpha_{1,k}$.

Implementation of NOMA (2/2)

- By applying SIC, user i_k^* can decode its partner's message by treating its own message as noise.
- If successful, user i_k^* can remove intra-NOMA interference and decode its own message with the following SINR

$$\mathrm{SINR}_k^{i_k^*} = \frac{|\mathbf{g}_{i_k^*}^H \mathbf{f}_k|^2 \alpha_{1,k}^2}{\sum_{l \in \mathcal{S}_1 \backslash k} |\mathbf{g}_{i_k^*}^H \mathbf{f}_l|^2 + \frac{M}{\rho}}. \tag{63}$$

- We use the following user selection criterion:

$$i_k^* = \arg\max\{\mathrm{SINR}_k^{k\to 1}, \cdots, \mathrm{SINR}_k^{k\to|\mathcal{S}_2|}\}. \tag{64}$$

- Note that this criterion selects a user which maximizes the probability for successfully removing the intra-NOMA interference, a key stage for SIC.

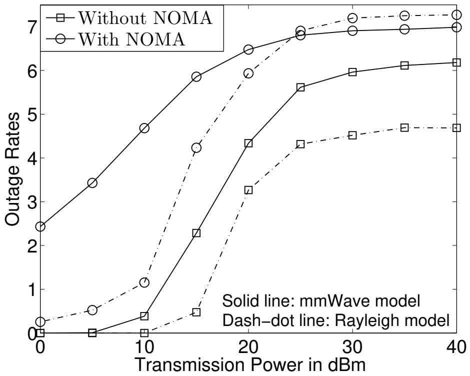
Performance Analysis

- With finite resolution analog beamforming, the user's effective channel gain can be expressed as follows:

$$|\mathbf{h}_k^H \mathbf{f}_k|^2 = \left| \sum_{m=1}^{M} h_{k,m} e^{-j \frac{(i_{k,m}^* - 1)2\pi}{N_q}} \right|^2 . \qquad (65)$$

- This expression is difficult to analyze.
- For a special case with one bit resolution, one beam, and Rayleigh fading, we can draw the following conclusions
    - The use of one bit resolution analog beamforming achieves a diversity gain of $\frac{M+1}{2}$ for the user in $\mathcal{S}_1$.
    - Recall that the full diversity gain is M since the base station has M antennas.
    - This result also applies to scenarios without NOMA.

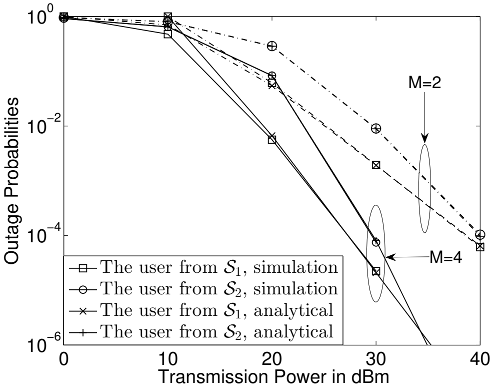
Numerical Results (1/2)



Outage rates achieved for different channel models. $M = 30$, $|\mathcal{S}_1| = 3$, $|\mathcal{S}_2| = 300$, $r_1 = 40$m, $r_y = r_1$, $\alpha = 3$, $R_0 = 1$ BPCU and $R_1 = 1.5$ BPCU.

- For OMA, a user's targeted data rate is $R_0 + R_1$.
- The use of NOMA can result in a significant performance gain compared to OMA.
- When the transmission power is 45dBm, the use of NOMA can offer a rate improvement of 4 BPCU over OMA, for both of the channel models.

Numerical Results (2/2)



Outage probabilities achieved for Rayleigh fading channel model. $|\mathcal{S}_1| = 1$, $|\mathcal{S}_2| = M$, $r_1 = 40$m, $r_y = \frac{r_1}{2}$, $\alpha = 3$, $R_0 = 1$ BPCU and $R_1 = 1$ BPCU.

- To facilitate this diversity analysis, we set $|\mathcal{S}_2| = M$, which means that the diversity gains for the users in $\mathcal{S}_1$ and $\mathcal{S}_2$ are $\frac{M+1}{2}$ and M, respectively.
- If perfect analog beamforming is used, a diversity gain of M is achieved by the user from $\mathcal{S}_1$.
- From the figure, one can clearly observe the loss of diversity gain for the user from $\mathcal{S}_1$.

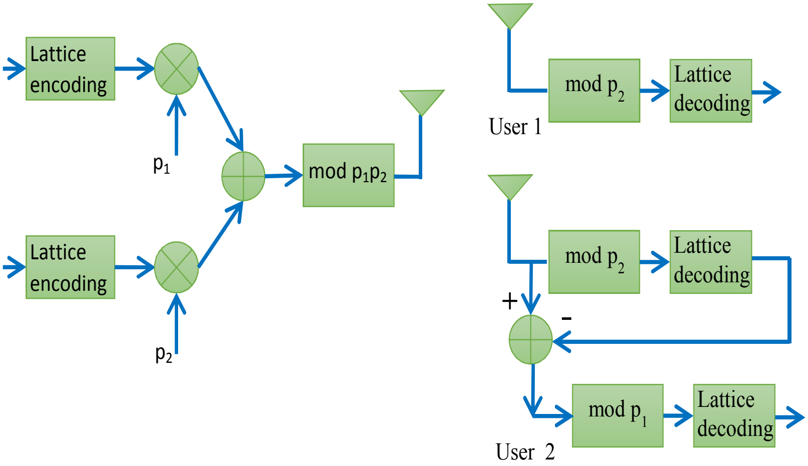
Outline

Coding and Modulation for NOMA (1/2)

- Effective channel coding and modulation schemes are crucial for NOMA, in order to ensure that the achievable rates predicted by theory can be realized in practice.
- More importantly, the integration of sophisticated channel codes with NOMA has also led to the development of new forms of NOMA.
- For example, a new form of NOMA, called lattice partition multiple access (LPMA), employs properties of lattice codes.
  - Consider a downlink scenario with two users.
  - Lattice encoding is applied at the transmitter.
  - The modulo operation is used to remove multiple access interference.

[23] D. Fang, Y.-C. Huang, Z. Ding, G. Geraci, S.-L. Shieh, and H. Claussen, "Lattice partition multiple access: A new method of downlink non-orthogonal multiuser transmissions," IEEE Globalcom, 2016.

Coding and Modulation for NOMA (2/2)



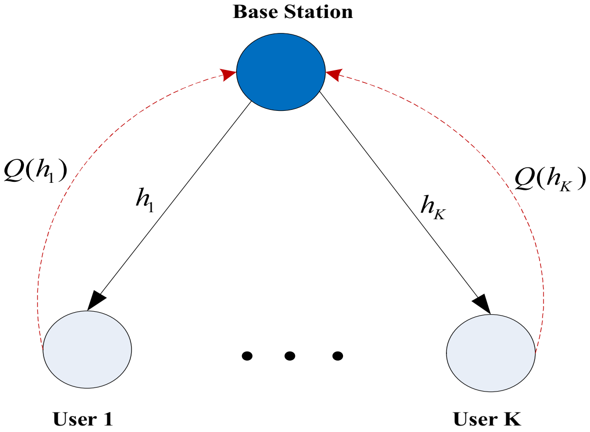
Outline

Imperfect Channel State Information (1/4)

- Imperfect CSI is one of the key obstacles in realizing the performance gain of NOMA in practice.
- It is worth pointing out that many NOMA designs have relatively weak requirements on the CSI.
  - For power domain NOMA, the base station needs to know only the ordering of the channels, but the exact values of the channel gains are not needed.
  - For ZF-MIMO-NOMA, the base station does not need to know users' channel matrices.
- The following types of CSI are useful in realizing the potential of NOMA in practice
  - CSI with channel estimation error [24]
  - Partial CSI, e.g., users' distance information [24]
  - Limited feedback

[24] Yang, Z. Ding, P. Fan and G. Karagiannidis, "On the Performance of Non-Orthogonal Multiple Access Systems with Partial Channel Information", IEEE TVT, 2016.

Imperfect Channel State Information (2/4)



Downlink NOMA with one-bit feedback.

[25] P. Xu, Y. Yuan, Z. Ding, X. Dai and R. Schober, "On the Outage Performance of Non-Orthogonal Multiple Access with One-Bit Feedback", IEEE TWC, 2016.

Imperfect Channel State Information (3/4)

- Prior to data transmission, the base station transmits one message α to each user in each block.
- User k feeds back in each fading block a single bit "$Q(h_k)$" to the base station, where $Q(h_k) = 1$ if $|h_k|^2 \geq \alpha$, and $Q(h_k) = 0$, otherwise
- Based on the feedback information $\{Q(h_k)\}$, the base station groups the users, and denotes the groups of the users with feedback "0" and "1" as $\mathcal{G}_{0|n}$ and $\mathcal{G}_{1|n}$, respectively.
- There is an ambuiguity about the users' CSI
  - The base station knows which group a user is in.
  - The base station cannot distinguish the users in the same group, which causes a problem for user ordering.

Imperfect Channel State Information (4/4)

- Because of this CSI ambiguity, a few challenging research problems rise
  - How to preserve the diversity gain?
    - A carefully chosen threshold can ensure that the same diversity gain as for perfect CSIT can be realized with one-bit feedback.
  - How to carry out power allocation?
    - Probabilistic optimization problems can be formulated based on either a short term power constraint or a long term power constraint.

---

[25] P. Xu, Y. Yuan, Z. Ding, X. Dai and R. Schober, "On the Outage Performance of Non-Orthogonal Multiple Access with One-Bit Feedback", IEEE TWC, 2016.
[26] Z. Ding, P. Fan and H. V. Poor, "Random Beamforming in Millimeter-Wave NOMA Networks", IEEE Access, 2017.

Outline

SWIPT—Background (1/2)

Wireless Energy Transfer (WET)

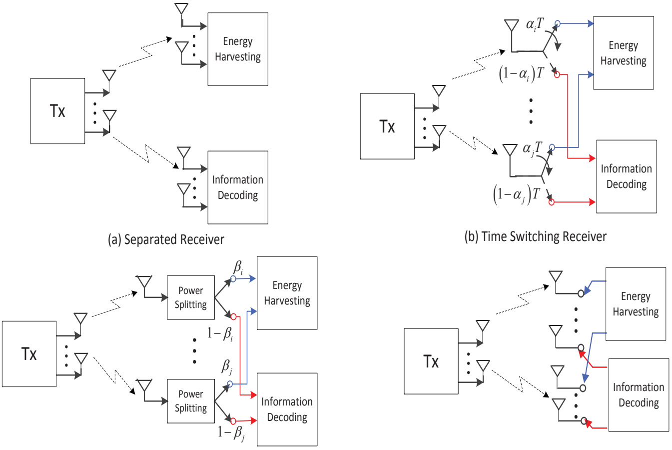
- Key Idea: Energy is transmitted from a power source to a destination over the wireless medium.
- Motivation: 1) Ambient radio frequency signals are everywhere; 2) WET could be the only means to increase the lifetime of energy constrained networks
- Tesla had already provided a successful demonstration to light electric lamps wirelessly in 1891, but WET has been forgotten for a long time due to its low energy efficiency.

What has changed now?

- More low power devices.
- Advanced antenna techniques for better energy efficiency.

[27]Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative Non-Orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer", IEEE JSAC, 2016.

SWIPT—Background (2/2)



(a) Separated Receiver

(b) Time Switching Receiver

(c) Power Splitting Receiver

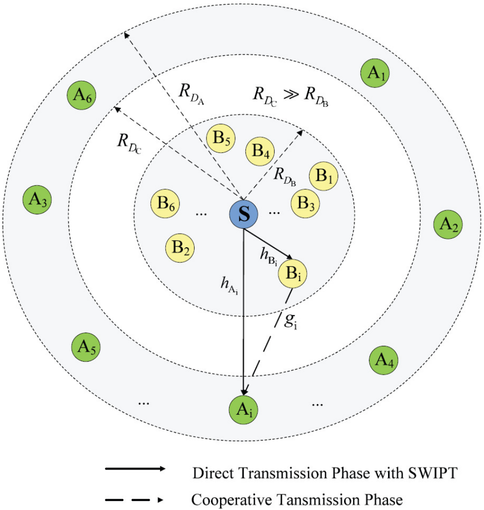
(d) Antenna Switching Receiver

[28]Z. Ding, C. Zhong, D. W. Ng, M. Peng, H. A. Suraweera, R. Schober and H. V. Poor, "Application of Smart Antenna Technologies in Simultaneous Wireless Information and Power Transfer", IEEE Commun. Magazine, 2015.

Motivation for SWIPT + Cooperative NOMA

- To improve the reliability of the far NOMA users without draining the near users' batteries, we consider the application of SWIPT to NOMA, where SWIPT is performed at the near NOMA users.
- Therefore, the aforementioned two communication concepts, cooperative NOMA and SWIPT, can be naturally combined.
- Cooperative SWIPT NOMA – a new wireless multiple access protocol that is both spectrally efficient and energy efficient .

Network Model



- Illustration of a downlink SWIPT NOMA system with one base station S (blue circle). The spatial distributions of the near users (yellow circles) and the far users (green circles) follow homogeneous PPPs.

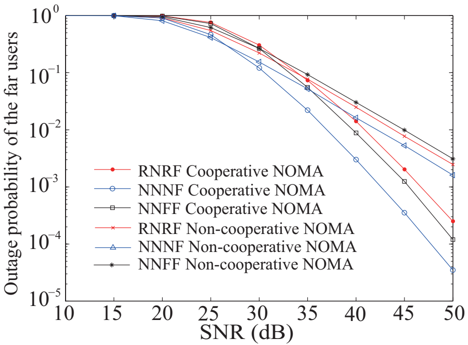
Non-Orthogonal Multiple Access with User Selection

A natural question arises: which near NOMA user should help which far NOMA user?

To investigate the performance of one pair of selected NOMA users, three opportunistic user selection schemes are studied, based on the locations of users to perform NOMA as follows:

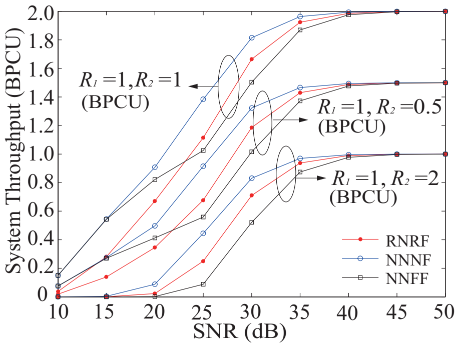
- Random near user and random far user (RNRF) selection, where both the near and far users are randomly selected from the two groups.
- Nearest near user and nearest far user (NNNF) selection, where a near user and a far user closest to the BS are selected from the two groups.
- Nearest near user and farthest far user (NNFF) selection, where a near user which is closest to the BS is selected and a far user which is farthest from the BS is selected.

Numerical Results (1/2)



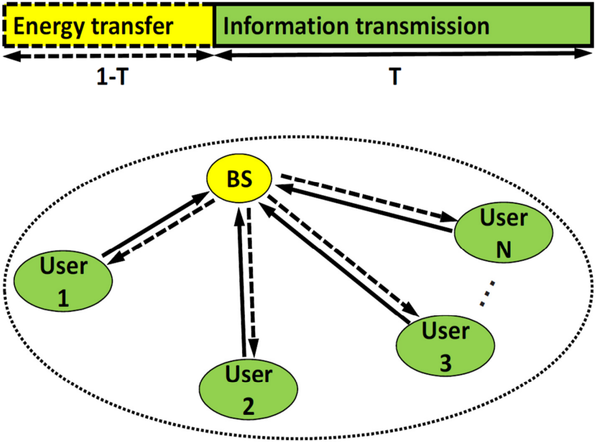
- Cooperative NOMA has a larger slope than non-cooperative NOMA.
- NNNF achieves the lowest outage probability.
- NNFF has higher outage probability than RNRF in non-cooperative NOMA, however, it achieves a lower outage probability than RNRF in cooperative NOMA.

Numerical Results (2/2)



- NNNF achieves the highest throughput since it has the lowest outage probability.
- There are ceilings in the high SNR region.
- Increasing R_2 from $R_2 = 0.5$ BPCU to $R_2 = 1$ BPCU can improve the throughput; however, for the case $R_2 = 2$ BPCU, the throughput is reduced.

Other types of SWIPT + NOMA



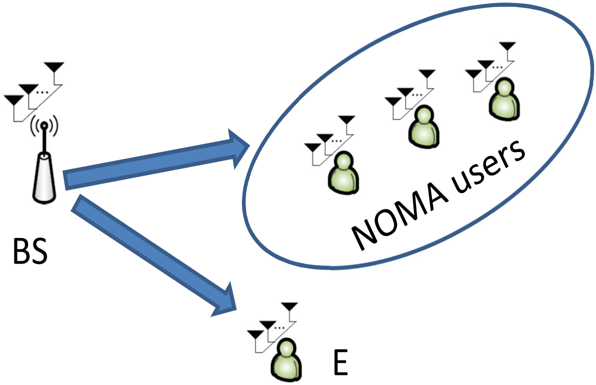
[29] P. D. Diamantoulakis, K. N. Pappi, Z.Ding, and G. K. Karagiannidis, "Wireless Powered Communications with Non-Orthogonal Multiple Access", IEEE TWC, 2016.

Outline

Security Provisioning for NOMA

- Security provisioning was not considered when the NOMA principle was developed first, similar to other multiple access techniques.
  - The strong NOMA user needs to decode the weak user's message in order to carry out SIC.
  - Such a security risk also exists for other multiple access techniques, e.g., a TDMA user can switch on during a time slot not allocated to it and attempt to decode another user's information.
  - In the current telecommunication systems, security is ensured based on encryption, instead of relying on multiple access techniques.
- In general, the use of the NOMA principle facilitates the implementation of physical layer security.

Security Provisioning for NOMA with External Eavesdroppers



[30] Y. Zhang, H. Wang, Q. Yang and Z. Ding, "Secrecy Sum Rate Maximization in Non-Orthogonal Multiple Access", IEEE CL, 2016.

Security Provisioning for NOMA with Untrusted Users

- It is difficult to prevent the user with a strong channel from eavesdropping.
- But with the help of NOMA power allocation and beamforming, the channel conditions of malicious and legitimate users can be controlled:
  - Consider a BS having two types of messages.
  - The multicasting message is for all the users, and the unicasting message is for a particular user.
  - A multicasting receiver is a potential eavesdropper for unicasting.
  - The use of NOMA improves spectral efficiency since two time slots are needed by OMA
  - The security of unicasting is also improved as power allocation and beamforming are designed to increase the difference between the users' effective channel gains.

[31] Z. Ding, Z. Zhao, M. Peng and H. V. Poor, "On the Spectral Efficiency and Security Enhancements of NOMA Assisted Multicast-Unicast Streaming", TCOM, 2017.

Outline

Promising Research Challenges

- Different variants of NOMA
- New coding and modulation for NOMA
- Hybrid multiple access
- User pairing/clustering
- Coexistence between mmWave and NOMA
- MIMO and cooperative NOMA
- Interplay between NOMA and cognitive radio
- Imperfect CSI and limited channel feedback
- Security provisioning in NOMA
- Efficient resource allocation for NOMA
- Implementation issues of NOMA
- Emerging applications of NOMA

Thank you for your attention!

[32] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends, IEEE JSAC, 2017
[33] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C-L. I and H. V. Poor, "Application of Non-orthogonal Multiple Access in LTE and 5G Networks", IEEE Communication Magazine, 2017.
[34] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in nonorthogonal multiple access networks for 5G and beyond", IEEE Networks, 2017

Email: zhiguo.ding@manchester.ac.uk, liangwei@nwpu.edu.cn.