NOMA: Principles and Recent Results

Jinho Choi
School of EECS
GIST

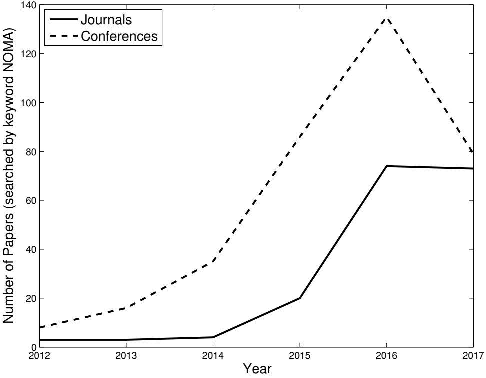
September 2017 (VTC-Fall 2017)

Abstract: Non-orthogonal multiple access (NOMA) becomes a key technology in 5G as it can improve the spectral efficiency in cellular systems by exploiting the power difference between users. In this talk, we review the principles of NOMA and discuss existing approaches as well as new results. In particular, we present optimization problems for resource allocation in downlink NOMA under various conditions and discuss NOMA based random access that can be used for uplink NOMA to support massive connectivity which is vital for machine-type communications in 5G.

1. Introduction

- Non-orthogonal multiple access (NOMA) is considered a key technology for 5G systems.
- However, NOMA is not a new concept and has been studied under different names from an information-theoretic point of view:
  - superposition coding (SC)
  - successive interference cancellation (SIC)
  - stripping, and onion peeling
- In particular, the notion of SC/SIC is employed to find the capacity of broadcast channel (for details, see Cover and Thomas, *Elements of Information Theory*).

- The application of NOMA to cellular systems is relatively new and has attracted a lot of attention recently.

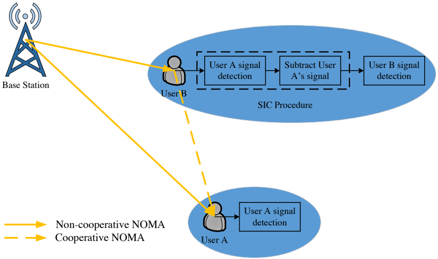


IEEE papers only (from ©IEEE Xplore) (2017: 2017 Jan. - Aug.)

- NOMA in cellular systems is to actively exploit power difference.
- Thus, the power allocation plays a crucial role in improving the spectral efficiency in conjunction with user allocation (or grouping).
- In standards, NOMA has been implemented by multi-user superposition transmission (MUST) schemes for downlink transmissions:

  3GPP R1-154999, TP for classification of MUST schemes, 2015.

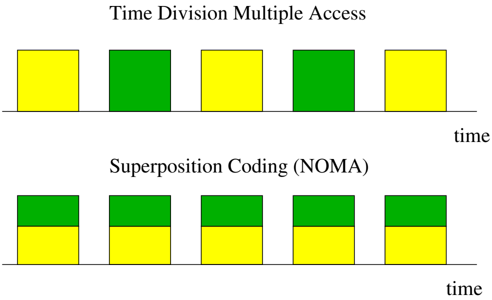
- A typical scenario is to support two users.
- One user is close to a base station (BS), called user 1.
- The other user is far away from the BS, called user 2.
- In NOMA, the BS supports two users with a shared channel.



This figure is taken from "Application of Non-orthogonal Multiple Access in LTE and 5G Networks, by Z. Ding, Y.Liu, J.Choi, M. Elkashlan, C.-L. I,and H.V. Poor, IEEE Commun. Magazine, February 2017.

2. NOMA and OMA

Example: downlink to support two users



In TDMA (as OMA), s_1 and s_2 are transmitted in different time slots. On the other hand, in OMA, $s_1 + s_2$ is transmitted in a shared time slot.

Capacity region of NOMA

- The signals received by users are

$$y_1 = h_1(s_1 + s_2) + n_1$$
$$y_2 = h_2(s_1 + s_2) + n_2,$$

where s_2 is usually stronger than s_1 as user 2 is far away from the BS.

- At user 1, s_2 can be decoded first and subtracted from y_1. Then, the signal becomes interference-free:

$$y_1 - h_1 s_2 = h_1 s_1 + n_1.$$

- Superposition coding: each signal is independently encoded at code rate R_k and the sum of two signals is transmitted as

$$s_1 + s_2.$$

- Decoding s_2 at user 1:

$$y_1 = \underbrace{h_1 s_2}_{=\text{signal}} + \underbrace{h_1 s_1 + n_1}_{=\text{interference+noise}}$$

- A condition based on information-theoretic view for successful decoding:

$$R_2 < \log_2\left(1 + \frac{|h_1|^2 P_2}{|h_1|^2 P_1 + N_0}\right),$$

where $P_k = \mathbb{E}[|s_k|^2]$.

- This condition is valid if each signal is encoded by capacity achieving code and s_k is Gaussian.
- The resulting assumption is *Gaussian codebook*.
- This simplifies the analysis. However, in practice, this assumption is not valid (due to finite-length codes and non-Gaussian signal constellation).
- After removing s_2 (by SIC), user 1 can decode s_1 if

$$R_1 < \log_2\left(1 + \frac{|h_1|^2 P_1}{N_0}\right).$$

- At user 2, no SIC is used as s_1 is weaker than s_2. Thus,

$$R_2 < \log_2\left(1 + \frac{|h_2|^2 P_2}{|h_2|^2 P_1 + N_0}\right).$$

- In summary, NOMA capacity is given by

$$\begin{aligned} R_1 &< \log_2\left(1+\frac{|h_1|^2 P_1}{N_0}\right) \text{ at user 1} \\ R_2 &< \log_2\left(1+\frac{|h_1|^2 P_2}{|h_1|^2 P_1 + N_0}\right) \text{ at user 1} \\ R_2 &< \log_2\left(1+\frac{|h_2|^2 P_2}{|h_2|^2 P_1 + N_0}\right) \text{ at user 2.} \end{aligned}$$

- The second and third regions can be combined as

$$R_2 < \log_2\left(1+\frac{\min\{|h_1|^2, |h_2|^2\} P_2}{\min\{|h_1|^2, |h_2|^2\} P_1 + N_0}\right).$$

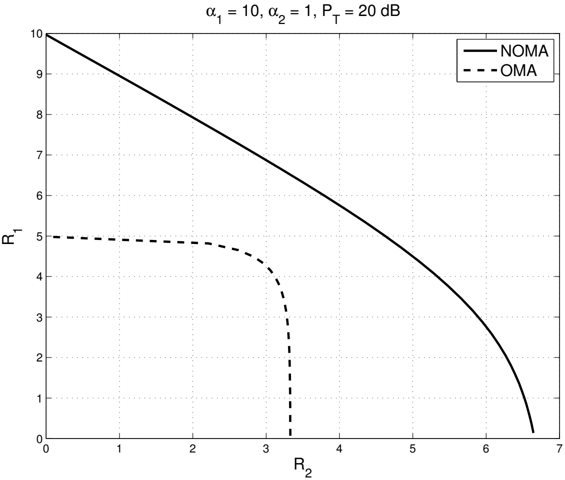
- For comparison, we consider TDMA for OMA. The capacity region becomes

$$\begin{aligned} R_1 &< \frac{1}{2}\log_2\left(1+\frac{|h_1|^2 P_1}{N_0}\right) \\ R_2 &< \frac{1}{2}\log_2\left(1+\frac{|h_2|^2 P_2}{N_0}\right). \end{aligned}$$

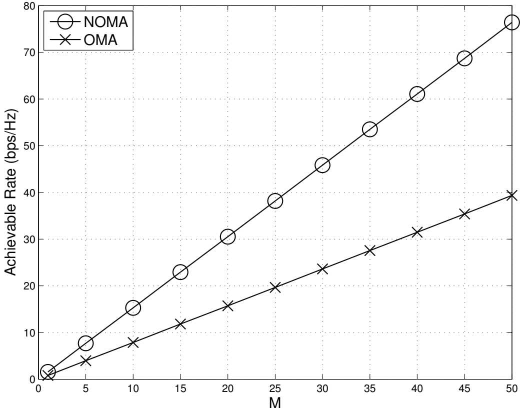
- In TDMA, to share a given channel between two users, the time slot becomes halved.
- For convenience, let

$$\alpha_k = |h_k|^2.$$

Capacity Regions of NOMA and OMA



MIMO-Capacity[1] Regions of NOMA and OMA

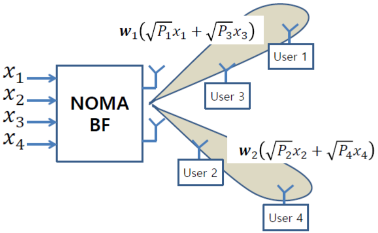


Scaling property of NOMA with spatial multiplexing.

---

[1] See Q. Sun, et al., "On the ergodic capacity of MIMO NOMA systems," IEEE Wireless Commun. Letters, Aug 2015.

3. Downlink Beamforming with NOMA



- In NOMA, a group of users can share the same beam.
- In general, the users in a group or cluster have similar spatial channels.
- Beams are designed to minimize the inter-cluster interference.
- NOMA is used within a cluster for multiple access with the power difference.

- NOMA beamforming with user clustering has been widely[2] studied.
- Suppose that a cluster consists of two users and the channel vectors are denoted by $\{\mathbf{h}_{m,1}, \mathbf{h}_{m,2}\}$ for the mth cluster.
- The beamforming vector for cluster m is denoted by $\mathbf{w}_m$.
- ZF inter-cluster beamforming:

$$\mathbf{w}_m \perp \mathrm{Range}(\mathbf{H}_{-m}), \ m = 1, \ldots, M,$$

where M is the number of clusters and

$$\mathbf{H}_{-m} = [\mathbf{H}_1 \ \ldots \mathbf{H}_{m-1} \ \mathbf{H}_{m+1} \ \ldots \ \mathbf{H}_M].$$

---

[2] For example, see B. Kim, et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in MILCOM 2013.

Minimum Transmit Power NOMA Beamforming

- Problem:

$$\min \sum_m P_{m,1} + P_{m,2}$$
$$\text{subject to } \mathrm{SINR}_{m,i} \ge G_{m,i}, \ i = 1, 2,$$

  where $G_{m,i}$ is the target SINR for user i in cluster m.
- A two-step approach:
  - (Inner Problem) For given $\mathbf{H}_m = [\mathbf{h}_{m,1} \ \mathbf{h}_{m,2}]$, $m = 1, \ldots, M$, find the optimal beams, $\mathbf{w}_m$, that minimize the total transmission powers.
  - (Outer Problem) Find the user clustering to minimize the total transmission power.
- In general, the inner problem can be easily solved, while the outer problem is difficult as it is a combinatorial problem.

- The solution to the inner problem (for a given user clustering) can found by noting that
$$\mathbf{w}_m \in \mathrm{Span}(\mathbf{P}_m \mathbf{h}_{m,1}, \mathbf{P}_m \mathbf{h}_{m,2}),$$
where $\mathbf{P}_m$ is the orthogonal projection that is given by
$$\mathbf{P}_m = \mathbf{I} - \mathbf{H}_{-m}(\mathbf{H}_{-m}^{\mathrm{H}} \mathbf{H}_{-m})^{-1} \mathbf{H}_{-m}^{\mathrm{H}}.$$
The orthogonal project is required for ZF with respect to inter-cluster interference.
- Then, the weight vector can be obtained to minimize the transmission power to cluster m subject to
$$\begin{aligned} |\mathbf{h}_1^{\mathrm{H}} \mathbf{w}|^2 &\ge |\mathbf{h}_2 \mathbf{w}|^2 \\ |\mathbf{h}_1^{\mathrm{H}} \mathbf{w}|^2 P_1 &\ge G_1 \sigma^2 \\ |\mathbf{h}_2^{\mathrm{H}} \mathbf{w}|^2 P_2 &\ge (|\mathbf{h}_2^{\mathrm{H}} \mathbf{w}|^2 P_1 + \sigma^2) G_2 \end{aligned}$$

- To solve the outer problem, for M clusters, the number of possible combinations is given by

$$\binom{2M}{2}\binom{2(M-1)}{2}\cdots\binom{2}{2} = \frac{(2M)!}{2^M}.$$

- For each user clustering, we can find the total transmission power. Then, we can choose the user clustering that minimizes the total transmission power.
- However, for a large M, this straightforward approach becomes computationally infeasible.
- To avoid a high computationally complexity, we can consider a greedy approach.

- For a greedy algorithm, let

$$||\mathbf{h}_{k(1)}||^2 \geq \ldots \geq ||\mathbf{h}_{k(2M)}||^2,$$

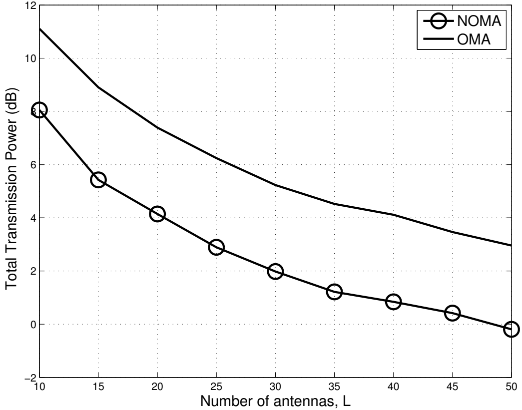
and

$$\begin{aligned} \mathcal{S}_1 &= \{k(1), \ldots, k(M)\} \text{ (strong user set)} \\ \mathcal{S}_2 &= \{k(M+1), \ldots, k(2M)\} \text{ (weak user set).} \end{aligned}$$

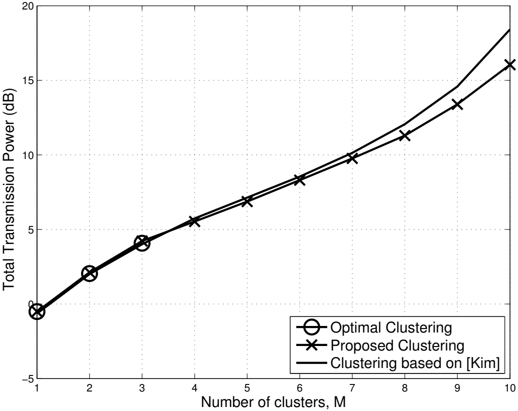
- For each user in $\mathcal{S}_1$, the best user in $\mathcal{S}_2$ can be found in terms of the transmit power as follows:

$$\bar{k}(i) = \arg \min_{m \in \mathcal{S}_2 \setminus \mathcal{I}_i} P^*(k(i), m),$$

where $\mathcal{I}_{i-1} = \mathcal{I}_i \cup \{\bar{k}(i)\}$, $\mathcal{I}_M = \emptyset$, $\bar{k}(i)$ represents the index of the weak user in $\mathcal{S}_2$ for the ith cluster.



The total transmit power for different values of the number of antennas, L, when there are $M = 3$ clusters with $(G_1, G_2) = (10 \text{ dB}, 6 \text{ dB})$.



The total transmit power for different values of the number of cluster, M, when $L = 20$, and $(G_1, G_2) = (10\ \mathrm{dB}, 6\ \mathrm{dB})$.

Remarks:

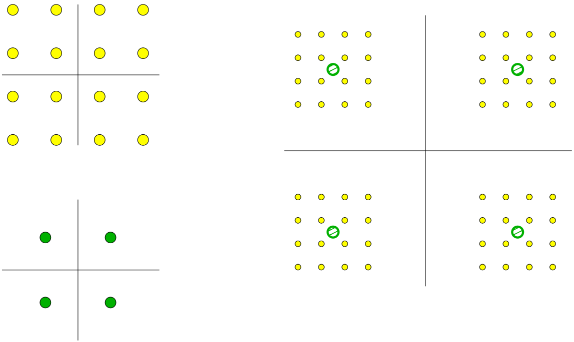
- Generalization: a generalization is required for any size of cluster.
- Optimal NOMA beamforming: while ZF inter-cluster beamforming is simple, it is not optimal. Optimal approaches are required.
- Rate maximization versus Power minimization: Since the sum rate maximization can result in a trivial solution (the strong user will take all the power to maximize the sum rate), it is often desirable to consider the total transmit power minimization with QoS constraints (e.g., SINR constraints).
- It is possible to consider sum rate maximization with constraints[3]

---

[3] See M. Hanif, et al., "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," IEEE TSP, Jan. 2016.

4. Power Allocation with a Practical NOMA Scheme

- Multiuser superposition transmission (MUST)
- Example: s_1 with 16-QAM and s_2 with 4-QAM



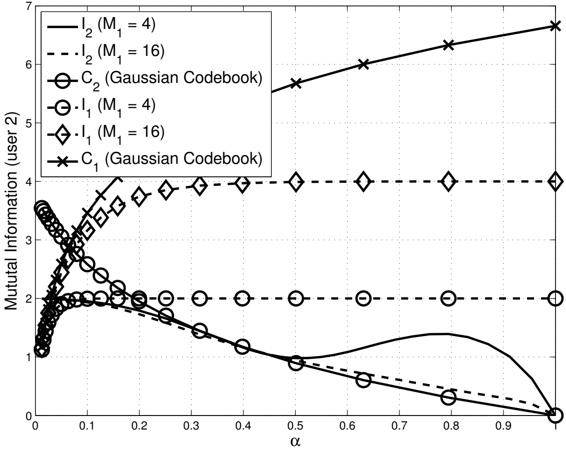
- Since channel capacity provides a bound, we may need more practical performance measure.
- We can use mutual information.
- Suppose that the received signals at users 1 and 2 are

$$\begin{aligned} y &= h_1(s_1 + s_2) + n_1 \\ z &= h_2(s_1 + s_2) + n_2. \end{aligned}$$

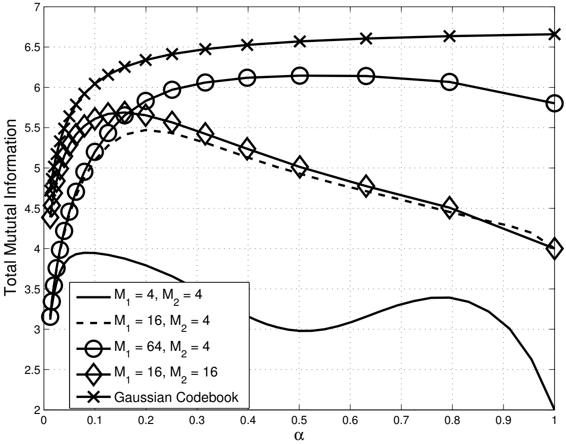
- The sum rate maximization becomes

$$\arg \max_{P_1 + P_2 \le P_T} \min\{\mathsf{I}(y; s_2), \mathsf{I}(z; s_2)\} + \mathsf{I}(y; s_1 \,|\, s_2),$$

where $\mathsf{I}(x; s)$ is the mutual information between x and s.



Mutual information versus $\alpha = \frac{P_1}{P_T}$ when $P_T = 20$ dB and $M_2 = 4$ (I_1 and I_2 stand for $\mathsf{I}(y; s_1|s_2)$ and $\mathsf{I}(z; s_2)$, respectively.)



Total mutual information versus $\alpha = \frac{P_1}{P_T}$ when $P_T = 20$ dB.

- With practical constellations, we have different power allocation results from the ideal one (i.e., Gaussian codebooks).
- The optimal power allocation depends on not only the total power, but also the constellation sizes.

5. Imperfect CSI in NOMA

- In NOMA, the power allocation plays a crucial role in ensuring successful SIC at receivers.
- The power allocation depends on CSI and perfect CSI is important.
- In practice, however, perfect CSI may not be affordable (due to the channel estimation error, feedback delay, and so on).
- To take into account imperfect CSI, the outage probability[4] has been considered in NOMA.

---

[4] See J. Cui, et al., "A novel power allocation scheme under outage constraints in NOMA systems," IEEE SPL, Sept 2016; S. Shi, et al., "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," IEEE TWC, July 2016; Z. Yang, et al., "On the performance of non-orthogonal multiple access systems with partial channel information," IEEE TCOM, Feb 2016.

- Outage probability:

  $$P_{\mathrm{out}} = \Pr(\mathrm{SINR}_k < \Gamma_k),$$

  where Γ_k is the target SINR at user k.
- For coded packets, the outage probability is the probability of decoding error:

  $$\Pr(\mathrm{SINR}_k \geq \Gamma_k) = \Pr(\log_2(1 + \mathrm{SINR}_k) < R_k)$$

  where the code rate R_k is given by

  $$R_k = \log_2(1 + \Gamma_k).$$

- The outage probabilities in NOMA:

$$\mathbb{P}_2 = \Pr\left(\log_2\left(1 + \frac{|h_2|^2 P_2}{|h_2|^2 P_1 + N_0}\right) < R_2\right)$$
$$\mathbb{P}_1 = \Pr\left(\log_2\left(1 + \frac{|h_1|^2 P_1}{N_0}\right) < R_1, \log_2\left(1 + \frac{|h_1|^2 P_2}{|h_1|^2 P_1 + N_0}\right) < R_2\right)$$

- We assume that the $|h_k|^2$'s are not known, but their pdfs are available at a transmitter.
- In some cases, it is possible to find closed-form expressions for the outage probabilities.

- Minimum Total Transmit Power Problem (of Two-User):

$$\min P_1 + P_2$$
$$\text{subject to } \mathbb{P}_k \leq \epsilon_k, \; k = 1, 2,$$

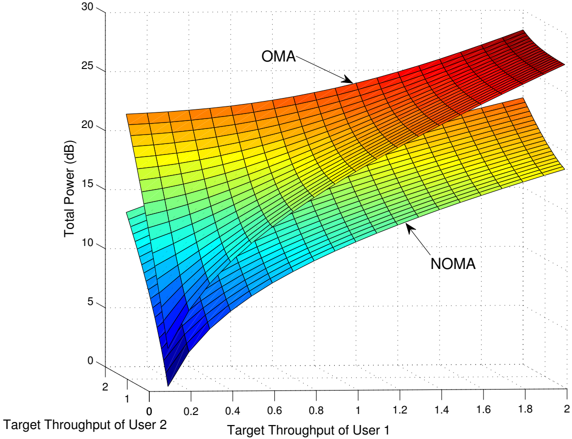
  where ϵ_k is the target maximum outage probability of user k.
- This problem was studied by Cui, et al. in 2016.
- While the total transmit power is minimized, no rate or throughput is guaranteed.
- Throughput with unknown CSI:

$$\eta_k = R_k(1 - \mathbb{P}_k)$$

- Problem: Joint Rate and Power Optimization:

$$\min_{R_k, P_k} \sum_k P_k$$
$$\text{subject to } \eta_k \geq \bar{\eta}_k, \ k = 1, \ldots, K,$$

  where $\bar{\eta}_k$ is the target minimum throughput of user k.
- This problem is a generalization of the minimum transmit power problem with outage probability constraints.
- An algorithm to solve the problem was found in J. Choi, "Joint Rate and Power Allocation for NOMA with Statistical CSI," IEEE TCOM 2017 (to appear).



Minimum total transmission powers, P_{T}, of OMA and NOMA for various values of target throughput when $\{\bar{\alpha}_1, \bar{\alpha}_2\} = \{1, 1/4\}$.

6. Uplink NOMA

- Code division multiple access (CDMA) and interleave-division multiple access (IDMA) for uplink can be seen as NOMA[5].
- However, in conventional CDMA, the power difference and user ordering are not actively exploited for SIC.
- In general, uplink NOMA can be inefficient if a large number of users are to be supported by a shared channel resource block.
- Moreover, if coordinated NOMA is used, it becomes quite involved for the case of a large number of users.
- Thus, it might be desirable to consider uplink NOMA for random access as uncoordinated NOMA.

---

[5] See P. Wang, et al., "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," IEEE Vehicular Technology Magazine, 2006.

Power domain multiple access[6]

- Suppose that there are L different power levels and each user can choose one of the power levels to transmit a packet.
- The received signal becomes

$$y = \sum_{k=1}^{K} \sqrt{P_k} s_k + n,$$

  where P_k and s_k represent the transmit power and signal from user k, respectively, and $n \sim \mathcal{N}(0, N_0)$ is the noise.
- It is assumed that

$$P_k \in \{\alpha_1, \ldots, \alpha_L\},$$

  where L is the number of power levels and α_l represents the lth power level.

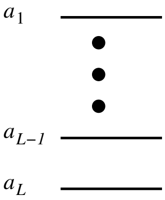
---

[6] For details, see J. Choi, "Re-Transmission Diversity Multiple Access based on SIC and HARQ-IR," IEEE Trans. Commun., November 2016.

- The SINR of layer l is defined as

$$\gamma_l = \frac{\alpha_l}{\sum_{m=l+1}^{L} \alpha_m + N_0},$$

  and the transmission rate of layer l is $R_l = \log_2(1 + \gamma_l)$.
- A user choosing the power level $P_k = \alpha_l$, the transmission rate has to be R_l.



- If $P_k = \alpha_k$, $k = 1, \ldots, K$, and $K = L$, all L users can successfully transmit signals to a receiver where SIC is used to decode all L signals.
- This approach can be employed for ALOHA.
- In power domain random access or NOMA-ALOHA, although multiple users transmit simultaneously, they can successfully transmit their packets without collision.
- Thus, the throughput can be improved by a factor of up to L.
- However, there can be collision if multiple users choose the same power level.

Throughput of NOMA-ALOHA

- Throughput of conventional ALOHA:

  $$T = Kp_a(1-p_a)^{K-1},$$

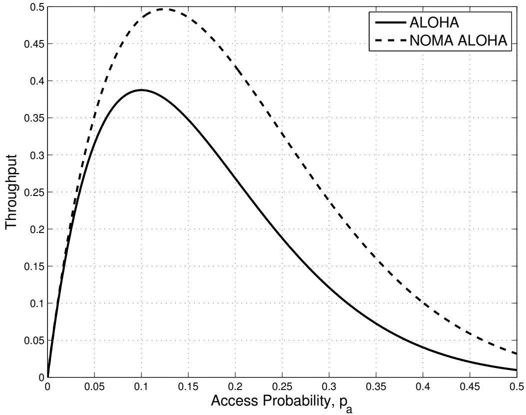
  where K is the number of users and p_a is the access probability.
- The maximum throughput is $e^{-1} \approx 0.3679$ which is achieved for a large K with $p_a = 1/K$ or $Kp_a = 1$.

- With $L = 2$, the throughput of NOMA-ALOHA is

  $$T = Kp_a(1-p_a)^{K-1} + \frac{1}{2}\binom{K}{2}p_a^2(1-p_a)^{K-2},$$

  where the second term is the probability that there are two users transmitting packets and they choose different power levels.
- For a large K, the maximum throughput can be achieved when $Kp_a = \sqrt{5} - 1 \approx 1.2361$. The resulting maximum throughput is $T \approx 0.4701$.
- The throughput increases by a factor of 1.2778 using NOMA for ALOHA.
- The more power levels, the higher throughput improvement at the cost of high transmission power.

Throughput curves of ALOHA and NOMA-ALOHA



Throughput of ALOHA and NOMA-ALOHA with 2 power levels when there are $K = 10$ *users.*

7. Conclusions

- NOMA is a promising transmission scheme that can improve the spectral efficiency.
- There are practical approaches to implement NOMA and some approaches are adopted in an LTE-A standard and proposed for 5G.
- NOMA has been applied to various existing systems and schemes such as beamforming, HARQ, two-way relay systems, coordinated multipoint (CoMP), and so on, in order to improve the spectral efficiency.
- In NOMA, since the power allocation and user clustering are important, they have been mainly studied, while there are also other important issues.

Further Study Topics

- Superposition coding (practical design)
- NOMA multicast
- Physical layer security with NOMA
- NOMA in multi-cell (CoMP)
- Massive MIMO with NOMA
- Massive MTC with NOMA
- Ultra-reliable low latency communications (URLLC) with NOMA

List of related publications

1. J. Choi, "Joint rate and power allocation for NOMA with statistical CSI," IEEE Trans. Commun., (accepted).
2. J. Choi, "NOMA based random access with multichannel ALOHA," IEEE J. Selected Areas in Communications, (accepted).
3. J. Choi, "On generalized downlink beamforming with NOMA," J. Communications and Networks (JCN), (accepted).
4. W. Cai, C. Chen, L. Bai, Y. Jin, and J. Choi, "Power Allocation Scheme and Spectral Efficiency Analysis for Downlink NOMA Systems," IET Signal Proc., (accepted).
5. J. Choi, "Effective Capacity of NOMA and a Suboptimal Power Control Policy with Delay QoS," IEEE Trans. Commun., vol. 65, No. 4, pp. 1849-1858, April 2017.
6. W. Cai, C. Chen, L. Bai, Y. Jin, and J. Choi, "Subcarrier and power allocation scheme for downlink OFDM-NOMA systems," IET Signal Proc., vol.11, Issue 1, pp. 51-58, (January) 2017.

7. J. Choi, "Power Allocation for Max-Sum Rate and Max-Min Rate Proportional Fairness in NOMA," IEEE Commun. Letters, vol.20, No.10, pp. 2055-2058, October 2016.
8. J. Choi, "On the Spectral Efficient Nonorthogonal Multiple Access Schemes," in *Proc. IEEE EuCNC 2016*, Athens, Greece, June 2016.
9. J. Choi, "On HARQ-IR for Downlink NOMA Systems," *IEEE Trans. Commun.*, vol.64, No.8, pp. 3576-3584, August 2016.
10. J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," IEEE Trans. Wireless Commun., vol.15, no.5, pp.3226-3237, May 2016.
11. J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," IEEE Commun. Letters, vol.20, No.3, pp.438-441, March 2016.
12. J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," IEEE Trans. Commun., vol.63, No.3, pp.791–800, March 2015.
13. J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," IEEE Commun. Letters, vol. 18, no. 2, pp. 313–316, February 2014.
14. J. Choi, "On multiple access using H-ARQ with SIC techniques for wireless ad hoc networks," Wireless Personal Communications, vol. 69, pp. 187–212, 2013.