A Survey of Non-Orthogonal Multiple Access for 5G

Linglong Dai, *Senior Member, IEEE*, Bichai Wang, Zhiguo Ding, *Member, IEEE*, Zhaocheng Wang, *Senior Member, IEEE*, Sheng Chen, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—In the 5th generation (5G) of wireless communication systems, hitherto unprecedented requirements are expected to be satisfied. As one of the promising techniques of addressing these challenges, non-orthogonal multiple access (NOMA) has been actively investigated in recent years. In contrast to the family of conventional orthogonal multiple access (OMA) schemes, the key distinguishing feature of NOMA is to support a higher number of users than the number of orthogonal resource slots with the aid of non-orthogonal resource allocation. This may be realized by the sophisticated inter-user interference cancellation at the cost of an increased receiver complexity. In this article, we provide a comprehensive survey of the original birth, the most recent development, and the future research directions of NOMA. Specifically, the basic principle of NOMA will be introduced at first, with the comparison between NOMA and OMA especially from the perspective of information theory. Then, the prominent NOMA schemes are discussed by dividing them into two categories, namely, power-domain and code-domain NOMA. Their design principles and key features will be discussed in detail, and a systematic comparison of these NOMA schemes will be summarized in terms of their spectral efficiency, system performance, receiver complexity, etc. Finally, we will highlight a range of challenging open problems that should be solved for NOMA, along with corresponding opportunities and future research trends to address these challenges.**

***Index Terms*—5G, non-orthogonal multiple access (NOMA), multi-user detection (MUD), spectral efficiency, massive connectivity, overloading, low latency.**

I. Introduction

THE rapid development of the mobile Internet and the Internet of things (IoT) leads to challenging requirements for the 5th generation (5G) of wireless communication systems, which is fuelled by the prediction of 1000-fold data traffic increase by the year 2020 [1]. Specifically, the key performance indicators (KPI) advocated for 5G solutions can be summarized as follows [2]: 1) The spectral efficiency is expected to increase by a factors of 5 to 15 compared to 4G; 2) To satisfy the demands of massive connectivity for IoT, the connectivity density target is ten times higher than that of 4G, i.e. at least $10^6/\mathrm{km}^2$; 3) 5G is also expected to satisfy the requirements of a low latency (radio latency $\leq$ 1 ms), low cost ($\geq$ 100 times the cost efficiency of 4G), and the support of diverse compelling services. In order to satisfy these stringent requirements, advanced solutions have to be conceived.

Over the past few decades, wireless communication systems have witnessed a "revolution" in terms of their multiple access techniques. Specifically, for 1G, 2G, 3G, and 4G wireless communication systems, frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA) have been used as the corresponding key multiple access technologies, respectively [3] [4]. From the perspective of their design principles, these multiple access schemes belong to the category of orthogonal multiple access (OMA), where the wireless resources are orthogonally allocated to multiple users in the time-, frequency-, code-domain or according in fact based on their combinations. We might collectively refer to these domains as "resources". In this way the users' information-bearing signals can be readily separated at a low complexity by employing relatively cost-efficient receivers. However, the number of supported users is limited by the number of available orthogonal resources in OMA. Another problem is that, despite the use of orthogonal time-, frequency- or code-domain resources, the channel-induced impairments almost invariably destroy their orthogonality. More specifically, considering that two signals $s_1(t)$ and $s_2(t)$, which are orthogonal either in time-, frequency- or code-domain, are transmitted over the dispersive channels $h_1(t)$ and $h_2(t)$, separately, the received signals $x_1(t) = s_1(t) \otimes h_1(t)$ and $x_2(t) = s_2(t) \otimes h_2(t)$ typically become non-orthogonal owing to the deleterious effects of dispersion. Hence, high-complexity "orthogonality-restoring measures", such as multi-user equalizers or space-time equalizers have to be invoked. Consequently, it remains a challenge for OMA to satisfy the radical spectral efficiency and massive connectivity requirements of 5G.

The innovative concept of non-orthogonal multiple access (NOMA) has been proposed in order to support more users than the number of available orthogonal time-, frequency-, or code-domain resources. The basic idea of NOMA is to

L. Dai, B. Wang, and Z. Wang are with Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing 100084, P. R. China (E-mail: daill@tsinghua.edu.cn, wbc15@mails.tsinghua.edu.cn, zcwang@tsinghua.edu.cn).

Z. Ding is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. (E-mail: z.ding@lancaster.ac.uk).

S. Chen is with Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (E-mail: sqc@ecs.soton.ac.uk), and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia.

L. Hanzo is with Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (E-mail: lh@ecs.soton.ac.uk).

This work was supported by the National Natural Science Foundation of China for Outstanding Young Scholars (Grant No. 61722109), the National Natural Science Foundation of China (Grant No. 61571270), and the Royal Academy of Engineering under the UK-China Industry Academia Partnership Programme Scheme (Grant No. UK-CIAPP\49). L. Hanzo would also like to acknowledge the financial support of the ERC Advanced Fellow Grant QuantCom.

Since this is a review article, no research data is available.

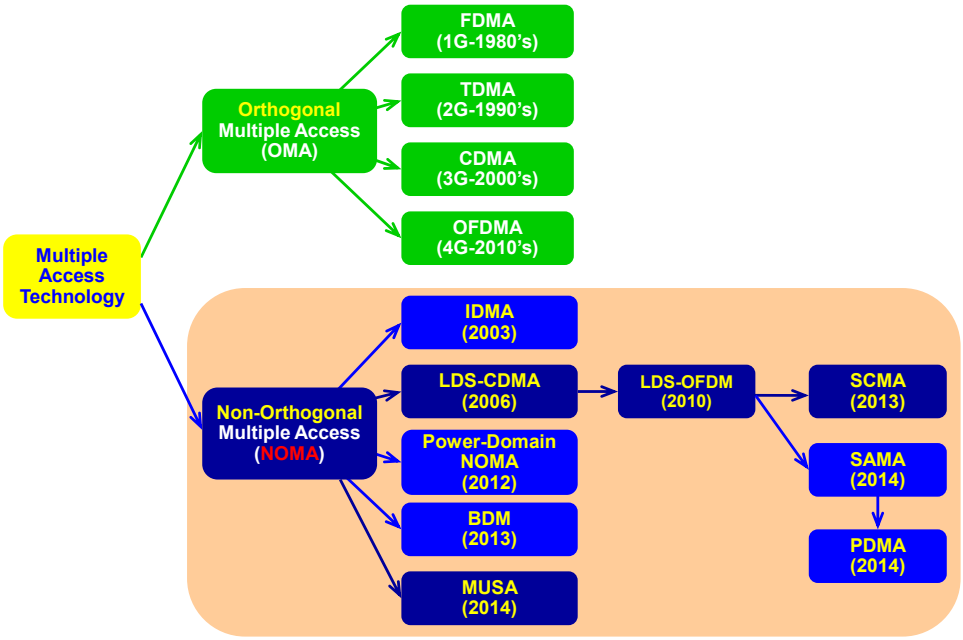

Fig. 1. The milestones of NOMA developments

support non-orthogonal resource allocation among the users at the ultimate cost of increased receiver complexity, which is required for separating the non-orthogonal signals. Recently, several NOMA solutions have been actively investigated [5]–[9], [13], which can be basically divided into two main categories, namely power-domain NOMA [14]–[17], [20]–[63] and code-domain NOMA [64]–[97], including multiple access solutions relying on low-density spreading (LDS) [64]–[73], sparse code multiple access (SCMA) [74]–[95], multi-user shared access (MUSA) [96], successive interference cancellation amenable multiple access (SAMA) [97], etc. Some other closely-related multiple access schemes, such as spatial division multiple access (SDMA) [98]–[111], pattern division multiple access (PDMA) [112] [113] and bit division multiplexing (BDM) [114] have also been proposed. The milestones of NOMA techniques are summarized in Fig. 1. Note that most the existing survey papers on NOMA [5]–[10] only focus on the subclass of power-domain NOMA schemes, even though the NOMA family is significantly broader. Some code-domain NOMA schemes are briefly introduced in [11] and [12]. By contrast, in this paper, both power-domain NOMA and code-domain NOMA, as well as the entire broad family of NOMA schemes proposed as part of the Rel-14 3GPP NR Study Item shown in Table I are introduced to provide a more comprehensive timely review. Furthermore, the comparison of 15 NOMA schemes proposed for Rel-14 3GPP NR Study Item are also included for sheding light on the unified design and implementation of NOMA. Moreover, in addition to the basic principles and theoretical analysis, prototype evaluations and field test results are also presented for quantifying the performance gain of NOMA, which are not discussed in the existing papers.

In this article, we will discuss the basic principles as well as pros and cons of NOMA in Section II. In Section III, the design principles and key features of these dominant NOMA solutions as well as the user grouping and resource allocation will be discussed, and a systematic comparison of these NOMA schemes will be provided in terms of their spectral efficiency, system performance, receiver complexity, etc. Performance evaluations and transmission experiments of NOMA are also introduced to verify the analytical results. In Section IV, we will highlight a range of challenging open problems that should be solved for supporting NOMA, such as their theoretical analysis, their sophisticated transmitter design, and the tradeoff between the attainable system performance versus receiver complexity. Accordingly, the corresponding opportunities and future research trends will be highlighted in order to provide some insights into potential future research in this promising field. Finally, our conclusions are offered in Section V. The structure of this article is shown in Fig. 2 at a glance.

II. Basic Principles and Advantages of NOMA

In this section, we will firstly compare the basic principles of OMA and NOMA. Then, the pros and cons of NOMA are contrasted to those of OMA in detail.

In conventional OMA schemes, such as FDMA, TDMA, CDMA and OFDMA used for 1G, 2G, 3G, and 4G, respectively, multiple users are allocated to orthogonal radio resources in the time-, frequency-, code-domain or to their combinations. More specifically, In FDMA for example, each user transmits a unique, user-specific signal over its unique frequency resource, hence the receiver can readily detect all users' data in their corresponding unique frequency bands, respectively. Similarly, in TDMA, an exclusive time slot is allocated to each user, hence it is easy to distinguish the

TABLE I
NOMA SCHEMES PROPOSED FOR THE REL-14 3GPP NR STUDY ITEM.

	NOMA schemes	Company	Full Name	Uplink (UL) Downlink (DL)
1	Power-domain NOMA [11]-[58]	DCM	Power-domain non-orthogonal multiple access	UL/DL
2	SCMA [69]-[90]	Huawei	Sparse code multiple access	UL/DL
3	MUSA [91]	ZTE	Multi-user shared access	UL/DL
4	PDMA [107] [108]	CATT	Pattern division multiple access	UL/DL
5	LSSA [121]	ETRI	Low code rate and signature based shared access	UL
6	RSMA [122]-[124]	Qualcomm	Resource spread multiple access	UL
7	IGMA [125]	Samsung	Interleave-grid multiple access	UL/DL
8	IDMA [126]	Nokia	Interleaver division multiple access	UL
9	NCMA [127] [128]	LGE	non-orthogonal coded multiple access	UL
10	NOCA [129]	Nokia	Non-orthogonal coded access	UL
11	GOCA [130]	MTK	Group orthogonal coded access	UL
12	LDS-SVE [137]	Fujitsu	Low density spreading-signature vector extension	UL/DL
13	FDS [138]	Intel	Frequency domain spreading	UL
14	LCRS [138]	Intel	Low code rate spreading	UL
15	RDMA [139]	MTK	Repetition division multiple access	UL

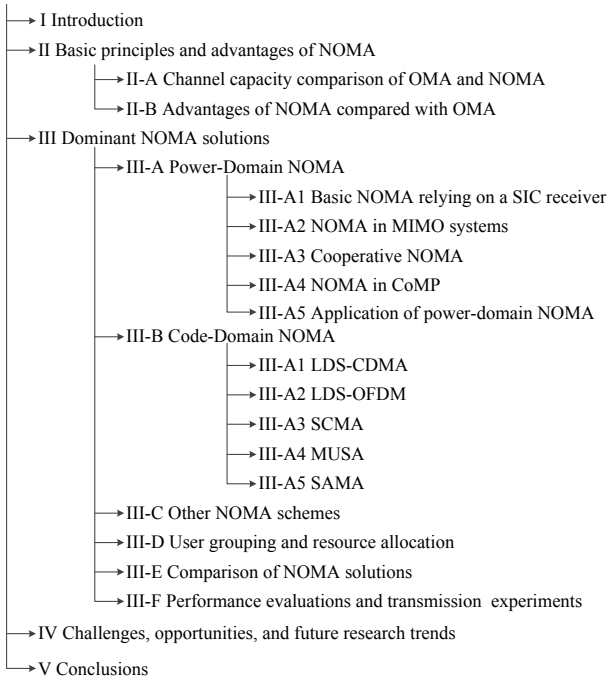

Fig. 2. The structure of this article.

different users' signals at the receivers in the time domain. In CDMA, multiple users can share the same time-frequency resources, while the transmitted symbols of different users may be mapped to orthogonal spreading sequences, such as Walsh-Hadamard codes. Hence, a low-complexity decorrelation receiver can be used for multi-user detection (MUD). OFDMA can be regarded as a smart integration of FDMA and TDMA, where the radio resources are orthogonally partitioned in the time-frequency grid. Theoretically, as a benefit of orthogonal resource allocation, there is no interference among users in OMA systems, hence low-complexity detectors having a linearly increasing complexity as a function of the number of users can be used to separate the different users' signals [115]. However, the maximum number of supportable users is rigidly restricted by the number of orthogonal resources available in conventional OMA schemes, which becomes a hard limit when massive connectivity is required for 5G. Additionally, it has been theoretically shown that OMA cannot always achieve the maximum attainable sum-rate of multi-user wireless systems, while NOMA is capable of achieving the multi-user capacity with the aid of time-sharing or rate-splitting if necessary [116], which will be detailed in the following subsection.

In order to circumvent the above limitation of OMA, NOMA has been recently investigated as a design alternative. The key distinguishing feature of NOMA is that of supporting a higher number of users than the number of orthogonal resource slots, which is achieved with the aid of sophisticated non-orthogonal resource allocation. More explicitly, carefully selected inter-user interference cancellation has to be employed at the cost of an increased receiver complexity. Furthermore, it is worth bearing in mind that even if we opt for OMA schemes, the time-domain signals are smeared by their convolution with the dispersive channel impulse response (CIR). In addition, a historic concept of NOMA is constituted by the family of CDMA systems relying on non-orthogonal spreading sequences. For example, we have many more m-sequences than the number of chips N_c in a specific m-sequence. Hence an m-sequence-based CDMA system is copable of supporting significantly more than N_c users, albeit at the cost of imposing gradually increased inter-user interference, as the number of users is increased. This increased interference can only be efficiently mitigated with the aid of powerful multi-user detectors. Hence, the NOMA concept is appealing. The family of NOMA schemes can be basically divided into two categories: power-domain NOMA and code-domain NOMA. In power-domain NOMA, different users are assigned different power levels according to their channel quality, while the same time-frequency-code resources are shared among multiple users. At the receiver side, power-domain NOMA exploits the users' power-difference, in order to distinguish different users based on successive interference cancellation (SIC). Code-domain NOMA is similar to

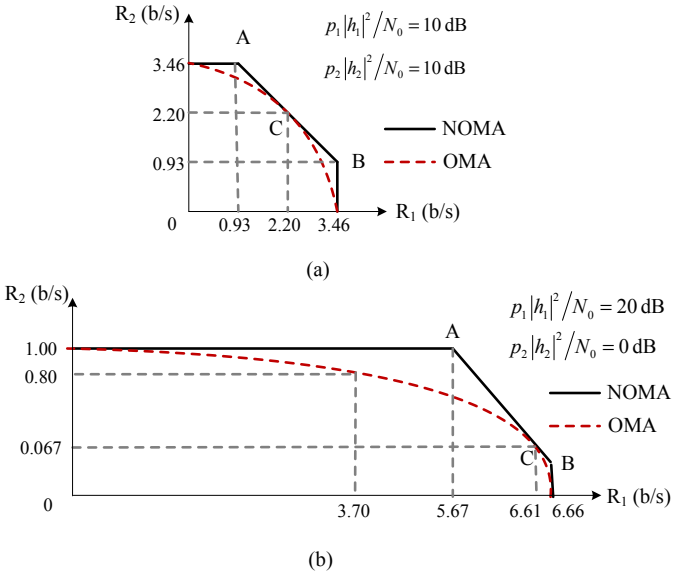

Fig. 3. Channel capacity comparison of OMA and NOMA in the uplink AWGN channel: (a) Symmetric channel; (b) Asymmetric channel [5] ©IEEE.

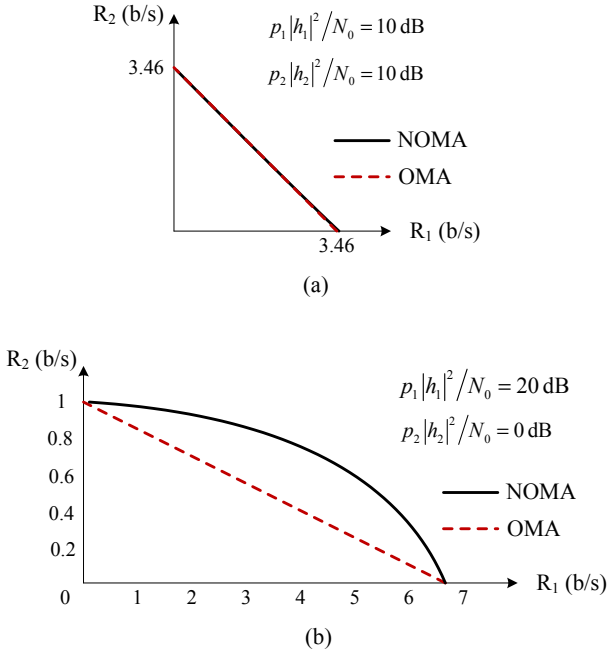

Fig. 4. Channel capacity comparison of OMA and NOMA in the downlink AWGN channel: (a) Symmetric channel; (b) Asymmetric channel [5] ©IEEE.

CDMA or multi-carrier CDMA (MC-CDMA), except for its preference for using low-density sequences or non-orthogonal sequences having a low cross-correlation.

A. Channel capacity comparison of OMA and NOMA

From the perspective of information theory, for the capacity of multiple access systems operating both in additive white Gaussian noise (AWGN) and fading scenarios, we have the following results for OMA and NOMA (applicable to both power-domain NOMA and code-domain NOMA):

- **AWGN channel**: In the uplink of an AWGN channel supporting K users (K can be larger than 2), the capacity of the multiple access channel can be formulated as [117]

$$\sum_{i=1}^{K} R_i \leq W \log\left(1 + \frac{\sum_{i=1}^{K} P_i}{N_0 W}\right), \tag{1}$$

where W is the bandwidth, P_i is the transmitted power, and N_0 is the power spectral density of Gaussian noise. More particularly, according to the capacity analysis found in the pioneering contribution of Tse and Viswanath [116], Fig. 3 and Fig. 4 from [5] portrays the channel capacity comparison of OMA and NOMA, where a pair of users communicating with a base station (BS) over an AWGN channel is considered as an example without loss of generality. Fig. 3 show that the uplink of NOMA is indeed capable of achieving the capacity region, while OMA is suboptimal in general, except at one specific point. However, at this optimal point, rate-fairness is not maintained, since the rate of the low-power user is much lower than that of the higher-power user, when the difference of the received powers of the two users is high. Note that the results for the simple two-user case can be extended to the general case of an arbitrary number of users [116]. Explicitly, it is shown in [116] that there are exactly $K!$ corner points when the K-user scenario is considered and the K-user NOMA system is capable of achieving the same optimal sum rate at all of these $K!$ corner points.
In the downlink, the boundary of the capacity region is given by the rate tuples [118]:

$$R_k = W \log\left(1 + \frac{P_k |h_k|^2}{N_0 W + \left(\sum_{j=k+1}^{K} P_j\right)|h_k|^2}\right), \tag{2}$$

which is valid for all possible splits $P = \sum_{k=1}^{K} P_k$ of total power at the BS. The optimal points can be achieved by NOMA with the aid of superposition coding at the transmitter and SIC at each of the receivers [116]. More particularly, Fig. 4 showed that the boundary of the rate pairs of NOMA is in general beyond that of OMA in asymmetric channels.
- **Fading channels**: In fading channels, the sum capacity in the uplink - provided that the channel state information (CSI) is only known at the receiver - can be represented as

$$C_{sum} = E\left\{\log\left(1 + \frac{\sum_{k=1}^{K} |h_k|^2 P_{ave}}{N_0}\right)\right\}, \tag{3}$$

where we assume that each user has the same average power P_{ave}. Hence OMA remains suboptimal in the uplink, while NOMA relying on MUD is optimal [116].
- **MIMO-NOMA and MIMO-OMA**: The versatile NOMA concept can be also extended to MIMO scenarios,

where the BS has M antennas and each user is equipped with N antennas. Additionally, multiple users can be randomly grouped into M clusters with two users in each cluster. It has been shown in [119] that MIMO-NOMA performs better than MIMO-OMA in terms of its sum channel capacity (except for transmission to a single user in MIMO systems), i.e., for any rate pair achieved by MIMO-OMA schemes, there is a specific power split for which MIMO-NOMA is capable of achieving rate pairs that are strictly higher.

B. Advantages of NOMA compared to OMA

We can see from the capacity analysis that it is feasible for NOMA to achieve a higher transmission rate than OMA. Specifically, the main advantages of NOMA compared to the classical OMA can be summarized as follows:

- **Improved spectral efficiency and cell-edge throughput**: The time-frequency resources are shared non-orthogonally among users both in the power-domain NOMA and in the code-domain of NOMA. As described above, in the uplink of AWGN channels, although both OMA and NOMA are capable of achieving the maximum attainable sum capacity, NOMA supports a more equitable user fairness. Additionally, the capacity bound of NOMA is higher than that of OMA in the downlink of AWGN channels. In multi-path fading channels subjected to inter-symbol-interference (ISI), although OMA is indeed capable of achieving the maximum attainable sum capacity in the downlink, NOMA relying on MUD is optimal, while OMA remains suboptimal, if the CSI is only known at the downlink receiver.
- **Massive connectivity**: Non-orthogonal resource allocation in NOMA indicates that the number of supportable users/devices is not strictly limited by the number of orthogonal resources available. Therefore, NOMA is capable of significantly increasing the number of simultaneous connections in rank-deficient scenarios, hence it has the potential to support massive connectivity. Of course, it should be noted that some practical implementation issues in NOMA systems, such as its hardware imperfections and computational complexity, may hinder the realization of massive connectivity, which will be detailed in Section IV.
- **Low transmission latency and signaling cost**: In conventional OMA relying on access-grant requests, a user first has to send a scheduling request to the base station (BS). Then, upon receiving this request, the BS schedules the user's uplink transmission by responding with a clear-to-send signal in the downlink channel. Thus, a high transmission latency and a high signaling overhead will be imposed, which becomes unacceptable in the case of massive 5G-style connectivity. Specifically, the access-grant procedure in LTE takes about 15.5 ms before the data is transmitted [120]. In this way, the radical requirement of maintaining a user delay below 1 ms cannot be readily satisfied [112]. By contrast, dynamic scheduling is not required in some of the uplink NOMA schemes. To elaborate a little further, in the uplink of a SCMA system, grant-free multiple access can be realized for users associated with pre-configured resources defined in the time- and frequency-domain, such as the codebooks, as well as the pilots. By contrast, at the receiver blind detection and compressive sensing (CS) techniques can be used for performing joint activity and data detection [91]. Hence again, beneficial grant-free uplink transmission can be realized in NOMA, which is capable of significantly reducing both the transmission latency and the signaling overhead. Note that in some NOMA schemes using SIC receivers, the SIC process may impose extra latency. Therefore, the number of users relying on SIC should not be excessive, and advanced MIMO techniques can be invoked for serving more users, as discussed in Section III.
- **Relaxed channel feedback**: The requirement of channel feedback will be relaxed in power-domain NOMA, because the CSI feedback is only used for power allocation. Hence there is no need for accurate instantaneous CSI knowledge. Therefore, regardless whether fixed or mobile users are supported, having a limited-accuracy outdated channel feedback associated with a certain maximum inaccuracy and delay will not severely impair the attainable system performance, as long as the channel does not change rapidly.

Given the above prominent advantages, NOMA has been actively investigated, with a views for employment in 5G as a promising solution. In the next section, we will discuss and compare the dominant NOMA solutions.

III. Dominant NOMA Solutions

In this section, we will discuss the families of prominant NOMA schemes by dividing them into two categories, namely power-domain and code-domain NOMA. Their design principles and key features will be highlighted, respectively. We will also provide their comparison in terms of their spectral efficiency, system performance, receiver complexity, etc. At the end of this section, performance evaluations and transmission experiments of NOMA will be discussed.

A. Power-Domain NOMA

In this subsection, we will discuss the first category of NOMA, namely, power-domain NOMA. In [14]–[17], the concept and key features of power-domain NOMA have been described in detail. In contrast to the multiple access schemes relying on the time-, frequency-, code-domain or on their combinations, NOMA can be realized in a recently emerged new domain, namely in the power domain. At the transmitter, different signals generated by different users are directly superimposed on each other after classic channel coding and modulation. Multiple users share the same time-frequency resources, and then are detected at the receivers by MUD algorithms such as SIC. In this way, the spectral efficiency can be enhanced at the cost of an increased receiver complexity compared to conventional OMA. Additionally, it is widely recognized based on information theory that non-orthogonal

multiplexing using superposition coding at the transmitter and SIC at the receiver not only outperforms classic orthogonal multiplexing, but it is also optimal from the perspective of achieving the capacity region of the downlink broadcast channels [116].

Some practical considerations for power-domain NOMA, such as multi-user power allocation, signalling overhead, SIC error propagation and user mobility, were discussed in [14]. To achieve a further enhancement of its spectral efficiency, the authors of [14]–[25] invoked a combination of NOMA with MIMO techniques. Particularly, the capacity comparison between MIMO-NOMA and MIMO-OMA has been investigated in [18] [19], and the superiority of MIMO-NOMA over MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity was shown analytically. Furthermore, in [20] [21], the potential gains of MIMO-NOMA were shown based on both link-level as well as on system-level simulations and using a NOMA test-bed developed. A hardware SIC receiver was used for taking into account the realistic hardware impairments quantified in terms of the error vector magnitude (EVM), the number of quantization bits in the analog/digital (A/D) converter, etc. The simulation results and the measurements obtained showed that in the variet of configurations considered, the cell throughput achieved by NOMA is about 30% higher than that of OFDMA. Furthermore, some open implementation issues were also discussed in [20] [21], including the granularity of the multi-user power allocation both in time and frequency, as well as the signaling overhead, feedback enhancements and receiver design were detailed. Additionally, the receiver design was discussed in [26]–[28]. A novel NOMA transmitter and receiver design was proposed in [26], where the signals of multiple users are jointly modulated at the transmitter side and detected at the receiver side. In this scheme, the desired signal of the cell center user can be directly detected without detecting the signal of the cell edge user, i.e., without SIC processing. Thus, a low complexity is achieved. Furthermore, the associated simulation results have shown that compared to the ideal SIC, the downlink NOMA link-level performance depends both on the actual receiver design and on the difference in the power ratio split between the cell edge user and cell center user. Besides, the design and performance of the SIC receiver for downlink NOMA combined with 2-by-2 open-loop SU-MIMO based on LTE TM3 (Transmission mode 3) were investigated in [27], where different receiver weight generation schemes were introduced both before SIC and after SIC according to the transmission rank combination between the users. The link-level simulation results showed that the codeword level SIC achieves higher performance than the symbol level SIC and in fact approaches the performance of ideal SIC. The impact of applying the SIC receiver for cell-edge users in downlink NOMA using SU-MIMO was investigated in [28]. The simulation results showed that there is an improvement of the NOMA gains over OMA in conjunction with SIC processing for the cell-edge users. Furthermore, in order to increase the attainable performance of the SIC receiver, cooperative NOMA transmission has been proposed in [29] [30]. A range of investigations related to multi-cell NOMA schemes were carried out in [31]–[33]. Moreover, since having an increased number of cell-edge users typically degrades the efficiency of coordinated multi-point (CoMP) transmissions, this limitation was circumvented by a promising NOMA solution proposed for a CoMP system in [34]. Additionally, the performance of NOMA techniques supporting randomly distributed users was evaluated in [35]. These simulation results demonstrated that the outage performance of NOMA substantially depended both on the users' target data rates and on their allocated power. In [36]–[45], the system-level performance of power-domain NOMA was evaluated and the associated simulation results showed that both the overall cell throughput and the cell-edge user throughput, as well as the degree of proportional rate-fairness of NOMA were superior to those of OMA. Furthermore, the impact of the residual interference imposed by realistic imperfect channel estimation on the achievable throughput performance was investigated in [46]–[48]. On the one hand, the channel estimation error results in residual interference in the SIC process, which hence reduces the achievable user throughput. On the other hand, the channel estimation error causes error in the transmission rate control for the respective users, which may result in decoding errors not only at the destination user terminal but also at other user terminals owing to the error propagation imposed by the SIC process. A simple transmission rate back-off algorithm was considered in [46] [47], and the impact of the channel estimation error was effectively mitigated. Simulation results showed that NOMA achieves beneficial user throughput gains over OMA in a scenario subject to channel estimation errors, which is similar to the case associated with perfect channel estimation.

Let us now elaborate on the power-domain NOMA techniques in this subsection. Firstly, the basic principle of power-domain NOMA relying on a SIC receiver will be discussed. Then, a promising extension relying on integrating NOMA with MIMOs will be discussed for the sake of increasing its attainable spectral efficiency. Another compelling extension to a cooperative NOMA transmission scheme will also be presented. Finally, the networking aspects of NOMA solutions will be discussed.

1) Basic NOMA relying on a SIC receiver: Firstly, we consider the family of single antenna systems relying on a single BS and K users.

In the downlink, the total power allocated to all K users is limited to P, and the BS transmits the signal x_i to the ith user subjected to the power-scaling coefficient p_i. In other words, the signals destined for different users are weighted by different power-scaling coefficients and then they are superimposed at the BS according to:

$$x = \sum_{i=1}^{K} \sqrt{p_i} x_i, \tag{4}$$

where $\mathrm{E}[|x_i|^2] = 1\,(i = 1, 2, \cdots, K)$ denotes the normalized power of the user signals, and we have $P = \sum_{i=1}^{K} p_i$. The received signal y_i of the ith user is

$$y_i = h_i x + v_i, \tag{5}$$

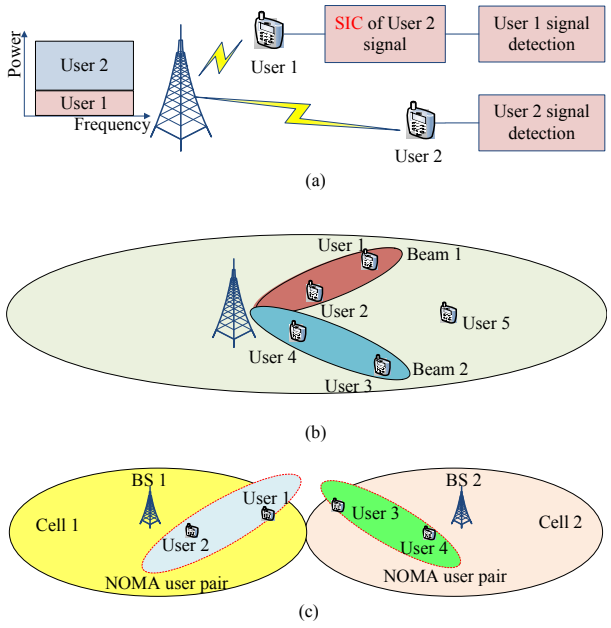

Fig. 5. Power-domain NOMA: (a) Basic NOMA relying on a SIC receiver; (b) NOMA in MIMO systems; (c) NOMA in CoMP.

where h_i denotes the channel gain between the BS and the ith user, while v_i associated with the power density N_i represents the Gaussian noise plus the inter-cell interference.

At the receiver, SIC is used for MUD. The optimal order of SIC detection relies on commencing with the detection of the highest-power user and proceeding to the weakest user (determined by $|h_i|^2/N_i, i = 1, 2, \cdots, K$). Based on this optimal SIC-detection order, any user can detect its information without substantial interference-contamination imposed by the other users whose normalized channel gain is lower than that of this user. In this way, the user having the strongest normalized channel gain can cancel the interference emanating from all the other users, and thus it remains almost uncontaminated. It is intuitive that the users associated with small normalized channel gains should be allocated higher power levels in order to improve their received signal-to-interference and noise ratio (SINR), so that a high detection reliability can be guaranteed. More particularly, it has been validated in [121] that, to maximize the sum rate, it is optimal for each user to decode the signals of users having poorer normalized channel gains first. Although the users having larger normalized channel gains require less power, they are capable of correctly detecting their data with a high probability, as a benefit of SIC. Without loss of generality, we assume that $|h_1|^2/N_1 \geq |h_2|^2/N_2 \geq \cdots \geq |h_K|^2/N_K$, and a descending-order-based power allocation $p_1 \leq p_2 \leq \cdots \leq p_K$ can be considered. Assuming perfectly error-free decoding of the interfering signals, the achievable rate of user i ($i = 1, 2, \cdots, K$) can be written as

$$R_i = W \log \left(1 + \frac{p_i |h_i|^2}{N_i W + \left(\sum_{j=i}^{i-1} p_j \right) |h_i|^2} \right). \quad (6)$$

In the case of two users as shown in Fig. 5 (a), we assume that the normalized channel gain of the second user is lower than that of the first one, i.e., $|h_1|^2/N_1 > |h_2|^2/N_2$, and thus $p_1 < p_2$. The second user detects its signal by regarding the signal of the first user as interference. The first user firstly detects the signal of the second user, and then subtracts its remodulated version from the received signal, so that the first user can detect its own signal without interference from the second user.

Assuming that the transmission bandwidth is normalized to 1Hz, the data rates of the first user and the second user can be represented as

$$R_1 = \log_2 \left(1 + \frac{p_1 |h_1|^2}{N_1} \right), \quad (7)$$

$$R_2 = \log_2 \left(1 + \frac{p_2 |h_2|^2}{p_1 |h_2|^2 + N_2} \right), \quad (8)$$

respectively. Thus, by tuning power allocation coefficients, the BS can adjust the data rate of each user. More particularly, it has been shown in Fig. 6 [14] that this NOMA scheme is capable of achieving higher rates than OFDMA. On the other hand, this NOMA scheme makes a full use of the natural difference of channel gains among the users, which implies that the near-far effect is effectively harnessed to achieve higher spectral efficiency. As a result, both the attainable sum capacity and the cell-edge user data rate can be improved [15].

In the uplink, the signal received at the BS is given by

$$y = \sum_{i=1}^{K} h_i \sqrt{p_i} x_i + v, \quad (9)$$

where p_i and x_i are the transmit power and signal transmitted by the i-th user, respectively. Furthermore, v associated with the power density N_0 represents the Gaussian noise plus the inter-cell interference at the BS. SIC is used for reliable signal detection at the BS. Without loss of generality, we assume that $p_1 |h_1|^2 \geq p_2 |h_2|^2 \geq \cdots \geq p_K |h_K|^2$, and accordingly the optimal decoding order for SIC is $x_1, x_2, \cdots, x_K$. Before the BS detects the i-th user's signal, it decodes the j-th ($j < i$) user's signal first and then removes $(i-1)$ users' signals from the observation y. The remaining $(K-i)$ signals are regarded as interference. As a result, the achievable data rate of the i-th user becomes

$$R_i = W \log \left(1 + \frac{p_i |h_i|^2}{N_0 W + \sum_{j=i+1}^{K} |h_j|^2 p_j} \right). \quad (10)$$

As illustrated in Section II-A, this NOMA scheme is capable of achieving the maximum attainable multi-user capacity in AWGN channels both in the uplink and downlink. Furthermore, this NOMA scheme has the potential of striking a more attractive tradeoff between the spectral efficiency and user-fairness.

When the number of users is sufficiently high, the SIC-induced error propagation may have a severe effect on the error probability in the absence of preventative measures. However,

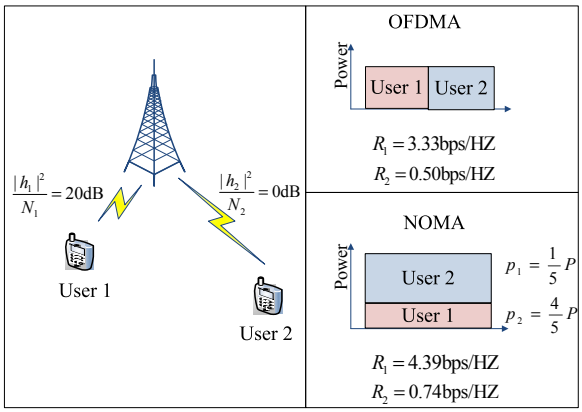

Fig. 6. Comparison of NOMA and OMA [14] ©IEEE.

some advanced user pairing and power allocation methods, as well as powerful channel coding schemes can be used for reducing the error probability. Indeed, it has been shown that error propagation only has a modest impact on the NOMA performance even under the worst-case scenario [14] [37].

2) NOMA in MIMO systems: Although the same time-frequency resources can be shared by multiple users in the basic NOMA employing SIC, the improvement of spectral efficiency still remains limited, hence may not satisfy the expected spectral efficiency improvements of 5G. An appealing solution is the extension of the basic NOMA using SIC by amalgamating it with advanced MIMO techniques [16] [17].

As illustrated in Fig. 5 (b), in downlink NOMA of MIMO systems, M_{BS} BS antennas are used for generating B different beams in the spatial domain with the aid of beamforming. Within each beam, the signals of multiple users may be transmitted by superimposing them, hence leading to the concept of intra-beam superposition modulation, which is similar to the basic NOMA using SIC, as discussed above. The bth $(1 \leq b \leq B)$ transmitter beamforming vector is denoted as $\mathbf{m}_b$. Let us assume that the number of users in the bth beam is k_b, the transmitted symbol of the ith user in the bth beam is $x_{b,i}$, and the corresponding power-scaling coefficient is $p_{b,i}$. Then, by accumulating all signals of the B different beams, the M_{BS}-dimensional transmitted downlink signal vector at the BS can be formulated as

$$\mathbf{x_0} = \sum_{b=1}^{B} \mathbf{m}_b \sum_{i=1}^{k_b} \sqrt{p_{b,i}} x_{b,i}. \tag{11}$$

Assuming that each user has N_r receiver antennas, the N_r-dimensional received signal vector of the ith user encapsulated in the bth beam can be represented as

$$\mathbf{y}_{b,i} = \mathbf{H}_{b,i}\mathbf{x_0} + \mathbf{v}_{b,i}, \tag{12}$$

where $\mathbf{H}_{b,i}$ denotes the channel matrix of size $N_r \times M_{BS}$ between the BS and the ith user in the bth beam, and $\mathbf{v}_{b,i}$ denotes the Gaussian noise plus the inter-cell interference.

At the receiver, a pair of interference cancellation approaches are used for removing the inter-beam interference and the intra-beam interference, respectively. The inter-beam interference can be suppressed by spatial filtering, which is similar to the signal detection algorithm of spatial division multiple access (SDMA) systems. Assuming that the spatial filtering vector of the ith user in the bth beam is $\mathbf{f}_{b,i}$, the signal $z_{b,i}$ after spatial filtering can be represented as

$$\begin{aligned} z_{b,i} =& \mathbf{f}_{b,i}^H \mathbf{y}_{b,i} \\ =& \mathbf{f}_{b,i}^H \mathbf{H}_{b,i} \mathbf{m}_b \sum_{j=1}^{k_b} \sqrt{p_{b,j}} x_{b,j} \\ &+ \mathbf{f}_{b,i}^H \mathbf{H}_{b,i} \sum_{\substack{b'=1 \\ b' \neq b}}^{B} \mathbf{m}_{b'} \sum_{j=1}^{k_b} \sqrt{p_{b',j}} x_{b',j} \\ &+ \mathbf{f}_{b,i}^H \mathbf{v}_{b,i}. \end{aligned} \tag{13}$$

By normalizing the aggregated power of the inter-beam interference and the receiver noise plus inter-cell interference to unity, we can rewrite (13) as

$$z_{b,i} = \sqrt{a_{b,i}} \sum_{j=1}^{k_b} \sqrt{p_{b,j}} x_{b,j} + q_{b,i}, \tag{14}$$

where $q_{b,i}$ is the normalized term representing the sum of the inter-beam interference and receiver noise plus inter-cell interference, while $a_{b,i}$ is formulated as

$$a_{b,i} = \frac{|\mathbf{f}_{b,i}\mathbf{H}_{b,i}\mathbf{m}_b|^2}{\left\{ \begin{array}{c} \sum\limits_{\substack{b'=1 \\ b' \neq b}}^{B} \sum\limits_{j=1}^{k_b} p_{b',j} |\mathbf{f}_{b,i}\mathbf{H}_{b,i}\mathbf{m}_{b'}|^2 \\ + \mathbf{f}_{b,i}^H E[\mathbf{v}_{b,i}\mathbf{v}_{b,i}^H]\mathbf{f}_{b,i} \end{array} \right\}}. \tag{15}$$

After spatial filtering, the system model (14) becomes similar to that of the basic NOMA combined with SIC, as described above. Therefore, the inter-beam interference can be suppressed, and then intra-beam SIC is invoked for removing the inter-user interference imposed by superposition coding within a beam.

Naturally, more users can also be simultaneously supported, because more than two users can share a single beamforming vector. To elaborate a little further, observe in Fig. 7 [15] that both the basic NOMA combined with SIC and the extended NOMA relying on MIMO are capable of achieving a higher sum-rate than OFDMA. Furthermore, the sum-rate of NOMA

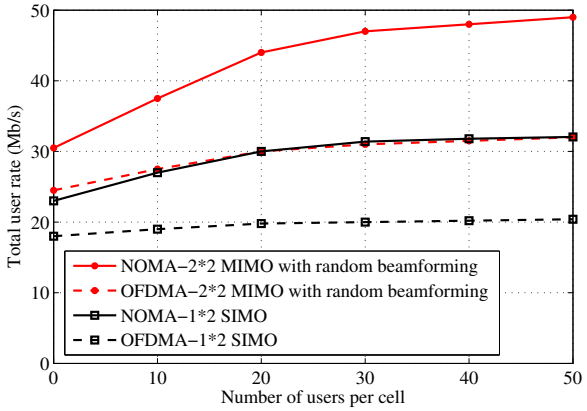

Fig. 7. System-level performance for NOMA applying opportunistic random beamforming in downlink [15] ©IEEE.

in the context of MIMO systems is higher than that of the basic NOMA using a single antenna at the BS. Additionally, in this NOMA scheme, the number of reference signals required is equal to the number of transmitter antennas, regardless of the number of non-orthogonal user-signals. In this way, when the number of users is increased beyond the number of transmitter antennas, the number of orthogonal downlink reference signals is not increased beyond the number of transmit antennas.

Observe from (14) and (15) that the design of both the beamforming vectors $\{\mathbf{m}_b\}$ and of the spatial filtering vectors $\{\mathbf{f}_{b,i}\}$ is crucial for interference cancellation, which should be carefully considered in MIMO NOMA systems. More particularly, the beamforming and spatial filtering optimization problem is usually formulated as a sum-rate maximization problem for perfect CSI scenarios. By contrast, it is typically formulated as a maximum outage probability (MMOP) minimization problem for realisitc imperfect CSI scenarios [122]. For the perfect CSI scenarios, an iterative minorization-maximization algorithm (MMA) was proposed for the efficient beamforming design of [123], where a second-order cone program (SOCP) based convex problem was solved in each iteration. Addition, the duality scheme concept has been introduced in [124], which can be regarded as the quasi-degraded solution [125] for the sum rate maximization problem, and a quadratic constrained quadratic programs (QCQP) convex problem should be solved. Furthermore, to realize low-complexity beamforming, a multiple-user CSI-based singular value decomposition (MU-CSI-SVD) algorithm was proposed in [122] for solving the sum rate maximization problem in idealized perfect CSI scenarios. This algorithm can be readily used for simplifying the outage probability expressions in realistic imperfect CSI scenarios. Therefore, in realistic imperfect CSI scenarios, the MU-CSI-SVD algorithm can also be beneficially used at a low complexity.

In the uplink systems associated with K users, where the BS is equipped with M_{BS} antennas and each user has a single transmitter antenna, the signal received at the BS can be represented as

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{h}_i \sqrt{p_i} x_i + \mathbf{v}, \tag{16}$$

where $\mathbf{h}_i$ is the M_{BS}-dimensional channel vector between user i and the BS. Furthermore, p_i and x_i are the transmit power and the signal transmitted by the i-th user, respectively, while $\mathbf{v}$ represents the Gaussian noise plus the inter-cell interference vector.

At the receiver, MMSE-SIC receiver can be used to realize signal detection [5]. It has been shown in [5] that for any decoding order, the sum throughput for all K users is equal to the maximum of the total user throughput given the received signal vector in (16). In practice, the decoding order can be adjusted according to the actual requirements, such as user fairness.

3) Cooperative NOMA: Recently, a cooperative NOMA transmission scheme was proposed in [29]. Similar to the basic NOMA, cooperative NOMA also uses a SIC receiver for detecting the multi-user signal. Therefore, the users associated with better channel conditions can be relied upon as relays in order to improve the reception reliability of the users suffering from poor channel conditions. For cooperative transmission, for example short-range communication techniques - such as Bluetooth and ultra-wideband (UWB) schemes - can be used for delivering signals from the users benefitting from better channel conditions to the users with poor channel conditions, which is the key difference with respect to the basic NOMA associated with SIC.

Without loss of generality, let us now consider a downlink cooperative NOMA system relying on a single BS for supporting K users, where the K users are ordered based on their channel qualities, with the first user having the worst channel condition, while the Kth user having the best channel condition. Cooperative NOMA relies on the following two phases [29].

The first phase, also termed the broadcast phase, represents the direct transmission. In this phase, the BS sends downlink messages to all the K users based on the principle of basic NOMA relying on SIC, where the superimposed information of the K users obeys the total power constraint. The SIC process is implemented at the user side. As a result, the users having better channel conditions have the knowledge of the signals intended to the users having poor channel conditions [126].

The second phase represents the cooperative transmission. During this phase, the cooperating users transmit their signals via their short range communication channels, such as Bluetooth or UWB. Particularly, the second phase includes $(K-1)$ time slots. In the first time slot, the Kth user broadcasts the superposition of the $(K-1)$ signals destined for the remaining users. Then the SIC process is invoked again at these $(K-1)$ users. The $(K-1)$th user combines the signals received during both phases by using maximum ratio combining (MRC), and it detects its own information at a higher SNR than that of the traditional SIC. Similarly, in the kth time slot, where $1 \leq k \leq K-1$, the $(K-k+1)$st user also broadcasts

the $(K-k)$ superposed signals for the remaining $(K-k)$ users, whose channel conditions are worse than that of this user. Then the $(K-k+1)$st user combines the observations gleaned from both phases and it detects its own information at a higher SNR than the traditional SIC. Therefore the emplyment of cooperation is indeed capable of enhancing the reception reliability. Note that it can only be invoked at low user loads, because the above-mentioned regime requires potentially excessive resources for cooperation. This might reduce the extra gain of NOMA.

To elaborate a little further, in practice the participation of all users in the cooperative NOMA cannot be readily realized due to the extra requirement of short-range communication resources as well as owing to the complex signal processing associated with a high signaling overhead. A promising solution to this problem is to reduce the number of cooperating users. Without loss of generality, we consider the appealingly simple case of having only two cooperative users as an example. Let us assume that the users are sorted in the order of improving channel qualities and that the mth and nth users are paired together, where we have $m < n$. It has been shown in [29] that the worst choice of m and n is $n = m + 1$, while the optimal choice is to group two users experiencing significantly different channel qualities.

Again, the cooperative NOMA further exploits the specific feature that the users having better channel qualities have the knowledge of other users' signals, whose channel qualities are poor. In this way, the maximum diversity gain can be achieved for all users by transmitting the signals of those specific users who have better channel qualities to other users.

4) NOMA in CoMP: NOMA also exhibits its own benefits in multi-cell applications, leading to the concept of NOMA in CoMP. However, by directly applying a single-cell NOMA design to multi-cell scenarios, NOMA in CoMP may result in severe inter-cell interference. As an example, a downlink cellular system having two cells and four users is depicted in Fig. 5 (c), where a two-user NOMA scheme is considered, with user 1 and user 2 being served by BS 1, while user 3 and user 4 being served by BS 2. However, at user 1, strong interference may be imposed by the signals transmitted from BS 2 (similarly, user 3 also suffers from severe interference caused by the signals transmitted from BS 1), which potentially leads to a significant performance degradation of the NOMA in CoMP scheme.

To mitigate the inter-cell interference in the downlink, joint transmit-precoding of all NOMA users' signals can be utilized. However, all users' data and channel information should be available at the BSs involved, and finding the optimal transmit-precoder is not trivial. Moreover, multi-user transmit precoding of single-cell NOMA may not be efficient in a NOMA in CoMP setting, since a beam generated via geographically separated BS antennas may not be capable of covering more than one angularly separated user for intra-beam NOMA. By exploiting that the CIRs of different users are likely to be rather different in the multi-cell scenario, a reduced-complexity transmit precoding scheme was proposed for NOMA in CoMP [31], where the precoder is applied only to the signals of the cell edge users, such as users 1 and 3 of Fig. 5 (c). Additionally, a multi-cell uplink NOMA system has been considered in [32], and the rate coverage probability (the probability that a given user's achievable rate remains above the target data rate) of a user who is at rank m (in terms of the distance from its serving BS) among all users in a cell and the mean rate coverage probability of all users in a cell were analyzed using the theory of order statistics and poisson cluster process. It has been shown that the average rate coverage of a NOMA cluster is better than that of its counterpart OMA cluster for higher number of users per cell and for higher target rate requirements. Furthermore, an up-to-date literature review of interference management techniques that apply NOMA in multi-cell networks has been provided in [33], including both NOMA using joint processing and NOMA relying on coordinated scheduling/beamforming. The major practical issues and challenges that arise in the implementation of multi-cell NOMA have also been highlighted in [33], such as the SIC implementation issues, imperfect CSI, as well as multi-user power allocation and clustering.

5) Application of power-domain NOMA: Recently, the concept of power-domain has been successfully applied to ATSC 3.0 [127], which is a new next-generation broadcasting standard in US, and this physical-layer non-orthogonal multiplexing technology is named layered-division-multiplexing (LDM).

Specifically, a two-layer LDM structure consisting of the upper layer (UL) and the lower layer (LL) is accepted by ATSC 3.0 to improve spectral efficiency and provide more versatile broadcasting services. The UL with higher power allocation is used to deliver mobile services to indoor, portable and handheld receivers, while the LL is designed to deliver high data rate services, such as UHDTV or multiple HDTV services to fixed reception terminals, where the operational SNR is usually high due to the large and possibly directional receive antennas [127]. At the transmitter, the data of each layer is firstly processed by its own physical-layer signal processing modules, including channel encoding, interleaving, modulation, etc., and then the signals from all layers are superimposed over the same time-frequency resources. At the receivers, to decode the UL signal, the lower-power LL service is treated as an additional interference. To decode the LL signal, the receiver firstly needs to cancel the UL signal, which is referred as SIC procedure in power-domain NOMA.

As shown in Fig. 8 from [127], the channel capacity for the mobile and fixed services of the LDM and TDM/FDM systems is compared. It is observed that LDM offers better performance than TDM/FDM in all scenarios, and the higher the SNR threshold of the fixed service, the larger the advantage of the LDM systems.

B. Code-Domain NOMA

The NOMA schemes discussed in the previous subsection realize multiplexing in the power domain. By contrast, in this subsection, we introduce the other main category of NOMA schemes, which achieves multiplexing in the code domain. The concept of code-domain NOMA is inspired by the classic CDMA systems, in which multiple users share the

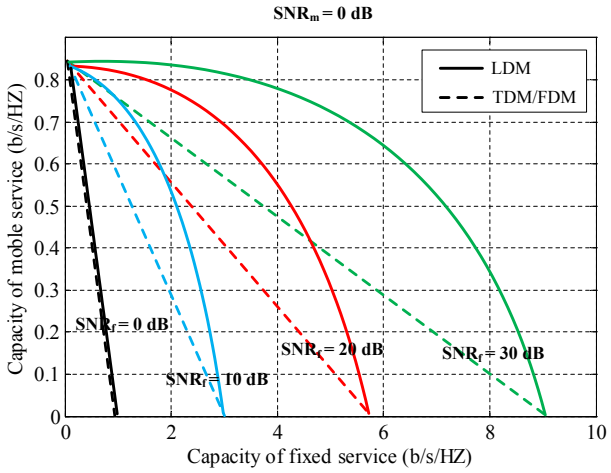

Fig. 8. Channel capacity advantage of LDM vs. TDM/FDM [127] ©IEEE.

same time-frequency resources, but adopt unique user-specific spreading sequences. However, the key difference compared to CDMA is that the spreading sequences are restricted to sparse sequences or non-orthogonal low cross-correlation sequences in NOMA. In this subsection, we first present the initial form of NOMA based on sparse spreading sequences, i.e., LDS-CDMA [64]–[67]. Then, the family of LDS aided multi-carrier OFDM systems (LDS-OFDM) [68]–[73] will be discussed, which retains all the benefits of OFDM-based multi-carrier transmissions in terms of its ISI avoidance, together with MUD-assisted LDS-CDMA operating at a lower complexity than that of the optimal maximum a posteriori probability (MAP) detector. Another important extension of LDS-CDMA is SCMA [74]–[95], which still enjoys the benefit of low-complexity reception, but has a better performance than LDS-CDMA. A suite of other improved schemes and special forms of CDMA, such as MUSA [96] and SAMA [97] will also be discussed in this subsection.

1) Low-density spreading CDMA (LDS-CDMA): Developed from the classic concept of CDMA, LDS-CDMA is designed for limiting the amount of interference imposed on each chip of conventional CDMA systems by using LDS instead of conventional spreading sequences. The basic principle of LDS-CDMA has been discussed in [64]–[66]. Additionally, [64] and [65] also discussed the iterative MUD based on the message passing algorithm (MPA) imposing a lower complexity than that of the optimal MAP detector. Specifically, in [64], the performance of LDS-CDMA communicating over memoryless Gaussian channels using BPSK modulation was analyzed. The simulation results showed that the performance of LDS-CDMA is capable of approaching the single-user performance for a normalized user-load as high as 200%. However, the performance of LDS-CDMA operating in multipath fading channels is still under investigation at the time of writing. The challenge is that the multipath fading channels will destroy the original LDS structure. On the other hand, a structured approach of designing LDS codes for LDS-CDMA has been proposed in [66], where the basic idea is to map the signature constellation elements to the spreading matrix hosting the spreading sequences. Furthermore, the capacity region of LDS-CDMA was calculated using information theoretic analysis in [67], and the accompanying simulation results showed how the attainable capacity depended on the spreading sequence density factor as well as on the maximum number of users associated with each chip, which provided insightful theoretical guidelines for practical LDS system designs.

Let us now consider a classic synchronous CDMA system operating in the uplink and supporting K users with the aid of N_c chips (N_c equals to the number of observations at the receiver). The transmitted symbol x_k of user k is firstly generated by mapping a sequence of independent information bits to a constellation alphabet, i.e., x_k is taken from a complex-valued constellation set $\mathbb{X}$. Then, the transmitted symbol x_k is mapped to a spreading sequence $\mathbf{s}_k$, such as the set of widely used PN sequences, which is unique for each user. The signal received during chip n can be represented by

$$y_n = \sum_{k=1}^{K} g_{n,k} s_{n,k} x_k + w_n, \tag{17}$$

where $s_{n,k}$ is the nth component of the spreading sequence $\mathbf{s}_k$, $g_{n,k}$ is the channel gain of user k on chip n, and w_n is a complex-valued Gaussian noise sample with a zero mean and a variance of σ^2. When we combine the signals received during all the N_c chips, the received signal vector $\mathbf{y} = [y_1, y_2, \cdots, y_{N_c}]^T$ is formulated as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{18}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$, $\mathbf{H}$ is the channel matrix of size $(N_c \times K)$, and the element $h_{n,k}$ in the nth row and the kth column of $\mathbf{H}$ is denoted by $g_{n,k}s_{n,k}$. Finally, $\mathbf{w} = [w_1, w_2, \cdots, w_{N_c}]^T$, and $\mathbf{w} \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}\right)$.

In classic CDMA systems, the elements of the spreading sequences $\mathbf{s}_k (k = 1, \cdots, K)$ are usually non-zero, i.e., the spreading sequences are not sparse. Consequently, the signals received from all the active users are overlaid on top of each other at each chip, and every user will be subjected to inter-user interference imposed by all the other users. If the spreading sequences are orthogonal, it is straightforward to eliminate the interferences, hence the information of all users can be accurately detected by a low-complexity correlation receiver. However, the classical orthogonal spreading sequences can only support as many users as the number of chips. By contrast, the above-mentioned PN-sequence family has many more codes than the number of chips in a sequence, but since the codes are non-orthogonal, they impose interference even in the absence of non-dispersive channels. Hence they require more complex MUDs. Another natural idea, which leads to LDS-CDMA, is to use sparse spreading sequences instead of the classic "fully-populated" spreading sequences to support more users, where the number of non-zero elements in the spreading sequence is much lower than N_c for the sake of reducing the interference imposed on each chip. Therefore, LDS-CDMA is potentially capable of improving the attainable system performance by using low-density spreading sequences [64], which is the key distinguishing feature between conventional CDMA and LDS-CDMA.

In LDS-CDMA, all transmitted symbols are modulated onto sparse spreading sequences. In this way, each user will only

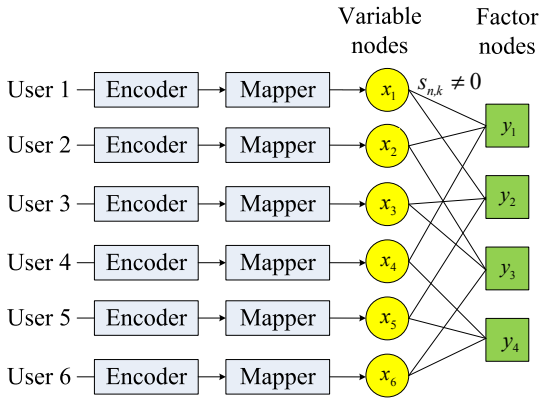

Fig. 9. Illustration of LDS-CDMA: 6 users only employ 4 chips for transmission, which implies that a normalized user-load of 150% can be achieved.

spread its data over a small number of chips, as shown in Fig. 9. As a result, the number of the superimposed signals at each chip will be less than the number of active users, which means that the interference imposed on each chip will be efficiently reduced, hence mitigating the multi-user intereference by carefully designing the spreading sequences. Therefore, the received signal at chip n in LDS-CDMA systems can be rewritten as

$$y_n = \sum_{k \in N(n)} g_{n,k} s_{n,k} x_k + w_n = \sum_{k \in N(n)} h_{n,k} x_k + w_n, \quad (19)$$

where $N(n)$ denotes the set of users whose sparse spreading sequences have a non-zero element at chip n, namely, $N(n) = \{k | s_{n,k} \neq 0\}$.

At the receiver, MUD based on message passing algorithm may be performed. Given the joint probability function $p(x_1, x_2, \ldots, x_E)$ for random variables $x_1, x_2, \ldots, x_E$, the message-passing algorithm is capable of simplifying the calculation of the marginal probability distribution for each variable as follows:

$$p(x_e) = \sum_{\sim \{x_e\}} p(x_1, x_2, \ldots, x_E), \quad (20)$$

where $\sim \{x_e\}$ represents all variables except for x_e. We assume that the joint probability function can be decomposed into the product of some positive functions, namely,

$$p(x_1, x_2, \ldots, x_E) = \frac{1}{Z} \prod_{d=1}^{D} f_d(X_d), \quad (21)$$

where Z is a normalized constant, X_d is a subset of $\{x_1, x_2, \ldots, x_E\}$, and $f_1(X_1), f_2(X_2), \ldots, f_D(X_D)$ are positive functions which are not necessarily the probability functions. Then, we can translate this form into the factor graph, which is a bipartite graph, as shown in Fig. 10 [128], where the circles represent variable nodes corresponding to $x_1, x_2, \ldots, x_E$, while the squares indicate observation nodes corresponding to $f_1(X_1), f_2(X_2), \ldots, f_D(X_D)$. An edge is present between a variable node x_e and an observation node $f_d(X_d)$ if and only if $x_e \in X_d$.

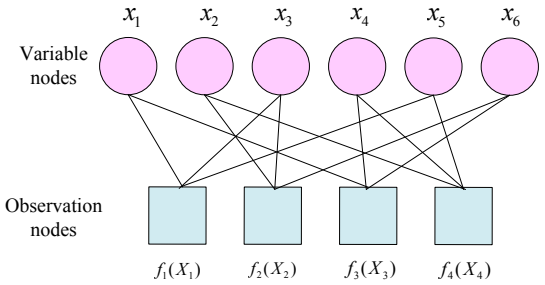

Fig. 10. Factor graph representation of MPA.

In general, the message passing algorithm relies on the factor graph representation of the problem as its input and returns the marginal distribution of all variable nodes. Messages can be passed between the variable node and the observation node through the edge between them, and the message can be interpreted as the soft-value that represents the reliability of the variable associated with each edge. The marginal distribution of a variable node can be interpreted as a function of the messages received by that variable node. The iterative form of the message passing algorithm can be represented as

$$m_{d \to e}^{(t)}(x_e) \propto \sum_{\{x_i | i \in N(d) \backslash e\}} f_d(X_d) \prod_{i \in N(d) \backslash e} m_{i \to d}^{(t-1)}(x_i), \quad (22)$$

$$m_{e \to d}^{(t)}(x_e) \propto \prod_{i \in N(e) \backslash d} m_{i \to e}^{(t-1)}(x_e), \quad (23)$$

where $m_{d \to e}^{(t)}(x_e)$ denotes the message transmitted from the observation node $f_d(X_d)$ to the variable node x_e at the tth iteration. Similarly, $m_{e \to d}^{(t)}(x_e)$ presents the message transmitted from the variable node x_e to the observation node $f_d(X_d)$. If the maximum number of iterations is T, the marginal probability distribution for each variable can be finally calculated as

$$p(x_e) \propto \prod_{d \in N(e)} m_{d \to e}^{(T)}(x_e). \quad (24)$$

It has been theoretically shown that the marginal distribution can be accurately estimated with the aid of a limited number of iterations, provided that the factor graph does not have loops [128]. However, in many practical situations the presence of loops cannot be avoided. Fortunately, the message passing algorithm is quite accurate for "locally tree like" graphs, which implies that the length of the shortest loop is restricted to $\mathcal{O}(log(E))$. Therefore, in most practical applications associated with a sparse structure, we can obtain an accurate merginal distribution estimate by an appropriate design of the factor graph.

In LDS-CDMA, the optimum MAP detection of $\mathbf{x}$ in (18) can be formulated as:

$$\hat{x}_k = \arg\max_{a \in \mathbb{X}} \sum_{\substack{\sim \{x_k\} \\ x_k = a}} p(\mathbf{x}|\mathbf{y}). \quad (25)$$

Without loss of generality, we assume that the transmitted symbols and noise are identically and independently distributed (i.i.d), and the transmitted symbols obey the uniform

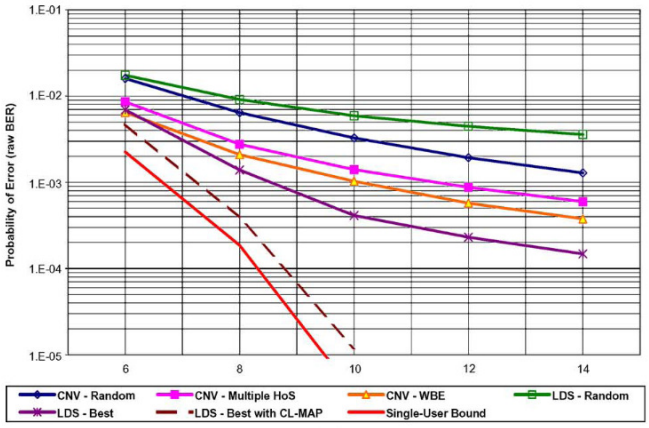

Fig. 11. Comparison of LDS-CDMA and DS-CDMA [64] ©IEEE.

distribution. Then according to Bayes' rule, (25) can be reformulated as

$$\hat{x}_k = \arg\max_{a\in\mathbb{X}} \sum_{\substack{\sim\{x_k\} \\ x_k=a}} \prod_{n=1}^{N_c} p(y_n|\mathbf{x}_{[n]}), \tag{26}$$

where

$$p\left(y_n|\mathbf{x}_{[n]}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_n - \sum_{k\in N(n)} h_{n,k}x_k)^2\right\}. \tag{27}$$

Observe that (26) represents a marginal process of $\prod_{n=1}^{N_c} p(y_n|\mathbf{x}_{[n]})$, which is similar to the form of the decomposable joint probability function of the message passing algorithm, apart from a normalization constant Z. To elaborate a little further, we can regard each term $p(y_n|\mathbf{x}_{[n]})$ as a positive function $f_d(X_d)$ in the message passing algorithm. Then, the factor graph shown in Fig. 9 can be constructed just like that of Fig. 10. Therefore, we can rewrite the iterative Equations (22) and (23) as follows:

$$m_{n\to k}^{(t)}(x_k) \propto \sum_{\{x_i|i\in N(n)\setminus k\}} \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y_n - h_{n,k}x_k - \sum_{i\in N(n)\setminus k} h_{n,i}x_i)^2\} \prod_{i\in N(n)\setminus k} m_{i\to n}^{(t-1)}(x_i), \tag{28}$$

$$m_{k\to n}^{(t)}(x_k) \propto \prod_{i\in N(k)\setminus n} m_{i\to k}^{(t-1)}(x_k). \tag{29}$$

Finally, the (approximate) marginal probability distribution of each variable after T iterations can be calculated by (24).

In the case of LDS, the number of edges in the factor graph is relatively low, hence less and longer loops can be expected based on a meritorious design of the factor graph based on a beneficial sparse spreading sequence design. Additionally, assuming that the maximum number of users superimposed at the same chip is w, the receiver complexity is on the order of $\mathcal{O}\left(|\mathbb{X}|^w\right)$ instead of $\mathcal{O}\left(|\mathbb{X}|^K\right)$ $(K > w)$ for conventional CDMA.

The performance of LDS-CDMA and direct sequence-CDMA (DS-CDMA), which is adopted by the 3G WCDMA systems, have been compared in [64]. More specifically, as shown in Fig. 11 [64], LDS-CDMA outperforms DS-CDMA using the best-found spreading sequences, where the MMSE-based partial parallel interference cancellation (PPIC) receiver has been adopted by both schemes. Furthermore, when LDS-CDMA relies on an MPA receivers, its performance approaches the single user bound within a small margin of 1.17 dB at a BER of 10^{-4}.

2) Low-density spreading aided OFDM (LDS-OFDM): OFDM and MC-CDMA are close relatives, especially when considering frequency-domain-spreading, which spreads each user's symbols across all the OFDM subcarriers, provided that the number of spreading-code chips is identical to the number of subcarriers. Then multiple users may be supported by overlaying the unique, user-specific spreading sequences of all users on top of each other across all subcarriers. As always, the spreading sequences may be chosen to be orthogonal Walsh-Hadamard codes or non-orthogonal m-sequences, as well as LDSs, for example.

Hence LDS-OFDM can be interpreted as an integrated version of LDS-CDMA and OFDM, where for example each user's symbol is spread across a carefully selected number of subcarriers and overlaid on top of each other in the frequency-domain. To elaborate a little further, in the conventional OFDMA system, only a single symbol is mapped to a subcarrier, and different symbols are transmitted on different subcarriers, which are orthogonal and hence do not interfere with each other. Therefore, the total number of transmitted symbols is restricted by the number of orthogonal subcarriers. By contrast, in the LDS-OFDM system, the transmitted symbols are firstly multiplied with LDS sequences, whose length is equal to the number of subcarriers and the resultant chips are transmitted on different subcarriers. When using LDS spreading sequences, each original symbol is only spread to a specific fraction of the subcarriers. As a result, each subcarrier carries chips related to a fraction of the original symbols. Suffice to say that apart from the already accentuated benefits, frequency-domain spreading is particularly advantageous in strongly frequency-selective channels, which would often obliterate some of the subcarriers and their information, whilst in the presence of frequency-domain spreading they would only affect some of the chips conveying the original symbols. This is likely to allow us to still recover the original symbols. We note in closing that the family of MUDs designed using the message passing algorithm for LDS-CDMA can also be used for LDS-OFDM in order to separate the overlaid symbols at the receiver.

At the time of writing a number of insightful LDS-OFDM investigations have already been disseminated in the literature. For example, the system model and properties of LDS-OFDM, including its frequency diversity order, receiver complexity, and its ability to operate under rank-deficient conditions in the presence of more users than chips have been presented in [68]. An upper limit was imposed on the number of users per subcarrier, in order to control the receiver complexity in [69]. Additionally, in [70] [71], the performance of LDS-

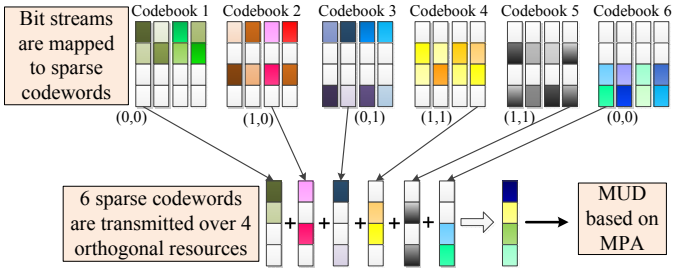

Fig. 12. SCMA encoding and multiplexing.

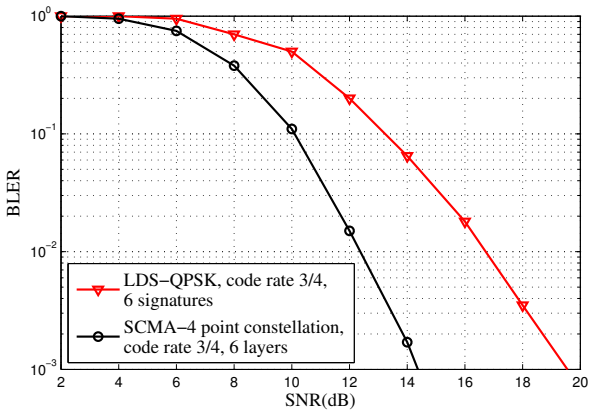

Fig. 13. Performance comparison between SCMA and LDS [74] ©IEEE.

OFDM was compared to both SC-FDMA and OFDMA in terms of the peak-to-average-power ratio (PAPR), as well as the link-level and system-level performance. The associated simulation results showed that LDS-OFDM is capable of significantly improving the attainable system performance at a given transmission power, spectral efficiency and fairness. Furthermore, in order to improve the achievable performance of LDS-OFDM, a joint subcarrier and power allocation method was proposed in [72], with the objective of maximizing the weighted sum-rate using an efficient greedy algorithm. As a further result, a pair of PAPR reduction techniques have been proposed for LDS-OFDM in [73].

3) Sparse code multiple access (SCMA): The recently proposed SCMA technique constitutes another important NOMA scheme, which relies on code domain multiplexing developed from the basic LDS-CDMA scheme. In [74], SCMA was extensively discussed in terms of its transmission and multiplexing aspects, as well as in terms of its factor graph representation and receiver architecture relying on the message passing algorithm. In contrast to the basic LDS-CDMA, as illustrated in Fig. 12, the bit-to-constellation mapping and spreading operations in SCMA are intrinsically amalgamated, hence the original bit streams are directly mapped to different sparse codewords, where each user has its own codebook. Without loss of generality, we assume that there are J codebooks, where each codebook contains M codewords of length K_l, and the number of non-zero elements in every codeword is N_{nz}. For example, in Fig. 12 we have $J = 6, M = 4, K_l = 4$, and $N_{nz} = 2$. We consider the rank-deficient scenario of $K_l < J$, which is capable of supporting the massive connectivity expected in 5G. All codewords in the same codebook contain zeros in the same $(K_l - N_{nz})$ dimensions, and the positions of zeros in the different codebooks are unique and distinct for the sake of facilitating collision avoidance for any pair of users. Therefore, the maximum number of codebooks is restricted by the selection of K_l and N_{nz}, which is equal to $\begin{pmatrix} K_l \\ N_{nz} \end{pmatrix}$. For each user, $\log_2 M$ bits are mapped to a complex codeword. The codewords of all users are then multiplexed onto K_l shared orthogonal resources, such as the OFDM subcarriers. Due to the sparsity of codewords, the signal received on subcarrier k can be represented by

$$y_k = \sum_{j \in N(k)} h_{kj} x_{kj} + w_k, \tag{30}$$

where x_{kj} is the kth component of the codeword $\mathbf{x}_j$ for user j, h_{kj} is the channel gain of user j at the kth subcarrier, and w_k denotes the complex-valued Gaussian noise with zero mean and variance σ^2. Similar to LDS-CDMA, the message passing algorithm can also be used for MUD at the SCMA receiver. However, the receiver complexity may become excessive. To circumvent this problem, improved variants of the message passing algorithm have been proposed in [75]–[83]. Specifically, a low-complexity logarithmic-domain message passing algorithm (Log-MPA) was proposed in [84]. The associated simulation results showed that the performance degradation of Log-MPA over the full-complexity MPA was negligible in practical applications, despite the fact that the Log-MPA achieved over 50% complexity reduction. For Log-MPA, the conditional channel probability calculation imposes up 60% of the total computational complexity in the whole decoding procedure. In [75], a dynamic search algorithm based on classic signal uncertainty theory was proposed for eliminating any unnecessary conditional channel probability calculation without degrading the decoding performance. On the other hand, in order to improve the BER performance of SCMA, the powerful turbo-principle has been invoked in [85] for exchanging extrinsic information between the SCMA detector and the channel decoder. By contrast, in [86], a low-complexity turbo-like combination of iterative detection and iterative decoding was conceived for striking a compelling performance versus complexity balance.

Quantitatively, the performance of SCMA relying on a multi-dimensional constellation having four points was compared to that of LDS using QPSK modulation in Fig. 13 [74]. These simulation results show that SCMA outperforms LDS in terms of its block error ratio (BLER). The key difference between LDS and SCMA is that SCMA relies on multi-dimensional constellations for generating its codebooks, which results in the so-called "constellation shaping gain" [74]. This gain is unavailable for other NOMA schemes. More explicitly, the "shaping gain" terminology represents the average symbol energy gain, when we change the shape of the modulation constellation. In general, the shaping gain is higher when the shape of the constellation becomes similar to a sphere. However, the SCMA codebook design is complex [74], since the different

layers are multiplexed with the aid of different codebooks. However, the best design criterion to be used for solving the multi-dimensional constellation problem is unknown at the time of writing. Having said this, it is anticipated that using the powerful semi-analytical tool of extrinsic information transfer-function (EXIT) charts for jointly designing the channel code and the constellation would lead to near-capacity performance. As a further solution, a multi-stage design approach has been proposed for finding a meritorious sub-optimal solution in [74]. More details concerning the codebook design can be found in [87].

Specifically, in order to simplify the optimization problem of the multi-dimensional constellation design, a mother-constellation can be generated first by minimizing the average alphabet energy for a given minimum Euclidian distance between any two constellation points. More particularly, an optimized design of the mother-constellation based on the classic star-QAM signaling constellation has been proposed in [88]. The resultant simulation results showed that the star-QAM based codebooks are capable of significantly enhancing the BER performance of the square-QAM based codebooks. Once the mother-constellation has been obtained, the codebook-specific operation can be applied to the mother-constellation in order to obtain specific constellations for each codebook. More specifically, the codebook-specific operations, such as phase rotation, complex conjugation and dimensional permutation, can be optimized for introducing correlation among the non-zero elements of the codewords, which is beneficial in terms of recovering the codewords contaminated by the interference imposed by other tones. Additionally, different power can be assigned to the symbols superimposed over the same time-frequency index for ensuring that the message passing algorithm can operate more efficiently by mitigating the interferences between the paired layers. Furthermore, inspired by the family of irregular low density parity check (LDPC) codes, an irregular SCMA structure has been proposed in [89], where the number of non-zero elements of the codewords can be different for different users. In this way, users having different QoS requirements can be simultaneously served.

Again, in the uplink of a SCMA system, grant-free multiple access can be realized by carefully assigning the codebooks and the pilots to the users based on [90]. As mentioned in Section II, a user does not have to send a scheduling request to the BS in the grant-free transmission scheme, thus a significant latency- and signaling overhead-reduction can be expected. As shown in Fig. 14, the pre-configured resource to be assigned to the users may be referred to as a contention transmission units (CTU). There are J codebooks defined over a time-frequency resource, and L pilot sequences are associated with each codebook. The grant-free multiple access regime allows contentions to occur, when multiple users are assigned to the same CTU. The network detects the uplink packets by attempting their reception using all possible access codes assigned to the predefined contention region. Then a random back-off procedure can be invoked, when collisions occur. At the receiver, blind detection and compressive sensing (CS) techniques can be used for performing joint activity and data detection, e,g., with the aid of the joint message passing algorithm (JMPA) of [91].

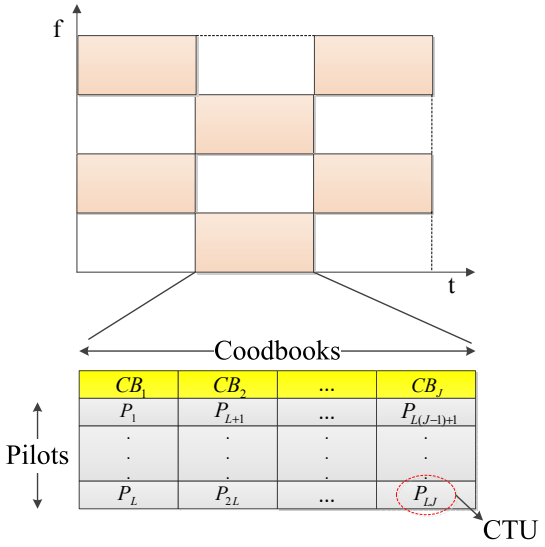

Fig. 14. Definition of a contention transmission unit (CTU) [90] ©IEEE.

SCMA can also be used in the downlink in order to improve the system throughput, leading to the multi-user SCMA (MU-SCMA) concept [92]. Based on a limited knowledge of the channel conditions of different users, the BS simply pairs the users, where the transmit power is appropriately shared among multiple users. This regime is hence reminiscent of the NOMA scheme relying on the previously mentioned power-domain multiplexing. Compared to MU-MIMO, MU-SCMA is more robust to channel-quality variations, and indeed, the provision of near-instantaneous CSI feedback is unnecessary for this open-loop multiple access scheme [92]. In [93], the concept of single-cell downlink MU-SCMA is extended to an open-loop downlink coordinated multipoint (CoMP) solution, which was termed as MU-SCMA-CoMP. In this scheme, the SCMA layers and transmits power are shared among multiple users within a CoMP cluster. The analysis and simulation results in [93] demonstrated its robustness both to user mobility and channel aging. Furthermore, the capacity of downlink Massive MIMO MU-SCMA was analyzed in [94] based on random matrix theory and it has been shown that compared to Massive MIMO MU-OFDMA systems, Massive MIMO MU-SCMA is capable of achieving a higher sum rate.

In a nutshell, the efficiency of SCMA has been verified both by simulations and real-time prototyping in [95]. Both the lab tests and the field tests demonstrated that SCMA is capable of supporting upto three times more users than the number of resource-slots, whilst still maintaining a link-integrity close to that of orthogonal transmissions.

4) Multi-user shared access (MUSA): MUSA is another NOMA scheme relying on code-domain multiplexing, which can be regarded as an improved CDMA-style scheme.

In the uplink of the MUSA system of Fig. 15, all transmitted symbols of a specific user are multiplied with the same spreading sequence (Note that different spreading sequences can also be used for different symbols of the same user, which results in beneficial interference averaging). Then, all symbols after spreading are transmitted over the same time-frequency

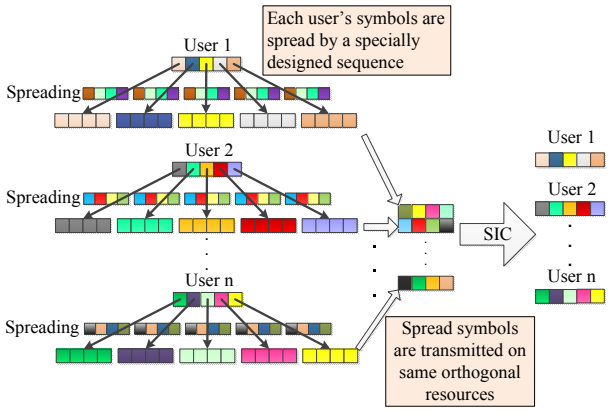

Fig. 15. Uplink MUSA system.

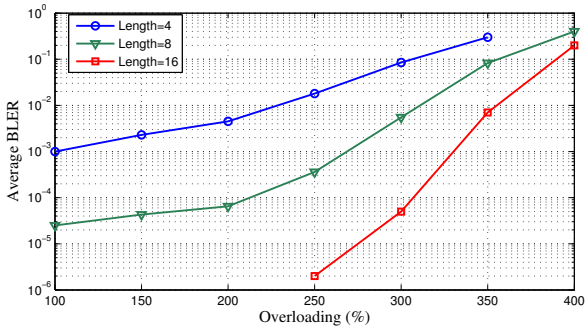

Fig. 16. The performance of MUSA with different-length spreading sequences at different normalized user-loads [96] ©IEEE.

resources, such as OFDM subcarriers. Without any loss of generality, we assume that each user transmits a single symbol every time, and that there are K users as well as N subcarriers. Rank-deficient scearios can also be supported by MUSA, i.e., $K > N$, which will impose interference amongst the users. At the receiver, linear processing and SIC are performed in order to separate the different users' data according to their channel conditions.

In the downlink of the MUSA system the users are separated into G groups. In each group, the different users' symbols are weighted by different power-scaling coefficients and then they are superimposed. Orthogonal sequences of length G can be used as spreading sequences in order to spread the superimposed symbols from G groups. More specifically, the users from the same group employ the same spreading sequence, while the spreading sequences are orthogonal across the different groups. In this way, the inter-group interferences can be removed at the receiver. Then, SIC can be used for carrying out intra-group interference cancellation by exploiting the associated power difference.

In MUSA, the spreading sequences should have low cross-correlation in order to facilitate near-perfect interference cancellation at the receiver. The MUSA technique is capable of improving the downlink capacity, which is an explicit benefit of the associated SINR difference and SIC. As a further compelling benefit, MUSA is capable of guaranteeing fairness amongst the multiplexed users without any capacity loss. In a nutshell, with the advent of advanced spreading sequences and powerful state-of-the-art SIC techniques, substantial gains can be obtained by MUSA, even for a normalized user-load as high as say 300%, which is shown in Fig. 16 [96].

5) Successive interference cancellation aided multiple access (SAMA): Let us consider an uplink SAMA system supporting K users with the aid of N orthogonal OFDM subcarriers, where we have $K > N$, i.e. when the system is rank-deficient. The system model of SAMA is similar to that of MUSA, but in SAMA, the non-zero elements of any spreading sequence $\mathbf{b}_k$ for user k are equal to one, and the spreading matrix $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_K)$ is designed based on the following principles [97]:

- The number of groups with different number of 1's in the spreading sequence should be maximized.
- The number of the overlapped spreading sequences which have the same number of 1's should be minimized.

Then the maximum number of user supported with the aid of N orthogonal subcarriers can be calculated as

$$\left(\begin{array}{c} N \\ 1 \end{array}\right) + \left(\begin{array}{c} N \\ 2 \end{array}\right) + \cdots + \left(\begin{array}{c} N \\ N \end{array}\right) = 2^N - 1. \tag{31}$$

For example, spreading matrices for $N = 2, K = 3, N = 3, K = 7$, and $N = 4, K = 15$ can be designed as follows:

$$\mathbf{B}_{2,3} = \left(\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right), \tag{32}$$

$$\mathbf{B}_{3,7} = \left(\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right), \tag{33}$$

$$\mathbf{B}_{4,15} = \left(\begin{array}{cccccccccccccc} 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right). \tag{34}$$

At the receiver, the message passing algorithm is invoked for separating the signals of the different users. The design objective used for determining the spreading matrices in SAMA is to facilitate convenient interference cancellation [97]. Consider $\mathbf{B}_{4,15}$ for example. The spreading sequence of the first user has 4 non-zero elements, hence the resultant diversity order is 4. Thus the first user's symbol is the most reliable one. Therefore, the first user's symbol can be readily determined in a few iterations, which is beneficial for the convergence of the symbol detection process of all the other users having lower diversity orders.

C. Other NOMA schemes

Apart from the prominent power-domain NOMA and code-domain NOMA solutions discussed in the previous two subsections, recently a range of alternative NOMA schemes have also been investigated, which will be discussed in this subsection.

1) Spatial division multiple access (SDMA): SDMA is one of the powerful NOMA schemes, and the philosophy of SDMA may be deemed to be related to that of classic CDMA, based on the following philosophy. Even if orthogonal Walsh-Hadamard spreading sequences are employed for distinguishing the users in CDMA systems, when they are transmitted over dispersive channels, their orthogonality is destroyed by their convolution with the CIR even in the absence of co-channel interference. Hence we end up with a potentially infinite variety of received sequences. This leads to the appealing concept of simply using the unique, user-specific CIRs for distinguishing the users, instead of unique, user-specific spreading sequences. Naturally, when the users transmitting in the uplink are close to each other, their CIRs become quite similar, which aggravates the task of the MUD in separating their signals. The beneficial properties of this family of solutions have attracted substantial research efforts, as detailed in [98]–[111].

To elaborate a little further, given the potentially infinite variety of CIRs, these sophisticated SDMA systems are capable of operating under highly rank-deficient conditions, namely when the number of mobile users transmitting in the uplink is much higher than the number of BS uplink-receiver antennas. This would avoid the hard-limited user-load of the Walsh-Hadamard code based CDMA systems, since the system-performance would only gracefully decay upon increasing the number of users. The resultant SDMA systems tend to exhibit a similar performance to their rank-deficient CDMA counterparts relying on m-sequences for example.

Since these SDMA systems rely on the CIR for distinguishing the users, they require accurate CIR estimation, which becomes extremely challenging, when the number of users is much higher than the number of BS receiver antennas. This logically leads to the concept of joint iterative channel and data estimation, which attracted substantial research interests [102] [108]. These high-end solutions often rely on powerful non-linear bio-inspired MUDs exchanging their soft-information with the channel estimator.

Below we will elaborate further on a variety of powerful solutions in a little more detail. In [98] the family of minimum bit error rate (MBER) MUDs was shown to be capable of outperforming the classic minimum mean-squared-error (MMSE) MUD in term of the achievable BER owing to directly minimizing the BER cost function. In this paper, genetic algorithms (GAs) were invoked for finding the optimum weight vectors of the MBER MUD in the context of multiple antenna aided multi-user OFDM. It was shown that the MBER MUD is capable of supporting more users than the number of receiver antennas available in highly rank-deficient scenarios.

A novel parallel interference cancelation (PIC) based turbo space time equalizer (STE) structure was designed in [99] for multiple antenna assisted uplink receivers. The proposed receiver structure allowed the employment of non-linear type of detectors such as the Bayesian decision feedback (DF) assisted turbo STE or the MAP STE, while operating at a moderate computational complexity. The powerful receivers based on the proposed structure tend to outperform the linear turbo detector benchmarker based on the classic MMSE criterion, even if the latter aims for jointly detecting all transmitters' signals. Additionally, the PIC based receiver is also capable of equalizing non-linear binary pre-coded channels. The performance difference between the presented algorithms was discussed using the powerful semi-analytical tool of extrinsic information transfer-function (EXIT) charts.

Wang *et al.* demonstrated [100] that the iterative exchange of extrinsic information between the K-best sphere detector (SD) and the channel decoder is appealing, since it is capable of achieving a near MAP performance at a moderate complexity. However, the computational complexity imposed by the K-best SD significantly increases when using a large value of K for the sake of maintaining a near-MAP performance in a high-throughput uplink SDMA/OFDM system supporting a large number of users and/or a high number of bits/symbols. This problem is further aggravated when the number of users/MSs exceeds that of the receive antennas at the BS, namely, in the challenging scenario of rank-deficient systems. It was demonstrated that the iterative decoding convergence of this two-stage system may be improved by incorporating a unity rate code (URC) having an infinite impulse response, which improves the efficiency of the extrinsic information exchange. Although this results in a slightly more complex three-stage system architecture, it allows us to use a low-complexity SD having a significantly reduced detection candidate list size. Alternatively, a reduced SNR is required. For example, given a target BER of 10^{-5} and a candidate list size of 32 for the SD, the three-stage receiver is capable of achieving a performance gain of 2.5 dB over its two-stage counterpart in a rank-deficient SDMA/OFDM 4-QAM system supporting eight co-channel users and employing for receive antennas at the BS, namely, in an (8×4) rank-deficient system having a normalized user-load of two. For the sake of further enhancing the three-stage concatenated receiver, the proposed iterative center-shifting SD scheme and the so-called irregular convolutional codes (IrCCs) were intrinsically amalgamated, which led to an additional performance gain of 2 dB.

In [101] Chen *et al.* proposed a space-time decision feedback equalization (ST-DFE) assisted MUD scheme for multiple receiver antenna aided SDMA systems. Again, a sophisticated MBER MUD design was invoked, which was shown to be capable of improving the achievable BER performance and enhancing the attainable system capacity over that of the standard MMSE design. An appealing adaptive implementation of the MBER ST-DFE assisted MUD was also proposed using a stochastic gradient-based least bit error rate algorithm, which was demonstrated to consistently outperform the classical least mean square (LMS) algorithm, while imposing a lower computational complexity than the LMS algorithm for the binary signalling scheme considered. It was demonstrated that the MBER ST-DFE assisted MUD is more robust to channel estimation errors as well as to potential error propagation imposed by decision feedback errors, than the MMSE ST-DFE assisted MUD.

The development of evolutionary algorithms (EAs) [102], such as GAs, repeated weighted boosting search (RWBS), particle swarm optimization (PSO), and differential evolution algorithms (DEAs) stimulated wide interests in the com-

munication research community. However, the quantitative performance-versus-complexity comparison of GA, RWBS, PSO, and DEA techniques applied to the joint channel estimation and turbo MUD/decoding in the context of SDMA/OFDM systems is a challenging problem, which has to consider both the channel estimation problem formulated over a continuous search space and the MUD optimization problem defined over a discrete search space. Hence the capability of the GA, RWBS, PSO, and DEA to achieve optimal solutions at an affordable complexity was investigated in this challenging application by Zhang *et al.* [102]. Their study demonstrated that the EA-assisted joint channel estimation and turbo MUD/decoder are capable of approaching both the Cramer-Rao lower bound of the optimal channel estimation and the BER performance of the idealized optimal Maximum likelihood (ML) turbo MUD/decoder associated with perfect channel estimation, respectively, despite imposing only a fraction of the idealized turbo ML-MUD/decoder's complexity.

From the discussions above, we can see that the concept of NOMA has already existed in various systems, such as SDMA, where users are distinguished using the unique, user-specific CIRs. Actually, these systems require accurate CIR estimation to successfully realize MUD, which becomes extremely challenging when the number of users is much higher than that of receiver antennas. Solving this CIR estimation problem logically leads to the concept of joint channel and data estimation, and these high-end solutions often rely on powerful non-linear MUDs. In fact, most of the studies focus on MUD design, and a series of non-linear MUD algorithms such as parallel interference cancellation (PIC) [99], and space-time decision feedback equalization (ST-DFE) [101] have been proposed. Further, with the development of evolutionary algorithms, algorithms like genetic algorithm (GA), and particle swarm optimization (PSO) may be explored to acquire accurate CIR estimation [102]. In contrast, power-domain NOMA transmits the superposition of multi-user signals with different power-allocation coefficients, and usually SIC is used at the receiver to detect multi-user signals. As the channel gain difference among users is translated into different multiplexing gains [5], both user grouping and resource allocation have substantial effects on the achievable throughput. As a result, most of the studies concerning power-domain NOMA focus on user grouping, resource (power) allocation, and performance analysis. Recently, with the development of mmWave communication and massive MIMO, combining power-domain NOMA with mmWave and massive MIMO has become a promising technique [129]–[131].

2) Pattern division multiple access (PDMA): Apart from the SDMA scheme mentioned above, the family of PDMA schemes [112] [113] constitutes another promising NOMA class that can be implemented in multiple domains. At the transmitter, PDMA employs non-orthogonal patterns, which are designed by maximizing the diversity and minimizing the correlations among the users. Then, multiplexing can be realized in the code-, power- or spatial-domains, or in fact in their combinations. Multiplexing in the code domain is reminiscent of SAMA [97]. Multiplexing in the power domain has a system model similar to multiplexing in the code domain,

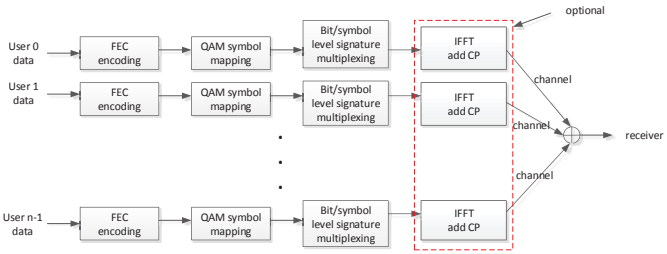

Fig. 17. The LSSA transmitter structure [132] ©IEEE.

but power-scaling has to be considered under the constraint of a given total power. Multiplexing in the spatial domain leads to the concept of spatial PDMA, which relies on multi-antenna aided techniques. In contrast to MU-MIMO, spatial PDMA does not require joint precoding for realizing spatial orthogonality, which significantly reduces the system's design complexity. Additionally, multiple domains can be combined in PDMA to make full use of the various wireless resources available. The simulation results of [113] demonstrated that compared to LTE, PDMA is potentially capable of achieving a 200% normalized throughput gain in the uplink, and more than 50% throughput gain may be attained in the downlink.

3) Signature-based NOMA: Signature-based NOMA schemes are also proposed as promising candidates for 5G. Low code rate and signature based shared access (LSSA) is one of them, and the transmitter structure conceived for uplink massive machine-type communication (mMTC) is depicted in Fig. 17 [132]. LSSA [132] multiplexes each user's data either at bit or symbol level with the aid of a specific signature pattern, which consists of a reference signal (RS), complex/binary sequence, and permutation pattern of a short length vector. All the users' signatures share the same short vector length, which can be chosen randomly by the mobile terminal or assigned to the user by the network. Moreover, LSSA can be optionally modified to have a multi-carrier variant in order to exploit the frequency diversity provided by a wider bandwidth, and to reduce the latency. It can also support asynchronous uplink transmissions, because the BS is capable of distinguishing/detecting the overlaid user signals by correlating them with the signature patterns, even if the transmission timing is different from each other.

Similar to LSSA, resource spread multiple access (RSMA) [133] [134] also assigns unique signatures to separate the different users and spreads their signals over all the available time and frequency resources. The unique signatures may be constituted by the power, spreading/scrambling codes with good correlation properties, interleavers, or their combinations, and interference-cancellation type receivers can be utilized. Depending on the specific application scenarios, RSMA may include [135]:

- **single carrier RSMA**: It is optimized for battery power consumption and link budget extension by utilizing single carrier waveforms and very low peak to average power ratio (PAPR) modulations. It allows grant-free transmission and potentially allows asynchronous access.
- **Multi-carrier RSMA**: It is optimized for low-latency access and allows for grant-free transmissions.

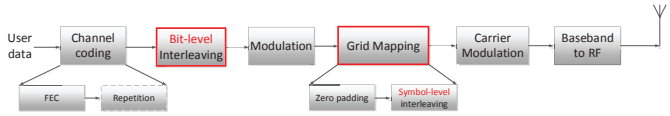

Fig. 18. The schematic of IGMA transmitter [136] ©IEEE.

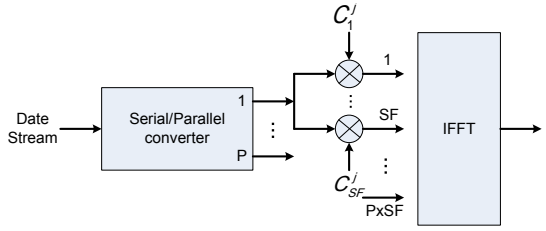

Fig. 19. NOCA transmitter structure [140] ©IEEE.

4) Interleaver-based NOMA: Interleave-grid multiple access (IGMA) is an interleaver-based multiple access scheme, which can distinguish the different users based on their different bit-level interleavers, different grid mapping patterns, or the combinations of these two techniques [136]. The typical transmitter structure of IGMA is shown in Fig. 18. Specifically, the channel coding process can be either simple repetition coding (spreading) of moderate coding rate for classic forward error correction (FEC) or low-rate FEC. By contrast, the grid mapping process of Fig. 18 may vary from sparse mapping based on zero padding to symbol-level interleaving, which could provide another dimension for user multiplexing. Whilst we need well-designed FEC codes and spreading code sequences, the design of bit-level interleavers and/or grid mapping patterns in Fig. 18 is somewhat more related. They provide scalability to support different connection densities, whilst striking a trade off between the channel coding gain and the benefit gleaned from sparse resource mapping. Moreover, the symbol-level interleaving of Fig. 18 randomizes the symbol sequence order, which may bring about further benefits in terms of combating the deleterious effects of frequency selective fading and inter-cell interference. Besides, a relatively low-complexity multi-user detector can be applied and the employment of a sparse grid mapping pattern could further reduce the detection complexity imposed.

Another interleaver-based multiple access scheme, namely interleave-division multiple access (IDMA), has also been proposed. Explicitly, IDMA interleaves the chips after the symbols have been multiplied by the spreading sequences. Hence, IDMA is effectively chip-interleaved CDMA. As shown in [137], compared to CDMA, IDMA is capable of achieving about 1 dB E_b/N_0 gain at a BER of 10^{-3} in highly loaded systems having a normalized user-load of 200%. The gain is mostly attributable to the fact that chip-interleaving results in an increased diversity-gain compared to conventional bit-interleaving.

5) Spreading-based NOMA: There are also many other NOMA schemes based on spreading codes, which are consistent with the concept of the aforementioned code-domain NOMA, and non-orthogonal coded multiple access (NCMA) is one of them [138]. NCMA is based on resource spreading by using non-orthogonal spreading codes having a low correlation. These codes can be obtained by finding the solutions of a Grassmannian line packing problem [139]. By imposing additional layers using superposition coding, it can provide an increased throughput and improve connectivity at a low block error ratio (BLER). Furthermore, since the receiver of the NCMA system adopts parallel interference cancellation (PIC), it has a scaleable performance vs complexity. Consequently, NCMA is eminently suitable for a larger number of connections exchanging small packets in massive machine-type communication (mMTC), or for reducing the collision probability in contention based multiple access.

Non-orthogonal coded access (NOCA) is also a spreading based multiple access scheme [140]. Similar to other spreading based schemes, the basic idea of NOCA is that the data symbols are spread using non-orthogonal sequences before transmission, which can be applied both in the frequency domain and/or in time domain based configurations. The basic transmitter structure is shown in Fig. 19, where SF denotes the spreading factors and C_j is the spreading sequence of the jth user. Specifically, the original modulated data sequence is first converted into P parallel sequences, then each of the P sequences is spread across SF number of subcarriers. In order to meet the requirements of different scenarios, the spreading factors can be adaptively varied.

6) Bit division multiplexing (BDM): The basic concept of BDM [114] relies on hierarchical modulation, but the resources of multiplexed users are partitioned at the bit level instead of the symbol level. Strictly speaking, the resource allocation of BDM is orthogonal in the bit domain, but multi-user signals can share the same constellation, which implies that they are superimposed in the symbol domain.

7) Compressive sensing (CS)-based NOMA: CS may be readily combined with NOMA schemes for exploiting either the user activity sparsity or data sparsity [143] [144]. A range of CS-based random access schemes have been conceived recently, such as the family of asynchronous random access protocols [145] and compressive random access arrangement of [146]. Additionally, in [147], random multiple access relying on CS was invoked for maximizing the system's total throughput. Furthermore, the attainable throughput associated with different amount of channel knowledge was discussed, which provided useful insights into the quantitative benefits of CS in the context of throughput maximization in random multiple access schemes. Furthermore, the joint detection of both node activity and of the received data was proposed in [148] for machine-type communication, which exploited the sporadic nature of the expected communication. All in all, CS is expected to play an important role in NOMA schemes.

8) Miscellaneous NOMA Schemes: Apart from the NOMA schemes introduced above, there are four other schemes proposed by different companies as part of the Rel-14 3GPP NR Study Item shown in Table I. Among these four NOMA schemes, three are spreading-based NOMA arrangements, namely the low density spreading-signature vector extension (LDS-SVE) proposed by Fujitsu [149], as well as the frequency domain spreading (FDS) [150] and the low code rate spreading (LCRS) [150] schemes proposed by Intel. They spread the user symbols to multiple RBs in order to attain an increased diversity gain. Repetition division multiple access

(RDMA) proposed by MTK [139] belongs to the family of interleaver-based NOMA schemes, which can readily separate the different users' signals, whilst exploiting both time- and frequency-diversity with the aid of the cyclic-shift repetition of the modulated symbols.

D. User grouping and resource allocation

In power-domain NOMA, the channel gain difference among users is translated into different multiplexing gains [5]. Therefore, both the user grouping and resource allocation techniques have a substantial impact on the achievable throughput [152]. In order to optimize the user grouping and resource allocation, the most appropriate optimization criterion has to be found first. Then, a compelling performance versus complexity trade-off has to be struck. Specifically, the classic proportional fair (PF) scheduler is known to strike an attractive tradeoff between the capacity and fairness attained. Hence it has been used both in the orthogonal 3G and 4G multiple access systems. Therefore, it has also been widely considered as a beneficial optimization criterion [5], [52]–[55] in NOMA systems. Specifically, both multi-user scheduling and power allocation per frequency block can be realized by maximizing the product of the average user throughput of all the users within a cell [5] [52]. In the uplink, the transmission power is usually independently determined for all users, thus scheduling a user set can be configured at a given power level. In the downlink, under the constraint of a fixed total power, user scheduling and power allocation should be jointly optimized. More particularly, for a given scheduling user set, the classic iterative water-filling based power allocation algorithm [5] [52] can be used, which achieves the maximum weighted sum of the user throughput, when exploiting the uplink-downlink duality. Given the optimal power allocation for each scheduling user set, the scheduling user set is selected by maximizing the optimization criterion.

More particularly, classic matching theory has been shown to constitute an efficient tool for user grouping in NOMA. Matching theory can be used for efficiently solving the challenging combinatorial problem of matching players from two distinct sets according to both the players' individual information and to their preferences [153] [154]. Due to its appealingly low complexity, matching theory has been widely used for solving diverse resource optimization problems in wireless NOMA networks [155]–[158]. The resource allocation in NOMA systems, such as user grouping and subchannel allocation, can be viewed as a classical matching problem, as exemplified by one-to-one [155], many-to-one [156] [157], and many-to-many [158] scenarios. The simplest matching scenarios is one-to-one matching, where each user from a set can be matched with at most one user from the opposite set. In a downlink cognitive radio aided NOMA network, a one-to-one matching problem was formulated for optimizing the user pairing in order to improve the system's throughput [155]. The pair of users belonging to the matched user pair can share the same spectrum in oredr to improve each user's data rate and the entire system's sum rate. Regarding heterogeneous NOMA networks, a many-to-one matching solution was invoked for solving a challenging spectrum allocation problem in [156], where a swap-operation aided matching algorithm was proposed for matching a small base stations with a suitable resource block, whilst aiming for maximizing the small cell users throughput. In many-to-many matching, at least one player in one set can be matched to multiple players in the opposite set. In [158], in order to maximize the system's sum rate, the subchannel allocation problem was formulated as a two-sided many-to-many matching process in the NOMA downlink. By contrast, a two-to-one matching was utilized for improving the energy efficiency of the NOMA downlink [157]. In this research, a low complexity algorithm was proposed for allocating multiple users supported by suitable subchannels for maximizing the system's energy efficiency. The subchannel allocation was considered to be a dynamic matching process between the user set and the subchannel set. According to the predefined preference lists, each user can send a matching request to its most preferred subchannel. However, the subchannel can accept or reject the user, depending on the energy efficiency the user can attain on this specific subchannel. The matching process will terminate, when no user is left to match. Furthermore, for uplink NOMA, user grouping based on exploiting the CSI knowledge was investigated by considering some predefined power allocation schemes [159], where the optimization of the user grouping formulated for achieving the maximum sum rate was analyzed in the system limit for various scenarios, and some optimum and sub-optimum algorithms associated with a polynomial-time complexity were proposed.

Additionally, the max-min fairness criterion associated with instantaneous CSI knowledge, i.e. maximizing the minimum user rate, and the min-max fairness criterion relying on the average-CSI knowledge, i.e. minimizing the maximum outage probability, have been considered in [56] for deriving the associated power allocation. Low-complexity algorithms have also been developed for solving the associated non-convex problems. The max-min fairness criterion has also been used in MIMO NOMA systems [57], where a dynamic user allocation and power optimization problem was investigated. Specifically, a sub-optimal two-step method has been proposed. In the first step, the power allocation is optimized by fixing a specific combination of user allocation according to the max-min fairness, while the second step considered all the legitimate user allocation combinations. Furthermore, the joint power and channel allocation optimization has been shown to be NP-hard in [58], and an algorithm combining Lagrangian duality and dynamic programming was proposed for delivering a competitive suboptimal solution. Furthermore, in [59], power allocation has been conceived for the NOMA downlink supporting two users when practical modulation schemes are employed. To elaborate a little further, the mutual information metric - rather than the Shannon-theoretic throughput metric - has been used for deriving a more accurate result. It has also been shown that the power allocation problem formulated for maximizing the total mutual information depends on the specific choice of the modulation schemes employed.

In addition to the spectral efficiency attained, the energy efficiency also constitutes a KPI for 5G, but a compelling

TABLE II
COMPARISON OF NOMA SCHEMES PROPOSED FOR THE REL-14 3GPP NR STUDY ITEM [142].

	NOMA Schemes	Signature	Collision Pattern	Candidate Receivers	Target Scenarios
1	Power-domain NOMA	Power domain superposition	Full collision	SIC	Low/medium SE per UE; High connection efficiency
2	SCMA	Symbol level spreading	Partial collision	MPA	Low/medium SE per UE; High connection efficiency
3	MUSA	Symbol level spreading	Partial collision	MMSE-SIC	Low/medium SE per UE; High connection efficiency
4	PDMA	Symbol level spreading	Partial collision	MPA; SIC-MPA; SIC	Low/medium SE per UE; High connection efficiency
5	LSSA	Symbol level spreading	Full collision	MMSE-SIC	Low/medium SE per UE; High connection efficiency
6	RSMA	Symbol level scrambling	Full collision	MMSE-SIC; ESE-PIC	Very low SE; Large coverage extension
7	IGMA	Symbol level interleaving	Partial collision	MPA; SIC-MPA; SIC	Low/medium SE per UE; High connection efficiency
8	IDMA	Bit level interleaving	Full collision	ESE-PIC	Low/medium SE per UE; High connection efficiency
9	NCMA	Symbol level spreading	Full collision	MMSE-SIC; ESE-PIC	Low/medium SE per UE; High connection efficiency
10	NOCA	Symbol level spreading	Full collision	MMSE-SIC; ESE-PIC	Low/medium SE per UE; High connection efficiency
11	GOCA	Symbol level spreading and scrambling	Full collision	MMSE-SIC	Low/medium SE per UE; High connection efficiency
12	LDS-SVE	Symbol level spreading	Partial collision	MPA; SIC-MPA; SIC	Low/medium SE per UE; High connection efficiency
13	FDS	Symbol level spreading	Full collision	MMSE-SIC	Low/medium SE per UE; High connection efficiency
14	LCRS	Bit level spreading	Full collision	ESE-PIC	Low/medium SE per UE; High connection efficiency
15	RDMA	Symbol level interleaving	Full collision	MMSE-SIC	Low/medium SE per UE; High connection efficiency

trade-off has to be struck between them. In [160], the perfect knowledge of the CSI was assumed at the BS and the associated energy efficiency maximization problem was investigated in the NOMA downlink by jointly optimizing both the subchannel assignment and the power allocation. Since perfect CSI knowledge cannot be readily acquired, the authors of [161] considered the energy efficiency maximization problem of the NOMA downlink in the presence of realistic imperfect CSI. As a further development, the authors of [162] studied the energy efficiency optimization of MIMO-aided NOMA systems in the face of fading channels in conjunction with statistical CSI at the transmitter, where the energy efficiency was defined in terms of the ergodic capacity attained at unity power consumption. In [163] the energy efficiency optimization problem of a cognitive radio aided multiuser downlink NOMA system was investigated subject to an individual quality of service constraint for each primary user. An efficient algorithm based on the classic sequential convex approximation method was conceived for solving the corresponding non-convex fractional programming problem formulated. Even wireless power transfer can also be integrated into NOMA systems for further improving their energy efficiency. For example, the authors of [164] considered the application of simultaneous wireless information and power transfer within a cooperative NOMA system for improving the energy efficiency of the system.

Considering the inter-cell interference, fractional frequency reuse (FFR), which allows the users having different channel conditions to rely on different frequency reuse factors, has been employed in [52], [58], [60] for further enhancing the performance of the cell-edge users. FFR-based power allocation strikes a tradeoff between the frequency bandwidth utilization per cell and the impact of inter-cell interference.

E. Comparison of NOMA Solutions

Based on the discussion above, Table II summarizes the comparison of existing dominant NOMA schemes.

From a theoretical perspective, code-domain NOMA is capable of achieving a beneficial "spreading gain" with the aid of using spreading sequences, which may also be termed as codewords. However, this benefit cannot be readily reaped by the above-mentioned power-domain NOMA regime. Achieving a "spreading gain" is an innate benefit of classic CDMA, which may also be viewed as a low-rate repetition-style channel coding scheme, where the code-rate is given by the spreading factor. In simple plausible terms, when for example one of the chips is corrupted by a high noise- or interference-sample, the specific spreading code may still be recovered by a matched filter or correlator based receiver. In contrast to CDMA, SCMA is capable of achieving an extra "shaping gain" due to the optimization of the associated multi-dimensional constellation [87].

We can also compare the dominant NOMA schemes in terms of their signalling techniques and complexity. In power-domain NOMA, SIC constitutes one of the popular interference cancellation techniques. The complexity of the SIC-MMSE is $\mathcal{O}(K^3)$, where K is the number of users supported. Therefore, the complexity of SIC is significantly lower than that of the optimal full-search-based MUD, especially when K is high, which is expected to be the case in practical 5G systems. Note that the implementation of NOMA in CoMP imposes a relatively high complexity, and a promising congenial solution to this problem has been discussed in [31]. However, side-information has to be transmitted in order to signal the associated power assignment, which imposes a signalling overhead. Additionally, in code-domain NOMA, the specific spreading sequences or codebooks have to be known at the receiver in order to support the MUD, which will increase the signalling cost, especially when the receiver does not know which users are active. On the other hand, in LDS-CDMA, LDS-OFDM, SCMA, and SAMA, the complexity of the MPA-based receiver is proportional to $\mathcal{O}(|\mathbb{X}|^w)$, where $|\mathbb{X}|$ denotes the cardinality of the constellation set $\mathbb{X}$, and w is the maximum number of non-zero signals superimposed on each chip or subcarrier. As a result, the complexity becomes high in typical scenarios of massive connectivity. The BER performance of three typical code-domain NOMA schemes is also compared in [165].

F. Performance Evaluations and Transmission Experiments of NOMA

We have provided some theoretical analysis of NOMA in the previous sections, which shows that NOMA yields a better performance than traditional OMA schemes, and this makes it a promising candidate for 5G wireless communication. In this part, we intend to present some performance evaluations and transmission experiments of NOMA, so as to verify the analytical results.

To assess the efficiency of NOMA, NTT DOCOMO performed performance evaluations and transmission experiments using prototype equipment [20], [21], [166]–[169]. Specifically, in their experiments, the radio frame configuration was designed based on LTE Release 8, and the targets of these evaluations were Transmission Mode 3 (TM3) and Transmission Mode 4 (TM4), operating without and with feeding back the user's precoding matrix index to the base station, respectively [170].

Researchers also performed experiments in an indoor radio-wave environment using the prototype equipment in [170]. In this experiment, both UE1 and UE2 are stationary, and the former was near the base station, while the latter was at a point about 50 meters from the base station. It has been shown that NOMA was capable of obtaining a throughput gain of approximately 80% over OFDMA when a 2-by-2 SU-MIMO is adopted.

Apart from NTT DOCOMO, Huawei Technologies have also developed a SCMA-based multi-user uplink prototype to verify the advantages of the SCMA technology in real communication systems [171]. The Huawei demo system consists of one base station using two antennas for diversity combined reception, and 12 single-antenna aided users for uplink access and data transmission.

TABLE III
SPECIFICATION FOR SCMA PROTOTYPE OF HUAWEI [171] ©IEEE.

Mode	Sparse code multiple access
Number of active UEs	12 out of 14
UE transmit power	23 dBm (max) with open-loop power control
Basic waveform	OFDM / F-OFDM
MIMO mode	1-by-2 SIMO
Center frequency/bandwidth	2.6GHz/20MHz
Scheduled resource	48RBs/4 RBs
Code rate	0.3-0.92
SCMA codebook	24-by-8, 4 points
Frame structure	TDD configuration 1, 4 Subframes for PUSCH

The basic system configurations of the demo are aligned with the current LTE TDD system. In particular, the researchers use the LTE TDD Configuration 1 and use OFDMA as the baseline for their performance comparison. The specifications of the prototype system are shown in Table III.

The in-lab prototype system relies on a software-defined baseband, which means that all the baseband processing is realized by a CPU instead of FPGA/DSP. At the base station side, a single server (Huawei Tecal RH2288) is responsible for all the baseband processing, which is then connected to standard commercial radio frequency components (Huawei product RRU3232). At the user side, the CPU of a laptop (MacBook Pro ME294CH/A) is used for modeling the baseband processing for two users, which is connected to two mobile RF modules for testing. A user interface (UI) is developed to show the real-time throughput of each UE, supporting the real-time change of both the user status and of the system operational modes as well.

The prototype can run in either OFDMA or SCMA mode, and supports real-time switching from one to the other. To ensure a fair comparison, same data rate is maintained for each user to guarantee the same quality of service. It is shown by the test that compared to the orthogonal multiple access baseline of 4G LTE, SCMA technology attains up to 300% throughput gain. For instance, 150% throughput gain can be observed by considering the fact that, if each user requires 12 physical resource blocks (RBs), a system having a total of 48 RBs can serve at most 4 users using orthogonal LTE OFDMA. However, with SCMA, the codebook design supports 6 users with the same throughput to simultaneously share 48 RBs, thus the equivalent delivered amount of data is actually $12 \times 6 = 72$ RBs rather than $12 \times 4 = 48$ RBs, which results in the throughput gain of about $72/48 = 150\%$. The 300% gain can be calculated in a similar way, but needs a different codebook associated with a larger spreading factor and larger number of data layers. In the prototype, a 24-by-8 SCMA codebook is used to support 12 users (each having 2 data streams) to transmit simultaneously. By contrast, for LTE OFDMA, only 4 users out of 12 can transmit data.

Apart from the fading simulator-based test in the Lab, a SCMA prototype has also been deployed in field trials to evaluate the performance. Specifically, four different test cases are designed, and UEs are deployed at different locations to

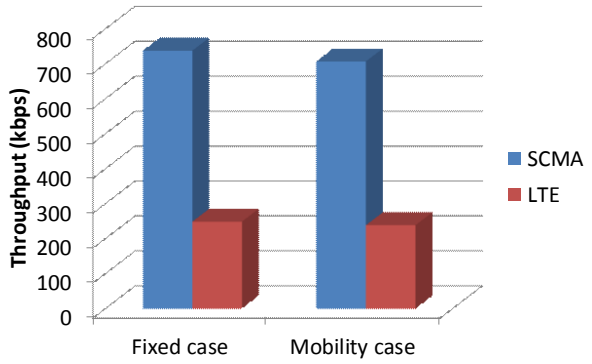

Fig. 20. SCMA throughput gain over OFDM in field testing [171] ©IEEE.

evaluate the performance of SCMA under different conditions. The four test cases are as follows:

- **Case 1:** 12 UEs closely located in an area without mobility.
- **Case 2:** 12 UEs located in an area with distant separation but no mobility
- **Case 3:** 12 UEs moving along a road about 120 meters away from the BS (open-loop power control maintains mediocre transmit power at UE))
- **Case 4:** 12 UEs are moving along a road about 180 meters away from the BS (open-loop power control provides a comparatively high power for the UE))

In all field trial tests, typical small packets of 20 bytes (METIS definition) are used as payload for both LTE and SCMA, and the scheduling resources in the whole system are limited to 4 RBSs in each subframe. The comparative test results of OFDM and SCMA are shown in Fig. 20, which indicates that SCMA achieves a nearly 300% throughput gain over OFDM.

IV. Challenges, Opportunities, and Future Research Trends

Existing NOMA schemes relying on either power-domain or code-domain multiplexing are capable of improving the spectral efficiency with the aid of non-orthogonal resource sharing. What's more, NOMA techniques are capable of operating in rank-deficient scenarios, which facilitates the support of massive connectivity. Therefore, NOMA solutions are considered as potentially promising 5G candidates. However, there are still numerous challenging problems to be solved. Hence below some of the key challenges of NOMA designs will be highlighted, along with opportunities and future research trends addressing these challenges.

A. Theoretical analysis

To elaborate a little further, in-depth theoretical analysis is required to provide additional insights to guide and inform the associated system design. The attainable capacity of multiple access schemes constitutes one of the most essential system performance criteria. Specifically, the capacity bounds of code-domain NOMA relying sophisticated spreading sequences has to be investigated. Similar methods and tools can also be conceived for MC-CDMA. On the other hand, the maximum normalized user-load that may be supported is limited both by the achievable interference cancellation capability and by the affordable receiver complexity, which is related to the specific design of both the spreading sequences and the receiver.

B. Design of spreading sequences or codebooks

In LDS systems, due to the non-orthogonal resource allocation, there is mutual interference amongst the users. The maximum number of superimposed symbols at each orthogonal resource "index" is determined by the particular spreading sequences or codewords of the users, which has a direct impact on the interference cancellation capability achieved at the receiver. Therefore, the factor graph of the message passing algorithm should be optimized to strike a compelling tradeoff between the normalized user-load supported and the receiver complexity imposed.

In addition, it has been shown that the message passing algorithm is capable of determining the exact marginal distribution in case of an idealized cycle-free factor graph, and an accurate solution can be obtained with the aid of "locally tree like" factor graphs, which implies that the cycle girth should be sufficiently high. Graph theory can be used to design cycle-free or "locally tree like" factor graphs for NOMA without any loss of spectral efficiency. On the other hand, realistic factor graphs exhibiting cycles can be decomposed into cycle-free graphs in some practical applications. In this way, the message passing based receiver is capable of attaining the optimal performance at the cost of a moderately increased receiver complexity. Additionally, the classic matrix design principle and low-density-parity-check (LDPC) code based design methods can be invoked for constructing the factor graph of NOMA solutions.

Apart from the challenge of factor graph design, we should also consider how to choose the non-zero values for each sequence. The non-zero values superimposed at the same resource indices should be distinct. A promising technique is to select different values from a complex-valued constellation for these non-zero elements in order to maintain the maximum possible Euclidean distance.

C. Receiver design

The complexity of an MPA-based receiver may still become excessive for massive connectivity in 5G. Therefore, some approximate solutions of the MPA can be used for reducing receiver complexity, such as a Gaussian approximation of the interference, which models the interference-plus-noise as a Gaussian distribution. This approximation becomes more accurate, when the number of connections becomes high, as expected in 5G. Additionally, the MPA can be used to jointly detect and channel-decode the received symbols, where the constructed graph consists of variable nodes, observation nodes and check nodes corresponding to the check equations of the LDPC code. In this way, extrinsic information can be more efficiently exchanged between the decoder and demodulator used at the receiver for improving the signal detection performance.

For an SIC-based receiver, the associated error propagation may degrade the performance of some users. Therefore, at each stage of SIC, a high-performance non-linear detection algorithm can be invoked for alleviating the influence of error propagation.

D. Channel estimation

In most NOMA contributions, perfect CSI is assumed for resource allocation or multi-user detection. However, it is not practical to obtain perfect CSI in practical systems, hence channel estimation errors exist in NOMA. The impact of the residual interference imposed by realistic imperfect channel estimation on the achievable throughput of NOMA systems has been investigated in [46] [47], and a low-complexity transmission rate back-off algorithm was conceived for mitigating the impact of the channel estimation errors. Furthermore, the design of the practical channel estimators conceived for NOMA was investigated in [172] [173], and some optimization algorithms have been proposed for reducing the channel estimation error. Nevertheless, with the increase of the number of users in future 5G systems, more grave inter-user interference will be caused, which in turn may result in severe channel estimation error. Therefore, more advanced channel estimation algorithms are required to achieve accurate channel estimation in NOMA systems.

E. Grant-free NOMA

As illustrated in Section II, a high transmission latency and a high signaling overhead are encountered by an access-grant based transmission scheme due to the uplink scheduling requests and downlink resource assignments required. It is expected that NOMA is capable of operating without grant-free transmissions at a low transmission latency, at a small signaling overhead, whilst supporting massive connectivity, especially in case transmitting short packets, as expected in 5G. Hence contention-based NOMA schemes constitute a promising solution, in which one or more pre-configured resources are assigned to the contending users. On the other hand, integrated protocols - including random back-off schemes - can be considered as a technique of resolving non-orthogonal collisions, whilst reducing the packet dropping rates. Additionally, without relying on any access-grant procedure, the BS cannot obtain any information on the associated user activity, which however can be fortunately detected by CS-aided recovery algorithms due to the sparsity of user activity.

F. Resource allocation

In power-domain NOMA, the interference cancellation capability of receivers is closely related to the accuracy of the power allocation scheme. On the other hand, the accuracy of allocating the power of each user directly affects the throughput of both power-domain NOMA and of code-domain NOMA. By carefully adjusting the power allocation under a specific total power constraint, the BS becomes capable of flexibly controlling the overall throughput, the cell-edge throughput and the rate-fairness of the users. The optimal resource allocation scheme has to search through the entire search space of legitimate solutions, and thus the complexity may become excessive. Both dynamic programming algorithms and greedy algorithms may be considered for realizing a near-optimal power allocation operating at a low complexity. Additionally, in order to support various applications, dynamic power allocation constitutes a promising research topic for future work.

G. Extension to MIMO

It is desirable to extend the existing NOMA schemes to their MIMO-aided counterparts, especially to large-scale MIMO systems, in order to further improve the attainable spectral efficiency by exploiting the spatial diversity gain and/or the multiplexing gain of MIMO systems. However, the design of MIMO-aided NOMA techniques is by no means trivial. Consider the power-domain NOMA as an example. Recall that the key idea of power-domain NOMA is to allocate the transmission power to users inversely proportionally to their channel conditions. For scenarios associated with single-antenna nodes, it is possible to compare the users' channel conditions, since channel gains/attenuations are scalars. However, in MIMO scenarios the channels are represented by a matrix. Hence it becomes difficult to decide, which user's channel is better. This dilemma leads to implementional difficulties for NOMA-solutions. This is still a promising open area at the time of writing, with very few solutions proposed in the open literature. A possible solution is to request the BS to form multiple beams, where NOMA techniques are invoked for supporting the users covered by the same directional beam and MIMO precoding/detection is used to cancel the inter-beam interference [174]. Another possible solution is to assign different beams to different users individually, where the NOMA power allocation constraint has to be taken into consideration for the design of beamforming [123].

H. Cognitive radio inspired NOMA

The advantage of NOMA techniques can be simply illustrated by exploiting the concept of cognitive radio networks. Specifically, the user associated with poorer channel conditions in a NOMA system can be viewed as a primary user in the context of cognitive radio networks. If conventional OMA is used, the bandwidth resources assigned to this primary user, such as the time slots or frequency slots are solely occupied by this user, and no other users can access these bandwidth resources, even if the primary user has a poor connection to the BS. The benefit of using NOMA is analogous to that of cognitive radio networks, where additional secondary users can be admitted into the specific beam occupied by the primary user. While these secondary users may impose performance degradations on the primary user, the overall system throughput can be significantly improved, particularly if the secondary users have better connections to the BS. By exploiting the appealing concept of cognitive radio networks we can readily illustrate the performance gain of NOMA over conventional OMA, which significantly simplifies the design of NOMA systems [174]. For example, for the scenarios

associated with MIMO schemes or in the presence of co-channel interference, the design of optimal power allocation is difficult, since it is challenging to decide the quality-order of the users' channel conditions. The exploitation of cognitive radio networks may impose new constraints on the power allocation, which has to strike a similar throughput and fairness tradeoff as conventional NOMA.

I. Further challenges

Some further challenges should also be resolved in the context of NOMA systems, including the associated reference signal design and channel estimation, the reduction of the PAPR in multi-carrier NOMA systems, such as LDS-OFDM, maintaining system scalability, the issues of channel-quality feedback design, BS cooperation, etc. Additionally, the existing multiple access design routinely assumes the employment of a single scheme for all applications, regardless of their diverse requirements. Accordingly, various system design factors have to be considered in order to accommodate the worst-case condition, which leads to inefficient multiple access design in many applications. Therefore, the software defined multiple access technology is expected to support the flexible configuration of multiple access schemes, and thus different services as well as applications can be supported in 5G. It is expected that NOMA solutions will achieve further performance improvements by addressing these challenges.

V. Conclusions

In this article, we have discussed the key concept and advantages of NOMA techniques, which constitute one of the promising technologies for future 5G systems. The dominant NOMA schemes have been introduced together with their comparison in terms of their operating principles, key features, receiver complexity, pros and cons, etc. We also highlighted a range of key challenges, opportunities and future research trends related to the design of NOMA, including its theoretical analysis, the design of spreading sequences or codebooks, their receiver design, the design issues of access-grant-free NOMA, resource allocation schemes, extensions to large-scale MIMOs and so on. It is expected that NOMA will play an important role in future 5G wireless communication systems supporting massive connectivity and low latency.

References

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[2] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[3] A. W. Scott and R. Frobenius, "Multiple access techniques: FDMA, TDMA, and CDMA," *RF Measurements for Cellular Phones and Wireless Data Systems*, pp. 413–429, Jan. 2008.

[4] H. Li, G. Ru, Y. Kim, and H. Liu, "OFDMA capacity analysis in MIMO channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4438–4446, Sep. 2010.

[5] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. E98-B, no. 3, pp. 403–414, Mar. 2015.

[6] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, May 2017.

[7] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Communications*, vol. 14, no. 4, 2016.

[8] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation in non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325-6343, Aug. 2016.

[9] M. S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565-577, Dec. 2016.

[10] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech Focus*, vol. 1, no. 2, pp. 1-6, Jun. 2017.

[11] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[12] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," to appear in *IEEE Commun. Surveys & Tutorials*, 2018.

[13] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[14] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE Intelligent Signal Processing and Communications Systems (IEEE ISPACS'13)*, Nov. 2013, pp. 770–774.

[15] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'13 Spring)*, Jun. 2013, pp. 1–5.

[16] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'13 Fall)*, Sep. 2013, pp. 1–5.

[17] N. Nonaka, Y. Kishiyama, and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'14 Fall)*, Sep. 2014, pp. 1–5.

[18] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. Vincent Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413-2424, Oct. 2017.

[19] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. Vincent Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534-537, Aug. 2017.

[20] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials," in *Proc. IEEE International Conference on Wireless Networks and Mobile Communications (IEEE WINCOM'15)*, Oct. 2015, pp. 1-6.

[21] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proc. IEEE Conference on Standards for Communications and Networking (IEEE CSCN'15)*, Oct. 2015, pp. 18-23.

[22] B. Kim, W. Chung, S. Lim, S. Suh, J. Kwun, S. Choi, and D. Hong, "Uplink NOMA with multi-antenna," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'15 Spring)*, May 2015, pp. 1-5.

[23] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537-552, Jan. 2016.

[24] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the internet of things," *IEEE Access*, vol. 4, pp. 1393-1405, Apr. 2016.

[25] Y. Lan, A. Benjebboiu, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop SU-MIMO," in *Proc. IEEE Signal Processing and Communication Systems (IEEE ICSPCS'14)*, Dec. 2014, pp. 1-5.

[26] C. Yan, A. Harada, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Receiver design for downlink non-orthogonal multiple access (NOMA)," in *Proc.*

IEEE Vehicular Technology Conference (IEEE VTC'15 Spring), May 2015, pp. 1-5.

[27] K. Saito, A. Benjebbour, Y. Kishiyama, Y. Okumura, and T. Nakamura, "Performance and design of SIC receiver for downlink NOMA with open-loop SU-MIMO," in *Proc. IEEE International Conference on Communication Workshop (IEEE ICCW'15)*, Jun. 2015, pp. 1161-1165.

[28] X. Chen, A. Beiijebbour, A. Li, H. Jiang, and H. Kayama, "Consideration on successive interference canceller (SIC) receiver at cell-edge users for non-orthogonal multiple access (NOMA) with SU-MIMO," in *Proc. IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (IEEE PIMRC'15)*, Aug. 2015, pp. 522-526.

[29] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Jun. 2015.

[30] Z. Ding, H. Dai, and H. V. Poor, "Relay Selection for Cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416-419, Jun. 2016.

[31] S. Han, C.-L. I, Z. Xu, and Q. Sun, "Energy efficiency and spectrum efficiency co-design: From NOMA to network NOMA," *IEEE MMTC E-Letter*, vol. 9, no. 5, pp. 21–24, Sep. 2014.

[32] H. Tabassum, E. Hossain, and M. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access (NOMA) in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555-3570, Aug. 2017.

[33] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," to appear in *IEEE Commun. Mag.*, 2017.

[34] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Jan. 2014.

[35] 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 Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Jul. 2014.

[36] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Personal Indoor and Mobile Radio Communications (IEEE PIMRC'13)*, Sep. 2013, pp. 611–615.

[37] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Global Communications Conference Workshops (IEEE Globecom Workshops'13)*, Dec. 2013, pp. 66–70.

[38] Y. Saito, A. Benjebbour, A. Li, K. Takeda, Y. Kishiyama, and T. Nakamura, "System-level evaluation of downlink non-orthogonal multiple access (NOMA) for non-full buffer traffic model," in *Proc. IEEE Conference on Standards for Communications and Networking (IEEE CSCN'15)*, Oct. 2015, pp. 94-99.

[39] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SU-MIMO for future LTE enhancements," in *Proc. IEEE Global Communications Conference Workshops (IEEE Globecom Workshops'14)*, Dec. 2014, pp. 706-710.

[40] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance of downlink non-orthogonal multiple access (NOMA) under various environments," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'15 Spring)*, May 2015, pp. 1-5.

[41] M. Kimura and K. Higuchi, "System-level throughput of NOMA with SIC in cellular downlink under FTP traffic model," in *Proc. IEEE International Symposium on Wireless Communication Systems (IEEE ISWCS'15)*, Aug. 2015, pp. 1-5.

[42] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *Proc. IEEE International Symposium on Wireless Communication Systems (IEEE ISWCS'12)*, Aug. 2012, pp. 261-265.

[43] P. Sedtheetorn and T. Chulajata, "Spectral efficiency evaluation for non-orthogonal multiple access in Rayleigh fading," in *Proc. IEEE International Conference on Advanced Communication Technology (IEEE ICACT'16)*, Jan. 2016, pp. 747-750.

[44] Z. Ding, F. Adachi, and H. V. Poor, "Performance of MIMO-NOMA downlink transmissions," in *Proc. IEEE Global Communications Conference (IEEE GLOBECOM'15)*, Dec. 2015, pp. 1-6.

[45] A. Li, A. Benjebbour, and A. Harada, "Performance evaluation of non-orthogonal multiple access combined with opportunistic beamforming," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'14 Spring)*, May 2014, pp. 1-5.

[46] K. Yamamoto, Y. Saito, and K. Higuchi, "System-level throughput of non-orthogonal access with SIC in cellular downlink when channel estimation error exists," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'14 Spring)*, May 2014, pp. 1-5.

[47] N. Nonaka, A. Benjebbour, and K. Higuchi, "System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists," in *Proc. IEEE International Conference On Communication Systems (IEEE ICCS'14)*, Nov. 2014, pp. 202-206.

[48] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654-667, Feb. 2016.

[49] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Apr. 2015.

[50] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Mar. 2015.

[51] K. Yakou, and K. Higuchi, "Downlink NOMA with SIC using unified user grouping for non-orthogonal user multiplexing and decoding order," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (IEEE ISPACS'15)*, Nov. 2015, pp. 508-513.

[52] J. Umehara, Y. Kishiyama, and K. Higuchi, "Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink," in *Proc. IEEE International Conference on Communication Systems (IEEE ICCS'12)*, Nov. 2012, pp. 324–328.

[53] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fairbased resource allocation," in *Proc. IEEE International Symposium on Wireless Communication Systems (IEEE ISWCS'12)*, Aug. 2012, pp.476–480.

[54] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'14 Spring)*, May 2014, pp. 1–5.

[55] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proc. IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (IEEE PIMRC'15)*, Aug. 2015, pp. 1127–1131.

[56] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[57] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Apr. 2016.

[58] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," in *Proc. IEEE Global Communications Conference (IEEE GLOBECOM'15)*, Dec. 2015, pp. 1–6.

[59] J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 483–441, Jan. 2016.

[60] Y. Hayashi, Y. Kishiyama, and K. Higuchi, "Investigations on power allocation among beams in non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'13 Fall)*, Sep. 2013, pp. 1–5.

[61] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with successive interference cancellation for future radio access," *IEEE Vehicular Technology Society, Asia Pacific Wireless communications symposium (IEEE APWCS'12)*, pp. 1–5, Aug. 2012.

[62] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *Proc. IEEE Wireless Communication Systems (IEEE ISWCS'12)*, Aug. 2012, pp. 476–480.

[63] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'11 Fall)*, Sep. 2011, pp. 1–5.

[64] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[65] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, Apr. 2008.

[66] J. Van De Beek and B. M. Popovic, "Multiple access with low-density signatures," in *Proc. IEEE Global Communications Conference (IEEE Globecom'09)*, Dec. 2009, pp. 1–6.

[67] R. Razavi, R. Hoshyar, M. A. Imran, and Y. Wang, "Information theoretic analysis of LDS scheme," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 798–800, Jun. 2011.

[68] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'10 Spring)*, May 2010, pp. 1–5.

[69] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th International Symposium on Wireless Communications Systems (IEEE ISWCS'14)*, Aug. 2014, pp. 781–785.

[70] M. Al-Imari, M. A. Imran, R. Tafazolli, and D. Chen, "Performance evaluation of low density spreading multiple access," in *Proc. IEEE Wireless Communications and Mobile Computing Conference (IEEE IWCMC'12)*, Aug. 2012, pp. 383–388.

[71] M. Al-Imari, M. A. Imran, and R. Tafazolli, "Low density spreading for next generation multicarrier cellular systems," in *Proc. IEEE Future Communication Networks (IEEE ICFCN'12)*, Apr. 2012, pp. 52–57.

[72] M. Al-Imari, M. A. Imran, R. Tafazolli, and D. Chen, "Subcarrier and power allocation for LDS-OFDM system," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'11 Spring)*, May 2011, pp. 1–5.

[73] M. Al-Imari and R. Hoshyar, "Reducing the peak to average power ratio of LDS-OFDM signals," in *Proc. IEEE Wireless Communication Systems (IEEE ISWCS'10)*, Sep. 2010, pp. 922–926.

[74] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (IEEE PIMRC'13)*, Sep. 2013, pp. 332–336.

[75] Y. Zhou, H. Luo, R. Li, and J. Wang, "A dynamic states reduction message passing algorithm for sparse code multiple access," in *Proc. IEEE Wireless Telecommunications Symposium (IEEE WTS'16)*, Apr. 2016, pp. 1-5.

[76] Y. Du, B. Dong, Z. Chen, J. Fang, and X. Wang, "A fast convergence multiuser detection scheme for uplink SCMA systems," *IEEE Commun. Lett.*, vol. 5, no. 4, pp. 388-391, May 2016.

[77] H. Mu, Z. Ma, M. Alhaji, P. Fan, and D. Chen, "A fixed low complexity message pass algorithm detector for uplink SCMA system," *IEEE Wireless Commun. Lett.*, vol. 4, no. 6, pp. 585-588, Aug. 2015.

[78] Z. Jia, Z. Hui, and L. Xing, "A low-complexity tree search based quasi-ML receiver for SCMA system," in *Proc. IEEE International Conference on Computer and Communications (IEEE ICCC'15)*, Oct. 2015, pp. 319-323.

[79] Y. Liu, J. Zhong, P. Xiao, and M. Zhao, "A novel evidence theory based row message passing algorithm for LDS systems," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[80] D. Wei, Y. Han, S. Zhang, and L. Liu, "Weighted message passing algorithm for SCMA," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[81] K. Xiao, B. Xiao, S. Zhang, Z. Chen, and B. Xia, "Simplified multiuser detection for SCMA with sum-product algorithm," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[82] Y. Du, B. Dong, Z. Chen, J. Fang, and L. Yang, "Shuffled multiuser detection schemes for uplink sparse code multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1231-1234, Jun. 2016.

[83] A. Bayesteh, H. Nikopour, M. Taherzadeh, H. Baligh, and J. Ma, "Low complexity techniques for SCMA detection," in *Proc. IEEE Global Communications Conference Workshops (IEEE Globecom Workshops'15)*, Dec. 2015, pp. 1-6.

[84] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *Proc. IEEE Global Communications Conference(IEEE Globecom'14)*, Dec. 2014, pp. 4782–4787.

[85] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE International Conference on Communications (IEEE ICC'15)*, Jun. 2015, pp. 2918-2923.

[86] B. Xiao, K. Xiao, S. Zhang, Z. Chen, B. Xia, and H. Liu, "Iterative detection and decoding for SCMA systems with LDPC codes," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[87] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'14 Fall)*, Sep. 2014, pp. 1–5.

[88] L. Yu, X. Lei, P. Fan, and D. Chen, "An optimized design of SCMA codebook based on star-QAM signaling constellations," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[89] S. Zhang, B. Xiao, K. Xiao, Z. Chen, and B. Xia, "Design and analysis of irregular sparse code multiple access," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[90] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma, and P. Zhu, "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Global Communications Conference (IEEE Globecom'14)*, Dec. 2014, pp. 1–5.

[91] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc. IEEE Wireless Communications Systems (IEEE ISWCS'14)*, Aug. 2014, pp. 853–857.

[92] H. Nikopour, E. Yi, A. Bayesteh, K. Au, M. Hawryluck, H. Baligh, and J. Ma, "SCMA for downlink multiple access of 5G wireless networks," in *Proc. IEEE Global Communications Conference (IEEE Globecom'14)*, Dec. 2014, pp. 1–5.

[93] U. Vilaipornsawai, H. Nikopour, A. Bayesteh, and J. Ma, "SCMA for open-loop joint transmission CoMP," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'15 Fall)*, Sep. 2015, pp. 1-5.

[94] T. Liu, X. Li, and L. Qiu, "Capacity for downlink massive MIMO MU-SCMA system," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'15)*, Oct. 2015, pp. 1-5.

[95] L. Lu, Y. Chen, W. Guo, H. Yang, Y. Wu, and S. Xing, "Prototype for 5G new air interface technology SCMA and performance evaluation," *China Communications*, Supplement No. 1, pp. 38-48, Dec. 2015.

[96] Z. Yuan, G. Yu, and W. Li, "Multi-user shared access for 5G," *Telecommunications Network Technology*, vol. 5, no. 5, pp. 28–30, May 2015.

[97] X. Dai, S. Chen, S. Sun, S. Kang, Y. Wang, Z. Shen, and J. Xu, "Successive interference cancelation amenable multiple access (SAMA) for future wireless communications," in *Proc. IEEE International Conference on Communication Systems (IEEE ICCS'14)*, Nov. 2014, pp. 1–5.

[98] M. Y. Alias, S. Chen, and L. Hanzo, "Multiple-antenna-aided OFDM employing genetic-algorithm-assisted minimum bit error rate multiuser detection," *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1713–1721, Sep. 2005.

[99] A. Wolfgang, S. Chen, and L. Hanzo, "Parallel interference cancellation based turbo space-time equalization in the SDMA uplink," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 609–616, Feb. 2007.

[100] L. Wang, L. Xu, S. Chen, and L. Hanzo, "Three-stage irregular convolutional coded iterative center-shifting K-best sphere detection for soft-decision SDMA-OFDM," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 2103–2109, May 2009.

[101] S. Chen, L. Hanzo, and A. Livingstone, "MBER space-time decision feedback equalization assisted multiuser detection for multiple antenna aided SDMA systems," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 3090–3098, Aug. 2006.

[102] L. Hanzo, S. Chen, J. Zhang, and X. Mu, "Evolutionary algorithm assisted joint channel estimation and turbo multi-user detection/decoding for OFDM/SDMA," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1204–1222, Mar. 2014.

[103] S. Chen, A. Wolfgang, C. J. Harris, and L. Hanzo, "Symmetric RBF classifier for nonlinear detection in multiple-antenna-aided systems," *IEEE Trans. Neural Networks*, vol. 19, no. 5, pp. 737–745, May 2008.

[104] A. Wolfgang, J. Akhtman, S. Chen, and L. Hanzo, "Reduced-complexity near-maximum-likelihood detection for decision feedback assisted space-time equalization," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2407–2411, Jul. 2007.

[105] J. Akhtman, A. Wolfgang, S. Chen, and L. Hanzo, "An optimized-hierarchy-aided approximate Log-MAP detector for MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1900–1909, May 2007.

[106] S. Chen, A. Livingstone, H. Q. Du, and L. Hanzo, "Adaptive minimum symbol error rate beamforming assisted detection for quadrature amplitude modulation," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1140–1145, Apr. 2008.

[107] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Turbo multi-user detection for OFDM/SDMA systems relying on differential evolution aided iterative channel estimation," *IEEE Trans. Commun.*, vol. 60, no. 6, pp. 1621–1633, Jun. 2012.

[108] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Joint channel estimation and multi-user detection for SDMA/OFDM based on dual repeated weighted boosting search," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3265–3275, Jun. 2011.

[109] C.-Y. Wei, J. Akhtman, S.-X. Ng, and L. Hanzo, "Iterative near-maximum-likelihood detection in rank-deficient downlink SDMA systems," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 653–657, Jan. 2008.

[110] A. Wolfgang, J. Akhtman, S. Chen, and L. Hanzo, "Iterative MIMO detection for rank-deficient systems," *IEEE Signal Process. Lett.*, vol. 13, no. 11, pp. 699–702, Nov. 2006.

[111] L. Xu, S. Chen, and L. Hanzo, "EXIT chart analysis aided turbo MUD designs for the rank-deficient multiple antenna assisted OFDM uplink," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2039–2044, Jun. 2008.

[112] "5G: Rethink mobile communications for 2020+," *FuTURE Mobile Communication Forum*, Nov. 2014.

[113] S. Kang, X. Dai, and B. Ren, "Pattern division multiple access for 5G," *Telecommunications Network Technology*, vol. 5, no. 5, pp. 43–47, May 2015.

[114] J. Huang, K. Peng, C. Pan, F. Yang, and H. Jin, "Scalable video broadcasting using bit division multiplexing," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 701–706, Dec. 2014.

[115] R. Zhang and L. Hanzo, "A unified treatment of superposition coding aided communications: Theory and practice," *IEEE Commun. Surveys & Tutorials*, vol. 13, no. 3, pp. 503-520, Jul. 2011.

[116] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press, 2005.

[117] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2006.

[118] P. P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 279–280, Mar. 1974.

[119] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, May 2016.

[120] L. Wang, X. Xu, Y. Wu, S. Xing, and Y. Chen, "Sparse code multiple access-Towards massive connectivity and low latency 5G communications," *Telecommunications Network Technology*, vol. 5, no. 5, pp. 6–15, May 2015.

[121] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649-1652, Jul. 2017.

[122] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2708-2722, Dec. 2017.

[123] 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 Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Sep. 2015.

[124] Q. Sun, S. Han, Z. Xu, S. Wang, I. Chih-Lin, and Z. Pan, "Sum rate optimization for MIMO non-orthogonal multiple access systems," in *Proc. IEEE Wireless Communications and Networking Conference (IEEE WCNC'15)*, Mar. 2015, pp. 747ÍC752.

[125] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174-6189, Dec. 2016.

[126] L. Li, L. Wang, and L. Hanzo, "Differential interference suppression aided three-stage concatenated successive relaying," *IEEE Trans. Commun.*, vol. 60, no. 8, pp. 2146–2155, May 2012.

[127] L. Zhang, W. Li, Y. Wu, X. Wang, S-I. Park, H. M. Kim, J-Y. Lee, P. Angueira, and J. Montalban, "Layered-division-multiplexing: Theory and practice," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 216–232, Mar. 2016.

[128] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[129] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.

[130] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy efficient beamspace MIMO-NOMA for Millimeter-Wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[131] A. Marcano, and H. L. Christiansen, "Performance of Non-Orthogonal Multiple Access (NOMA) in mmWave wireless communications for 5G networks," in *Proc. IEEE International Conference on Computing, Networking and Communications (IEEE ICNC' 17)*, Jan. 2017, pp. 26-29.

[132] 3GPP, "Low code rate and signature based multiple access scheme for New Radio," TSG RAN1 #85, Nanjing, China, 23rd-27th, May. 2016.

[133] 3GPP, "Discussion on multiple access for new radio interface," TSG RAN WG1 #84bis, Busan, Korea, 11th-15th, Apr. 2016.

[134] 3GPP, "Initial views and evaluation results on non-orthogonal multiple access for NR uplink," TSG RAN WG1 #84bis, Busan, Korea, 11th-15th, Apr. 2016.

[135] 3GPP, "Candidate NR multiple access schemes," TSG RAN WG1 #84b, Busan, Korea, 11th-15th, Apr. 2016.

[136] 3GPP, "Non-orthogonal multiple access candidate for NR," TSG RAN WG1 #85, Nanjing, China, 23rd-27th, May. 2016.

[137] K. Kusume, G. Bauch, and W. Utschick, "IDMA vs. CDMA: Analysis and comparison of two multiple access schemes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 78–87, Jan. 2012.

[138] 3GPP, "Considerations on DL/UL multiple access for NR," TSG RAN WG1 #84bis, Busan, Korea, 11th-15th, Apr. 2016.

[139] A. Medra and T. N. Davidson, "Flexible codebook design for limited feedback systems via sequential smooth optimization on the grassmannian manifold," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1305–1318, Mar. 2014.

[140] 3GPP, "Non-orthogonal multiple access for New Radio," TSG RAN WG1 #85, Nanjing, China, 23rd-27th, May. 2016.

[141] 3GPP, "New uplink non-orthogonal multiple access schemes for NR," TSG RAN WG1 #86, Gothenburg, Sweden, 22nd-26th, Aug. 2016.

[142] 3GPP, "Categorization and analysis of MA schemes," TSG RAN WG1 Meeting #86bis, Lisbon, Portugal, 10th-14th, Oct. 2016.

[143] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC'15 Fall)*, Sep. 2015, pp. 1–5.

[144] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473-1476, Jul. 2016.

[145] V. Shah-Mansouri, S. Duan, L.-H. Chang, V. W. Wong, and J.-Y. Wu, "Compressive sensing based asynchronous random access for wireless networks," in *Proc. IEEE Wireless Communications and Networking Conference (IEEE WCNC'13)*, Apr. 2013, pp. 884–888.

[146] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *Proc. IEEE International Conference on Communications Workshops (IEEE ICC'14)*, Jun. 2014, pp. 539–544.

[147] J.-P. Hong, W. Choi, and B. D. Rao, "Sparsity controlled random multiple access with compressed sensing," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 998–1010, Feb. 2015.

[148] "Components of a new air interface - building blocks and performance," *Mobile and wireless communications Enablers for the Twenty-twenty Information Society METIS*, Mar. 2014.

[149] 3GPP, "Initial LLS results for UL non-orthogonal multiple access," TSG RAN WG1 #85, Nanjing, China, 23rd-27th, May. 2016.

[150] 3GPP, "Multiple access schemes for new radio interface," TSG RAN WG1 #84bis, Busan, South Korea, 11th-15th, Apr. 2016.

[151] 3GPP, "New uplink non-orthogonal multiple access schemes for NR," TSG RAN WG1 #86, Gothenburg, Sweden, 22nd-26th, Aug. 2016.

[152] M. A. Sedaghat and R. R. MÍźller, "On user pairing in NOMA uplink", https://arxiv.org/abs/1707.01846, Jul. 2017.

[153] A. Roth and M. Sotomanyor, *Two Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. 1st ed. Cambridge, U.K.: Cambridge University. Press, 1989.

[154] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52-59, May 2015.

[155] W. Liang, Z. Ding, Y. Li, and L. SongH, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319-5332, Dec. 2017.

[156] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for nonorthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825-5837, Sep. 2017.

[157] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access (NOMA) network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722-3732, Jul. 2016.

[158] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory," in *Proc. IEEE Global Communications Conference (IEEE Globecom'15)*, Dec. 2015, pp. 1-6.

[159] H. Tabassum, M. S. Ali, E. Hossain, M.. J. Hossain, and Dong In Kim, "Non-orthogonal multiple access (NOMA) in cellular uplink and downlink: Challenges and enabling techniques", https://arxiv.org/abs/1608.05783, Aug. 2016.

[160] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722-3732, Sep. 2016.

[161] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA

systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874-2885, Dec. 2017.

[162] Q. Sun, S. Han, C. L. I, and Z. Pan, "Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems," in *Proc. IEEE International Conference on Communications (IEEE ICC'15)*, Jun. 2015, pp. 2668-2673.

[163] Y. Zhang, Q. Yang, T. X. Zheng, H. M. Wang, Y. Ju, and Y. Meng, "Energy efficiency optimization in cognitive radio inspired non-orthogonal multiple access," in *Proc. IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (IEEE PIMRC'16)*, Sep. 2016, pp. 1ÍC6.

[164] Y. Xu, C. Shen, Z. Ding, X. Sun, S. Yan, G. Zhu, and Z. Zhong, "Joint beamforming and power splitting control in downlink cooperative swipt noma systems," *IEEE Trans. Signal Process.*, vol. 15, no. 18, pp. 4874ÍC4886, Sep. 2017.

[165] B. Wang, K. Wang, Z. Lu, T. Xie, and J. Quan, "Comparison study of non-orthogonal multiple access schemes for 5G," in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (IEEE BMSB'15)*, Jun. 2015, pp. 1–5.

[166] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annu. Symp. PIMRC*, London, U.K., Sep. 2013, pp. 611–615.

[167] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA) under various environments," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC-Spring' 15)*, May. 2015, pp. 1–5.

[168] K. Saito, A. Benjebbour, A. Harada,Y. Kishiyama, and T. Nakamura, "Link-level performance evaluation of downlink NOMA with SIC receiver considering error vector magnitude," in *Proc. IEEE Vehicular Technology Conference (IEEE VTC-Spring' 15)*, May. 2015, pp. 1–5.

[169] K. Saito, A. Benjebbour, Y. Kishiyama, Y. Okumura, and T. Nakamura, "Performance and design of SIC receiver for downlink NOMA with open-Loop SU- MIMO," in *Proc. IEEE International Conference on Communication Workshop (IEEE ICCW' 15)*, June. 2015, pp. 1161–1165.

[170] A. Benjebbour, K. Saito, Y. Saito, and Y. Kishiyama, "5G radio access technology," *NTT DOCOMO Technical Journal*, vol. 17, no. 4, pp. 16–28, 2015.

[171] L. Lu, Y. Chen, W. Guo, H. Yang, Y. Wu, and S. Xing, "Prototype for 5G new air interface technology SCMA and performance evaluation", *China Communications*, vol. 12, no. supplement, pp. 38–48, Dec. 2015.

[172] Y. Tan, J. Zhou, and J. Qin, "Novel channel estimation for non-orthogonal multiple access systems," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1781-1785, Dec. 2016.

[173] K. Struminsky, S. Kruglik, D. Vetrov, and I. Oseledets, "A new approach for sparse Bayesian channel estimation in SCMA uplink systems," in *Proc. IEEE International Conference on Wireless Communications & Signal Processing (IEEE WCSP'16)*, Oct. 2016, pp. 1-5.

[174] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Sep. 2015.