Non-Orthogonal Multiple Access: A Unified Perspective

Qi Wang, Rong Zhang, Lie-Liang Yang, and Lajos Hanzo

***Abstract*—Non-orthogonal multiple access (NOMA) is a promising technique for future mobile communication systems, which can approach multiuser channel capacity by sharing the same time-frequency resources with multiple users. In this article, we provide a unified framework for NOMA and review the principles of various NOMA schemes in different domains with the objective of creating a unified framework. A systematic performance comparison of different NOMA schemes regarding their peak-to-average power ratio, receiver complexity, latency, grant-free access, user load and peak throughput is also provided for different application scenarios. Relying on our unified framework, we generalize the current understanding of NOMA principle from the conventional code and power domains to the spatial domain as well as to their hybrids and to the networking domain. Finally, the challenges in terms of resource allocation, channel estimation, security, system flexibility and implementation issues are also discussed.**

I. Introduction

As a key technique of improving the system capacity, multiple access solutions have witnessed a substantial evolution over the past decades. Both orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) have been used in mobile communication systems. Frequency-division multiple access (FDMA) is employed in the first generation (1G) mobile communication systems, while time-division multiple access (TDMA) is used in most of the 2G systems. Both FDMA and TDMA belong to the OMA family since orthogonal resources are allocated to multiple users in the frequency or time domain to avoid multiuser interference (MUI). In the 2G and 3G mobile systems, such as IS-95, CDMA2000 and the uplink of WCDMA, code-division multiple access (CDMA) is adopted, which belongs to NOMA since the spreading sequences of the users do not have to be orthogonal. Although typically a Rake receiver is used for low-complexity single-user detection, multiuser detection (MUD) is required for CDMA to approach the single-user performance for mitigating the MUI and the channel induced intersymbol interference (ISI). In the LTE and LTE-Advanced (4G) systems, orthogonal frequency-division multiple access (OFDMA) is used in the downlink and single-carrier frequency division multiple access (SC-FDMA) is employed in the uplink, which belong to the OMA family.

The authors are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: qw1e16@soton.ac.uk, {rz,lly,lh}@ecs.soton.ac.uk).

The financial support of the EPSRC project EP/P034284/1 and that of the European Research Councils Advanced Fellow Grant is gratefully acknowledged.

IEEE Wireless Communications, 2018

In recent years, the explosive expansion of the mobile Internet has imposed new demands on the mobile networks by supporting enhanced mobile broadband services and wide area Internet of Things (IoT), which expedited the development of the next generation (5G) mobile networks. Compared to 4G networks, 5G will support a significantly higher system capacity, high peak data rate, massive number of mobile device connections, low power consumption and latency. In IoT applications, grant-free access is required in the uplink to avoid a high signalling overhead and system latency, where the conventional request-grant procedure can be omitted, while the user identities and the user data can be transmitted to the base station (BS) simultaneously. Therefore, various NOMA techniques have been proposed, which can be classified into code-domain and power-domain NOMA. The code-domain NOMA family is developed from classic CDMA, including interleave-division multiple access (IDMA), low density spreading aided CDMA (LDS-CDMA), multiuser shared access (MUSA), resource spread multiple access (RSMA), and sparse code multiple access (SCMA), all of which distinguish multiple users with the aid of some generalized "codewords" [1]–[5]. The power-domain NOMA was originally proposed as power division multiple access (PDMA) by assigning different users different power [6]. It was then theoretically studied in [7] in conjunction with superposition coding (SPC) at the transmitter and successive interference cancellation (SIC) at the receiver.

Given the diverse instantiations of NOMA techniques, their unification is expected to lead to new insights and to inspire further advances. Against this backdrop, in this article, a unified framework is proposed for revealing the relationships of different NOMA techniques, where the common principles of different NOMA techniques are waved together according to the specific dimension used for user separation. Based on our unified framework, a systematic evaluation of different NOMA schemes is provided in terms of their peak-to-average power ratio (PAPR), receiver complexity, latency, grant-free access, user load and peak throughput. We further generalize the current understanding of the NOMA principle, expanding it from the conventional code and power domains to the spatial domain as well as to their hybrid versions and to the networking domain. Finally, the associated open issues and opportunities of NOMA are discussed, focusing on their resource allocation, channel estimation, security, system flexibility and implementation issues.

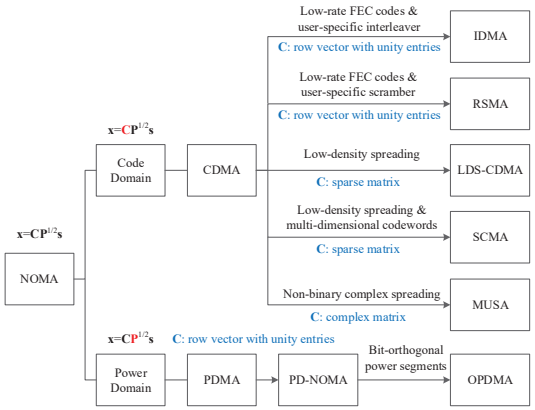

Fig. 1. Classification of different NOMA techniques.

II. From Unification to Specification

A. A Unified Model

Viewed from a unified perspective, in all the NOMA schemes, the signals of multiple users share the same time-frequency resources. The information of each user is conveyed by user-specific spreading codes and/or mapped to different power levels for user separation. Hence, the canonical base-band model of the transmitted signal vector $\mathbf{x}$ can be written in a unified form as

$$\mathbf{x} = \mathbf{C}\mathbf{P}^{1/2}\mathbf{s}, \tag{1}$$

where $\mathbf{s} \in \mathcal{C}^{K\times 1}$ is the symbol vector of K users, $\mathbf{P} = \text{diag}\left(p_1, p_2, \ldots, p_K\right)$ is the power allocation matrix, while $\mathbf{C} \in \mathcal{C}^{N_s\times K}$ denotes the spreading matrix. Several NOMA techniques have been proposed in the literature, most of which can be categorized to two types according to their properties, namely code-domain NOMA and power-domain NOMA. For code-domain NOMA, the spreading matrix $\mathbf{C}$ can be specifically designed for supporting particular user-load and complexity requirements. For power-domain NOMA, the spreading matrix $\mathbf{C}$ is simply a row vector having unity entries, while the optimization is mainly based on the specific choice of $\mathbf{P}$. It is also worth noting that when $\mathbf{C}$ is an identity matrix, the above-mentioned unified model represents OMA. In a nutshell, the different NOMA techniques are summarized in Fig. 1 according to our unified model. Their principles are detailed in Section II-B and II-C.

At the receiver, the spread multiuser signal carries all the information needed for detection. For the uplink channel, the equivalent received signal can be written as $\bar{y} = \sum_{k=1}^{K} \bar{h}_k s_k + \bar{w}$, which can be obtained via various signal processing techniques according to the specific spreading methods used, where $\bar{h}_k$ is the equivalent channel gain of user k and $\bar{w} \sim \mathcal{CN}\left(0, N_0\right)$ is the equivalent additive white Gaussian noise (AWGN). The k-user capacity region is described by $2^K - 1$ constraints. For each non-empty subset $\mathcal{S} \subset \{1, 2, \ldots, K\}$ of users, the achievable rates satisfy [8]

$$\sum_{k\in\mathcal{S}} R_k \leq \log_2\left(1 + \frac{\sum_{k\in\mathcal{S}} p_k |\bar{h}_k|^2}{N_0}\right). \tag{2}$$

Theoretical results show that the $K!$ corner points of the rate-region can be achieved by SIC, and the decoding process would start from the specific user with the strongest power and gradually proceed to the weaker ones [8].

In the downlink, let $\bar{w}_k \sim \mathcal{CN}\left(0, N_0\right)$ denote the equivalent AWGN for user k. Without loss of generality, we assume $|\bar{h}_1| \leq |\bar{h}_2| \leq |\bar{h}_k| \cdots \leq |\bar{h}_K|$. The equivalent received signal at user k is represented by $\bar{y}_k = \bar{h}_k \sum_{j=1}^{K} s_j + \bar{w}_k$, $k = 1, 2, \ldots, K$. For all possible power splits of the total power at the BS, the boundary of the capacity region of user k is given by [8]

$$R_k = \log_2\left(1 + \frac{p_k |\bar{h}_k|^2}{N_0 + \left(\sum_{j=k+1}^{K} p_j\right)|\bar{h}_k|^2}\right), \; k = 1, 2, \ldots, K, \tag{3}$$

which can be achieved theoretically by SPC at the transmitter and SIC at the receiver. The cancellation order at each receiver always starts from the user having a lower channel gain, since a higher power is assigned to it in SPC.

In its conceptually simplest guise, optimal detection is based on the maximum *a posteriori* (MAP) principle, which calculates the *a posteriori* probabilities (APPs) of all the possible transmitted symbols for all users. Hence it has a huge complexity even for a low number of users. To reduce the complexity, the message passing algorithm (MPA) based on the classic factor graph representation was proposed for approximating the MAP solution. Specifically, by denoting the transmitted symbols and the observations as variable nodes and check nodes, the real-valued messages are exchanged by the connected edges between the variable nodes and the check nodes. Alternatively, the MUI can be estimated and subtracted from the received signal, followed by a single-user matched filter, decorrelator or linear minimum mean-square-error (MMSE) filter. According to the specific order in which computation is carried out through the nodes, the interference cancellation can be serial (successive) or parallel, where the former structure is much simpler to implement. When the soft-in soft-out (SISO) decoders' *extrinsic* information is obtained, the detector can iteratively estimate the MUI by either hard or soft symbol-by-symbol decision, which are hence termed as hard or soft interference cancellation.

B. Specification - Code-Domain NOMA

The code-domain NOMA was developed from the classic CDMA, which distinguishes multiple users with the aid of unique, user-specific spreading sequences in the columns of $\mathbf{C}$. Multiple users can share the same time and frequency resources. If the spreading sequences assigned are perfectly orthogonal and synchronous for the different users, the multiuser channel can be decoupled into a set of single-user channels with the aid of matched filters. In practical systems, however, the orthogonality cannot be satisfied for various reasons. For instance, the number of users may be higher than the number of orthogonal spreading sequences, i.e. the length of the spreading sequences. Even if orthogonal spreading sequences are employed in synchronous downlink

transmission, the orthogonality will be destroyed by the delay spread of the channel. It is also widely recognized that in uplink transmission, non-orthogonal pseudo-random spreading sequences are required owing to the system's asynchronous nature. Therefore, CDMA constitutes a classic NOMA technique, which leads to the various code-domain NOMA solutions summarized below.

- **Interleave-division multiple access**: IDMA can be considered as a special case of CDMA with a spreading gain of one and using low-rate FEC codes. Effectively it is a chip-interleaved CDMA scheme. Inspired by the theoretical results of [9], namely that the optimal spectral efficiency can be achieved by a low spreading factor combined with low-rate FEC codes, a user-specific interleaver acts as the only means of distinguishing multiple users in IDMA systems [2]. If the interleavers do not guarantee orthogonality, IDMA constitutes a NOMA scheme. Since no spreading is used, the MUD can be simplified to a chip-by-chip elementary signal estimator (ESE) at the IDMA receiver. For each user, the ESE estimates transmitted chip based on the mean and variance of the received signal as well as on the basis of the distortion. The FEC decoders then calculate the APP of each chip and feed them back to the ESE for iterative detection. Therefore, the complexity of MUD per user increases only linearly with the number of users. *Naturally, the spreading matrix* $\mathbf{C}$ *of IDMA is simply a row vector having unity entries*.
- **Resource spread multiple access**: RSMA uses low-rate FEC codes and spreading/scrambling codes with good correlation properties for separating the symbols of multiple users. Similar to IDMA, low rate FEC codes are used for achieving a high coding gain, and long spreaders/scramblers are employed to reduce the MUI. The employment of different interleavers is optional for RSMA, based on the particular system requirements. Depending on the specific application scenarios, either single-carrier RSMA (SC-RSMA) or multi-carrier RSMA (MC-RSMA) may be used. The SC-RSMA is typically used for short data transmission bursts with battery power saving and coverage extension, which utilizes low-PAPR single-carrier waveforms (e.g., filtered $\frac{\pi}{2}$-BPSK, filtered offset QPSK and constant envelope modulations) and allows grant-free transmission as well as asynchronous access. By contrast, MC-RSMA is optimized for low-latency access. Grant-free transmission is also possible. *Since both the spreaders and scramblers operate at the bit level before symbol mapping,* $\mathbf{C}$ *is also a row vector with unity entries*.
- **Low density spreading aided CDMA**: LDS-CDMA is also a special case of CDMA, which employs LDS sequences for different users. In order to reduce the interference imposed on each chip, the number of non-zero elements in the spreading sequences should be much lower than the number of chips, so that a chip-level iterative SISO MUD relying on the *sum-product* algorithm can be employed at a reasonable complexity. However, the sparsity will be destroyed in multipath fading channels, which significantly increases the detection complexity. An option to solve this problem is a combination of LDS-CDMA and OFDM, resulting in low-density spreading aided OFDM (LDS-OFDM), in which the symbols are mapped to the OFDM subcarriers according to the LDS sequence assigned to each user. Since only a few subcarriers carry information symbols, the MUD based MPA developed for LDS-CDMA can also be used in the frequency domain for LDS-OFDM. *In LDS-CDMA,* $\mathbf{C}$ *is a sparse matrix in the unified model, which requires specific designs*.

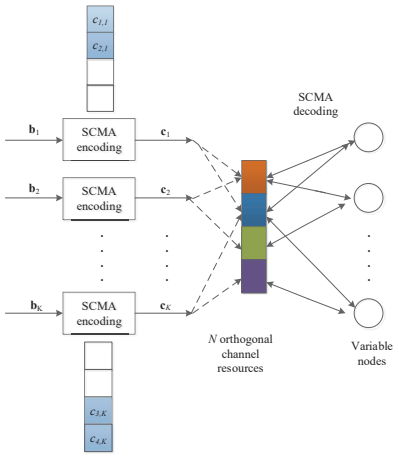

Fig. 2. SCMA encoding and decoding.

- **Sparse code multiple access**: SCMA is based on LDS-CDMA, where LDS sequences are used for user separation. However, unlike LDS-CDMA, SCMA directly maps the transmitted bits to complex-valued multi-dimensional codewords for each user. The conceptual diagram of the SCMA system is illustrated in Fig. 2, where $\mathbf{c}_k$ denotes the codeword assigned to the transmitted bit vector $\mathbf{b}_k$ of user k. Each user has a unique codebook and the codewords in the same codebook share the same sparse pattern. The codebooks assigned to the different layers are non-orthogonal, thus the number of users can be higher than the number of orthogonal resources and hence a massive number of connections can be supported [4]. At the receiver, SCMA has to eliminate the interference by an iterative MUD based on the MPA, which has a low computational complexity as a benefit of sparse codewords. The usage of multi-dimensional constellations results in considerable "shaping gain", which is the main difference between SCMA and LDS-CDMA. A higher shaping gain can be achieved, when the constellation is more close to a sphere, which is reminiscent of the sphere packing principle. *In SCMA, the spreading matrix* $\mathbf{C}$ *is a sparse matrix composed of* $\mathbf{c}_k$. *Since multi-dimensional constellations are used, the codeword of each user is jointly determined by the transmitted bit vector and the user-specific codebook, while the symbol vector* $\mathbf{s}$ *in the unified model* (1) *is a row vector with unity entries*.

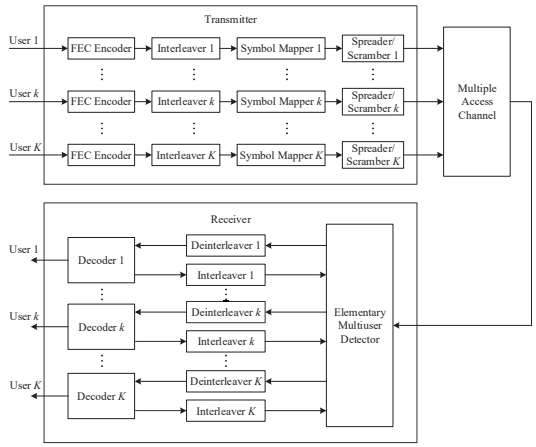

Fig. 3. General transceiver structure of code-domain NOMA.

- **Multiuser shared access**: MUSA adopts complex-valued non-binary spreading sequences to distinguish the data of different users. In the classic CDMA scheme, long pseudo-random spreading sequences are used to reduce the correlation, hence leading to a high computational complexity and to a potentially long delay at the receiver. In MUSA, the length of the complex-valued spreading sequences can be short due to the additional freedom provided by the imaginary part. The elements of the complex spreading sequences can be chosen from two-dimensional constellations, such as quadrature amplitude modulation (QAM) and additionally including the origin. For example, when rectangular 8-QAM is used, the set is $\{0, 1, 1+i, i, -1+i, -1, -1-i, -i1-i\}$ and 9^L sequences can be generated to support a large number of users, where L is the length of the spreading sequences. At the receiver, SIC is used to cancel the interference between users, under the assumption of perfect synchronization. Each user can choose its spreading sequence autonomously, thus grant-free access is achieved, which reduces the signaling overhead and latency. *In this case, the spreading matrix* $\mathbf{C}$ *is a complex-valued matrix of a small size*.

As discussed above, all the code-domain NOMA schemes can be viewed as variants of CDMA. Therefore, the transmitter and receiver of any code-domain NOMA scheme can be represented by the general schematic of Fig. 3. The elementary MUD deals with the multiuser interference before FEC decoding, whose outputs are then passed to the SISO decoders for the sake of achieving a coding gain. In IDMA and RSMA, low-rate FEC codes are used for achieving a coding gain, while in LDS-CDMA and SCMA, LDSs are used for facilitating low-complexity MAP detection by exploiting their sparsity. For NOMA schemes using a SIC receiver, the elementary MUD directly subtracts the estimated interference from the upper layers in order to attain the low complexity. In MPA-based receivers, however, the elementary MUD iteratively exchanges the information among multiple users via the factor graph to achieve an improved performance.

C. Specification - Power-Domain NOMA

The idea of power-domain NOMA (PD-NOMA) dates back to the conception of PDMA proposed in 1998 [6], where multiple users have different receive power in the uplink so that SIC can be employed. Recently, the PD-NOMA with SPC at the transmitter and SIC receiver was analyzed theoretically and attracted extensive attention [7], [10]–[12]. *In PD-NOMA, the symbol vector of multiple users is directly multiplied by the power allocation matrix* $\mathbf{P}$ *without any need for spreading, and the system model of PD-NOMA is identical to the aforementioned unified model. In other words, PD-NOMA employs the same codeword "1" for different users at a different power, and its spreading matrix can be considered to be the row vector* $\mathbf{CP}^{1/2} = \left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K}\right)^T$. Therefore, the transceiver structure of Fig. 3 is also suitable for PD-NOMA, when a SIC receiver is used in the elementary MUD. In a single-cell downlink scenario, we assume that the noise plus inter-cell interference is the same for each user for simplicity. The optimal decoding order relies on decoding the user having the weaker channel gain first, while treating the signals having higher channel gains as interference. Therefore, the user having a higher channel gain is assigned a lower power to support the SIC. In the uplink, SIC is performed at the BS, and the optimal decoding order ranges from x_1 to x_K if $p_1|h_1|^2 \geq p_2|h_2|^2 \geq p_k|h_k|^2 \cdots \geq p_k|h_K|^2$, where h_k is the channel gain of user k.

To a certain extent, the power-domain NOMA can be viewed as being "orthogonal" if power divison multiplexing (PDM) is considered [13]. By dividing the transmit power into multiple power segments (PSs), multiple information streams can be transmitted at different PSs. For a specific normalized noise power and channel gain, the first PS is set to $e_1 = 2^r - 1$ in order to achieve a spectral efficiency of r bit/s/Hz according to the Shannon capacity. When considering the first PS as interference, the second PS is given by $e_2 = (1 + e_1)(2^r - 1)$ for the spectral efficiency of r bit/s/Hz, which ensures bit-orthogonality in the power domain. Naturally, the PS associated with a higher index is higher than the one with a lower index. Based on this observation, orthogonal power division multiple access (OPDMA) was proposed, which relied on bit-orthogonal PSs capable of maintaining rate-fairness, which used SPC at the transmitter and SIC at the receiver. In fact, OPDMA can be regarded as an alternative interpretation of PD-NOMA [13].

D. Comparisons of NOMA Techniques

Theoretically, NOMA techniques are capable of approaching the multiuser capacity with the aid of an optimal MUD. However, there are different constraints, which limit their performances in practical systems, such as the complexity of the receiver, energy efficiency, peak data rate as well as the synchronization and latency requirements. For example, in massive machine type communication (mMTC) systems, a large number of users have to be supported, who are usually equipped with limited battery power. Therefore, low-rate FEC codes have to be used to mitigate the interference and path loss. Since the mMTC uplink traffic is typically

constituted by short data bursts, grant-free access should be supported for reducing the overhead and latency. Furthermore, the battery life of mMTC terminals has to be very long, which requires a low-complexity receiver and low-PAPR signals to increase the efficiency of the power amplifier. By contrast, in enhanced mobile broadband (eMBB) scenarios, high peak spectral efficiency is the major concern, whilst the complexity of the BS can be high and the PAPR of transmitted signals as well as the complexity of the mobile receiver should be low. Moreover, in ultra-reliable and low-latency communication (URLLC) applications, such as user-specific 3D video rendering and augmented reality, vehicle-to-vehicle communications, and remote control of cars, drones and robots, low latency and low packet loss ratios are the prime concerns. Based on the different application scenarios, different NOMA schemes are suitable for the application-specific constraints.

Table I compares different NOMA techniques in terms of their PAPR, receiver complexity, latency, grant-free access, user load and peak throughput. The PAPRs of MC-RSMA, MUSA and PD-NOMA are higher due to their superposition nature, which are more vulnerable to nonlinearity and require a high-linearity power amplifier at the transmitter. The SIC structure of the PD-NOMA receiver leads to a high complexity and latency. However, when an MPA-based receiver is used, which is the case for IDMA, LDS-CDMA and SCMA, both the complexity and the latency of the receiver are higher, even when the spreading matrix is sparse. Code-domain NOMA is capable of supporting grant-free access, but its peak throughput is limited due to its spread-spectrum nature. For code-domain NOMA, the user load is determined by the number of available spreading codes, which is relatively low in RSMA. In PD-NOMA, a high user load may be maintained at the cost of a complex resource allocation strategy and high receiver complexity.

III. Generalization and Research Opportunities

A. Spatial-Domain NOMA

Multiuser multiple-input multiple-output (MIMO) systems, including space division multiple access (SDMA), massive MIMO, coordinated multipoint (CoMP), cooperative MIMO, etc. belong to the category of spatial-domain NOMA. Multiuser MIMO systems support multiple users in different geographical locations, who are allowed to simultaneously share the same time/frequency/code resources. Here, the unique, user-specific channel impulse responses (CIRs) can be viewed as codewords, which serve the same purpose as the codewords in the other code-domain NOMA systems. Specifically, in an uplink SDMA system where each user having a single antenna simultaneously communicates with a BS equipped with an array of receive antennas, the MIMO channel's input-output relationship can be described as $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is the complex-valued Gaussian noise. In this equation, the data symbol x_k sent by the kth user is related to $\mathbf{y}$ through the kth column of $\mathbf{H}$, and $\mathbf{H}$ is usually a dense matrix having complex-valued random elements. By contrast, in a cooperative MIMO system, where several BSs cooperatively transmit/receive, $\mathbf{H}$ may not be full of nonzero elements, due to the users' limited transmit power and owing to the signals' propagation path loss. This is especially the case, when the BSs employ cooperative distributed antennas.

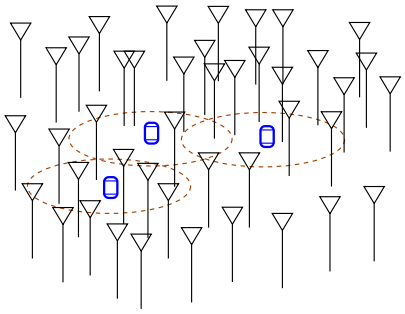

Fig. 4. In ultra-dense wireless network, a user communicates via a few antennas distributed around it.

Future global 5G and 5G+ wireless systems are expected to rely on an ultra-dense heterogeneous wireless network associated with densely distributed antennas, similar to the scenario shown in Fig. 4. In this ultra-dense network, each user only has access to a few antennas around it, typically within its line-of-sight (LoS) region. In this case, the MIMO channel matrix $\mathbf{H}$ relating the users to the distributed antennas within a region may be sparse, with each column only having a small fraction of nonzero random elements. Consequently, the MPA algorithm developed for other code-domain NOMA schemes may be also employed for achieving near-optimal performance.

B. Hybrid NOMA Schemes

The various NOMA schemes discussed above may also be combined in order to form hybrid NOMA systems. For example, the combination of SDMA and CDMA yields a space-code division based NOMA scheme, which relies on applying sparse codes and/or distributed antennas, in order to support more users than the number of unique resources, whilst facilitating MPA-aided detection to achieve near-MAP performance. The distributed antenna aided CDMA system proposed in [14] constitutes another hybrid NOMA scheme, which is sparse in the spatial-domain. The code-domain NOMA philosophy can also be amalgamated with multi-carrier systems relying on OFDM and on its variants, where the information bits are encoded by user-specific spreading codes and then spread over the corresponding OFDM subcarriers, yielding a combination of NOMA and OFDMA. Additionally, the integration of MIMO and NOMA can be used for improving the user coverage and data rate [15].

C. Research Opportunities

1) Resource Allocation: Resource allocation plays an important role in the overall system's performance optimization. In power-domain NOMA, since SIC supporting a large number of users leads to a high complexity and latency, multiple users may be allocated to different groups, where non-orthogonal resources are shared within each group and orthogonal resources

TABLE I
COMPARISON OF DIFFERENT NOMA TECHNIQUES

NOMA techniques	PAPR	Receiver complexity	Latency	Grant-free access	User load	Peak throughput
IDMA	Low	High	High	Yes	High	Low
RSMA	Low (SC-RSMA) High (MC-RSMA)	Low	High (SC-RSMA) Low (MC-RSMA)	Yes	Low	Low
LDS-CDMA	Low	High	High	Yes	Moderate	Low
SCMA	Low	High	High	Yes	Moderate	Low
MUSA	High	Low	Low	Yes	High	Low
PD-NOMA	High	Moderate	Moderate	No	High	High

are split among different groups. The optimal resource allocation can only be obtained via exhaustive search, which has a high complexity especially in the scenario where the topology of users changes quickly. When a fixed resource allocation scheme is used, the performance of NOMA degrades and may even become inferior to OMA. Therefore, robust yet low-complexity resource allocation is required for PD-NOMA. In code-domain NOMA, the spreading sequences and codebooks should be optimized, while appropriate codeword assignment is required to minimize the interference.

2) Implementation Issues: In NOMA schemes, both SIC and MUD receiver schemes can be used to deal with the multiuser interference, whose complexity increases at a polynomial order or even at an exponential order with the number of users. In order to have a low complexity and latency, only a small number of users can be grouped to share the same time-frequency resources, which limits the attainable system performance. In addition, the SPC requires costly high-resolution analog-to-digital converters (ADCs) to reduce the quantization error, which is also unfeasible for a large number of users in NOMA. Moreover, the PD-NOMA requires instantaneous channel state information (CSI) for power allocation, and SIC requires the knowledge of both the constellation and of the power allocation of each user, which increases the signaling as well as processing overhead and hence reduces the spectral efficiency. An efficient scheme should be developed for reducing the dependence on the CSI. Furthermore, non-coherent schemes should be investigated, which can dispense with pilot-aided channel estimation. Finally, most research on NOMA only considered synchronous transmissions, which is impractical, especially for the uplink due to the users' random distribution and movements.

3) Channel Estimation: The achievable performance of NOMA systems is heavily dependent on reliable channel estimation. Otherwise, a significant performance loss should be expected. For example, in PD-NOMA systems, in addition to its own channel, a user has to estimate the channels of all the other users having a higher power than the power of itself, in order to invoke SIC. Any unreliable estimation of these channels will degrade the signal detection performance, due to the potentially unreliable detection of the stronger users as well as owing to the resultant error propagation. Future NOMA systems are expected to support a large number of users. Without any doubt, reliable channel estimation under the constraint of limited pilot overhead remains an open challenge for practical NOMA systems.

4) System Flexibility: Since future networks have to operate in diverse application scenarios, it is necessary to ensure that the NOMA schemes have the flexibility to support diverse requirements, such as a high user overload, wide coverage area, high data rate and low latency. It should also be compatible with single-carrier and multi-carrier modulation as well as with other new waveforms conceived for 5G. Inspired by the software defined radio concept, a software defined multiple access (SoDeMA) was proposed in [1] to allow different NOMA schemes to coexist, so that flexible configurations can be used to support different services and applications. However, it remains an open question as to how we can exploit the benefits of different domains for improved flexibility.

5) Security: In certain NOMA systems, a user has to decode some of the other users' information, in order to detect its own. For example, in the PD-NOMA systems, a user has to first decode all the stronger users before detecting its own signal. Explicitly, this imposes a security concern on the other users. In the code-domain NOMA systems, a user can also decode other users' information by exploiting the knowledge of their spreading codes. Additionally, NOMA systems also experience all the typical security threats of all the other types of wireless systems. Therefore, it is critical to design secure NOMA systems and this challenge might be solved with the aid of physical layer security techniques.

IV. CONCLUSIONS

In this article, a unified framework was proposed for NOMA to analyze its performance. The principles of code-domain and power-domain NOMA techniques were introduced and their performances were compared in terms of PAPR, receiver complexity, latency, grant-free access, user load and peak throughput. Depending on their specific properties, they can be used in different scenarios under application-specific constraints. Based on our unified framework, we stretched the current concept of the NOMA principle from conventional code and power domains to the spatial domain as well as to their hybrids and to the networking domain. Finally, we have listed some open challenges for NOMA, including resource allocation, channel estimation, security, system flexibility and implementation issues, which can be useful for the deployment of NOMA for future networks.

REFERENCES

[1] 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.

[2] P. Li, "Interleave-division multiple access and chip-by-chip iterative multi-user detection," *IEEE Radio Commun.*, vol. 43, no. 6, pp. S19–S23., Jun. 2005.

[3] *Resource Spread Multiple Access*, 3GPP R1-164688, May 2016.

[4] H. Nikopour and H. Baligh, "Sparse Code Multiple Access," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2013, pp. 332–336.

[5] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, K. Niu, "Pattern division multiple access-A novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.

[6] G. Mazzini, "Power division multiple access," in *Proc. IEEE Int. Conf. Universal Pers. Commun. (ICUPC)*, vol. 1. Oct. 1998, pp. 543–546.

[7] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, Jun. 2013, pp. 1–5.

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

[9] S. Verdú, and S. Shlomo, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. on Inform. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[10] 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.

[11] 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, Dec. 2014.

[12] 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. 99, no. PP, pp. 1–42, Oct. 2016.

[13] W. Han, Y. Zhang, X. Wang, J. Li, M. Sheng, and X. Ma, "Orthogonal power division multiple access: A green communication perspective," *IEEE J. Sel. Area. Commun.*, vol. 34, no. 12, pp. 3828–3842, Dec. 2016.

[14] L.-L. Yang and W. Fang, "Performance of distributed antenna DS-CDMA systems over composite lognormal shadowing and Nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 2872–2883, Jul. 2009.

[15] 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," to appear in *IEEE J. Sel. Areas Commun.*, 2017.