Signal Processing and Learning for Next Generation Multiple Access in 6G

Wei Chen, *Senior Member, IEEE*, Yuanwei Liu, *Fellow, IEEE*, Hamid Jafarkhani, *Fellow, IEEE*,
Yonina C. Eldar *Fellow, IEEE*, Peiying Zhu, *Fellow, IEEE*, Khaled B Letaief, *Fellow, IEEE*

Abstract—**Wireless communication systems to date primarily rely on the orthogonality of resources to facilitate the design and implementation, from user access to data transmission. Emerging applications and scenarios in the sixth generation (6G) wireless systems will require massive connectivity and transmission of a deluge of data, which calls for more flexibility in the design concept that goes beyond orthogonality. Furthermore, recent advances in signal processing and learning, e.g., deep learning, provide promising approaches to deal with complex and previously intractable problems. This article provides an overview of research efforts to date in the field of signal processing and learning for next-generation multiple access, with an emphasis on massive random access and non-orthogonal multiple access. The promising interplay with new technologies and the challenges in learning-based NGMA are discussed.**

I. INTRODUCTION

THE fifth generation (5G) wireless systems are currently reaching commercial maturity, and next-generation wireless networks are expected to support extremely high data rates and radically new applications, which require massive connectivity, such as Internet of Everything (IoE), metaverse, augmented reality and industry 4.0. The International Data Corporation (IDC) forecasts that by 2025 there will be 55.7 billion connected devices, most of which will be Internet of Things (IoT) devices, generating 73.1 zettabytes (ZB) of data [1].

Motivated by these demands, substantial breakthroughs must be achieved by the sixth generation (6G) communication systems [2], [3], including i) one terabit per second (Tbps) peak data rate to satisfy the data rate requirements of emerging applications and Augmented Reality (AR) / Virtual Reality (VR) communication, ii) higher spectral efficiency (SE) for massive connectivity, e.g., Internet of Things (IoT) devices that are ten times greater than that of 5G, with low latency and low cost, iii) hyper reliable and low-latency communication, e.g., in an industrial environment for full automation, control, and operation, iv) the support of distributed compute and

Wei Chen is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China (email:weich@bjtu.edu.cn).

Yuanwei Liu is with Queen Mary University of London, U.K (Email: yuanwei.liu@qmul.ac.uk).

Hamid Jafarkhani is with University of California, Irvine, US (Email: hamidj@uci.edu). His work was supported in part by the NSF Award CCF-2328075.

Yonina C. Eldar is with Weizmann Institute of Science, Israel (Email: yonina.eldar@weizmann.ac.il).

Peiying Zhu is with Huawei Technologies Co., Ltd, Canada (Email: peiying.zhu@huawei.com).

Khaled B Letaief is with The Hong Kong University of Science and Technology, China (Email: eekhaled@ust.hk).

AI-powered applications, which requires integrated AI and communication, v) the demand for multiple functions, e.g., sensing and navigation [4]. To date, wireless communication systems primarily rely on the orthogonality of resources to facilitate design and implementation, from user access to data transmission. The challenging new requirements in 6G call for more flexible design principles that go beyond orthogonality. Especially, breakthroughs in physical-layer random access and multiple-access technologies are critical, in order to provide effective support to all upper-layer services. It is desired to design new techniques beyond the frequency division multiple access (FDMA) in the first generation (1G), time division multiple access (TDMA) in the second generation (2G), and orthogonal frequency division multiple access (OFDMA) in the fourth generation (4G) and 5G. To achieve high SE and high energy efficiency (EE) in the next-generation wireless networks, next generation multiple access (NGMA) technologies will likely deviate from orthogonality as the design principle.

In orthogonal multiple access (OMA) techniques, such as TDMA, FDMA, and OFDMA implemented in 1G through 5G systems, each resource block is allocated to a single user. In contrast to conventional multiple-access techniques, the key concept of NGMA is to intelligently accommodate multiple users in the allotted resource blocks, e.g., time slots, frequency bands, spreading codes and beams, in the most effective manner for different applications. By integrating with other advanced transmission concepts, the performance of multiple access can be further enhanced. Such advanced transmission concepts include RISs [5], [6], cell-free massive MIMO [7]–[12], full duplex relaying, heterogeneous networks, mmWave communications [13]–[15], and THz communications [16]. More specifically, by properly integrating NGMA with these advanced concepts, synergy effects can be realized, providing extra benefits for network performance. For example, RISs can provide additional channel paths to build stronger combined channels with apparent strength differences and also artificially re-align users' combined channels to obtain more gain [17]. RIS-assisted medium access control protocol is designed, analyzed and optimized in [18]–[20] to improve the system throughput for NGMA. In [21], a joint active and passive beamforming scheme is designed in distributed RIS-aided massive access multiple-input single-output systems with supporting non-orthogonal multiple access (NOMA) and OMA transmissions simultaneously. In cell-free massive MIMO, since access points (APs) serve all users on the network, colliding users located far away can also be detected by some

APs based on the distinction of the pilot powers received [22].

The goals of what people envision being possible within the communication bands are expanding. In particular, there is a goal to use communication signals to perform sensing (or at least assist in sensing). In this case, signal waveform and signal processing are normally jointly designed, e.g., to enable integrated sensing and communication (ISAC) and integrated navigation and communication (INAC). The superposition of the communication and sensing signals in ISAC is a kind of non-orthogonal resource sharing, which shares a similar idea with NOMA. The prominent features of NOMA in efficient interference management and flexible resource allocation inspire sensing interference cancellation for ISAC [23]–[25]. Moreover, our goal in communicating might not be to communicate raw data, but might be to communicate meaning. Recently, with the rapid development of machine learning and artificial intelligence, semantic communications have emerged as a hot research topic, which addresses the semantic- and effectiveness-level communication problems [26]–[31]. Efficient multiple access schemes designed for semantic communication are required to accommodate the new semantic users given the limited radio resources.

Massive random access (MRA) and NOMA[1] are key technologies in NGMA. Random access is a necessary process for establishing wireless links between terminals and networks, which is extremely challenging in the scenario of a large number of devices, i.e., MRA. On the other hand, by employing superposition coding [32] at the transmitter and successive interference cancellation (SIC) [36] at the receiver, NOMA allows more users to be served than conventional multiple access techniques. Advances in computing and digital signal processing technologies now enable us to deal with some complex and previously intractable problems arising in NGMA [37]–[41]. Recent research has shown improvements in connectivity [37]–[39], system capacity (and SE) [40], [42], [43] and reduction in overall latency [44], [45]. In particular, artificial intelligence and machine learning (AI / ML) have become one of the main flagship activities in the 3rd generation partnership project (3GPP), which includes a wider use of AI / ML for optimizing radio access networks (RAN) and air interfaces [46]–[48]. AI-empowered multiple access for spectrum sensing, protocol designs, and optimizations is reviewed in [49]. The mining, classification, recognition, prediction, and learning capabilities of advanced signal processing technology make wireless networks more intelligent, and it is also essential to study the implementative and interpretative nature of technology to improve the performance of the NGMA.

The organization of this article is as follows. Section II reviews the background of MRA and NOMA for NGMA, including both fundamental limits and practical schemes. We provide an overview of the current research contributions using advanced signal processing and machine learning for MRA

and NOMA in Section III and Section IV, respectively. The interplay between NGMA and other new next-generation technologies and scenarios, e.g., age of information (AoI), near-field communications (NFC), ISAC, simultaneously transmitting and reflecting surfaces (STARS) and semantic communications, is discussed in Section V. In Section VI, we highlight several implementation challenges in using learning-based methods for NGMA. Finally, Section VII concludes the paper. Table I provides a list of acronyms used in this paper.

II. Background of Massive Random Access and Non-Orthogonal Multiple Access

In this section, we introduce the background of MRA and NOMA, with the focus on the non-orthogonal use of wireless resources to enhance massive connectivity and data transmission.

A. Massive Random Access

Massive machine-type communication (MMTC) is one of the three major application scenarios in 5G. It is still of significant importance in Beyond 5G (B5G) and 6G, making it one of the most popular research directions in recent years. Conventional communication systems are mainly designed for human-type communication (HTC), with the goal of supporting the transmission of a large amount of data for a small number of users at high data rates. MMTC has distinct characteristics from existing HTCs, such as uplink dominant, massive user access, sporadic communication, small packet transmission, low communication rate and low cost devices. The indicator for access devices in 5G is 10^6 users per square kilometer, while the number of devices that need to be supported in 6G has been increased to 10^7 users per square kilometer [50]. If the number of users is large, there are issues related to pilot collision for reliable channel estimation. Some NOMA schemes have been extensively studied in 5G standards [51]. This is still a huge challenge for the design of next-generation communication systems, as the existing access mechanism for 5G MMTC is not sufficient to efficiently and reliably support the transmission of small data packets for massive devices. As shown in Fig. 1, the goal of MRA is to detect a few active devices from a large number of devices, and then NOMA can be used to support those multiple active devices. Due to the complexity of SIC in NOMA, it is advisable to divide users into groups with orthogonal resources, and apply NOMA for each group independently.

Various coordinated multiple access technologies, e.g., FDMA, TDMA, CDMA[2], OFDMA, space division multiple access (SDMA), and NOMA, require specific protocols to efficiently manage communication among users with access to the same systems, whose fundamental limits are usually characterized as the classical multiple-access channel (MAC) capacity. With an increasing number of randomly activated users, the overhead of coordinating active users would overwhelm the system. Uncoordinated multiple access allows a set of users to transmit over a common wireless medium opportunistically and independently, resulting in different theoretical analyses.

---

[1]In this paper we will use the acronym "NOMA" to denote the class of non-orthogonal communication techniques that employ a single transmission codebook per user and SIC at the receiver. Some other classes of non-orthogonal techniques use multiple codebooks for each user, including time-sharing schemes [32] and rate-splitting schemes [33]–[35].

[2]Here we mean the synchronous CDMA case.

TABLE I: LIST OF ACRONYMS

|			
AA-MF-SIC	Activity-Aware Low-Complexity Multiple Feedback Successive Interference Cancellation	MAP	Maximum A Posteriori
ADMM	Alternating Direction Multiplier Method	MC-NGMA	Multi-Concept-Oriented Next Generation Multiple Access
AI	Artificial Intelligence	MF-NGMA	Multi-Functional-Oriented Next Generation Multiple Access
AMP	Approximate Message Passing	MIMO	Multiple-Input Multiple-Output
ANOMA	Asynchronous Non-Orthogonal Multiple Access	ML	Machine Learning
AP	access point	MLE	Maximum Likelihood Estimation
AR	Augmented Reality	MMSE	Minimum Mean Squared Error
AUD	Active User Detection	MMTC	Massive Machine-Type Communication
AWGN	Additive White Gaussian Noise	MMV	Multiple Measurement Vector
BC	Broadcast Channel	mmWave	Millimeter-wave
BCS	Bayesian Compressive Sensing	MnAC	Many-Access Channel
BER	Bit Error Rate	MPA	Message Passing Algorithm
BS	Base Station	MRA	Massive Random Access
BOMP	Block Orthogonal Matching Pursuit	MRC	Maximum Ratio Combining
BSAMP-CP	Backward Sparsity Adaptive Matching Pursuit with Checking and Projecting	MS	Mode Switching
BSBL	block sparse Bayesian learning	M-SP	Modified Subspace Pursuit
CCS	Code Compressed sensing	MT-NGMA	Multi-Tool-Oriented Next Generation Multiple Access
CDMA	Code Division Multiple Access	NFC	Near-Field Communications
CD-NOMA	Code-Domain Non-Orthogonal Multiple Access	NGMA	Next Generation Multiple Access
CE	Channel Estimation	NN	Neural Network
CF mMIMO	Cell-Free massive Multiple Input Multiple Output	NOMA	Non-Orthogonal Multiple Access
C-NOMA	Cooperative Non-Orthogonal Multiple Access	OAMP	Orthogonal Approximate Message Passing
CRAN	Cloud Radio Access Network	OFDMA	Orthogonal Frequency Division Multiple Access
CS	Compressed Sensing	OLS	Orthogonal Least Square
CSA	Coded Slotted Aloha	OMA	Orthogonal Multiple Access
CSI	Channel State Information	OMP	Orthogonal Matching Pursuit
CS-MUD	Compressive Sensing based Multi-User Detection	PDMA	Pattern Division Multiple Access
CTSMA	Coded Tandem Spread Spectrum Multiple Access	PD-NOMA	Power-Domain Non-Orthogonal Multiple Access
DDPG	Deep Deterministic Policy Gradient	PtrNet	Pointer Network
DL	Deep Learning	PUPE	Per-User Probability of Error
DL	Downlink	RAN	Radio Access Networks
DNN	Deep Neural Network	RAR	Random Access Response
DoFs	Degrees-of-Freedom	RIS	Reconfigurable Intelligent Surface
DQN	Deep Q-Network	RL	Reinforcement Learning
DRL	Deep Reinforcement Learning	SBL	Sparse Bayesian Learning
EE	Energy Efficiency	SC	Spatial Coupling
ELAA	Extremely Large-scale Antenna Array	SCA	Successive Convex Approximation
ES	Energy Splitting	SCMA	Sparse Code Multiple Access
FDMA	Frequency Division Multiple Access	SDMA	Space Division Multiple Access
FISTA	Fast Iterative Shrinkage-Thresholding Algorithm	SE	Spectral Efficiency
FMCW	Frequency Modulated Continuous Wave	SI	Side Information
GGAMP	Gaussian Generalized Approximate Message Passing	SIC	Successive Interference Cancellation
GGSO	Group Gram-Schmidt Orthogonalization	SISD	Structured Iterative Support Detection
GOMP	Group Orthogonal Matching Pursuit	SISO	Single-Input Single-Output
HTC	Human-Type Communication	SMV	Single Measurement Vector
ICT	Information and Communication Technology	SNR	Signal-to-Noise Ratio
IDC	International Data Corporation	SSM	Structured Sparse Model
INAC	Integrated Navigation And Communication	STARS	Simultaneously Transmitting And Reflecting Surfaces
IoE	Internet of Everything	T&R	Transmission and Reflection
IoT	Internet of Things	TA-BSASP	Threshold Aided Block Sparsity Adaptive Subspace Pursuit
ISAC	Integrated Sensing And Communication	TBM	Tensor-Based Modulation
ISI	Inter-Symbol Interference	TDMA	Time Division Multiple Access
ISTA	Iterative Shrinkage-Thresholding Algorithm	THz	Terahertz
IUI	Inter-User Interference	TS	Time Switching
LDMA	Location-Division Multiple Access	UAV	Unmanned Aerial Vehicle
LDPC	Low-Density Parity-Check	UL	Uplink
LDS-CDMA	Low-Density Signature Code Division Multiple Access	USL	Unsupervised Learning
LIP	Linear Inverse Problem	VAE	Variational Autoencoder
LoS	Line of Sight	VAMP	Vector Approximate Message Passing
LSTM	Long Short-Term Memory	VR	Virtual Reality
MAC	Multiple-Access Channel	ZB	Zettabytes
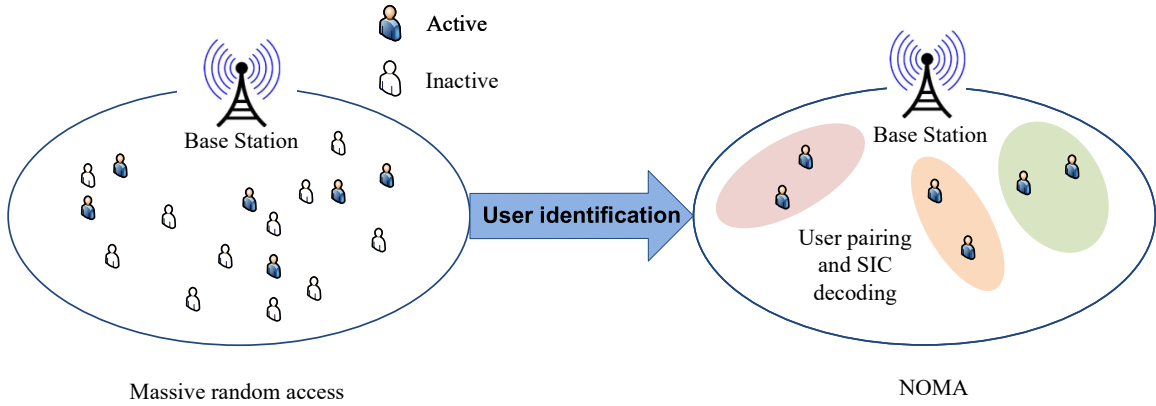
Fig. 1: Illustration of the MRA and NOMA scenarios.

1) Basic Principles: From the perspective of information theory, the fundamental limits of MRA have been characterized in various settings. The many-access channel (MnAC) is studied in [52], which provides the asymptotic regime where the total number of users n increases as the blocklength m tends to infinity. Given the stability requirement, the total number of users n increases with the frame length [53], [54]. In [55], the age of information for random access is derived, revealing its relation with the number of total nodes and their activation probability. Finite-blocklength achievability bounds for the random access channel under average-error and maximal-power constraints are presented in [56]. Non-asymptotic random coding achievability bounds with the criterion of per-user probability of error (PUPE) are derived [57], [58]. Under the PUPE criterion, non-asymptotic achievability and converse bounds on the minimum required energy per bit for MRA in MIMO quasi-static Rayleigh fading channels are established.

A fundamental problem in MRA is to detect the activity of users. Assume that there are in total n users sharing a wireless channel, and both the transmitter and the receiver have a single antenna. Each user can be assigned a unique dedicated codeword represented by an m-dimensional vector in $\mathbf{A}$, leading to a way of coordinated multiple access. Alternatively, user could randomly select codewords from the codebook $\mathbf{A}$ to transmit, leading to uncoordinated multiple access. The signal at the BS contributed by codeword $\mathbf{a}_j$ can be modeled as $x_j \mathbf{a}_j$ where x_j is a complex scalar that represents the product of the transmitted symbol and channel gain. If the codeword is not transmitted, $x_j = 0$. The total received signal at the BS can be expressed as

$$\mathbf{y} = \sum_{j=1}^{n} x_j \mathbf{a}_j + \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{z}, \qquad (1)$$

where $\mathbf{z}$ represents noise and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ denotes the codebook. Note that this is a general mathematical model, and can be easily solved if $m \geq n$. In the case of users

with unique codewords, active users and their corresponding channel are obtained according to the estimated $\mathbf{x}$. For users randomly selecting codewords from the common codebook, some collisions may occur if multiple users select the same codeword, and further collision resolution techniques are required to correctly identify all access users. When $m \ll n$, i.e., the total number of observations is much smaller than the total number of codewords, it is not possible to solve for a unique $\mathbf{x}$ even if the noise term is equal to zero, as there is a large number of solutions satisfying equation (1). We are interested in this case, as it is desired to support a large number of potential users with limited resources. Since most codewords are inactive, a suitable sparsity constraint on $\mathbf{x}$ may rule out all solutions except the one that is expected. Therefore, the activity detection problem in MRA is equivalent to the sparse recovery problem in compressed sensing (CS), and the identities of active users are obtained according to the activated codewords. Furthermore, with some variation of assumptions, different problems in MRA boil down to the same sparse recovery problem in equation (1). For example,

- if all users transmit with the unit power, recovering $\mathbf{x}$ in equation (1) leads to joint activity detection and channel estimation;
- for the Gaussian MAC channel, the message of user j can be carried by x_j and recovering $\mathbf{x}$ decodes the messages of active users.

Searching for an $\mathbf{x}$ in equation (1), that has these sparsity properties, leads to the problem of minimizing the number of selected codewords that could have generated the received signal $\mathbf{y}$, which involves minimizing the "0-semi-norm" of $\mathbf{x}$. It is an ill-posed sparse linear inverse problem, which has a non-continuous objective function and is NP-hard [59]. An alternative approach is to employ a smoothed objective function, e.g., the ℓ_1 norm of $\mathbf{x}$, and the CS theory claims that $\mathbf{x}$ can be recovered with a very high possibility when the

length of the codeword satisfies

$$m = \mathcal{O}\left(s \log \frac{n}{s}\right), \qquad (2)$$

where s denotes the number of active users. The above results assume a single BS antenna. Multiple antennas at the BS can bring additional improvements in MRA [60], [61]. In addition to these CS-based methods, structured access protocols, e.g., coded tandem spread spectrum multiple access (CTSMA) [62], which solve the access problem from the coding perspective, have also attracted considerable attention.

From a signal processing standpoint, there are two primary approaches to improving the performance of MRA. One is to employ advanced signal processing techniques that extract and leverage various prior information or structures embedded in the signal. More details will be presented in Section III.A. Another way is to employ the power of data-driven methods, or namely, machine learning methods. Plenty of empirical evidence has shown that deep learning can significantly increase convergence speed and accuracy in solving optimization problems, and even exploit knowledge embedded in complex problems that is usually difficult to model. More research results along this line will be presented in Section III.B.

2) Grant-Free Random Access: Most of the wireless communication systems currently in operation use a grant-based random access mechanism. Each active device utilizes a preamble sequence to notify the BS that it has become active, after which the BS sends a response in the form of a grant for transmitting the device's data. Therefore, active devices must engage in multiple signaling interactions with the BS to establish connections and allocate specific resources for data transmission. However, due to the majority of small data packets transmitted in MMTC, the negligible signaling overhead in large data packet transmission accounts for a relatively large proportion in MMTC, which cannot be ignored and often leads to an excessive proportion of signaling overhead, resulting in low EE and SE for the transmission of small data packets. Furthermore, the cumbersome handshake process in grant-based random access inevitably leads to significant latency, which makes the grant-based mechanism unsuitable for low latency applications.

A promising technology to support MMTC is grant-free random access, where channel resources are accessed by active devices without obtaining permissions from the BS. The grant-base random access and grant-free random access schemes are illustrated in Fig. 2. Grant-free random access brings significant benefits in reducing the signaling overhead and access latency. It has the potential to be applied in many emerging scenarios. A comprehensive review of grant-free random access protocols and techniques in satellite-based IoT networks is showed in [63]. A typical grant-free random access scheme is to assign a unique codeword, also called preamble, pilot or signature in the literature, to each device. To support massive devices with limited resources, non-orthogonal codewords need to be employed. Active devices initiate access by sending their codewords, and the BS conducts activity detection via detecting these codewords from the received signal. Another type of grant-free random access is unsourced random access, also called uncoordinated multiple access, in which all devices share the same codebook to transmit messages directly and the BS does not have the identity information of the active device. The goal of unsourced random access is to decode the information transmitted by the devices. If the device needs to be identified, the identity information can be included in the transmitted message. The benefits brought by grant-free random access comes at an increased likelihood of collisions resulting from uncoordinated channel access and inter-user interference due to the non-orthogonal access resources. Therefore, novel reliability enhancement techniques are needed.

B. Non-Orthogonal Multiple Access

In OMA, a single wireless resource can be allocated to only one user, such as by frequency division or time division. However, with the development of 5G, B5G, and the arrival of 6G, there will be an explosive increase in data traffic and access demand and traditional OMA cannot meet a series of challenging indicators, such as connectivity density and spectral efficiency. NOMA is proposed to improve spectrum efficiency and access capacity and may encompass a variety of use cases or deployment scenarios [51]. Compared to OMA, different NOMA users are multiplexed without the constraint of strict orthogonal radio resource allocation. With the help of optimized resource allocation and advanced receivers, the overall throughput of the system can be greatly improved.

1) Basic Principles: The capacity region of the multiaccess fading uplink channel is characterized in [64], [65]. MolavianJazi and Laneman establish the second-order achievable rate regions for the Gaussian MAC in [66], and Yavas et al. give the finite-blocklength achievability bounds for the Gaussian MAC and random access channel (RAC) under average-error and maximal-power constraints in [56]. The achievable rate region of OMA is a subset of the capacity region for single-input single-output (SISO) Gaussian MACs and broadcast channels (BCs), since each user can only use a portion of the resource blocks in OMA [32]. NOMA can achieve reliable communication at Synchronous a larger set of rates in the SISO cases. Consider synchronous power-domain NOMA (PD-NOMA), where users' signals are superimposed via superposition coding and then receivers adopt SIC techniques for interference cancellation. Fig. 3 illustrates the downlink and uplink signal transmission and processing of PD-NOMA. For downlink transmission with two users, the transmitted signal of the BS using superposition coding can be expressed as

$$x = \sqrt{p_1}s_1 + \sqrt{p_2}s_2, \qquad (3)$$

where s_1 and s_2 are the transmitted signals for the two users, and p_1 and p_2 are the corresponding transmit power. Let h_1 and h_2 denote the channel coefficients of User 1 and User 2, respectively, and σ^2 denote the noise power. Under the assumption of $|h_1|^2 > |h_2|^2$, User 1 first decodes the signal of User 2 and subtracts it from the received signal through SIC and then decodes its intended signal. On the other hand, User 2 directly decodes its intended signal by treating User

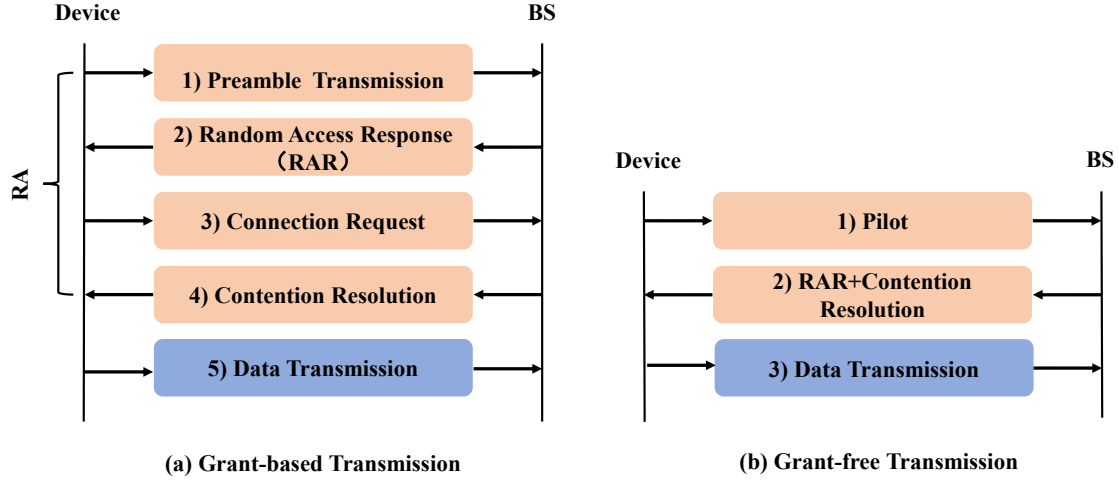**(a) Grant-based Transmission**　　　　　**(b) Grant-free Transmission**

Fig. 2: Grant-base random access and grant-free random access schemes.
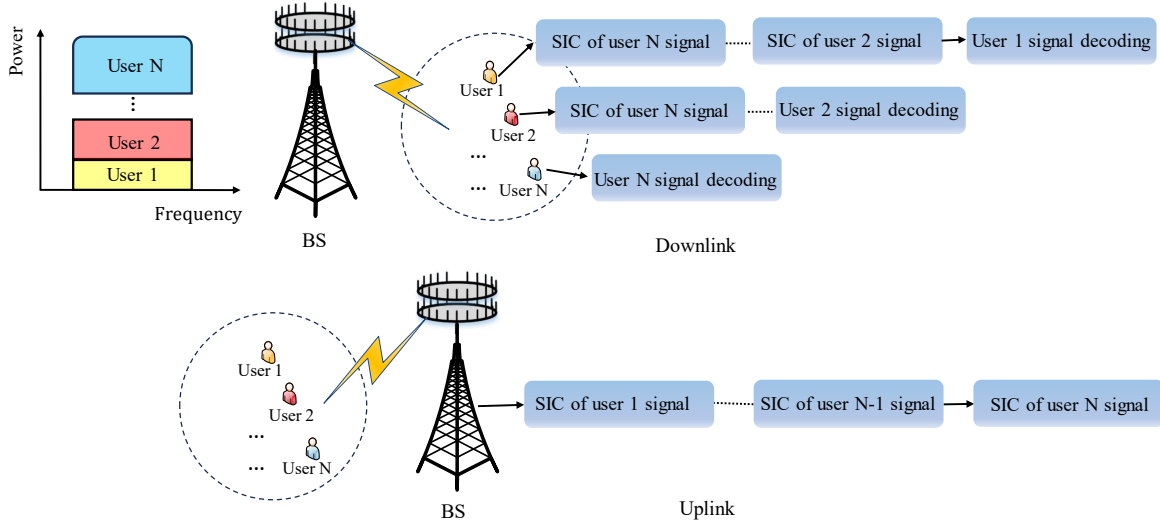

Fig. 3: An illustration of downlink and uplink PD-NOMA.

1's signal as noise. Since User 1's channel is stronger, User 1's achievable rate is larger than that of User 2 and User 1 can decode User 2's signal and cancel it. Therefore, the achievable rates of User 1 and User 2 are given by

$$R_1 = \log_2\left(1 + \frac{p_1|h_1|^2}{\sigma^2}\right) \tag{4}$$

and

$$R_2 = \log_2\left(1 + \frac{p_2|h_2|^2}{p_1|h_2|^2 + \sigma^2}\right), \tag{5}$$

respectively. For synchronous uplink NOMA transmission, the two users send their signals to the BS. At the BS, users 1' signal which experiences stronger channel is first decoded and then users 2' signal. In both the downlink and the uplink, OMA

is only able to achieve some specific points on the unicast capacity region, e.g., the points where only one user is active, while NOMA is more flexible [67]. For example, NOMA with joint decoding achieves the "corner points" of the multiple access region in the two-user single-antenna uplink case, and points on the "dominant face" can be achieved by time sharing between the two "NOMA with sequential decoding" schemes or rate splitting [68]. For the Gaussian single antenna downlink channel, the standard superposition coding approach is indeed a NOMA scheme, and can achieve an arbitrary point on the boundary of the unicast capacity region. For the multiple antenna case, the channel is not degraded, and one needs a more sophisticated form of transmission, e.g., dirty-paper coding [69].

Although NOMA has advantages such as high bandwidth

efficiency and connectivity, there are still many challenges to be tackled. As mentioned above, SIC is the key technology for demodulating superimposed signals. However, if errors occur during the demodulation of weak user's signals in the first stage, it will subtract the wrong modulated symbol from the received superimposed signal and cause a decrease in the performance of the demodulation in the second stage. Accurate channel state information (CSI) is crucial in NOMA. Error in estimating the CSI may lead to user grouping, ordering, and decoding errors. The increasing number of connected users and larger scale antennas in the next generation wireless networks will make the acquisition of CSI more challenging. Another challenge in NOMA is the security issue. Existing research on secure communication to avoid untrusted users and external eavesdroppers usually assumes the availability of full, partial, or statistical CSI. However, it is difficult to obtain CSI in the case of passive eavesdropping. Some interesting techniques, e.g., multi-dimensional directional modulation, cyclic feature suppression-based techniques, and channel-based inter-leaving may be potential solutions for enhancing security [70].

2) Asynchronous NOMA: Because of the distributed nature of multi-user networks and the effects of multipath and propagation delays, signals from different users experience different time delays to the receiver [71]. Perfect symbol synchronization requires feedback and coordination, which complicates the system greatly. Having multiple antennas at the receiver makes it even more complicated. Assuming a multiple antenna receiver in uplink NOMA, one can only synchronize the symbols of the received signals perfectly at one of the receive antennas and the rest of antennas may experience imperfect timing synchronization among received signals. Even if such a complete symbol synchronization is possible, it is not clear if it is desirable. In fact, the MAC's capacity-region for a system with time asynchrony and rectangular waveforms in CDMA is larger than that of a perfectly synchronized system [72]. Also, intentionally adding symbol-level timing offsets to make the signals asynchronous has shown advantages in managing interference in MIMO [71], [73], [74] and in differential decoding for MACs [75], [76]. The possible advantages along with the difficulties in achieving perfect synchronization motivate a thorough analysis and design of NOMA systems under imperfect timing synchronization.

For uplink asynchronous NOMA (ANOMA), the capacity-region for two users utilizing different pulse shaping methods has been derived in [77]. The analysis shows that the asynchronous transmission enlarges the capacity-region. In fact, not only does asynchronous transmission, i.e., timing mismatch, improve the performance of NOMA systems, but also SIC is not optimal for asynchronous NOMA [77]. Let us assume User k transmits $\sqrt{P_k}s_k[n]$, where $s_k[n]$ denotes the nth normalized transmitted symbol and P_k denotes the transmit power. Then, under imperfect timing synchronization among different users, the received signal at the BS is given by

$$y(t) = \sum_{n=1}^{N}\sum_{k=1}^{K} h_k \sqrt{P_k}s_k[n]p(t - nT - \tau_k T) + \eta(t), \quad (6)$$

where N represents the frame length, K is the number

of users, T is the symbol interval, τ_k denotes the relative time delay at User k, $p(.)$ is the pulse-shaping filter, and $\eta(t)$ denotes the normalized additive white Gaussian noise (AWGN). The set of sufficient statistics can be found by proper filtering at the receiver and over-sampling K times, each time synched with one of the users [71]. The resulting input-output model is

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{H}\boldsymbol{s} + \boldsymbol{\eta}, \quad (7)$$

where $\boldsymbol{y}, \boldsymbol{R}, \boldsymbol{H}, \boldsymbol{s}$, and $\boldsymbol{n}$ represent the set of samples, the timing offsets matrix, the effective channel matrix, the transmited symbols, and the noise vector, respectively. Note that the perfect synchronization results in the conventional system model of $y[n] = \sum_{k=1}^{K} h_k \sqrt{P_k}s_k[n] + n[n]$.

In a perfectly symbol-synchronized NOMA system, at each time instant, only the inter-user interference (IUI) degrades the performance. In an asynchronous NOMA, not only IUI but also inter-symbol interference (ISI) degrades the performance. Therefore, the conventional SIC is not optimal anymore and efficient joint detection methods may be needed. In addition, because of the timing asynchrony, sufficient statistics results in over-sampling and the corresponding sampling diversity [71] can improve the overall performance. While the conventional wisdom suggests that the imperfect timing increases the overall interference and the overall performance is degraded, surprisingly, imperfect timing in fact decreases the overall interference, because of the benefits of sampling diversity [77], as the reduction in IUI outweighs the addition of ISI [78].

For downlink asynchronous NOMA, the superposition is performed at the transmitter and the transmitted signal is received by the intended users [79], [80]. Adding intentional time delays at the transmitter can in fact improve the performance. The transmitted signal, including the added intentional time delays, can be written as $s(t) = \sum_{n=1}^{N}\sum_{k=1}^{K}\sqrt{P_k}s_k[n]p(t - nT - \tau_k T)$, and the received signal by User k is represented as

$$y_k(t) = h_k s(t) + \eta_k(t). \quad (8)$$

Similar to the case of uplink ANOMA, the set of sufficient statistics can be found by proper filtering and over-sampling as

$$\boldsymbol{y_k} = h_k \boldsymbol{R_k}\boldsymbol{s} + \boldsymbol{\eta_k}, \quad (9)$$

where $\boldsymbol{y_k}, \boldsymbol{R_k}, h_k$, and $\boldsymbol{\eta_k}$ represent the set of samples, the timing offsets matrix, the effective channel coefficient, and the noise vector at User k, respectively. Note that the perfect synchronization results in the conventional system model of $y_k[n] = h_k \sum_{j=1}^{K}\sqrt{P_j}s_j[n] + n_k[n]$. Note that any User j with $|h_j| < |h_k|$ will have a smaller capacity compared with that of User k. Therefore, its transmission rate will be chosen lower than the capacity of User k and from an information theoretical point of view, User K can decode the signal of a weaker user with no error. Using SIC, each user decodes all signals from weaker users and removes them from the received signal. Then, it considers the remaining interference from the stronger users as noise and decodes its own message. Downlink ANOMA has to deal with similar imperfect timing

issues discussed for uplink ANOMA; however, it can perform coordinated resource management. For instance, power allocation can be performed globally with collective CSI from all users. More importantly, because of the virtual-MIMO nature of the system model in (9), enhanced joint processing can be performed at the transmitter to exploit the virtual-MIMO structure. The joint processing at the transmitter can be realized by intelligent correlated precoding and circumvents the need for SIC, as done in referece [80]. The idea of adding intentional time delays have been extended to intentionally adding frequency offsets in NOMA multi-carrier and OFDM systems [81], [82].

III. MASSIVE RANDOM ACCESS

In this section, we introduce model-based methods, using advanced signal processing, and learning empowered methods for MRA.

A. Model-Based Methods for MRA

For active user detection in MRA, model- methods are classified into sparse recovery based methods, structured sparse recovery based methods, and low-rank matrix/tensor recovery based methods, as shown in Fig. 4.

1) Sparse Recovery Based Methods for MRA: As discussed in the previous section, the grant-free access scheme is attractive in MMTC, where active devices directly transmit their pilot and data to the BS without grant. Simultaneously conducting accurate active user detection (AUD) and CE with non-orthogonal pilots is critically important and challenging for device to successfully establish reliable communication with the BS. Fortunately, the sporadic nature of MMTC traffic results in sparse user activity patterns. By using sparsity, the non-orthogonal pilot detection and CE problem can be modeled as a sparse linear inverse problem (LIP), thus reminiscent of the application of CS. By solving the LIP, not only do we realize highly reliable AUD and accurate CE, but also we alleviate the multi-user interference problem.

CS Based Typical Methods: Various CS based algorithms could be used to solve the AUD and CE problem, e.g., iterative shrinkage-thresholding algorithm (ISTA) [83], fast ISTA (FISTA) [84], orthogonal matching pursuit (OMP) [85], message passing [86], and sparse Bayesian learning (SBL) [87], [88].

- *OMP*: Schepker and Dekorsy used OMP [89] and orthogonal least squares (OLS) for MUD, which iteratively selected the most probable active users and subsequently estimated their data [90]. Group orthogonal matching pursuit (GOMP) was further exploited to deal with the asynchronicity case, where chips are not received at the same time [91]. As CS based methods have high computational complexity, most literature considers $\leq 10^4$ devices in their experiments. To achieve good scalability, a contention-based MRA scheme was proposed in [92] to support massive access, e.g., $\geq 10^6$, within a certain time-frequency resource, and a one-step access scheme was adopted to save energy and spectrum resources.

- *AMP*: The approximate message passing algorithm (AMP) [86] is attractive in AUD and CE owing to its high computational efficiency. Reference [93] considered two cases depending on whether the large-scale component of channel fading is known in the BS and designed the minimum mean squared error (MMSE) denoiser for AMP according to the channel statistics. It also provided an analytical characterization of the probabilities of false alarm and missed detection through state evolution. Although AMP is an iterative algorithm with high computational efficiency, it is sensitive to the distribution of the pilot. Small deviations from the independent and identically distributed (i.i.d.) sub-Gaussian random variable model of the pilot can cause the algorithm to diverge. To address this issue, reference [94] developed a vector AMP algorithm (VAMP) and proved that VAMP has strict scalar state evolution, which is applicable to a wider class of pilot distributions.

- *SBL*: SBL is a popular approach for single sparse signal recovery from a Bayesian perspective, and has empirically shown superior recovery performance [95], [96]. Gaussian generalized approximate message passing (GGAMP) is incorporated into SBL in [87], and the approach is further extended from the single measurement vector (SMV) to the multiple measurement vectors (MMV). Numerical experiments show that the proposed algorithm has great robustness and computational advantages. Moreover, in [88] considered a general joint-sparse model and provided a Bayesian approach that extends the SBL for the joint-sparse model.

- *Other CS-based methods*: Taking advantage of the total squared coherence of pilot sequences and convex relaxation, a novel pilot design approach was proposed for joint AUD and CE in [97]. The proposed design exhibits superior performance in relation to the random pilot design and the design based on tight frames [98]. According to CS theory, the overhead of CS-based methods highly depends on the sparsity and the number of devices. A more sparse representation of the active device's channel can be obtained by using dictionary learning. In view of the distinct wireless propagation environments in different cells, a dictionary-learning-enhanced AUD and CE method for massive MIMO systems was proposed in [99], which uses a dictionary learned from the historical channel information of the specific cell. A dictionary learning-based uncoordinated access method was proposed to jointly identify the set of active devices and detect data symbols embedded in signals transmitted by active devices in [100].

CS-Based Methods Using Additional Information: Additional information has been used for AUD and CE to achieve better performance. Various additional information could be exploited, including constellation information, active user temporal correlation, channel spatial correlation, and so on.

- *Use prior information:* Devices might have different service requirements, which in turn leads to different data
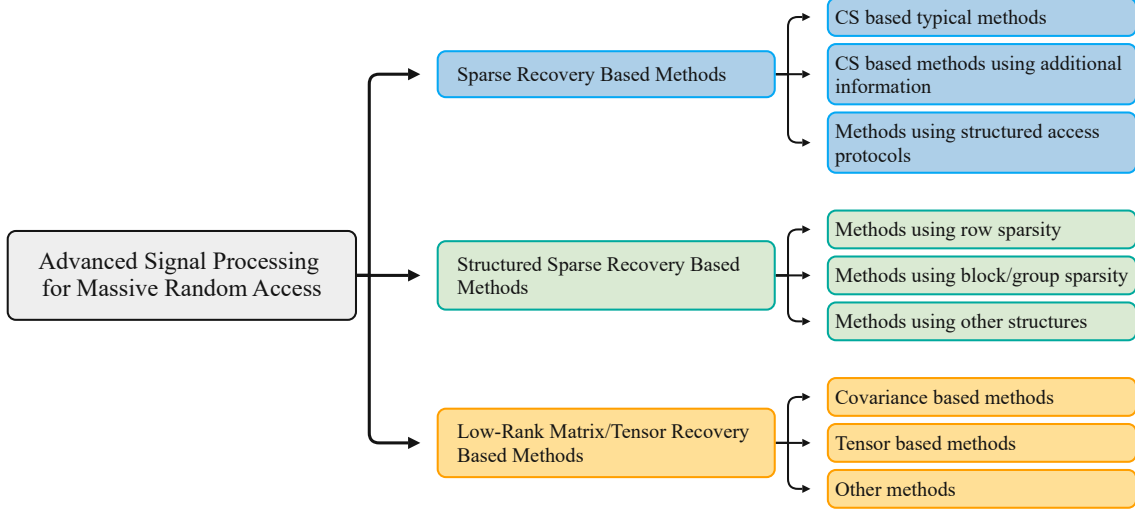
Fig. 4: Advanced signal processing for MRA.

lengths for transmission. The data length diversity was exploited in [101] to enhance AUD and CE. In [102], different prior information was used simultaneously to improve the performance of MRA. In specific, the joint sparsity between the channel and the data was used to improve the accuracy of pilot detection, and the modulation information and the cyclic redundancy check were used for channel correction to improve the performance of data recovery. Reference [103] improved the performance of AUD and CE using side information (SI) brought from data recovery, and improved the performance of data recovery in SIC using the SI brought from AUD.

- *Use temporal correlation:* Modified subspace pursuit (M-SP) was proposed to exploit a prior support and the prior support quality information in massive MIMO CE [104]. Instead of blindly using the prior support, the proposed approach utilizes it adaptively by taking the quality information into account. Taking advantage of the temporal correlation of the channel, the required pilot overhead was further reduced. Moreover, they proved that an appropriate prior support can lead to better CS recovery performance and further propose some robustness designs to combat incorrect prior support quality information. The channel response estimated in the previous slots was used as SI when estimating the channel response for the current slots in [105]. An algorithm called AMP-SI was proposed to embed the SI into the denoiser design under the SMV-AMP framework for single-antenna systems. In [106], it was proposed to leverage the SI of the temporal correlation in the activity of the device, that is, the explicit quantification of the correlation between the activity mode estimated by the AMP algorithm in the former coherent block and the actual activity mode in the current coherent block. The SI was applied to the AMP denoiser design, and a new SI-aided AMP framework

was derived. Numerical results showed that the proposed method can significantly improve the performance of AUD and CE.

- *Use channel spatial correlation:* An alternating direction multiplier method (ADMM) using prior information on second-order statistics, that is, spatial channel covariance of each user channel, was proposed to improve performance in [107]. In specific, the original problem of AUD and CE can be formulated as:

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{SX} - \mathbf{Y}\|_{\mathrm{F}}^2 + \beta_1 \|\mathbf{X}\|_{2,1}, \qquad (10)$$

where $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{X}$ denote the received signal matrix, the pilot matrix/sensing matrix and the effective channel matrix, respectively. $\|\cdot\|_F$ represents the Frobenius norm. $\|\cdot\|_{2,1}$ is a matrix norm that sums up the ℓ_2 norm of all rows in the matrix. β_1 denotes a trade-off between the fidelity term and the row-sparsity level of $\mathbf{X}$. By using the information embedded in the estimated channel covariance matrix, the problem can be reformulated as:

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{SX} - \mathbf{Y}\|_{\mathrm{F}}^2 + \beta_1 \|\mathbf{X}\|_{2,1}$$
$$+ \beta_2 \sum_{i=1}^{N} \mathbb{I}(\mathbf{x}_i) \left\| \mathbf{x}_i \mathbf{x}_i^{\mathrm{H}} - \tilde{\mathbf{R}}_i \right\|_{\mathrm{F}}^2, \qquad (11)$$

where $\tilde{\mathbf{R}}_i$ is the estimated scaled covariance matrix, $\mathbf{x}_i \mathbf{x}_i^{\mathrm{H}}$ is the sample covariance matrix, and β_1 and β_2 are parameters that control the trade-off. $\mathbb{I}(\mathbf{x}_i)$ is defined as follows:

$$\mathbb{I}(\mathbf{x}_i) = \left\{ \begin{array}{ll} 1, & \|\mathbf{x}_i\|_2 > 0, \\ 0, & \|\mathbf{x}_i\|_2 = 0, \end{array} \right. \quad \forall i \in \mathcal{N} \qquad (12)$$

where $\mathcal{N} = \{1, \ldots, N\}$. The extra channel space correlation in (11) improves the AUD and CE performance with lower signaling overhead.

- *Using constellation information:* Constellation information has been used to improve AUD and CE performance in literature. In [108], a low complexity detection algorithm for MMTC is proposed, namely, activity-aware low complexity multiple feedback successive interference cancelation (AA-MF-SIC). The main idea of this method is to evaluate the reliability of each estimated symbol. If the current estimate fails to satisfy some given constraint, other possible constellation points are then considered with the cost of increased computational complexity. In [101], backward sparsity adaptive matching pursuit with checking and projecting (BSAMP-CP) was proposed to jointly carry out sparsity level estimation, AUD, CE and data recovery in two phases. Specifically, in the first phase, it conducted the sparsity level estimation in a backward manner, exploiting the data length diversity of active users. In the second phase, it jointly conducted AUD, CE and data recovery, taking the joint sparsity information of pilot and data symbols, the error checking information and the modulation constellation information into account. Using the prior information, the problem formulated in [101] can be expressed as:

$$\min_{\mathbf{H},\mathbf{D}} \|\mathbf{Y} - \mathbf{S}\mathbf{X}\|_{\mathrm{F}} \tag{13a}$$

$$\text{s.t. } \|\mathbf{X}\|_{\mathrm{row},0} \leq K, \tag{13b}$$

$$\mathbf{X} = \mathbf{H}[\mathbf{P}, \mathbf{D}], \tag{13c}$$

$$\forall k = 1 \dots N, \tag{13d}$$

$$\mathrm{g}\left(\mathbf{D}_{(\{k\},:)}\right) = 1, \tag{13e}$$

$$\mathbf{D}_{(\{k\},:)} \in (\mathcal{A} \cup \mathcal{O})^{n_d}, \tag{13f}$$

where $\| \cdot \|_{\mathrm{row},0}$ represents $\ell_{\mathrm{row},0}$-semi norm that outputs the number of non-zero rows of the input matrix, and K denotes the number of active users. $\mathbf{H}$ denotes the channel coefficients of all devices, and $\mathbf{P}$ and $\mathbf{D}$ denote the pilot matrix and data matrix of device k, respectively. The constraints (13b) and (13c) describe the joint sparse characteristics of pilot and data. The constraint (13e) describes the checking mechanism when data recovery is implemented, where the function $\mathrm{g}(\cdot)$ denotes the error checking procedure. The constraint (13f) uses constellation point information, where $\mathcal{A}$, $\mathcal{O}$ and n_d denote the transmitted symbol set, the set of zeros and the maximum number of transmitted data symbols, respectively. By exploiting backward activity level estimation that exploits data length diversity, this method achieved superior performance.

Methods Using Structured Access Protocols: Reference [109] exploited coded slotted aloha (CSA) and proposed the paradigm of coded random access, in which the structure of the access protocol can be mapped to a structure of an erasure-correcting code defined on a graph. This provided the possibility of designing efficient random access protocols using coding theory. A novel grant-free NOMA method called CTSMA was proposed in [62]. This method divides each user's data packet into multiple segments for processing, so that conflicts occur only on segments using the same extension code, which can be resolved with redundant segments. For the

MMTC scenario in industrial IoT, a multislot design scheme was proposed with joint CTSMA and CSA to further promote collision resolution capability [110]. Performance analysis has shown the superiority of multislot design in the short uplink contention period. A coded CS (CCS) was proposed in [111], in which each active device divides its data into several subblocks and then adds redundancy using system linear block codes. CS was used to restore the order of the subblocks and splice them together to obtain the original information. A pilot-based coherent scheme was proposed in [112] for unsourced random access in the MIMO scenario, where the user messages are divided into two parts. Specifically, several bits are used to select a shorter code word from the codebook, and then encode the remaining message bits. The receiver uses the MMV-AMP algorithm to estimate the channel from the pilot part, and then decodes the second part with the maximum ratio combining (MRC). A signal scrambling based joint blind channel estimation, activity detection, and data decoding (SS-JCAD) scheme is proposed for coded massive random access in [113]. Signal scrambling technique is used and treated as the user-specific signature. A simple yet efficient receiver is designed which integrates the blind CSI estimation module with the forward error correction decode to achieve joint blind channel estimation, activity detection, and data decoding. The authors of [114] propose a high-efficiency massive uncoupled unsourced random access scheme for 6G wireless networks without requiring extra parity check bits. A low-complexity Bayesian joint decoding algorithm was designed to implement codeword detection and stitching based on channel statistical information.

A summary of existing work on sparse recovery based methods for MRA is given in Table II.

2) Structured Sparse Recovery-Based Methods for MRA: Unlike traditional sparse recovery methods that focus solely on identifying individual sparse elements, structured sparse recovery goes a step further by exploiting the inherent structures or patterns presented in the signal. Structured sparse recovery-based methods represent a compelling alternative that harnesses the power of structures to unlock deeper insights and achieve superior performance for MRA. Moreover, it enhances the efficiency of the recovery algorithm, reduces computational complexity, and enables scalable solutions for MRA problems.

Methods Using Row Sparsity: For the BS equipped with a single antenna, the joint AUD and CE can be expressed as an SMV problem in CS. In the case of the BS with multiple antennas, the joint AUD and CE turns into an MMV problem. The user activity pattern is the same for all signals received by different BS antennas, which results in a row sparsity structure in the joint AUD and CE problem. In [60], [93], [115], AMP algorithms were applied to the multi-antenna scenario. AMP with vector denoiser and parallel AMP-MMV were used for joint AUD and CE in [93], with simulation results that demonstrate a significant performance improvement caused by the deployment of multiple antennas in BS. According to the theoretical analysis in [60], [115], when the number of BS antennas tends to infinity, perfect AUD can be achieved, which indicates that the sparse signal recovery-based method is suitable for the MRA problem. Reference

TABLE II: Summary of Existing Work on Sparse Recovery Based Methods for MRA

Category	Algorithm	Characteristics	Ref.
CS-based typical methods	ISTA,FISTA	Classic method for solving LIP	[83], [84]
	SBL	Provide an efficient algorithm that extends SBL for a joint-sparse model	[88]
	OMP	Using OMP algorithm for sparse signal recovery	[89]
	OMP	Use OMP and orthogonal least squares (OLS) for MUD	[90]
	GOMP	Use GOMP to deal with asynchronous case	[91]
	OMP	Propose a contention-based MRA scheme to support massive access	[92]
	AMP	Use AMP algorithm for AUD and CE and provides analytical characterization	[93]
	VAMP	VAMP has strict state evolution and stronger robustness compared with AMP	[94]
	Pilot design algorithm	Propose a novel pilot design approach for joint AUD and CE	[97]
	Dictionary learning	Use dictionary learning to enhance AUD and CE in massive MIMO system	[99]
	Dictionary learning	Propose dictionary learning-based uncoordinated access method to reduce the pilot overhead	[100]
CS-based methods of using additional information	M-OMP	Use various prior information to enhance performance	[102]
	FSP-RIC	Use SI brought from the data recovery	[103]
	M-SP	Use channel time correlation to reduce pilot overhead	[104]
	AMP-SI	Use channel response in the previous slots as SI to estimate the cerrent channel response	[105]
	AMP	Use the temporal correlation in device activity as SI and apply SI into AMP denoiser design	[106]
	ADMM	Utilize prior information on the second-order statistics to improve performance	[107]
	AA-MF-SIC	Use constellation point information to evaluate the reliability of the estimation	[108]
	CSAMP-CP	Divide AUD, CE and data recovery into two phases	[101]
Methods using structured access protocols	CSA	Propose the paradigm of coded random access	[109]
	CTSMA	Divide data packet into multiple segments for processing	[62]
	CTSMA	Propose a scheme joint CTSMA and CSA to promote the collision resolution capability	[110]
	CCS	Divide the data into subblocks and use algorithms to recover and splice it	[111]
	MMV-AMP	Split the user messages into two parts and use MMV-AMP and MRC for decoding	[112]
	SS-JCAD	Use the scrambling pattern as a user-specific signature	[113]
	OAMP	Implement codeword detection in the case of unknown priors and codeword stitching with the assistance of channel statistics	[114]

[116] considered joint AUD and CE in uplink cloud radio access network (CRAN) systems with few active users and proposed a modified Bayesian compressive sensing algorithm (BCS), which exploits not only the sparsity of the active users, but also the prior channel information of path loss effects and the joint sparsity structures. Moreover, in [117], Cheng et al. proposed an MMV-based orthogonal AMP (OAMP), and derived the group Gram-Schmidt orthogonalization (GGSO) procedure to implement OAMP. It has been shown that OAMP is superior to AMP when pilot sequences are generated using the Hadamard pilot matrix. Most previous work considered narrow-band MRA scenarios assuming single-carrier transmission. Reference [118] investigated grant-free MRA in massive MIMO-OFDM systems and proposed a Turbo-GMMV-AMP algorithm running in an alternating manner for joint AUD and CE, which jointly exploits the space-frequency-angular-domain sparsity of the MRA channel matrix to enhance performance. In [119], Chen et al. proposed a contention-based joint AUD and CE method for massive MIMO systems with angular domain enhancement. The problems of AUD and CE were presented as LIP with simultaneous row sparse and cluster sparse structures, in which angle of arrival (AOA) information was used to enhance AUD and CE. They used SBL to formulate the optimization problem, which is given as:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{S}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x})$$
$$\text{s.t. } f(\mathbf{x}) = \min_{\boldsymbol{\omega} \geq 0} \mathbf{x}^T \boldsymbol{\Omega} \mathbf{x} + \log |\boldsymbol{\Sigma}|, \tag{14}$$

where λ denotes the variance of Gaussian noise, $\boldsymbol{\Omega} \triangleq \text{diag}(\omega_1, \ldots, \omega_N) \in \mathbb{R}^{N \times N}$, $\boldsymbol{\omega}$ is a vector of hyperparameters governing the prior variance of the elements in signal and $\boldsymbol{\Sigma} = \lambda \mathbf{I} + \mathbf{S}\boldsymbol{\Omega}\mathbf{S}^T$. By solving the bilevel optimization problem in (14), multiple active users can be detected successfully, together with accurate CE, even when they use the same time-frequency-code resource for random access. For bilevel optimization, interested readers may refer to [120].

Methods Using Block/Group Sparsity: Block/group sparsity generalizes the standard sparse signal model by separating the elements into a number of blocks/groups. By combining CS based multi-user detection (CS-MUD) with multi-carrier access schemes, a new solution was proposed to leverage the group sparsity structure through the GOMP [121]. In [122], the joint AUD and CE in grant-free NOMA systems was formulated as a block CS-based sparse signal recovery problem. The authors made explicit use of the block sparsity inherent in the equivalent block-sparse model and proposed two enhanced block CS-based greedy algorithms, namely, threshold aided block sparsity adaptive subspace pursuit (TA-

BSASP) and cross-validation aided block sparsity adaptive subspace pursuit (CVA-BSASP). In [123], the authors proposed block orthogonal matching pursuit (BOMP) algorithm for AUD and CE. The algorithm transforms the original model, and then exploit the block sparsity of matrix $\mathbf{X}$. Based on the maximal Euclidean norm, the objective function in the l-th iteration of the algorithm was given as:

$$\mathcal{I}_l = \underset{\mathcal{I}_B \in \mathcal{I}_{\text{Block}}}{\arg\max} \left\| \overline{\mathbf{P}}_{[\mathcal{I}_B]}^H \mathbf{r}_{l-1} \right\|_2, \tag{15}$$

where $\mathcal{I}_l$ is the selected index set, $\mathcal{I}_{\text{Block}}$ is the union of all block indices, $\overline{\mathbf{P}}$ is obtained from $\mathbf{X}$ through a series of transformations and $\mathbf{r}_{l-1}$ is the residue of the $(l-1)$-th iteration. Furthermore, a block sparse Bayesian learning (BSBL) algorithm was proposed, which outperformed BOMP in terms of computational complexity and recovery performance. As AMP is sensitive to the choice of pilot sequence while having low complexity, AMP-BSBL algorithm was proposed in [124], which retains the advantages of AMP and is robust to the choice of pilot sequence. In [125], [126], a goal-oriented detection approach is proposed to exploit the continuous angular group sparsity feature of massive MIMO wireless channels. For stationary users with known angular information, their activity can be identified according to the active angular signals without reconstructing the corresponding channels and messages.

Methods Using Other Structures: There are methods that use other structures to reduce the degree of freedom in the problem. A low-complexity multiuser detector based on structured CS, namely, the structured iterative support detection (SISD) algorithm, utilized the inherent structural sparsity of user activities in a NOMA system to realize joint AUD and CE [127]. Simultaneous SBL for joint sparse approximation with two structured sparse models (SSMs) were proposed in [128], where one is row-sparse with embedded element-sparse and the other one is row-sparse plus element-sparse, which can be extended for AUD and CE in some special cases. Taking advantage of a more accurate prior distribution to characterize the structured sparsity of access signals, efficient joint activity and data detection algorithms, namely, the OAMP MMV algorithm with simplified structure learning and accurate structure learning, were proposed in [129]. Furthermore, it can be extended to SIC-based detection with channel coding to achieve highly reliable random access. An expectation-maximization-aided generalized AMP algorithm with a Markov random field support structure, namely, EM-MRF-GAMP, was proposed in [130], which captures the inherent clustered sparsity structure of the angular domain channel. Then, message reconstruction in the form of a clustering decoder was performed by recognizing slot-distributed channels of each active user based on similarity. The proposed scheme achieved better error performance compared to traditional CS-based random access schemes.

3) Low-Rank Matrix/Tensor Recovery-Based Methods for MRA: Low-rank matrix/tensor recovery-based methods, leveraging the inherent structure and sparsity present in the channel, offer a highly effective way to address the challenges in MRA. In [131]–[135], by exploiting the low rank and statistical char-

acteristics of the signal, covariance-based methods for grant-free MRA were used in AUD. The sparse activity detection problem was formulated as a large scale fading estimation problem. Let $a_n \in \{1, 0\}$ denote the activity of device n, $\mathbf{h}_n$ denote the channel vector between device n and the BS, where $\mathbf{h}_n \in \mathbb{C}^{M \times 1}$ include both Rayleigh fading component and large scale fading component, M denotes the number of antennas and L is the pilot length. The received signal $\mathbf{Y} \in \mathbb{C}^{L \times M}$ at the BS can be expressed as

$$\mathbf{Y} = \sum_{n=1}^{N} a_n \mathbf{s}_n \mathbf{h}_n^T + \mathbf{Z} \triangleq \mathbf{S} \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{H} + \mathbf{Z}, \tag{16}$$

where $\mathbf{S} \triangleq [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N] \in \mathbb{C}^{L \times N}$, $\mathbf{\Gamma} \triangleq \text{diag}(\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^{N \times N}$ with $\gamma_n = (a_n g_n)^2$ is a diagonal matrix and $\mathbf{H} \triangleq \left[\mathbf{h}_1/\sqrt{g_1}, \mathbf{h}_2/\sqrt{g_2}, \cdots, \mathbf{h}_N/\sqrt{g_N} \right]^T \in \mathbb{C}^{N \times M}$ is the normalized channel matrix with g_n is the large scale fading component. The user activity detection can be formulated as a maximum likelihood estimation (MLE) problem. After normalization and simplification, the objective function can be written as follow:

$$\min_{\boldsymbol{\gamma}} \ \log |\mathbf{\Sigma}| + \text{tr}\left(\mathbf{\Sigma}^{-1} \hat{\mathbf{\Sigma}}\right)$$
$$\text{s.t.} \ \ \boldsymbol{\gamma} \geq 0, \tag{17}$$

where $\hat{\mathbf{\Sigma}} \triangleq \frac{1}{M} \mathbf{Y} \mathbf{Y}^H = \frac{1}{M} \sum_{m=1}^{M} \mathbf{y}_m \mathbf{y}_m^H$ is the average sample covariance matrix of the received signal on different antennas, $\mathbf{\Sigma} = \mathbf{S} \mathbf{\Gamma} \mathbf{S}^H + \sigma_w^2 \mathbf{I}$ is the covariance matrix, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^T$ is the indicator of activity and large-scale fading of all devices. The coordinate descent method can be used to find good solutions of this optimization problem. An effective active set algorithm was proposed following reference [132], [133] to reduce the computational complexity in [136]. Furthermore, in [132], [133] it was shown that covariance-based methods can significantly shorten the minimum pilot sequence length for AUD when CE is not required. In [137], [138], tensor-based modulation (TBM) was introduced to unsourced random access, where data decoding exploits low-rank tensor structure and tensor decomposition. A dimension reduction method was proposed to use the sparse and low-rank structure to project the original device state matrix into a low-dimension space in [139]. Then an optimized design framework with coupled full column rank constraints was developed to reduce the size of the search space, leading to a shorter pilot sequence than the original sparse recovery-based methods.

A summary of existing work on structured sparse recovery-based methods for MRA is given in Table III.

B. Machine Learning Empowered MRA

Deep learning (DL) is a promising approach to solving complex optimization problems. In particular, by unfolding an iterative algorithm into a neural network (NN), we can learn the parameters of iterative algorithms from training data, which differs from traditional algorithms that employ predetermined parameters. This approach is also called deep unfolding. Using DL for MRA has several advantages. One

TABLE III: Summary of Existing Work on Structured Sparse Recovery-Based Methods for MRA

| Category	Algorithm	Characteristics	Ref.
Methods using the row sparsity	AMP	Employ two different AMP algorithms, namely, the AMP with vector denoiser and the parallel AMP-MMV	[93]
	AMP	Use AMP algorithm to solve MMV problem, joint AUD, CE and data transmission	[60], [115]
	BCS	Joint sparsity structures in multi-antenna uplink CRAN systems	[116]
	OAMP	Propose the OAMP-MMV form, derive GGSO procedure to implement OAMP	[117]
	GMMV-AMP	Use angular domain sparsity to enhanced performance in MIMO-OFDM system	[118]
	SBL	Use simultaneous row sparse and cluster sparse structures to enhance AUD and CE in massive MIMO system	[119]
Methods using the block/group sparsity	GOMP	Combine CS-MUD with multi-carrier access schemes	[121]
	BSP	Exploit the block sparsity and propose two algorithms, namely, TA-BSASP and CVA-BSASP	[122]
	MP-BSBL	Combine BSBL with message passing and easily used to data detection	[123]
	AMP-BSBL	Retain the advantages of BSBL and has low complexity	[124]
	ADMM and Clustering	Utilize wireless channels' angular continuous group-sparsity feature and knowledge of stationary users' angular information	[125]
Methods using other structures	SISD	Exploit the inherent structured sparsity of user activity naturally existing in NOMA systems	[127]
	SBL	Propose two SSMs based on SBL	[128]
	OAMP-MMV	Exploit a more accurate prior distribution to characterize the structured sparsity of signals	[129]
	EM-MRF-GAMP	Exploit the inherent clustered sparsity structure in the angular domain in MIMO system	[130]
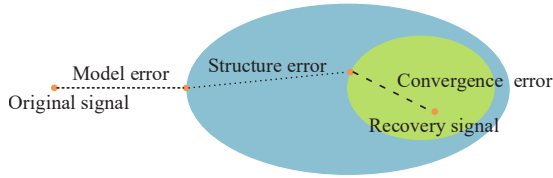
Fig. 5: The decomposition of the error in model-driven approaches [59].

prominent advantage is the increase in the speed of convergence. A fast DL-based approach for CS-based multi-user detection in massive MTC systems shows more than a 10-fold decrease in computing time, compared to traditional iterative algorithms [140]. Reducing the computing time is critically important for latency sensitive applications in 5G and 6G. In addition, appropriately trained DL networks are capable of increasing the detection/estimation accuracy. As shown in Fig. 5, the detection/estimation error of the traditional model-driven approach results is composed of model error, structure error and convergence error in general. The model error comes from imperfect modeling of the original engineering problem, i.e., oversimplification o f t he r eal p roblem. T he structure error comes from approximation of the original objective function, e.g., using convex relaxation to facilitate algorithm derivation for complex problems. The optimal solution of the approximated problem would deviate from the true "optimal" one of the original problem. The convergence error comes from iterative algorithms that fall into local optima in the function landscape. DL-based methods learn the mapping from the input to the output directly and has the potential to relieve challenges brought by the model error, the structure error and the convergence error in model-driven approaches. Reference [59] provided a comprehensive review of recent progress in the development of DL for solving linear inverse problems.

Significant efforts have been put into designing a powerful receiver with the assistance of DL in the past few years. A fast DL-based approach for AUD in MRA was first proposed in [140], where a nonlinear block restrictive activation unit was proposed to capture the block sparse structure in wide-band wireless communication systems (or multi-antenna systems). Reference [141] proposed to aggregate two Zadoff-Chu preamble sequences from two different roots to obtain a larger set of preambles and used DL to decode the aggregate preambles. DL-based activity detection for cell-free (CF) massive multiple input multiple output (mMIMO) systems was investigated in [142], and a transfer learning-based ensemble model was established in the central processing unit to achieve a better global detection decision. Various DL-based methods [143]–[148] have been developed for joint AUD and CE have been developed. Reference [38] introduced an adaptive tuning module and designs a feature-assisted adaptive tuning DL method that includes inner and outer networks to solve the massive device detection problem. In [145], [148], a deep unfolding method based on the traditional ADMM algorithm was developed, where the iteration operation was unfolded into the network layer. Reference [148] further considered a model-driven neural network architecture based on vector AMP. The long short-term memory (LSTM) was exploited for the detection of active devices and the estimation of the channel of active devices in [144]. In [146], the joint AUD and CE problem was formulated as a group-sparse matrix estimation problem, and a simplified unrolled recurrent neural network with reduced training parameters was proposed to solve the problem. Correlated antennas at the BS was considered in the design of the DL method for MRA in [147]. Joint activity detection, CE, and data recovery were investigated in [38], [149]. A novel dual network was proposed in [149] to utilize the sparse consistency between the channel vector and the data matrix of all active users. Information distilled from the initial

data recovery phase was employed as prior information to improve DL methods for activity detection and CE, which in turn improves data recovery performance. Theoretical analysis demonstrated that the proposed method can converge faster and support more users [38].

Pilot sequences can be jointly designed in DL-based methods for MRA. In [150], two DL models, using the standard autoencoder structure in DL, were investigated for the joint design of pilot sequences and CE methods, and the joint design of pilot sequences and device activity detection in MIMO-based grant-free random access, respectively. For small packet transmission, CE becomes costly and inevitably introduces a huge amount of overhead. Several existing works have attempted to investigate non-coherent schemes, which do not need CSI at the receiver. In [151], the transmitted messages were embedded in the index of the transmitted pilot sequence of each active device, and a DL-modified AMP network was developed to effectively exploit the correlation of pilot activity. In [152], inspired by the algorithm based on AMP, a model-driven DL method was proposed for noncoherent communications to jointly detect user activity and desired data.

DL-based methods have also been investigated for the MRA problem with imperfect synchronization. In [153], a learned AMP network was designed for joint user activity detection, delay detection and CE. Asynchronous access of machine-type devices that send data packets of different frame sizes, called data length diversity, was considered in [154]. The composite problem of AUD, CE, and data recovery was formulated as a structured sparse recovery problem, and a DL method unfolding from AMP with a backward propagation algorithm (AMP-BP) was developed to this end.

The above DL-based methods for MRA consider a fixed number of layers. However, they ignore a key character in traditional iterative algorithms, where the number of iterations required for convergence changes with varying sparsity levels, i.e., "easy" tasks should be solved with fewer iterations. In MRA, the number of active users is difficult to predict and could vary in a large range depending on service type and user mobility. For a relatively small number of active users, using a fixed number of iterations leads to the waste of computing power and increases communication latency. For a relatively large number of active users, using a fixed number of iterations leads to poor user detection accuracy, as the algorithm is not likely to converge yet. To this end, a novel depth adaptive approach was developed to effectively reduce the computational complexity of DL-based methods [39]. The proposed method with adaptive depth provided the solution to adapt the neural network to the varying conditions in the number of access users.

Another line of research to improve MRA efficiency is the tuning of access configurations, which controls the probability of user access and alleviates access contention. DL-based transmit power control schemes were proposed in [155], [156]. Unlike conventional transmit power control schemes in which complex optimization problems have to be solved in an iterative manner, the DL-based scheme determines the transmit power with a relatively low computation time. The joint design of the transmit power control functions and the estimation

algorithm was proposed in [157], which improves the accuracy of AUD and CE. A DL-based scheme was proposed to detect and resolve preamble collisions in [158]. Given the detection result, the preamble collision can be resolved according to the different timing advance, which is also realized by DL. In [159], DL was employed to detect random access collisions by learning the features of the signals received in the BS and then an additional contention process was scheduled for users with collisions. A summary of existing work on machine learning for MRA is given in Table IV.

IV. Non-Orthogonal Multiple Access

In this section, we introduce progress in model-based methods and learning empowered methods for NOMA.

A. Model-Based Methods for NOMA

NOMA allows users to share the same resources and distinguishes them in power, code, or some other emerging domains. In this subsection, we will introduce methods for several prominent NOMA schemes, including PD-NOMA, code-domain NOMA (CD-NOMA), and pattern division multiple access (PDMA). Then, we focus on methods dealing with challenges including asynchronous NOMA and NOMA with imperfect CSI.

1) Practical NOMA Schemes:

- *PD-NOMA* is realized in the power-domain [160], in contrast to the traditional multiple access schemes relying on the time, frequency, code-domain or on their combinations. Users send messages using different power depending on their respective channel quality after classic channel coding and modulation. Multiple users share the same time-frequency resources, and then multiuser detection algorithms, such as SIC, are used at the receiver to detect different user signals from the superimposed signals. In this way, the SE can be enhanced at the cost of increased receiver complexity compared to conventional OMA. With the rise of multi-antenna technology, there is an increasing number of studies that combine PD-NOMA with multi-antenna to further improve SE [161]–[165]. In the multi-antenna scenario, the user clustering, beamforming design, and power allocation subject to the quality of service constraints need to be simultaneously considered and jointly optimized. For instance, the authors of [165] proposed a two-step user clustering and power control algorithm. Considering the intra-cluster interference and inter-cluster interference, K-means algorithm was used for user clustering, and then power control was done. Simulation results showed that user clustering by K-means can achieve lower energy consumption and higher EE.
- *CD-NOMA* is inspired by CDMA systems, in which users share the same time-frequency resources but adopt unique user-specific spreading sequences. User-specific scrambling/interleaving, symbol spreading and symbol to resource mapping could be used to distinguish different

TABLE IV: Summary of Existing Work on Machine Learning for MRA.

Category	Characteristics	Ref.
AUD	Exploit block restrictive activation nonlinear unit to capture the block sparse structure	[140]
	Establish a transfer learning-based ensemble model in CF mMIMO system	[142]
Joint AUD and CE	Exploit LSTM	[144]
	Use deep unfolding method based on ADMM	[145], [148]
	Exploit group-sparse property to simplify network	[146]
	Employ initial data recovery phase as prior infomation	[149]
Joint AUD and pilot design	Use standard auto-encoder structure	[150]
	Develop a DL-modified AMP network to exploit the pilot activtity correlation	[151]
	Propose GAMP-inspired DL method	[152]
Imperfect synchronization	Design a learned AMP network for joint user activity detection, delay detection, and CE	[153]
	Consider data length diversity and formulate a structured sparse recovery	[154]
Unknown sparsity level	Develop a novel depth adaptive approach to adapt different number of access users	[39]
Access configuration	Use DL scheme to replace conventional power control schemes	[155], [156]
	Jointly design the transmit power control functions and the estimation algorithm	[157]
	Resolve random access collision by DL	[158], [159]

users at the transmitter side [51]. Low-density signature (LDS)-CDMA [166] uses a specially designed non-orthogonal spreading sequence with sparse characteristics. Each user extends its own information according to the signature sequence to the corresponding code chip for transmission. At the receiver, a near-optimum chip-level iterative multiuser decoding based on the message passing algorithm (MPA) is proposed to approximate optimum detection by efficiently exploiting the LDS structure. LDS-OFDM [167] retains the benefits of OFDM-based multicarrier transmissions such as intersymbol interference avoidance and operates at a lower complexity than the optimal maximum a posteriori (MAP) detector of LDS-CDMA. Sparse code multiple access (SCMA) [168] directly maps the bits to multidimensional codewords in the complex domain after channel coding at the transmitter, and the transmitted codewords of different users are non-orthogonally superimposed in a sparse spreading manner on the same time-frequency resources, and this sparse characteristic is used at the receiver for low-complexity multiuser joint detection. It is crucial to design the spreading sequences or the codewords in CD-NOMA. A set of user-specific non-orthogonal binary spreading sequences for uplink grant-free NOMA is studied in [169] to improve the performance of the joint channel estimation (CE) and multi-user detection. An improved low-density spreading sequence design based on projective geometry is proposed in [170]. In [171], reinforcement learning was exploited to design the SCMA codebooks, which generates codebooks with superior quality and has a low complexity. Another breakthrough to achieve ow-complexity CD-NOMA is rearranging OFDM subcarrier sequences in a symbol to efficiently supports multiuser access of lightweight IoT nodes [172].

- *PDMA* [173] relies on uniquely designed multi-user diversity patterns to recognize non-orthogonal transmissions in the power, time, frequency, spatial and code domains. PDMA considers the joint design of the transmitter and receiver. On the transmitter side, transmitted data is mapped to a resource group that can consist of time, frequency, and spatial resources, or any combination of these resources, which is called PDMA pattern and can be represented by a binary vector. The pattern is important for separating the signals of users sharing the same resources, and its design should take into account the complexity of the detection. The general SIC algorithm is used at the receiver side to perform multiuser detection, to differentiate overlapping user signals based on different user patterns.

2) Asynchronous NOMA: As discussed in the previous section, the capacity region of the asynchronous uplink channel is larger than that of the synchronous uplink channel. As such, the burden of perfectly synchronizing different users does not apply to NOMA. In addition, it is beneficial to add intentional time asynchrony to improve the performance of NOMA systems.

Uplink ANOMA: Uplink ANOMA systems can be created by adding intentional timing mismatch among the symbols of different users. Assuming the timing mismatch is known, over-sampling generates sufficient statistics and provide sampling diversity. ANOMA offers a larger capacity region, compared to synchronous NOMA, but requires dealing with IUI and ISI. A simple SIC is not optimal anymore and maximum-likelihood sequence detection is needed to manage the ISI [71]. One approach to manage ISI in ANOMA is to represent the ANOMA system with a Toeplitz system model and apply the trellis-based detection method to the corresponding Ungerboeck model for ISI channels. Then, as proposed in [77], the channel diagonalization and turbo principle can be used to design an ANOMA transceiver. The optimal timing mismatch for a system with two users is half of the symbol interval. For systems with K users, a good timing mismatch choice among different users is the symbol interval divided by K. The operational rate point for two users can be outside of the synchronous uplink channel's capacity region, i.e., it will outperform any possible synchronous NOMA system. In [78], the sum throughput of two-user ANOMA systems is calculated

analytically. It is shown that by intentionally adding a timing mismatch, equal to half of the symbol interval between the symbols of the two users and over-sampling, the performance of ANOMA is always better than that of synchronous NOMA for a sufficiently large frame. Even when the exact timing mismatch between the two users is not known at the receiver, using an arbitrary equal-distance over-sampling, ANOMA's sum throughput is still more than that of the synchronous NOMA. The average symbol error rate of ANOMA using QAM over Rayleigh fading channels is analytically calculated in [174]. Also, an iterative detection method that takes into account the correlated nature of the noise and the error propagation is proposed. The approach is more effective for lower-order QAMs. In [175], another iterative method using message passing detection for ANOMA is designed and its performance is analysed. A massive device-to-device network utilizing ANOMA to reduce the decoding complexity, energy consumption, and bit error rate is considered in [176]. To manage co-channel interference, an optimization framework is defined to maximize the number of successfully decoded symbols.

Downlink ANOMA: The main idea behind downlink ANOMA is to add intentional sub-symbol time delays between the symbols of different users before implementing the superposition coding at the transmitter. At the receivers, over-sampling provides sufficient statistics for decoding. Because of the over-sampling, the noise is correlated and needs to be whitened. Singular-value decomposition and precoding can be used to create independent parallel channels that do not need SIC [79]. In fact, adding intentional time delays will induce frequency-selectivity that can improve the performance in conjunction with beamforming and precoding methods [177]. The performance of ANOMA, with or without precoding, is better than that of synchronous NOMA.

In NOMA, the stronger user decodes the signal of the weaker user and cancel its interference. Since the stronger user has access to the data from the weaker use, it can forward it to the weaker user, acting as a relay. Then, the weaker user has access to two copies of its symbols and can use them to improve the performance. In [178], the achievable average rate of such a cooperative NOMA (C-NOMA) scheme is analyzed. In [179], C-NOMA is designed for multiple antennas. Using MRC at the receivers and antenna selection at the transmitter, the outage probability is calculated analytically. It is shown in [180] that, similar to a regular NOMA, adding intentional time delays to C-NOMA schemes will improve the performance. The throughput of both users in such a C-ANOMA system is calculated and it is shown analytically that C-ANOMA outperforms C-NOMA. In addition, the optimal timing mismatch is half of the symbol length for large frame sizes. In [181], a system similar to C-ANOMA is used for satellite communication. An optimization problem to jointly allocate resources to improve the fairness is defined and solved for a multi-satellite system.

3) NOMA With Imperfect CSI: While NOMA can provide many advantageous over OMA, the benefits of NOMA are usually shown under perfect assumptions, such as the availability of perfect CSI at both transmitter and receiver. For the multi-user/multi-cell environments for which NOMA is an essential technology, acquiring perfect CSI is not an easy task (if not impossible). Thus, designing systems to deal with imperfect CSI is very important when shifting from the OMA paradigm to a NOMA paradigm. Channel state information plays a crucial role in wireless communications. In a single-user system, communication system design heavily depends on how much channel information is available [182]. Although the effect of imperfect CSI at the receiver on emerging NOMA techniques is still an ongoing research topic, similar to OMA systems, the estimated CSI can be modeled as the perfect CSI plus a Gaussian residual error. The achievable rates in NOMA depend on not only the individual estimates of channel coefficients, but also their relation. In uplink NOMA, the BS first decodes the signal of the strong user and then, after removing its interference, decodes the signal of the weaker user. In downlink NOMA, the stronger user decodes the weaker user's signal first and after removing it, decodes its own signal. On the other hand, the weaker user only decodes its own message considering the other user's signal as noise. Therefore, the decoding order which depends on the relative relation between channel coefficients plays a significant role in system performance. If the imperfect CSI at the receiver results in incorrect decoding orders, it can have catastrophic consequences.

Assuming perfect CSI at the receiver, the estimated channel values need to be communicated with other nodes. In some scenarios, channel reciprocity is used for this purpose. However, in most practical cases, the estimated channel values should be quantized and sent back to other nodes through a limited feedback channel. Downlink NOMA with limited feedback is analyzed in [183], where variable-length quantizers that can approach the perfect CSI performance are designed. It is proved that the relationship between the feedback rates and the losses in rate and outage probability is at least exponential. In [184], downlink NOMA systems in which the transmitter does not know the perfect CSI is studied. It is shown that if the power allocation strategy is not well-designed, outages can occur. In addition, for specific target rates, a power allocation method that achieves the OMA outage performance is designed. In [185], the performance of a limited feedback system using NOMA in a hybrid UHF/mmWave downlink network is analyzed. The achieved outage probability is much lower than that of a network with no feedback. In [186], optimal quantizer algorithms for NOMA and ANOMA are designed. A two-user downlink ANOMA system with limited feedback is considered and its max-min rate's upper and lower bounds are expressed in closed-form. For the same max-min rate, the required feedback rate of ANOMA is less than that of NOMA. The design of a limited feedback that includes reconfigurable intelligent surfaces in NOMA systems is considered in [187]. The rate loss resulted from the quantization error is calculated.

B. Machine Learning for NOMA

The capability of DL can greatly aid in resolving diverse complex problems in NOMA such as CSI acquisition [188], signal constellation design [189], resource allocation, user

pairing and multiuser detection. Generally, two categories of learning techniques have been applied in the literature to solve the above problems, including DL and reinforcement learning (RL). DL has been widely used for estimating CSI [42], throughput prediction [190], transceiver design [191], [192], and resource management [193], [194]. Based on the agent-learning control mechanism, RL and deep RL (DRL) are used in resource management, especially in dynamic resource management [195]–[197].

Resource allocation is essential in NOMA in order to achieve the optimal total sum rate. Traditional methods try to maximize the EE or the sum rate under a series of constraints. Most optimization problems are nonconvex and cannot be equivalently transformed into convex optimization problems, so it can only be solved suboptimally using methods such as successive convex approximation (SCA) [198], which has been used in power allocation. In [193], a DL-based framework for subchannel and power allocation was proposed to improve NOMA EE in mmWave networks. The subchannel allocation was solved by semi-SL, in which some labeled data was initially generated by a two-sided matching scheme. Using the generated data along with unlabeled data, the network was trained and was able to accurately predict the real label of an unlabeled input. The power allocation was solved using a DL scheme, which replaces the complex iterative process and only uses simple multiplication and addition of matrices to compute the output. The simulation results showed that the proposed framework achieved a higher EE and a lower complexity. Considering a relay-aided device to device system, a deep neural network (DNN)-based algorithm was designed in [194] to replace intractable coupled joint optimization and acquire a joint power loading solution at the source and relay nodes. Their results revealed that the DNN model leads to promising performance in terms of sum rate and computational complexity. In [199], a two-step algorithm was designed to perform sub-band assignment and power allocation successively for each NOMA-based IoT user to minimize the transmit power sum. And unsupervised learning (USL) was implemented for power control.

In addition to DL methods, RL was applied for dynamic resource allocation in time-varying communication channels [195]–[197], [200]–[202]. In [202], a model-free centralized and distributed approach based on DRL was proposed in multiband hybrid OMA-NOMA to jointly optimize user association, transmit power allocation, sub-channel assignment, and multiple access technique selection. In [195], a two-step DRL-based algorithm was proposed to solve the dynamic optimization problem of EE. To be specific, using the current channel conditions as input, a deep q-network (DQN) was designed to obtain the optimal subcarrier allocation policy, and a deep deterministic policy gradient (DDPG) network was used to dynamically output the transmit power for all users. To maximize the average performance of sum rates in uplink NOMA-IoT, two RL algorithms, i.e., SARSA-learning and DRL, were proposed in [196] to dynamically allocate users and balance resource blocks and network traffic. Numerical results demonstrated that the two algorithms have a lower complexity and outperformed OMA systems. In [197], DQN

and DDPG were proposed for sub-carrier assignment and power allocation in RIS assisted semi-grant-free NOMA transmission, respectively. Two DDPGs were integrated to assign amplitude and phase shift to RIS's reflecting elements. A distributed cooperative channel assignment and power control approach based on multi-agent RL is presented in [203] to solve the massive access management problem.

The design of the transceiver is another important issue in NOMA. In practice, there are multiple practical schemes, e.g., PD-NOMA and CD-NOMA, and the signal formats of each scheme are significantly different, requiring different transceiver designs, which hinders the standardization and implementation of NOMA. The DL method emerges and makes possible a unified transceiver architecture. In [204], a data-driven model was proposed for SCMA. The traditional decoding algorithm in SCMA, such as MPA, was replaced by a fully connected DNN (FC-DNN), which is trained using a synthetic dataset to minimize the symbol error rate. It also reduced the computation time to a large extent. The variational autoencoder (VAE) was exploited to perform signal detection with jointly designed spreading signatures [191]. An end-to-end learning-based SCMA framework was introduced in [205] to jointly design the codebooks and low-complexity decoder. The above data-driven DNNs lead to a unified framework for different NOMA schemes, while being trained using different datasets. However, data-driven methods do not take advantage of the structure of NOMA signals and thus lead to low training efficiency. Model-driven methods bring structured DNN based on traditional algorithms and domain knowledge. In [40], [206], an SIC-inspired DNN was designed for multi-user detection. The network was inspired by the conventional SIC detection structure, in which DNNs replaced the detection layer in SIC, and the output of the first branch was fed to the second branch to cancel inter-user interference like SIC. Numerical results demonstrated the effectiveness and superior performance of the model-driven method. The DL methods mentioned above consider block-wise optimization. A unified end-to-end optimization framework called DeepNOMA was proposed in [192], which consists of a channel model, a multiple access signature mapping module, and a multiuser detection model. End-to-end optimization is enabled by minimizing the overall reconstruction loss. What is more, each users' signals in non-orthogonal transmissions are treated as multiple distinctive but correlated tasks, and multitask learning and balancing techniques are used to improve the performance, guarantee fairness among users, and avoid local optima.

In NOMA, there are also other important problems that can be solved efficiently by DL. User pairing is essential in NOMA systems, as a reasonable user pairing strategy can improve system performance. In [207], after obtaining the power allocation strategy, a user pairing matrix was designed based on the channel gain of all users and a DQN-based algorithm was utilized to solve the user pairing problem to get the maximum sum rate. Reference [208] expanded the scenario to multicell NOMA and jointly optimized the user pairing and association problem. The pointer network (PtrNet) was used to solve the combinatorial optimization problem, and DRL was utilized at the training phase. Simulation results showed

that the scheme achieved near-optimal performance, which is better than the random user pairing and association heuristic by up to 30%. A DL-based adaptive user pairing scheme, which decides the two optimal user pairings, was proposed in RIS-aided MIMO-NOMA to maximize the overall spectral efficiency in [209].

Machine learning has also been leveraged to improve transmission security in NOMA networks. In [210], the interaction between the source station and the unmanned aerial vehicle (UAV) was viewed as a dynamic game and a DRL algorithm was applied for dynamic power allocation at the source station to suppress the attack motivation of the UAV. The simulation results showed that this strategy improves the system data rate and suppresses the attack probability. The NOMA transmission game against jamming was investigated in [211], where a fast Q-based NOMA power allocation scheme that combines the hotbooting technique and Dyna architecture is proposed for a dynamic game to accelerate the learning and thus improve the communication efficiency against smart jamming. The summary of the existing work on machine learning for NOMA is given in Table V.

V. INTERPLAY WITH OTHER NEXT-GENERATION WIRELESS TECHNOLOGIES AND SCENARIOS

Having introduced the fundamentals and advanced signal processing and learning techniques of MRA and NOMA, in this section, we focus our attention on the promising interplay with other new next-generation technologies and scenarios, e.g., AoI, NFC, ISAC, STARS and semantic communications.

A. NGMA with New Scenarios

In the future, with the rapid development of IoT, the diversity of IoT devices will increase, leading to varying communication needs among different devices. For example, in the context of Internet of Vehicles, in addition to the massive access characteristic of MMTC, it is also necessary to ensure low latency and high reliability for certain sensors. In this subsection, we introduce the consideration of AoI and heterogeneous services.

1) Age of information: Upcoming use cases and applications will soon impose new and stricter demands on MRA. In the design of 6G systems, new KPIs will become pivotal, e.g., AoI. In future 6G networks, information freshness will be increasingly important for applications of various types of real-time monitoring systems (e.g., smart driving) and state update systems (e.g., industrial control). The use of outdated information could profoundly compromise the precision and dependability of systematic decision-making processes. To quantify information freshness, a new metric, i.e., AoI, is introduced to measure the information freshness at the receiver [212]. To reduce the AoI of the system, an optimized age based random access scheme is proposed in [213], and a neural network where AoI is aided as the prior information is further proposed. To improve the overall status update performance, a joint access control and resource allocation scheme is proposed in [214], taking AoI as the performance metric. To achieve a low average AoI and large throughput in satellite-based IoT, a

grant-free age-optimal random access protocol is proposed in [215], where the number of access time slots in each transmission frame is optimized. In [216], an RL-based transmission strategy is proposed in which the AoI of the users and the total throughput of the system are jointly optimized. Based on the AoI levels of users, they are divided into two groups and assigned distinct access patterns. According to [217], in the slotted-ALOHA framework, a device is allowed to access the channel with a constant probability if its AoI exceeds a specified limit. The long-term expected AoI is calculated, followed by an optimization of the access probability and the threshold.

2) Heterogeneous services: In future wireless networks, numerous distinct application scenarios are encountered. In these contexts, users with different Quality of Service (QoS) needs must coexist. Emerging in heterogeneous radio access networks, a blend of enhanced mobile broadband (eMBB), MMTC, and ultra-reliable low-latency communication (URLLC) devices could be accommodated within the same spectrum resource. In [218]–[220], the coexistence of eMBB and URLLC is studied. The joint resource allocation problem to maximize the minimum expected achieved rate of eMBB users while satisfying the low latency and high reliability constraints of URLLC users is studied in [218]. Furthermore, URLLC traffic is overlapped on eMBB by adopting the NOMA superposition technique. In [219], the non-orthogonal coexistence between eMBB and URLLC in the downlink of a multi-cell massive MIMO system is rigorously analyzed. Puncturing and superposition coding are considered as alternative downlink coexistence strategies, and the eMBB spectral efficiency and the URLLC error probability are analyzed. In [220], the optimal frequency and power allocation scheme and eMBB-URLLC pairing policy is derived for scheduling URLLC traffic in a downlink system in the presence of eMBB traffic. A preamble allocation scheme using hierarchical reinforcement learning is proposed for the coexistence scenario of MMTC and URLLC in [221]. A promising scheme based on RSMA for network slicing is proposed in [222] to guarantee the performance of heterogeneous devices. Analysis shows that RSMA can outperform its NOMA counterpart, and obtain significant gains over OMA in some regions. In [223], a novel hierarchical hybrid access class barring (ACB) and backoff (BO) scheme is proposed, where the hybrid ACB-BO is exploited to balance the delay-energy tradeoff, and the hierarchical structure is proposed to prioritize communication services. Deep reinforcement learning is applied to the proposed random access scheme to dynamically adjust the ACB factors and BO indicators in an online manner.

B. Near-field Non-Orthogonal Multiple Access

In 6G, paradigm shifts are currently taking place in the architecture to satisfy the stringent requirements. Specifically, by increasing the aperture size of transceivers and using extremely high-frequency bands, the achievable data rate and connectivity can be massively improved. In the meantime, near-field signal propagation in future wireless networks will become important, which brings new opportunities for NOMA.

TABLE V: Summary of Existing Work on Machine Learning for NOMA

Category	Scenarios	Algorithm	Characteristics	Ref.
Resource allocation	mmWave NOMA	DNN	Use semi-SL for subchannel allocation	[193]
	Relay-aided NOMA	DNN	Propose a joint power loading solution at source and realying nodes	[194]
	NOMA-IoT	USL-based DNN	Implement USL for power control	[199]
	NOMA-aided MEC	DQN	Use DQN-based offloading strategy to minimize the system cost	[200]
	Hybrid NOMA	Actor-critic	Use proximal policy optimization and recurrent neural networK	[201]
	Multi-band hybrid OMA-NOMA	DRL	Propose model-free centralized and distributed approaches based on DRL	[202]
	RIS assisted NOMA	DQN, DDPG	Propose a two-step DRL based aligorithm	[195], [197]
	NOMA-IoT	SARSA-learning,DRL	Design novel 3D state and action spaces	[196]
	NOMA-IoT	DQN	Build a QoS-aware reward to cover both network EE and devices QoS	[203]
Transceiver design	Grant-free NOMA	FC-DNN	Propose a unified NOMA framework	[191], [204], [205]
		SIC-inspired DNN	Design a two branch network and the structure is SIC-liked	[40], [206]
		DeepNOMA	Propose an end-to-end optimization framework	[192]
User pairing	NOMA	DQN	Design and optimize a user pairing matirx	[207]
	Muticell NOMA	PtrNet	Use a reinforce-based method to perform parameter optimization	[208]
	RIS-aided MIMO-NOMA	DNN	Propose DL-based adaptive user pairing scheme	[209]
Secure issue	NOMA	Q-learning	View the interaction between BS an UAV as dynamic game	[210]
	MIMO NOMA	Hotbooting Q-learning	Apply Dyna architecture and hotbooting techniques	[211]

In the following, we first introduce the key features of NFC. Then, we highlight the *beamfocusing* property in NFC which can benefit NOMA. Finally, we introduce the beamfocusing-empowered NOMA communications.

1) Key Features of NFC: In wireless communication, the *near-field* region refers to the area close to the transceivers so that the electromagnetic field in that region exhibit near-field effects. One widely used rule-of-thumb for determine the near-field region is the *Rayleigh distance*, which depends on the aperture size of the transceivers and the carrier frequency. The larger the aperture size and the higher the carrier frequency, the larger the near-field region becomes. Within the near-field region, signal propagation has different characteristics compared to that within the far-field region. In the following, we elaborate on the new features of NFC.

- **Distance-Dependant Radiation Pattern**: The near-field region in wireless communication deviates from the free-space path loss observed in far-field communications. In particular, the near-field radiation pattern, which represents the angular distribution of signal power, exhibits variations as a function of the communication distance. This means that even in a single direction, the path loss may oscillate instead of following a monotonically decreasing trend, as described by the far-field Friis' path loss formula [224]. As we will elaborate in Section. V-B2, this near-field property gives rise to the possibility of *beamfocusing*.
- **Spherical and Irregular Wavefronts**: In contrast to the planar waves in the far-field region, near-field radiation has a spherical or irregular wavefront. Close to the Rayleigh distance, the wavefront appears spherical, with the energy spreading equally in all directions. For regions close to the transmitter, the wavefront can become irregular or distorted. This is because the antenna's physical structure and geometry can introduce variations in the propagation path and cause interference effects.
- **Faster Decay and Reactive Field**: Near-field also exhibits a faster decay due to the dominance of spherical waves and evanescent waves. Evanescent is a type of electromagnetic field whose energy decrease exponentially with distance. The region where evanescent wave dominates is referred to as the reactive near-field. It is important to note that the reactive near-field is confined to just a few wavelengths from the antennas or antenna arrays, even if they are infinitely large [225]. As a result, this region is typically disregarded in near-field communication and signal processing design.
- **Enhanced Degrees-of-Freedom (DoFs)**: The line-of-sight (LoS) MIMO channel in NFC typically exhibits a higher number of DoFs. In contrast to the LoS MIMO channel in conventional far-field communications, where the achievable DoF is 1, the enhanced spatial multiplexing gain can be achieved in NFC even without a rich scattering environment, which is the major benefit of NFC.
- **Non-Uniform Pathloss**: The communication channel in NFC can exhibit high variability due to quick changes in the relative positions and orientations between the transmitter and receiver. As a result, the subchannels be-
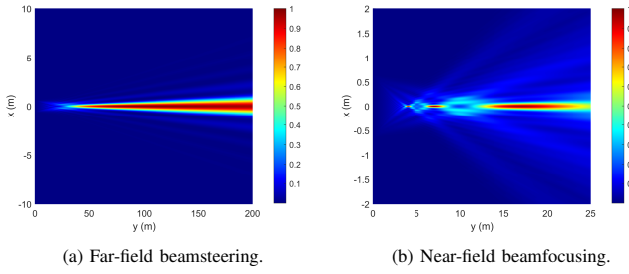
(a) Far-field beamsteering. (b) Near-field beamfocusing.

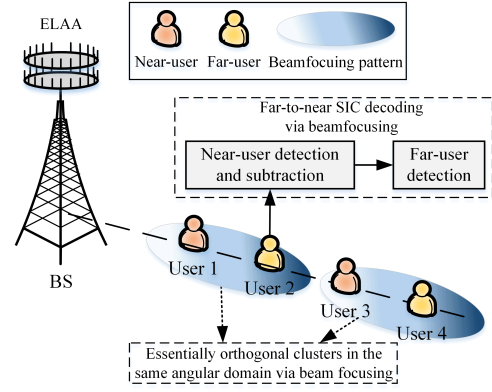Fig. 6: Beampattern of far-field beamseering and near-field beamfocusing.

Fig. 7: Illustration of near-field beamfocusing-empowered NOMA, where the near-field beamfocusing vectors divide users in the same angular direction into two clusters and the signal energy is focused on the far-user in each cluster.

tween two multi-antenna transceivers exhibit non-uniform pathloss. In channel modeling, non-uniform spherical wave models should be used characterize these near-field channels.

2) Near-field Beamfocusing Property: As previously discussed, the signal propagation in the near-field region exhibits distinct characteristics, primarily manifesting as spherical waves, in contrast to the planar waves observed in conventional far-field communications. Consequently, the fundamental principles underlying beamforming techniques diverge between near-field and far-field communication scenarios within multi-antenna systems. Specifically, as shown in Fig. 6a, in the far-field region, the angular field distribution remains unaffected by the distance from the transmit antenna array, resulting in the capability of the far-field beamforming process to solely steer towards a predetermined direction, commonly denoted as *beamsteering*. Conversely, in the near-field region, the angular field distribution is contingent upon the distance from the transmit antenna array. Thus, as shown in Fig. 6b, the near-field beamforming technique exhibits the ability to concentrate the beam energy at a specific distance along a designated direction, referred to as *beamfocusing*.

Note that near-field beamfocusing is not a new research topic and can date back more than 60 years [226]–[228]. However, it was mainly investigated in the field of physics and antenna propagation. In recent years, there has been a notable trend towards the utilization of extremely large-scale antenna arrays (ELAAs) and ultra-high frequencies, such as mmWave and Terahertz (THz) bands, to meet the rigours communication requirements of future wireless systems. Consequently, the near-field region in wireless systems becomes non-negligible, with its spatial extent reaching tens or even hundreds of meters with respect to the BS [224]. Given this very large near-field region, many researchers began to investigate the potential benefits of near-field beamfocusing in wireless communications [229]–[231]. Specifically, the basic principle of near-field beamfocusing in wireless communications was discussed in [229], where several distances distinguishing the near-field and far-field regions of an antenna array were defined. Considering a multi-user communication scenario, the authors of [230] investigated the beamfocusing capability of different antenna architectures. It was demonstrated that in the near-field region, different beams can be generated to focus on different communication users located in the same direction without causing significant inter-user interference. As a further advance, a novel concept of location-division multiple access

(LDMA) was proposed in [231]. By exploiting the additional distance dimensional provided by near-field beamfocusing, multiple users can be supported in the location domain in the near-field region, which provides much higher DoFs than the angular domain in the far-field region.

3) Near-field Beamfocusing-empowered NOMA: With the aid of the promising beamfocusing property in NFC, there are two main benefits that can be achieved by the near-field beamfocusing-empowered NOMA.

- **Far-to-near SIC decoding**: With the properly designed beamfocusing vectors at the BS, the signal can be more focused on the users far from the BS than those near-users, i.e., the effective channel gain of the far-user can be larger than that of the near-user. As a result, it enables the far-user to carry out SIC to detect and remove the signal of the near-user, thus realizing the far-to-near SIC decoding order, as shown in Fig. 7. This is almost impossible to realize for conventional far-field NOMA communications, where the SIC decoding order among users is generally determined by their distance to the BS. Therefore, the near-field beamfocusing-empowered NOMA provides enhanced flexibility in the SIC decoding order design, which helps to better satisfy the communication requirements of each user.
- **Distance-domain user clustering**: Moreover, with the aid of beamfocusing, users located in the same angular domain can be further distinguished via the distance domain into multiple (nearly) orthogonal clusters, as shown in Fig. 7. This is greatly different from the far-field NOMA, where the user clustering is mainly determined by angular similarity. By doing so, on the one hand, the inter-user interference can be further mitigated due to the orthogonality between user clusters. On the other hand, with the reduced number of users in each cluster, the number of SIC operations reduces accordingly, which reduces the complexity of using NOMA.

To exploit the above benefits of the near-field beamfocusing-empowered NOMA, two novel frameworks were proposed

in [232] for a BS employing the hybrid beamforming structure, namely, single-location-beamfocusing near-field NOMA and multiple-location-beamfocusing near-field NOMA. For the single-location-beamfocusing near-field NOMA, users located in the same angualr direction but having distinct communication requirements are clustered into one group, which are served by one analog beamformer focusing on one specific user. For the multiple-location-beamfocusing near-field NOMA, users in the same group can have different angular directions, which is served by one analog beamformer focusing on multiple users by exploiting the beam-splitting technique. It shows that compared to conventional far-field NOMA, the proposed near-field beamfocusing-empowered NOMA can achieve higher spectral efficiency and enhanced inter-user interference mitigation. Moreover, the authors of [233] investigated the NOMA-based coexistence of near-field and far-field communications, where the near-field beamfocusing vector is shown to be able to serve an additional far-field user via NOMA. Note that the investigation of near-field/hybrid-field NOMA communications is still in an early stage, more research efforts are required to unlock the full benefits.

C. NOMA for Integrated Sensing and Communications

ISAC has been recognized as a crucial facilitator of ubiquitous intelligence in 6G [25], [234]. ISAC systems aim to integrate communication and sensing functionalities, leveraging shared spectrums, hardware platforms, and signal processing modules. Nonetheless, the coexistence of these two functions poses significant challenges in terms of interference cancellation and resource allocation within the ISAC framework. Consequently, the subsequent section presents a comprehensive examination of ISAC systems from a multiple-access perspective, while also exploring the potential advantages of NOMA in enhancing ISAC systems.

1) ISAC from the Multiple Access Perspective: As depicted in Fig. 8, the ISAC system can be conceptualized as a multiple-access framework consisting of two distinct users: the communication user and the sensing user (target). The communication user operates in either an uplink (UL) mode, transmitting communication signals to the BS, or a downlink (DL) mode, receiving communication signals from the BS. In contrast, the sensing target behaves more like an inherent full-duplex user, concurrently "receiving" and "transmitting" signals. This distinction arises from the divergent operational mechanisms between sensing and communication. More specifically, while communication users are capable of actively engaging in coding and decoding processes to convey or recover information, the sensing target is passive. Consequently, the BS is responsible for transmitting probing signals to the sensing target and subsequently receiving the sensing echo signals reflected by the target. This enables the BS to acquire crucial parameters such as location, speed, and shape. It is worth noting that in some cases, such as the sensing-assisted beamforming framework described in [235], the communication user and sensing target may be the same physical object, which is out of the scope of our discussion.

2) Motivation of Employing NOMA for ISAC: Based on the above discussion, the concept of mitigating inter-user interference within multiple-access frameworks can be applied to solve inter-function interference in ISAC systems. A direct approach entails assigning "orthogonal" resource blocks in the frequency, time, or code domain to communication and sensing functions, resembling the concept of OMA where "orthogonal" resource blocks are allocated to different users. This orthogonal allocation strategy avoids inter-function interference in ISAC systems. However, while such an interference-free design reduces implementation complexity, it may result in inefficient resource utilization. As a remedy, non-orthogonal ISAC is a more popular design principle [25]. In a manner analogous to NOMA, this approach allows for the joint utilization of resource blocks by both communication and sensing functions, resulting in a substantial improvement in resource efficiency. However, it is imperative to effectively mitigate the occurrence of inter-function interference. Therefore, in the following, we discuss the property of inter-function interference in non-orthogonal ISAC systems in both uplink and downlink non-orthogonal ISAC systems.

- **Uplink ISAC**: The uplink ISAC system is depicted in Fig. 8a, where an uplink communication user and a sensing target are present. In this configuration, the sensing target can be considered as a virtual uplink communication user, "transmitting" echo signals to the BS. Consequently, the BS receives a superimposed uplink communication signal and sensing echo signal, leading to mutual interference between the communication and sensing functions.
- **Downlink ISAC**: The downlink ISAC system is illustrated in Fig. 8b, involving a downlink communication user and a sensing target. Unlike uplink ISAC systems, the sensing target cannot be entirely regarded as a virtual downlink communication user in the downlink scenario. This is because it has been demonstrated that, in the downlink case, all communication signals can be utilized for sensing, while the sensing signals should be exclusively dedicated to achieving the full sensing DoFs [236]. Consequently, only sensing-to-communication interference is present in this case.

Given the resemblance between non-orthogonal ISAC and NOMA in terms of their underlying principles, it is natural to leverage NOMA techniques to address inter-function interference in ISAC. However, due to the inherent distinction between the original purpose of NOMA in enabling efficient communication and the requirements of ISAC systems, it becomes necessary to redesign NOMA specifically for ISAC, resulting in the development of *multi-functional NOMA*. Consequently, numerous studies have emerged for studying NOMA-ISAC systems, which will be detailed in the following.

3) Existing Works on NOMA-ISAC: The performance of uplink NOMA-ISAC systems was initially explored in [237], wherein the authors employed the SIC technique to mitigate inter-functional interference. Building upon this concept, the authors of [238] proposed a NOMA-aided protocol that facilitates the seamless integration of communication, sensing, and computing functions. It was demonstrated that the pro-

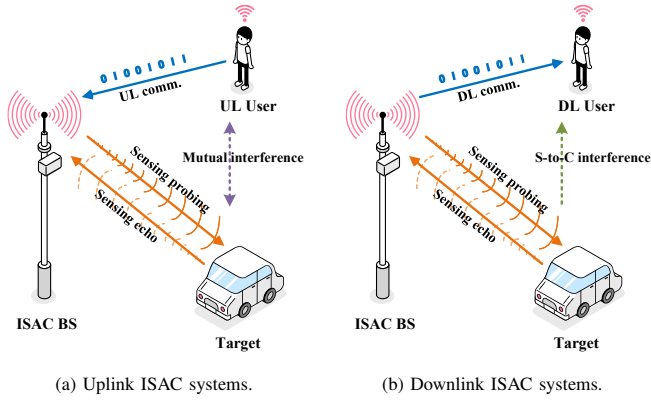(a) Uplink ISAC systems.   (b) Downlink ISAC systems.

Fig. 8: Illustration of ISAC systems from the multiple access perspective.

posed NOMA-aided protocol outperforms the conventional protocol based on SDMA due to its superior interference cancellation capabilities. As a novel contribution, a semi-NOMA framework was conceived for uplink ISAC systems in [24]. This framework involves the sharing of only a subset of resource blocks between the communication and sensing functions, thereby enabling a more adaptable resource allocation scheme and enhanced interference cancellation capabilities. Furthermore, a recent study [239] proposed a model and pattern based environment sensing technique utilizing CSI, which can be executed concurrently with communication functions through NOMA techniques. Regarding downlink ISAC systems, the potential of NOMA was initially explored in [23]. While focusing on NOMA within the communication function, this study demonstrated that NOMA can still offer additional DoFs for ISAC, particularly in scenarios with system overload or highly correlated communication channels. Building upon this foundation, the authors of [240] extended the concept of NOMA to address the sensing-to-communication interference between the two functions. They proposed a NOMA-inspired ISAC framework that effectively eliminates sensing-to-communication interference and enables the simultaneous utilization of dedicated sensing signals for communication. Furthermore, the integration of the sensing function into multicast-unicast communication systems was effectively achieved with the aid of NOMA, as illustrated in [241]. To address security concerns in NOMA-ISAC systems, [242] introduced a secure precoding approach, offering a solution to the security problem. Lastly, the issue of sensing scheduling in NOMA-ISAC systems was investigated in the most recent work [243], which was jointly optimized with the beamforming. Some other works (e.g., [244]) also investigate the application of NOMA in full-duplex ISAC systems, where the UL and DL communication users coexist. It is important to note that, in addition to the resource allocation and interference management issues discussed in the aforementioned works, waveform design is also crucial in ISAC system development. Specifically, sensing functions typically require deterministic waveforms, while communication functions rely on random signals to convey information [245]. Further research is needed to explore how NOMA techniques can be utilized to address this inherent tradeoff from the waveform perspective.

D. Simultaneously Transmitting and Reflecting Surfaces (STARS) for NOMA

In contrast to wireless networks ranging from the initial 1G to the latest 5G, which were primarily designed to overcome challenges posed by unpredictable radio conditions such as signal fading and blockages, RISs enable the establishment of a "smart radio environment" to facilitate the realization of 6G in a flexible and sustainable manner [246]. Significantly, the two-dimensional configuration and nearly passive operational characteristics of RISs contribute to their exceptional compatibility with prevailing wireless technologies. Recently, a novel type of STARS was proposed, which integrates both transmission and reflection (T&R) functions into a single RIS and achieves full-space coverage of the smart radio environment. In the following, we will first give a brief introduction to STARS and then discuss the benefits of STARS-enhanced NOMA.

1) Basis of STARS: At its core, STARS is a periodic structure with adjustable T&R phase-shift coefficients. STARS builds upon the capabilities of RISs by introducing the capability to simultaneously transmit signals while reflecting them. This means that in addition to modifying the wireless channel by reflecting the incident signals, STARS can also let signals penetrate through the surface. By controlling the T&R coefficients of the reconfigurable elements, STARS can dynamically adjust the phase and amplitude of both the transmitted and reflected signals. The simultaneous transmissions and reflection capabilities of STARS offer several advantages in wireless communication systems. Firstly, they provide enhanced signal coverage and connectivity by simultaneously serving users on both sides of the surface. This enables improved communication links and increased network capacity. Secondly, STARS can enable cooperative communication and multi-user interference management. By intelligently controlling T&R coefficients, STARS can coordinate the signals of multiple users, manage interference, and optimize the overall system performance. Furthermore, STARS is compatible with existing wireless technologies, making it a promising candidate for integration into future wireless networks. It can be seamlessly integrated with other communication technologies, including NOMA and MIMO, to further enhance network performance. In terms of hardware modeling, each element of a STARS can be treated as a lumped circuit with electric and magnetic impedances. The basic hardware and channel models for STARS were proposed in [247]. By configuring the values of two complex-valued impedances, the T&R coefficients of each STAR element can be controlled in terms of the amplitude and phase-shift responses. In [248], three practical operating protocols for STAR-RISs are proposed, namely, energy splitting (ES), mode switching (MS), and time switching (TS). Moreover, for STARS with passive-lossless elements, the phase shifts of the T&R coefficients are coupled. In [249], a correlated T&R phase-shift model was proposed for STARS.

2) STARS enhanced NOMA: Compared to conventional transmitting/reflecting-only RISs, STARS provides more DoFs for facilitating the T&R NOMA. Recalling the fact that the performance gain of NOMA over OMA relies on the channel
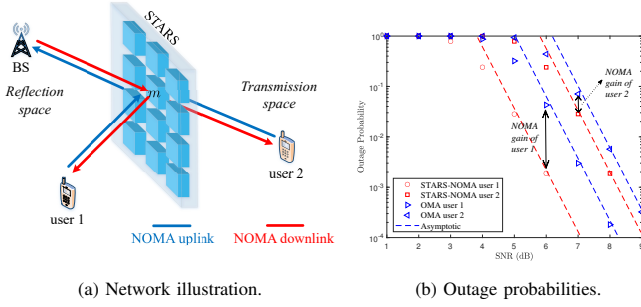
(a) Network illustration. (b) Outage probabilities.

Fig. 9: STARS-NOMA network and outage probabilities for NOMA users.

Fig. 10: Illustration of an end-to-end semantic communication framework.

condition disparity between paired users. The key idea of T&R NOMA is to pair a transmitted user and a reflected user together for NOMA transmission, as shown in Fig. 9a. The advantage of T&R NOMA is that by adjusting the T&R amplitude and phase-shift coefficients, a distinct channel difference can be achieved between users located at each side of STARS, which leads to high NOMA performance gain. Such a T&R NOMA operation addresses the issue when employing transmitting/reflecting-only RISs in NOMA, where the users surrounding the transmitting/reflecting-only RIS might have a similar channel condition. To further illustrate the benefits of T&R NOMA, Fig. 9b depicts the outage probabilities for STAR-NOMA and STAR-OMA users [249]. It can be observed that with the aid of STARS, NOMA outperforms OMA for both users' outage probabilities. Moreover, the performance gain of STAR-NOMA over STAR-OMA is more pronounced in the high signal-to-noise ratio (SNR) region.

3) Existing Works on STARS-NOMA: Motivated by the potential benefits brought by STARS for NOMA, increasing research interests have been devoted to the performance analysis and optimization of STARS-NOMA. The STARS-NOMA system was first proposed in [250] where the achievable sum rate was maximized by jointly optimizing the decoding order, power allocation coefficients, active beamforming, T&R beamforming. In terms of performance analysis of the STAR-NOMA network, various studies have been conducted to evaluate the bit error rate (BER) performance [251], the impact of the correlated T&R phase-shift [249], the ergodic rate for users over Rician fading channels [252], the outage probability of users over spatially correlated channels [253], and the performance of an uplink STARS-NOMA network [254]. Apart from these studies, recent works focused on STARS-NOMA networks with the more complex distribution of users and scatters in the environment. For example, In [255], a fitting method was proposed to model the distribution of STARS-NOMA channels composing small-scale fading power as the tractable Gamma distribution. Based on the proposed model, a unified analytical framework based on stochastic geometry was provided to capture the random locations of STARS, BSs, and users. In [256], the ergodic rate of the STARS-NOMA system was investigated, where the direct links from the BS to cell-edge users are non-line-of-sight due to obstacles, and STARS is used to provide line-of-sight links to these cell-edge
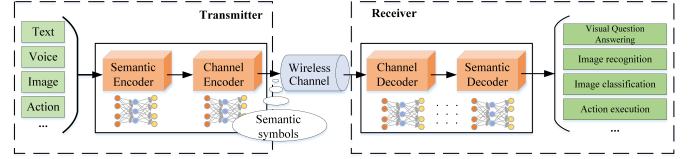
users.

In terms of optimization, existing works focused on the beamforming design and power allocation scheme for STARS-NOMA. In [257], the framework for applying NOMA to STARS networks was first proposed and a cluster-based beamforming design was adopted to jointly optimize the decoding order, power allocation coefficients, and beamforming at both the BS side and STAR-RIS side. In [258], a joint power and discrete amplitude allocation scheme was proposed for the STARS-NOMA system, which can reduce CE workload and hardware complexity. In addition, recall that one of the benefits of STARS is that it can be seamlessly integrated with other communication technologies. Exploiting STARS with other technologies is a major interest for the research community. For example, in [259], the concept of index modulation (IM) was first incorporated into the STARS-NOMA system to improve spectral efficiency. Specifically, the proposed IM-aided STARS-NOMA system enables extra information bits to be transmitted by allocating subsurfaces to different users in a pre-defined subsurface allocation pattern. In [260], [261], a new active STARS-NOMA network was investigated. In particular, simple active devices are introduced into the STARS to overcome the "double-fading" effect. Furthermore, STARS was also exploited in a MIMO-enabled NOMA system [262]. Among others, exploiting STARS-NOMA in secured communication is an appealing option. In [263], [264], the secrecy performance of STARS-NOMA systems over general cascaded $\kappa - \mu$ fading and Rayleigh fading channels was investigated.

E. Interplay between NOMA and Semantic Communications

The Shannon classical information theory guides the development of communication systems from 1G to 5G. Thanks to the great efforts of information and communication technology (ICT) researchers in the past few decades, many efficient communication technologies were proposed to address the technical-level communication problem [26] and thus approach the fundamental Shannon limit. However, the communication requirements in the upcoming 6G, such as extremely high data rate and massive connectivity, are far from being satisfied, which calls for a new information transmission paradigm to be developed. In this subsection, we first provide a brief overview of semantic communications and then discuss the interplay between semantic communications and NOMA.

1) Overview of Semantic Communications: Compared to conventional information transmission, which treats the source data equally into the bit sequences and ignores the intrinsic meaning, semantic communications are task- and goal-oriented. As shown in Fig. 10, in semantic communications,

only the information (conveyed by semantic symbols) which is related to the specific meaning/actions/goals required by the receiver should be transmitted [28], [31]. This is achieved by the semantic and channel encoders at the transmitter, which is empowered by advanced machine learning tools, e.g., deep learning, for the semantic feature extraction from the source data (e.g., text, voice, image, etc.). When receiving the semantic symbols through the wireless channel, the receiver employs the channel and semantic decoders to accomplish the required tasks and/or achieve the transmission goals (e.g., visual question answering [265], image recognition/classification [266]–[268], action execution [269], CSI acquisition [270]–[272], molecular communications [273], etc.). It can be observed that only the key information needed to perform the function or deliver the meaning needs to be transmitted in semantic communications, the required radio resources can be greatly reduced. Therefore, semantic communications are more energy-efficient and sustainable than conventional bit-level communications. More importantly, existing research contributions showed that the superiority of semantic communications over conventional bit-level communications is more significant in b ad wireless conditions, which enhances the reliability of communication systems. It can be observed that semantic communications provides a promising information transmission option for future wireless networks. In the following, we focus our attention on the interplay between NOMA and semantic communications in multi-user networks, namely, NOMA-enabled semantic communications and semantic communications-enhanced NOMA.

2) NOMA-enabled Semantic Communications: On the one hand, when employing semantic communications in multi-user networks, one of the most fundamental problems is how to develop efficient multiple access schemes to accommodate the new semantic users given the limited radio resources. As a first step, the authors of [274] studied the NOMA design for facilitating the heterogeneous semantic and bit multi-user communication, where a transmitter simultaneously serves one semantic user and one conventional bit user. Recalling the fact that the transceivers of semantic communications have to be jointly trained in advance, it is almost impossible for the conventional bit user to employ SIC to detect the semantic user's signal, which leads to a fixed *bits-to-semantics SIC ordering* in the heterogeneous transmission. To manage the interference received by the bit user and maximize the resource efficiency, a novel semi-NOMA scheme was proposed in [274]. The total available bandwidth is divided into one non-orthogonal sub-band for the heterogeneous transmission and one orthogonal sub-band for the bit-only transmission. By adjusting the bandwidth and power allocation at the transmitter, the proposed semi-NOMA can unify both conventional NOMA and OMA schemes and thus achieves the maximum semantic-versus bit rate region, i.e., being the optimal multiple access scheme for the considered heterogeneous semantic and bit multi-user communication. Moreover, the authors of [275] explored the employment of NOMA to support pure semantic multi-user communication, which is capable of serving multiple users with different modalities of data via semantic communications. The results showed that the proposed NOMA-enabled semantic multi-user communication has strong robustness and can achieve higher spectral and power efficiency.

3) Semantic Communications-enhanced NOMA: On the other hand, how to employ semantic communications to enhance the NOMA performance is another interesting research topic. One important observation from existing research is that for achieving the same transmission goal, semantic communications generally require fewer radio resources (e.g., transmit power and bandwidth) than conventional bit-level communications. Motivated by this, an opportunistic semantic and bit communication strategy was proposed in [276] for the secondary user in uplink NOMA. The key idea is that the secondary user can appropriately select a semantic or bit communication option to control the resulting co-channel interference when reusing the resource block of the primary NOMA user. The advantages of the proposed opportunistic semantic and bit communication strategy can be explained as follows. When the primary NOMA user has a high communication requirement (e.g., a high-resolution video user), the transmit power that can be used by the secondary NOMA user is strictly capped. In this case, employing bit-level communications at the secondary NOMA user cannot achieve a satisfactory communication performance due to the limited power budget. As a result, semantic communications comes to the rescue, which guarantees the performance of the secondary user in the low SNR regime. Otherwise, when the primary NOMA user has a low communication requirement (e.g., an IoT user), the secondary NOMA user can use the bit-level communication option to achieve the best performance using a sufficiently high power (i.e., high SNR regime). Fig. 11 provides simulation results to further demonstrate the benefits of the proposed opportunistic semantic and bit communication strategy. It can be observed that with the employment of semantic communications, the performance of the secondary NOMA user can be greatly improved, especially when the communication requirement of the primary NOMA user is high.

VI. Towards Learning-Based Next Generation Multiple Access: Challenges and Future Work

Model-driven methods take advantage of various domain knowledge in the problems of NGMA. However, the model-driven methods usually suffer from high computational complexity and are restricted by the modeling of domain knowledge. As a feasible alternative, data-driven methods learn from data to solve detection and estimation problems, providing a new way to improve performance and reduce computational complexity. The use of learning-based methods for NGMA still requires a lot of in-depth exploration. Some of the challenges and future work are discussed in this section.

A. Construction of Training Dataset

The foundation of data-driven methods is the source of training data. The performance of data-driven methods is highly dependent on the consistency between the training and testing sets. Most existing data-driven methods for NGMA use data generated according to some channel models in both training and evaluation. Given the historical CSI data accumulated in
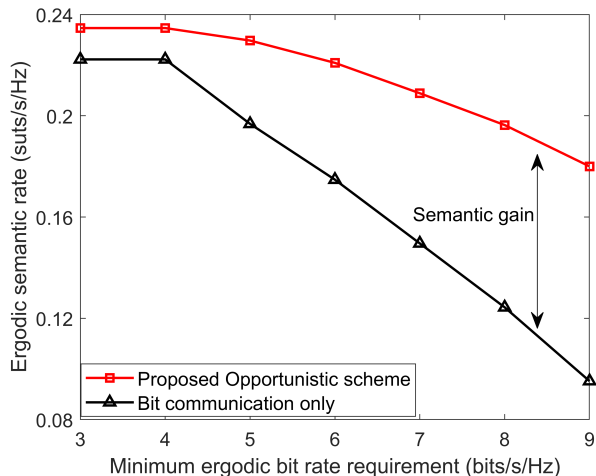
Fig. 11: The ergodic (equivalent) semantic rate achieved by the secondary NOMA user versus the ergodic bit rate required by the primary NOMA user over fading channels. The system parameter settings can be found in [276].

BS, it is possible to customize the algorithms for a specific cell by using real dataset of the cell. In situations where real data cannot be obtained, ray tracing could be employed to generate more realistic channel data. Other features beyond channel characteristics would be exploited in learning-based methods. For example, data-driven methods can learn the access probability of users, and use the prior information to enhance random access.

If only a small amount of real data is available, a potential approach is to use simulated data to train the learning model, and then use transfer learning to fine-tune the learning model on the real data. In practice, it is beneficial to use newly obtained data to update the model. Another way is to generate the dataset using generative AI methods, such as VAE and generative adversarial network.

B. Specialized Neural Network Design

The convolutional neural network explores structural information in images, and the long short-term memory network explores the relevance of sentences in natural language. Therefore, it is necessary to design specialized networks for different tasks in communication systems. In NGMA, domain knowledge should be exploited in the design of learning-based methods. A potential solution is to use deep unfolding methods, which construct the neural network according to some iterative optimization algorithm. The resulting neural network inherits the domain knowledge embedded in the traditional optimization problem.

Furthermore, computational complexity and storage requirement of the learning-based methods are strictly restricted in our tasks. Light-weight networks are desired, especially for low-cost devices. One approach to speed up neural network deployments, while reducing model storage size, is model pruning. Other techniques for model compression, e.g., quantization, knowledge distillation, and low-rank factorization,

would also be useful tools in building the specialized neural networks for our task.

C. Scalability and Generalization

Scalability and generalization are major concerns in the implementation of learning-based methods for NGMA. The existing work on machine learning for NGMA has shown reduced computational complexity in the inference stage, in comparison to the traditional optimization-based iterative methods. However, the training phase of learning-based methods is time-consuming. Given various communication scenarios, e.g., different environments, SNRs, number of active users and so on, applying one learning model for one specific scenario is costly in both model training and model storage. To relieve the burden in training a number of models, one could employ meta-learning, which is a process that helps models learn new and unseen tasks on their own with little effort. Furthermore, deviations between training and testing data can lead to a decrease in performance. For example, a model trained at one SNR may experience performance degradation at another SNR. Mixing training data with different settings may improve the robustness of data-driven methods. However, the ability of the network is limited, which can also lead to unsatisfactory training results. It would be interesting to investigate the mechanisms that can adjust the learning model to adapt to different scenarios.

VII. CONCLUSION

In this article, we have presented a comprehensive overview of the research efforts to date in signal processing and learning for NGMA, with a focus on MRA and NOMA. In particular, the fundamental limits and practical schemes are reviewed first, and then the state-of-the-art research contributions in the use of advanced signal processing and machine learning techniques are provided. We further discuss the promising interplay between NOMA and other new next-generation technologies including NFC, ISAC, STARS, and semantic communications. Finally, some challenges and future work in the use of learning-based methods for NGMA are listed. We believe that the in-depth overview of intelligent signal processing and learning for NGMA will provide inspiration for the development of new solutions for NGMA.

REFERENCES

[1] I. D. C. (IDC). (2021) Future of industry ecosystems: Shared data and insights. [Online]. Available: https://blogs.idc.com/2021/01/06/future-of-industry-ecosystems-shared-data-and-insights/
[2] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.
[3] ITU-R1 Working Party (WP) 5D, "The ITU-R framework for IMT-2030," 2023.
[4] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, 2021.
[5] T. Hou, J. Wang, Y. Liu, X. Sun, A. Li, and B. Ai, "A joint design for STAR-RIS enhanced NOMA-CoMP networks: A simultaneous-signal-enhancement-and-cancellation-based (SSECB) design," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 1043–1048, 2021.

[6] A. Khaleel and E. Basar, "A novel NOMA solution with RIS partitioning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 70–81, 2021.

[7] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 742–758, 2014.

[8] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. on Wireless Commun.*, vol. 19, no. 1, pp. 77–90, 2020.

[9] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017.

[10] Y. Han, B. D. Rao, and J. Lee, "Massive uncoordinated access with massive MIMO: A dictionary learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1320–1332, 2019.

[11] M. A. Aref and S. K. Jayaweera, "Deep learning-aided successive interference cancellation for MIMO-NOMA," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–5.

[12] J. Li, Z. Wu, P. Zhu, D. Wang, and X. You, "Scalable pilot assignment scheme for cell-free large-scale distributed MIMO with massive access," *IEEE Access*, vol. 9, pp. 122 107–122 112, 2021.

[13] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, 2016.

[14] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, 2017.

[15] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surv. Tut.*, vol. 22, no. 4, pp. 2192–2235, 2020.

[16] X. Xu, Q. Chen, X. Mu, Y. Liu, and H. Jiang, "Graph-embedded multi-agent learning for smart reconfigurable THz MIMO-NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 259–275, 2021.

[17] G. Yang, X. Xu, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3137–3151, 2021.

[18] X. Cao, B. Yang, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Han, H. V. Poor, and L. Hanzo, "Massive access of static and mobile users via reconfigurable intelligent surfaces: Protocol design and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1253–1269, 2022.

[19] X. Cao, B. Yang, C. Huang, C. Yuen, M. Di Renzo, Z. Han, D. Niyato, H. V. Poor, and L. Hanzo, "AI-assisted MAC for reconfigurable intelligent-surface-aided wireless networks: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 21–27, 2021.

[20] X. Cao, B. Yang, H. Zhang, C. Huang, C. Yuen, and Z. Han, "Reconfigurable-intelligent-surface-assisted MAC for wireless networks: Protocol design, analysis, and optimization," *IEEE Internet of Things J.*, vol. 8, no. 18, pp. 14 171–14 186, 2021.

[21] S. Yang, J. Zhang, Y. Zhu, S. Jin, and C. Yuen, "The application of distributed RIS to massive access miso systems: NOMA or OMA?" *IEEE Trans. Wirel. Commun.*, vol. 23, no. 5, pp. 4364–4378, 2024.

[22] Q. Zhang, J. Zhang, and S. Jin, "Grant-free random access in cell-free massive MIMO systems with UE detection thresholds: A stochastic geometry approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8117–8121, 2023.

[23] Z. Wang, Y. Liu, X. Mu, Z. Ding, and O. A. Dobre, "NOMA empowered integrated sensing and communication," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 677–681, 2022.

[24] C. Zhang, W. Yi, Y. Liu, and L. Hanzo, "Semi-integrated-sensing-and-communication (Semi-ISaC): From OMA to NOMA," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 1878–1893, 2023.

[25] X. Mu, Z. Wang, and Y. Liu, "NOMA for integrating sensing and communications towards 6G: A multiple access perspective," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 316–323, 2024.

[26] W. Weaver and C. E. Shannon, *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[27] D. Gunduz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting Bits: Context, semantics and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, 2023.

[28] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, 2023.

[29] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, 2022.

[30] J. Xu, B. Ai, W. Chen, N. Wang, and M. Rodrigues, "Deep joint source-channel coding for image transmission with visual protection," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 6, pp. 1399–1411, 2023.

[31] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[32] T. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ: Wiley-Interscience, 2006.

[33] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surv. Tut.*, vol. 24, no. 4, pp. 2073–2126, 2022.

[34] B. Clerckx, Y. Mao, E. A. Jorswieck, J. Yuan, D. J. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, 2023.

[35] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, and G. Caire, "Is NOMA efficient in multi-antenna networks? a critical look at next generation multiple access techniques," *IEEE Open J. of the Commun. Society*, vol. 2, pp. 1310–1343, 2021.

[36] S. P. Weber, J. G. Andrews, X. Yang, and G. de Veciana, "Transmission capacity of wireless Ad Hoc networks with successive interference cancellation," *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2799–2814, 2007.

[37] Y. Bai, W. Chen, F. Sun, B. Ai, and P. Popovski, "Data-driven compressed sensing for massive wireless access," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 28–34, 2022.

[38] Y. Bai, W. Chen, B. Ai, Z. Zhong, and I. J. Wassell, "Prior information aided deep learning method for grant-free NOMA in mMTC," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 112–126, 2022.

[39] W. Chen, B. Zhang, S. Jin, B. Ai, and Z. Zhong, "Solving sparse linear inverse problems in communication systems: A deep learning approach with adaptive depth," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 4–17, 2021.

[40] J.-M. Kang, I.-M. Kim, and C.-J. Chun, "Deep learning-based MIMO-NOMA with imperfect SIC decoding," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3414–3417, 2020.

[41] R. K. Senapati and P. J. Tanna, "Deep learning-based NOMA system for enhancement of 5G networks: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2022.

[42] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, 2018.

[43] W. Saetan and S. Thipchaksurat, "Application of deep learning to energy-efficient power allocation scheme for 5G SC-NOMA system with imperfect SIC," in *2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2019, pp. 661–664.

[44] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things," *IEEE Trans. Industr. Inform.*, vol. 15, no. 5, pp. 2995–3005, 2019.

[45] Y. Fu, W. Wen, Z. Zhao, T. Q. Quek, S. Jin, and F.-C. Zheng, "Dynamic power control for NOMA transmissions in wireless caching networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1485–1488, 2019.

[46] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2022.

[47] W. Chen, W. Wan, S. Wang, P. Sun, G. Y. Li, and B. Ai, "CSI-PPPNet: A one-sided one-for-all deep learning framework for massive MIMO CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7599–7611, 2024.

[48] W. Wan, W. Chen, S. Wang, G. Y. Li, and B. Ai, "Deep plug-and-play prior for multitask channel reconstruction in massive MIMO systems," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4149–4162, 2024.

[49] X. Cao, B. Yang, K. Wang, X. Li, Z. Yu, C. Yuen, Y. Zhang, and Z. Han, "AI-empowered multiple access for 6G: A survey of spectrum sensing, protocol designs, and optimizations," 2024.

[50] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.

[51] Y. Chen, A. Bayesteh, Y. Wu, B. Ren, S. Kang, S. Sun, Q. Xiong, C. Qian, B. Yu, Z. Ding *et al.*, "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 19–27, 2018.

[52] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[53] J. Yu and L. Chen, "Stability analysis of frame slotted Aloha protocol," *IEEE Trans. Mob. Comput.*, vol. 16, no. 5, pp. 1462–1474, 2017.

[54] J. Yu, P. Zhang, L. Chen, J. Liu, R. Zhang, K. Wang, and J. An, "Stabilizing frame slotted Aloha-based IoT systems: A geometric ergodicity perspective," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 714–725, 2021.

[55] J. Wang, J. Yu, X. Chen, L. Chen, C. Qiu, and J. An, "Age of information for frame slotted Aloha," *IEEE Trans. on Commun.*, vol. 71, no. 4, pp. 2121–2135, 2023.

[56] R. C. Yavas, V. Kostina, and M. Effros, "Gaussian multiple and random access channels: Finite-blocklength analysis," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 6983–7009, 2021.

[57] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[58] I. Zadik, Y. Polyanskiy, and C. Thrampoulidis, "Improved bounds on Gaussian MAC and sparse regression via Gaussian inequalities," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 430–434.

[59] Y. Bai, W. Chen, J. Chen, and W. Guo, "Deep learning methods for solving linear inverse problems: Research directions and paradigms," *Signal Processing*, vol. 177, p. 107729, 2020.

[60] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

[61] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receive," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, 2021.

[62] G. Ma, B. Ai, F. Wang, X. Chen, Z. Zhong, Z. Zhao, and H. Guan, "Coded tandem spreading multiple access for massive machine-type communications," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 75–81, 2018.

[63] T. T. T. Le, N. U. Hassan, X. Chen, M.-S. Alouini, Z. Han, and C. Yuen, "A survey on random access protocols in direct-access LEO satellite-based IoT communication," *IEEE Commun. Surv. Tutor.*, pp. 1–1, 2024.

[64] D. Tse and S. Hanly, "Multiaccess fading channels. i. polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, 1998.

[65] S. Hanly and D. Tse, "Multiaccess fading channels. ii. delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, 1998.

[66] E. MolavianJazi and J. N. Laneman, "A second-order achievable rate region for Gaussian multi-access channels via a central limit theorem for functions," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6719–6733, 2015.

[67] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[68] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, 1996.

[69] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[70] Y. Wu, A. Khisti, C. Xiao, G. Caire, K.-K. Wong, and X. Gao, "A survey of physical layer security techniques for 5G wireless networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 679–695, 2018.

[71] M. Ganji and H. Jafarkhani, "Interference mitigation using asynchronous transmission and sampling diversity," in *Proc. IEEE Globecom*, Dec. 2016.

[72] S. Verdu, "The capacity region of the symbol-asynchronous Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 35, no. 4, pp. 733–751, July 1989.

[73] S. Shao, T. Y., T. Kong, K. Deng, and Y. Shen, "Performance analysis of a modified V-BLAST system with delay offsets using zero-forcing detection," *IEEE Trans. Veh. Technol.*, vol. 56, no. 6, pp. 3827–3837, Nov. 2007.

[74] A. Das and B. Rao, "MIMO systems with intentional timing offset," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–14, Mar. 2011.

[75] S. Poorkasmaei and H. Jafarkhani, "Asynchronous orthogonal differential decoding for multiple access channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 481–493, Jan. 2014.

[76] M. Avendi and H. Jafarkhani, "Differential distributed space-time coding with imperfect synchronization in frequency-selective channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1811–1822, April 2014.

[77] M. Ganji, X. Zou, and H. Jafarkhani, "Asynchronous transmission for multiple access channels: Rate-region analysis and system design for uplink NOMA," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4364–4378, July 2021.

[78] X. Zou, B. He, and H. Jafarkhani, "An analysis of two-user uplink asynchronous non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 1404–1418, Feb 2019.

[79] J. Cui, G. Dong, S. Zhang, H. Li, and G. Feng, "Asynchronous NOMA for downlink transmissions," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 402–405, 2017.

[80] M. Ganji and H. Jafarkhani, "Time asynchronous NOMA for downlink transmission," in *IEEE Wireless Communications and Networking Conference (WCNC-19)*, Apr. 2019.

[81] ——, "Improving NOMA multi-carrier systems with intentional frequency offsets," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1060–1063, 2019.

[82] J. Son, H. Lee, I. Jung, and D. Hong, "Exploiting intentional frequency offset in NOMA-OFDM systems: From basic to practical," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 7087–7097, 2023.

[83] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[84] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[85] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[86] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, NOV 10 2009.

[87] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "A GAMP-based low complexity sparse Bayesian learning algorithm," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 294–308, 2018.

[88] W. Chen, "Simultaneous sparse Bayesian learning with partially shared supports," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1641–1645, 2017.

[89] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[90] H. F. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," in *2011 8th International Symposium on Wireless Communication Systems*. IEEE, 2011, pp. 291–295.

[91] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 389–400, 2013.

[92] Y. Bai, W. Chen, B. Ai, and Z. Zhong, "Contention based massive access scheme for B5G: A compressive sensing method," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 1854–1859.

[93] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, 2018.

[94] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[95] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. Jun, pp. 211–244, 2001.

[96] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[97] W. Chen and F. Xiao, "Optimized pilot design for joint compressive sensing multi-user and channel detection in massive MTC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–5.

[98] W. Chen, M. R. D. Rodrigues, and I. J. Wassell, "Projection design for statistical compressive sensing: A tight frame based approach," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2016–2029, 2013.

[99] Y. Bai, W. Chen, Y. Bai, and B. Ai, "Dictionary learning based channel estimation and activity detection for mMTC with massive MIMO," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 2272–2277.

[100] Y. Han, B. D. Rao, and J. Lee, "Massive uncoordinated access with massive MIMO: A dictionary learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1320–1332, 2020.

[101] H. Xiao, W. Chen, J. Fang, B. Ai, and I. J. Wassell, "A grant-free method for massive machine-type communication with backward activity level estimation," *IEEE Trans. Signal Process.*, vol. 68, pp. 6665–6680, 2020.

[102] Y. Bai, W. Chen, B. Ai, and Z. Zhong, "Contention-based nonorthogonal massive access with massive MIMO," *China Commun.*, vol. 17, no. 11, pp. 79–90, 2020.

[103] H. Xiao, B. Ai, and W. Chen, "A grant-free access and data recovery method for massive machine-type communications," in *ICC 2019 -2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[104] X. Rao and V. K. Lau, "Compressive sensing with prior support quality information and application to massive MIMO channel estimation with temporal correlation," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4914–4924, 2015.

[105] A. Ma, Y. Zhou, C. Rush, D. Baron, and D. Needell, "An approximate message passing framework for side information," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1875–1888, 2019.

[106] Q. Wang, L. Liu, S. Zhang, and F. C. M. Lau, "Exploiting temporal side information in massive IoT connectivity," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1432–1447, 2023.

[107] H. Djelouat, M. Leinonen, L. Ribeiro, and M. Juntti, "Joint user identification and channel estimation via exploiting spatial channel covariance in mMTC," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 887–891, 2021.

[108] R. B. Di Renna and R. C. de Lamare, "Activity-aware multiple feedback SIC for massive machine-type communications," in *SCC 2019; 12th International ITG Conference on Systems, Communications and Coding*. VDE, 2019, pp. 1–6.

[109] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, 2015.

[110] G. Ma, B. Ai, F. Wang, and Z. Zhong, "Joint design of coded tandem spreading multiple access and coded slotted ALOHA for massive machine-type communications," *IEEE Trans. Industr. Inform.*, vol. 14, no. 9, pp. 4064–4071, 2018.

[111] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6509–6533, 2020.

[112] A. Fengler, O. Musa, P. Jung, and G. Caire, "Pilot-based unsourced random access with a massive MIMO receiver, interference cancellation, and power control," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1522–1534, 2022.

[113] Y. Zhang, G. Song, Y. Li, Z. Zhang, Y. L. Guan, and C. Yuen, "Signal scrambling based joint blind channel estimation, activity detection, and decoding for massive random access," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2024.

[114] F. Tian, X. Chen, Y. L. Guan, and C. Yuen, "Design and analysis of massive uncoupled unsourced random access with Bayesian joint decoding," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 10 350–10 364, 2024.

[115] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, 2018.

[116] X. Xu, X. Rao, and V. K. Lau, "Active user detection and channel estimation in uplink CRAN systems," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 2727–2732.

[117] Y. Cheng, L. Liu, and L. Ping, "Orthogonal AMP for massive access in channels with spatial and temporal correlations," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 726–740, 2021.

[118] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive

[119] W. Chen, H. Xiao, L. Sun, and B. Ai, "Joint activity detection and channel estimation in massive MIMO systems with angular domain enhancement," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 2999–3011, 2022.

[120] Y. Zhang, P. Khanduri, I. Tsaknakis, Y. Yao, M. Hong, and S. Liu, "An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning," *IEEE Signal Process. Mag.*, vol. 41, no. 1, pp. 38–59, 2024.

[121] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–5.

[122] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, 2018.

[123] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, 2018.

[124] A. Rajoriya and R. Budhiraja, "Joint AMP-SBL algorithms for device activity detection and channel estimation in massive MIMO mMTC systems," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2136–2152, 2023.

[125] S. Daei and M. Kountouris, "Blind goal-oriented massive access for future wireless networks," *IEEE Trans. Signal Process.*, vol. 71, pp. 901–916, 2023.

[126] S. Daei, S. Razavikia, M. Kountouris, M. Skoglund, G. Fodor, and C. Fischione, "Blind asynchronous goal-oriented detection for massive connectivity," in *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2023, pp. 167–174.

[127] 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, 2016.

[128] W. Chen, D. Wipf, Y. Wang, Y. Liu, and I. J. Wassell, "Simultaneous Bayesian sparse approximation with structured sparse models," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6145–6159, 2016.

[129] Y. Mei, Z. Gao, Y. Wu, W. Chen, J. Zhang, D. W. K. Ng, and M. Di Renzo, "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1851–1869, 2021.

[130] X. Xie, Y. Wu, J. An, J. Gao, W. Zhang, C. Xing, K.-K. Wong, and C. Xiao, "Massive unsourced random access: Exploiting angular domain sparsity," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2480–2498, 2022.

[131] S. Haghighatshoar, P. Jung, and G. Caire, "Improved scaling law for activity detection in massive MIMO systems," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 381–385.

[132] Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu, "Phase transition analysis for covariance-based massive random access with massive MIMO," *IEEE Trans. Inf. Theory*, vol. 68, no. 3, pp. 1696–1715, 2021.

[133] ——, "Covariance based joint activity and data detection for massive random access with massiveMIMO," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[134] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, 2021.

[135] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection in multi-cell massive MIMO exploiting channel large-scale fading," *IEEE Trans. Signal Process.*, vol. 69, pp. 3768–3781, 2021.

[136] Z. Wang, Z. Chen, Y.-F. Liu, F. Sohrab, and W. Yu, "An efficient active set algorithm for covariance based joint data and activity detection for massive random access with massive MIMO," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4840–4844.

[137] A. Decurninge, I. Land, and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 552–556, 2020.

[138] ——, "Tensor decomposition bounds for TBM-based massive access," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2021, pp. 346–350.

[139] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420–435, 2019.

[140] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–5.

[141] A. E. Mostafa, V. W. Wong, Y. Zhou, R. Schober, Z. Luo, S. Liao, and M. Ding, "Aggregate preamble sequence design and detection for massive IoT with deep learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3800–3816, 2021.

[142] L. Diao, H. Wang, J. Li, P. Zhu, D. Wang, and X. You, "A scalable deep-learning-based active user detection approach for SEU-assisted cell-free massive MIMO systems," *IEEE Internet of Things J.*, vol. 10, no. 22, pp. 19 666–19 680, 2023.

[143] X. Shao, X. Chen, Y. Qiang, C. Zhong, and Z. Zhang, "Feature-aided adaptive-tuning deep learning for massive device detection," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1899–1914, 2021.

[144] Y. Ahn, W. Kim, and B. Shim, "Active user detection and channel estimation for massive machine-type communication: Deep learning approach," *IEEE Internet of Things J.*, vol. 9, no. 14, pp. 11 904–11 917, 2022.

[145] Z. Mao, X. Liu, M. Peng, Z. Chen, and G. Wei, "Joint channel estimation and active-user detection for massive access in internet of things—a deep learning approach," *IEEE Internet of Things J.*, vol. 9, no. 4, pp. 2870–2881, 2022.

[146] Y. Shi, H. Choi, Y. Shi, and Y. Zhou, "Algorithm unrolling for massive access via deep neural networks with theoretical guarantee," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 945–959, 2022.

[147] U. K. S. Shiv, S. Bhashyam, C. R. Srivatsa, and C. R. Murthy, "Learning-based sparse recovery for joint activity detection and channel estimation in massive random access systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2295–2299, 2022.

[148] J. Johnston and X. Wang, "Model-based deep learning for joint activity detection and channel estimation in massive and sporadic connectivity," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9806–9817, 2022.

[149] Y. Bai, W. Chen, Y. Ma, N. Wang, and B. Ai, "Dual-net for joint channel estimation and data recovery in grant-free massive access," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[150] Y. Cui, S. Li, and W. Zhang, "Jointly sparse signal recovery and support recovery via deep learning with applications in MIMO-based grant-free random access," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 788–803, 2021.

[151] Z. Ma, W. Wu, F. Gao *et al.*, "Model-driven deep learning for non-coherent massive machine-type communications," *arXiv preprint arXiv:2301.00516*, 2023.

[152] J. Huang, H. Zhang, C. Huang, L. Yang, and W. Zhang, "Noncoherent massive random access for inhomogeneous networks: From message passing to deep learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1457–1472, 2022.

[153] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, 2021.

[154] Y. Bai, W. Chen, B. Ai, and P. Popovski, "Deep learning for asynchronous massive access with data frame length diversity," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 5529–5540, 2024.

[155] W. Lee, M. Kim, and D.-H. Cho, "Deep learning based transmit power control in underlaid device-to-device communication," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2551–2554, 2019.

[156] ——, "Transmit power control using deep neural network for underlay device-to-device communication," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 141–144, 2019.

[157] Z. Sun, N. Yang, C. Li, J. Yuan, and T. Q. S. Quek, "Deep learning-based transmit power control for device activity detection and channel estimation in massive access," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 183–187, 2022.

[158] H. S. Jang, H. Lee, T. Q. S. Quek, and H. Shin, "Deep learning-based cellular random access framework," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7503–7518, 2021.

[159] C. Zhang, X. Sun, W. Xia, J. Zhang, H. Zhu, and X. Wang, "Deep learning based double-contention random access for massive machine-type communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1794–1807, 2023.

[160] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, 2017.

[161] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, 2016.

[162] M. R. G. Aghdam, B. M. Tazehkand, and R. Abdolee, "Joint optimal power allocation and beamforming for MIMO-NOMA in mmwave communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 938–941, 2022.

[163] G. Li, M. Zeng, D. Mishra, L. Hao, Z. Ma, and O. A. Dobre, "Energy-efficient design for IRS-empowered uplink MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9490–9500, 2022.

[164] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, 2016.

[165] M. Liu, J. Zhang, K. Xiong, M. Zhang, P. Fan, and K. B. Letaief, "Effective user clustering and power control for multiantenna uplink NOMA transmission," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8995–9009, 2022.

[166] 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, 2008.

[167] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *2010 IEEE 71st Vehicular Technology Conference*. IEEE, 2010, pp. 1–5.

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

[169] N. Y. Yu, "Binary golay spreading sequences and reed-muller codes for uplink grant-free NOMA," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 276–290, 2020.

[170] G. Millar, M. Kulhandjian, A. Alaca, S. Alaca, C. D'Amours, and H. Yanikomeroglu, "Low-density spreading design based on an algebraic scheme for NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 698–702, 2022.

[171] Y.-M. Chen, C. D. S. González, P.-H. Wang, and K.-P. Chen, "Reinforcement learning-based SCMA codebook design for uplink Rayleigh fading channels," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1717–1721, 2021.

[172] C. Du, J. Yu, R. Zhang, J. Ren, and J. An, "Orthcatter: High-throughput in-band OFDM backscatter with Over-the-Air code division," in *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. Santa Clara, CA: USENIX Association, Apr. 2024, pp. 1301–1314.

[173] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and 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, 2016.

[174] C. Liu, N. C. Beaulieu, J. Cheng, S. Wu, C. Jiang, and H. Yang, "An analysis of the error rate performance for uplink asynchronous signal detection in non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4591–4606, 2022.

[175] J. Liu, Y. Li, G. Song, and Y. Sun, "Detection and analysis of symbol-asynchronous uplink NOMA with equal transmission power," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1069–1072, 2019.

[176] V. Basnayake, D. N. K. Jayakody, H. Mabed, A. Kumar, and T. D. Ponnimbaduge Perera, "M-ary QAM asynchronous-NOMA D2D network with cyclic triangular-SIC decoding scheme," *IEEE Access*, vol. 11, pp. 6045–6059, 2023.

[177] M. Ganji, X. Zou, and H. Jafarkhani, "Exploiting time asynchrony in multi-user transmit beamforming," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3156–3169, 2020.

[178] J.-B. Kim and I.-H. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1949–1952, 2015.

[179] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686–1689, 2015.

[180] X. Zou, M. Ganji, and H. Jafarkhani, "Cooperative asynchronous non-orthogonal multiple access with power minimization under QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1503–1518, 2020.

[181] M. Zhao, N. Ye, Q. Ouyang, Y. Jin, Y. Jin, and L. Zhao, "Multi-satellite cooperative communication: Exploiting time asynchrony in non-orthogonal transmissions," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6868–6873, 2023.

[182] H. Jafarkhani, *Space-Time Coding: Theory and Practice*. Cambridge University Press, 2005.

[183] X. Liu and H. Jafarkhani, "Downlink non-orthogonal multiple access with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6151–6164, 2017.

[184] J. A. Oviedo and H. R. Sadjadpour, "Fundamentals of power allocation strategies for downlink multi-user NOMA with target rates," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1906–1917, 2020.

[185] P. Swami, M. K. Mishra, V. Bhatia, and T. Ratnarajah, "Performance analysis of NOMA enabled hybrid network with limited feedback," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4516–4521, 2020.

[186] X. Zou, M. Ganji, and H. Jafarkhani, "Downlink asynchronous non-orthogonal multiple access with quantizer optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1606–1610, 2020.

[187] M. Almasi and H. Jafarkhani, "Reconfigurable intelligent surface-aided NOMA with limited feedback," in *IEEE International Conference on Communications (ICC-23)*, May 2023.

[188] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, 2018.

[189] L. Jiang, X. Li, N. Ye, and A. Wang, "Deep learning-aided constellation design for downlink NOMA," in *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, 2019, pp. 1879–1883.

[190] T.-H. Vu, T.-V. Nguyen, D. B. da Costa, and S. Kim, "Performance analysis and deep learning design of underlay cognitive NOMA-based CDRT networks with imperfect SIC and co-channel interference," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8159–8174, 2021.

[191] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of things," *IEEE Trans. Industr. Inform.*, vol. 15, no. 5, pp. 2995–3005, 2019.

[192] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, "DeepNOMA: A unified framework for NOMA using deep multi-task learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2208–2225, 2020.

[193] H. Zhang, H. Zhang, K. Long, and G. K. Karagiannidis, "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2406–2415, 2020.

[194] Z. Ali, G. A. S. Sidhu, F. Gao, J. Jiang, and X. Wang, "Deep learning based power optimizing for NOMA based relay aided D2D transmissions," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 3, pp. 917–928, 2021.

[195] Y. Zhang, X. Wang, and Y. Xu, "Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.

[196] W. Ahsan, W. Yi, Z. Qin, Y. Liu, and A. Nallanathan, "Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5083–5098, 2021.

[197] J. Chen, L. Guo, J. Jia, J. Shang, and X. Wang, "Resource allocation for IRS assisted SGF NOMA transmission: A MADRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1302–1316, 2022.

[198] A. Nedić, J.-S. Pang, G. Scutari, Y. Sun, G. Scutari, and Y. Sun, "Parallel and distributed successive convex approximation methods for big-data optimization," *Multi-agent Optimization: Cetraro, Italy 2014*, pp. 141–308, 2018.

[199] H. J. Park, H. W. Kim, and S. H. Chae, "Deep-learning-based resource allocation for transmit power minimization in uplink NOMA IoT cellular networks," *IEEE Trans. Cogn. Commun.*, vol. 9, no. 3, pp. 708–721, 2023.

[200] J. Ling, J. Xia, F. Zhu, C. Gao, S. Lai, and V. Balasubramanian, "DQN-based resource allocation for NOMA-MEC-aided multi-source data stream," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, p. 44, 2023.

[201] J. Zheng, X. Tang, X. Wei, H. Shen, and L. Zhao, "Channel assignment for hybrid NOMA systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1370–1374, 2021.

[202] C. Chaieb, F. Abdelkefi, and W. Ajib, "Deep reinforcement learning for resource allocation in multi-band and hybrid OMA-NOMA wireless networks," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 187–198, 2023.

[203] H. Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977–2990, 2021.

[204] M. Kim, N.-I. Kim, W. Lee, and D.-H. Cho, "Deep learning-aided SCMA," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 720–723, 2018.

[205] Q. Luo, Z. Liu, G. Chen, Y. Ma, and P. Xiao, "A novel multitask learning empowered codebook design for downlink SCMA networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1268–1272, 2022.

[206] H. Yu, Z. Fei, Z. Zheng, and N. Ye, "Finite-alphabet signature design for grant-free NOMA: A quantized deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10 975–10 987, 2020.

[207] F. Jiang, Z. Gu, C. Sun, and R. Ma, "Dynamic user pairing and power allocation for NOMA with deep reinforcement learning," in *2021 IEEE Wireless Commun and Networking Conference (WCNC)*, 2021, pp. 1–6.

[208] M. Ma and V. W. S. Wong, "Joint user pairing and association for multicell NOMA: A pointer network-based approach," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[209] R. H. Y. Perdana, T.-V. Nguyen, and B. An, "Adaptive user pairing in multi-IRS-aided massive MIMO-NOMA networks: Spectral efficiency maximization and deep learning design," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 4377–4390, 2023.

[210] C. Li, Z. Gao, J. Xia, D. Deng, and L. Fan, "Cache-enabled physical-layer secure game against smart uAV-assisted attacks in b5G NOMA networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, pp. 1–9, 2020.

[211] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, 2017.

[212] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *2011 8th Annu. IEEE Commun. Society Conf. on Sensor, Mesh and Ad Hoc Commun. and Netw.*, 2011, pp. 350–358.

[213] Z. Sun, W. Chen, Y. Sun, X. Qin, T. Hou, L. Li, and B. Ai, "Deep learning-based grant-free massive access with age-of-information minimization," *arXiv preprint arXiv*, 2024.

[214] B. Yu, Y. Cai, and D. Wu, "Joint access control and resource allocation for short-packet-based mMTC in status update systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 851–865, 2021.

[215] T. Yang, J. Jiao, S. Wu, R. Lu, and Q. Zhang, "Grant free age-optimal random access protocol for satellite-based internet of things," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3947–3961, 2022.

[216] H. Zhao, H. Yu, Z. Zhang, M. Zeng, and Z. Fei, "Deep reinforcement learning for the joint aoi and throughput optimization of the random access system," in *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2022, pp. 695–700.

[217] O. T. Yavascan and E. Uysal, "Analysis of slotted ALOHA with an age threshold," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1456–1470, 2021.

[218] Y. Prathyusha and T.-L. Sheu, "Coordinated resource allocations for eMBB and URLLC in 5G communication networks," *IEEE Trans. on Veh. Technol.*, vol. 71, no. 8, pp. 8717–8728, 2022.

[219] G. Interdonato, S. Buzzi, C. D'Andrea, L. Venturino, C. D'Elia, and P. Vendittelli, "On the coexistence of eMBB and URLLC in multi-cell massive MIMO," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1040–1059, 2023.

[220] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Superposition-based URLLC traffic scheduling in 5G and beyond wireless networks," *IEEE Trans. on Commun.*, vol. 70, no. 9, pp. 6295–6309, 2022.

[221] J. Wang, C. Xing, and J. Liu, "Intelligent preamble allocation for coexistence of mMTC/URLLC devices: A hierarchical Q-learning based approach," *China Commun.*, vol. 20, no. 8, pp. 44–53, 2023.

[222] Y. Liu, B. Clerckx, and P. Popovski, "Network slicing for eMBB, URLLC, and mMTC: An uplink rate-splitting multiple access approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2140–2152, 2024.

[223] W. Fan, P. Fan, and Y. Long, "Joint delay-energy optimization for multi-priority random access in machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1416–1431, 2024.

[224] Y. Liu, Z. Wang, J. Xu, C. Ouyang, X. Mu, and R. Schober, "Near-field communications: A tutorial review," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1999–2049, 2023.

[225] C. Ouyang, Z. Wang, B. Zhao, X. Zhang, and Y. Liu, "On the impact of reactive region on the near-field channel gain," *IEEE Commun. Lett.*, vol. 28, no. 10, pp. 2417–2421, 2024.

[226] R. W. Bickmore, "On focusing electromagnetic radiators," *Can. J. Phys.*, vol. 35, no. 11, pp. 1292–1298, 1957.

[227] J. Sherman, "Properties of focused apertures in the Fresnel region," *IRE Trans. Antennas Propag.*, vol. 10, no. 4, pp. 399–408, Jul. 1962.

[228] J. Musil, "Properties of antennas focused in the Fresnel zone," *Czech. J. Phys.*, vol. 17, no. 10, pp. 874–888, 1967.

[229] E. Björnson, Ö. T. Demir, and L. Sanguinetti, "A primer on near-field beamforming for arrays and reconfigurable intelligent surfaces," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput.* IEEE, 2021, pp. 105–112.

[230] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, M. F. Imani, and Y. C. Eldar, "Beam focusing for near-field multiuser MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7476–7490, Sep. 2022.

[231] Z. Wu and L. Dai, "Multiple access for near-field communications: SDMA or LDMA?" *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1918–1935, 2023.

[232] J. Zuo, X. Mu, and Y. Liu, "Non-Orthogonal Multiple Access For Near-Field Communications," *arXiv preprint arXiv:2304.13185*, 2023.

[233] Z. Ding, R. Schober, and H. V. Poor, "NOMA-based coexistence of near-field and far-field massive MIMO communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 8, pp. 1429–1433, 2023.

[234] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges," *IEEE Netw.*, vol. 35, no. 5, pp. 158–167, Sep./Oct. 2021.

[235] F. Liu, W. Yuan, C. Masouros, and J. Yuan, "Radar-assisted predictive beamforming for vehicular links: Communication served by sensing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7704–7719, Nov. 2020.

[236] X. Liu, T. Huang, N. Shlezinger, Y. Liu, J. Zhou, and Y. C. Eldar, "Joint transmit beamforming for multiuser MIMO communications and MIMO radar," *IEEE Trans. Signal Process.*, vol. 68, pp. 3929–3944, Jun. 2020.

[237] C. Ouyang, Y. Liu, and H. Yang, "On the performance of uplink ISAC systems," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1769–1773, Aug. 2022.

[238] Z. Wang, X. Mu, Y. Liu, X. Xu, and P. Zhang, "NOMA-aided joint communication, sensing, and multi-tier computing systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 3, pp. 574–588, Mar. 2023.

[239] E. Memisoglu, H. Türkmen, B. A. Ozbakis, and H. Arslan, "CSI-based NOMA for integrated sensing and communication," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1086–1090, Jun. 2023.

[240] Z. Wang, X. Mu, Y. Liu, and Z. Ding, "Exploiting Sensing Signal in ISAC: A NOMA Inspired Scheme," *arXiv preprint arXiv:2201.04547*, 2022.

[241] X. Mu, Y. Liu, L. Guo, J. Lin, and L. Hanzo, "NOMA-aided joint radar and multicast-unicast communication systems," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1978–1992, Jun. 2022.

[242] Z. Yang, D. Li, N. Zhao, Z. Wu, Y. Li, and D. Niyato, "Secure precoding optimization for NOMA-aided integrated sensing and communication," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 8370–8382, Dec. 2022.

[243] C. Dou, N. Huang, Y. Wu, L. Qian, and T. Q. S. Quek, "Sensing-efficient NOMA-aided integrated sensing and communication: A joint sensing scheduling and beamforming optimization," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13 591–13 603, 2023.

[244] K. Chen, R. Li, L. Wang, L. Xu, and A. Fei, "Beamforming optimization for full-duplex NOMA enabled integrated sensing and communication," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2022, pp. 264–268.

[245] Y. Xiong, F. Liu, Y. Cui, W. Yuan, T. X. Han, and G. Caire, "On the fundamental tradeoff of integrated sensing and communications under Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 69, no. 9, pp. 5723–5751, Sep. 2023.

[246] M. Di Renzo, A. Zappone, M. Debbah, M. S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[247] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 3134–3138, May, 2021.

[248] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, 2022.

[249] J. Xu, Y. Liu, X. Mu, R. Schober, and H. V. Poor, "STAR-RISs: A correlated T&R phase-shift model and practical phase-shift configuration strategies," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 1097–1111, 2022.

[250] J. Zuo, Y. Liu, Z. Ding, and L. Song, "Simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[251] M. Aldababsa, A. Khaleel, and E. Basar, "STAR-RIS-NOMA networks: An error performance perspective," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1784–1788, 2022.

[252] X. Yue, J. Xie, Y. Liu, Z. Han, R. Liu, and Z. Ding, "Simultaneously transmitting and reflecting reconfigurable intelligent surface assisted NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 189–204, 2023.

[253] T. Wang, M.-A. Badiu, G. Chen, and J. P. Coon, "Outage probability analysis of STAR-RIS assisted NOMA network with correlated channels," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1774–1778, 2022.

[254] J. Xu, X. Mu, J. T. Zhou, and Y. Liu, "Simultaneously transmitting and reflecting (STAR)-RISs: Are they applicable to dual-sided incidence?" *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 129–133, 2023.

[255] Z. Xie, W. Yi, X. Wu, Y. Liu, and A. Nallanathan, "STAR-RIS aided NOMA in multicell networks: A general analytical framework with gamma distributed channel modeling," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5629–5644, 2022.

[256] B. Zhao, C. Zhang, W. Yi, and Y. Liu, "Ergodic rate analysis of STAR-RIS aided NOMA systems," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2297–2301, 2022.

[257] J. Zuo, Y. Liu, Z. Ding, L. Song, and H. V. Poor, "Joint design for simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 611–626, Jan. 2023.

[258] S. Yang, J. Zhang, W. Xia, H. Gao, and H. Zhu, "Joint power and discrete amplitude allocation for STAR-RIS-aided NOMA system," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 13 382–13 386, 2022.

[259] J. Zhu, P. Gao, G. Chen, P. Xiao, and A. Quddus, "Index modulation for STAR-RIS assisted NOMA system," *IEEE Commun. Lett.*, vol. 27, no. 2, pp. 716–720, 2023.

[260] J. Xu, J. Zuo, J. T. Zhou, and Y. Liu, "Active simultaneously transmitting and reflecting (STAR)-RISs: Modeling and analysis," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2466–2470, 2023.

[261] H. Luo, L. L. Lv, Q. Wu, Z. Ding, N. Al-Dhahir, and J. Chen, "Beamforming design for active IOS aided NOMA networks," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 282–286, 2023.

[262] F. Fang, B. Wu, S. Fu, Z. Ding, and X. Wang, "Energy-efficient design of STAR-RIS aided MIMO-NOMA networks," *IEEE Trans. on Commun.*, vol. 71, no. 1, pp. 498–511, 2023.

[263] Z. Xie, Y. Liu, W. Yi, X. Wu, and A. Nallanathan, "Physical layer security for STAR-RIS-NOMA in large-scale networks," *arXiv preprint arXiv:2304.06128*, 2023.

[264] S. Xu, C. Liu, H. Wang, M. Qian, and J. Li, "STAR-RIS-assisted scheme for enhancing physical layer security in NOMA systems," *IET Communications*, vol. 16, no. 19, pp. 2328–2342, 2022.

[265] J. Guo, W. Chen, Y. Sun, J. Xu, and B. Ai, "VideoQA-SC: Adaptive semantic communication for video question answering," 2024. [Online]. Available: https://arxiv.org/abs/2406.18538

[266] E. Bourtsoulatze, D. Burth Kurka, and D. Gunduz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019.

[267] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[268] L. Guo, W. Chen, Y. Sun, and B. Ai, "Digital-SC: Digital semantic communication with adaptive network split and learned non-linear quantization," 2024. [Online]. Available: https://arxiv.org/abs/2305.13553

[269] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Wireless Commun.*, vol. 70, no. 8, pp. 5181–5192, 2022.

[270] J. Xu, B. Ai, N. Wang, and W. Chen, "Deep joint source-channel coding for CSI feedback: An end-to-end approach," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 260–273, 2023.

[271] Y. Guo, W. Chen, F. Sun, J. Cheng, M. Matthaiou, and B. Ai, "Deep learning for CSI feedback: One-sided model and joint multi-module learning perspectives," 2024. [Online]. Available: https://arxiv.org/abs/2405.05522

[272] Y. Guo, W. Chen, J. Xu, L. Li, and B. Ai, "Deep joint CSI feedback and multiuser precoding for MIMO OFDM systems," 2024. [Online]. Available: https://arxiv.org/abs/2404.16289

[273] C. Yukun, C. Wei, and A. Bo, "Building semantic communication system via molecules: An end-to-end training approach," *China Commun.*, vol. 21, no. 7, pp. 113–124, 2024.

[274] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, 2023.

[275] W. Li, H. Liang, C. Dong, X. Xu, P. Zhang, and K. Liu, "Non-orthogonal multiple access enhanced multi-user semantic communication," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 6, pp. 1438–1453, 2023.

[276] X. Mu and Y. Liu, "Exploiting semantic communication for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, 2023.