ABSTRACT

As a prominent member of the Next Generation Multiple Access (NGMA) family, Non-Orthogonal Multiple Access (NOMA) has been recognized as a promising multiple access candidate for the Sixth-Generation (6G) networks. This article focuses on applying NOMA in 6G networks, with an emphasis on proposing the so-called One Basic Principle Plus Four New concept. Successive Interference Cancellation (SIC) importance becomes evident, starting with the basic NOMA principle. In particular, this article discusses the advantages and drawbacks of channel-state-information-based SIC and quality-of-service-based SIC. In addition, it explores applying NOMA to meet the new 6G performance requirements, especially for massive connectivity. Further, this article considers integrating NOMA with new physical layer techniques, followed by introducing new application scenarios for NOMA toward 6G. Finally, the article investigates applying machine learning in NOMA networks, ushering in the machine learning empowered NGMA era.

INTRODUCTION

According to the latest Visual Network Index report [1], the global mobile data traffic in 2022 will be seven-fold larger compared to 2017. In addition to this explosive data demand, emerging bandwidth-thirsty applications, for example, space-air-ground-integrated-networks (SAGINs), virtual reality (VR), augmented reality (AR), and so on, requiring strict quality-of-service (QoS) create further challenges for next-generation wireless networks.

To solve these challenges, the sixth-generation (6G) networks need breakthroughs beyond the current fifth-generation (5G) networks [2]. The expected performance targets of 6G are:

- The connectivity density is ten-fold larger, compared to 5G.
- The peak data rate reaches 1 terabit per second.
- The energy efficiency is a hundred times higher than that of 5G.
- The air interface latency decreases to 0.1 millisecond.
- The reliability increases to 99.99999 percent.

In addition, unlike 5G focusing on a single objective, most 6G scenarios need to optimize multiple objectives simultaneously [3]. To this end, highly efficient next-generation multiple access (NGMA) techniques are vital for 6G.

Due to introducing an additional degree of freedom in the power domain, non-orthogonal multiple access (NOMA) has attracted significant attention from both academia and industry. For decoding different signals in the same orthogonal (time/frequency) resource block (ORB), NOMA employs superposition coding at the transmitters and successive interference cancellation (SIC) at the receivers. Consequently, NOMA is able to provide additional access for overloaded traffic structures — a common scenario in 6G with heterogeneous ultra-dense networks [4]. Although NOMA has already been thoroughly investigated in the 5G and beyond networks, previous research focused on static devices and the data rate of broadband users. This ignores several fundamental problems for NGMA, for example, the effect of mobility, the design tradeoffs in terms of connectivity, reliability and latency.

Our goal in this article is to fill this gap by investigating NOMA via the One Basic Principle Plus Four New concept as illustrated in Fig. 1. More specifically, the basic principle is to deeply investigate the non-orthogonality from the SIC perspective, which is rarely discussed in existing NOMA-related magazines. Building on this basic principle, we explore the following four new directions:

- New requirements: Supporting massive connectivity by considering various QoS requirements, including latency and reliability
- New techniques: Integrating NOMA with other 6G physical layer techniques
- New applications: Application in heterogeneous scenarios to support emerging 6G applications
- New tools: Integration with artificial intelligence (AI) to design an adaptive resource allocation.

Basic Principle: Rethinking SIC in NOMA

Various forms of NOMA use the SIC process as the key component. For these conventional NOMA, SIC decoding has been designed by using a single metric, to ensure low implementation complexity. Take power-domain uplink NOMA as an example, where the channel state information (CSI) is used as a metric for deter-
mining SIC decoding order (i.e., users are distinguished by their channel conditions). Without loss of generality, consider a special case with a strong user and a weak user. The CSI-based SIC decodes the strong user’s signal first. If SIC’s first stage is successful, the base station (BS) then decodes the weak user’s signal. The CSI-based SIC is an intuitive decoding strategy; and it is able to effectively explore the users’ dynamic channel conditions. Another SIC implementation is based on the users’ QoS requirements, as in cognitive-radio inspired NOMA.

Consider a two-user case as an example. Assume that the primary user is an Internet-of-Things (IoT) sensor, and the secondary user is a broadband user. The QoS-based SIC detects the primary user first, because the data rate achievable during the first SIC stage can be extremely low, due to strong co-channel interference. Therefore, it is reasonable to first decode the signal of the sensor which is to be served with a low data rate.

While these conventional SIC decoding schemes can be implemented with low complexity, and have their own advantages in different application scenarios, it is important to point out that they are not optimal. For example, both the CSI- and QoS-based SIC decoding orders suffer from the same drawbacks (i.e., an outage probability error floor is inevitable), as explained in the following. For CSI-based SIC, at high signal-to-noise ratio (SNR), the strong user’s signal-to-interference-plus-noise ratio (SINR) becomes a constant. This means that the outage probability for the first SIC stage will never go to zero, regardless of how large the user’s transmit power is. Following a similar reasoning, one can also conclude that QoS-based SIC also suffers from this outage probability error floor — which cannot be avoided by simply increasing the users’ transmit power levels. Surprisingly, this inevitable floor can be avoided by using a simple SIC decoding strategy, termed hybrid SIC, as discussed in [5]. In particular, the hybrid SIC takes both the users’ QoS requirements and their channel conditions into consideration. Again, consider the simple case of one primary user and one secondary user as an example. Unlike CSI-based NOMA which compares the two users’ channel conditions, the hybrid SIC evaluates the quality of a user’s channel condition by taking the users’ QoS requirements into consideration. For example, if the secondary user’s channel is strong enough to guarantee the primary user’s QoS requirement, the secondary user’s signal is decoded during the first SIC stage. Otherwise, the primary user’s signal is decoded first. The performance gain of the hybrid SIC over conventional SIC schemes is shown in Fig. 2, which demonstrates that the outage probability error floor suffered by the two conventional schemes can be avoided. It is important to point out hybrid NOMA is still not optimal, and an important direction for future research is to design more sophisticated SIC for optimizing the reception reliability and data rates of NOMA transmission.

**NEW REQUIREMENTS: MASSIVE CONNECTIVITY FOR NOMA**

As a multiple access technique, the most important task for NGMA in 6G is to provide massive connectivity. Due to obtaining diverse QoS requirements in most 6G applications, the massive connectivity provided by NGMA needs to be constrained by different QoS requirements — in terms of user experience — which include not only data rates, but also latency and reliability. After that, NGMA is capable of maximizing the spectral efficiency of 6G networks. It is worth mentioning that a high spectral efficiency also contributes to energy conservation for 6G. To satisfy these performance metrics, and hence guarantee massive connectivity, the design of downlink NOMA becomes QoS-oriented, and semi-grant-free (semi-GF) transmissions are preferable for uplink NOMA.

**QoS-BASED NOMA FOR DOWNLINK TRANSMISSION**

Note that multiple SIC iterations introduce long waiting times and high outage probabilities. For downlink NOMA in 6G, the main challenge is to overcome high latency and low reliability due to SIC, especially for the massive users scenario. In addition to applying the hybrid SIC for enhancing the spectral efficiency, QoS-based NOMA needs to design principles of QoS-based user clustering (Q-UC) and QoS-based power allocation (Q-PA) — in terms of connectivity.

For Q-UC, two fundamental questions should be answered:

- Which is better, grouping the same or different QoS users in a NOMA cluster?
- How can SIC be avoided for latency-reliability-
FIGURE 3. Connectivity performance of semi-GF NOMA with 10 ORBs, where each NOMA cluster supports a maximum of two users [7, 8].

For the single-ORB case, it can be assumed that the ORB for each GF user is pre-allocated; accordingly, the number of potential GF users (PGFUs) in each ORB is known in advance. If the GB user is delay-tolerant, several SIC iterations are acceptable. Consequently, the broadcast threshold contains two kinds of information:
• Minimal transmit power (MTP)
• Maximal tolerance to interference (MTI).
If GF users have transmit power higher than MTP, they upload messages, and the successful GF users are decoded before the GB user. To avoid frequent collisions, the MTP-threshold is suitable for GF users with a relatively low activation rate. If GF users have transmit power lower than MTI, they are also able to upload messages that are decoded after the GB user. Note that the MTI needs to take all PGFUs into account to avoid the transmission outage of the GB user. This restriction can be relaxed when the GB user is based on long-term communications. In this case, the MTI only needs to consider the average number of active GF users, which is significantly smaller than the total number of PGFUs. If the GB user is delay-sensitive, it needs to be decoded firstly, so the threshold only contains the MTI. For the multi-ORB case, the number of PGFUs for each ORB is unknown, since GF users are capable of choosing ORBs at random, resulting in a complicated threshold design. To reduce collisions, it is better to control the Semi-GF NOMA transmit power. Based on channel gains, a power pool design with layered transmit power levels [7] can be applied to decrease outage probabilities. By restricting the number of GF users in each ORB, an access class barring (ACB) [8] aided user barring approach can be utilized to further enhance the connectivity. However, the protocol design of semi-GF NOMA under multi-ORB cases still faces numerous challenges. As illustrated in Fig. 3, the connectivity of semi-GF NOMA with ideal power pool and user barring is the highest. Without controlling the transmit power, semi-GF NOMA experiences more collisions than traditional GB transmission.

NEW TECHNIQUES: INTEGRATION OF NOMA WITH EMERGING PHYSICAL LAYER TECHNIQUES

In 6G, numerous emerging techniques for the physical layer will be gradually developed to address fundamental problems in future networks. For example, the orthogonal-time-frequency-space (OTFS) modulation aims to enhance the performance in a challenging communication scenario with high mobility. Terahertz (THz) communications are able to provide multi-GHz bandwidth to address the lack of spectrum resources. The integration of NOMA with OTFS, THz, and other promising techniques is an exciting research challenge for 6G.

OTFS-NOMA

Channels for traditional slow-mobile (or static) users in a coherence time can be regarded as invariable. With the aid of orthogonal frequency-division multiplexing (OFDM), communication performance can be evaluated in the time frequency plane with ORBs. However, delay...
and Doppler effects mainly decide fast-mobile users’ communication quality. It is worth pointing out that the traditional OFDM subcarriers in the delay-Doppler plane are not orthogonal, introducing strong inter-symbol interference and inter-carrier interference for fast-mobile users. To solve this problem, a new modulation scheme, referred to as OTFS [9], was proposed to provide nearly non-fading channels for these doubly dispersive channels.

Since the channel state for fast-mobile users is commonly worse than that of slow-mobile users, NOMA can be an efficient technique by exploiting this channel disparity. In this part, we discuss a general scenario that the slow-mobile user has a stronger channel than the fast-mobile user. In OTFS-NOMA systems, one slow-mobile user and one fast-mobile user can be grouped into a NOMA cluster. As the slow-mobile user has a strong channel, the SIC process is carried out by this user. Unlike conventional NOMA systems, the SIC at the slow-mobile user has two unique stages:
• The slow-mobile user first decodes the fast-mobile user’s signal in the delay-Doppler plane.
• After cancelling the decoded signal in the first stage from the received signal, the slow-mobile user decodes its signal in the time-frequency plane.

For the fast-mobile user, it decodes its own signal directly in the delay-Doppler plane. The proposed OTFS-NOMA scheme benefits both the fast- and slow-mobile users in terms of latency and spectral efficiency. More specifically, fast-mobile users obtain additional ORBs to enhance performance. Slow-mobile users have the opportunity to access spectrum resources solely allocated to fast-mobile users, under the conventional OTFS modulation, with orthogonal multiple access (OMA).

**THz-NOMA**

THz communication is a highly-promising technology for 6G, whose bandwidth is three to four orders of magnitude larger than current wireless systems. Due to its short wavelength, super massive multiple-input multiple-output (SM-MIMO) becomes an indispensable component for THz communications to exploit spatial diversity. In general, if the number of users with independent channels is much smaller than the number of antennas, the channels between different users are nearly orthogonal. In this case, NOMA is not a more efficient multiple access technique than traditional solutions for multiuser MIMO (MU-MIMO) systems. However, the low-rank channels of THz users are highly correlated because of the limited-scattering transmission. Therefore, NOMA becomes a promising technique to improve the spectral efficiency in THz aided networks.

For THz communications, NOMA aims to group all highly correlated users into one NOMA cluster; and allocate different transmit power levels to them for decoding, according to the hybrid SIC, as mentioned in earlier. After that, each NOMA cluster has one beam and different NOMA clusters are served by MU-MIMO techniques.

As shown in Fig. 4, since the line-of-sight link is the most important transmission path in THz communications, grouping users located in the same geographic area into one cluster becomes possible. This user clustering design has low complexity since it only needs the full CSI of one reference user (i.e., cluster head), in each NOMA cluster. A recent work [10] proposed a machine learning aided solution for THz-NOMA, based on a hybrid precoding scheme with a sub-connection structure, where the K-means machine learning algorithm is applied for user clustering; and a distributed alternating direction method of multipliers algorithm is applied for power allocation. Based on this solution, the energy efficiency of THz-NOMA systems can be maximized. This interesting work can be further extended by considering location-based user clustering to reduce the channel estimation complexity.

**Integration of NOMA with Other Techniques**

In addition to OTFS-NOMA and THz-NOMA, other promising physical layer techniques will also benefit from NOMA. Some typical use cases for NGM systems in 6G are listed as follows.

**Integrated Sensing and Communication (ISaC):** NOMA is capable of helping receivers to distinguish communication and radar-return signals. The communication signal can be decoded first and cancelled from the received signal via SIC. Then, the radar signal can be analyzed via conventional sensing algorithms.

**Visible Light Communication (VLC):** NOMA is able to provide high spectral efficiency for VLC to solve the small modulation bandwidth problem of light-emitting diodes, especially for multi-user cases. NOMA also benefits from the high SNR that VLC offers.

**Index Modulation (IM):** NOMA is capable of handling the strong inter-user interference issue in IM. The users with similar spatial features can be grouped into one NOMA cluster, hence increasing both spectral efficiency and energy efficiency.

**New Scenarios: Application of NOMA to Heterogeneous Scenarios**

6G networks are expected to be transformative and revolutionary encompassing applications—like data-driven, ubiquitous wireless connectivity, and connected intelligence. Although we are still in the early stages of defining 6G networks, some scenarios with heterogeneous traffics are recognized as potential candidates toward 6G. This section discusses the application of NOMA by considering the basic principle to new scenarios toward 6G [11].
**FIGURE 5.** Application of NOMA in scenarios toward 6G.

**NOMA IN INTEGRATED TERRESTRIAL AND AERIAL NETWORKS**

SAGINs, intrinsically integrating unmanned aerial vehicle (UAV)-aided terrestrial networks with sky-platforms and satellites, have become a focal point in wireless communications research, as they address a wide range of challenges encountered in single-tier networks. As shown in Fig. 5, based on the cooperation of high-altitude platforms (HAPs) and low altitude platforms (LAPs), drones, aircrafts and satellites, the conceived multi-tier SAGINs are capable of filling the coverage-holes of single-tier networks, by integrating the advantages of each network. In contrast to terrestrial wireless networks, the challenging optimization problems encountered in SAGINs have to provide massive connectivity under multiple objectives (delay, throughput, bit error rate (BER), power) to arrive at an attractive solution. For example, in the case of low-earth orbit systems, a large number of satellites are needed to cover the globe. SAGINs rely on the seamless integration of heterogeneous network segments with the goal of providing uninterrupted and ubiquitous connectivity to everyone, everything and everywhere. As mentioned earlier, the hybrid SIC-enabled NOMA can be applied in SAGINs — to enhance not only data rates, but also delay and BER — since hybrid SIC is promising to meet heterogeneous QoS requirements in multi-tier SAGINs, while guaranteeing ubiquitous and massive connectivity.

**NOMA IN RECONFIGURABLE-INTELLIGENT-SURFACE-ENHANCED NETWORKS**

Owing to their ability of smartly reconfiguring the wireless propagation environment via passive reflecting elements, reconfigurable intelligent surfaces (RISs) are capable of providing potential received SINR enhancements, by constructing a software-defined wireless environment. Compared to OMA-RIS networks, NOMA-RIS networks can achieve higher spectral efficiency and enhanced massive connectivity. More importantly, NOMA-RIS networks are more flexible than the conventional power-domain NOMA-based networks — because RISs can improve or degrade the channel quality of individual users, by adjusting the phase shifts of reflecting elements and their positions. In turn, adjusting these phase shifts smartly controls the users’ channel conditions.

Additionally, by integrating RISs in MIMO-RIS networks, the strict constraint on the number of antennas at the transmitters and receivers can be relaxed, due to the multiple reflecting elements on RISs. For RIS-enabled networks, traditional CSI-based SIC brings high implementation complexity — since perfect CSI for cascaded channels at the transmitter is desired. Fortunately, QoS-based SIC can be used to relax this limit. By integrating hybrid-SIC into NOMA-RIS, networks are able to handle the tradeoff between channel estimation complexity and high spectral efficiency.

**NOMA IN ROBOTIC COMMUNICATIONS/AUTONOMOUS NETWORKS**

Autonomous robotics have brought tremendous changes in various socio-economic aspects in our society, and have attracted significant attention in the wireless communications field. Autonomous robots are classified into three categories: aerospace robots (e.g., airship/airplane, HAP and UAV); ground robots (e.g., autonomous vehicles, smart home robots and mobile robots); marine robots (e.g., unmanned ships, unmanned submarines, underwater robots). Before fully reaping the benefits of autonomous robotics, challenges such as delay sensitivity and collision tolerance have to be tackled. With the aid of NOMA, the communication signal is designed to be ultra-reliable and low-latency, to fit the requirement of autonomous robotic systems. Additionally, since robots often serve in dynamic scenarios, the channel conditions of robots are varying at each timeslot, also improving the design flexibility of NOMA-based networks. By dynamically assigning the robots and other mobile users with heterogeneous mobility states into the same NOMA cluster to activate hybrid-SIC-based OTSF-NOMA schemes, the
received SINR of both robots and mobile users can be enhanced. Accordingly, NOMA can be helpful for the efficiency and safety of autonomous robotics.

**NOMA in VR/AR Multi-Layer Video Transmission**

VR and AR provide a better experience for users because of their immersive scenes, and have been recognized as one of the key scenarios in 6G networks. In the conventional OMA schemes, resources are allocated by blocks of a fixed size. It is non-trivial to divide them into small sizes, meaning that OMA users have to wait for access to resources. However, in VR and AR multi-layer video transmission networks, services to users are often delay-sensitive — while high transmit rate and reliable communications are highly demanded — highlighting the importance of high-spectral efficiency in wireless communication. Since hybrid-SIC enabled NOMA is capable of cancelling the outage probability error floor of primary users for providing a reliable transmission, applying hybrid-SIC enabled NOMA in VR and AR multi-layer video transmission networks is necessary.

**NOMA in E-Health Systems**

Due to its benefits of supporting enhanced spectrum efficiency and massive connectivity, NOMA has attracted increased attention for emerging remote e-Health services in recent years. Besides the power-domain NOMA, the QoS-based NOMA scheme can also be applied in communications-aware E-health systems. The QoS-based SIC scheme can be applied to first detect the inquired user’s (delay-sensitive) signal, and then decode the monitored user’s (delay-tolerant) information. Consequently, the precious communication resources can be efficiently utilized. Since the number of smart devices (e.g., smart watches) in communications-aware E-health systems has been growing exponentially, hybrid-SIC enabled NOMA is helpful for providing massive connectivity with strict QoS requirements in E-health systems. Further, it provides additional access channels for offloading heavy computation tasks to edge devices or edge servers.

---

**NEW TOOLS: MACHINE LEARNING-EMPowered NOMA-BASED NETWORKS**

Since scenarios toward 6G networks are heterogeneous, dynamic and more complex, new approaches are needed to tackle more challenging problems in these scenarios [12]. Most conventional approaches adopted in the current research contributions have a limited number of controllable variables and fail to optimize Markovian problems efficiently. Most conventional approaches adopted in the current research contributions have a limited number of controllable variables and fail to optimize Markovian problems efficiently. Moreover, comparing conventional NOMA scheme with single-objective (CSI or QoS) user clustering and power allocation, the hybrid SIC (as discussed earlier) introduces more controllable variables — as it has multi-objective (CSI and QoS) user clustering and power allocation. Therefore, machine learning becomes a competitive candidate for designing NGMA. In contrast to conventional algorithms, existing AI-based algorithms are capable of addressing problems in a more general framework, with a large number of controllable variables and temporal correlated events.

The general advantages/disadvantages of applying machine learning approaches in NOMA-based networks are listed in Table 1. Additionally, machine learning (ML) empowers NOMA-based networks to realize the interaction between BSs and the dynamic environment. More importantly, in the conventional solutions, the control policy aims for instantaneously achieving the current benefits for networks — without considering the long-term network evolution — the objective of ML algorithms. This section introduces the role of ML in NOMA-based networks.

**REinFORCEMENT LEARNING for NOMA-BASED NETWORKS**

Reinforcement learning (RL) models enable BSs/ access points (APs) to learn from the real-time feedback of dynamic/uncertain environment and mobile users, as well as from their historical experiences [13]. Similarly, in RL-empowered/ NOMA-based networks, BSs/APs are capable of

### Table 1: Pros and cons of machine learning algorithms in NOMA-based networks.

| Environment       | Conventional Solutions                                                                 | ML-Empowered Solutions                                                                 |
|-------------------|----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Non-Interactive   | Ignore User behavior/peculiarity                                                       | User behavior/peculiarity considered                                                   |
| Interactive       | Cannot adapt to the dynamic environment                                                | Rapidly adapting to the dynamic environment (Dynamic NOMA user re-clustering and resource allocation) |
| Resource Allocation | Focus on benefits of the current timeslot                                               | Can incorporate farsighted system evolution                                            |
| Long-term         |                                                                                   |                                                                                        |
| Resource Allocation |                                                                                  |                                                                                        |
| Short-term        |                                                                                   |                                                                                        |
| User Mobility     |                                                                                   |                                                                                        |
| Static User       |                                                                                   |                                                                                        |
| Dynamic User      |                                                                                   |                                                                                        |
| User Behavior     |                                                                                   |                                                                                        |

1. QL represents Q-learning, DQN is short for Deep Q-Network, DDPG is the acronym for deep deterministic policy gradient-based algorithms.
In FL-enabled, NOMA-based networks, BSs act as distributed learners — training their generated data and transferring their local model parameters, instead of the raw training dataset to an aggregating unit. Moreover, based on well-trained DL and DRL, transfer learning can be used to update the full/partial neural network parameters — for fast convergence in a new environment, with similar tasks — but limited data or training time. Rapidly adapting their control policy to the dynamic environment and keep on improving their decision-making ability. Application examples of RL in NOMA-based networks include the following.

**Dynamic User Clustering/Pairing:** The optimal policies for both CSI-based and QoS-based user clustering in NOMA-based networks are challenging, especially when user mobility is considered. By adopting RL-based algorithms for user re-clustering, the performance of such networks is improved in dynamic scenarios.

**Long-term Resource Allocation:** RL aims at achieving long-term benefits by deciding the resource allocation (e.g., power resource, caching resource, computing resource, spectrum resource) policy in NOMA-based networks. RL-based algorithms are capable of outperforming the conventional algorithms in dynamic scenarios, by interacting with the environment. However, since the optimality of RL-based algorithms cannot be strictly guaranteed, their superiority is reduced in some simple static scenarios.

RL models are suitable for Markovian problems, which means that the formulated problem in NOMA-based networks has to be reformulated as a Markovian one.

**Deep Learning for NOMA-Based Networks**

Deep learning (DL) uses cascaded neural network layers to extract features from the input data automatically and make a decision. DLs have shown great potential to revolutionize NOMA-based networks [14]. Application examples of DL in such networks include the following.

**CSI Acquisition/Channel Estimation:** As introduced above, CSI-based SIC is a fundamental principle of power-domain NOMA. However, acquiring CSI is challenging, especially when integrating NOMA with other techniques, such as MIMO and RIS. With the aid of DL, the CSI can be acquired automatically by extensive training of the input data using existing channel models. Additionally, DL can also be applied for detecting/estimating the dynamically fluctuating channel.

**Resource Allocation:** Since the performance of NOMA-based networks is significantly affected by the resource allocation policy, resource management is a principal problem in such networks. However, resource allocation under dynamic channel conditions is challenging. Thus, DL is attractive for dynamic resource allocation in NOMA-based networks, including power allocation, subchannel assignment, and subcarrier assignment.

DL-based algorithms are also attractive for solving other problems in such networks, such as signal constellation design, signal detection, and signal decoding. Although DL has been proved to be efficient, and to provide effective performance improvement, the systematic DL architecture design (e.g., the number of layers, the connections between layers) to reduce the computational complexity is still an open problem.

**Other Machine Learning for NOMA-Based Networks**

In addition to DL-based and RL-based based algorithms, unsupervised learning algorithms can be also adopted for user clustering. By exploiting the correlation feature of users’ channel responses, the K-means algorithm can be adopted for NOMA user clustering. By learning the inherent structures and correlations of users, the expectation maximization algorithm can be used for user clustering, in both fixed and dynamic user scenarios. Federated learning (FL), relying on directly training statistical models on remote devices, at the edge of distributed networks, is very attractive to preserve users’ data privacy in wireless networks [15]. NOMA is able to reduce the aggregation latency in FL model updating, by providing multiple access in the same channel. In FL-enabled, NOMA-based networks, BSs act as offloading learners — training their generated data and transferring their local model parameters, instead of the raw training dataset to an aggregating unit. Moreover, based on well-trained DL and DRL, transfer learning can be used to update the full/partial neural network parameters — for fast convergence in a new environment, with similar tasks — but limited data or training time.

**Conclusion and Outlook**

In this article, we described our view on the advances of NOMA for 6G based on the One Principle Plus Four New concept. As a core principle, the hybrid-SIC generalized CSI-based decoding order to a joint CSI- and QoS-based one, is able to remove the conventional outage probability error floor for joining users. Building on this principle, four new directions — including massive connectivity, the integration with other physical layer techniques, the application to heterogeneous scenarios, and the AI enhancement of NOMA, are presented in detail. These discussions demonstrate that NOMA has tremendous benefits for the NGMA family, especially for networks with distinct QoS requirements.

**Acknowledgment**

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), U.K., under Project EP/W035588/1.

**References**

[1] Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2017–2022 white paper,” Dec. 2019.

[2] P. Yang et al., “6G Wireless Communications: Vision and Potential Techniques,” IEEE Network, vol. 33, no. 4, July 2019, pp. 70–75.

[3] G. Gui et al., “6G: Opening New Horizons for Integration of Comfort, Security, and Intelligence,” IEEE Wireless Commun., vol. 27, no. 5, Oct. 2020, pp. 126–32.

[4] L. Liu et al., “Tractable Coverage Analysis for Hexagonal Macrocell-Based Heterogeneous UDNs With Adaptive Interference-Aware Comp.,” IEEE Trans. Wireless Commun., vol. 18, no. 1, Jan. 2019, pp. 503–17.

[5] Z. Ding, R. Schober, and H. V. Poor, “Unveiling the Importance of SIC in NOMA Systems—Part 1: State of the Art and Recent Findings,” IEEE Commun. Lett., vol. 24, no. 11, Nov. 2020, pp. 2373–77.

[6] Z. Ding et al., “Simple Semi-Grant-Free Transmission Strategies Assisted by Non-Orthogonal Multiple Access,” IEEE Trans. Commun., vol. 67, no. 6, June 2019, pp. 4464–78.

[7] J. Choi, “NOMA-Based Random Access With Multi-Channel Aloha,” IEEE SAC, vol. 35, no. 12, Dec. 2017, pp. 2736–43.

[8] 3GPP TS 22.201, “Technical Specification Group Services and System Aspects,” 3GPP, Tech. Rep., Release 13, 2016.

[9] R. Hadani et al., “Orthogonal Time Frequency Space Modulation,” Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[10] H. Zhang et al., “Energy Efficient User Clustering, Hybrid Precoding and Power Optimization in Terahertz MIMO-NOMA Systems,” IEEE JSAC, vol. 38, no. 9, Sept. 2020, pp. 2074–85.

[11] H. Zhang et al., “Energy-Efficient Resource Allocation in NOMA Heterogeneous Networks,” IEEE Wireless Commun,
[12] W. Saad, M. Bennis, and M. Chen, “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems,” IEEE Network, vol. 34, no. 3, Oct. 2020, pp. 134–42.
[13] K. B. Letaief et al., “The Roadmap to 6G: AI Empowered Wireless Networks,” IEEE Commun. Mag., vol. 57, no. 8, Aug. 2019, pp. 84–90.
[14] Q. Mao, F. Hu, and Q. Hao, “Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey,” IEEE Commun. Surv. Tutor., vol. 20, no. 4, 4th Qtr. 2018, pp. 2595–2621.
[15] S. Niknam et al., “Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges,” IEEE Commun. Mag., vol. 58, no. 6, June 2020, pp. 46–51.

**Biographies**

**Yuanwei Liu** [SM’19] (yuanwei.liu@qmul.ac.uk) is a senior lecturer (associate professor) at Queen Mary University of London. His research interests include 5G/6G networks, NOMA, RIS, UAV communications, and machine learning. He is a Web of Science Highly Cited Researcher 2021. He is currently a Senior Editor of IEEE Communications Letters, and an Editor of IEEE Trans. Wireless Communications and IEEE Trans. Communications. He received the IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020. He also received the 2020 IEEE Signal Processing and Computing for Communications Technical Early Achievement Award; and the IEEE Communication Theory Technical Committee 2021 Early Achievement Award. In addition, he received the IEEE ComSoc Young Professional Outstanding Nominee Award in 2021.

**Wenqiang Yi** [M’20] (w.yi@qmul.ac.uk) has been a postdoctoral researcher at Queen Mary University of London since 2020. His research interests include 5G/6G networks, NOMA, RIS, UAV communications, and machine learning. He is a Web of Science Highly Cited Researcher 2021. He is currently a Senior Editor of IEEE Communications Letters, and an Editor of IEEE Trans. Wireless Communications and IEEE Trans. Communications. He received the IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020. He also received the 2020 IEEE Signal Processing and Computing for Communications Technical Early Achievement Award; and the IEEE Communication Theory Technical Committee 2021 Early Achievement Award. In addition, he received the IEEE ComSoc Young Professional Outstanding Nominee Award in 2021.

**Zhiguo Ding** [F’20] (zhiguo.ding@manchester.ac.uk) is currently a Professor at the University of Manchester. From September 2012 to September 2022, he has also been an academic visitor at Princeton University. His research interests are 5G networks and statistical signal processing. He is an area editor for IEEE OJ-COMS; an Editor for IEEE TVT; and he has served as an editor for IEEE TCOM, WCL and IEEE CL. He received the 2018 IEEE ComSoc Heinrich Hertz Award, IEEE VTS Jack Neubauer Memorial Award, Friedrich Wilhelm Bessel Research Award 2020, and IEEE SPCC Technical Recognition Award 2021. He is a Distinguished Lecturer of IEEE ComSoc, and was named a Web of Science Highly Cited Researcher in two categories 2021.

**Xiao Li** (x.liu@qmul.ac.uk) received his Ph.D. degree from Queen Mary University of London in 2021. His research interests include UAV-aided networks, machine learning, NOMA techniques, RIS. He is the exemplary reviewer of IEEE Trans. Communication and IEEE Communications Letters, 2019.

**Octavia A. Dobre** [F] (odobre@mun.ca) is a professor and research chair at Memorial University, Canada. Her research interests include technologies for beyond 5G, as well as optical and underwater communications. She has published more than 400 refereed papers in these areas. She serves as the Editor-in-Chief (EiC) of the IEEE Open J. on Communications, an editor with prestigious journals, as well as general chair and technical co-chair of flagship conferences in her area of expertise. She is the recipient of diverse awards, such as Best Paper Awards at IEEE ICC, IEEE Globecom, IEEE WCNC, and IEEE PIMRC conferences. She is an elected member of the European Academy of Sciences and Arts, a fellow of the Engineering Institute of Canada, and a fellow of the Canadian Academy of Engineering.

**Naofal Al-Dhahir** (aldhahir@utdallas.edu) is Erik Jonsson Distinguished Professor and ECE Dept. Associate Head at UT-Dallas. He earned his Ph.D. degree from Stanford University; and he was a principal member of technical staff at GE Research Center and AT&T Shannon Laboratory from 1994 to 2003. He is co-inventor of 43 issued patents; co-author of approximately 500 papers; and co-recipient of five IEEE best paper awards. He is an IEEE Fellow; received the 2019 IEEE SPCC technical recognition award and 2021 Qualcomm faculty award. He served as Editor-in-Chief of IEEE Transactions on Communications from January 2016 to December 2019. He is a Fellow of the National Academy of Inventors.