Wireless Communications and Mobile Computing

Achieving Sustainable 5G

Lead Guest Editor: Kai Yang
Guest Editors: Jinsong Wu and Nan Yang

Achieving Sustainable 5G

Wireless Communications and Mobile Computing

Achieving Sustainable 5G

Lead Guest Editor: Kai Yang
Guest Editors: Jinsong Wu and Nan Yang

Editorial Board

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 8245319, 2 pages
https://doi.org/10.1155/2018/8245319

Editorial

Achieving Sustainable 5G

Kai Yang[1], **Jinsong Wu**[2] **and Nan Yang**[3]

[1]*The School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China*
[2]*The Department of Electrical Engineering, Universidad de Chile, Santiago, Chile*
[3]*The Research School of Engineering, Australian National University, Canberra, ACT 2601, Australia*

Correspondence should be addressed to Kai Yang; yangkai@bit.edu.cn

Received 27 September 2018; Accepted 27 September 2018; Published 18 October 2018

Due to the exponentially increased demands of mobile data traffic, e.g., a 1000-fold increase in traffic demand from 4G to 5G, and the explosive growth in connected mobile devices, dramatic changes in the designs of network architecture are required to meet the 5G requirements, and the opportunities and challenges of 5G rapidly attract great attention from academics, industries, and governments. According to the trend in cellular networks evolution, 5G networks will be heterogeneous ones consisting of macrocells along with a large number of small cells, device-to-device pairs, and machine type communication devices based communication tiers [1]. Indeed, to achieve sustainable 5G and to accelerate the launch of 5G networks, various promising technologies have been proposed and investigated as essential enablers for the operators to achieve a more efficient use of available radio resource and network infrastructure and to reduce both the capacity expenditure and operation expenditure in the network deployment and operations. The motivation behind this special issue is to solicit cutting-edge research results on achieving sustainable 5G.

The paper "Achievable Rates of Gaussian Interference Channel with Multi-Layer Rate-Splitting and Successive Simple Decoding" proposes a scheme which employs multi-layer rate-splitting (RS) at the transmitters and successive simple decoding (SSD) at the receivers in the two-transmitter and two-receiver Gaussian interference channel (IC) model and then studies the achievable sum capacity of this scheme. Numerical simulations are presented to validate that multi-layer RS and SSD are not generally weaker than simultaneous decoding with respect to the achievable sum capacity, at least for some certain channel gain conditions of IC.

The paper "The Rayleigh Fading Channel Prediction via Deep Learning" presents a multi-time channel prediction system based on backpropagation (BP) neural network with multi-hidden layers, which can predict channel information effectively and benefit for massive multiple-input multiple-output performance, power control, and artificial noise-aided physical-layer security scheme design. Meanwhile, an early stopping strategy to avoid the overfitting of BP neural network is introduced.

The paper "General Multimedia Trust Authentication Framework for 5G Networks" proposes a novel multimedia authentication framework based on trusted content representation (TCR) for 5G networks. The general framework is suitable for various multimedia contents, e.g., text, audio, and video. The generality of the framework is guaranteed by the TCR technique, which authenticates the contents semantics at both high and low levels.

The paper "A Sparse Temporal Synchronization Algorithm of Laser Communications for Feeder Links in 5G Nonterrestrial Networks" addresses the temporal synchronization problem in laser communications. In this paper, a new sparsity-aware algorithm for temporal synchronization is proposed without carrier aid through sparse discrete polynomial-phase transformation and sparse discrete fractional Fourier transformation.

The paper "Uplink Nonorthogonal Multiple Access Technologies Toward 5G: A Survey" aims to provide a comprehensive overview about the promising nonorthogonal multiple access (NOMA) schemes. The state-of-the-art NOMA schemes are analyzed by comparing the operations applied at the transmitter, and typical multiuser detection algorithms

corresponding to these NOMA schemes are introduced. In addition, the implementation issues of NOMA are discussed for practical deployment.

The paper "A Novel Query Method for Spatial Data in Mobile Cloud Computing Environment" presents a memory-based spatial data query method that uses the distributed memory file system Alluxio to store data and build a two-level index based on the Alluxio key-value structure. According to the characteristics of Spark computing framework, a data input format for spatial data query is discussed.

The paper "Physical-Layer Channel Authentication for 5G via Machine Learning Algorithm" develops a novel authentication method to detect spoofing attacks without a special test threshold while a trained model is used to determine whether the user is legal or illegal. In addition, a two-dimensional test statistic features authentication model is presented for further improvement of detection rate.

Conflicts of Interest

The guest editors declare that they have no possible conflicts of interest or private agreements with companies.

Acknowledgments

We would like to thank all the reviewers who have participated in reviewing the articles submitted to this special issue.

Kai Yang
Jinsong Wu
Nan Yang

References

[1] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving sustainable ultra-dense heterogeneous networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6039878, 10 pages
https://doi.org/10.1155/2018/6039878

Research Article

Physical-Layer Channel Authentication for 5G via Machine Learning Algorithm

Songlin Chen,[1] **Hong Wen**,[1] **Jinsong Wu**,[2] **Jie Chen**,[1] **Wenjie Liu**,[1] **Lin Hu**,[3] **and Yi Chen**[1]

[1]*The National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*Department of Electrical Engineering, Universidad de Chile, Santiago 833-0072, Chile*
[3]*Chongqing Key Laboratory of Mobile Communication Technology, Chong Qing University of Post & Telecommunication of China, Chongqing, China*

Correspondence should be addressed to Hong Wen; wcdma_2000@hotmail.com

Received 26 January 2018; Accepted 19 September 2018; Published 2 October 2018

Academic Editor: Vicente Casares-Giner

By utilizing the radio channel information to detect spoofing attacks, channel based physical layer (PHY-layer) enhanced authentication can be exploited in light-weight securing 5G wireless communications. One major obstacle in the application of the PHY-layer authentication is its detection rate. In this paper, a novel authentication method is developed to detect spoofing attacks without a special test threshold while a trained model is used to determine whether the user is legal or illegal. Unlike the threshold test PHY-layer authentication method, the proposed AdaBoost based PHY-layer authentication algorithm increases the authentication rate with one-dimensional test statistic feature. In addition, a two-dimensional test statistic features authentication model is presented for further improvement of detection rate. To evaluate the feasibility of our algorithm, we implement the PHY-layer spoofing detectors in multiple-input multiple-output (MIMO) system over universal software radio peripherals (USRP). Extensive experiences show that the proposed methods yield the high performance without compromising the computing complexity.

1. Introduction

5G mobile communication system puts forward the requirements that are high-speed, high efficiency, and high security under three typical application scenarios: enhanced Mobile Broadband (eMBB), Large-Scale Internet of Things (IoT), and ultra Reliable & Low-latency Connections (uRLLC) [1, 2]. The specific application scenarios that enhance the need for mobile broadband including high-traffic and high-density wireless networks are densely used in indoors or urban areas, in which large-area signals of wireless mobile networks are continuously covered in rural areas. Meanwhile, 5G involves the interconnection and communication between a large number of machines and equipment, which is a necessary condition for the operation of IoT [3]. Many mobile devices access the wireless network at the same time, which results in heavy burden of authentication computing in the wireless network. Therefore, lightweight access methods are required for intensive application scenarios of 5G wireless communication networks.

In response to this need, scholars have successively carried out researches on light-weight security measures based on computational cryptography [4, 5]. However, it is still very difficult to use the cipher algorithm that meets the resource-constrained application scenarios such as wireless mobile terminals, IoT, and sensor networks. Therefore, there is a need to find new technologies to construct the lightweight security scheme. In the last decade, the research of PHY-layer security technology has brought new vitality to the wireless mobile communication industry [6–10]. The physical layer of the characteristics is difficult to be counterfeit, which can provide high level security with low cost to overcome the lack of the cipher based security technologies. Consequently, physical layer characteristics which can be used to improve

the security of wireless communications have been widely concerned for researchers.

Several PHY-authentication techniques are proposed. In [11–17], the received signal strength (RSS) and channel impulse response (CIR), as well as channel state information (CSI), are utilized to detect identity-based attacks in wireless networks, such as man-in-the middle and denial-of-service (DoS) attacks. The work [18] presents a PHY-authentication framework that can be adapted for multicarrier transmission. In order to detect Sybil attacks, [19, 20] present a PHY-authentication protocol that combines with high-layer authentication based on the channel response decorrelations rapidly in space, and channel-based detection of Sybil attacks in wireless networks is implemented. In [21], Peng Hao et al. developed a practical authentication scheme by monitoring and analyzing the packet error rate (PER) and received signal strength indicator (RSSI) at the same time to enhance the spoofing attack detection capability. In [22–24], the authors analysed the spatial decorrelation property of the channel response and validated the efficacy of the channel-based authentication for spoofing detection in MIMO system by the comparison between channel information "difference" of two or several frames.

However, in above-mentioned works, artificial thresholds are needed to detect spoofing attack. In fact, threshold range cannot be accurately confirmed, resulting in spoofing detection with low precision. In this paper, a machine learning based PHY-layer authentication is developed, which provides an intelligent decision method instead of a one-dimension test threshold. Specifically, Adaboost [25, 26] based algorithm with one-dimensional feature is employed to detect spoofing attacks. To enhance authentication performance, the two-dimensional feature is carried out. The major contributions of this paper are summarized as follows:

(1) An AdaBoost based PHY-layer authentication algorithm is proposed to increase the authentication rate.

(2) The authentication model based on two-dimensional feature is established, which has a stronger performance for cheating detection than the one-dimensional authentication method.

(3) The proposed PHY-layer channel authentication scheme is implemented in a real world environment, based on MIMO-OFDM systems. The simulation results show that the detection rate is greatly increased.

The rest of this paper is organized as follows. Section 2 describes system model and problem formulation. Our proposed algorithm for PHY-layer authentication is presented in Section 3. The system experiment and simulation results are presented in Section 4. In Section 5, we conclude this paper.

2. System Model

In this section, we provide a system model of physical layer authentication and hypothesis testing.

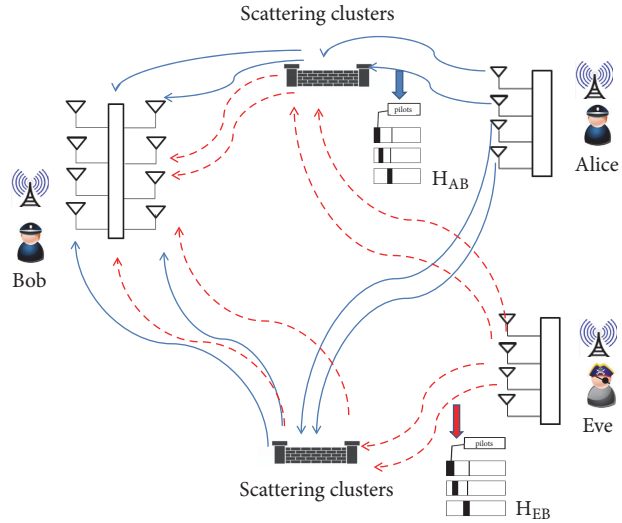


Figure 1: Alice-Bob-Eve model in MIMO system.

2.1. MIMO Three Parts System Model. As shown in Figure 1, our analysis is based on an Alice-Bob-Eve model in MIMO system, where Alice and Bob are legitimate users equipped with N_T and N_R antennas, respectively. Eve with N_T antennas attempts to spoof Alice by using her identity. They are assumed to be located in spatially separated positions. In order to address this spoofing detection, Bob tracks the uniqueness of wireless channel responses to discriminate between legitimate signals from Alice and illegitimate signals from Eve. That is a physical layer authentication. The detailed physical layer authentication process is as follows: Signals with the pilots which can be used to estimate the channel response of the corresponding transmitter are transmitted over the wireless multipath channel to the receiver. The i-th transmission data contains N_f-frames, while each frame consists of N_s OFDM symbols.

Bob is assumed to obtain the Alice-Bob channel information for any frame index $k > 1$, $\widehat{H}_k^{AB}$, and save it which extracted by the channel estimation. After a while, when Bob receives the next data frame, the $k+1$th data frame, $\widehat{H}_{k+1}^{AB}$, which is extracted and estimated by Bob the unknown channel response information. Bob compares $\widehat{H}_{k+1}^{AB}$ with the channel of Alice, $\widehat{H}_k^{AB}$, to determine whether the corresponding signal is actually send by Alice.

If the values of $\widehat{H}_k^{AB}$ and $\widehat{H}_{k+1}^{AB}$ are approaching, Bob considers the sender's identity as valid and stores it. On the contrary, Bob determines that the sender's identity is invalid and directly abandons the data frame.

Channel information is detected by the channel estimation algorithm, denoted by $\widehat{H}_k^{AB}$ and $\widehat{H}_{k+1}^{AB}$. Each data frame contains N_s OFDM symbols. Thus, the channel information is given by

$$\widehat{H}_k^{AB} = \left[\widehat{H}_{k,1}^{AB}, \widehat{H}_{k,2}^{AB}, \ldots, \widehat{H}_{k,N_s}^{AB}\right] \quad (1)$$

where $\widehat{H}_{k,x}^{AB}$ $(x = 1, 2, \ldots, N_s)$ denotes the x-th OFDM symbol of channel information.

2.2. Hypothesis Testing. A binary hypothesis testing is performed to determine the identity authentication in the continuous data frames. Let the receiver Bob verify that the kth data frame originates from the legitimate sender Alice, and the extracted channel information is H_k^{AB}; the sender of the $k+1$ th data frame is still unknown and the channel information is H_{k+1}^{AB}: the null hypothesis H_0 indicates that the packet is indeed sent by the Alice. The alternative hypothesis H1 is that the real client of the packet is not Alice. The spoofing detection builds the hypothesis test given by

$$\begin{aligned} H_0 &: H_{k+1}^{AB} \longrightarrow H_k^{AB} \\ H_1 &: H_{k+1}^{AB} \longrightarrow H_k^{AB} \end{aligned} \tag{2}$$

where all elements of N_k and N_{k+1} are i.i.d. complex Gaussian noise samples $CN(0, \delta^2)$. Therefore, if channel information for hypothesis testing is directly used, the need of considering the impact of noise variables will increase the certification complexity. To this end, since N_k and N_{k+1} are with the same statistical characteristics, the "difference" of channel information can eliminate the influence of noise variables. The physical layer authentication translates into the comparison between the "difference" of the channel information and the set threshold. Equation (2) can be expressed as

$$H_0 : diff\left(H_{k+1}^{AB}, H_k^{AB}\right) < \eta$$

$$H_1 : diff\left(H_{k+1}^{AB}, H_k^{AB}\right) > \eta \tag{3}$$

where $diff(A, B)$ denotes the calculating result of the difference between A and B and η is the test threshold.

The null hypothesis, $\boldsymbol{H}_0$, is that the identity is legitimate and Bob accepts this hypothesis if the test statistic he computes, $diff(A, B)$, is below some threshold η. Otherwise, Bob accepts the alternative hypothesis, $\boldsymbol{H}_1$, that the identity is illegitimate. The channel response "difference" is recorded as T, and (3) can be also written as

$$T = diff\left(H_{k+1}^{AB}, H_k^{AB}\right) \begin{matrix} > H_1 \\ < H_0 \end{matrix} \eta \tag{4}$$

As shown in (4), the physical layer authentication is actually a comparison between channel information "difference" and authentication threshold. Thus, the difference between channel information and authentication threshold is the key of physical layer authentication. The test statistics can measure the similarity of channel information and calculate the channel information difference. In this paper, we use two kinds of test statistic T_A and T_B, respectively. In particular, assuming Bob obtains two consecutive frame channel response of $\widehat{H}_{k-1,x}^{AB}$ and $\widehat{H}_{k,x}^{AB}$, respectively, from Alice. We build test statistics of T_A and T_B based on the two frames for the purpose of discrimination identity of Alice or Eve. Subsequently, Bob acquires the k+1th frame channel information as $\widehat{H}_{k+1,x}^{AB}$.

The test statistics are calculated as

$$T_A(k) = \left|\frac{diff\left(\widehat{H}_{k+1,x}^{AB} - \widehat{H}_{k,x}^{AB}\right)}{diff\left(\widehat{H}_{k,x}^{AB} - \widehat{H}_{k-1,x}^{AB}\right)}\right| = \left|\frac{\sum_{x=1}^{N_s}\sum_{m=1}^{N}\sum_{n=1}^{N}\left|\widehat{H}_{k+1,x}^{AB}(m,n) - \widehat{H}_{k,x}^{AB}(m,n)\, e^{j\widehat{\theta}(m,n)}\right|}{\sum_{x=1}^{N_s}\sum_{m=1}^{N}\sum_{n=1}^{N}\left|\widehat{H}_{k,x}^{AB}(m,n) - \widehat{H}_{k-1,x}^{AB}(m,n)\, e^{j\widehat{\theta}(m,n)}\right|}\right| \begin{matrix} > H_1 \\ < H_0 \end{matrix} \eta_A, \tag{5}$$

where $\widehat{\theta}(m,n)$ is the phase offset and can be denoted by

$$\widehat{\theta}(m,n) = \arg\left(\widehat{H}_{k,x}^{AB}(m,n)\left[H_{k+1,x}^{XB}(m,n)\right]^*\right) \tag{6}$$

From (5), T_A can be taken as the difference of the subcarrier amplitude, which avoids the effect of $\widehat{\theta}(m,n)$.

Two consecutive data frames, $\widehat{H}_{k,x}^{AB}$ and $\widehat{H}_{k+1,x}^{AB}$, represent measurement errors in the phase of the channel response. Each channel response value consists of N_s frequency domain channel matrix, which is OFDM symbol of N dimensional square matrix and n denotes the mth row and n denotes the column element phase offset.

$$T_B(k) = \left|\frac{diff\left(\widehat{H}_{k+1,x}^{AB} - \widehat{H}_{k,x}^{AB}\right)}{diff\left(\widehat{H}_{k,x}^{AB} - \widehat{H}_{k-1,x}^{AB}\right)}\right| = \left|\frac{\sum_{x=1}^{N_s}\sum_{m=1}^{N}\sum_{n=1}^{N}\left|\widehat{H}_{k+1,x}^{XB}(m,n) - \widehat{H}_{k,x}^{AB}(m,n)\right|^2}{\sum_{x=1}^{N_s}\sum_{m=1}^{N}\sum_{n=1}^{N}\left|\widehat{H}_{k,x}^{AB}(m,n) - \widehat{H}_{k-1,x}^{AB}(m,n)\right|^2}\right| \begin{matrix} > H_1 \\ < H_0 \end{matrix} \eta_A \tag{7}$$

where T_B is the test statistic based on amplitude and phase information. We use T_A and T_B as the one-dimensional test statistic, respectively, for detecting spoofing attack. Unfortunately, it is hard to find the best threshold for achieving high accuracy authentication detection rate. To tackle this problem, we propose a learning algorithm based on AdaBoost to achieve physical layer authentication, in which T_A and T_B are used as training features.

3. Physical Authentication with AdaBoost Algorithm

In this section, we propose a learning algorithm based on AdaBoost for physical authentication.

3.1. AdaBoost Algorithm. AdaBoost is the abbreviation of adaptive boosting and developed by Yoav Freund [24] and is

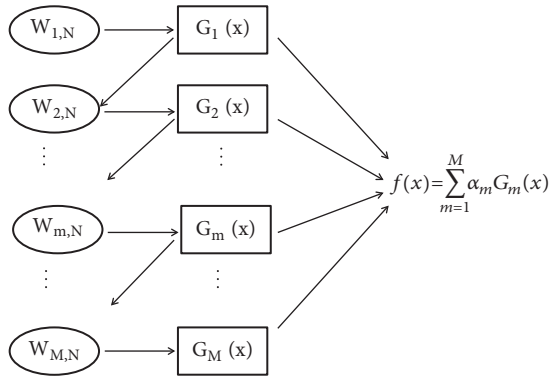


Figure 2: AdaBoost algorithm.

the most widely used form of boosting algorithm. Boosting is a powerful technique combined with base classifiers [25] to produce a form of committee whose performance can be significantly better than other base classifiers. The principal of AdaBoost algorithm is that this algorithm improves its performance by the iterative algorithm, which is adaptive in the sense that subsequent weak classifiers, called as learners, are adjusted to improve those instances misclassified by previous classifiers. AdaBoost can be seen as a particular method of training a boosted classifier. A boost classifier is a classifier as follows:

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x) \tag{8}$$

where each $G_m(x)$ is a weak classifier that takes x as input and returns a value y_m indicating the class of x. The weak classifiers, each of classifiers is trained by using a weighted coefficient $w_{m,i}$ from the data set where the weighting coefficient associated depending on the performance of the weak classifiers such as decision tree (support vector machine) SVM, are trained in sequence. More specially, data points which are misclassified by one of the weak classifiers are being given greater weight, which are used to train the next weak classifier. As illustrated in Figure 2, once all the classifiers have been trained until there are no misclassified data points, then their final model is generated via a weight majority voting scheme.

3.2. Physical Authentication with AdaBoost Algorithm. The physical authentication with AdaBoost algorithm is proposed for detection spoofing. The performance chart of the algorithm is illustrated in Figure 3. Bob collects the channel matrix, $\widehat{\mathbf{H}}_1^{AB}$, which obtained by channel estimation using the pilot from Alice and records it. When Bob receives the next data frame from the Alice, the Bob collects channel information, $\widehat{\mathbf{H}}_2^{AB}$. Similarly, Bob collects continuous N-frames channel information from Alice and stores as $\widehat{\mathbf{H}}^{AB} = [\widehat{\mathbf{H}}_1^{AB}, \widehat{\mathbf{H}}_2^{AB}, \ldots, \widehat{\mathbf{H}}_N^{AB}]$. In the same time an Eve sends the data frames to the Bob and claims that he is Alice. In practical communication scenarios, we do not know where and who Eves are. But in proposed scheme Eves are needed to be test training purpose. Therefore, one or several Eve nodes are set for this purpose. Bob continuously extracts the continuous N frames channel information from Eve and stores as $\widehat{\mathbf{H}}^{EB} = [\widehat{\mathbf{H}}_1^{EB}, \widehat{\mathbf{H}}_2^{EB}, \ldots, \widehat{\mathbf{H}}_N^{EB}]$.

The data set is preprocessed by Bob. Firstly, Bob calculates the value of data set, $\widehat{\mathbf{H}}^{AB}$, $\widehat{\mathbf{H}}^{EB}$. Secondly, Bob calculates the test statistics based on test statistics T_A, T_B as

$$T_A^{XB}(k) = \left| \frac{diff\left(\widehat{H}_{k+1,x}^{XB} - \widehat{H}_{k,x}^{XB}\right)}{diff\left(\widehat{H}_{k,x}^{XB} - \widehat{H}_{k-1,x}^{XB}\right)} \right| = \left| \frac{\sum_{x=1}^{N_s} \sum_{m=1}^{N} \sum_{n=1}^{N} \left| \widehat{H}_{k+1,x}^{XB}(m,n) - \widehat{H}_{k,x}^{XB}(m,n)\, e^{j\widehat{\theta}(m,n)} \right|}{\sum_{x=1}^{N_s} \sum_{m=1}^{N} \sum_{n=1}^{N} \left| \widehat{H}_{k,x}^{XB}(m,n) - \widehat{H}_{k-1,x}^{XB}(m,n)\, e^{j\widehat{\theta}(m,n)} \right|} \right| \begin{matrix} > H_1 \\ < H_0 \end{matrix} \eta_A^X, \tag{9}$$

$$T_B^{XB}(k) = \left| \frac{diff\left(\widehat{H}_{k+1,x}^{XB} - \widehat{H}_{k,x}^{XB}\right)}{diff\left(\widehat{H}_{k,x}^{XB} - \widehat{H}_{k-1,x}^{XB}\right)} \right| = \left| \frac{\sum_{x=1}^{N_s} \sum_{m=1}^{N} \sum_{n=1}^{N} \left| \widehat{H}_{k+1,x}^{XB}(m,n) - \widehat{H}_{k,x}^{XB}(m,n) \right|^2}{\sum_{x=1}^{N_s} \sum_{m=1}^{N} \sum_{n=1}^{N} \left| \widehat{H}_{k,x}^{XB}(m,n) - \widehat{H}_{k-1,x}^{XB}(m,n) \right|^2} \right| \begin{matrix} > H_1 \\ < H_0 \end{matrix} \eta_B^X \tag{10}$$

Finally, Bob generates training data set of two categories. The first one is

$$T_A^{AB} = \{x_1, \ldots, x_i, \ldots, x_N, y_A\}, \tag{11a}$$

$$T_B^{AB} = \{x_1, \ldots, x_i, \ldots, x_N, y_A\}, \tag{11b}$$

where $x_i \in T_A^{AB}(k)$ or $x_i \in T_B^{AB}(k)$, $y_A = +1$, by substituting $\widehat{\mathbf{H}}^{AB}$, into (9), (10), yields T_A^{AB},T_B^{AB}, and the value of y_A represents that the transmitter is the legal transmitter from Alice. And the second training set is

$$T_A^{EB} = \left\{x_1^E, \ldots, x_i^E, \ldots, x_N^E, y_B^E\right\} \tag{12a}$$

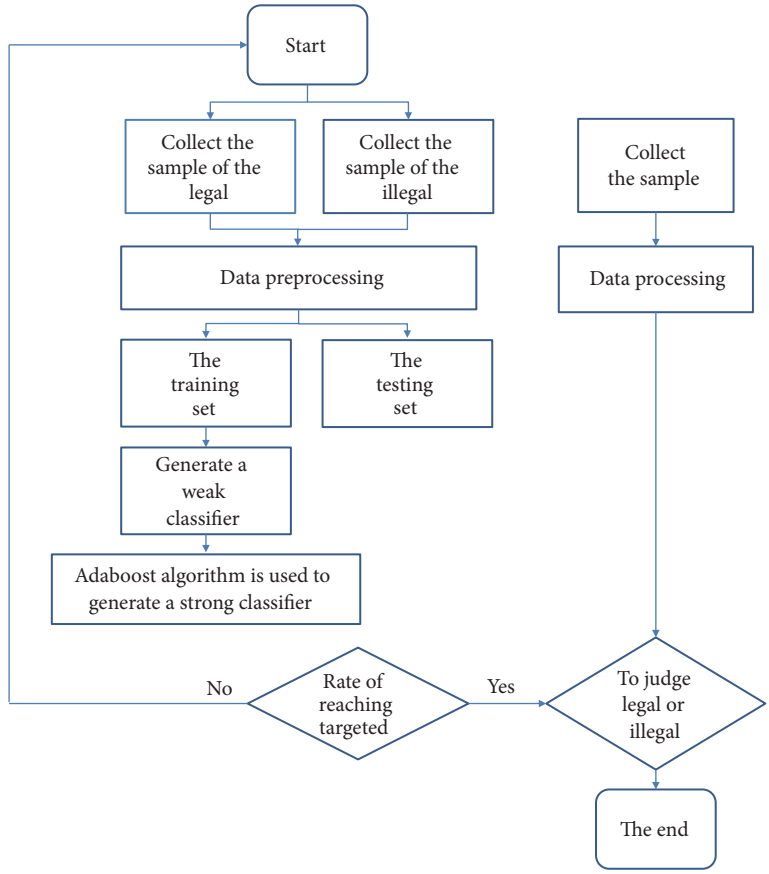


Figure 3: Physical authentication with AdaBoost algorithm.

$$T_B^{EB} = \left\{x_1^E, \ldots, x_i^E, \ldots, x_N^E, y_B^E\right\} \tag{12b}$$

where $x_i^E \in T_A{}^{EB}(k)$ or $x_i^E \in T_B{}^{EB}(k)$, $y_B^E = -1$, by substituting $\widehat{\mathbf{H}}^{EB}$, into (9) and (10), yields T_A^{EB} and T_B^{EB}, and the value of y_i represents that the transmitter is the illegal transmitter from Eve. Bob uses the two classification training data set $T_A^{AB}, T_B^{AB}, T_A^{EB}$, and T_B^{EB} as input training set.

Spoofing detection is essentially a two-classification problem, which is considered to be solved through AdaBoost algorithm. The training data is made up of a bunch of sample points. Each sample point comprises input sample x_i and label y_i where $y_i \in \{-1, 1\}$. Each sample point is given an associated weight parameter $w_{m,i}$, m means *m-th* training, and i means the number of sample points, which is initially set $1/i$ for all sample points. We suppose that we have a procedure available for training a weak classifier using weighted sample points. At each iteration of the training process, AdaBoost trains a new weak classifier by using the sample points in which the weighting coefficients are adjusted according to the performance of the previously trained weak classifier, so as to give greater weight to the misclassified data points, in which the classification error rate e_m is used to evaluate misclassified data set D_m

$$e_m = P\left(G_m\left(x_i\right) \neq y_i\right) = \sum_{i=1}^{2t} w_{mi} I\left(G_m\left(x_i\right) \neq y_i\right) \tag{13}$$

Then the coefficient α_m of G_m is calculate as

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{14}$$

Finally, we generate a final model that different weight is being given to different weak classifiers in (8). The AdaBoost algorithm is given as in Algorithm 2, in which the point of the training data can be doubled by combining with the one-dimension test statistics T_A and T_B together and become a new two-dimensional features authentication model for spoofing detection. Therefore, in the AdaBoost algorithm, the input training data set T is following two optional sets:

(1) One-dimension test statistics training data set:

$$\begin{aligned} T &= \left\{T_A^{AB}, T_A^{EB}\right\} \\ \text{or } T &= \left\{T_B^{AB}, T_B^{EB}\right\} \end{aligned} \tag{15}$$

(2) Two-dimension test statistics training data set:

$$T = \left\{\left(T_A^{AB}, T_B^{AB}\right), \left(T_A^{EB}, T_B^{EB}\right)\right\} \tag{16}$$

Input:
The channel information of legal transmitter or illgal transmitter:
Process:
1: Bob calculates the value of data set $\widehat{\mathbf{H}}^{AB}$ and $\widehat{\mathbf{H}}^{EB}$ from Alice and simulated Eve:
$\widehat{\mathbf{H}}^{AB} = \left[\widehat{\mathbf{H}}_1^{AB}, \widehat{\mathbf{H}}_2^{AB}, \ldots, \widehat{\mathbf{H}}_N^{AB}\right]$
$\widehat{\mathbf{H}}^{EB} = \left[\widehat{\mathbf{H}}_1^{EB}, \widehat{\mathbf{H}}_2^{EB}, \ldots, \widehat{\mathbf{H}}_N^{EB}\right]$
2: The data set are preprocessed by Bob:
3: The data set are divided into two parts, and the one is training data set and the other is testing data set:
4: Use training data set to get the weak classifier:
5: Use the Adaboost algorithm to generate a strong classifer:
6: The testing data set is used to verify whether the claasifier can achieve the target detection rate, otherwise it will return to the first step:
7: The final classifier is the authenticaton decision model, which can judge whether the new packets are legitimate or illegal:
End

Algorithm 1: Physical authentication.

Input:
training data set T:
Process:
1: Initialize the weight distribution of the sample points:

$$D_1 = (w_{11}, \ldots, w_{1i}, \ldots, w_{1,2t}), \quad w_{1i} = \frac{1}{2t}, \ i = 1, 2, \ldots, 2t$$

2: for $m = 1$ to M do, m means m-th training
3: Use the training data set of D_m to learn and get the weak classifier:

$$G_m(x) : x_i \longrightarrow \{-1, +1\}$$

4: Calculate the classification error rate of D_m on the training data set:

$$e_m = P(G_m(x_i) \neq y_i) = \sum_{i=1}^{2t} w_{mi} I\left(G_m\left(x_i\right) \neq y_i\right)$$

5: Calculate the coefficient of G_m:

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$$

6: Update the weight distribution of the training data set:

$$D_{m+1} = (w_{m+1,1}, \ldots, w_{m+1,i}, \ldots, w_{m+1,2t}),$$
$$w_{m+1,i} = \frac{w_{mi}}{Z_m} \exp\left(-\alpha_m y_i G_m\left(x_i\right)\right), \quad i = 1, 2, \ldots, 2t$$
$$Z_m = \sum_{i=1}^{2t} w_{mi} \exp\left(-\alpha_m y_i G_m\left(x_i\right)\right)$$

7: Construct a linear combination of weak classifiers:

$$f(x) = \sum_{i=1}^{M} \alpha_m G_m(x)$$

End for
return: $G(x) = \text{sign}(f(x))$

Algorithm 2: AdaBoost.

4. Experimental Verification

In this section, we will describe the system setup and the test process of measuring the Algorithm 1 for detecting Alice and Eve.

4.1. System Setup. We consider the spoofing detection of a receiver called Bob, the legal transmitter called Alice, and the spoofing node called Eve. They were placed in three separate locations in a room, surrounded by many other devices such as printers, desktops, and other types of equipment as shown in Figure 4. There are scattering and refraction phenomena in the room due to the presence of obstacles in the wireless channel from Alice to Bob and Eve to Bob. As shown in Figure 5, we set up experimental platform which implemented on USRPs, and experiments were performed in an indoor environment. Bob is equipped with an 8∗8 MIMO system, Alice is equipped with a 2∗2 MIMO system, and the spoofing node called Eve is equipped with a 2∗2 MIMO system. The signals are sent over 2 antennas each at center frequency 3.5GHz with bandwidth 2MHz.

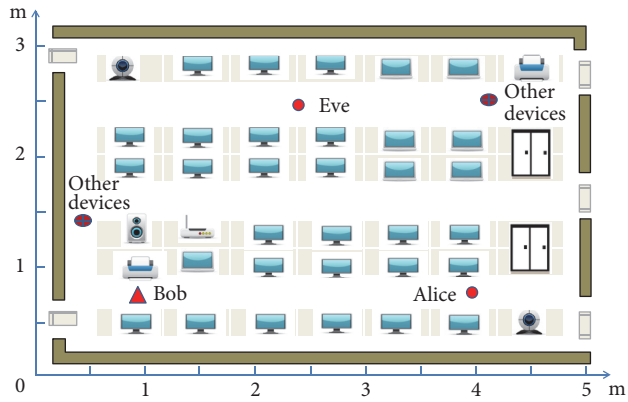


Figure 4: The experiments consisted of Alice, Bob, and Eve.

Figure 5: Real MIMO communication platform consisted of Alice, Bob, and Eve.

4.2. Experiment. In the experiment, the following steps are taken.

Step 1. Bob extracts channel information from Alice and Eve by the existing channel estimation mechanisms, respectively.

Step 2. Bob preprocesses the dataset according to (5), (7), (9), and (10) while the threshold is between [0, 1] (normalization).

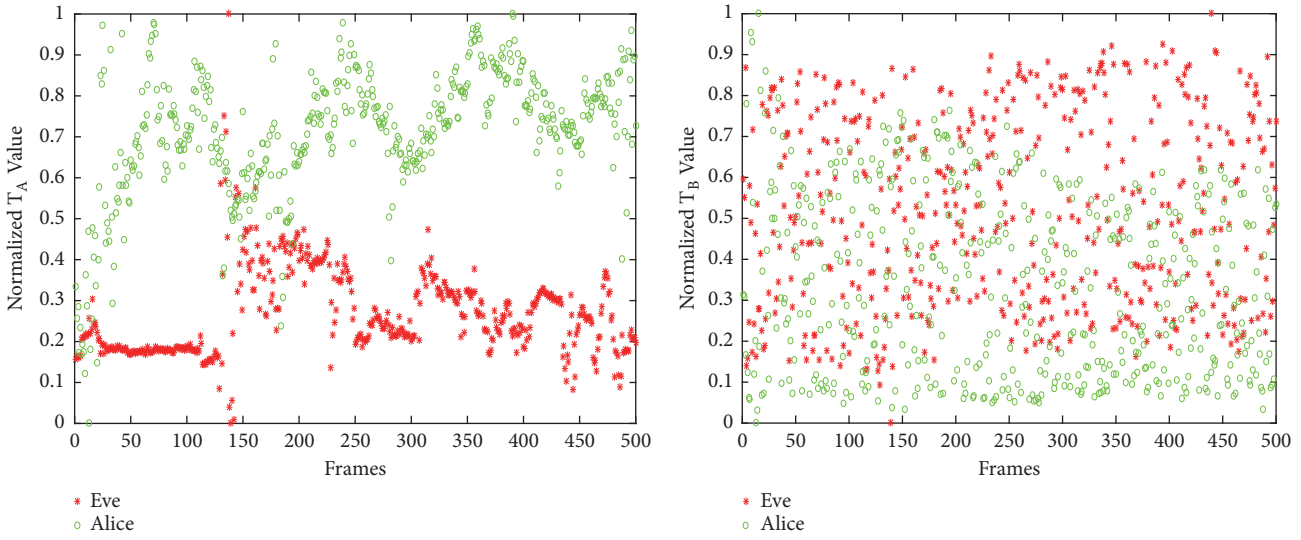
Step 3. Bob generates a training data set of two classifications according to (11a), (11b), (12a), and (12b).

Step 4. The two classification training data set T is generated according to (15) or (16).

Step 5. Bob is trained to generate a strong classifier based on the training data set of two classifications by using AdaBoost algorithm under Matlab program.

Step 6. Bob uses a strong classifier to judge the test set and obtain the authentication detection rate.

In the experiment, we consider that the collection frames are five hundred frames and the value of test statistic was normalized between 0 and 1. The test statistic T_A of channel information of the Alice and Bob as a function of frames is shown in Figure 6(a), in which the red points is $T_A(k)$ in (5) and green points is $T_A^E(k)$ in (9). As can be seen, there is the overlapped area. Meanwhile, from Figure 6(b), the overlapped area is large, when we chose the test statistic T_B of channel information in which the red points is $T_A(k)$ in (7) and green points is $T_A^E(k)$ in (10). It is clearly shown that it is difficult to acquire the best manual test threshold for the accuracy of authentication. Moreover, we use T_A, T_B, and the number of frames, respectively, to draw a three-dimensional plot. As shown in



(a) Normalized T_A of Alice and Eve

(b) Normalized T_B of Alice and Eve

FIGURE 6: Normalized T_A and T_B value of the legal transmitter Alice and the spoofing node Eve for spoofing detection with center frequency 3.5GHz with bandwidth 2MHz.

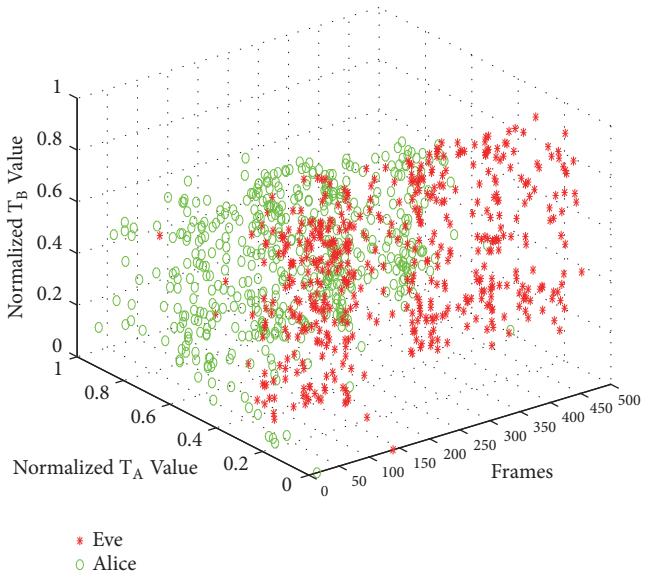


FIGURE 7: Normalized T_A value and T_B value of the legal transmitter Alice and the spoofing node Eve drawing three dimensional plot.

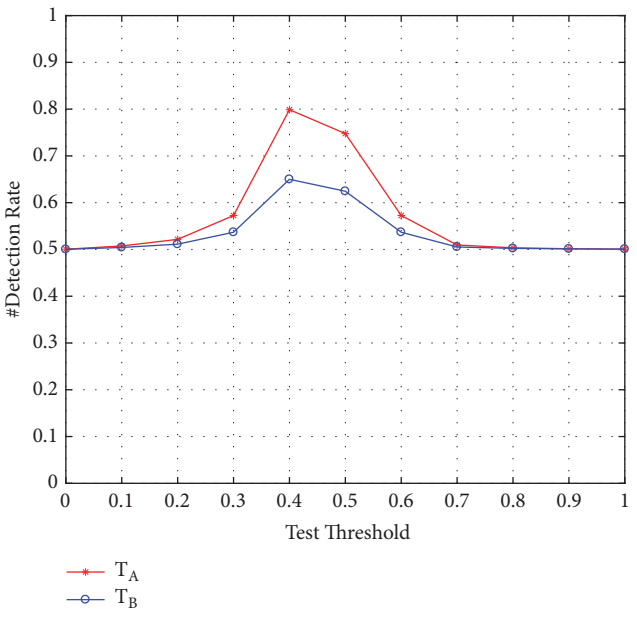


FIGURE 8: Correct classification rate of T_A and T_B.

Figure 7, obviously, it is hard to use the traditional manual threshold method to identify the identity of data sets in the three-dimensional condition. However, machine learning algorithm based the authentication model can effectively settle this problem and a dividing curved surface can perform the identification by the AdaBoost adaptive adjustment algorithm.

4.3. Simulation Results. In this section, simulation results are provided to demonstrate the performance of the proposed authentication scheme.

As a comparison, we considered the PHY-layer spoofing detection [15] with a varied test threshold. From the Figure 8, we can see that when test threshold equals 0.4, the best authentication detection rate results of using T_A or T_B reached 79.8% and 65.4%, respectively. In addition, our proposed method which combined two test statistics T_A and T_B as a two-dimensional feature can improve the accuracy of detection. We use T_A, T_B, and the number of frames, respectively, to draw a three-dimensional plot. Figure 9 illustrates the comparison of spoofing detection among the three methods, from which we can conclude that manual threshold method based on T_A test statistics can achieve 79.8% detection rate

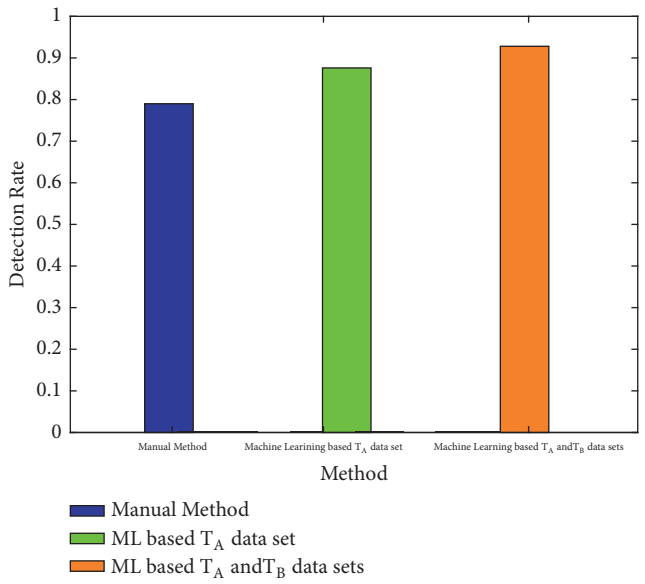


Figure 9: The simulation result with the different method of authentication scheme.

while machine learning based authentication method with T_A test statistic can acquire 87.1% detection rate and machine learning based authentication method with two-dimensional features T_A and T_B can achieve 91.3% accuracy rate with an additional 10% more computation complexity.

To sum up, the proposed authentication scheme achieves a superior performance over manual threshold strategy [15]. Based on the above observation, the proposed machine learning based authentication scheme with tow-dimensional feature not only exhibits excellent performance than manual method but also has higher authentication rate than that of the same algorithm with one-dimensional feature.

5. Conclusions

In this paper, machine learning algorithm based physical-layer channel authentication for the 5G wireless communication security is proposed. A machine learning authentication method could draw a conclusion whether the received packets are from a legitimate transmitter or from a counterfeiter by using one-dimension or two-dimensional joint features. The effectiveness of the proposed authentication scheme is validated by widely simulations. All the data used in the simulation are derived from real OFDM-MIMO communication platform, which provides a real communication environment. Moreover, the authentication results show that the novel methods provide a higher rate in detecting the spoofing attacks than those of the manual threshold based physical layer authentication schemes. The training of the classifier can be done offline. Therefore, the novel method can perform authentication fast. In addition, whether we can use more machine learning algorithms to further optimize our authentication model and find a better statistical test of large difference in channel information is issue that we need to deal with in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by NSFC (no. 61572114), Sichuan Sci & Tech. Achievements Transformation Project (no. 2016CC003), Sichuan Sci & Tech. Service Development Project (no. 18KJFWSF0368), Hunan Provincial Nature Science Foundation Project 2018JJ2535, Chile CONICYT FONDECYT Regular Project 1181809, and National Key R&D Program of China (2018YFB0904900 and 2018YFB0904905).

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[2] J. Thompson, X. Ge, and H.-C. Wu, "5G wireless communication systems: prospects and challenges [Guest Editorial]," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 62–64, 2014.

[3] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A survey on internet of things from industrial market perspective," *IEEE Access*, vol. 2, pp. 1660–1679, 2014.

[4] A. Bogdanov, M. Knežević, G. Leander, D. Toz, K. Varıcı, and I. Verbauwhede, "SPONGENT: the design space of lightweight cryptographic hashing," *IEEE Transactions on Computers*, vol. 62, no. 10, pp. 2041–2053, 2013.

[5] R. Zhang, L. Zhu, C. Xu, and Y. Yi, "An Efficient and secure RFID batch authentication protocol with group tags ownership Transfer," *IEEE Collaboration and Internet Computing*, pp. 168–175, 2015.

[6] L. Hu, H. Wen, B. Wu et al., "Cooperative Jamming for Physical Layer Security Enhancement in Internet of Things," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 219–228, 2018.

[7] L. Hu, H. Wen, B. Wu, J. Tang, F. Pan, and R.-F. Liao, "Cooperative-Jamming-Aided Secrecy Enhancement in Wireless Networks with Passive Eavesdroppers," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2108–2117, 2018.

[8] H. Wen, *Physical layer approaches for securing wireless communication systems*, Springer, New York, NY, USA, 2013.

[9] L. Hu, H. Wen, B. Wu, J. Tang, and F. Pan, "Adaptive Base Station Cooperation for Physical Layer Security in Two-Cell Wireless Networks," *IEEE Access*, vol. 4, pp. 5607–5623, 2016.

[10] L. Hu, H. Wen, B. Wu, J. Tang, and F. Pan, "Adaptive Secure Transmission for Physical Layer Security in Cooperative Wireless Networks," *IEEE Communications Letters*, vol. 21, no. 3, pp. 524–527, 2017.

[11] D. B. Faria and D. R. Cheriton, "Detecting identity-based attacks in wireless networks using signalprints," in *Proceedings of the ACM Workshop on Wireless Security*, pp. 43–52, Los Angeles, Calif, USA, 2006.

[12] M. Demirbas and Y. Song, "An RSSI-based scheme for sybil attack detection in wireless sensor networks," in *Proceedings of*

the WoWMoM 2006: 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks, pp. 564–568, June 2006.

[13] Y. Chen, W. Trappe, and R. P. Martin, "Detecting and localizing wireless spoofing attacks," in *Proceedings of the 2007 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON*, pp. 193–202, San Diego, Calif, USA, June 2007.

[14] N. Patwari and S. K. Kasera, "Robust location distinction using temporal link signatures," in *Proceedings of the ACM International Conference on Mobile Computing and NETWORKING*, pp. 111–122, 2007.

[15] H. Wen, Y. Wang, X. Zhu, J. Li, and L. Zhou, "Physical layer assist authentication technique for smart meter system," *IET Communications*, vol. 7, no. 3, pp. 189–197, 2013.

[16] J. K. Tugnait, "Wireless user authentication via comparison of power spectral densities," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1791–1802, 2013.

[17] Z. Jiang, J. Zhao, X. Li, J. Han, and W. Xi, "Rejecting the attack: Source authentication for Wi-Fi management frames using CSI Information," in *Proceedings of the IEEE INFOCOM 2013 - IEEE Conference on Computer Communications*, pp. 2544–2552, Turin, Italy, April 2013.

[18] X. Wu and Z. Yang, "Physical-layer authentication for multi-carrier transmission," *IEEE Communications Letters*, vol. 19, no. 1, pp. 74–77, 2015.

[19] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-authentication protocol for spoofing detection in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10037–10047, 2016.

[20] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-Layer Spoofing Detection with Reinforcement Learning in Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10037–10048, 2016.

[21] P. Hao, X. Wang, and A. Refaey, "An enhanced cross-layer authentication mechanism for wireless communications based on PER and RSSI," in *Proceedings of the 2013 13th Canadian Workshop on Information Theory, CWIT 2013*, pp. 44–48, Canada, June 2013.

[22] S. Chen et al., "Machine-to-Machine communications in ultra-dense networks—A survey," *IEEE Communications Surveys & Tutorials*, vol. 1, no. 1, 99 pages, 2017.

[23] L. Xiao, L. J. Greenstein, N. B. Mandayam, and W. Trappe, "Channel-based detection of sybil attacks in wireless networks," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 492–503, 2009.

[24] H. Wen, P.-H. Ho, C. Qi, and G. Gong, "Physical layer assisted authentication for distributed ad hoc wireless sensor networks," *IET Information Security*, vol. 4, no. 4, pp. 390–396, 2010.

[25] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 8547620, 13 pages
https://doi.org/10.1155/2018/8547620

Research Article

Achievable Rates of Gaussian Interference Channel with Multi-Layer Rate-Splitting and Successive Simple Decoding

Hanxiao Yu and Zesong Fei

School of Information and Electronics, Beijing Institute of Technology, Beijing, China

Correspondence should be addressed to Zesong Fei; feizesong@bit.edu.cn

Received 26 January 2018; Revised 5 June 2018; Accepted 16 July 2018; Published 1 August 2018

Academic Editor: Jinsong WU

The capacity bound of the Gaussian interference channel (IC) has received extensive research interests in recent years. Since the IC model consists of multiple transmitters and multiple receivers, its exact capacity region is generally unknown. One well-known capacity achieving method in IC is Han-Kobayashi (H-K) scheme, which applies two-layer rate-splitting (RS) and simultaneous decoding (SD) as the pivotal techniques and is proven to achieve the IC capacity region within 1 bit. However, the computational complexity of SD grows exponentially with the number of independent signal layers, which is not affordable in practice. To this end, we propose a scheme which employs multi-layer RS at the transmitters and successive simple decoding (SSD) at the receivers in the two-transmitter and two-receiver IC model and then study the achievable sum capacity of this scheme. Compared with the complicated SD, SSD regards interference as noise and thus has linear complexity. We first analyze the asymptotic achievable sum capacity of IC with equal-power multi-layer RS and SSD, where the number of layers approaches to infinity. Specifically, we derive the closed-form expression of the achievable sum capacity of the proposed scheme in symmetric IC, where the proposed scheme only suffers from a little capacity loss compared with SD. We then present the achievable sum capacity with finite-layer RS and SSD. We also derive the sufficient conditions where employing finite-layer RS may even achieve larger sum capacity than that with infinite-layer RS. Finally, numerical simulations are proposed to validate that multi-layer RS and SSD are not generally weaker than SD with respect to the achievable sum capacity, at least for some certain channel gain conditions of IC.

1. Introduction

Due to the broadcast nature of wireless channel, the interference greatly affects the performance of wireless communication when multiple signal streams are transmitted on the same time/frequency resources. As a general description, the interference channel (IC) model has been proposed to describe the channel statistics where multiple transmitters and multiple receivers share the same physical resources [1]. In recent decades, IC has been regarded as an important building block in cognitive radio [2], multicell network [3, 4], and massive input massive output system [5].

A basic IC model is illustrated in Figure 1, where two transmitters, i.e., Tx-1 and Tx-2, aim to simultaneously transmit their signals to two receivers, i.e., Rx-1 and Rx-2, respectively. Before analyzing the capacity bound and the capacity approaching techniques of IC, we may first recall the other two well-studied multiuser channel models, i.e., the multiple access channel (MAC) model and the broadcast channel (BC) model, where rate-splitting (RS), superposition coding (SC), and successive simple decoding (SSD) are usually required to approach the capacity bounds of MAC and BC. However, different from MAC and BC which either has a single transmitter or a single receiver, IC has at least two independent links, i.e., Tx-1-to-Rx-1 and Tx-2-to-Rx-2 as shown in Figure 1, which may interfere with each other. Each receiver in IC receives multiple signal streams, which constitute an MAC. Symmetrically, each transmitter in IC broadcasts the signal to multiple users which exactly follows a BC. Therefore, the IC model can be regarded as a composition of MAC and BC, and this fact makes the analysis in either MAC or BC not sufficient in the IC.

In the past several decades, the problem of finding the exact capacity region of Gaussian IC has been proven to

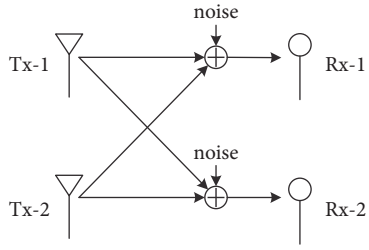


Figure 1: The interference channel model with two transmitters and two receivers.

be pretty hard. The exact capacity region of IC with strong interference is derived in [6]. Meanwhile, [7] analyzes the capacity region of discrete-memoryless IC. Some researchers have focused on the capacity region of degraded IC, e.g., Z-IC [8–12].

Nevertheless, the capacity region of a general IC has not been clearly revealed yet. One best known achievable rate region of a general IC is Han-Kobayashi (H-K) inner bound [13]. The original H-K bound is hard to be analytically described and depicted; therefore, H. Etkin in [14] proposes a simplified H-K scheme, which approaches the IC capacity region within 1 bit. The capacity region achieved by the simplified H-K scheme is termed simple H-K region. To prove the achievability, the simultaneous decoding (SD) of more than one codeword is required, which may lead to exponential computational complexity. Recently, simultaneous nonunique decoding (SND) is proved to be rate optimal when random coding is applied [15]. However, the decoding complexity is still high. Hence, one key question to be studied in the era of IC is

(i) Are there any simple decoding and encoding methods which can be used to achieve the H-K capacity region?

To find the answer, we may look at the capacity approaching techniques in MAC, due to the fact that MAC can be regarded as a degraded version of IC. To approach every rate pair in MAC capacity region, the transmitters employ RS and SC and the receiver employs SSD and successive interference cancellation (SIC) [16]. While SD/SND generally has exponential computational complexity with respect to the number of independent coding layers in the transmission signals, SSD/SIC only requires linear computational complexity. Hence, SSD/SIC are pretty simple compared with SD or SND and have been attractive to the researchers from many fields [17–19]. For example, J. Cao in [19] proves that, with infinite number of rate splits at each transmitter, applying RS and SSD can asymptotically achieve capacity of MAC bound in a distributed manner.

In view of the benefits of RS, two RS-based schemes are proposed in [20, 21], separately, to achieve the H-K inner bound in IC. However, Omar in [22] shows that joint decoding is still required in [20, 21] (instead of SSD) and the receiving complexity of the methods in [20, 21] is actually not reduced. Therefore, whether RS and SSD can achieve the H-K inner bound remains a question. Y. Zhao in [23] studies the maximum achievable rate with SSD in the deterministic model for IC. However, the deterministic model only works in high SNR region. Still, [23] does not answer the aforementioned question. In [24], the authors point out that any finite-layer RS and SSD cannot achieve the corner points of the SD bound of the symmetric Gaussian IC, where the interference is strong but not very strong. Alternatively, a sliding window superposition coding method is proposed in [24] to achieve the simultaneous decoding inner bound where interference cancellation is available at different time slots. However, this method suffers from performance loss since the first and last blocks are not fully loaded with messages. Therefore, with general channel gain settings, the question that whether RS and SSD can be used to achieve the SD achievable rate region is still unsolved.

As conjectured by Omar in [22], multiple-layer RS may be required such that SSD can achieve the bound close to H-K capacity. Following this conjunction, in this paper, we conduct an asymptotic analysis of the achievable rate with multi-layer RS and SSD in IC and aim at answering the question whether infinite-layer RS can achieve the SD inner bound [22, 24]. We note that once multi-layer RS and SSD are able to approach the SD achievable rate region, they can be directly applied in the H-K scheme to achieve the utmost capacity region of IC. We assume that RS is conducted by splitting the transmission power and assigning suitable rate for each split. We start with equal-power RS with infinite number of layers and then find that infinite-layer equal-power RS and SSD cannot approach the SD bound in general. Especially, we derive the closed-form formula of the performance gap between the proposed scheme and the SD bound in symmetric IC. We note that the performance gap is pretty small. Based on the above results, we then analyze the achievable rate with finite-layer RS. Surprisingly, with certain channel gain conditions, we show that employing finite-layer RS and SSD can achieve even better sum capacity than SD. To sum up, the main result of this paper is that employing multiple layers RS and SSD/SIC can nearly approach the SD bound in IC. The result can be further exploited as a guideline in designing capacity approaching technologies in practical scenarios, such as designing good inter-cell interference cancellation method.

This paper is organized as follows. Section 2 describes the system model and the useful notations. Section 3 presents the achievable rate with SSD when infinite-layer RS is applicable. In Section 4, we analyze the achievable sum capacity when finite-layer RS is assumed. The numerical results are presented in Section 5. Section 6 concludes this paper.

2. System Model

2.1. Multi-Layer Rate-Splitting for Interference Channel. We consider an IC model with two transmitters, i.e., Tx-i, $i = 1, 2$, and two receivers, i.e., Rx-j and $j = 1, 2$, where Rx-j is the target receiver of Tx-i when $j = i$. We assume the additive white Gaussian noise (AWGN) channel, as shown in Figure 2, where the channel gain between Tx-i and Rx-j is $h_{j,i}$, and the noise variance is N_0. The transmission power at Tx-i is P_i.

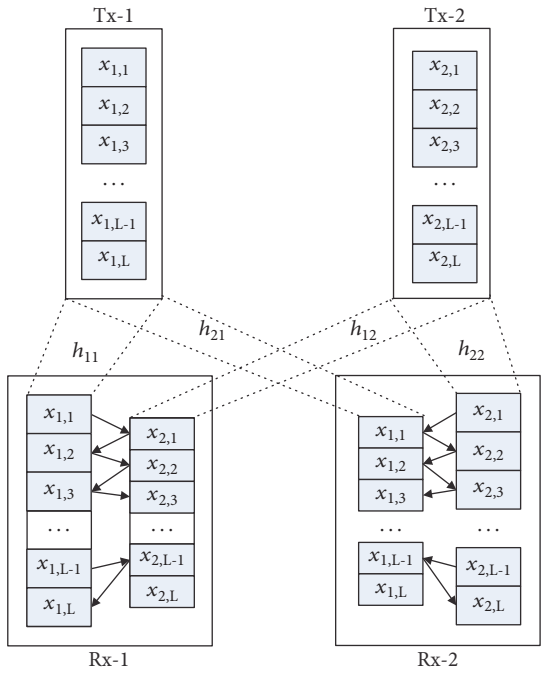


FIGURE 2: An illustration of multi-layer RS and SSD in IC, with the decoding order Π_{12}.

Without loss of generality, we assume $P_1 = P_2 = P$. Then, the received signals are given by

$$y_1 = \sqrt{h_{1,1}}x_1 + \sqrt{h_{1,2}}x_2 + n_1,$$
$$y_2 = \sqrt{h_{2,1}}x_1 + \sqrt{h_{2,2}}x_2 + n_2, \tag{1}$$

where x_i is the transmitting signal of Tx-i, with $\|x_i\|_2^2 = P_i$.

To exploit the potential of IC, we employ multi-layer RS at the transmitters. In the proposed scheme, Tx-i's total transmission data rate R_i is split into L_i *splits* by splitting the total transmission power into $[p_{i,1}, \cdots, p_{i,k} \cdots, p_{i,L_i}]$, where $p_{i,k}$ is the allocated power to k-th split of Tx-i and $\sum_{k=1}^{L_i} p_{i,k} = P$. We assume equal-power RS throughout this paper, i.e., $p_{i,k} = P/L_i = p_i$, unless otherwise stated, since equal-power allocation is usually applied as a baseline in analyzing the achievable capacities in different systems with RS [25, 26]. Correspondingly, the transmission signal of Tx-i can be represented as

$$x_i = \sum_{k=1}^{L_i} x_{i,k}, \tag{2}$$

where $\|x_{i,k}\|_2^2 = p_{i,k}$. We note that the elaborately designed power allocation among the message splits may further improve the system performance [27], which is also a promising future direction.

Without loss of generality, we assume $L_1 = L_2 = L$. It should be noted that IC can be regarded as two MACs from the point of view of the receivers and that the SD bound is derived by taking the minimum of the sum capacity of the two MACs, i.e.,

$$r_{\mathrm{SD}} = \min\left\{\log\left(1 + \frac{h_{1,1}P_1 + h_{1,2}P_2}{N_0}\right), \log\left(1 + \frac{h_{2,1}P_1 + h_{2,2}P_2}{N_0}\right)\right\}. \tag{3}$$

To maintain low computational complexity, we apply SSD as well as interference cancellation at each receiver to sequentially decode the signal splits, where the interference splits are regarded as additive Gaussian white noise. Each receiver may first detect a signal split and then an interference split one after another. The successfully decoded splits are then cancelled from the received signal. In our system, we consider a fixed decoding order for a certain $\Pi_{m,l}$. The optimal decoding order and power allocation will be investigated in the future work. We define the notation $\Pi_{l,m}$, $l, m = 1, 2$, to represent the decoding order where at Rx-1, the successive decoding starts from $x_{l,1}$ and, at Rx-2, the successive decoding starts from $x_{m,1}$. Afterwards, the splits of both transmitters are decoded one after another. As an example, when $\Pi_{1,2}$ is assumed, Tx-1 successively decodes and cancels $x_{1,1}$, $x_{2,1}$, $x_{1,2}$, $x_{2,2}$..., and Tx-2 successively decodes and cancels $x_{2,1}$, $x_{1,1}$, $x_{2,2}$, $x_{1,2}$... Therefore, there are a total number of four decoding orders, i.e., $\Pi_{1,1}$, $\Pi_{1,2}$, $\Pi_{2,1}$, and $\Pi_{2,2}$. We visually illustrate this example in Figure 2, where $\Pi_{1,2}$ is assumed and the arrows indicate the decoding order.

The received signal to interference and noise ratio (SINR) of the kth split transmitted from Tx-i to Rx-j with decoding order $\Pi_{l,m}$ is denoted as $s_{j,i}^{\Pi_{l,m},(k)}$. Accordingly, we define the achievable rate of the kth split transmitted from Tx-i at Rx-j with decoding order $\Pi_{l,m}$ as $r_{j,i}^{\Pi_{l,m},(k)}$, which is given by

$$r_{j,i}^{\Pi_{l,m},(k)} = \log\left(1 + s_{j,i}^{\Pi_{l,m},(k)}\right). \tag{4}$$

The coding rate of the message in each split should be equal to the corresponding achievable channel capacity derived in (4) to ensure successful decoding. Assuming perfect interference cancellation, the SINR of each power split is calculated by dividing the received power of this split by noise plus all residual interference. As an example, when decoding order $\Pi_{1,2}$ is assumed, the SINR of each power split is formulated as follows:

$$s_{1,1}^{\Pi_{1,2},(k)} = \frac{h_{1,1}\left(P_1/L\right)}{h_{1,1}\left(P_1/L\right)(L-k) + h_{1,2}\left(P_2/L\right)(L-k+1) + N_0}, \tag{5}$$

$$s_{1,2}^{\Pi_{1,2},(k)} = \frac{h_{1,2}\left(P_2/L\right)}{h_{1,1}\left(P_1/L\right)(L-k) + h_{1,2}\left(P_2/L\right)(L-k) + N_0}, \tag{6}$$

$$s_{2,1}^{\Pi_{1,2},(k)} = \frac{h_{2,1}\left(P_1/L\right)}{h_{2,1}\left(P_1/L\right)(L-k) + h_{2,2}\left(P_2/L\right)(L-k) + N_0}, \tag{7}$$

$$s_{2,2}^{\Pi_{1,2},(k)} = \frac{h_{2,2}\left(P_2/L\right)}{h_{2,2}\left(P_2/L\right)(L-k) + h_{2,1}\left(P_1/L\right)(L-k+1) + N_0}. \quad (8)$$

Furthermore, we define the sum achievable rate of Tx-i's signal at Rx-j with decoding order $\Pi_{l,m}$ as

$$r_{j,i}^{\Pi_{l,m},[L]} = \sum_{k=1}^{L} r_{j,i}^{\Pi_{l,m},(k)}. \quad (9)$$

The decoding order may affect the receiving SINR of each split as well as the achievable rate. Therefore, at transmitter, it is necessary to consider the effect of the decoding order when assigning the rate to each split. Besides, due to the fact that some splits will be decoded by both receivers, the rates of these splits should be carefully assigned such that successful interference cancellations at two receivers can be carried out. With a fixed decoding order $\Pi_{l,m}$, we define a rate matching (RM) operation in this paper, which ensures that the maximum affordable transmission rate is assigned to each split such that the split can be successfully recovered by both Rx-1 and Rx-2. For example, when RM is employed, the transmission rate of kth split of Tx-i, with the decoding order $\Pi_{l,m}$, is defined as $\widehat{r}_i^{\Pi_{l,m},(k)}$, which is given by

$$\widehat{r}_i^{\Pi_{l,m},(k)} = \min\left\{r_{1,i}^{\Pi_{l,m},(k)}, r_{2,i}^{\Pi_{l,m},(k)}\right\}. \quad (10)$$

2.2. Notations. Recall that, in this paper, we aim to study whether RS and SSD can achieve the SD bound, when large even infinite number of layers is available. However, it is nontrivial to directly compare their performances. Hence, the analysis in this paper is organized in incremental steps, as illustrated in the following.

We start with the case where RM is not conducted; i.e., the data rate of each split is only decided by the received SINR of the target receiver with a given SSD order. We denote this scheme where infinite-layer RS and SSD are applied without RM as **EPRSO** (as a short notation of *Equal-Power Rate Splitting without RM*). We note that this scheme is not realistic, since RM is not employed to ensure the success of SSD. To analyze EPRSO, we propose a genie-aided model, where the interference splits are decoded with the help of genie transmitters. Then we study the realistic settings by considering in RM operations. And we denote the scheme applying infinite-layer RS and SSD with RM as **EPRSW** (as a short notation of *Equal-Power Rate Splitting with RM*). Obviously, EPRSO achieves the upper bound capacity performance of EPRSW. Besides, we define the scheme named **f-EPRSW** (as a short notation of *finite-layer Equal-Power Rate Splitting with RM*) where finite-layer RS and SSD with RM are assumed. In the following sections, we first analyze the gap of the achievable sum capacity between SD and EPRSO and then analyze the gap between EPRSO and EPRSW by taking EPRSO as a bridge in comparing SD and EPRSW. Finally, we compare the performance between SD and f-EPRSW.

The performance metric used to compare SD, EPRSO, EPRSW, and f-EPRSW is the achievable sum capacity at the receivers [22, 24]. Note that when the sum rate of the proposed scheme, i.e., EPRSW, is exactly the same as SD, then through time sharing technique and regarding interference as noise, the proposed scheme can also reach other points in the capacity region. Therefore, it is sufficient to study the achievable sum capacity.

3. Performance Analysis of Infinite-Layer RS

In this section, we analyze whether EPRSO and EPRSW can approach SD bound when the splitting number approaches infinity. To begin with, we present some preliminary Lemmas and Theorems.

3.1. Preliminary. In this paragraph, we do not assume that $P_1 = P_2 = P$, since the results derived in the following Lemmas and Theorems still hold with arbitrary P_i. We first show the existence of the limit of $r_{j,i}^{\Pi_{l,m},[L]}$ when $L \longrightarrow +\infty$ according to Lemmas 1 and 2.

Lemma 1 (monotonicity). *Given the decoding order $\Pi_{l,m}$, $r_{j,i}^{\Pi_{l,m},[L]}$ increases with L if $j = 1$ and $l \neq i$, or if $j = 2$ and $m \neq i$, and $r_{j,i}^{\Pi_{l,m},[L]}$ decreases with L if $j = 1$ and $l = i$, or if $j = 2$ and $m = i$.*

Proof. Without loss of generality, we take $\Pi_{1,2}$ as an example, and aim to prove that $r_{1,1}^{\Pi_{1,2},[L]}$ increases with L by mathematical deduction method. The proof consists of two steps, i.e., the base step and the induction step, where $r_{1,1}^{\Pi_{1,2},[L]} < r_{1,1}^{\Pi_{1,2},[L+1]}, \forall L$.

In base step, we aim to prove that $r_{1,1}^{\Pi_{1,2},[1]} < r_{1,1}^{\Pi_{1,2},[2]}$. We first calculate $r_{1,1}^{\Pi_{1,2},[1]}$ and $r_{1,1}^{\Pi_{1,2},[2]}$ by assuming $L = 1$ and 2, respectively. $r_{1,1}^{\Pi_{1,2},[1]}$ and $r_{1,1}^{\Pi_{1,2},[2]}$ are given, respectively, by

$$\log\left(\frac{h_{1,1}P_1}{h_{1,2}P_2 + N_0}\right), \quad \log\left(\frac{h_{1,1}P_1/2}{h_{1,1}P_1/2 + h_{1,2}P_2/2 + N_0}\right) + \log\left(\frac{h_{1,1}P_1/2}{h_{1,2}P_2/2 + N_0}\right). \quad (11)$$

Hence, $r_{1,1}^{\Pi_{1,2},[2]} - r_{1,1}^{\Pi_{1,2},[1]}$ is given by

$$r_{1,1}^{\Pi_{1,2},[2]} - r_{1,1}^{\Pi_{1,2},[1]} = \log\left(\frac{(1/2)\,h_{1,2}h_{11}P_1P_2 + \mho}{(1/4)\,h_{1,2}h_{11}P_1P_2 + \mho}\right) > 0, \quad (12)$$

where $\mho$ is the same term appeared in both numerator and denominator. The proof of the induction step is similar, which is omitted for brevity. Therefore, by mathematical deduction, the statement in Lemma 1 holds. □

Lemma 2 (upper bound). *Given the decoding order $\Pi_{l,m}$, $\lim_{L\longrightarrow+\infty} r_{j,i}^{\Pi_{l,m},[L]}$ is upper bounded by $h_{j,i}P_i/N_0$, if $j = 1$ and $l \neq i$, or if $j = 2$ and $m \neq i$ (or lower bounded by $h_{j,i}P_i/N_0$, if $j = 1$ and $l = i$ or $j = 2$ and $m = i$).*

Proof. We take $r_{1,1}^{\Pi_{1,2},[L]}$ as an example. $r_{1,1}^{\Pi_{1,2},[L]}$ is upper bounded by $\hat{r}_{1,1}^{\Pi_{1,2},[L]}$, which is given by

$$\begin{aligned}\hat{r}_{1,1}^{\Pi_{1,2},[L]} &= \sum_{k=1}^{L} \log\left(1 + \frac{h_{1,1}P_1/L}{N_0}\right) \\ &= \log\left(1 + \frac{h_{1,1}P_1/L}{N_0}\right)^{L}.\end{aligned} \tag{13}$$

Let $L \longrightarrow +\infty$; then we have

$$r_{1,1}^{\Pi_{1,2},[L]} < \hat{r}_{1,1}^{\Pi_{1,2},[L]} = \frac{h_{11}P_1}{N_0}, \tag{14}$$

where $\lim_{x\longrightarrow+\infty} \log(1 + 1/x)^x = 1$. □

According to Lemmas 1 and 2 and the theorem of supremum, there exists a limit of $r_{j,i}^{\Pi_{l,m},[L]}$ when $L \longrightarrow +\infty$, with decoding order $\Pi_{l,m}$. Define

$$r_{j,i}^{\Pi_{l,m},[+\infty]} = \lim_{L\longrightarrow+\infty} r_{j,i}^{\Pi_{l,m},[L]}. \tag{15}$$

where $r_{j,i}^{\Pi_{l,m},[+\infty]}$ is given by the following Theorem 3.

Theorem 3 (limit). *When $L \longrightarrow +\infty$, $r_{j,i}^{\Pi_{l,m},[L]}$ converges to $r_{j,i}^{[+\infty]}$, $i, j = 1, 2$, with any choice of $\Pi_{l,m}$,*

$$r_{1,1}^{[+\infty]} = \frac{h_{1,1}P_1}{h_{1,1}P_1 + h_{1,2}P_2} \log\left(1 + \frac{h_{1,1}P_1 + h_{1,2}P_2}{N_0}\right), \tag{16}$$

$$r_{1,2}^{[+\infty]} = \frac{h_{1,2}P_2}{h_{1,1}P_2 + h_{1,2}P_2} \log\left(1 + \frac{h_{1,1}P_1 + h_{1,2}P_2}{N_0}\right), \tag{17}$$

$$r_{2,1}^{[+\infty]} = \frac{h_{2,1}P_1}{h_{2,1}P_1 + h_{2,2}P_2} \log\left(1 + \frac{h_{2,1}P_1 + h_{2,2}P_2}{N_0}\right), \tag{18}$$

$$r_{2,2}^{[+\infty]} = \frac{h_{2,2}P_2}{h_{2,1}P_1 + h_{2,2}P_2} \log\left(1 + \frac{h_{2,1}P_1 + h_{2,2}P_2}{N_0}\right). \tag{19}$$

Proof. We take $r_{1,1}^{\Pi_{1,m},[+\infty]}$ as an example. We note that $\log(1 + x) \longrightarrow x$ when $x \longrightarrow 0$. Hence, we can remove the log terms in (4), and the limit of $r_{1,1}^{\Pi_{1,m},[L]}$ is given by (20). We define $Z_1 = (h_{1,1}P_1 + h_{1,2}P_2)/(h_{1,1}P_1 + h_{1,2}P_2 + N)$, and $Z_2 = h_{1,1}P_1/(h_{1,1}P_1 + h_{1,2}P_2 + N)$,

$$\begin{aligned}\lim_{L\longrightarrow+\infty} r_{1,1}^{\Pi_{1,m},[L]} &= \lim_{L\longrightarrow+\infty} \sum_{l=0}^{L-1} \frac{h_{1,1}P_1}{L\left(h_{1,1}P_1 + h_{1,2}P_2 + N_0\right) - l\left(h_{1,1}P_1 + N_0 + h_{1,2}P_2\right) - h_{1,1}P_1} \\ &= \lim_{L\longrightarrow+\infty} \sum_{l=0}^{L-1} \frac{h_{1,1}P_1}{L\left(h_{1,1}P_1 + h_{1,2}P_2 + N_0\right)} \frac{1}{1 - (l/L)\left(\left(h_{1,1}P_1 + h_{1,2}P_2\right)/\left(h_{1,1}P_1 + h_{1,2}P_2 + N_0\right)\right) - \left(h_{1,1}P_1/L\left(h_{1,1}P_1 + h_{1,2}P_2 + N_0\right)\right)}\end{aligned} \tag{20}$$

Equation (20) can be rewritten as

$$\begin{aligned}\lim_{L\longrightarrow+\infty} r_{1,1}^{\Pi_{1,m},[L]} &= \lim_{L\longrightarrow+\infty} \sum_{l=0}^{L-1} \frac{Z_2}{L} \frac{1}{1 - ((l+1)/L) Z_1} \\ &= \frac{Z_2}{Z_1} \lim_{L\longrightarrow+\infty} \sum_{l=0}^{L-1} \frac{Z_1}{L} \times \frac{1}{1 - (Z_1/L)(1 + l)} \\ &= \frac{Z_2}{Z_1} \int_0^{Z_1} \frac{1}{1 - x - o(x)} dx \\ &= -\frac{Z_2}{Z_1} \log(1 - x - o(x))\Big|_0^{Z_1} \\ &= -\frac{h_{1,1}P_1}{h_{1,1}P_1 + h_{1,2}P_2} \log\left(1 - \frac{h_{1,1}P_1 + h_{1,2}P_2}{h_{1,1}P_1 + h_{1,2}P_2 + N_0}\right) \\ &= \frac{h_{1,1}P_1}{h_{1,1}P_1 + h_{1,2}P_2} \log\left(1 + \frac{h_{1,1}P_1 + h_{1,2}P_2}{N_0}\right) \\ &= r_{1,1}^{[+\infty]}.\end{aligned} \tag{21}$$

Interestingly, we see that $\Pi_{1,m}$ does not affect the asymptotic behavior of $r_{1,1}^{\Pi_{1,m},[L]}$ when $L \longrightarrow +\infty$. Thus, we can omit the notation of $\Pi_{l,m}$. With the similar approach, we can derive $r_{1,2}^{[+\infty]}$, $r_{2,1}^{[+\infty]}$, and $r_{2,2}^{[+\infty]}$. □

3.2. Analysis of EPRSO. As aforementioned, it is nontrivial to directly find the relationship between the achievable sum capacity between EPRSW and SD, so we firstly study EPRSO as a bridge. As shown in Figure 3, the original IC is decomposed into two virtual MACs with the help of genie Tx-1 and genie Tx-2, and the two MACs do not interfere with each other. We note that the genie transmitters are introduced to convert the original IC to two virtual MACs, where the achievable rates are easier to be computed. With the given channel conditions and decoding order $\Pi_{l,m}$, the achievable rate of the two transmitters in virtual MAC-j, $j = 1, 2$, is given by $r_{j,1}^{\Pi_{l,m},[L]}$ and $r_{j,2}^{\Pi_{l,m},[L]}$, respectively. Meanwhile, Tx-i will have two capacities, i.e., $r_{1,i}^{\Pi_{l,m},[L]}$ and $r_{2,i}^{\Pi_{l,m},[L]}$, dedicated for virtual MAC-1 and 2, respectively. Therefore, the total achievable rate is derived as

$$\begin{aligned}r_{\text{EPRSO}}^{[L]} = &\min\left\{r_{1,1}^{\Pi_{l,m},[L]}, r_{2,1}^{\Pi_{l,m},[L]}\right\} \\ &+ \min\left\{r_{1,2}^{\Pi_{l,m},[L]}, r_{2,2}^{\Pi_{l,m},[L]}\right\},\end{aligned} \tag{22}$$

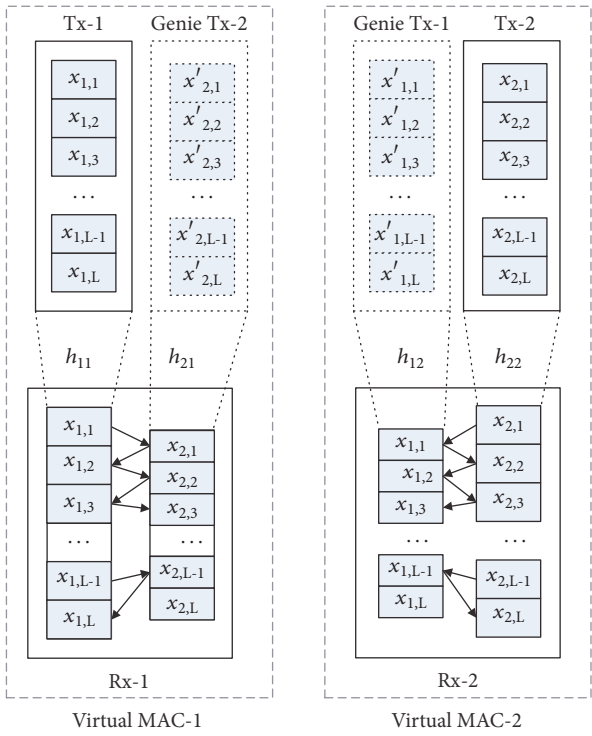


Figure 3: Decomposition of IC into two MACs, with the help of two genie transmitters.

when $L \longrightarrow +\infty$, $r_{\text{EPRSO}}^{[+\infty]}$ is derived as

$$r_{\text{EPRSO}}^{[+\infty]} = \min\left\{r_{1,1}^{[+\infty]}, r_{2,1}^{[+\infty]}\right\} + \min\left\{r_{1,2}^{[+\infty]}, r_{2,2}^{[+\infty]}\right\}. \quad (23)$$

The following theorem demonstrates the sufficient conditions of the channel coefficients, where EPRSO asymptotically approaches the performance of SD.

Theorem 4 (EPRSO). *EPRSO achieves the same performance as SD if both* $(h_{1,1} \leq h_{2,1})$ *and* $(h_{1,2} \leq h_{2,2})$ *hold, or if both* $(h_{1,1} \geq h_{2,1})$ *and* $(h_{1,2} \geq h_{2,2})$ *hold.*

Proof. We take the first condition as an example. When $(h_{1,1} \leq h_{2,1})$ and $(h_{1,2} \leq h_{2,2})$ hold, we have

$$\begin{aligned} r_{1,1}^{\Pi_{l,m},[+\infty]} &\leq r_{2,1}^{\Pi_{l,m},[+\infty]}, \\ r_{1,2}^{\Pi_{l,m},[+\infty]} &\leq r_{2,2}^{\Pi_{l,m},[+\infty]}, \end{aligned} \quad (24)$$

Therefore, the achievable rate is derived as

$$r_{\text{EPRSO}}^{[+\infty]} = r_{1,1}^{[+\infty]} + r_{1,2}^{[+\infty]} = \log\left(1 + \frac{h_{1,1}P + h_{1,2}P}{N_0}\right), \quad (25)$$

which exactly follows the expression of the SD bound in (3). □

The gap between EPRSO and SD is also calculated as

$$r_{\text{SD}} - r_{\text{EPRSO}}^{[+\infty]} = \min\left\{\left|r_{i,j}^{[+\infty]}, r_{j,j}^{[+\infty]}\right|, \left|r_{i,i}^{[+\infty]}, r_{j,i}^{[+\infty]}\right|\right\}, \quad (26)$$

where $i, j = 1, 2$.

Remark 5. When symmetric IC is assumed, i.e., $h_{1,1} = h_{2,2}$ and $h_{1,2} = h_{2,1}$, the gap between EPRSO and SD is derived as

$$r_{\text{SD}} - r_{\text{EPRSO}}^{[+\infty]} = \frac{h_{1,2} - h_{1,1}}{h_{1,2} + h_{1,1}} \log\left(1 + \frac{(h_{1,2} + h_{1,1})P}{N_0}\right). \quad (27)$$

As an example, when $h_{1,1} = 1$ and $h_{1,2} = 0.9$, the loss of EPRSO is about 5% compared to SD.

Remark 6. In symmetric IC, $r_{\text{EPRSO}}^{[+\infty]}$ equals r_{SD} if and only if $h_{1,1} = h_{1,2} = h_{2,1} = h_{2,2}$.

3.3. Analysis of EPRSW. Compared with EPRSO, EPRSW satisfies the individual rate constraint for each power split by employing RM operation. Therefore, the sum rate constraint in (23) is not sufficient. The achievable sum rate of IC with infinite-layer RS, SSD, and RM; i.e., EPRSW, is denoted as $r_{\text{EPRSW}}^{[L]}$

$$\begin{aligned} r_{\text{EPRSW}}^{[L]} = \sum_{k=1}^{L} \Big(&\min\left\{r_{1,1}^{\Pi_{l,m},(k)}, r_{2,1}^{\Pi_{l,m},(k)}\right\} \\ &+ \min\left\{r_{1,2}^{\Pi_{l,m},(k)}, r_{2,2}^{\Pi_{l,m},(k)}\right\}\Big), \end{aligned} \quad (28)$$

where L is the splitting number. The following theorem demonstrates the sufficient conditions where EPRSW approaches EPRSO.

Theorem 7 (EPRSW). *EPRSW achieves the same performance of EPRSO if there exists* $\Pi_{l,m}$*, such that, for any* k_1, k_2*, the following two conditions are satisfied:*

(1)

$$\left(r_{1,1}^{\Pi_{l,m},(k_1)} - r_{2,1}^{\Pi_{l,m},(k_1)}\right)\left(r_{1,1}^{\Pi_{l,m},(k_2)} - r_{2,1}^{\Pi_{l,m},(k_2)}\right) \geq 0, \quad (29)$$

(2)

$$\left(r_{1,2}^{\Pi_{l,m},(k_1)} - r_{2,2}^{\Pi_{l,m},(k_1)}\right)\left(r_{1,2}^{\Pi_{l,m},(k_2)} - r_{2,2}^{\Pi_{l,m},(k_2)}\right) \geq 0, \quad (30)$$

Proof. When the above conditions are satisfied, the min operator in (28) can be taken out of $\sum$ and then (28) is exactly the same as (23). □

Remark 8. In symmetric IC where $h_{11} \geq h_{12}$, we readily see that $r_{1,1}^{\Pi_{l,m},(k_1)} \geq r_{2,1}^{\Pi_{l,m},(k_1)}$ and $r_{2,2}^{\Pi_{l,m},(k_1)} \geq r_{1,2}^{\Pi_{l,m},(k_1)}$, when $\Pi_{1,2}$ is applied. In this case, the sufficient conditions in Theorem 7 are satisfied, which means that no gap exists between EPRSW and EPRSO in symmetric IC and the gap between EPRSW and SD also follows (27). Otherwise, in symmetric IC where $h_{11} \leq h_{12}$, the conditions in Theorem 7 also hold, if the decoding order $\Pi_{2,1}$ is assumed. According to the above analysis, we are ready to say that, infinite-layer RS and SSD can achieve the same capacity region as SD, if the sufficient conditions in both Theorems 4 and 7 are satisfied.

However, due to the implicit expressions of the conditions given in Theorem 7, it is not straightforward to conclude the channel gain settings where EPRSWs achieve the same

performance as SD. In the following, we show that, in most channel gain settings, the conditions of Theorems 4 and 7 cannot be satisfied simultaneously.

When L-layer RS is employed, the gap between EPRSO and EPRSW is derived in (31),

$$\begin{aligned}\Delta^{L}_{\Pi_{l,m}} &= \sum_{k=1}^{L}\Big(\Big|r_{1,1}^{\Pi_{l,m},(k)} - r_{2,1}^{\Pi_{l,m},(k)}\Big| \\ &\quad\cdot \mathrm{I}\Big(-\Big(r_{1,1}^{\Pi_{l,m},(k)} - r_{2,1}^{\Pi_{l,m},(k)}\Big)\Big(r_{1,1}^{+\infty} - r_{2,1}^{+\infty}\Big)\Big) \\ &\quad+\Big|r_{1,2}^{\Pi_{l,m},(k)} - r_{2,2}^{\Pi_{l,m},(k)}\Big| \\ &\quad\cdot \mathrm{I}\Big(-\Big(r_{1,2}^{\Pi_{l,m},(k)} - r_{2,2}^{\Pi_{l,m},(k)}\Big)\Big(r_{1,2}^{+\infty} - r_{2,2}^{+\infty}\Big)\Big)\Big) \\ &= \sum_{k=1}^{L}\sum_{c=1}^{2}\Big|r_{1,c}^{\Pi_{l,m},(k)} - r_{2,c}^{\Pi_{l,m},(k)}\Big|\cdot \mathrm{I}\Big(-\Big(r_{1,c}^{\Pi_{l,m},(k)} \\ &\quad- r_{2,c}^{\Pi_{l,m},(k)}\Big)\Big(r_{1,c}^{+\infty} - r_{2,c}^{+\infty}\Big)\Big),\end{aligned} \tag{31}$$

where $\mathrm{I}(\cdot)$ is an indication function, i.e., $\mathrm{I}(+) = 1$ and $\mathrm{I}(-) = 0$. Without loss of generality, we assume $h_{1,1} > h_{2,1}$, $h_{1,2} > h_{2,2}$, which is the sufficient condition of EPRSO achieving the SD bound. Furthermore, we assume that the receivers follow the decoding order $\Pi_{1,2}$. Since $r_{1,1}^{[+\infty]} \geq r_{2,1}^{[+\infty]}$, to ensure that EPRSW and EPRSW have the same performance, $\Delta^{L}_{\Pi_{1,2}}$ must be equal to 0; i.e., the following condition must hold:

$$r_{1,1}^{\Pi_{1,2},(k)} \geq r_{2,1}^{\Pi_{1,2},(k)}, \quad 1 \leq k \leq K, \tag{32}$$

i.e.,

$$\frac{h_{1,2}}{h_{1,1}}(L-l+1) + \frac{L}{h_{1,1}} \geq \frac{h_{2,2}}{h_{2,1}}(L-l) + \frac{L}{h_{2,1}}, \quad 1 \leq k \leq K. \tag{33}$$

When $k = L-1$, (33) is simplified to

$$2\frac{h_{1,2}}{h_{1,1}} + \frac{L}{h_{1,1}} \geq \frac{h_{2,2}}{h_{2,1}} + \frac{L}{h_{2,1}}. \tag{34}$$

For $L \longrightarrow +\infty$, the terms $h_{1,2}/h_{1,1}$ and $h_{2,2}/h_{2,1}$ can be ignored, and a necessary condition of (34) is $h_{2,1} \geq h_{1,1}$, which contradicts the assumption.

According to the above analysis, with equal-power infinite-layer RS, the sufficient conditions in Theorems 4 and 7 are usually contradictory; i.e., when EPRSO achieves the SD bound, the gap between EPRSO and EPRSW, i.e., $\Delta^{L}_{\Pi_{l,m}}$ is always non-zero. Hence, EPRSW performs slightly worse than SD in general, even if the split number approaches infinite. Besides, in symmetric IC, the achievable sum capacities of EPRSO/EPRSW are no larger than that of SD. Nevertheless, we find that the performance gap between EPRSO/EPRSW and SD is usually pretty small, as further illustrated in Section 5, which makes its finite-layer variant a good tradeoff between complexity and performance.

4. Performance Analysis of Finite-Layer RS

In the previous section, we have analyzed the asymptotic achievable rate of multi-layer RS and SSD in IC, by making an unrealistic assumption where each transmitter employs infinite-layer RS. In this section, we consider the achievable rate where only finite-layer RS is allowed.

4.1. Achievable Sum Capacity with Finite-Layer RS. To analyze the achievable sum capacity with finite-layer RS, we first define the discrete sequence

$$\boldsymbol{r}_i^{\Pi_{l,m}} = \left[r_i^{\Pi_{l,m}}[1], \cdots, r_i^{\Pi_{l,m}}[L], \cdots\right], \quad L \in \mathbb{Z}^+, \tag{35}$$

where its Lth element is given by

$$r_i^{\Pi_{l,m}}[L] = \min\left\{r_{1,i}^{\Pi_{l,m},[L]}, r_{2,i}^{\Pi_{l,m},[L]}\right\}. \tag{36}$$

Then we define the discrete sequence

$$\boldsymbol{r}_{\mathrm{sum}}^{\Pi_{l,m}} = \left[r_{\mathrm{sum}}^{\Pi_{l,m}}[1], \cdots, r_{\mathrm{sum}}^{\Pi_{l,m}}[L], \cdots\right], \tag{37}$$

where $r_{\mathrm{sum}}^{\Pi_{l,m}}[L]$ represents the achievable sum capacity with L-layer equal-power RS and is given by

$$r_{\mathrm{sum}}^{\Pi_{l,m}}[L] = r_1^{\Pi_{l,m}}[L] + r_2^{\Pi_{l,m}}[L]. \tag{38}$$

We note that $r_{\mathrm{sum}}^{\Pi_{l,m}}[L]$ describe the relationship between the achievable sum capacity and the layer number of RS. For illustration purpose, we, respectively, interpolate $r_1^{\Pi_{l,m}}[L]$ and $r_2^{\Pi_{l,m}}[L]$ into two continuous functions, namely, $r_1^{\Pi_{l,m}}(\widehat{L})$ and $r_2^{\Pi_{l,m}}(\widehat{L})$, $\widehat{L} \in \mathbb{R}^+$. And we define $r_{\mathrm{sum}}^{\Pi_{l,m}}(\widehat{L}) = r_1^{\Pi_{l,m}}(\widehat{L}) + r_2^{\Pi_{l,m}}(\widehat{L})$. $r_{\mathrm{sum}}^{\Pi_{l,m}}(\widehat{L})$ can be either monotonically increasing, monotonically decreasing, or convex with an extreme point.

According to Lemma 1, $r_{1,i}^{\Pi_{l,m},[L]}$ increases/decreases with L if $i = l/i \neq l$. Thus, by varying decoding order (note that we have four decoding orders), a total number of four cases exist with respect to the monotonicity of $r_{1,i}^{\Pi_{l,m},[L]}$ and $r_{2,i}^{\Pi_{l,m},[L]}$; i.e., the two variables both increase and both decrease, the former increases and the latter decreases, or the former decreases and the latter increases, with respect to L. We illustrate this in Figure 4. The first two cases in Figure 4 may have two subvariants, according to whether $r_{1,i}^{\Pi_{l,m},[L]}$ and $r_{2,i}^{\Pi_{l,m},[L]}$ intersect. For cases 1-3, $r_i^{\Pi_{l,m}}(\widehat{L})$ is monotone. However, for case 4, there exists an extreme point, denoted as L^*.

As an example, we assume $\Pi_{1,2}$ is applied and the possible shapes of $r_{\mathrm{sum}}^{\Pi_{1,2}}(\widehat{L})$ are illustrated in Figure 5. From Figure 5, we observe that $r_{\mathrm{sum}}^{\Pi_{l,m}}(\widehat{L})$ increases with L when $L \leq L_1^*$. However, when $L > L_1^*$, increasing the number of splitting layers does not always lead to capacity gain (as shown in the right part of Figure 5). According to the above qualitative analysis, we conclude that infinite-layer RS is not always better than finite-layer RS.

4.2. Analysis of EPRSW with Finite-Layer RS. Recall that, in last section, we concluded that the achievable sum capacity of

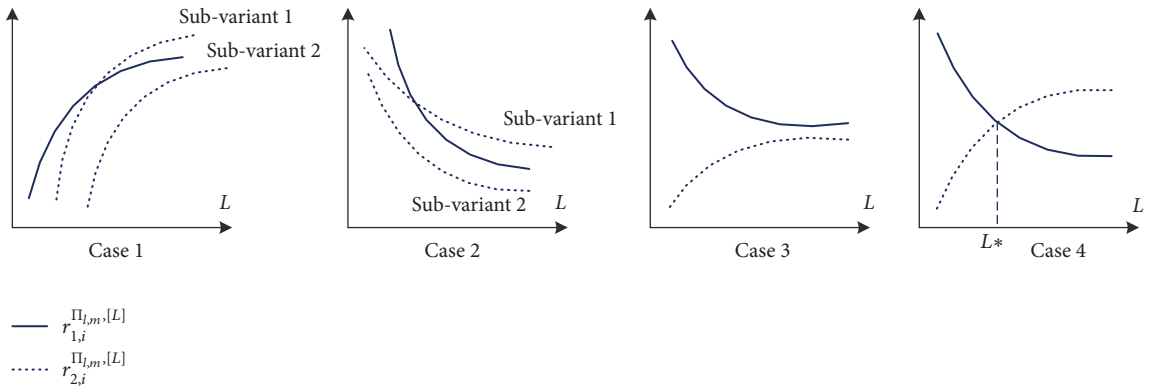


FIGURE 4: Illustrations of the increasing/decreasing property of $r_{1,i}^{\Pi_{l,m},[L]}$ and $r_{2,i}^{\Pi_{l,m},[L]}$ with various decoding orders. For example, if $i = 1$, the decoding orders of the four cases are $\Pi_{1,1}$, $\Pi_{2,2}$, $\Pi_{1,2}$, and $\Pi_{2,1}$, respectively.

EPRSW is usually smaller than that of SD. However, we show in the following that EPRSW with finite-layer RS can actually outperform SD in certain channel conditions.

Consider EPRSW with L-layer RS, as shown in Figure 2. Since the decoding order is set as $\Pi_{1,2}$, the last split of Tx-2, i.e., $x_{2,L}$, does not need to be decoded by Rx-1. Similarly, the last split of Tx-1, i.e., $x_{1,L}$, does not need to be decoded by Rx-2. Therefore, there is no need to conduct RM on the transmission rates of these two splits as defined in (10); i.e., the transmission rates are directly given by $\widetilde{r}_i^{\Pi_{l,m},(k)} = r_{i,i}^{\Pi_{l,m},(k)}$, $i = 1, 2$, which is strictly larger than $\widetilde{r}_i^{\Pi_{l,m},(k)}$. We note that when $L \longrightarrow +\infty$, the gap between $\widetilde{r}_i^{\Pi_{l,m},(k)}$ and $\widehat{r}_i^{\Pi_{l,m},(k)}$ approaches zero, due to the infinitely-small SNR. However, this gap cannot be ignored with finite value of L.

Observing this fact, the achievable sum capacity of EPRSW with finite-layer (in short f-EPRSW), i.e., $r_{\text{f-EPRSW}}^{\Pi_{l,m},[L]}$, is given i

$$r_{\text{f-EPRSW}}^{\Pi_{l,m},[L]} = \begin{cases} \sum_{k=1}^{L-1} \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} + \min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right) + r_{2,2}^{\Pi_{l,m},(L)} + \min\left\{ r_{1,1}^{\Pi_{l,m},(L)}, r_{2,1}^{\Pi_{l,m},(L)} \right\}, & l = 1,\ m = 1 \\ \sum_{k=1}^{L-1} \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} + \min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right) + r_{2,2}^{\Pi_{l,m},(L)} + r_{1,1}^{\Pi_{l,m},(L)}, & l = 1,\ m = 2 \\ \sum_{k=1}^{L} \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} + \min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right), & l = 2,\ m = 1 \\ \sum_{k=1}^{L-1} \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} + \min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right) + r_{1,1}^{\Pi_{l,m},(L)} + \min\left\{ r_{1,2}^{\Pi_{l,m},(L)}, r_{2,2}^{\Pi_{l,m},(L)} \right\}, & l = 2,\ m = 2. \end{cases} \tag{39}$$

The expression of $r_{\text{f-EPRSW}}^{\Pi_{l,m},[L]}$, for arbitrary $l, m \in \{1, 2\}$, can be further synthesized as

$$\begin{aligned} r_{\text{f-EPRSW}}^{\Pi_{l,m},[L]} = & \sum_{k=1}^{L-1} \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} \right. \\ & \left. + \min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right) + \left(r_{2,2}^{\Pi_{l,m},(l)} \right)^{2-l} \\ & \cdot \left(\min\left\{ r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)} \right\} \right)^{l-1} + \left(r_{1,1}^{\Pi_{l,m},(l)} \right)^{m-1} \\ & \cdot \left(\min\left\{ r_{1,1}^{\Pi_{l,m},(l)}, r_{2,1}^{\Pi_{l,m},(l)} \right\} \right)^{2-m}. \end{aligned} \tag{40}$$

In the following, the sufficient channel conditions where f-EPRSW outperforms SD are given in Theorem 9.

Theorem 9 (f-EPRSW). *The proposed f-EPRSW outperforms EPRSO when the conditions (1)-(3) and one of the conditions (4)-(5) are satisfied:*

(1) $\Pi_{l,m} \neq \Pi_{2,1}$;

(2) *The channel coefficients can satisfy the conditions in Theorem 4;*

(3) *The channel coefficients can satisfy the conditions in Theorem 7;*

(4)

$$\frac{1}{\left(P + L/h_{1,2}\right)\left(P + L/h_{1,1}\right)} > \frac{h_{1,2}}{h_{1,1} + h_{1,2}} \log\left(1 + \left(h_{1,1} + h_{1,2}\right) P\right)$$

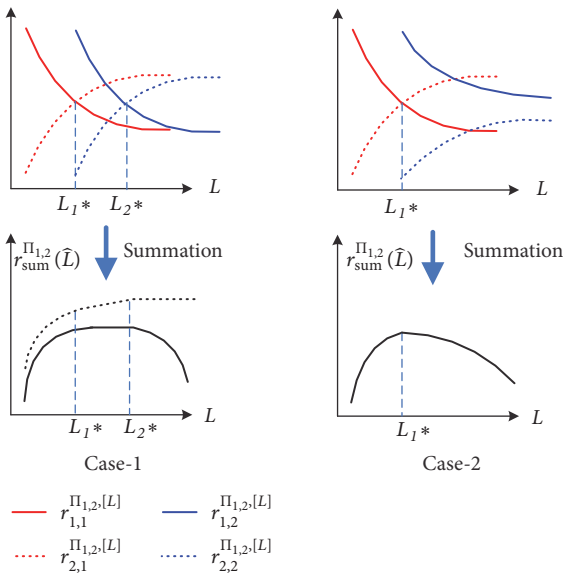


FIGURE 5: Illustrations of $r_{\text{sum}}^{\Pi_{l,m}}(\widehat{L})$. Case-1 and Case-2 show two examples where increasing L does not increase the achievable sum capacity.

$$-\sum_{k=1}^{L}\log\left(1+\frac{h_{1,1}P}{(h_{1,1}+h_{1,2})(L-k)P+L}\right) \tag{41}$$

(5)

$$\frac{1}{(P+L/h_{2,2})(P+L/h_{2,1})} > \frac{h_{2,2}}{h_{2,2}+h_{2,1}}\log\left(1+(h_{2,2}+h_{2,1})P\right) - \sum_{k=1}^{L}\log\left(1+\frac{h_{2,2}P}{(h_{2,2}+h_{2,1})(L-k)P+L}\right) \tag{42}$$

Proof. Without loss of generality, we assume conditions (1)-(4) are satisfied. According to (1) and (2), $\forall\epsilon$, $\exists L^*$, when $L > L^*$, $|r_{\text{EPRSW}}^{[L]} - r_{\text{SD}}| < \epsilon$. Furthermore, according to (4), we readily see that

$$r_{\text{EPRSW}}^{[L^*]} + \left(r_{2,2}^{\Pi_{l,m},(l)}\right)^{2-l}\left(\min\left\{r_{1,2}^{\Pi_{l,m},(l)}, r_{2,2}^{\Pi_{l,m},(l)}\right\}\right)^{l-1} - \min\left\{r_{1,2}^{\Pi_{l,m},(k)}, r_{2,2}^{\Pi_{l,m},(k)}\right\} > r_{\text{SD}}, \tag{43}$$

which indicates that f-EPRSW achieves better capacities than SD. □

The reason why f-EPRSW outperforms SD is because RS and SSD can *transform* the underlying physical channel conditions by recovering and cancelling some signal splits. Furthermore, some splits of the transmitters may not be decoded by the unexpected receivers f-EPRSW, which relaxes the RM requirements and thus enables higher data rates on these splits. As an instance, assume that multi-layer RS is applied, with several iterations of SSD, and the achievable sum capacity is already quite close to the SD bound. Then if we remove the RM requirement on the last split, as is done in f-EPRSW, the achievable rate can surpass the SD bound. Nevertheless, it is still pretty hard to provide the exact expression of the capacity gain, which may be an interesting future research aspect.

5. Numerical Results

In this section, we present some numerical results to verify the above analysis. We assume a two-transmitter two-receiver IC model, where the transmission power of each transmitter is $P = 10$ and the noise variance is $N_0 = 1$. First of all, we show the relationship between the achievable rate of EPRSW and the number of layers in RS, with some typical channel gain settings. Then we compare the achievable sum capacities of EPRSW and SD with general channel gain settings. Finally, we show some special cases where f-EPRSW outperforms SD.

Figure 6 shows the relationship between the achievable rate and the number of layers in RS in different decoding orders and compares the achievable sum capacities of EPRSO, EPRSW, and SD. The channel coefficients are set as $h_{1,1} = h_{2,2} = 1, h_{1,2} = h_{2,1} = 0.9$, which constitute a typical symmetric IC with strong but not very strong interference. Taking Figure 6(a) as an example, we can observe that $r_{2,1}^{\Pi_{1,2},[L]}$ and $r_{1,2}^{\Pi_{1,2},[L]}$ increase with L and are upper bounded by $h_{2,1}(P/L)/N_0$ and $h_{1,2}(P/L)/N_0$, respectively. Meanwhile, $r_{1,1}^{\Pi_{1,2},[L]}$ and $r_{2,2}^{\Pi_{1,2},[L]}$ decrease with L and are lower bounded by $h_{1,1}(P/L)/N_0$ and $h_{2,2}(P/L)/N_0$, respectively. From Figure 6, we also see that $r_{j,i}^{\Pi_{l,m},[L]}$ is independent of $\Pi_{l,m}$ when $L \longrightarrow +\infty$. These results exactly follow the analysis in Section 3.1. Besides, with appropriately decided decoding orders, i.e., $\Pi_{1,2}$ and $\Pi_{2,1}$, the achievable sum capacities increase with L in both EPRSO and EPRSW schemes. However, with $\Pi_{1,1}$ or $\Pi_{2,2}$, RS does not increase the achievable sum capacity. At some L, EPRSO can outperform EPRSW due to the absence of RM. However, there is always a gap between EPRSW and SD, which coincides the conclusion derived in Section 3.3.

In Figure 7, we present the achievable sum capacities of EPRSW versus different L (as shown by the red cycles) as well as the SD capacity region (as shown by the black dotted area). The black arrow indicates the direction of the growth of L. Figure 7(a) shows that larger L leads to better capacity. Meanwhile, Figure 7(b) also shows that better fairness can be achieved with larger L, if $\Pi_{1,1}$ or $\Pi_{2,2}$ is assumed as the decoding orders. The orthogonal multiple access (OMA) achievable rate region is highlighted in Figure 7, which shows that OMA is strictly suboptimal compared with SD and EPRSW.

Figure 8 shows the performance gap between EPRSW and SD with extensive channel gain settings, where $h_{1,1} = 1$ and $h_{1,2}, h_{2,1}$, and $h_{2,2}$ take values within [0.5, 1.5]. Each black cycle represents a set of channel coefficients where EPRSW can achieve the SD bound with a maximum of X% loss, where X=0.5, 1, 5, and 10, respectively. We can observe that when 5% loss is allowed, EPRSW performs almost as good as SD in

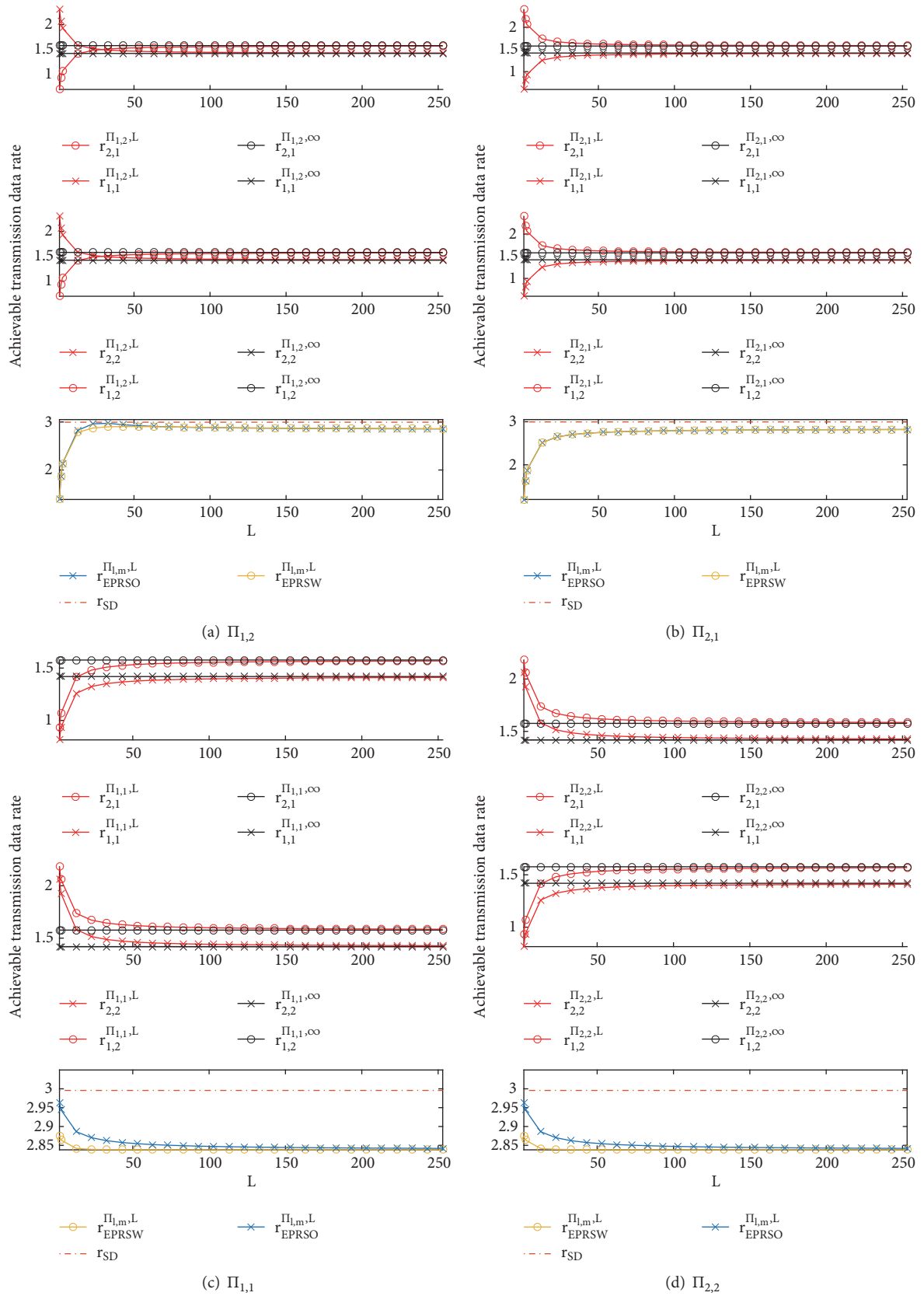
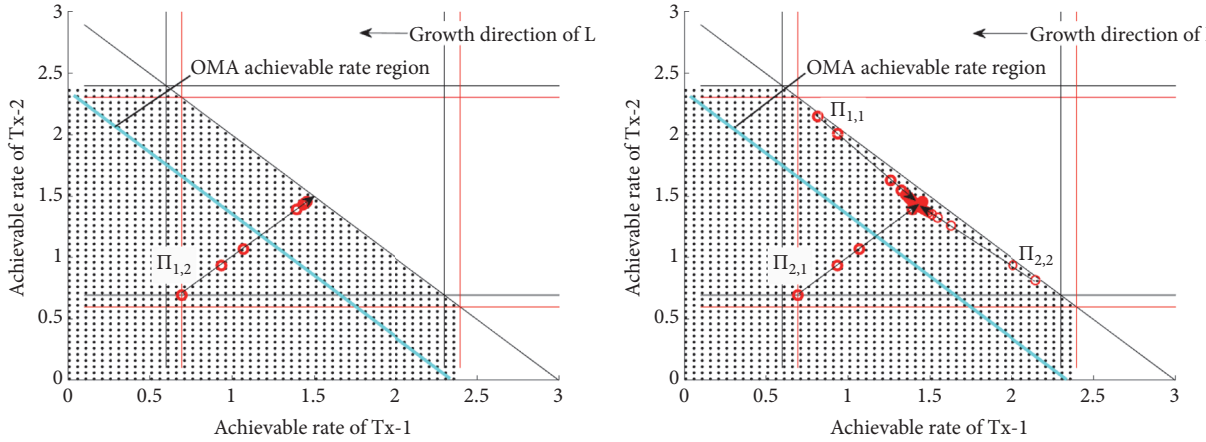


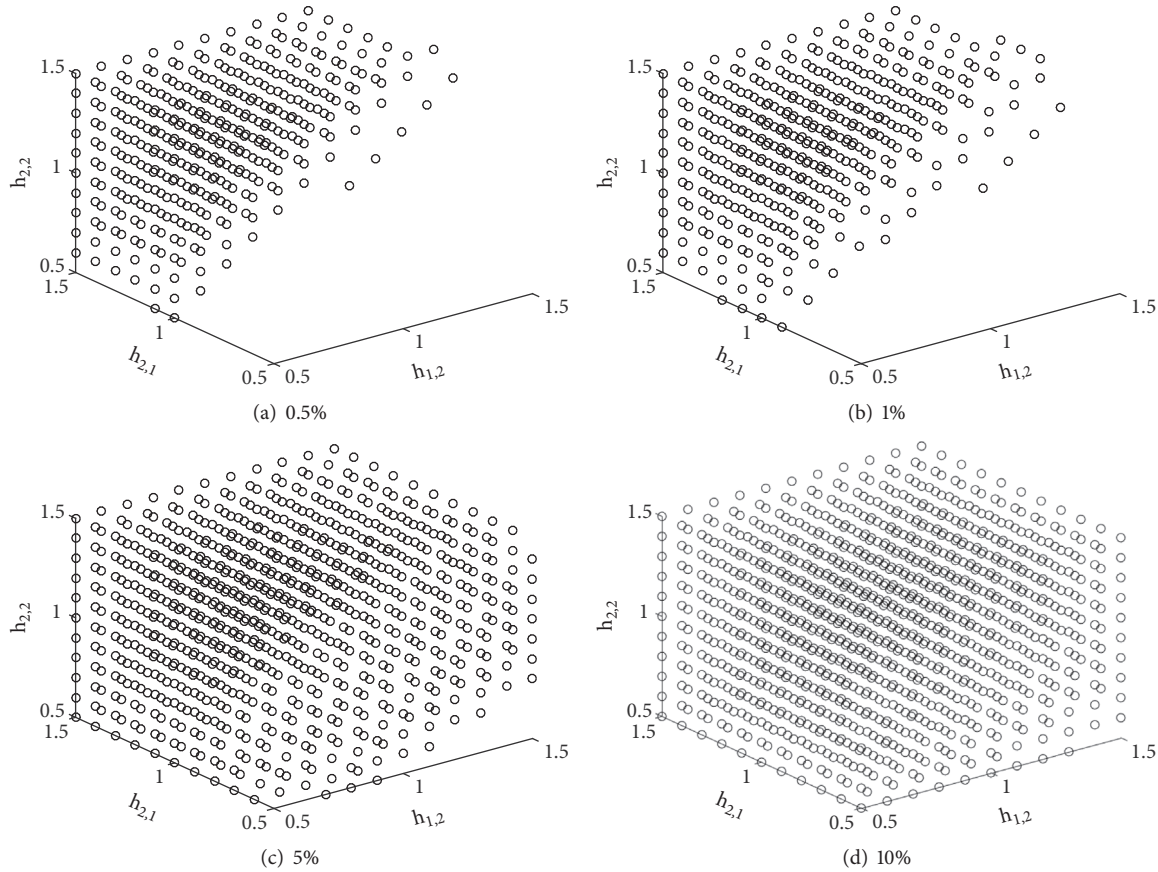
FIGURE 6: Achievable rate vs. the number of layers in RS, with varying decoding order. The channel coefficients are set with $h_{1,1} = h_{2,2} = 1, h_{1,2} = h_{2,1} = 0.9$.



(a) Achievable rate vs. the number of layers in RS with $\Pi_{1,2}$

(b) Achievable rate vs. the number of layers in RS with $\Pi_{1,1}$, $\Pi_{2,1}$, and $\Pi_{2,2}$

FIGURE 7: Achievable rate regions of SD and EPRSW, with different decoding orders. The black dotted area indicates the SD bound, and the black arrows indicates the growth direction of L.



(a) 0.5%

(b) 1%

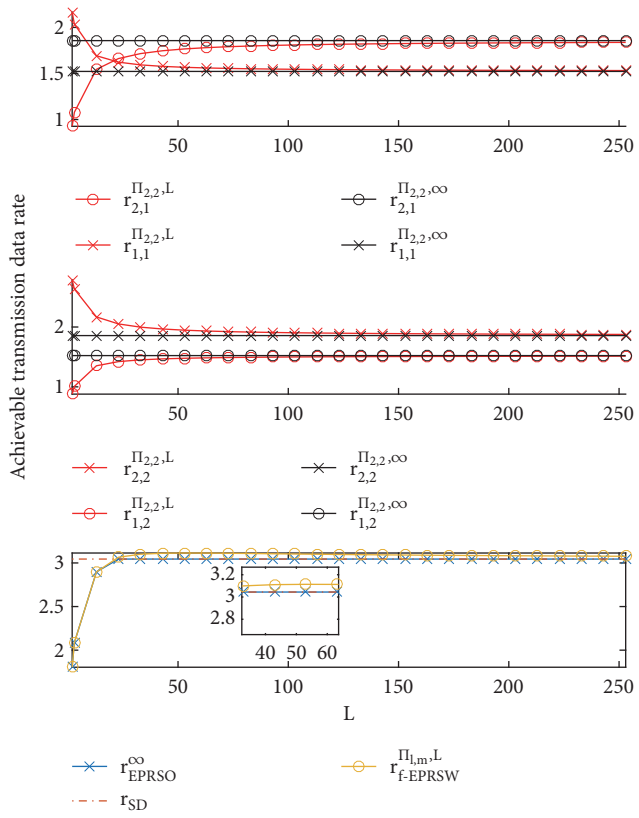
(c) 5%

(d) 10%

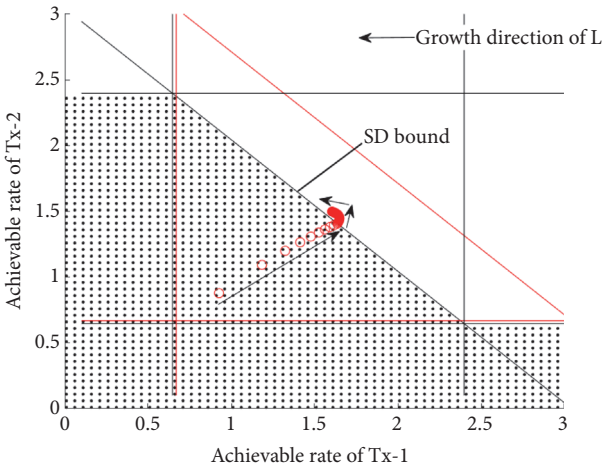
FIGURE 8: A plot of the channel gain settings where EPRSW can achieve the SD bound with a maximum of X% loss (X=0.5, 1, 5, and 10). The channel coefficients are set as follows: $h_{1,1} = 1$ and $h_{1,2}, h_{2,1}, h_{2,2}$ take values within [0.5, 1.5].

most channel gain settings. This reflects that the performance gap between EPRSW and SD is rather small.

In Figure 9, we present a case where f-EPRSW can outperform SD. The channel coefficients are set as $h_{1,1} = h_{1,2} = 2$ and $h_{2,1} = h_{2,2} = 1$, which indicates the strong interference situation. The achieved capacity of f-EPRSW with SSD surpasses the boundary of SD region, with the increase of L. When channel coefficients vary slightly from



(a) Achievable rate vs. the number of layers in RS with $\Pi_{1,2}$



(b) Comparison between the achievable rate region of SD and that of f-EPRSW, varying L

FIGURE 9: A case where f-EPRSW outperforms SD, with $h_{1,1} = 2, h_{1,2} = 2, h_{2,1} = 1$, and $h_{2,2} = 1$.

this setting, it is also observed that f-EPRSW can outperform SD. Hence, the case in Figure 9 is not an isolated evidence.

6. Conclusion and Future Work

In this paper, we have studied a fundamental problem in the Gaussian IC: whether multi-layer RS and SSD can achieve the SD capacity bound. The analysis in this paper shows that the achievable sum capacity of the EPRSW scheme with equal-power infinite-layer RS and SSD cannot reach, but can be pretty close to the SD achievable bound in IC. The exact capacity loss of EPRSW compared with SD was derived in symmetric IC. Nevertheless, the proposed f-EPRSW scheme, which employs equal-power finite-layer RS, SSD, and suitable transmission rate assignment, can even outperforms SD in certain channel gain settings. Therefore, we can conclude that applying RS and SSD is not always weaker than SD, at least when multiple layers and suitable assignment method are employed. At last, we note that extending the proposed scheme and the analysis into the multiuser case would be an interesting future research direction.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by Beijing Major Science and Technology Projects (D171100006317001) and in part by 111 Project of China under Grant B14010.

References

[1] A. B. Carleial, "Interference Channels," *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 60–70, 1978.

[2] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: an information theoretic perspective," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 894–914, 2009.

[3] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.

[4] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving Sustainable Ultra-Dense Heterogeneous Networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[5] S. Lagen, A. Agustin, and J. Vidal, "Coexisting linear and widely linear transceivers in the MIMO interference channel," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 652–664, 2016.

[6] H. Sato, "The Capacity of the Gaussian Interference Channel Under Strong Interference," *IEEE Transactions on Information Theory*, vol. 27, no. 6, pp. 786–788, 1981.

[7] A. A. El Gamal and M. H. Costa, "The capacity region of a class of deterministic interference channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 28, no. 2, pp. 343–346, 1982.

[8] R. Kolte, A. Ozgur, and H. Permuter, "The capacity region of a class of deterministic state-dependent Z-interference channels," in *Proceedings of the 2014 IEEE International Symposium on Information Theory, ISIT 2014*, pp. 656–660, USA, July 2014.

[9] L. Zhou and W. Yu, "Gaussian Z-interference channel with a relay link: achievability region and asymptotic sum capacity," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 58, no. 4, pp. 2413–2426, 2012.

[10] S. Zhao, T. Zhang, Z. Zeng, and Y. Cao, "The Diversity-Multiplexing Tradeoff of One-Side Interference Channel with Relay," in *Proceedings of the 2010 IEEE Vehicular Technology Conference (VTC 2010-Fall)*, pp. 1–5, Ottawa, ON, Canada, September 2010.

[11] K. Mohanty and M. K. Varanasi, "The generalized degrees of freedom region of the MIMO Z-inteference channel with delayed CSIT," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 64, no. 1, pp. 531–546, 2018.

[12] C. Hellings and W. Utschick, "Improper Signaling versus Time-Sharing in the SISO Z-Interference Channel," *IEEE Communications Letters*, 2017.

[13] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 27, no. 1, pp. 49–60, 1981.

[14] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, 2008.

[15] B. Bandemer, A. E. Gamal, and Y.-H. Kim, "Simultaneous nonunique decoding is rate-optimal," in *Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012*, pp. 9–16, USA, October 2012.

[16] A. J. Grant, B. Rimoldi, R. L. Urbanke, and P. A. Whiting, "Rate-splitting multiple access for discrete memoryless channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 47, no. 3, pp. 873–890, 2001.

[17] N. Ye, A. Wang, X. Li, W. Liu, X. Hou, and H. Yu, "On Constellation Rotation of NOMA With SIC Receiver," *IEEE Communications Letters*, vol. 22, no. 3, pp. 514–517, 2018.

[18] Z. Wei, D. W. Ng, and J. Yuan, "Joint Pilot and Payload Power Control for Uplink MIMO-NOMA With MRC-SIC Receivers," *IEEE Communications Letters*, vol. 22, no. 4, pp. 692–695, 2018.

[19] J. Cao and E. M. Yeh, "Asymptotically optimal multiple-access communication via distributed rate splitting," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 53, no. 1, pp. 304–319, 2007.

[20] E. Sasoglu, "Successive cancellation for cyclic interference channels," in *Proceedings of the 2008 IEEE Information Theory Workshop (ITW)*, pp. 36–40, Porto, Portugal, May 2008.

[21] H. Yagi and H. V. Poor, "Multi-level rate-splitting for synchronous and asynchronous interference channels," in *Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings, ISIT 2011*, pp. 2080–2084, Russia, August 2011.

[22] O. Fawzi and I. Savov, *Rate-splitting in the presence of multiple receivers*, 2012, http://arxiv.org/abs/1207.0543.

[23] Y. Zhao, C. W. Tan, A. S. Avestimehr, S. N. Diggavi, and G. J. Pottie, "On the maximum achievable sum-rate with successive decoding in interference channels," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 58, no. 6, pp. 3798–3820, 2012.

[24] L. Wang, E. Sasoglu, and Y. Kim, "Sliding-window superposition coding for interference networks," in *Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT)*, pp. 2749–2753, Honolulu, HI, USA, June 2014.

[25] C. Hao, B. Rassouli, and B. Clerckx, "Achievable DoF regionso of MIMO networks with imperfect CSIT," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 63, no. 10, pp. 6587–6606, 2017.

[26] E. Piovano and B. Clerckx, "Optimal DoF region of the K-User MISO BC with Partial CSIT," *IEEE Communications Letters*, 2017.

[27] Z. Chen, Y. Dong, P. Fan, D. O. Wu, and K. B. Letaief, "Multiple-Layer Power Allocation for Two-User Gaussian Interference Channel," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9161–9176, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6497340, 11 pages
https://doi.org/10.1155/2018/6497340

Research Article

The Rayleigh Fading Channel Prediction via Deep Learning

Run-Fa Liao,[1] Hong Wen,[1] Jinsong Wu,[2] Huanhuan Song,[1] Fei Pan,[1] and Lian Dong[1]

[1]*The National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*Department of Electrical Engineering, Universidad de Chile, Santiago 833-0072, Chile*

Correspondence should be addressed to Hong Wen; wcdma_2000@hotmail.com

Received 25 January 2018; Revised 26 April 2018; Accepted 4 June 2018; Published 25 July 2018

Academic Editor: Dajana Cassioli

This paper presents a multi-time channel prediction system based on backpropagation (BP) neural network with multi-hidden layers, which can predict channel information effectively and benefit for massive MIMO performance, power control, and artificial noise physical layer security scheme design. Meanwhile, an early stopping strategy to avoid the overfitting of BP neural network is introduced. By comparing the predicted normalized mean square error (NMSE), the simulation results show that the performances of the proposed scheme are extremely improved. Moreover, a sparse channel sample construction method is proposed, which saves system resources effectively without weakening performances.

1. Introduction

The future wireless communications (5G) put forward the demands of high-speed transmission, quick access, high reliability, and strong security communications [1]. Hence, new technologies should be adopted to meet the high-speed and high-efficiency transmissions and access demands of 5G [2]. Massive MIMO, non-orthogonal multiple access (NOMA), and tight cooperation for wireless senor nodes are expected to become key technologies for the future 5G systems. A major limitation for massive MIMO, NOMA, and coordinated multipoint (CoMP) systems is the channel state information (CSI) knowledge at the transmitter, which can be obtained partly by the channel prediction techniques. Meanwhile, the physical layer security methods utilize channel reciprocity and diversity to accomplish the so-called "encryption" in the physical layer [3, 4]. Compared with the conventional cryptographic technologies, under the same security requirement, the key length in physical layer security is greatly reduced, and even not required, which is especially suitable for the quick access system. Unfortunately, physical layer security transmission is only dependent on physical CSI. For wireless fading channels, the change of channel information is not conducive to the implementation of the physical MIMO, cooperation, and the security. As shown in Figure 1, the channel information is constantly changing due to the change of the location of the legal receiver, which makes the base station unable to perform robust precoding or beamforming etc. Therefore, the channel prediction is a key point for such problems.

There are considerable research results on the channel parameter prediction. The literatures [5–8] employ the optimal linear algorithm and autoregressive tracing algorithm to predict the flat fading channel, in which channel impulse response prediction is performed by linearly combining the current CSI with the past one. In [5], performance analysis is carried out for long range prediction (LRP) under the actual channel model and the stationary random phase model. Complex-valued neural networks are discussed by T. Ding and A. Hirose to predict time-varying channels and applying them on the hardware [9]. The error rates, compared to the traditional methods, have made improvements. The article [10] utilizes the echo state network to predict channel and proposes a fixed weight method, to reduce computing complexity. Literature [11] proposed a novel support vector machine method to predict a more sophisticated environment. The MUSIC algorithm for channel prediction is investigated in [12]. However, the above-mentioned algorithms either have the lack of high estimation error rate or suffer from high complexity. The major flaws of these methods are

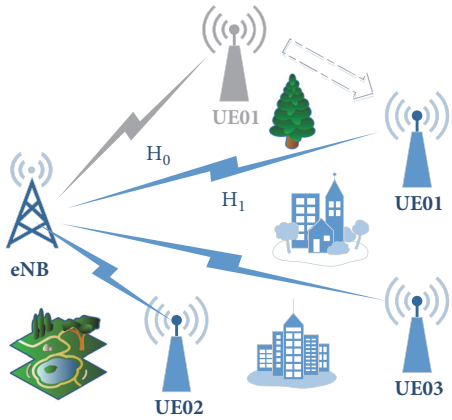


Figure 1: Predicting scene.

that all of them only predict the parameters for the next moment without providing the prediction of CSI after the multiple moments.

In this paper, the multi-time channel prediction system is proposed by taking advantage of the BP neural network with the single hidden layer. Hinton puts forward the concept of deep learning in the year of 2006, which is actually the multi-hidden-layer multi-sensor neural network, including BP neural network and convolutional network [13]. Deep learning is often used in computer vision, pattern recognition, and image classification [13–15]. In this paper, we employ deep learning for wireless channel prediction, while the early stopping strategy is adopted to avoid overfitting [16]. In addition, two-sample construction schemes, namely, the sparse sample construction scheme (SSCS) and normal samples construction scheme (NSCS), are proposed, which can reduce the computational cost and guarantee the prediction accuracy.

We adopt LTE standard frame structure, and the length of the frame is set to 10ms. Specifically, each frame is stratified into 10 subframes, and each subframe consists of two time slots. We assume that each time slot uses 1/3 overhead comb-type pilots. The single-time channel prediction system can save 1/10 pilot resources, and it can save 1/60 system resources. The multi-time algorithm we proposed can save 1/2 or 2/3 pilot resources, which can save 1/6 or 2/9 system resources. So the algorithm we proposed in this paper is very meaningful and useful.

This article is structured as follows. Section 2 introduced the Rayleigh fading channel model and BP neural network, which will be the basis of our novel mothed. The multi-time channel prediction system that can predict the channel information at multiple moments is presented in Section 3. Section 4 includes the simulation results and analyses. Conclusions are given in Section 5.

2. Preliminary

The symbols used in this article will be briefly described. Uppercase bold letters are used for the matrix and lowercase bold letters for vectors. The elements are represented by the letters with subscripts and not bold. The n^{th} vector and the n^{th} samples are presented by the superscripts with round brackets.

2.1. Rayleigh Fading Channel Model. The propagation in any wireless channel is either a line-of-sight (LOS) propagation or a non-line-of-sight (NLOS) propagation. The probability density function (PDF) of a received signal in LOS environment obeys the Rician distribution, while the PDF of the received signal in the NLOS environment obeys the Rayleigh distribution. We can form a Rayleigh channel by scattering components without a direct path, which can be expressed as follows [17–19]:

$$h(t) = \sum_{n=1}^{N} a_n e^{j(2\pi f_n t + \varphi_n)} \tag{1}$$

where N is the number of multipaths and a_n is the amplitude of the n^{th} path. f_n, φ_n represent the Doppler frequency shift and the phase of the n^{th} path, respectively. The Doppler frequency shift is expressed as $f_n = (v/c) f_c \cos\theta_n$, where v is the moving speed of the user. c is the speed of light, f_c is the carrier frequency, and θ_n is the angle between the user's moving direction and the incident radio wave angle.

The sharp Rayleigh fading channels conforming to a given Doppler spectrum are generated by complex sine wave synthesis, just like the Jakes' channel model [20]. The final channel information of the Jakes model is complex-valued, which is given by the following:

$$h(t) = h_I(t) + jh_Q(t) \tag{2}$$

In this paper, the deep learning samples are sampled at different transmission time slots, and the associated complex-valued CSIs of h (t) can be divided into real and imaginary parts. Accordingly, we predict the real value and the imaginary value of channel state information separately. The related processing procedure is given by the following:

$$h(kT_s) = h_I(kT_s) + jh_Q(kT_s) \tag{3}$$

Then, we construct the deep learning samples by capturing channel information $h_I(kT_s)$, $h_Q(kT_s)$. Finally, BP neural network is adopted to predict the channel information at a later time based on learning from the channel information of the past time slots.

2.2. Back Propagation (BP) Neural Network. Hornik proved that the multi-layer feed forward network containing enough neurons in the hidden layer can approach a continuous function of arbitrary complexity and precision [21]. The deep learning technology has been extensively used in computer vision, pattern recognition, and image classification [13–15]. In this paper, the deep learning is exploited to match the fading channel changing trajectory and to achieve channel prediction. Backpropagation (BP) algorithm multi-layer feed forward neural network prediction model is used to predict the fading channel. Figure 2 illustrates a typical multi-input neural network, which includes an input layer, a hidden layer,

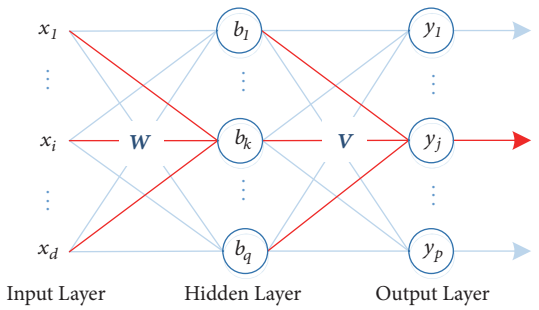


FIGURE 2: Single hidden layer BP neural network.

and an output layer. In the field of machine learning, the neural network as shown in Figure 2 is generally called two-layer neural network (the input layer does not count), or a single hidden layer neural network. We will adopt this statement in this paper.

Given the training sample set,

$$D = \left\{\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right), \left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right), \ldots, \left(\mathbf{x}^{(\mathrm{m})}, \mathbf{y}^{(\mathrm{m})}\right)\right\}, \quad \mathbf{x}^{(k)} \in \mathbb{R}^d, \ \mathbf{y}^{(\mathrm{k})} \in \mathbb{R}^p \tag{4}$$

where $\mathbf{x}^{(k)}, \mathbf{y}^{(k)}$ denote the k^{th} input sample and the k^{th} output sample, respectively. d and p ar the dimension of the input sample and output sample, respectively. As discussed above, we use the lowercase bold letters with parentheses superscript, for example, $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ represents the k^{th} input-output sample.

Then we get the input sample matrix,

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(\mathrm{m})}\right], \quad \mathbf{X} \in \mathbb{R}^{d \times m}, \tag{5}$$

and the output sample matrix,

$$\mathbf{Y} = \left[\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(\mathrm{m})}\right], \quad \mathbf{Y} \in \mathbb{R}^{p \times m} \tag{6}$$

The input value of any node in neural network is the previous neuron multiplied by the weight plus the threshold and then activated by the activation function. Without loss of generality, taking the k-th hidden neuron as an example, $b_k = g_1(\sum_{i=1}^{d} x_i W_{ik} + \xi_k)$ [13]. This paper will use the vectorization description to the neural network transmission formula. The neural network forward propagation vectorization is expressed as follows:

$$\begin{aligned} \mathbf{Z}_1 &= \mathbf{W}^T \mathbf{X} + \boldsymbol{\Xi} \\ \mathbf{B} &= g_1\left(\mathbf{Z}_1\right) \\ \mathbf{Z}_2 &= \mathbf{V}^T \mathbf{B} + \boldsymbol{\Theta} \\ \hat{\mathbf{Y}} &= g_2\left(\mathbf{Z}_2\right) \\ L\left(\hat{\mathbf{y}}, \mathbf{y}\right) &= \frac{1}{2} \sum_{j=1}^{p} \left(\hat{y}_j - y_j\right)^2 \\ J\left(\hat{\mathbf{Y}}, \mathbf{Y}\right) &= \frac{1}{m} \sum_{i=1}^{m} L\left(\hat{\mathbf{y}}^{(\mathrm{i})}, \mathbf{y}^{(\mathrm{i})}\right) \end{aligned} \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d \times q}$ is the weight matrix connected the input layer and the hidden layer, and $\mathbf{V} \in \mathbb{R}^{q \times p}$ is the weight matrix connected the hidden layer and the output layer. $\boldsymbol{\Xi} = \underbrace{[\boldsymbol{\xi}, \boldsymbol{\xi}, \cdots, \boldsymbol{\xi}]}_{m}$ is the hidden layer threshold matrix which consists of the hidden layer threshold vector $\boldsymbol{\xi}$ and $\boldsymbol{\Theta} = \underbrace{[\boldsymbol{\theta}, \boldsymbol{\theta}, \cdots, \boldsymbol{\theta}]}_{m}$ is the output layer threshold matrix which consists of the output layer threshold vector $\boldsymbol{\theta}$. $\mathbf{Z}_1$ is the hidden layer input matrix and $\mathbf{B}$ is the hidden layer output matrix. $\mathbf{Z}_2$ is the input matrix for the output layer. $\hat{\mathbf{Y}}$ is the output vector of the output layer which is also the final output of the neural network. g_1, g_2 are the hidden layer and output layer activation functions, respectively. Note that g_1 is always the sigmoid activation function and g_2 is the purelin activation function. The functions operating on vector or matrix mean act on each element separately (e.g., $f(\mathbf{x}) = (f(1), f(2))$, $\mathbf{x} = (1\ 2)$). Additionally, its dimension is the same as the original vector or matrix. $L(\hat{\mathbf{y}}, \mathbf{y})$ is the loss function. We adopt the mean square error (MSE) of the output as the loss function. $J(\hat{\mathbf{Y}}, \mathbf{Y})$ is the cost function, and it is equal to the average of the loss function with m samples. The neural network iteratively updates the network weight matrix and the threshold vector by minimizing the cost function $J(\hat{\mathbf{Y}}, \mathbf{Y})$.

We get $\hat{\mathbf{y}}$ by forward propagation and then update the weight matrix $\mathbf{W}$, $\mathbf{V}$ and threshold vector $\boldsymbol{\xi}$, θ by backpropagation with the gradient descent method. For convenience, the partial derivative of the cost function $J(\hat{\mathbf{Y}}, \mathbf{Y})$ to output $\hat{\mathbf{Y}}$ is denoted by $d\hat{\mathbf{Y}}$, which is $d\hat{\mathbf{Y}} = \partial J(\hat{\mathbf{Y}}, \mathbf{Y}) / \partial \hat{\mathbf{Y}}$. The vectorization representation of neural network backpropagation iteration formulas is given by the following:

$$\begin{aligned} d\mathbf{Z}_2 &= \frac{\partial J\left(\hat{\mathbf{Y}}, \mathbf{Y}\right)}{\partial \hat{\mathbf{Y}}} * g_2{}'\left(\mathbf{Z}_2\right) \\ d\mathbf{V} &= \frac{1}{m} \mathbf{B} \left(d\mathbf{Z}_2\right)^T \\ d\boldsymbol{\theta} &= \frac{1}{m} d\mathbf{Z}_2 \cdot \mathbf{e}_1 \\ d\mathbf{Z}_1 &= \mathbf{V} \cdot d\mathbf{Z}_2 * g_1{}'\left(\mathbf{Z}_1\right) \\ d\mathbf{W} &= \frac{1}{m} \mathbf{X} \left(d\mathbf{Z}_1\right)^T \\ d\boldsymbol{\xi} &= \frac{1}{m} d\mathbf{Z}_1 \cdot \mathbf{e}_1 \end{aligned} \tag{8}$$

$$\tag{9}$$

where the symbol $*$ denotes the elements in matrix (or vector) $\boldsymbol{A}$ and $\mathbf{B}$ multiplied correspondingly. $(\cdot)^T$ means the matrix transpose. $g'(\cdot)$ represents the function derivation, and vector $\mathbf{e}_1$ satisfies equation $\mathbf{e}_1 = \underbrace{(1, 1, \cdots, 1)}_{m}$. The parameter update rule is expressed as $\tau \longleftarrow \tau - \Delta\tau$. For example, the parameter $\mathbf{W}$ is adaptively updated in the form of $\mathbf{W} = \mathbf{W} - \eta d\mathbf{W}$, where η is the learning rate.

As we know, the regularization, dropout, and early stopping strategies are employed to prevent the overfitting of the neural network [21, 22]. This paper adopts the early stopping strategy to avoid the risk of overfitting; we divide the input

and output samples into training set, verification set, and test set. The training set is used to calculate the gradient and update the link weight and threshold. The verification set is used to estimate the error. If the training set error decreases with the validation set error increasing, training process will stop and the related weights and thresholds with the smallest validation set error will be returned.

In summary, combined with early stopping strategy, the gradient descent method is used to continuously update the neural network information, i.e., weight matrix $\mathbf{W}$, $\mathbf{V}$ and threshold vector $\boldsymbol{\xi}$, $\boldsymbol{\theta}$. Once the parameters update is completed, the neural network will be ready to predict the channel state information (CSI).

3. Rayleigh Fading Channel Prediction Method

The multi-time channel prediction system with single hidden layer effectively predict the channel information at multiple moments, and the proposed deep learning prediction system of the multi-layer neural network can cope with a more sophisticated channel information prediction.

3.1. Prediction Scheme

3.1.1. Single-Time Prediction Scheme. Literature [10] proposed the prediction scheme of channel prediction through the echo state network (ESN). In [10], the channel information of the first n moments is regarded as the input samples and the channel information at the $(n+1)^{th}$ moment as the output samples. And the samples construction scheme of prediction system in this section will follow this program. Let the output layer's dimension be, which is mentioned above, equal to 1. That would be a single-time channel prediction system. Moreover, we add the estimation error at channel state information, which is

$$\widetilde{h}(t) = h(t) + n(t) \tag{10}$$

where $n(t)$ is the Gaussian white noise, and its variance is σ_n^2. We construct the following neural network training sample [10]:

$$\begin{aligned} \mathbf{x}^{(k)} &= \left[\widetilde{h}(kT_s), \widetilde{h}((k+1)T_s), \ldots, \widetilde{h}((k+d-1)T_s)\right] \\ y^{(\mathrm{k})} &= \widetilde{h}((k+d)T_s) \end{aligned} \tag{11}$$

For the k^{th} input and output sample set $(\mathbf{x}^{(k)}, y^{(k)})$, the input sample $\mathbf{x}^{(k)}$ represents the channel information samples at the k^{th} time and the next $(d-1)$ times, and the output sample $y^{(k)}$ is the channel information samples at the $(k+n)^{th}$ time. We choose n_T training samples to train the neural network. And the training set matrix is given by

$$\begin{aligned} \mathbf{X}_{nT} &= \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n_{\mathrm{T}})}\right] \\ \mathbf{y}_{nT} &= \left[y^{(1)}, y^{(2)}, \cdots, y^{(n_T)}\right] \end{aligned} \tag{12}$$

The test set matrix consisting of the next n_R samples will test the neural network, as shown below:

$$\begin{aligned} \mathbf{X}_{nR} &= \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(n_{\mathrm{R}})}\right] \\ \mathbf{y}_{nR} &= \left[\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(n_{\mathrm{R}})}\right] \end{aligned} \tag{13}$$

The training set $(\mathbf{X}_{nT}, \mathbf{y}_{nT})$ is used to train the neural network, while the test set $(\mathbf{X}_{nR}, \mathbf{y}_{nR})$ will be utilized to measure the performance of neural network.

3.1.2. Multi-Time Prediction System. The existing researches of channel prediction [5–12] are either not accurate enough or too complicated to perform in resource-constrained sensor networks. The echo state network (ESN) channel prediction [10] greatly improves the system predictive performance and reduces the system complexity properly. But it only can predict channel information at the next moment. In this section, a multi-time prediction system is exploited to predict channel information at multiple time slots. It achieves a stronger engineering performance. Furthermore, we propose two-sample construction methods, i.e., *sparse sample construction scheme* (SSCS) and *normal samples construction scheme* (NSCS). In the following, we will discuss these two schemes in detail.

(a) Normal Samples Construction Scheme (NSCS). The normal samples construction scheme is a continuous sample construction method. It just adds more outputs on Y. *Zhao's* scheme. The k^{th} input and output samples are expressed as follows:

$$\begin{aligned} \mathbf{x}^{(k)} &= \left[\widetilde{h}(kT_s), \widetilde{h}((k+1)T_s), \ldots, \widetilde{h}((k+d-1)T_s)\right] \\ \mathbf{y}^{(\mathrm{k})} &= \left[\widetilde{h}((k+d)T_s), \widetilde{h}((k+d+1)T_s), \ldots, \right. \\ &\quad \left. \widetilde{h}((k+d+q-1)T_s)\right] \end{aligned} \tag{14}$$

The NSCS takes full use of the channel information, but it increases the amount of computation. For example, when $d = 10, q = 10$, we can construct 4990 training samples from 5000 channel information sample values. There is only one channel data difference between two adjacent training samples.

(b) Sparse Samples Construction Scheme (SSCS). In order to reduce the computational complexity, we propose a sparse sample construction scheme. That is, there is no duplicate channel state information for any two input sample information, as follows:

$$\begin{aligned} \mathbf{x}^{(k)} &= \left[\widetilde{h}(k \cdot dT_s), \widetilde{h}((k \cdot d+1)T_s), \ldots, \right. \\ &\quad \left. \widetilde{h}((k \cdot d+d-1)T_s)\right] \\ \mathbf{y}^{(k)} &= \left[\widetilde{h}(((k+1) \cdot d)T_s), \widetilde{h}(((k+1) \cdot d+1)T_s), \ldots, \right. \\ &\quad \left. \widetilde{h}(((k+1) \cdot d+q-1)T_s)\right] \end{aligned} \tag{15}$$

The SSCS only needs 500 samples to traverse 5000 channel information sample values, if $d = 10$, $q = 10$.

1. The weight matrices $\mathbf{W}$, $\mathbf{v}$ are initialized randomly from 0 to 1. The threshold vectors $\boldsymbol{\xi}$, θ initialized to 0. Set the training goal $\varepsilon_{\text{goal}}$ and learning rate η a reasonable value, respectively;
2. Input the channel information training set $(\mathbf{X}_{nT}, \mathbf{y}_{nT})$.
3. while $J(\hat{\mathbf{y}}_{nT}, \mathbf{y}_{nT}) > \varepsilon_{\text{goal}}$ do:
4. Calculate $\mathbf{Z}_1, \mathbf{B}, \mathbf{z}_2, \hat{\mathbf{y}}_{nT}$, and the $L(\hat{\mathbf{y}}_{nT}, \mathbf{y}_{nT})$, $J(\hat{\mathbf{y}}_{nT}, \mathbf{y}_{nT})$ of the loss function and cost function according to equation (7);
5. According to equation (8), the gradient of the output layer weight matrix $d\mathbf{v}$ and the gradient of the threshold $d\theta$ are calculated respectively;
6. The weight matrix of the hidden layer $d\mathbf{W}$ and the gradient of the threshold vector $d\boldsymbol{\xi}$ are calculated are calculated according to (9);
7. Update the weight matrix of the hidden matrix and the output layer $\mathbf{W}, \mathbf{V}$ and the threshold vectors $\boldsymbol{\xi}, \theta$;
8. End while
9. Input the channel information test set $(\mathbf{X}_{nR}, \mathbf{y}_{nR})$ and calculate the NMSE according to (16).

Algorithm 1

1. The weight matrices $\mathbf{W}$, $\mathbf{V}$ are initialized randomly from 0 to 1. The threshold vectors $\boldsymbol{\xi}, \boldsymbol{\theta}$ are initialized to 0. Set the training goal ε_{goal} and the learning rate η to a reasonable value, respectively. The intermediate variable ω is initialized to 1;
2. Input the channel information training set $(\mathbf{X}_{nT}, \mathbf{Y}_{nT})$ and verification set $(\mathbf{X}_{nV}, \mathbf{Y}_{nV})$ to train the neural network.
3. For $J(\hat{\mathbf{Y}}_{nT}, \mathbf{Y}_{nT}) > \varepsilon_{\text{goal}}$:
4. Calculate the hidden layer $\mathbf{Z}_1, \mathbf{B}$, output layer data $\mathbf{Z}_2, \hat{\mathbf{Y}}$ and cost function $J(\hat{\mathbf{Y}}, \mathbf{Y})$ of training set and verification set according to equation (7), respectively;
5. If $J(\hat{\mathbf{Y}}_{nV}, \mathbf{Y}_{nV}) > \omega$;
6. Quit
7. Else do:
8. $\omega = J(\hat{\mathbf{Y}}_{nV}, \mathbf{Y}_{nV})$;
9. According to (8), calculate the gradient of the output layer weight matrix $d\mathbf{V}$ and threshold vector $d\boldsymbol{\theta}$, respectively;
10. According to (9), calculate the gradient of hidden layer weight matrix $d\mathbf{W}$ and threshold vector $d\boldsymbol{\xi}$, respectively;
11. Update the weight matrix of hidden layer and output layer $\mathbf{W}, \mathbf{V}$, and the threshold vector $\boldsymbol{\xi}, \boldsymbol{\theta}$;
12. End for
13. Input the channel information test set $(\mathbf{X}_{nR}, \mathbf{Y}_{nR})$, and calculate the NMSE according to (16)

Algorithm 2

The simulation and analysis of the two schemes will be carried out, respectively. The prediction performance metric is expressed by the normalized mean squared error (NMSE). The NMSE at the q^{th} time slot is given by the following:

$$NMSE_q = \frac{\sum_{k \in n_R} \left| \hat{y}_q^{(k)} - y_q^{(k)} \right|^2}{\sum_{k \in n_R} \left| y^{(k)} \right|^2} \tag{16}$$

In addition, we adopt the early stopping strategy for the multi-time channel prediction system to avoid overfitting. More specifically, the early stopping divides the sample set into training set $(\mathbf{X}_{nT}, \mathbf{y}_{nT})$, validation set $(\mathbf{X}_{nV}, \mathbf{y}_{nV})$, and test set $(\mathbf{X}_{nR}, \mathbf{y}_{nR})$. The training set $(\mathbf{X}_{nT}, \mathbf{y}_{nT})$ is used to calculate the gradient and update the link weight matrixes $\mathbf{W}$, $\mathbf{V}$ and threshold $\boldsymbol{\xi}$, $\boldsymbol{\theta}$. The verification set is only used to estimate the cost $J(\hat{\mathbf{Y}}_{nV}, \mathbf{Y}_{nV})$. If the training set error decreases and the validation set error increases, the training will stop and the connection weights $\mathbf{W}_{epoch}$, $\mathbf{V}_{epoch}$ and thresholds $\boldsymbol{\xi}_{epoch}$, $\boldsymbol{\theta}_{epoch}$ will return. The specific algorithm is shown in Algorithm 1 and Algorithm 2.

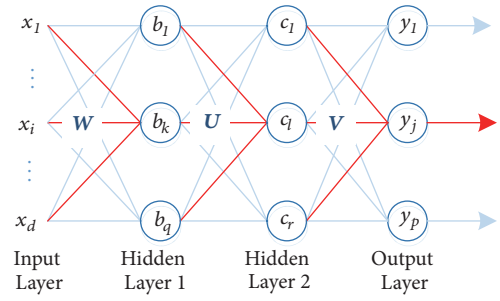


Figure 3: Three-layer multi-time prediction system.

3.1.3. Multi-Input and Multi-Output Multi-Layer Neural Network Channel Prediction System. Deep neural network can effectively predict the channel information when channel environment is complicated. More importantly, the deep neural network has better performance than other processing means. Meanwhile, deep neural networks achieve the same performance with fewer neurons. As shown in Figure 3, there

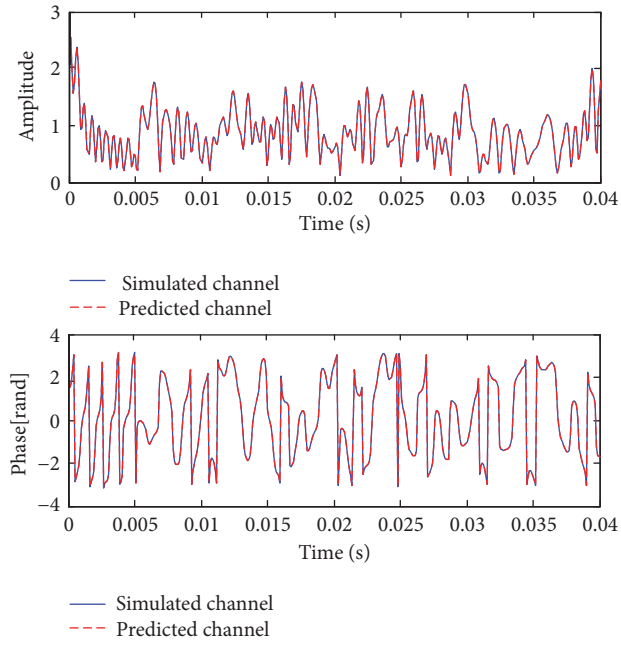


FIGURE 4: Simulated channel and the predicted channel amplitude, phase.

is a three-layer neural network structure with double hidden layer, which is used to predict a more complicated channel.

The parameter update of the three-layer neural network is almost the same as that of the two-layer neural network. Both adopt the backpropagation algorithm, except for the former needing to update three weight matrixes and three threshold vectors in one epoch.

3.2. Complexity Analysis. In this section, we compare the computational complexity of existing algorithms with deep learning methods. The existing methods, such as AR methods [6], DWT-AR-LR methods [9], the ESN method [10], and SVM prediction methods [10], will be mentioned below. The computing complexity of AR method is $O(N_{AR})$, where N_{AR} denotes the order of AR. The complexity of ESN prediction method is $O(\max(M, N_{nz}, N))$, where M is the number of variables, N_{nz} is the number of nonzero elements of middle layer weight matrix, and N is the number of variables in the middle layer. DWT-AR-LR method's complexity is $O(\max(N_{DWT}, N_{AR}, N_{LR}))$ where N_{DWT}, N_{AR}, and N_{LR} represent the number of samples and the order of AR and LR, respectively.

Note that the propagation weight matrix $\mathbf{W}$, $\mathbf{V}$ and threshold vector $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ of the neural network prediction algorithm proposed in this paper are calculated offline. In addition, its overhead is very small. The computational complexity of the mathematical operations of $\mathbf{W}^T\mathbf{x}^{(k)}$, $\mathbf{V}^T\mathbf{b}$ and $g_2(\mathbf{V}^T\mathbf{b} + \boldsymbol{\theta})$ in the neural network are $O(d \times q)$, $O(q \times p)$ and $O(p)$, respectively. Accordingly, the computational complexity of the neural network channel prediction system is $O(\max(d \times q, q \times p))$, where d, q, p are the number of neurons in input layer, hidden layer, and output layer, respectively. In this paper, the number of neurons is very small (e.g., d=10, q=10, p=10), especially in the multi-layer neural network. And there comes the low complexity.

4. Simulations

4.1. Single-Time Channel Prediction. Firstly, we use the Jakes model to simulate three channel predicted systems [20]. The channel power is fixed to $E_0^2 = 1$. In simulation, we set 34 ($N = 34$) scattering components, 500 ($Ns = 500$) channel information samples, and the sampling interval to be $Ts = 1 \times 10^{-4}s$. The maximum Doppler frequency shift is $f_d = 926Hz$. Phase φ observes uniform distribution, i.e., $\varphi \sim U(-\pi, \pi)$. We obtain 400 neural network samples through 500 channel information samples. Set the training samples $n_T = 200$, test samples $n_R = 200$, learning rate $\eta = 0.001$, and the target error $\varepsilon_{goal} = 1 \times 10^{-4}$. Figure 4 depicts the amplitude and phase of the simulated and predicted channels under the Jakes model. We can see that the channel predicted by the BP neural network is almost identical to the simulated channel of Jakes model.

NMSE is the performance measure; Figure 5 is the comparison of different prediction methods. The x-axis is the signal-to-noise ratios (SNR) of the channel information $h(t)$ and noise $n(t)$, while the y-axis is the NMSE. The red line with triangle is the performance of single-time predict system employing two-layer BP neural network. Rich neuron information gives the BPNN long-term channel memory capability. Thus, it can effectively perform channel prediction. As shown in Figure 5, the NMSE of BP neural network prediction algorithm gradually decreases with the increase of SNR and eventually reaches zero. With a low computational complexity, the accuracy of BPNN method is better than other methods (i.e., SVM, ESN, and DWT-AR-LR).

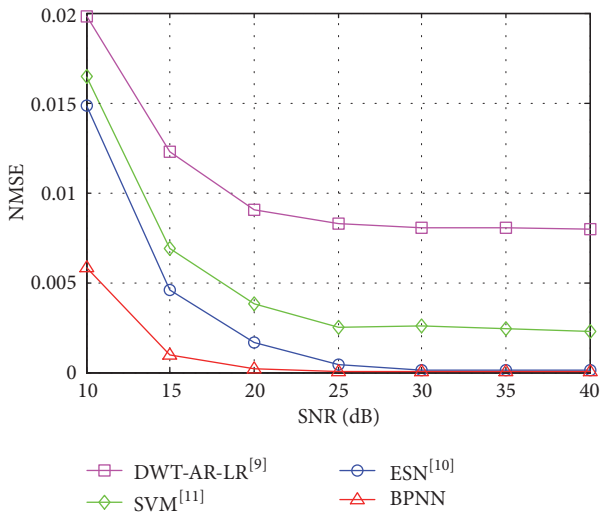


Figure 5: The prediction accuracy, BPNN versus existing methods.

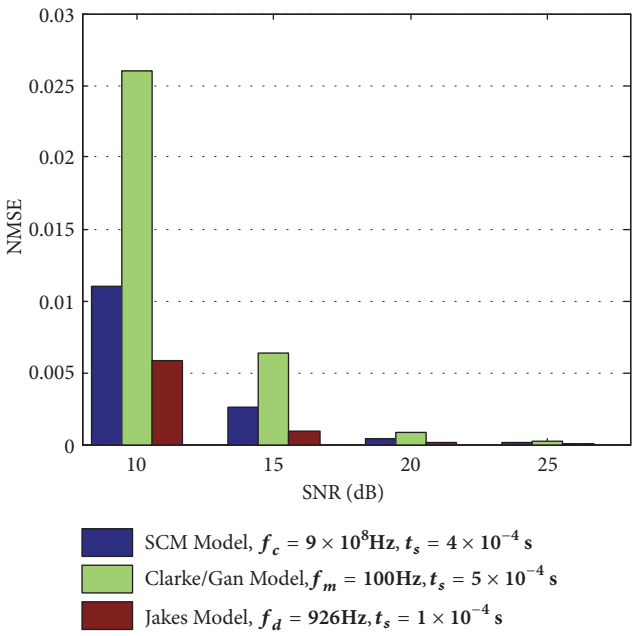


Figure 6: The prediction accuracy under different channel models.

In order to verify the robustness of the algorithm, we present various simulations under different fading Rayleigh channels, for example, the fast fading channel Clarke/Gan's model [23] and the well-known 3GPP Spatial Channel Model (SCM) [24] for MIMO systems.

Figure 6 demonstrates the predicted normalized mean square error (NMSE) under different Rayleigh fading channels. As we know, the autocorrelation of CSI satisfies the zero-order Bessel function over time. Thus, a bad time domain correlation leads to a more difficult channel prediction. For the poor time domain correlation, the CSI of Clarke/Gan's model is sampled in the frequency domain and transformed to the time domain by IFFT, leading to a poor channel prediction performance. Owing to the strong time domain correlation, Jakes model has the best prediction performance under different values of SNR.

Undeniably, the BP neural network also faces the problem, as other algorithms, which is that the poorer time domain correlation of CSI, the more difficulty to predict future channel. In short, the BPNN algorithm performs better than the other two algorithms, and the prediction performance of the Jakes channel model is the best.

4.2. Multi-Time Channel Prediction System. Similarly, the CSI is generated by the Jakes model for multi-time channel prediction system. The major difference is that we use 5000 channel information samples, i.e., $Ns = 5000$. Other parameters are the same as the single-time channel prediction system.

We research two-sample construction methods (NSCS and SSCS) of multi-time channel prediction system. Under the two strategies, we select 4000 samples and 400 samples, respectively, and 75% of which are training samples, 15% of which are verification samples, and the remaining samples are the test samples. The dimension of input layer sample is = 10. The number of hidden layer neurons is $q = 10$. Output layer neurons are $p = 10$ and = 20, respectively.

4.2.1. Comparison of Two-Sample Construction Schemes with 10-Input and 10-Output (d=10, p=10). For multi-time channel prediction system, as the prediction time increases, the corresponding error increases exponentially. We compare the prediction of two sampling methods, i.e., NSCS and SSCS.

Figure 7 shows that prediction accuracy generally improves as the number of epochs increases. Owing to the early stopping strategy, it is shown that the normal sample construction scheme stops after 224 epochs and the sparse sample prediction scheme stops after 42 epochs. The SSCS scheme has fewer iterations than the NSCS scheme, which can increase the speed of operation and save system resources.

Figure 8 is the NMSE performance of two-sample construction schemes. We can see that the performance of NSCS is better than that of SSCS at the cost of computation complexity. On the other hand, the performance difference between NSCS and SSCS is less than 10^{-4} which can be ignored. Moreover, in order to achieve the same target of NMSE, the latter has less epochs than the former, which is more practical. Notice that an epoch SSCS takes less time than NSCS. To summarize, the NSCS and SSCS we proposed both meet the requirement in [25] of a low estimated error. The SSCS effectively reduces the resource consumption without degrading system performance.

4.2.2. Comparison of Two-Sample Construction Schemes with 10-Input and 20-Output (d = 10, p = 20). Figure 9 shows that the normal samples construction scheme stops after 86 epochs. The sparse sample prediction scheme stops after 61 epochs. Figure 10 is the NMSE of two-sample construction schemes. The multi-time prediction, just like the 10-input and 10-output prediction system, error increases exponentially

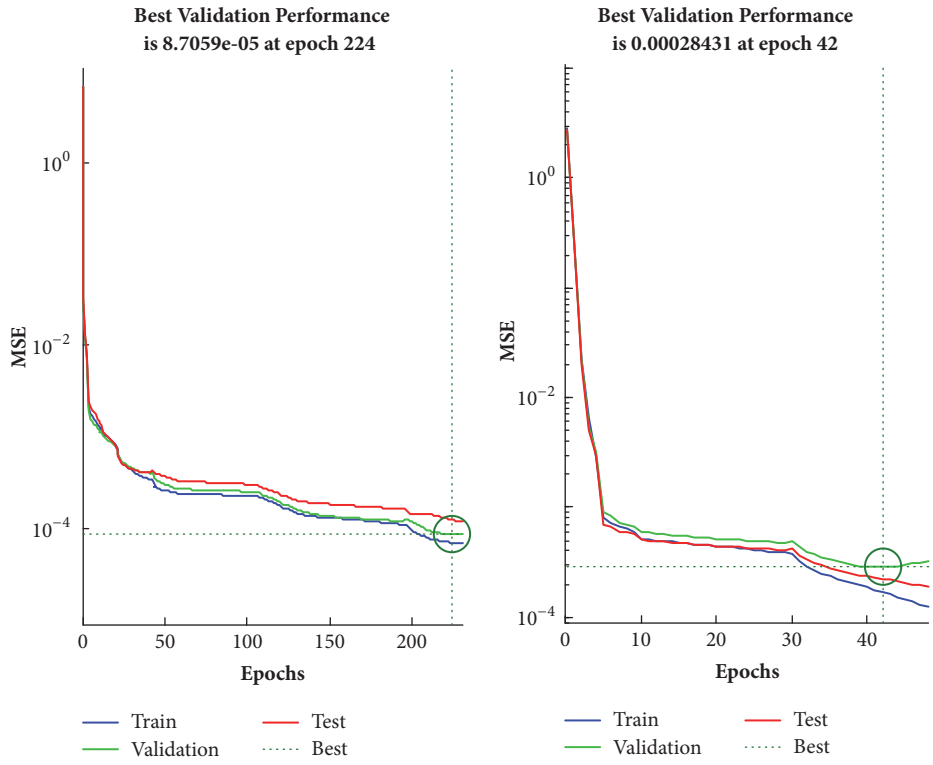


FIGURE 7: The epochs of NSCS and SSCS (d=10, p=10).

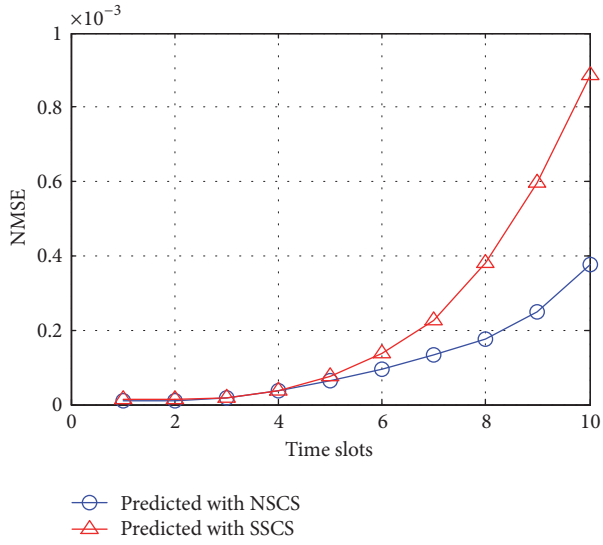


FIGURE 8: The NMSE of NSCS and SSCS (d=10, p=10).

with the different timeslot. The performance of SSCS is slightly worse than that of NSCS.

Nevertheless, the calculation and time costs of SSCS are much smaller than NSCS's.

4.2.3. The Performance of SSCS with Different Power of Noise. Figure 11 demonstrates the predicted NMSE at different SNR in the multi-time prediction system. With the weakening of the time domain correlations, the prediction errors of different time slot increase exponentially.

4.3. Multi-Hidden Multi-Moment Prediction System. Figure 12 compares the prediction performance of three-layer neural networks and two-layer neural networks. It reveals that the three-layer neural network outperforms the two-layer neural

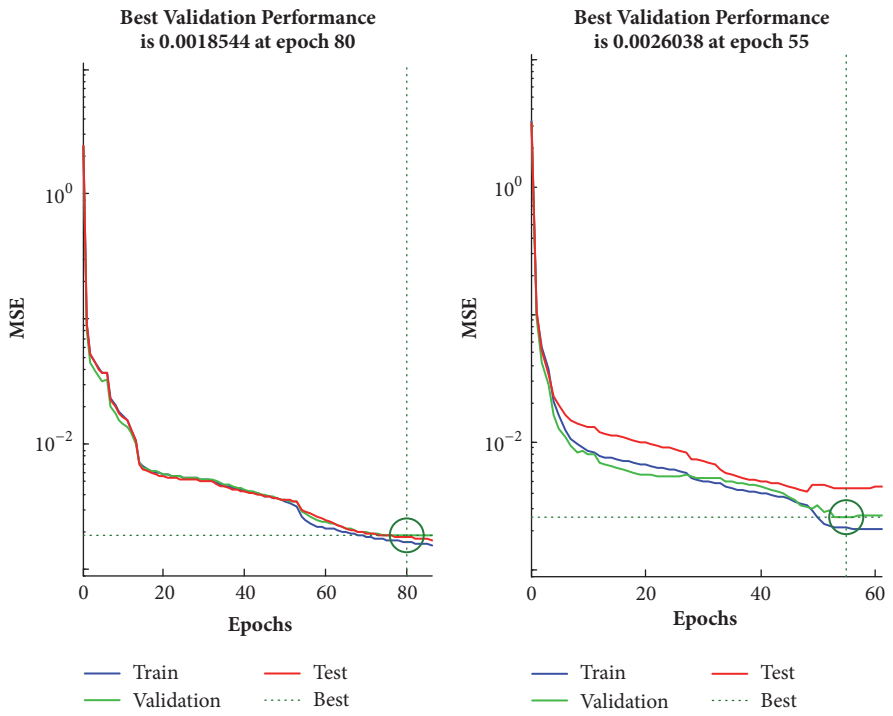


FIGURE 9: The epochs of NSCS and SSCS (d=10, p=20).

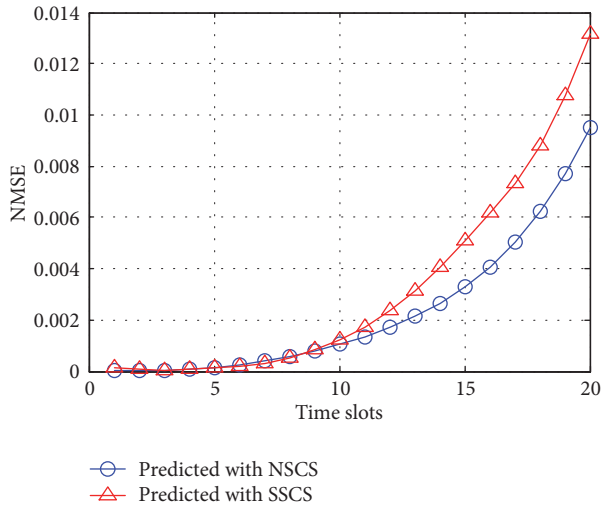


FIGURE 10: The NMSE of NSCS and SSCS (d=10, p=20).

network. However, its effectiveness is not obvious since the channel information is not very complicated.

5. Conclusions

The BP neural network with multi-hidden layer is introduced into the channel prediction application. A novel multiple moment CSI prediction scheme is proposed for improving the performance of the massive MIMO, NOMA, CoMP, and physical layer security schemes. The proposed prediction scheme can perform effectively with a short pilot overhead, which is suitable for resource-constrained communication scenes. Meanwhile, we proposed two significant sample construction methods, which extremely improves the prediction performance and reduces the computing complexity. Wide experiences verified the effectiveness of our proposed scheme

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

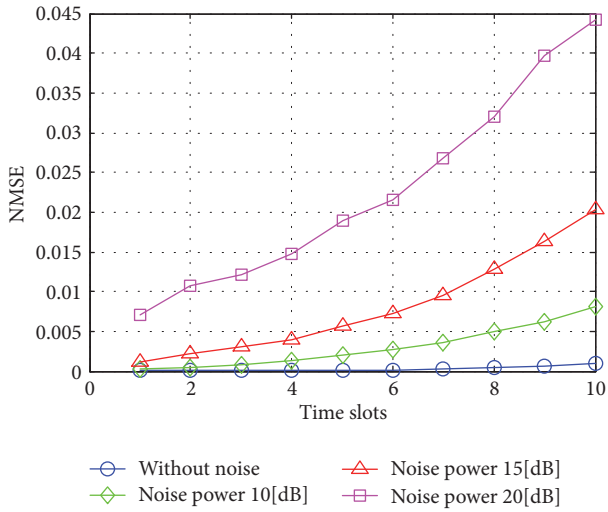


Figure 11: The multi-time prediction with noise.

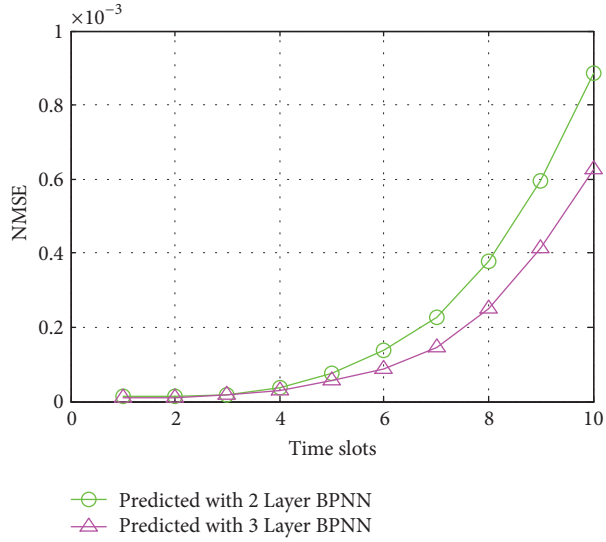


Figure 12: Two-layer neural network and three-layer neural network performance comparison.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by NSFC (no. 61572114), National Major R&D Program (no. 2018YFB0904905), and Chile Conicyt Fondecyt Project no. 1181809.

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[2] W. Cheng, X. Zhang, and H. Zhang, "Statistical-QoS Driven Energy-Efficiency Optimization over Green 5G Mobile Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3092–3107, 2016.

[3] L. Hu, H. Wen, B. Wu, J. Tang, and F. Pan, "Adaptive Secure Transmission for Physical Layer Security in Cooperative Wireless Networks," *IEEE Communications Letters*, vol. 21, no. 3, pp. 524–527, 2017.

[4] L. Hu, H. Wen, B. Wu et al., "Cooperative Jamming for Physical Layer Security Enhancement in Internet of Things," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 219–228, 2018.

[5] A. Duel-Hallen, H. Hallen, and T.-S. Yang, "Long range prediction and reduced feedback for mobile radio adaptive OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 10, pp. 2723–2732, 2006.

[6] P. Sharma and K. Chandra, "Prediction of state transitions in Rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 416–425, 2007.

[7] A. Arredondo, K. R. Dandekar, and G. Xu, "Vector channel modeling and prediction for the improvement of downlink received power," *IEEE Transactions on Communications*, vol. 50, no. 7, pp. 1121–1129, 2002.

[8] M. Sternad and D. Aronsson, "Channel estimation and prediction for adaptive OFDM downlinks," in *Proceedings of the Proc. IEEE Veh. Technol*, vol. 2, pp. 1283–1287, 2003.

[9] T. Ding and A. Hirose, "Fading channel prediction based on combination of complex-valued neural networks and chirp Z-transform," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1686–1695, 2014.

[10] Y. Zhao, H. Gao, NC. Beaulieu, Z. Chen, and H. Ji, "Echo State Network for Fast Channel Prediction in Rice of Fading Scenarios," *IEEE Communications Letters*, vol. 21, no. 3, pp. 672–675, 2017.

[11] X. Zhao, C. Hou, and Q. Wang, "A new SVM-based modeling method of cabin path loss prediction," *International Journal of Antennas and Propagation*, vol. 2013, Article ID 279070, 7 pages, 2013.

[12] K.-T. Kim, D.-K. Seo, and H.-T. Kim, "Efficient radar target recognition using the MUSIC algorithm and invariant features," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 3, pp. 325–337, 2002.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] G. Arulampalam and A. Bouzerdoum, "A generalized feedforward neural network architecture for classification and regression," *Neural Networks*, vol. 16, no. 5-6, pp. 561–568, 2003.

[15] G. A. Carpenter, *Neural Network Models for Pattern Recognition and Associative Memory*, vol. 2, Elsevier Science Ltd., 1989.

[16] L. Prechelt, "Automatic early stopping using cross validation: Quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[17] N. C. Beaulieu and Y. Chen, "Maximum likelihood estimation of local average SNR in ricean fading channels," *IEEE Communications Letters*, vol. 9, no. 3, pp. 219–221, 2005.

[18] Y. Chen and N. C. Beaulieu, "Estimation of ricean K parameter and local average SNR from noisy correlated channel samples," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 640–648, 2007.

[19] G. L. Stuber, *Principles Mobile Communication*, Kluwer, Boston, Massachusetts, Mass, USA, 2nd edition, 2001.

[20] W. C. Jakes, *Microwave Mobile Communications*, Wiley, New York, NY, USA, 1974.

[21] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[22] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computation*, vol. 7, no. 2, pp. 219–269, 1995.

[23] R. H. Clarke, "A statistical theory of mobile-radio reception," *Bell Labs Technical Journal*, vol. 47, no. 6, pp. 957–1000, 1968.

[24] 3GPP, Spatial channel model for MIMO simulations 25.996 V6.1.0, 2003, http://www.3gpp.org/.

[25] J. Tang, H. Wen, L. Hu et al., "Associating MIMO beamforming with security codes to achieve unconditional communication security," *IET Communications*, vol. 10, no. 12, pp. 1522–1531, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 8974802, 9 pages
https://doi.org/10.1155/2018/8974802

Research Article

General Multimedia Trust Authentication Framework for 5G Networks

Ling Xing,[1] Qiang Ma,[2] Honghai Wu,[1] and Ping Xie[1]

[1]*School of Information Engineering, Henan University of Science and Technology, Luoyang, China*
[2]*School of Information Engineering, Southwest University of Science and Technology, Mianyang, China*

Correspondence should be addressed to Qiang Ma; maqiang_my@163.com

Received 12 January 2018; Revised 24 April 2018; Accepted 16 May 2018; Published 28 June 2018

Academic Editor: Jinsong Wu

Due to the varieties of services and the openness of network architectures, great challenges for information security of the 5G systems are posed. Although there exist various and heterogeneous security communication mechanisms, it is imperative to develop a more general and more ubiquitous authentication method for data security. In this paper, we propose for the 5G networks a novel multimedia authentication framework, which is based upon the trusted content representation (TCR). The framework is general and suitable for various multimedia contents, e.g., text, audio, and video. The generality of the framework is achieved by the TCR technique, which authenticates the contents' semantics in both high and low levels. Analysis shows that the authentication framework is able to authenticate multimedia contents effectively in terms of active and passive authenticating ways.

1. Introduction

The requirements for higher data transmission rate and better user services experiences promote the development of the 5G cellular networks. Compared with its precedent wireless communication standards (i.e., 1G, 2G, 3G, and 4G networks), the 5G system is proposed to support wider range of connected devices types and various service applications [1, 2]. It is estimated that the 5G would become the ubiquitous information infrastructure network in the near future [3]. Along with the growth of the users and services in 5G networks come the security problems, e.g., data privacy, data integrity, and data authentication [4–6]. Ensuring the secure wireless communication, together with the security of data transferred, is mandatory for the success of the 5G networks.

Various types of multimedia content proliferate over the wireless networks. Especially equipped with the social media techniques, people find it is quite convenient to communicate on the mobile devices. Although the multimedia contents bring much convenience for the information sharing over the 5G networks, security concerns of the contents cannot be overlooked [7–9]. Due to openness feature of the underlying network, the received multimedia contents should be carefully examined. For example, questions should be raised whether the obtained video or image has been attacked by malicious purposes, or whether the video has been deliberately altered to screw its original meaning. Thus the security of multimedia contents must be ensured to keep the integrity of the contents as intact as the original.

As far as the formats of the multimedia are concerned, methods to authenticate the contents are different in terms of purpose, efficiency, and validity. Those methods or algorithms for authenticating the multimedia contents can be roughly classified into four groups, i.e., watermarking-based, encryption-based, streaming-based, and robust-hashing-based methods.

The watermarking-based authentication methods for multimedia usually adopt the watermark embedding and extraction approaches. For example, Md. Asikuzzaman et al. [10] proposed an imperceptible and robust blind watermarking for video, in which the watermark was embedded into the levels of the dual-tree complex wavelet transform coefficients. There are other multimedia watermarking methods which emphasize the features of the watermark, e.g., the content dependent video watermark [11] and wavelet-based multiple watermarks [12]. Although the watermarking methods can

provide the authentication of contents in some degree, the embedding process itself harms the integrity of the content. In addition, one watermarking method can only deal with some specific attacks, which hinders these methods from being widely adopted to cope with the various attacks existing in the 5G networks.

The encryption-based multimedia authentication methods usually employ the encryption algorithms, e.g., RSA algorithm and DES algorithm, to encrypt the whole or parts of the contents. For example, Zafar Shahid et al. [13] presented a selective encryption (i.e., the truncated rice code and the Exp-Golomb code) approach for video of High Efficiency Video Coding (HEVC), which shows the characteristics of format compliant and real-time property. Similarly, Glenn Van Wallendael et al. [14] proposed the encryption for HEVC by utilizing the differences of intraprediction mode, the sign of motion vector difference, and the residual sign. Note that the primary goal of the encryption-based methods is to ensure the confidentiality. The time overhead incurred by the encryption and decryption process is often a nonnegligible issue for the social media devices.

The aim of the streaming-based multimedia authentication methods is to ensure that the received contents have the integrity property, which is verified by the recipients. Currently many research works have been focused on this area; e.g., Kang et al. [15] studied the pollution attack detection and prevention method by trust for peer-to-peer streaming, Lu et al. [16] proposed a privacy protection method based on trust for peer-to-peer data sharing network, and Cheng et al. [17] authenticated the live streaming media by means of TESLA- (Timed Efficient Stream Loss-tolerant Authentication-) based protocol. One category of streaming-based authentication methods is the graph structure type, e.g., the chain line approach, the tree approach, and the butterfly approach, which treats the content packets as individual ones and exploits various packets' hash attaching means [18]. The other category is based on the multimedia coding structure. For example, Kianoosh Mokhtarian et al. [19] proposed the authentication for Scalable Video Coding (SVC) streams, whose hash appendence follows the coding structure of SVC.

The robust-hashing-based methods to authenticate the multimedia have the best performance in terms of robustness when the authentication is applied under the network circumstances. The robustness of the hashes means that hashes remain the same or change a little after the perceptual content preserving operations are being conducted. For example, Lokanadham Naidu Vadlamudi et al. [20] proposed a robust hash algorithm by using features of histogram for image authentication. Due to the robustness and sensitivity of this kind of method, it is much superior to the other three methods when it comes to the network packet loss phenomenon in 5G networks. Since certain packets of multimedia contents may be lost because of traffic jams or network failure, contents integrity verifying process for the recipients should take into consideration the robustness of multimedia's representation.

Another aspect we should bear in mind is that all the four methods are for the low level semantics authentication, which treats only the integrity either in pixels level or in frames level. This absolutely lacks completeness for the authentication of multimedia contents of the 5G networks, since high level semantics of the contents are overlooked by the current methods. For instance, the contents' high level semantics including but not limited to title, name, author, and format should also be authenticated. Therefore, in this paper we endeavor to tackle the multimedia authentication by proposing combined authentication method for multimedia contents for 5G network. We propose a generalized multimedia trust authentication model based on the trusted content representation (TCR) method, where the generalized term means that it can be applied to various multimedia content formats. Also we provide the features needed to authenticate the integrity of the contents.

The rest of the paper is organized as follows. We describe the architecture of asymmetric wireless communication channel in Section 2, where the asymmetric term means that one channel is safe and is used for authentication information transfer while the other one is open and used for multimedia contents transfer. In Section 3 we illustrate the TCR indexing technique and show the features of TCR. In Section 4 we explain the framework of the general multimedia trust authentication for 5G networks, in which the philosophy of asymmetric wireless channel is adopted. Then we conclude our work in Section 5.

2. Architecture of Asymmetric Wireless Communication Channel

2.1. Security Threats to 5G Networks. Since the 5G networks are open and insecure, the multimedia contents, which are transmitted over them, are prone to various attacks by malicious purposes. From the perspectives of network communication system, we summarize security threats to multimedia into three groups.

(1) The Originality of the Multimedia May Lack Trust. Currently users can upload from the mobile devices the various kinds of multimedia contents up to the video or audio sharing websites by various means. However, the effective and efficient mechanisms to supervise the legacy of users' behaviors are still unavailable. While people find it quite easy to collect, edit, and distribute the multimedia contents by modern multimedia processing tools (e.g., Photoshop, Illustrator), people should think twice about the originality of the received multimedia. In other words, the user, who publishes the multimedia and claims the author of the multimedia, may likely not be the "true" author. Originality of the contents should be checked in some manner. This incurs the problem of intellectual property infringement. Therefore, it is urgent to find a good way to verify the originality of the multimedia.

(2) The Channel to Convey the Multimedia May Be Attacked. The openness of the 5G network protocols poses serious security problems to the contents being transmitted over it. The notorious attack, "man-in-the-middle", is the most common threat to the contents. The attacker maliciously

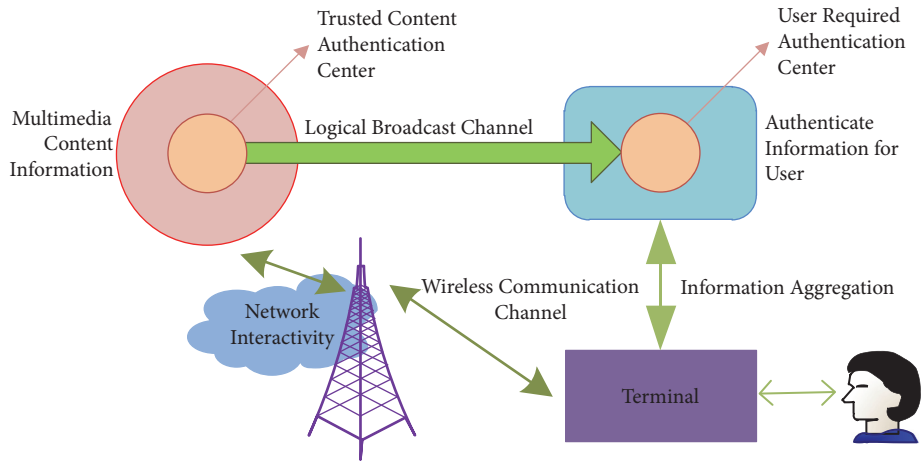


Figure 1: Asymmetric wireless communication channel architecture.

hurts the multimedia via many means, e.g., video frames' rearrangement, dropping, inserting, and altering. The purpose of the attack is to destroy the contents' integrity, which distorts the recipient's understanding of the contents. Note that there are many free packets sniffer tools on the Internet, which makes the attacks on the transmission process of multimedia much easier. We state that the threats to the multimedia on its journey from the sender to the receiver should be taken into consideration and the integrity of the contents needs to be ensured.

(3) The Identity of the Multimedia Recipients May Lack Trust. Recall that current authentication for the mobile recipient's identity is mainly based on the user-password method. This brings security problems once the legal recipient's platform is attacked. For instance, an attacker may break into and obtain the legal recipient's password to certain multimedia server, which results in the illegal copy or broadcasting of the contents. Another scenario is that the multimedia sender requires checking the identity of the receiver and the integrity of contents received by the receiver, e.g., video conferencing. Therefore, both the platform security and the received multimedia integrity are necessary to be authenticated.

2.2. The Asymmetric Wireless Communication Channel. The concept of asymmetric wireless communication channel for 5G networks is shown in Figure 1. We use the term asymmetric to describe the unbalanced properties of two channels existing in the architecture. One channel is the ordinary wireless communication channel, which is reserved for the multimedia contents sharing. The channel is dual and its capacity is large enough to support various kinds of services. It is not transparent to the users and not attacks-free due to its openness characteristics. The other channel, which is more akin to logical broadcasting communication channel, is introduced to transfer the authentication information used for multimedia contents. This channel is single-way and transparent. It does not exist in current networks and should be created by means of broadcasting techniques. Its capacity is much less compared to the former one. However it is closed and safe since the information transmitted over this channel is confidential.

The Trusted Content Authentication center (TCA center) in Figure 1 serves as a trusted third party. It collects the multimedia content information from the 5G network base stations. For instance, it analyzes the history records of users' behaviors and extracts multimedia content information. Then it generates the trusted content representation for each content resource. The representations of authentication are the safety benchmarks for the multimedia contents. We apply the logical broadcasting channel to safeguard these representations. The term logical means that it can be realized by virtual technologies, such as the network tunnels.

Suppose the representations are safely transferred to the User Required Authentication center (URA center), then the remaining task is to authentication the multimedia contents according to users' authentication requirements. The terminal is the wireless devices and is responsible for communicating with the center. It can have an application installed within its platform and initiate the authentication request to the URA center. The center returns the authentication information, together with the multimedia content profile, to the terminal, which are wrapped as information aggregation. If the center is not capable of providing relevant authentication information, it sends to the terminal a replay indicating that the terminal asks the TCA center through the wireless communication channel for the authentication information. The TCA center forwards the corresponding authentication information to the UCA center, which further sends it to the terminal. Then the terminal checks the multimedia contents' integrity by comparing the representation with the content itself.

According to the proposed asymmetric channels for the 5G networks, there are two ways for the users or terminals to obtain and authenticate the multimedia content information. The terminal connects to the wireless base station, which provides the access point to the Internet. The users upload or download contents through these points. In order to ensure the security of the terminal, the access points apply authenticity test to the terminals. The TCA center forms the TCR

index for the multimedia content. Considering the long-tail features of the popularity distribution of the content, we only generate the indexes for those "popular" contents, which can satisfy almost all users' authentication requirements [21, 22].

The other channel for the user to obtain the authentication information is the logical broadcasting band. The application on the terminals asks the URA center for contents authentication information by specifying the contents profile. Then the center searches its authentication information database for the specified terms. If it finds the required one, then it returns back to the terminal. Note that in order to protect the security of wireless channel between the terminal and the center, encryption methods are to be adopted. We have researched the asymmetric channel in our previous work and its detail is referred to [23, 24]. The database, which stores the authentication benchmarks, receives the authentication information from the TCA center periodically. The security of this logical channel must be ensured in order to provide the basis of trust for terminal authentication test.

3. Multimedia Trusted Content Representation

3.1. Trusted Content Representation Description. We first clarify what information the recipients can obtain from the multimedia contents for 5G networks. As for the contents recipients, they can get two kinds of information, i.e., the high level semantics and the low semantics. The former means that the descriptions about the contents, while the latter means the perceptual understanding of the contents. We define these two terms in detail as follows.

(1) *Multimedia High Level Semantics* (HLS). This kind of semantics is the description or explanation information about the multimedia, which is usually generated by the content author and is mainly used for the indexing and searching. The semantics are in the form of plaintext and can be added before the contents. Examples of HLS are title, author, keywords, and date. Note that HLS belongs to the conceptual level regarding recipients understanding of the contents.

(2) *Multimedia Low Level Semantics* (LLS). The LLS denotes the multimedia data which provides perceptual information for the recipients. They often exist in the form of binary data, e.g., the pixel matrix of an image and the frame series of a video. Note that information of LLS needs no summary of conceptual levels by contents authors or publishers. The perceptual understandings are formed by the contents recipients themselves.

We believe that the authentication of multimedia contents should cover both HLS and LLS information. Specifically, the HLS of contents should remain the same as the one of original contents and the LLS should also be as intact as the original one. The undergoing operations on the LLS, which are without malicious purposes, should preserve the perceptual understanding as the original one. Considering the 5G network packets loss to the data of LLS, we propose to authenticate the LLS robustly. We believe the robust-hashing is more appropriate for 5G networks and choose the robust-hashing methods to generate the LLS of contents.

Note that, as for the HLS, we suggest no robustness to its description. One reason is that the HLS information is quite sensitive and one bit change may completely destroy its meaning. The other reason is that we use the HLS to locate the multimedia contents and the HLS serves as the authentication starting points for contents verification. Therefore, we propose the TCR method, which represents the HLS and LLS robustly. The model is described as follows.

We denote the multimedia contents space by $\mathbf{M}$ and the multimedia content by $\mathbf{m}$. Let the HLS of $\mathbf{M}$ be denoted by $\mathbf{S}$ and we separate the space $\mathbf{S}$ into several subspaces, which are mutually independent of each other; i.e., the space $\mathbf{S}$ is the cross product of its subspaces. Each subspace represents certain conceptual description of $\mathbf{M}$. Then the HLS for $\mathbf{m}$ is depicted by $\mathbf{s}$, which is defined as

$$\mathbf{s} = (s_1, s_2, \ldots, s_n) \tag{1}$$

The terms of $\mathbf{s}$ are the instances of the subspaces of $\mathbf{S}$. Note that the number n is an integer which represents the granularity of the HLS spaces.

Regarding the robustness representation of the LLS for multimedia contents, we allow the perception preserving operations on the contents on condition that these operations do not change the representation too much. However, the degree of change is determined by the authentication requirements relating to the network circumstances. In order to clearly characterize the LLS feature, we here examine the categories of operational results of content $\mathbf{m}$.

We adopt the robust-hashing authentication method for the LLS representation. For the content $\mathbf{m}$, the results of operations being conducted on $\mathbf{m}$ can be thought of having three groups, i.e., $\mathbf{m}_s$, $\mathbf{m}_c$, $\mathbf{m}_d$, which are explained as follows.

(1) The set $\mathbf{m}_s$ denotes the multimedia contents, which are produced by the perception preserving operations on the content $\mathbf{m}$, e.g., the scaling of image and the brightness change on the video frame.

(2) The set $\mathbf{m}_c$ denotes the multimedia contents, which are produced by the content changing operations on the content $\mathbf{m}$, e.g., the frames reordering of video and the image blocks covering.

(3) The set $\mathbf{m}_d$ denotes the multimedia contents, which are different from and independent of content $\mathbf{m}$. The LLS of $\mathbf{m}_d$ is completely and fundamentally distinguished from $\mathbf{m}$.

We apply the robust-hashing methods to all those above sets. Let the hashing function be denoted by $\mathrm{H}(\cdot)$. Thus we obtain the corresponding hashes as follows.

$$\mathbf{H} = \mathrm{H}(\mathbf{m}) \tag{2}$$

$$\mathbf{H}_s = \mathrm{H}(\mathbf{m}_s) \tag{3}$$

$$\mathbf{H}_c = \mathrm{H}(\mathbf{m}_c) \tag{4}$$

$$\mathbf{H}_d = \mathrm{H}(\mathbf{m}_d) \tag{5}$$

Obviously we have the relation $\mathbf{H} = \mathbf{H} \cup \mathbf{H}_s \cup \mathbf{H}_c \cup \mathbf{H}_d$, where $\mathbf{H}$ denotes the LLS space for multimedia contents.

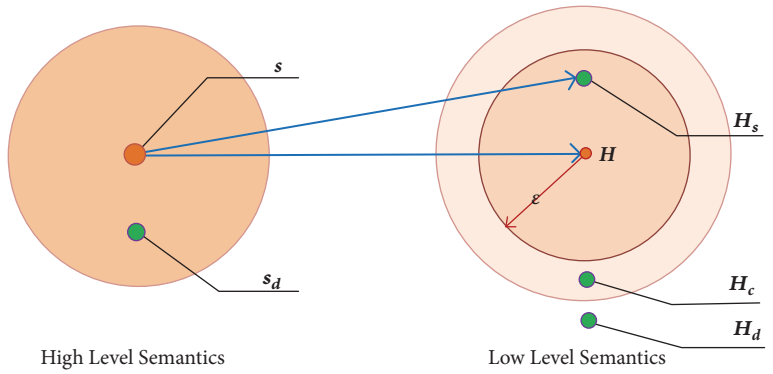


FIGURE 2: Model of trusted content representation.

We propose the trusted content representation method to describe the HLS and LLS for content **m**. Therefore the content **m** is represented by **T**, which is defined as

$$\mathbf{T} = (\mathbf{S}\ \mathbf{H}) \tag{6}$$

The TCR model is depicted in Figure 2, in which the HLS space is composed of **S** and $\mathbf{S}_d$. The $\mathbf{S}_d$ is the HLS for contents which is different from and independent of content **m**. Note that, as for the LLS, we allow the robustness, which means that the hash of the content being verified against **m** is secure if it falls into the set $\mathbf{H} \cup \mathbf{H}_s$. The parameter ε decides the distance between the original hash and the allowed maximum hash, which also draws the line for the robustness of the LLS for content **m**.

In regard to the authentication for multimedia content **m**, we apply the TCR technique to verify the received content. We first check the extracted HLS to see if it equals **S**. If no, we conclude that the received content is not safe since the high semantics are not secure; if yes, we further check the extracted LLS to see whether it is within the range of $\mathbf{H}_s$. If yes, we state that the content being verified is authenticated; if no, we assert that the low level semantics are being attacked and the content received is not secure.

3.2. Properties of Trusted Content Representation. We in this part summarize the properties of the model when it is employed in the 5G networks to authenticate multimedia contents. There are five properties, i.e., collision-free, security, robustness, sensitivity, and compactness.

(1) Collision-Free Property. The collision-free property means that, for every two dissimilar multimedia contents, they have different trusted content representations. Suppose that $\mathbf{m} \in \mathbf{M}$ and $\mathbf{m}_d \in \mathbf{M}$, and contents **m** and $\mathbf{m}_d$ are different, then we have

$$\mathbf{S} \neq \mathbf{S}_d \wedge pr\left(\left\|\mathbf{H} - \mathbf{H}_d\right\| > \tau\right) \approx 1 \tag{7}$$

where symbol $\wedge$ means logic operation AND and *pr* means the probability. The parameter τ is determined theoretically or empirically and it defines the robustness of the LLS for the contents. In other words, the distance for every two different multimedia contents is greater than the threshold τ with probability one.

(2) Security Property. This property ensures that the trusted content representation for multimedia contents should be resilient against various malicious attacks and be detectable after being attacked. It also states that the representation **T** for certain content cannot be regenerated by attackers. Usually this is achieved through key controlled representation generation. Specifically, the LLS should be hashed by keys. The security property has two aspects; i.e., one is the one-way hash, and the other is the unpredictability.

The one-way hash means that it is easy to generate the hash for LLS from the contents, while it is extremely difficult and impossible to recover the LLS of contents from the hash value. Suppose the contents **m** has the representation (**S H**), and from **T** can be inferred the content **m**', then we have

$$pr\left(\mathrm{H}\left(\mathbf{m}'\right) = \mathbf{H}\right) \approx 2^{-L} \approx 0 \tag{8}$$

where the integer L is the length of the hash and we assume the one and zero in the hash follow uniform distribution.

The unpredictability means that given a multimedia content **m**, it is empirically impossible to infer its hash for LLS. Since the hash for LLS generation is controlled by key, suppose key K_1 is used for the legal generating LLS hash for **m** and key K_2 is another key which is illegal, then we have

$$pr\left(\mathrm{H}\left(\mathbf{m} \mid K_1\right) == \mathrm{H}\left(\mathbf{m} \mid K_2\right)\right) \approx 2^{-L} \approx 0 \tag{9}$$

where $\mathrm{H}(\mathbf{m} \mid K_1)$ means that the hash generation is controlled by K_1. Note that K_1 does not equal K_2.

(3) Robustness Property. The robustness property of this model is embodied in the LLS for the contents; i.e., some conception preserving operations on the LLS are permitted. Suppose we have a content **m** and its similar version $\mathbf{m}_s$, then we have the following relation

$$\mathbf{S} == \mathbf{S}_s \wedge \left\|\mathbf{H} - \mathbf{H}_s\right\| \leq \varepsilon \tag{10}$$

where threshold ε draws the boundary between similarity and difference for contents comparison.

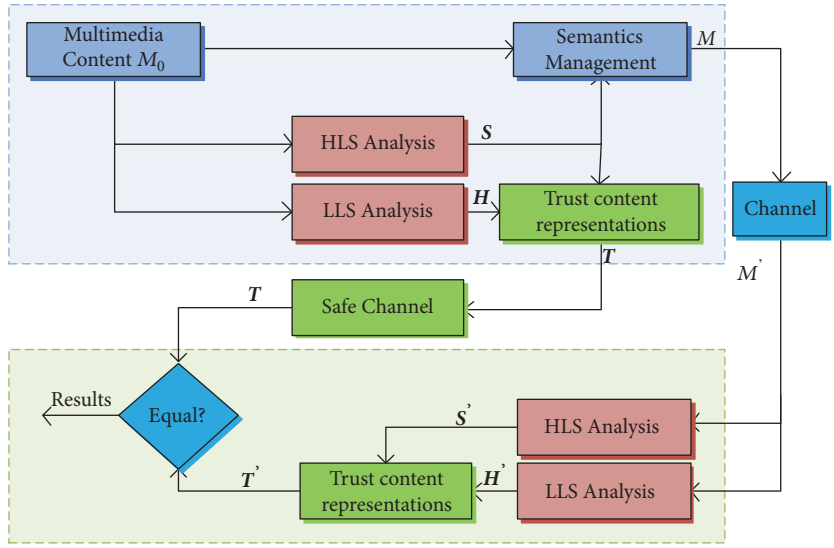


Figure 3: Framework of multimedia trust authentication based on trusted content representation.

(4) Sensitivity Property. This one is quite important for the contents security verification. It means that the model can detect those attacks which result in the changes of trusted content representation. There are two kinds of attacks regarding the semantics of multimedia. One is attacks on the HLS and the other is on the LLS. The latter attacks harm recipients' perceptual understanding on the LLS. We denote the altered contents of $\mathbf{m}$ by $\mathbf{m}_{diff}$, which belongs to the set $\mathbf{m}_c \cap \mathbf{m}_d$. Then the following relation holds.

$$\mathbf{S} \neq \mathbf{S}_{diff} \vee \left\| \mathbf{H} - \mathbf{H}_{diff} \right\| > \varepsilon \tag{11}$$

where the symbol $\vee$ denotes OR logic operation.

(5) Compactness Property. Compactness property means that the shorter the trusted content representation is, the better it is. If the length of the representation is shorter, it requires less storage and less transmission overhead for 5G networks. However, we believe that there is a balance between the compactness and correctness of the representation. For instance, if the contents' HLS representation $\mathbf{S}$ has much larger length, then $\mathbf{S}$ can represent more detailed information. It means the HLS space is divided into subspaces of more granularities. On the contrary, if the length is smaller, the representation $\mathbf{S}$ is much more compact. But more HLS information cannot be included. Therefore, it needs tactics and empirics to decide the length of the trusted content representation. Note that this balance also applies to the LLS representation.

Although these five properties exist in the TCR model, we claim that not all these five properties can be met at one time. For example, if the model is more emphasized on the security of the multimedia contents, then the security and sensitivity properties should come first. However, if the model is more concerned from the perspective of the transmission efficiency, then the compactness property should be weighed more. Thus in real-life applications of 5G system, the properties of the model should be determined accordingly.

4. General Multimedia Trust Authentication Framework

When we apply the TCR model in the 5G networks, we break this authentication into two processes, i.e., the semantics analysis and adding by the sender, the semantics extraction and comparison by the recipient, which is shown in Figure 3.

We denote the raw multimedia content generated by sender by M_0. First we analyze the content to obtain its HLS and LLS semantics, which are represented by symbols $\mathbf{S}$ and $\mathbf{H}$, respectively. The HLS is added to M_0 through semantics management. The adding method can be realized by the appendence to the content's head or tail. The multimedia M_0, together with $\mathbf{S}$, is conveyed to the recipient across the channel. The channel is an open network, i.e., the 5G network, whose security is not ensured. On the other hand, the HLS and LLS are synthesized into trusted content representation $\mathbf{T}$, which is securely transmitted by safe channel. The safe channel can be achieved by PKI (Public Key Infrastructure) technique, which encrypts the representation $\mathbf{T}$ through the certificates provided by an authentication center.

The safe channel is also used to test the recipient's identity. Note that the safe channel serves as the logical broadcasting channel depicted in Figure 1. Before the recipient can access any contents from the 5G networks, its authenticity must be ensured by the base stations. Some modern identity method, e.g., trust computing, can be applied to check the security profile of the terminals' platform. Thus the base station first initiates the identity authentication requirement when the recipient asks for connecting to the wireless networks. It starts the challenges and responses mechanism to perform the credit checking. For instance, the terminal hardware stores certain credit issued by the base station parties. The base station sends the nonce to the terminal. Then the terminal returns the cipher message of credit and nonce, digitally encrypted by the station's public key, to the station. The base station decrypts the cipher to compare the recipient's

```
<?xml version="1.0"encoding="utf-8"?>
<HighLevelSemantics>
   <title>"The Mystery of Van Gogh's Ear"</title>      <!--title-->
   <keywords>"Van Gogh,Ear"</keywords">           <!--keyword-->
   <classification>"Documentary"</classification>   <!--classification-->
   <description>
      "What happened on the December night in 1888 when Vincent van
Gogh took a blade to his own ear"
   </description>                                  <!--description-->
   <source>"BBC"</source>                           <!--source-->
   <language>"English"</language>                   <!--language-->
    <coverage>"19th century"</coverage>           <!--coverage-->
   <identifier>
     "http://www.bbc.co.uk/programmes/b07nswft"
  </identifier>                                      <!--identifier-->
   <creator>"Bill Locke"</creator>                   <!--creator-->
   <publisher>"Lion Television"</publisher>       <!--publisher-->
   <contributor>"Jack MacInnes"</contributor>   <!--contributor-->
   <rights>"copyright reserved"</rights>              <!--rights-->
   <date>"2016"</date>             <!--date-->
   <type>"VidVideo"</type>     <!--type-->
</HighLevelSemantics>
```

Figure 4: High level semantics instance based on XML.

credit with its stored one and decides whether the recipient is legal or not. It also compares the decrypted nonce to its previously sent one. If they match, then it concludes that the message is new and free of replay attack. Otherwise, it believes the message is generated by attackers and rejects the access requirement. We use the base station to actively verify the authenticity of the recipients in order to prevent some malicious users from deliberately sabotaging the security of the 5G networks.

The recipient obtains from the channel the content denoted by M', whose integrity is to be verified by the recipient. The recipient extracts from M' the HLS and LLS semantics and form the trusted content representation $\mathbf{T}$'. Then the recipient compares the representations $\mathbf{T}$' and $\mathbf{T}$ and judges whether the content M' is secure or not. If those two terms are the same, the recipient can be ensured that the content M' is of integrity. Otherwise, it is not secure and it may have been attacked. Note that when the recipient compares the LLS, the robustness should be taken into consideration.

Regarding the semantics management, we propose a metadata based method to append the HLS to the multimedia content. The metadata is organized in XML format. It presents its semantics to the recipient before the content is presented to the recipient. For example, the packets of supplemental enhancement information within the H.264 stream can be used to store the HLS; the label unit of "APPn" within the JPEG file can be used for the HLS. We provided a method to divide the HLS space into 14 subspaces based on our previous work [25]. The instance of XML based HLS representation is shown in Figure 4, in which fourteen concepts are used for characterizing the high semantics.

We observe from the framework of Figure 3 that there are both the active and passive authentication modes in our model. The HLS information authentication is active, since it adds or appends additional semantics information into the multimedia contents. However, the LLS authentication is passive, because we extract the content's low level conceptual description akin to the digest. The superiority of our framework is that it does not harm the content itself and the trusted content representation can be transferred separately from the content, which can be protected by advanced encryption algorithms.

5. Security Analysis of the Framework

Suppose that recipient obtains from the 5G networks the multimedia content lacking trust. This could be caused by attackers modifying the semantics of the content. We denote the "original" content uploaded by an attacker to the networks by TCR $\mathbf{T}$', which generates the HLS and LLS as $\mathbf{S}'$ and $\mathbf{H}'$, respectively. The real and secure content relating to the $\mathbf{T}$' is denoted by $\mathbf{T}$, which consists of HLS and LLS, respectively. The purpose of the attacking is to distort the semantics of the content. It can be realized from three perspectives. The first one is that the attacker modifies the high semantics of the content while keeping the low semantics intact. The second case is opposite compared to the first, attacking the low semantics of the content while keeping the high semantics intact. The third is that attacker modifies both the high and low semantics.

Note that in our proposed trust authentication framework the safe channel is to transfer the authentication information. Thus we assume this channel is free of attacks and the destination of the channel can obtain the secure and intact information. We analyze the security of the framework from three aspects according to the types of attacks to multimedia of the 5G networks.

For the first scenario, the high semantics is attacked only. Thus we have $\mathbf{S}' \neq \mathbf{S}$ and $\mathbf{H}' \neq \mathbf{H}$. The terminal obtains

the TCR from the URA center in Figure 1 as ($\mathbf{S}'$ $\mathbf{H}$). The $\mathbf{H}$ could be valid or void, because there is a chance that no content exists under the high semantics $\mathbf{S}'$. If $\mathbf{H}$ is void, the terminal can claim that the content it received is not trusted. If $\mathbf{H}$ is valid, then it should have $\mathbf{H}' \neq \mathbf{H}$ according to the collision-free property. Thus terminal also can find the attacks. Therefore, the authentication is successful under this kind of attack scenario.

For the second type attack, only the low semantics is altered. Thus it holds that $\mathbf{S}' = \mathbf{S}$ and $\mathbf{H}' \neq \mathbf{H}$. This can be easily be detected since the terminal can obtain the TCR for the received content as ($\mathbf{S}$ $\mathbf{H}$). According to the robustness and sensitivity properties, the terminal can be alarmed that $\mathbf{H}' \neq \mathbf{H}$. Note that the aim of this attack is to distort the low semantics and thus the difference between $\mathbf{H}'$ and $\mathbf{H}$ is large enough to be detected. Otherwise, the $\mathbf{H}'$ falls into the robustness property of $\mathbf{H}$ and the low semantics of the attacked version could be almost the same as the original one. This attack scenario can also be successfully detected.

For the third kind of attack, both the high and low semantics are changed. We have $\mathbf{S}' \neq \mathbf{S}$ and $\mathbf{H}' \neq \mathbf{H}$. It means that the attacked multimedia content is completely different from the original one. By comparing the HLS of the content, the URA center could return TCR ($\mathbf{S}'$ $\mathbf{H}'$) to the terminal if by chance the attacked content is the same as other legal contents. But this chance could be almost impossible since we suppose the attack is to distort the semantics and not to change the content into another legal one. Similar to the first attack scenario, $\mathbf{H}$ could be void. Thus the terminal can know that the content lacks trust.

The above analysis shows that the general multimedia trust authentication framework can detect the attacks which have already happened. It authenticates the contents in a passive way. However, it can also authentication the content actively, which means that it can prevent the attacks from happening. Note that the base station first verifies the credit of the terminal before it can access the resources of the 5G networks. If certain attacker pretends legal user and asks for permission to the network, it lacks the credit and will fail to be part of the wireless network. However, we state that our framework cannot stop the legal terminal from harming the 5G network security. For instance, the legal user could download the legal content, deliberately alter the content, and reupload the content. However, we believe the legal user or terminal could behave securely and legally and this scenario is beyond our analysis for the authentication framework.

6. Conclusion

The secure and sustainable architecture of the 5G networks is vital for the networks health developments. We propose a general multimedia content trust authentication framework to verify the various categories of multimedia. The framework is novel in that it adopts two channels to transmit the information; i.e., one is for the usual multimedia and the other is for the authentication information. The former one is open and prone to attacks while the latter is closed and free of attacks. The framework adopts the trusted content representation technique, which authenticates the contents in both high and low level semantics. The high level semantics are conceptual terms generated to locate the contents from humans' understanding. The low level semantics are robust and perceptual terms to measure the integrity of perception. We analyze the security of the framework and show that it can authenticate the multimedia contents actively and passively. In our future work, we will look at the implementation of the general multimedia contents trust authentication by simulating upon 5G systems. Attention should also be focused on the attacks and security experiments on the proposed model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant no. 61771185, Grant no. 61772175), Doctoral Research Foundation of Southwest University of Science and Technology (Grant no. 17zx7158), Science and Technology Research Project of Henan Province (Grant no. 182102210044, Grant no. 182102210708), and Key Scientific Research Program of Henan Higher Education (Grant no. 18A510009, Grant no. 17A520005).

References

[1] A. Tzanakaki, M. Anastasopoulos, I. Berberana et al., "Wireless-optical network convergence: Enabling the 5G architecture to support operational and end-user services," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 184–192, 2017.

[2] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, 2014.

[3] L. Yan, X. Fang, and Y. Fang, "A Novel Network Architecture for C/U-Plane Staggered Handover in 5G Decoupled Heterogeneous Railway Wireless Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3350–3362, 2017.

[4] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. Di Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 20–27, 2015.

[5] R. Chaudhary, N. Kumar, and S. Zeadally, "Network Service Chaining in Fog and Cloud Computing for the 5G Environment: Data Management and Security Challenges," *IEEE Communications Magazine*, vol. 55, no. 11, pp. 114–122, 2017.

[6] P. Gandotra and R. K. Jha, "A survey on green communication and security challenges in 5G wireless communication networks," *Journal of Network and Computer Applications*, vol. 96, pp. 39–61, 2017.

[7] L. Xing, Q. Ma, and L. Jiang, "Microblog user recommendation based on particle swarm optimization," *China Communications*, vol. 14, no. 5, pp. 134–144, 2017.

[8] Q. Ma, L. Xing, and L. Zheng, "Authentication of Scalable Video Coding Streams Based on Topological Sort on Decoding Dependency Graph," *IEEE Access*, vol. 5, pp. 16847–16857, 2017.

[9] L. Xing, Z. Zhang, H. Lin, and F. Gao, "Content Centric Network with Label Aided User Modeling and Cellular Partition," *IEEE Access*, vol. 5, pp. 12576–12583, 2017.

[10] M. Asikuzzaman, M. J. Alam, A. J. Lambert, and M. R. Pickering, "Imperceptible and robust blind video watermarking using chrominance embedding: a set of approaches in the DT CWT domain," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1502–1517, 2014.

[11] I. Setyawan and I. K. Timotius, "Content-dependent spatio-Temporal video watermarking using 3-dimensional discrete cosine transform," in *Proceedings of the 2013 5th International Conference on Information Technology and Electrical Engineering, ICITEE 2013*, pp. 79–83, Yogyakarta, October 2013.

[12] B. Sridhar and C. Arun, "An enhanced approach in video watermarking with multiple watermarks using wavelet," *Journal of Communications Technology and Electronics*, vol. 61, no. 2, pp. 165–175, 2016.

[13] Z. Shahid and W. Puech, "Visual protection of HEVC video by selective encryption of CABAC binstrings," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 24–36, 2014.

[14] G. Van Wallendael, A. Boho, J. De Cock, A. Munteanu, and R. Van De Walle, "Encryption for high efficiency video coding with video adaptation capabilities," in *Proceedings of the 2013 IEEE International Conference on Consumer Electronics, ICCE 2013*, pp. 31-32, Las Vegas, USA, January 2013.

[15] X. Kang and Y. Wu, "A trust-based pollution attack prevention scheme in peer-to-peer streaming networks," *Computer Networks*, vol. 72, pp. 62–73, 2014.

[16] Y. Lu, W. Wang, B. Bhargava, and D. Xu, "Trust-based privacy preservation for peer-to-peer data sharing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 36, no. 3, pp. 498–502, 2006.

[17] C. Cheng, T. Jiang, and Q. Zhang, "TESLA-based homomorphic MAC for authentication in P2P system for live streaming with network coding," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 291–298, 2013.

[18] A. Habib, D. Xu, M. Atallah, B. Bhargava, and J. Chuang, "A tree-based forward digest protocol to verify data integrity in distributed media streaming," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 7, pp. 1010–1013, 2005.

[19] K. Mokhtarian and M. Hefeeda, "Authentication of scalable video streams with low communication overhead," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 730–742, 2010.

[20] L. N. Vadlamudi, R. P. V. Vaddella, and V. Devara, "Robust hash generation technique for content-based image authentication using histogram," *Multimedia Tools and Applications*, vol. 75, no. 11, pp. 6585–6604, 2016.

[21] Y.-J. Park, "The adaptive clustering method for the long tail problem of recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1904–1915, 2013.

[22] M. O'Mahony and M. Bray, "The long tail of content," *Communications Engineer*, vol. 4, no. 4, pp. 20–25, 2006.

[23] L. Xing, J. Ma, X.-H. Sun, and Y. Li, "Dual-mode transmission networks for DTV," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 474–480, 2008.

[24] L. Xing, L. Jiang, G. Yang, and B. Wen, "A novel trusted computing model for network security authentication," *Journal of Networks*, vol. 9, no. 2, pp. 339–343, 2014.

[25] L. Xing, Q. Ma, and M. Zhu, "Tensor semantic model for an audio classification system," *Science China Information Sciences*, vol. 56, no. 6, pp. 1–9, 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 8284617, 17 pages
https://doi.org/10.1155/2018/8284617

Research Article

A Sparse Temporal Synchronization Algorithm of Laser Communications for Feeder Links in 5G Nonterrestrial Networks

Lichen Zhu, Hangcheng Han, Xiangyuan Bu, and Jichao Wang

School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Correspondence should be addressed to Hangcheng Han; hanhangcheng@bit.edu.cn

Received 26 January 2018; Accepted 1 May 2018; Published 20 June 2018

Academic Editor: Nan Yang

To foster the rollout of 5G in unserved areas, 3GPP has kicked off a study item on new radio to support nonterrestrial networks (NTNs). Due to ultra-wideband of laser, laser communication is very promising for the feeder links of NTNs; however, imprecise temporal synchronization hinders its deployment, which results from a combination of propagation delay, velocity, acceleration, and jerk of NTN platform. The prior synchronization algorithms are inapplicable to the temporal synchronization in laser communications due to the extremely high data rate and Doppler shift. This paper is devoted to addressing the temporal synchronization problem in laser communications. In particular, we first observe the sparsity of laser signal in time-frequency domain. On top of this observation, we propose a new sparsity-aware algorithm for temporal synchronization without carrier aid through sparse discrete polynomial-phase transform and sparse discrete fractional Fourier transform. Subsequently, we implement the proposed algorithm via designing a hardware prototype. To further evaluate its performance, we conduct extensive simulations, and the results demonstrate the effectiveness of the proposed algorithm in terms of good accuracy, low power consumption, and low computational complexity, suggesting its attractiveness for the feeder links of 5G NTNs.

1. Introduction

With the exponential growth of wireless traffic volume, new radio (NR) becomes the foundation of 5G to provide universal coverage [1, 2]. To foster the rollout of 5G in unserved areas, nonterrestrial networks (NTNs) have been kicked off by 3GPP [3–5]. In this context, the major challenge is that the feeder links between the gateway and the NTN platforms must be broadband; i.e., the data rate should be on the order of gigabits per second (Gbps) [6–8]. Fortunately, laser is innately of the ultra-wideband, and the laser communication is thus very promising for the feeder links in satisfying the data rate requirement, while it is difficult for laser communications to achieve precise temporal synchronization because of the high data rate, e.g., the order of Gbps, and the high Doppler resulting from the movement of NTN platforms [9, 10].

To address the temporal synchronization problem, a large body of works have been proposed, which can be classified into three categories: tracking loop-based methods, transformation-domain methods, and sparse transformation-domain methods.

Specifically, tracking loop is a type of conventional temporal synchronization algorithms [11–14]. In [11], the authors developed a novel architecture for tracking loop to accurately measure the aircraft's speed, which is helpful for temporal synchronization. In [12, 13], the performance of tracking loop on ultra-wideband impulse radio was analyzed. A common tracking loop called digital delay locked loop (DDLL) was discussed in [14] for tracking direct sequence spread spectrum signal with high Doppler. As analyzed in these works, tracking loop algorithms are suitable for systems with low Doppler and low data rate. While the convergence rate degrades rapidly with the increase of Doppler and Doppler changing rate. Furthermore, high data rate system requires parallel architecture of tracking loop, which needs more resources than the serial one because multiple samples have to be handled during one clock cycle [15, 16] and

cannot meet the low power consumption requirement on NTN platforms [17].

The transformation-domain algorithms [18–21] follow open loop principle and are able to maintain a constant processing delay in high Doppler environments. In [18], the discrete fractional Fourier transform (DFrFT) was adopted for acceleration and velocity estimation. In [19], the discrete polynomial-phase transform (DPT) was used to estimate the aircraft dynamic parameters. In [20], the authors studied the time-frequency characteristics from the perspective of signal time-frequency distribution, i.e., Wigner-Ville distribution, to facilitate the temporal synchronization. In [21], the keystone transform was employed to solve the target range migration problem in temporal synchronization. Although transformation-domain algorithms are suitable for the scenario with high Doppler, their computational complexity is usually very high, which violates the low power consumption requirement as well.

To reduce the complexity, the sparse transformation-domain algorithm is introduced into high Doppler and high data rate systems [22–25]. The authors in [22] proposed a sparse algorithm based on the fast Fourier transform (FFT), namely, sparse FFT (SFFT). The applications of SFFT were introduced in [23], and the authors in [24] further evaluated the implementation performance of SFFT. The paper [25] designed a sparse using the discrete fractional Fourier transform (DFrFT) for the temporal synchronization in the scenario with high Doppler changing rate. With low complexity, the sparse algorithm is recommended to low power consumption platforms. However, these sparse transformation-domain algorithms require carrier recovery. It is pointed out that carrier recovery is difficult to implement on laser communications for NTN platforms because of the complex structure [26] and then restricts the application of the sparse transformation-domain algorithm on the feeder links of 5G NTNs.

Motivated from the observations above, we employ an incoherent laser transmission system called intensity modulation direct detection (IMDD) for its simplicity and low cost [27, 28]. The IMDD system can be modeled as a Gaussian channel with the positive real input signal and input-dependent noise [26, 29]. This paper designs a sparse transformation-domain algorithm in IMDD-based laser communications for 5G NTNs, which is of good accuracy, low power consumption, and low computational complexity. The primary contributions of the paper are summarized as follows.

(1) We develop a temporal synchronization algorithm based on the sparse pilot. Armed with the sparse DPT (SDPT) and the sparse DFrFT (SDFrFT), the proposed algorithm is able to estimate the propagation delay, velocity, acceleration, and jerk without carrier aid in high dynamic environment.

(2) We analyze the accuracy and complexity of the proposed algorithm and compare its performance with the nonsparse algorithms and the DDLL algorithms. The analytical and experimental results show that the proposed algorithm performs similar to the nonsparse algorithm and superior to the DDLL algorithm in terms of accuracy. While the proposed algorithm can achieve lower complexity, making it the most suitable for scenarios with high Doppler and low power consumption among these three algorithms.

(3) We discuss the implementation issues including module reuse analysis and clock rate analysis. According to the analysis, the proposed algorithm can be implemented with less clock rate than the conventional DDLL algorithm.

(4) We implement the proposed algorithm via designing a hardware prototype, and the results agree well with the theoretical analysis, demonstrating that the algorithm can be implemented on resource constrained platforms, e.g., 5G NTN platforms.

The rest of the paper is organized as follows. In Section 2, the system model is described. The proposed temporal synchronization algorithm is then presented in Section 3 and analyzed in Section 4. The related implementation issues are shown in Section 5. Simulation and experiment results are given in Sections 6 and 7 to demonstrate the effectiveness of the proposed algorithm, respectively, which are followed by the conclusions drawn in Section 8.

2. System Model

The system model of NTNs is presented in Figure 1, where the IMDD-based laser communications serve for the feeder link between spaceborne platform and gateway.

The laser signal is received by photodetection and is transformed to an electrical signal with positive and negative level, which is expressed as

$$\begin{aligned} s(t) &= A_0 g(t - \tau(t)) + w(t) \\ &= A_0 g\left(t - \left(a_0 + a_1 t + a_2 t^2 + a_3 t^3\right)\right) + w(t), \end{aligned} \tag{1}$$

where $g(t)$ is the waveform of signal with amplitude A_0, $\tau(t)$ denotes the propagation delay, which contains four components related to distance (a_0), velocity (a_1), acceleration (a_2), and jerk (a_3), respectively, and $w(t)$ is the zero-mean additive white Gaussian noise with variance σ^2. The experimental results show that $a_1 \gg a_2 \gg a_3$ [11]. The nature of temporal synchronization is to estimate $\tau(t)$, namely, a_0, a_1, a_2, and a_3, which can be obtained with the aid of pilot.

Figure 2 shows the pilot position in the transmitted frame, where L_{frm}, L_h, and L_p denote the lengths of the entire frame, the frame header, and the pilot, respectively. The pilot waveform, denoted by $g_p(t)$, is a sequence of periodic square waves and is expressed as

$$g_p(t) = \begin{cases} \frac{1}{2}, & kT_{syb} < t \le \left(k + \frac{1}{2}\right) T_{syb}, \\ -\frac{1}{2}, & \left(k + \frac{1}{2}\right) T_{syb} < t \le (k+1) T_{syb}, \end{cases} \tag{2}$$

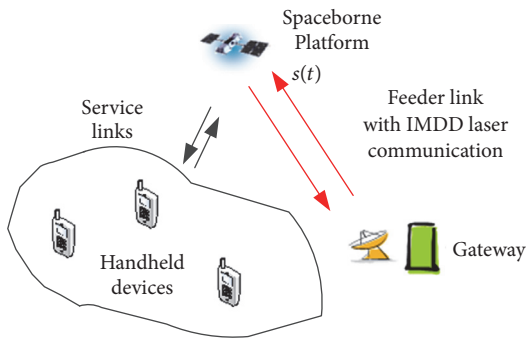


FIGURE 1: The system model of NTNs.

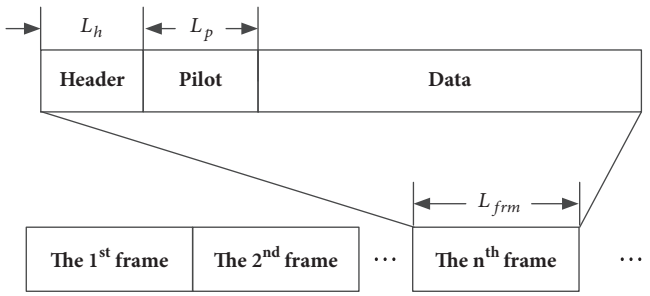


FIGURE 2: The pilot in the transmitted frame.

where T_{syb} denotes the symbol duration and $k = 0, 1, 2, \ldots, L_p - 1$. According to the Fourier series, we rewrite (2) as

$$g_p(t) = \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin\left(2(2k-1)\pi f_c t\right), \tag{3}$$

where $f_c = 1/(2T_{syb})$. Through low-pass filtering, (3) is converted to

$$g_p(t) = \sum_{k=1}^{\infty} \frac{A_k}{2k-1} \sin\left(2(2k-1)\pi f_c t\right), \quad A_k \in (0,1), \tag{4}$$

where A_k denotes the amplitude of the kth order harmonic, which decreases rapidly with the increase of k.

After analog-digital conversion, from (1) and (4), we have

$$\begin{aligned} s(n) &= r(n) + w(n) \\ &= A_0 g\left(nT_s - \tau\left(nT_s\right)\right) + w(n), \\ g_p\left(nT_s\right) &= \sum_{k=1}^{\infty} \frac{A_k}{2k-1} \sin\left(2(2k-1)\pi f_c nT_s\right), \end{aligned} \tag{5}$$

where T_s denotes the sample interval.

To simplify the computation, we assume $A_k = 0$ when $k > 3$; then $g_p(nT_s)$ in (5) is rewritten as

$$g_p\left(nT_s\right) = \sum_{k=1}^{3} \frac{A_k}{2k-1} \sin\left(2(2k-1)\pi f_c nT_s\right). \tag{6}$$

3. The Proposed Temporal Synchronization Algorithm

As $g_p(nT_s)$ is the sum of sinusoidal functions, the parameters estimation can be expressed as follows according to the maximum likelihood principle [30, 31],

$$\begin{aligned} &\{\hat{a}_1, \hat{a}_2, \hat{a}_3\} \\ &= \underset{a_1,a_2,a_3}{\operatorname{argmax}} \left| \sum_n s_c(n)\, e^{-j2\pi((f_c - a_1)nT_s - a_2(nT_s)^2 - a_3(nT_s)^3)} \right|, \\ &\hat{a}_0 = \arg\left(\sum_n s_c(n)\, e^{-j2\pi((f_c - \hat{a}_1)nT_s - \hat{a}_2(nT_s)^2 - \hat{a}_3(nT_s)^3)} \right), \end{aligned} \tag{7}$$

where $s_c(n) = s(n) + j \cdot s_{\pi/4}(n)$, and $s_{\pi/4}(n)$ represents $s(n)$ with $\pi/4$ phase shift. Based on (7), we will elaborate how to perform the temporal synchronization in four steps, as shown in Figure 3. We conduct the preprocess in the first step to facilitate the following three steps. In the second step, as estimating $\hat{a}_3$ is difficult within limited process duration in high Doppler environment, the parameter $\hat{a}_2$ will be preferentially derived based on a two-order SDPT (SDPT$_2$). In the third step, the parameters $\hat{a}_0$ and $\hat{a}_1$ will be obtained based on SDFrFT, and $\hat{a}_2$ will also be improved accordingly. The parameter $\hat{a}_3$ will be presented based on a three-order SDPT (SDPT$_3$) after resampling in the fourth step. Finally, we output all the estimated results, namely, $\hat{\tau}(nT_s)$.

3.1. The First Step: Preprocess. To prepare data for estimation, preprocess contains three stages.

First, the pilot is located with the aid of frame header. Based on the cross-correlation between the received signal and the template in header, the pilot is located by

$$\begin{aligned} \hat{n}_p = \arg\max_n &\left(\sum_{b=0}^{\lfloor f_s L_h T_{syb} \rfloor - 1} s(n) \cdot g_h\left((n-b)T_s\right) \right) \\ &+ \left\lfloor f_s L_h T_{syb} \right\rfloor, \end{aligned} \tag{8}$$

where $\lfloor \cdot \rfloor$ denotes the floor function, $f_s = 1/T_s$ denotes the sample rate, and $g_h(nT_s)$ is the header waveform.

Second, we derive the complex-form of $s(n)$ to meet the requirement of (7). Considering the $\pi/4$ phase shift of $g_p(nT_s)$, namely,

$$\begin{aligned} g_{p-\pi/4}\left(nT_s\right) &= g_p\left(nT_s - \frac{T_{syb}}{2}\right) \\ &= \sum_{k=1}^{3} \frac{A_k}{(2k-1)} \sin\left(2(2k-1)\pi f_c\left(nT_s - \frac{T_{syb}}{2}\right)\right) \\ &= \sum_{k=1}^{3} \frac{(-1)^{k-1} A_k}{(2k-1)} \cos\left(2(2k-1)\pi f_c nT_s\right), \end{aligned} \tag{9}$$

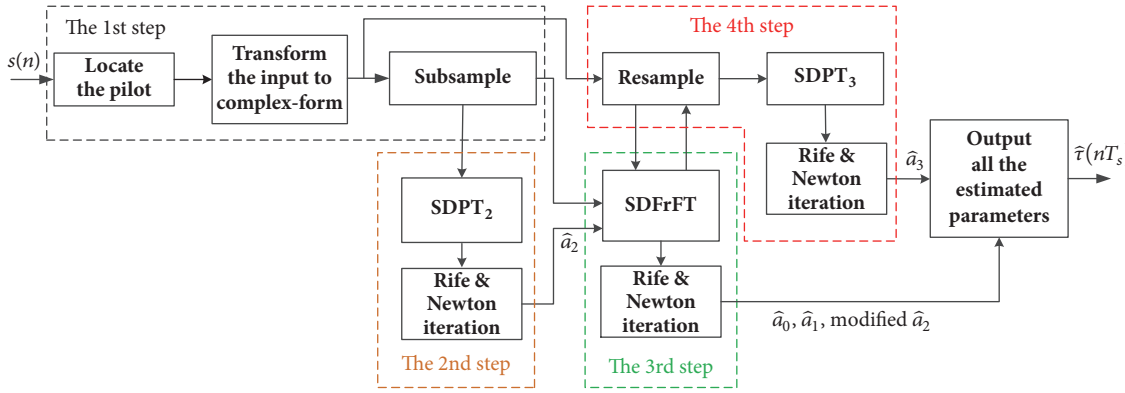


Figure 3: The diagram of the sparsity-based scheme.

and the complex-form of $g_p(nT_s)$ is expressed as

$$\begin{aligned} g_c(nT_s) &= g_{p-\pi/4}(nT_s) + j \cdot g_p(nT_s) \\ &= \sum_{k=1}^{3} \frac{(-1)^{k-1} A_k}{2k-1} \exp\left((-1)^{k-1} j2(2k-1)\pi f_c nT_s\right) \\ &= \mathbf{A}^T \mathbf{P}, \end{aligned} \tag{10}$$

where $\mathbf{A} = [A_1, -A_2/3, A_3/5]^T \in \mathbb{R}^3$, and $\mathbf{P} = [P_1, P_2, P_3]^T \in \mathbb{C}^3$ with $P_k = \exp(j(-1)^{k-1}2(2k-1)\pi(f_c nT_s - \tau(nT_s)))$ and $k = 1, 2, 3$. From (10), the complex-form of $s(n)$ is

$$s_c(n) = A_0 g_c(nT_s) + w(n) = \mathbf{A_0}^T \mathbf{P} + w(n), \tag{11}$$

where $\mathbf{A_0} = [A_0A_1, -A_0A_2/3, A_0A_3/5]^T \in \mathbb{R}^3$.

Third, since the pilot has been located, we obtain N_{seq} entries by subsampling the pilot data start from $\hat{n}_p$ with subsample rate $r_{smp} = \beta L_{frm} T_{syb}/T_s$, where β is a positive integer. The challenge here lies in subsampling. The subsequent samples will drift away from the original position caused by Doppler, and the sample position may be out of the pilot range. Hence, a sufficient length of pilot, i.e., L_p, is required to prevent the sampling position out of range. In the following, we discuss the minimum value of L_p.

Letting L_m denote the maximum range of drift, we have

$$L_m = N_{seq} \frac{|a_{1m}| \beta L_{frm}}{c}, \tag{12}$$

where a_{1m} denotes the maximum velocity. The reason for ignoring a_2 and a_3 is that the platform cannot accelerate any more after it reaches the maximum velocity. Generally, as the drift direction is unknown, the lower limit of L_p is $\lceil 2L_m \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. An improved method can reduce the lower limit of L_p to $\lceil L_m \rceil$ by simultaneously obtaining 3 sample sets that start from $\hat{n}_p$, $\hat{n}_p + (\lceil L_m \rceil/2) f_s T_{syb}$, and $\hat{n}_p + \lceil L_m \rceil f_s T_{syb}$, respectively, among which at least a set will not be out of range. Then we select the best one among these sets with the following method. Since the frame header is exploited to locate the frame, it assists in determining the drift direction and magnitude, as shown in Figure 4, where L_{1st} and L_{last} denote the intervals from the sample position to the header in the first frame and to the last frame, respectively, and L_d denotes the drift value with $L_d = L_{1st} - L_{last}$. The first, or second, or third sample set is selected when $L_d > L_p/2$, $L_d \in [-L_p/2, L_p/2]$, or $L_d < -L_p/2$, respectively.

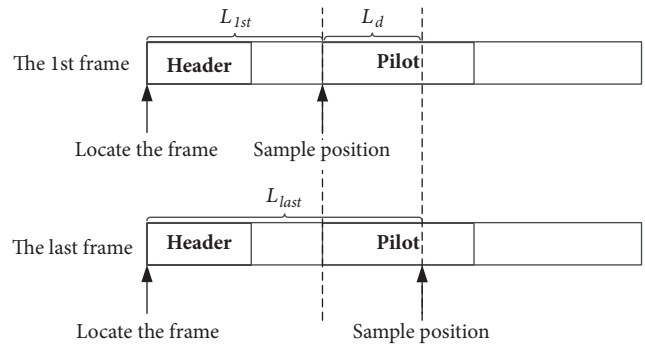


Figure 4: The method of determining drift direction and magnitude.

After subsampling, $s_c(n)$ is converted to $s_c(m)$ with $m = 0, 1, \ldots, N_{seq} - 1$, and the sample rate and sample interval turn into that $f_s' = f_s/r_{smp}$ and that $T_s' = r_{smp} \cdot T_s$, respectively. In the following section, we focus on the signal processing of the proposed temporal synchronization.

3.2. The Second Step: Derive $\hat{a}_2$ Based on SDPT$_2$. According to (7), it is possible to jointly estimate all the four unknown parameters. Unfortunately, the complexity of joint estimation is unaffordable. To reduce the complexity, we decompose the parameter estimation problem into several subproblems to estimate the parameters one by one, which is essentially the same as the joint estimation because the parameters to be estimated have a weak association with

each other and the separate estimation results in no performance loss. Generally, the highest-order parameter, a_3, is estimated firstly [32]. However, as mentioned above, high Doppler makes the sample position drift away and leads to a limited process duration. Moreover, as a_3 represents the third derivative of distance, it cannot be accurately estimated in such short duration. In comparison, the acceleration estimation is much easier, and we thus preferentially estimate a_2.

As a reduced-order method to estimate a_2, DPT_2 is defined as [32]

$$\begin{aligned}\mathrm{DPT}_2\left(s_c(m),f,\xi\right) &\triangleq \mathrm{DFT}\left(\mathrm{DP}_2\left(s_c(m),\xi\right),f\right)\\ &= \sum_{m=0}^{N_{seq}-1}\mathrm{DP}_2\left(s_c(m),\xi\right)e^{-j2\pi fmT_s'},\end{aligned} \tag{13}$$

where DFT denotes the discrete Fourier transform with rotation factor $e^{-j2\pi fmT_s'}$, ξ is a positive integer, and DP_2 is given by

$$\mathrm{DP}_2\left(s_c(m),\xi\right) \triangleq s_c(m)\,s_c^*(m-\xi), \tag{14}$$

where $(\cdot)^*$ denotes the conjugate operation. For simplicity, let $x(m) = \mathrm{DP}_2(s_c(m),\xi)$, and we have

$$\begin{aligned}x(m) &= s_c(m)\,s_c^*(m-\xi)\\ &= \left(r(m)+w(m)\right)\left(r^*(m-\xi)+w^*(m-\xi)\right)\\ &= r(m)\,r^*(m-\xi)+r(m)\,w^*(m-\xi)\\ &\quad + r^*(m-\xi)\,w(m)+w(m)\,w^*(m-\xi),\end{aligned} \tag{15}$$

where

$$\begin{aligned}r(m)\,r^*(m-\xi) &= \left(\mathbf{A}_0^T\mathbf{P}\mathbf{P}_\xi{}^T\mathbf{A}_0\right) = \mathbf{A}_0^T\mathbf{P}\left(\mathbf{P}_\xi^*\right)^T\mathbf{A}_0\\ &= \mathbf{A}_0^T\begin{bmatrix}P_1P_1^*(\xi) & P_1P_2^*(\xi) & P_1P_3^*(\xi)\\ P_2P_1^*(\xi) & P_2P_2^*(\xi) & P_2P_3^*(\xi)\\ P_3P_1^*(\xi) & P_3P_2^*(\xi) & P_3P_3^*(\xi)\end{bmatrix}\mathbf{A}_0\end{aligned} \tag{16}$$

is the principal component of $x(m)$; $\mathbf{P}_\xi$ and $P(\xi)$ denote $\mathbf{P}$ and P with ξ-delay, respectively. The other components of $x(m)$ are related to $w(m)$ and can be regarded as the noise terms. Moreover, $r(m)r^*(m-\xi)$ is dominated by $P_1P_1^*(\xi)$, $P_2P_2^*(\xi)$, and $P_3P_3^*(\xi)$. The amplitude of $P_1P_1^*(\xi)$ is much larger than that of $P_2P_2^*(\xi)$ and $P_3P_3^*(\xi)$. Consequently, the components of $r(m)r^*(m-\xi)$ except for $P_1P_1^*(\xi)$ are regarded as the disturbance terms, and (15) can be rewritten as

$$\begin{aligned}x(m) &= \left(A_0A_1\right)^2P_1P_1^*(\xi)+\ell\left(mT_s'\right)+\lambda\left(mT_s'\right)\\ &= \left(A_0A_1\right)^2e^{j2\pi\left(\left(-2a_2\xi T_s'^2+3a_3\xi^2T_s'^3\right)m-3a_3\xi T_s'^3m^2-\phi\right)}\\ &\quad +\ell\left(mT_s'\right)+\lambda\left(mT_s'\right),\end{aligned} \tag{17}$$

where $\ell(mT_s')$ and $\lambda(mT_s')$ denote the disturbance terms and noise terms, respectively, and ϕ is a constant. Letting $\widehat{X}(k) = \mathrm{DFT}(x(m))$ and $\widehat{k}_0 = \mathrm{argmax}\{|\widehat{X}(k)|\}$, since $3a_3\xi(T_s')^3m^2$ and $3a_3\xi^2(T_s')^3m$ are small enough to be ignored, $\widehat{a}_2$ can be expressed as

$$\widehat{a}_2 = -\frac{1}{2\xi\left(T_s'\right)^2}\widehat{k}_0. \tag{18}$$

Considering the sparsity of DFT result, SFFT can be employed to reduce the complexity, as shown in Algorithm 1.

As the accuracy of SDPT_2 is limited by the resolution of DFT, which is given by $\Delta f = f_s'/M$, accurately estimating a_2 cannot be achieved with small M. Thus, we adopt Rife and Newton methods to improve the estimation accuracy by updating $\widehat{k}_0$ and recalculating $\widehat{a}_2$ due to (18), as shown in Algorithm 2, whose effectiveness is demonstrated in Figure 5.

3.3. The Third Step: Obtain $\widehat{a}_0$, $\widehat{a}_1$, and the Improved $\widehat{a}_2$. In this part, $\widehat{a}_0$, $\widehat{a}_1$, and the improved $\widehat{a}_2$ are obtained by Pei sampling-type DFrFT [22], which is given in the following equation:

$$\begin{aligned}X_\alpha(k) &\triangleq \begin{cases}\sqrt{\dfrac{\sin\alpha-j\cos\alpha}{M}}e^{j(\cot\alpha/2)(kU_s)^2}\cdot\displaystyle\sum_{m=0}^{M-1}s_c(m)\,e^{-j(2\pi km/M)}e^{j(\cot\alpha/2)(mT_s')^2}, & \alpha\geq 0\\ \sqrt{\dfrac{j\cos\alpha-\sin\alpha}{M}}e^{j(\cot\alpha/2)(kU_s)^2}\cdot\displaystyle\sum_{m=0}^{M-1}s_c(m)\,e^{-j(2\pi km/M)}e^{j(\cot\alpha/2)(mT_s')^2}, & \alpha<0\end{cases}\\ &\approx \sqrt{\frac{1}{M}}e^{j(\cot\alpha/2)(kU_s)^2}\cdot\mathrm{DFT}\left(s_c(m)\,e^{j(\cot\alpha/2)(mT_s')^2}\right).\end{aligned} \tag{19}$$

with $m\in[0,M-1]$ and $k\in[0,M-1]$.

In (19), α and U_s denote the rotation angle and the sample interval of the output, $X_\alpha(k)$, respectively. Besides, $T_s'\times U_s = (2\pi|\sin\alpha|)/M$ shall be satisfied. Then we derive the optimal values of k and α by

$$\left\{\widehat{k}_m,\widehat{\alpha}_m\right\} = \arg\max_{\alpha,k}\left|X_\alpha(k)\right|. \tag{20}$$

Input: $x(m)$.
Output: $\widehat{X}(k)$.
1: Zero padding and the length of $x(m)$ is changed from N_{seq} to M.
2: Let $L = \log_2 M$.
3: For $i = 0; i < L; i{+}{+}$ **do**
4: Tear apart the spectrum randomly by $x_1(m) = x((\sigma_i \cdot m) \bmod M)$ and $X_1(k) = X((\sigma_i^{-1} \cdot m) \bmod M)$, where $X(k)$ and $X_1(k)$ denote the expressions of $x(m)$ and $x_1(m)$ in frequency-domain, respectively, and $m, \sigma_i, k \in [1, M]$.
5: Apply a flat window function to expand the spectrum range.
(1) Let $G(k) \in \begin{cases} [1-\delta, 1+\delta], & k \in [-\varepsilon' M, \varepsilon' M] \\ [0, \delta], & k \notin [-\varepsilon M, \varepsilon M] \end{cases}$ be the window function in frequency-domain,
where $\varepsilon' \in (0,1)$ and $\varepsilon \in (0,1)$ denote the cutoff frequencies of the passband and stopband, respectively, and δ denotes the ripple amplitude, whose reference value is $1/M^c$, where c is a positive integer.
(2) Compute $x_2(m) = g(m) \cdot x_1(m)$, where $g(m)$ denotes the expression of $G(k)$ in time-domain. The length of $g(m)$ is $\omega = o(B\log_2(M/\sigma_i))$.
6: Subsample in frequency-domain. Letting B be the data length after subsampling, we have $X_3(k) = \mathrm{FFT}(\sum_{j=0}^{\lfloor M/B-1 \rfloor} x_2(m + j \cdot B))$ with $k \in [1, B]$.
7: Map with a hash function.
(1) Define a hash function $h_\sigma(k) = \lfloor \sigma_i \cdot kB/M \rfloor$.
(2) Define an offset function $o_{\sigma_i}(k) = \sigma_i \cdot k - h_{\sigma_i}(k) \cdot B/M$.
(3) Let Γ_i denote the support set of the largest l coefficients in $X_3(k)$. The preimage set of Γ_i is I_i, whose size is lM/B, where $\mathrm{I}_i = \{k \in [1, M] \mid \mathrm{h}_{\sigma_i}(k) \in \Gamma_i\}$ and $B = \sqrt{Ml/\log_2(M/\delta)}$.
(4) Obtain the l largest spectrum coefficients as
$$\widehat{X}_{4,i}(k) = \begin{cases} \dfrac{X_3(h_\sigma(k)\mathrm{e}^{-j\pi o_{\sigma_i} k\omega/M})}{G(o_{\sigma_i}(k))}, & k \in \mathrm{I}_i, \\ 0, & k \in [1, \mathrm{M}] \cap \bar{\mathrm{I}}_i. \end{cases}$$
(5) Record the nonzero position of $\widehat{X}_{4,i}(k)$.
8: End for
9: Letting v_k be the occurrence times of coordinate k in the sets and only retaining the coordinate whose occurrence times are larger than $L/2$, we have $\mathrm{I}' = \{k \in \mathrm{I}_1 \cup \cdots \cup \mathrm{I}_L \mid v_k > L/2\}$, and the rest terms are zeroes.
10: For each coordinate in I', we obtain the corresponding spectrum coefficients $\widehat{X}^r(k)$ with $k \in \mathrm{I}'$ and $r = 1, 2, \cdots L$.
11: Return $\widehat{X}(k) = \begin{cases} \mathrm{median}\{\widehat{X}^r(k)\}, & k \in \mathrm{I}' \\ 0, & k \notin \mathrm{I}' \end{cases}$, where median$(\cdot)$ denotes computing the median of a sequence.

Algorithm 1: Estimate $\widehat{a}_2$ based on the SFFT algorithm.

1: Modify the DFT result by Rife method.
Since $\widehat{X}(k)$ denotes the output of SFFT, the modified result is
$$\widehat{k}_1 = \frac{1}{T_s'}\left[\widehat{k}_0 + r\frac{\left|\widehat{X}\left(\widehat{k}_0 + r\right)\right|}{\left|\widehat{X}\left(\widehat{k}_0\right)\right| + \left|\widehat{X}\left(\widehat{k}_0 + r\right)\right|}\right],$$
where
$$r = \begin{cases} 1, & |\widehat{X}(\widehat{k}_0 + 1)| \geq |\widehat{X}(\widehat{k}_0 - 1)| \\ -1, & |\widehat{X}(\widehat{k}_0 + 1)| < |\widehat{X}(\widehat{k}_0 - 1)|. \end{cases}$$
2: Further modify the result by the iteration of Newton method.
For a positive integer i, the $(i+1)$-th iteration result is expressed as $\widehat{k}_{i+1} = \widehat{k}_i - \lambda_i(\widehat{X}'(\widehat{k}_i)/\widehat{X}''(\widehat{k}_i))$, where $\widehat{X}'(\widehat{k})$ and $\widehat{X}''(\widehat{k})$ denote the first-order and the second-order derivatives of $\widehat{X}(\widehat{k})$, respectively, and λ denotes the tuning step.
3: Replace $\widehat{\mathbf{k}}_0$ by the latest iteration result, $\widehat{\mathbf{k}}_\mathbf{N}$.

Algorithm 2: Modify the DFT result based on Rife and Newton methods.

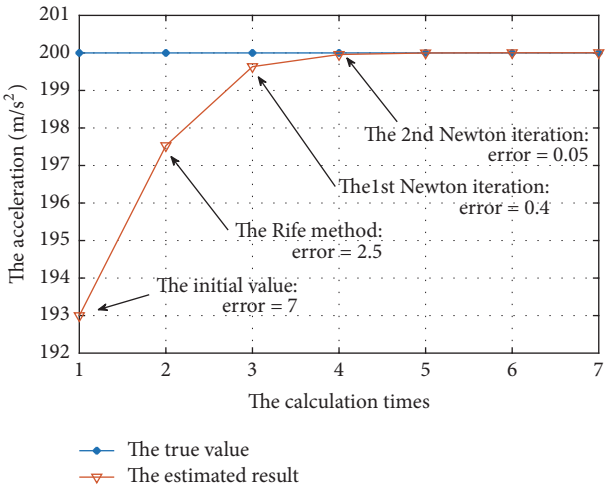


Figure 5: The effectiveness of Rife and Newton methods.

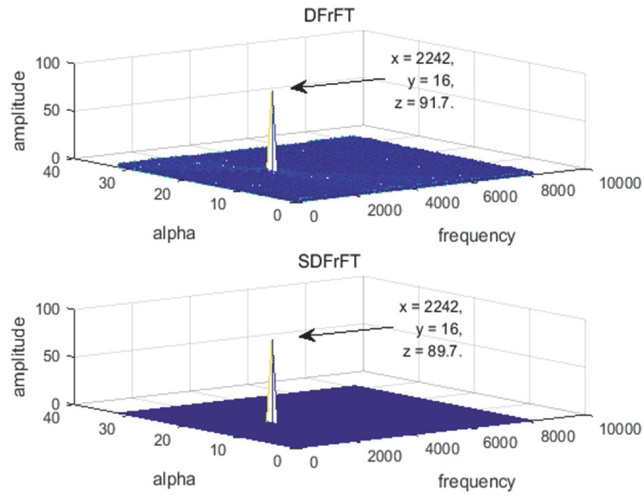


Figure 6: The comparison between DFrFT and SDFrFT.

From (19) and (20), we obtain the estimated results of $\hat{a}_0$, $\hat{a}_1$, and $\hat{a}_2$ in the way

$$\hat{a}_2 = \frac{\cot\left(\hat{\alpha}_m\right)}{4\pi},$$

$$\hat{a}_1 = \frac{\hat{k}_m f_s'}{M}, \tag{21}$$

$$\hat{a}_0 = \arg\left\{\frac{X_{\hat{\alpha}_m}\left(\hat{k}_m\right)}{-\operatorname{sgn}\left(\hat{\alpha}_m\right) e^{j(1/2)[\cot\hat{\alpha}_m(\hat{k}_m U_s)^2+\hat{\alpha}_m+\pi/2]}}\right\}.$$

To reduce the computational complexity, we replace DFT in the implementation of Pei sampling-type DFrFT with SFFT due to the sparse spectrum. The implementation of DFrFT based on SFFT is called SDFrFT, which exhibits a similar performance to DFrFT, as shown in Figure 6.

In addition, Rife and Newton methods can be employed again to replace $\hat{k}_m$ with $\hat{k}_N$, whose procedure is similar to that in Algorithm 2 and is omitted for brevity. Then we substitute $\hat{k}_N$ into (19), (20), and (21) to recalculate $\hat{a}_1$ and $\hat{a}_0$.

3.4. The Fourth Step: Resample and Derive $\hat{a}_3$ Based on SDPT$_3$. As mentioned above, high dynamic, especially high velocity brings the drift of sample position and results in an insufficient process duration, which is contrary to accurately estimating a_3. A method to mitigate the drift by updating the sample rate is introduced as follows.

Note that the velocity, $\hat{a}_1$, has been derived from (21); it can be substituted into the following formula to update the sample rate:

$$f_s'' = \frac{\left[2\hat{a}_1 + R_b\right] f_s'}{R_b}. \tag{22}$$

The sample interval turns into that $T_s'' = 1/f_s''$ accordingly. As f_s'' matches with the change of data rate caused by Doppler, the drift of sample position is mitigated, and estimating a_3 becomes easier with the increase of process duration.

In the following, SDPT$_3$ is introduced to estimate a_3. Assuming N_{seq}' entries are obtained after resampling, we have

$$\begin{aligned}\mathrm{SDPT}_3\left(s_c(m), f, \xi\right) &\triangleq \mathrm{SFFT}\left(\mathrm{DP}_3\left(s_c(m), \xi\right), f\right)\\ &= \sum_{m=0}^{N_{seq}'-1} \mathrm{DP}_3\left(s_c(m), \xi\right) e^{-j2\pi f m T_s''},\end{aligned} \tag{23}$$

where DP$_3$ is defined as

$$\begin{aligned}\mathrm{DP}_3\left(s_c(m), \xi\right) &\triangleq \mathrm{DP}_2\left(\mathrm{DP}_2\left(s_c(m), \xi\right), \xi\right)\\ &= \mathrm{DP}_2\left(s_c(m)\, s_c^*(m-\xi), \xi\right)\\ &= s_c(m)\, s_c^*(m-\xi)^2\, s_c(m-2\xi).\end{aligned} \tag{24}$$

Letting $\mathbf{P}_{2\xi}$ and $P(2\xi)$ denote $\mathbf{P}$ and P with 2ξ-delay, respectively, $x_2(m) = \mathrm{DP}_3(s_c(m), \xi)$, and $\lambda'(mT_s'')$ denote the noise terms, (24) can be further expanded as

$$\begin{aligned}x_2(m) &= s_c(m)\, s_c^*(m-\xi)^2\, s_c(m-2\xi)\\ &= \left(\mathbf{A}_0^T \mathbf{P}\left(\mathbf{P}_\xi^*\right)^T \mathbf{A}_0\right) \cdot \left(\mathbf{A}_0^T \mathbf{P}_\xi^*\left(\mathbf{P}_{2\xi}\right)^T \mathbf{A}_0\right)\\ &\quad + \lambda'\left(mT_s''\right)\\ &= \mathbf{A}_0^T \cdot \left(\mathbf{P}\left(\mathbf{P}_\xi^*\right)^T\right) \cdot \left(\mathbf{A}_0 \mathbf{A}_0^T\right) \cdot \left(\mathbf{P}_\xi^*\left(\mathbf{P}_{2\xi}\right)^T\right)\\ &\quad \cdot \mathbf{A}_0 + \lambda'\left(mT_s''\right)\end{aligned}$$

$$= \left(\mathbf{A}_0^T \otimes \mathbf{A}_0^T\right) \cdot \left(\mathbf{P}\left(\mathbf{P}_\xi^*\right)^T \otimes \mathbf{P}_\xi^*\left(\mathbf{P}_{2\xi}\right)^T\right) \cdot \left(\mathbf{A}_0 \otimes \mathbf{A}_0\right) + \lambda'\left(mT_s''\right), \tag{25}$$

where $\otimes$ denotes the Kronecker product. Obviously, $x_2(m)$ is dominated by $P_1P_1^*(\xi)P_1^*(\xi)P_1(2\xi)$, $P_2P_2^*(\xi)P_2^*(\xi)P_2(2\xi)$, and $P_3P_3^*(\xi)P_3^*(\xi)P_3(2\xi)$. Furthermore, as the amplitude of $P_1P_1^*(\xi)P_1^*(\xi)P_1(2\xi)$ is much larger than that of the other two components, (25) can be rewritten as

$$\begin{aligned} x_2(m) &= P_1P_1^*(\xi)P_1^*(\xi)P_1(2\xi) + \ell'\left(mT_s''\right) + \lambda'\left(mT_s''\right) \\ &= \left(A_0A_1\right)^4 e^{-j2\pi\left(6a_3\xi^2T_s''^3m + 2a_2\xi^2T_s''^2 - 6a_3\xi^3T_s''^3\right)} + \ell'\left(mT_s''\right) + \lambda'\left(mT_s''\right), \end{aligned} \tag{26}$$

where $\ell'(mT_s'')$ denotes the disturbance terms. From (26), $\hat{a}_3$ is obtained by

$$\hat{a}_3 = -\frac{1}{6\xi^2\left(T_s''\right)^3}\operatorname{argmax}\left\{\left|\mathrm{SFFT}\left(x_2(m)\right)\right|\right\}. \tag{27}$$

The estimation accuracy can be further improved by Rife and Newton methods, whose procedure is similar to that in Algorithm 2 and is omitted for brevity.

In addition, as longer duration is obtained after resampling, we conduct SDFrFT again to further improve the accuracy of $\hat{a}_2$, $\hat{a}_1$, and $\hat{a}_0$, whose procedure is similar to that in the third step and is also omitted for brevity.

4. Performance Analysis

To further present the advantages of the proposed temporal synchronization algorithm, we will theoretically analyze its performance in terms of estimation accuracy and algorithm complexity in this section.

4.1. Estimation Accuracy Analysis. As $\hat{a}_3$, $\hat{a}_2$, and $\hat{a}_1$ are derived from DPT_3, DPT_2, and DFrFT, respectively, we discuss the performance of these three algorithms as follows.

According to (17) and (26), the result of DPT contains the disturbance terms and the noise terms. The ratios of the principal components to the maximum disturbances in DPT_2 and DPT_3 are $9{A_1}^2/{A_2}^2$ and $81{A_1}^4/{A_2}^4$, respectively. As $A_1 \gg A_2$, the disturbance terms are negligible. The noise terms follow a zero-mean Gaussian distribution with variances $\sigma^2_{\lambda,\mathrm{DPT}_2}$ and $\sigma^2_{\lambda,\mathrm{DPT}_3}$, as shown in (28) and (29), respectively,

$$\sigma^2_{\lambda,\mathrm{DPT}_2} = E\left[\left|\left(\mathbf{A}_0^T\mathbf{P} + w(m)\right)\left(\mathbf{A}_0^T\mathbf{P} + w^*(m-\xi)\right) - \left(\mathbf{A}_0^T\mathbf{P}\right)^2\right|^2\right] = 2A^2\sigma^2 + \sigma^4 \tag{28}$$

$$\sigma^2_{\lambda,\mathrm{DPT}_3} = E\left[\left|\left(\mathbf{A}_0^T\mathbf{P} + w(m)\right)\left(\mathbf{A}_0^T\mathbf{P} + w^*(m-\xi)\right)^2 \cdot \left(\mathbf{A}_0^T\mathbf{P} + w(m-2\xi)\right) - \left(\mathbf{A}_0^T\mathbf{P}\right)^4\right|^2\right] = 6A^6\sigma^2 + 10A^4\sigma^4 + 4A^2\sigma^6 + \sigma^8 \tag{29}$$

with $A = A_0(A_1 + A_2/3 + A_3/5)$.

Considering the impact of sample position of pilot signal on the valid signal length, we discuss the SNRs of DPT_3, DPT_2, and DFrFT when the pilot length is insufficient as follows. Ignoring the impact of a_2 and a_3, the valid signal length is

$$N_{valid} = \frac{L_p \cdot c}{L_{frm} \cdot \left|a_{1m}\right|}. \tag{30}$$

The rest of signal is approximate random and can be regarded as noise, whose length is $(N_{seq} - N_{valid})$. The SNRs of DPT_3, DPT_2, and DFrFT are

$$\begin{aligned} \mathrm{SNR}_{\mathrm{DPT}_3} &= \frac{N_{valid}^2\left(A_0A_1\right)^8}{N'_{seq}\left(6A^6\sigma^2 + 10A^4\sigma^4 + 4A^2\sigma^6 + \sigma^8\right) + \left(N'_{seq} - N_{valid}\right)\left(A_0A_1\right)^8} \\ \mathrm{SNR}_{\mathrm{DPT}_2} &= \frac{N_{valid}^2\left(A_0A_1\right)^4}{N_{seq}\left(2A^2\sigma^2 + \sigma^4\right) + \left(N_{seq} - N_{valid}\right)\left(A_0A_1\right)^4} \\ \mathrm{SNR}_{\mathrm{DFrFT}} &= \frac{N_{valid}^2\left(A_0A_1\right)^2}{N_{seq}\sigma^2 + \left(N_{seq} - N_{valid}\right)\left(A_0A_1\right)^2}, \end{aligned} \tag{31}$$

respectively. The SNRs will be much lower than 0 dB when $N_{valid} \ll N_{seq}$, which leads to an inaccurate estimation result. Fortunately, N_{valid} can be obtained by (30) with a given a_{1m}, and $N_{valid} \approx N_{seq}$ can be ensured by selecting the data length.

Thus, we assume the pilot length is sufficient for estimation, and (31) can be simplified as follows in the high SNR regime:

$$\begin{aligned}\mathrm{SNR}_{\mathrm{DPT}_3} &\approx \frac{N'_{seq}\left(A_0A_1\right)^8}{6A^6\sigma^2+10A^4\sigma^4+4A^2\sigma^6+\sigma^8}\\ &\approx \frac{N'_{seq}\left(A_0A_1\right)^2}{6\sigma^2},\\ \mathrm{SNR}_{\mathrm{DPT}_2} &\approx \frac{N_{seq}\left(A_0A_1\right)^4}{2A^2\sigma^2+\sigma^4} \approx \frac{N_{seq}\left(A_0A_1\right)^2}{2\sigma^2},\\ \mathrm{SNR}_{\mathrm{DFrFT}} &\approx \frac{N_{seq}\left(A_0A_1\right)^2}{\sigma^2}.\end{aligned} \tag{32}$$

With the aid of Newton method, the accuracy of $\widehat{a}_3$, $\widehat{a}_2$, and $\widehat{a}_1$ based on DPT_3, DPT_2, and DFrFT is improved and given by

$$\begin{aligned}\sigma^2_{\widehat{a}_3} &= \frac{6}{(2\pi)^2 SNR_{\mathrm{DPT}_3}T_s''^2\left(N_{seq}'^2-1\right)}\\ &\approx \frac{36\sigma^2}{(2\pi)^2\left(A_0A_1\right)^2T_s''^2N'_{seq}\left(N_{seq}'^2-1\right)},\\ \sigma^2_{\widehat{a}_2} &= \frac{6}{(2\pi)^2 SNR_{\mathrm{DPT}_2}T_s''^2\left(N_{seq}'^2-1\right)}\\ &\approx \frac{12\sigma^2}{(2\pi)^2\left(A_0A_1\right)^2T_s'^2N_{seq}\left(N_{seq}^2-1\right)},\\ \sigma^2_{\widehat{a}_1} &= \frac{6}{(2\pi)^2 SNR_{\mathrm{DFrFT}}T_s''^2\left(N_{seq}'^2-1\right)}\\ &\approx \frac{6\sigma^2}{(2\pi)^2\left(A_0A_1\right)^2T_s'^2N_{seq}\left(N_{seq}^2-1\right)}.\end{aligned} \tag{33}$$

With a given SNR, the accuracy of $\widehat{a}_3$, $\widehat{a}_2$, and $\widehat{a}_1$ can be improved by lengthening data or increasing sample interval. Assuming $N'_{seq} = N_{seq}$ and $T_s'' \approx T_s'$, the accuracy relationship among $\widehat{a}_3$, $\widehat{a}_2$, and $\widehat{a}_1$ is

$$\sigma^2_{\widehat{a}_3} \approx 3\sigma^2_{\widehat{a}_2} \approx 6\sigma^2_{\widehat{a}_1}. \tag{34}$$

Note that the accuracy of $\widehat{a}_2$ can be further improved by α-search of DFrFT. Letting the estimation error before α-search and the search range of α-search be $x_{\widehat{a}_2}$ and $[\widehat{a}_2 - 3\sigma_{\widehat{a}_2}, \widehat{a}_2 + 3\sigma_{\widehat{a}_2}]$, respectively, and assuming that the search times are sufficient, the estimation error after α-search is

$$x'_{\widehat{a}_2} = \begin{cases} x_{\widehat{a}_2} + 3\sigma_{\widehat{a}_2}, & x_{\widehat{a}_2} < -3\sigma_{\widehat{a}_2},\\ 0, & x_{\widehat{a}_2} \in \left[-3\sigma_{\widehat{a}_2}, 3\sigma_{\widehat{a}_2}\right],\\ x_{\widehat{a}_2} - 3\sigma_{\widehat{a}_2}, & x_{\widehat{a}_2} > 3\sigma_{\widehat{a}_2}.\end{cases} \tag{35}$$

The expectation and variance of $x'_{\widehat{a}_2}$ are presented in the following equation:

$$\begin{aligned}E\left(x'_{\widehat{a}_2}\right) &= \int_{-\infty}^{-3\sigma_{\widehat{a}_2}}\left(x_{\widehat{a}_2}+3\sigma_{\widehat{a}_2}\right)\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2} + \int_{-3\sigma_{\widehat{a}_2}}^{3\sigma_{\widehat{a}_2}} 0\\ &\quad\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2} + \int_{3\sigma_{\widehat{a}_2}}^{+\infty}\left(x_{\widehat{a}_2}-3\sigma_{\widehat{a}_2}\right)\\ &\quad\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2} = 0,\\ \mathrm{var}\left(x'_{\widehat{a}_2}\right) &= \int_{-\infty}^{-3\sigma_{\widehat{a}_2}}\left(x_{\widehat{a}_2}+3\sigma_{\widehat{a}_2}\right)^2\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2}\\ &\quad+\int_{-3\sigma_{\widehat{a}_2}}^{3\sigma_{\widehat{a}_2}} 0\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2} + \int_{3\sigma_{\widehat{a}_2}}^{+\infty}\left(x_{\widehat{a}_2}-3\sigma_{\widehat{a}_2}\right)^2\\ &\quad\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2} = 2\left(\int_{-\infty}^{-3\sigma_{\widehat{a}_2}} x_{\widehat{a}_2}^2\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2}\right.\\ &\quad+6\sigma_{\widehat{a}_2}\int_{-\infty}^{-3\sigma_{\widehat{a}_2}} x_{\widehat{a}_2}\cdot p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2}\\ &\quad\left.+9\sigma^2_{\widehat{a}_2}\int_{-\infty}^{-3\sigma_{\widehat{a}_2}} p\left(x_{\widehat{a}_2}\right)dx_{\widehat{a}_2}\right) = 10\left(1\right.\\ &\quad\left.-\mathrm{erf}\left(\frac{3}{\sqrt{2}}\right)\right)\sigma^2_{\widehat{a}_2} \approx 0.027\sigma^2_{\widehat{a}_2},\end{aligned} \tag{36}$$

with $p(x_{\widehat{a}_2}) = (1/\sqrt{2\pi\sigma^2_{\widehat{a}_2}})e^{-x^2_{\widehat{a}_2}/2\sigma^2_{\widehat{a}_2}}$ and $\mathrm{erf}(x) = (2/\sqrt{\pi})\int_0^x e^{-t^2}dt$. If the search range becomes to $[-2\sigma_{\widehat{a}_2}, 2\sigma_{\widehat{a}_2}]$, the variance turns into $0.35\sigma^2_{\widehat{a}_2}$ accordingly.

Then we discuss the accuracy of $\widehat{a}_0$, which is derived from DFrFT due to (21). In the high SNR regime, $\widehat{a}_0$ follows a zero-mean Gaussian distribution with variance

$$\sigma^2_{\widehat{a}_0} = \frac{\sigma^2}{(2\pi)^2 N_{seq}\left(A_0A_1\right)^2\mathrm{sinc}^2\left(\Delta f\right)}, \tag{37}$$

where Δf denotes the offset between the DFT result and the true value.

Finally, the performance of SFFT is discussed as follows. Letting $\widehat{X}_{-\Gamma}(k) = \{\widehat{X}(k) \mid k \notin \Gamma\}$, the estimation error probability based on SFFT is [22]

$$\begin{aligned}&P\left[\left|\widehat{X}(k)-X(k)\right|^2 \geq \frac{\varepsilon}{l}\left\|\widehat{X}_{-\Gamma}(k)\right\|_2^2 + 3\delta^2\left\|\widehat{X}(k)\right\|_1^2\right]\\ &\quad< O\left(\frac{l}{\varepsilon B}\right).\end{aligned} \tag{38}$$

SFFT performs similar to FFT with a sufficient small error probability if $l \ll B$. Consequently, the accuracies of $\widehat{a}_3$, $\widehat{a}_2$, $\widehat{a}_1$, and $\widehat{a}_0$ based on SFFT are approximately equal to those based on FFT, which demonstrates the agreement between the proposed sparse algorithm and nonsparse algorithm.

Table 1: The algorithm complexity.

Approach	Multiplies	Adds
Correlation	$2f_sL_hT_{syb}$	$2\left(f_sL_hT_{syb}-1\right)$
DP2	N_{seq}	0
DFT		
FFT	$\frac{M}{2}\log_2 M$	$M\log_2 M$
SFFT	$\left(\omega+\frac{B}{2}\log_2 B+\frac{lM}{B}\right)L$	$\left(B\log_2 B+\frac{lM}{B}\right)L$
DFrFT		
DFrFT	$\left(M+\frac{M}{2}\log_2 M\right)\lambda_\alpha$	$(M\log_2 M)\,\lambda_\alpha$
SDFrFT	$\left(M+\left(\omega+\frac{B}{2}\log_2 B+\frac{lM}{B}\right)L\right)\lambda_\alpha$	$\left(B\log_2 B+\frac{lM}{B}\right)L\lambda_\alpha$
Rife & Newton methods	$\left(3N_{seq}+1\right)\lambda_{itrt}+2$	$\left(2N_{seq}-1\right)\lambda_{itrt}+2$

4.2. Algorithm Complexity Analysis. The algorithm complexity is shown in Table 1, where λ_{itrt} and λ_α denote the iteration times of Newton method and the α-search times, respectively. Referring to the values of ω, B, and L in Algorithm 1, the complexities of SFFT and FFT can be expressed as $O(\log_2 M\sqrt{Ml\log_2 M})$ and $O(M\log_2 M)$, respectively. In the large data size regime, the complexity of FFT is about M times more than that of SFFT.

In the condition of $N_{seq}=M$, considering DP_2 and Rife and Newton methods are employed three times, respectively, and SDFrFT and SFFT are employed twice, respectively, the total numbers of multiplications, C_{mul}, and additions, C_{add}, in the proposed sparse algorithm are

$$\begin{aligned} C_{mul} &\approx \left(9\lambda_{\text{itrt}}+2\lambda_\alpha\right)N_{seq} \\ &\quad +2\lambda_\alpha\log_2\left(N_{seq}\sqrt{N_{seq}l\log_2 N_{seq}}\right), \\ C_{add} &\approx 6\lambda_{\text{itrt}}N_{seq}+2\lambda_\alpha\log_2\left(N_{seq}\sqrt{N_{seq}l\log_2 N_{seq}}\right). \end{aligned} \tag{39}$$

Similarly, the total numbers of multiplications, $\dot{C}_{mul}$, and additions, $\dot{C}_{add}$, in the nonsparse algorithm are

$$\begin{aligned} \dot{C}_{mul} &\approx 9\lambda_{\text{itrt}}N_{seq}+\lambda_\alpha N_{seq}\log_2 N_{seq}, \\ \dot{C}_{add} &\approx 6\lambda_{\text{itrt}}N_{seq}+2\lambda_\alpha N_{seq}\log_2 N_{seq}. \end{aligned} \tag{40}$$

Figure 7 presents the complexity comparison of the proposed sparse algorithm with the nonsparse algorithm and the DDLL algorithm. The results indicate that the proposed algorithm has a much lower complexity than both the nonsparse algorithm and the DDLL algorithm.

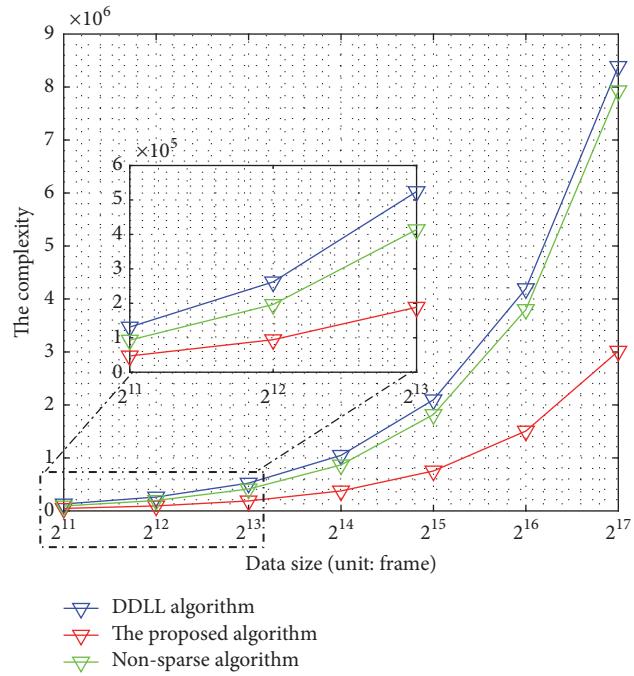


Figure 7: The comparison in terms of complexity.

5. Implementation Issues

In this section, we discuss the implementation issues of the proposed algorithm including the module reuse analysis and the clock rate analysis.

5.1. Module Reuse Analysis. In NTN platforms, with the requirement of low power consumption and low cost, the chip size is strictly limited, which means that the reuse of system module is necessary.

Figure 8 presents the implementation procedure of the proposed algorithm. According to Section 3, it mainly contains four steps, among which some modules can be reused. First, DP_3 can be conducted by twice DP_2 due to (24). Thus, $SDPT_3$ and $SDPT_2$ share a DP_2 module. Second, SDPT and SDFrFT share an SFFT module. Third, $SDPT_2$, $SDPT_3$, and SDFrFT share a "Rife and Newton methods" module. Consequently, a correlation module, a DP_2 module, an SFFT module, and a "Rife and Newton methods" module are sufficient for the whole procedure. We just need $f_sL_hT_{syb}$ real

Table 2: The comparison between DDLL and the proposed algorithm.

Items	The DDLL Algorithm	The proposed Algorithm	Ratio
Process clock rate	156.25 MHz	610.35 kHz	256:1
Degree of parallelism	64	1	64:1

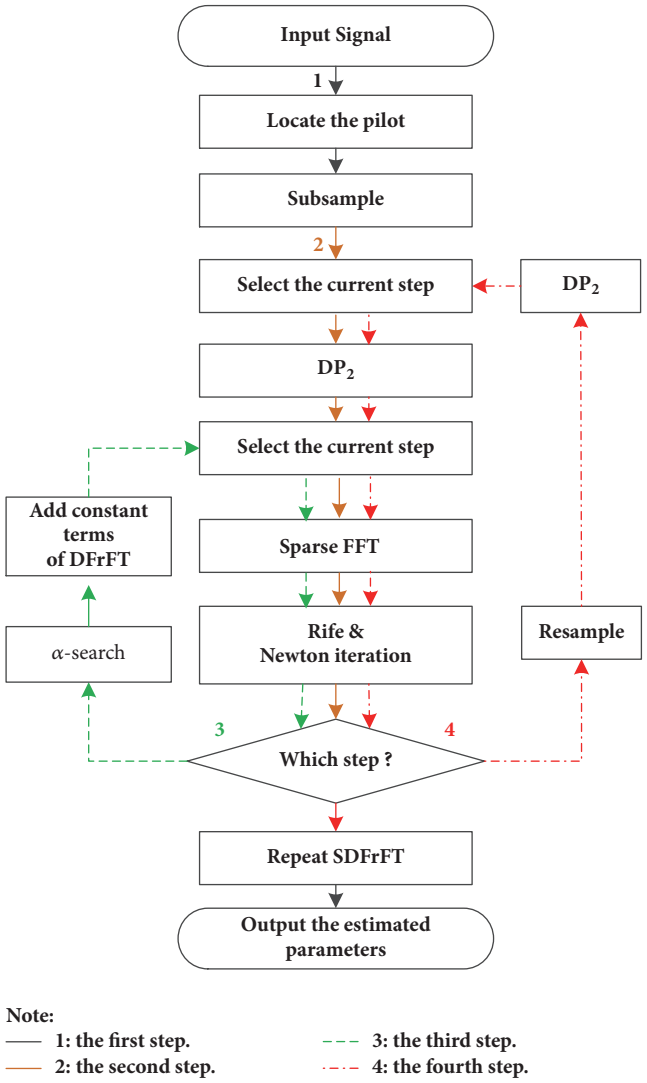


Figure 8: The implementation of the proposed algorithm.

multipliers, about $(B\log_2 B + lM/B + 2N_{seq} + f_s L_h T_{syb})$ adders, and about $(\omega + (B/2)\log_2 B + (l/B + 1)M + 4N_{seq})$ complex multipliers in total, which is easy to implement on chip, e.g., on the Field Programmable Gate Array (FPGA).

5.2. Clock Rate Analysis. The clock rate is a crucial parameter in high speed system. A 5 Gbps communication system requires a clock rate of $f_s \geq 10$ GHz, which is difficult to implement on FPGA as the maximum rate of FPGA is on the order of hundreds of MHz. To achieve this target clock rate, the parallel processing is required in the conventional DDLL algorithm, which would lead to high resource consumption. In comparison, the clock rate and the degree of parallelism can be decreased in the proposed sparse algorithm thanks to the subsampling, and the power consumption is accordingly decreased to a great extent.

Table 3: The critical parameters.

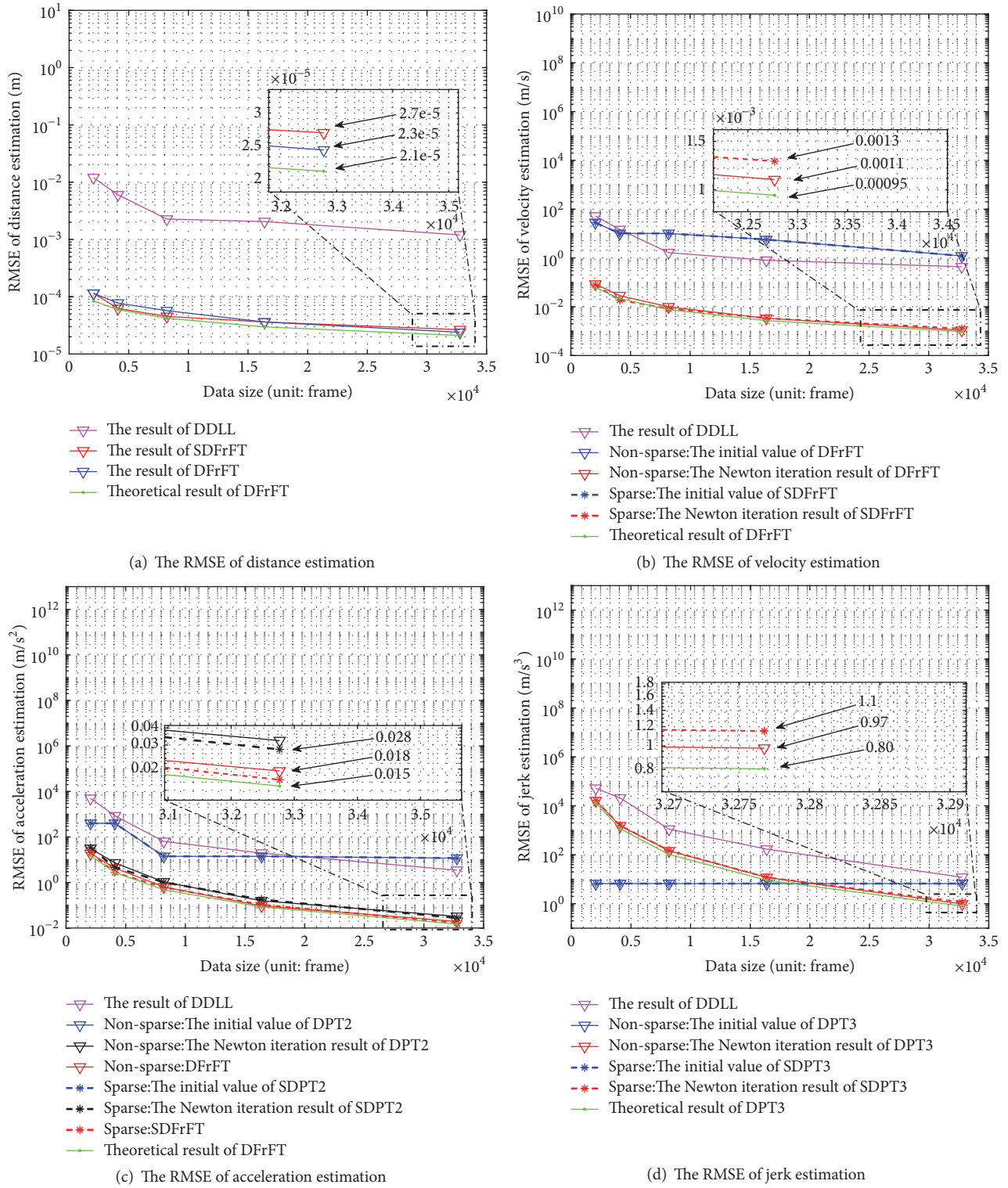
Parameter	Value
Modulation	OOK
Demodulation	Direct detection
Data rate	5 Gbps
SNR	14 dB
The maximum velocity	7 km/s
The maximum acceleration	800 m/s^2
The maximum jerk	60 m/s^3
L_{frm}	8192 bit
L_{syn}	32 bit
L_p	49 bit

The comparison between the DDLL algorithm and the proposed algorithm in terms of clock rate and degree of parallelism are shown in Table 2 with $L_{frm} = 8192$ bit, $\beta = 1$, $f_s = 10$ GHz, and $r_{smp} = \beta L_{frm} T_{syb}/T_s = 16384$. The numerical results demonstrate the superiority of the proposed algorithm over the DDLL algorithm in terms of power consumption. From the above discussions, the proposed algorithm can be applied to the spaceborne platform of 5G NTNs.

6. Simulation Results

In this section, we evaluate the performance of the proposed sparse algorithm in comparison with the nonsparse algorithm and the DDLL algorithm. The main parameters are listed in Table 3. In the simulation, we first verify the root-mean-square error (RMSE) of these three algorithms versus data size, N_{seq}. As shown in Figure 9, the theoretical results derived from (33), (36), and (37) are presented by the green lines in the figures. It is observed that the simulation results match well with theoretical results.

Specifically, Figure 9(a) plots the RMSE of distance estimation $(\widehat{a}_0)$ versus N_{seq}. Although there exists a small gap between the results of DFrFT and SDFrFT, the proposed sparse algorithm can still accurately estimate a_0. Moreover, it outperforms the DDLL algorithm, especially in the small data size regime. The reason is that DDLL needs long convergence time to achieve high-precision estimation in high dynamic environment. Although the convergence rate can be



(a) The RMSE of distance estimation

(b) The RMSE of velocity estimation

(c) The RMSE of acceleration estimation

(d) The RMSE of jerk estimation

FIGURE 9: The comparison of the three algorithms versus N_{seq}.

improved by increasing the loop bandwidth, it results in a high steady-state error and violates the accuracy requirement.

Figure 9(b) plots the RMSE of velocity estimation ($\widehat{a}_1$) versus N_{seq}. SDFrFT exhibits a similar performance to DFrFT on estimating a_1, which is similar to that on estimating a_0. Furthermore, with the aid of Newton method, SDFrFT outperforms the DDLL algorithm, especially in the small data size regime. From the result, a small L_p is sufficient for

estimating a_1, which indicates that the proposed algorithm can be implemented without enormous channel cost.

Figure 9(c) plots the RMSE of acceleration estimation ($\widehat{a}_2$) versus N_{seq}. As shown in the figure, the accuracy of DPT_2-based result is improved by Newton method and further improved by the α-search of DFrFT, which can be explained by the analysis in Section 4. The improved estimated result has a better accuracy than the DDLL-based result. In addition, in the small data size regime, there exists a larger gap between the results of the proposed algorithm and the DDLL algorithm in Figure 9(c) than that in Figure 9(b). This is because the transformation-domain algorithm is designed for estimating the acceleration, and DDLL responds more slowly in tracking the acceleration compared with tracking the velocity.

Figure 9(d) plots the RMSE of jerk estimation ($\widehat{a}_3$) versus N_{seq}. The comparison among DPT_3, $SDPT_3$, and DDLL indicates that there are a huge gap and a minor gap between the results of the proposed algorithm and the DDLL algorithm and between the results of the proposed sparse algorithm and the nonsparse algorithm, respectively. Moreover, in the small data size regime, in comparison to the acceleration estimation results in Figure 9(c), there exists a larger gap between the results of the proposed algorithm and the DDLL algorithm in terms of the jerk estimation. The reason is that $SDPT_3$ is designed for estimating the jerk, and DDLL responds more slowly in tracking the jerk than the acceleration, which presents the advantage of the proposed algorithm in high dynamic environment.

Next, we focus on the comparison between the sparse algorithm and nonsparse algorithm versus parameter β in the formula $r_{smp} = \beta L_{frm} T_{syb}/T_s$ with $N_{seq} = 2048$. The simulation results are plotted in Figure 10 and match well with the theoretical results.

Specifically, Figure 10(a) presents the RMSE of distance estimation ($\widehat{a}_0$) versus β. There exists no pronounced difference between the results of SDFrFT and of DFrFT, which agrees with the results in Figure 9(a). However, the variation tendency of the results in Figure 10(a) is different from that in Figure 9(a). The RMSE decreases with the increase of N_{seq} in Figure 9(a) and remains constant no matter the value of β in Figure 10(a). It can be explained by (37) that the accuracy of $\widehat{a}_0$ is related to data size and is independent of the sample interval.

Figures 10(b) and 10(c) plot the RMSEs of velocity estimation ($\widehat{a}_1$) and acceleration estimation ($\widehat{a}_2$) versus β, respectively. The results indicate that the performance is almost the same no matter the sparse method is used or not, which verifies the rationality of the sparse algorithm. Compared with lengthening data, increasing sample interval only has a slighter effect on estimating a_1 and a_2 in the condition of the same process duration. It can be explained by (33) that lengthening data contributes more to the accuracy than increasing sample interval. Fortunately, the difference between these two cases is so small that an accurate estimation still can be achieved by increasing the interval. As increasing sample interval improves the estimation accuracy without increasing computational burden, we prefer it rather than lengthening data for the sake of power consumption.

From the results in Figures 9(d) and 10(d), lengthening data performs better than increasing sample interval in the same process duration, which is similar to the comparison results of Figures 9(b), 9(c), 10(b), and 10(c). The gap can be filled up by slightly increasing the data size, e.g., as shown in Figure 11 with $N_{seq} = 4096$, whose results indicate that the configuration of $N_{seq} = 4096$ and $\beta = 16$ exhibits a prior performance to the configuration of $N_{seq} = 32768$ and $\beta = 1$ in Figure 9(d), and a much better performance than that of $N_{seq} = 2048$ and $\beta = 16$ in Figure 10(d).

From the above discussion, as increasing sample interval improves the estimation accuracy without increasing computation resource, it is a better choice than lengthening data in resource constrained platforms.

7. Experiment Results

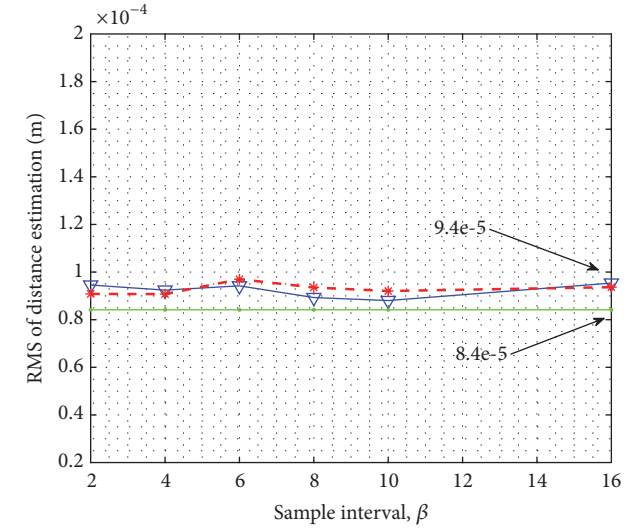
Referring to the parameters listed in Table 3, we have built a hardware platform for the experiment to verify the implementability of the proposed algorithm. The topology of hardware platform is given in Figure 12. The laser signal is received by telescope and is transformed to an electrical signal in the optical receiver. After that, the electrical signal is transmitted to channel simulator, which is employed to tune the transmission delay a_0 and delay rate a_1 with the control of PC, via the LAN extension for instrumentation (LXI) protocol. The output of channel simulator is sampled by an analog-digital converter (ADC). Then FPGA utilizes the samples to implement the proposed temporal synchronization algorithm according to Figure 8. The synchronization results are reported to PC, via the peripheral component interconnect express (PCI-E) protocol.

We tune the transmission delay and delay rate by channel simulator and estimate them by the proposed algorithm with $M = N_{seq} = 2048$ and $\beta = 2$. The experiment is repeated 50 times for each configuration of a_0 and a_1, and the statistical results are presented in Table 4, where the RMSEs of transmission delay and delay rate are on the orders of 0.01 ns and 0.3 ns/s, respectively. Due to the performance loss caused by the nonideal hardware and the nonideal channel, the experimental result performs inferior to the simulation result. Fortunately, the accuracy of experimental result still meets the demand of 5G NTNs, and the performance can be further improved by lengthening the process duration.

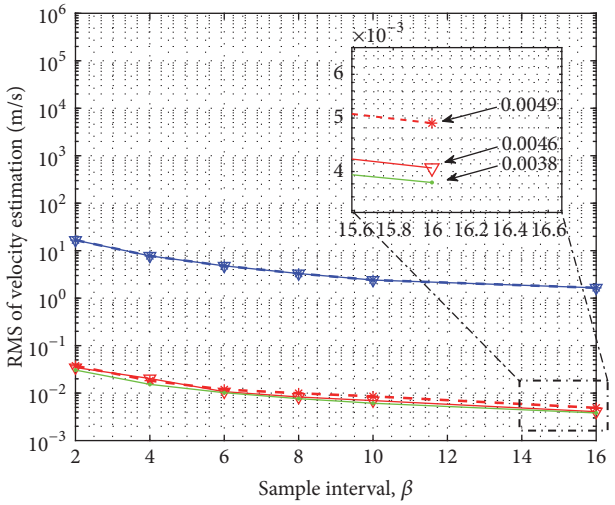
Next, we repeat the proposed temporal synchronization algorithm once per 50 milliseconds to continuously estimate the transmission delay, and the result is depicted in Figure 13. The result in Figure 13 indicates that the proposed algorithm can be implemented to continuously estimate the temporal parameters in real time, and the accuracy achieves about 0.01

Table 4: The experiment results of temporal synchronization.

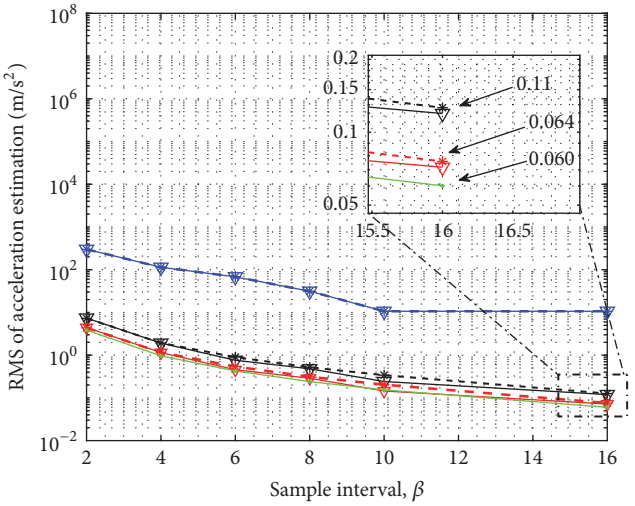
True value of a_0 (ns)	True value of a_1 (ns/s)	RMSE of $\widehat{a}_0$ (ns)	RMSE of $\widehat{a}_1$ (ns/s)
1	100	0.0103	0.264
1	1000	0.0100	0.311
10	100	0.0168	0.192
10	1000	0.0270	0.220
100	100	0.0138	0.268
100	1000	0.0133	0.306
1000	100	0.0185	0.212
1000	1000	0.0265	0.109



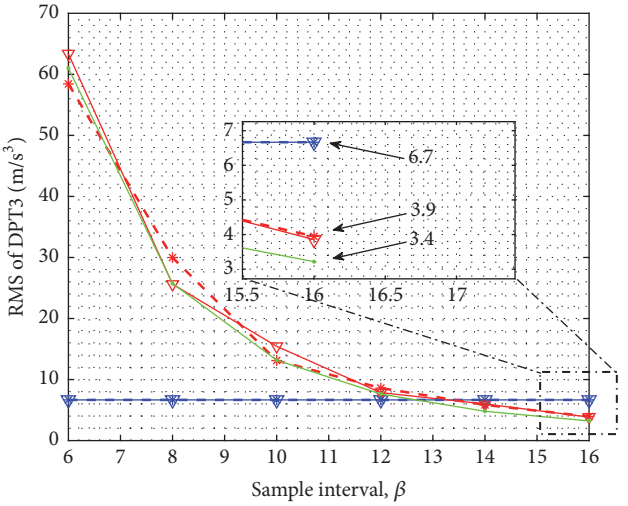
(a) The RMSE of distance estimation



(b) The RMSE of velocity estimation



(c) The RMSE of acceleration estimation



(d) The RMSE of jerk estimation

Figure 10: The comparison of the sparse algorithm and nonsparse algorithm versus β with $N_{seq} = 2048$.

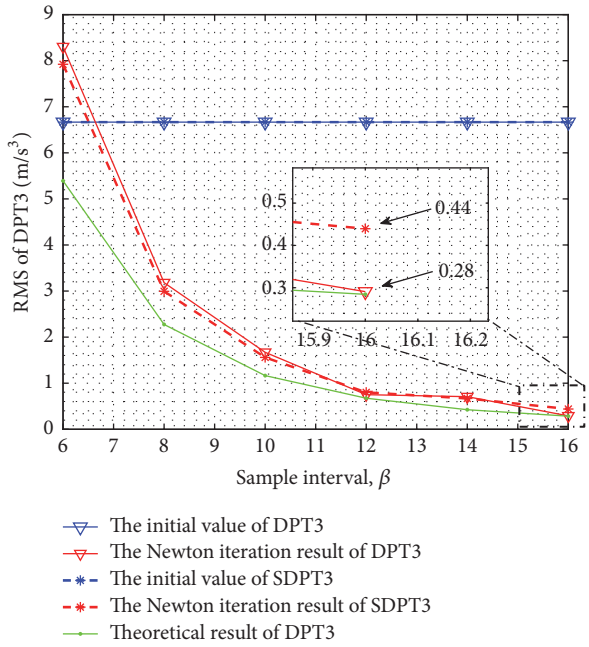


Figure 11: The jerk estimation versus β with $N_{seq} = 4096$.

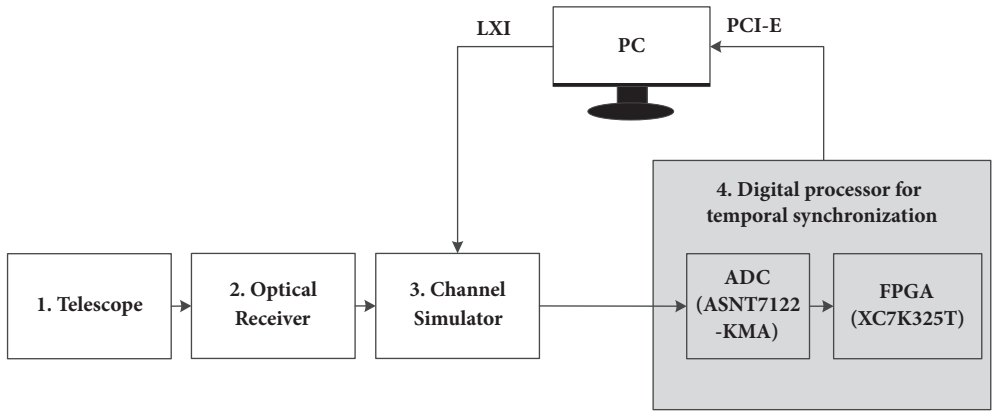


Figure 12: The topology of hardware platform.

ns in experimental environment. As the employed channel simulator cannot simulate the acceleration and jerk currently, the performance on estimating a_2 and a_3 will be verified in follow-up experiments.

8. Conclusions

In this paper, we have developed a temporal synchronization algorithm based on the sparse pilot and the sparse transform method in the scenario with high Doppler and high data rate. The performance of the proposed algorithm has been analyzed in comparison with the conventional DDLL algorithm and the nonsparse algorithm. The analytical and simulation results demonstrate that the proposed algorithm performs best in terms of a good accuracy and the lowest complexity. In addition, we have implemented the proposed algorithm, which confirms its implementability in resource constrained platforms, e.g., the IMDD-based laser communications for feeder links in 5G NTN platforms.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

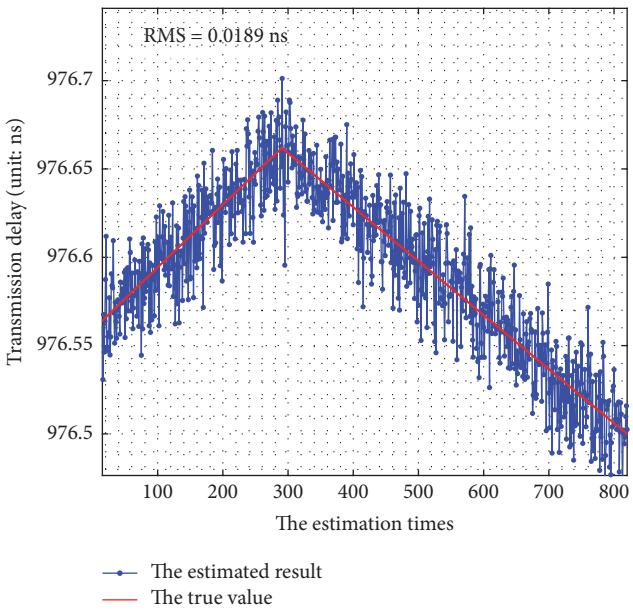


FIGURE 13: The experiment result of continuous estimation.

Acknowledgments

This work was funded by Research on Space-Oriented Terahertz Wideband Communication Theory and Technology [6161001093].

References

[1] B. Soret, A. De Domenico, S. Bazzi, N. H. Mahmood, and K. I. Pedersen, "Interference Coordination for 5G New Radio," *IEEE Wireless Communications Magazine*, pp. 1–7, 2017.

[2] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving Sustainable Ultra-Dense Heterogeneous Networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[3] D. B. Mckenna and B. J. Cox, "Doppler insensitive non-terrestrial digital cellular communications network," *Patent US6377802*, vol. 23, 2002.

[4] B. A. Lauer, K. Wang, T. Lamarca et al., "Systems and methods for facilitating communications originating from a non-terrestrial network," *Patent US9503956*, Nov 2016.

[5] B. A. Lauer, K. Wang, T. Lamarca et al., "Systems and methods for facilitating communications destined for a non-terrestrial network," *Patent WO/2015/184339*, Article ID 184339, Dec 2015.

[6] S. Poulenard, M. Crosnier, and A. Rissons, "Ground segment design for broadband geostationary satellite with optical feeder link," *IEEE/OSA Journal of Optical Communications & Networking*, vol. 7, no. 4, pp. 325–336, Apr 2015.

[7] B. Roy, S. Poulenard, S. Dimitrov et al., "Optical feeder links for high throughput satellites," in *Proceedings of the IEEE International Conference on Space Optical Systems and Applications (ICSOS '15)*, pp. 1–6, New Orleans, LA, USA, October 2015.

[8] N. Perlot, T. Dreischer, C. M. Weinert et al., "Optical GEO feeder link design," in *Proceedings of the FutureNetw*, pp. 1–8, Berlin, Germany, 2013.

[9] H. Someya, I. Oowada, H. Okumura, T. Kida, and A. Uchida, "Synchronization of bandwidth-enhanced chaos in semiconductor lasers with optical feedback and injection," *Optics Express*, vol. 17, no. 22, pp. 19536–19543, 2009.

[10] D. O. Caplan, "Laser communication transmitter and receiver design," *Journal of Optical Fiber Communications Reports*, vol. 4, no. 4-5, pp. 225–362, Sept 2007.

[11] J. Li, H. Ruan, and F. Liu, "The design and analysis of PD radar aid by GPS/INS ultra-tightly coupled navigation," in *Proceedings of the IET International Radar Conference 2013*, China, April 2013.

[12] Y. Li, Z. Tao, and Z. X. Niu, "Ultra-wide bandwidth Communication signal tracking Algorithm based on del ay-locked Loop," *Transactions of Beijing Institute of Technology*, vol. 26, no. 11, pp. 1019–1021, Nov 2006.

[13] R. Alhakim, K. Raoof, and E. Simeu, "Design of tracking loop with dirty templates for UWB communication systems," *Signal Image & Video Processing*, vol. 8, no. 3, pp. 461–477, Mar 2014.

[14] Z. Y. Zhan, J. S. Ding, W. Y. Wu, and J. G. Hou, "A new tracking method for high dynamic DS-SS signal," in *Proceedings of the IET International Radar Conference '15*, pp. 1–5, Hangzhou, China, 2015.

[15] C. Lin, B. Shao, and J. Zhang, "A high data rate parallel demodulator suited to FPGA implementation," in *Proceedings of the International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS '10)*, pp. 1–4, Chengdu, China, December 2010.

[16] D. Schmidt and B. Lankl, "Parallel architecture of an all digital timing recovery scheme for high speed receivers," in *Proceedings of the CSNDSP Newcastle upon Tyne*, pp. 31–34, UK, 2010.

[17] C. Lin, J. Zhang, and B. Shao, "A High Speed Parallel Timing Recovery Algorithm and Its FPGA Implementation," in *Proceedings of the 2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC '11)*, pp. 63–66, Wuhan, China, October 2011.

[18] W. C. Du, X. Q. Gao, and G. H. Wang, "Using FRFT to estimate target radial acceleration," in *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition '07*, pp. 442–447, Beijing, China, November 2007.

[19] R. Brcich and A. Zoubir, "The use of the DPT in passive acoustic aircraft flight parameter estimation," in *Proceedings of the IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications*, pp. 819–822, Brisbane, Qld, Australia, 1997.

[20] F. Lu, H. Wang, X. Liu et al., "Time-frequency characteristics of PSWF with Wigner-Ville Distributions," in *Proceedings of the ICSIP*, pp. 568–582, Beijing, China, 2016.

[21] F. Pignol, F. Colone, and T. Martelli, "Lagrange polynomial interpolation based Keystone Transform for passive radar," *IEEE Transactions on Aerospace & Electronic Systems*, vol. 99, 2017.

[22] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Simple and practical algorithm for sparse Fourier transform," in *in Proceedings of the Acm-Siam Symposium on Discrete Algorithms*, pp. 1183–1194, San Francisco, CA, USA.

[23] A. C. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, "Recent developments in the sparse fourier transform: a compressed fourier transform for big data," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 91–100, 2014.

[24] J. Schumacher, *High Performance Sparse Fast Fourier Transform*, Dept. Computer Science, ETH Zurich, Zurich, Switzerland, 2013.

[25] S. Liu, T. Shan, R. Tao et al., "Sparse discrete fractional fourier transform and its applications," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6582–6595, 2014.

[26] A. Chaaban and M. S. Alouini, "Optical intensity modulation direct detection versus heterodyne detection: A high-SNR capacity comparison," in *Proceedings of the 5th International Conference on Communications and Networking (COMNET '15)*, pp. 1–5, Tunis, Tunisia, November 2015.

[27] G. N. Liu, L. Zhang, T. Zuo, Q. Zhang, J. Zhou, and E. Zhou, "IM/DD Transmission Techniques for Emerging 5G Fronthaul, DCI and Metro Applications," in *Proceedings of the Optical Fiber Communication Conference*, pp. 1–3, Los Angeles, California, 2017.

[28] M. A. Khalighi and M. Uysal, "Survey on Free Space Optical Communication: A Communication Theory Perspective," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2231–2258, 2014.

[29] S. Arnon, J. Barry, G. Karagiannidis, R. Schober, and M. Uysal, *Advanced Optical Wireless Communication Systems*, Cambridge University Press, New York, NY, USA, 1st edition, 2012.

[30] M. Y. Hong, "A fast sinusoidal signal analysis technique for the determination of complex frequencies," in *Proceedings of the EEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), '02*, pp. 1469–1472, Orlando, FL, USA, 2002.

[31] T. J. Abatzoglou, "A Fast Maximum Likelihood Algorithm for Frequency Estimation of a Sinusoid Based on Newton's Method," *IEEE Transactions on Acoustics Speech & Signal Processing*, vol. 33, no. 1, pp. 77–89, Mar 1985.

[32] S. Peleg, B. Porat, and B. Friedlander, "The discrete polynomial transform (DPT), its properties and applications," in *Proceedings of the Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems & Computers*, pp. 116–120, Pacific Grove, CA, USA, 2002.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6187580, 26 pages
https://doi.org/10.1155/2018/6187580

Review Article

Uplink Nonorthogonal Multiple Access Technologies Toward 5G: A Survey

Neng Ye,[1,2] **Hangcheng Han**,[1] **Lu Zhao**,[1] **and Ai-hua Wang**[1]

[1]*School of Information and Electronics, Beijing Institute of Technology, Beijing, China*
[2]*China Electronics Technology Group Corporation (CETC), Key Laboratory of Aerospace Information Applications, China*

Correspondence should be addressed to Hangcheng Han; hanhangcheng@bit.edu.cn

Received 26 January 2018; Accepted 14 May 2018; Published 12 June 2018

Academic Editor: Giovanni Stea

Owing to the superior performance in spectral efficiency, connectivity, and flexibility, nonorthogonal multiple access (NOMA) is recognized as the promising access protocol and is now undergoing the standardization process in 5G. Specifically, dozens of NOMA schemes have been proposed and discussed as the candidate multiple access technologies for the future radio access networks. This paper aims to make a comprehensive overview about the promising NOMA schemes. First of all, we analyze the state-of-the-art NOMA schemes by comparing the operations applied at the transmitter. Typical multiuser detection algorithms corresponding to these NOMA schemes are then introduced. Next, we focus on grant-free NOMA, which incorporates the NOMA techniques with uplink uncoordinated access and is expected to address the massive connectivity requirement of 5G. We present the motivation of applying grant-free NOMA, as well as the typical grant-free NOMA schemes and the detection techniques. In addition, this paper discusses the implementation issues of NOMA for practical deployment. Finally, we envision the future research challenges deduced from the recently proposed NOMA technologies.

1. Introduction

In the past several decades, the wireless communication system has evolved from the first generation, an analog communication network which only transfers voice messages, to LTE networks, which satisfies the great demands on mobile broadband data transmissions. Recently, the development of 5G has raised new challenges with respect to peak data rate, user experience data rate, spectral efficiency (SE), energy efficiency (EE), massive connectivity, low latency, and ultra-reliability, etc.

Nonorthogonal multiple access (NOMA) technologies have been recognized by both industry and academia as one promising tendency and progress, ever since the deployment of orthogonal frequency-division multiple access (OFDMA) in LTE, to meet the wide-ranging requirements for 5G and beyond under the strict constraint of the limited radio resources [1]. The idea of NOMA can trace back to the information-theoretic researches about multiuser information theory [2]. In downlink broadcast channel (BC), superposition coding and successive interference cancelation (SIC) receiving are employed to approach the entire capacity region of BC. Meanwhile, in uplink multiple access channel (MAC), the signals of different transmitters are overlapped and SIC receiver is applied to achieve the corner points of MAC capacity region. In 1990s, multiple access protocols, which exploited the differences between the power levels of the received packets, were proposed and studied by Shimamoto (1992) [3], Pedersen (1996) [4], and Mazzini (1998) [5], respectively. In 2008, Y. Yan and A. Li in [6] proposed a superimposed radio resource sharing (SRRS) scheme which utilizes the near-far effect to enhance the uplink throughput performance. SRRS superimposes different uplink data streams on the same radio resources and applies SIC at the receiver, which can be regarded as a prototype of NOMA,

Despite all the related researches, NOMA is still not commercialized in the past decades due to the concern of high computational complexity of SIC-type receiver. However, the rapid growth of processing power of the microprocessors in these years has provided an opportunity to the standardization and commercialization of NOMA technologies. Recently, downlink nonorthogonal transmissions, featured

Table 1: Summery of existing surveys about NOMA.

Survey	Scope	Contributions
[13]	Power-domain NOMA	A comprehensive survey about power domain-NOMA, as well as the related designs.
[14]	NOMA schemes	Review of power-domain and code-domain NOMA schemes
[15]	NOMA schemes and waveforms	Review of some NOMA schemes and nonorthogonal waveforms.
[16]	NOMA schemes	Review of some NOMA schemes towards 5G, as well as the application scenarios and typical receivers.
[17]	Theoretical analysis of NOMA	Review of the theoretical analysis about power-domain NOMA and cognitive radio inspired NOMA.
[18]	Downlink NOMA	Industrial view about downlink NOMA in 5G.

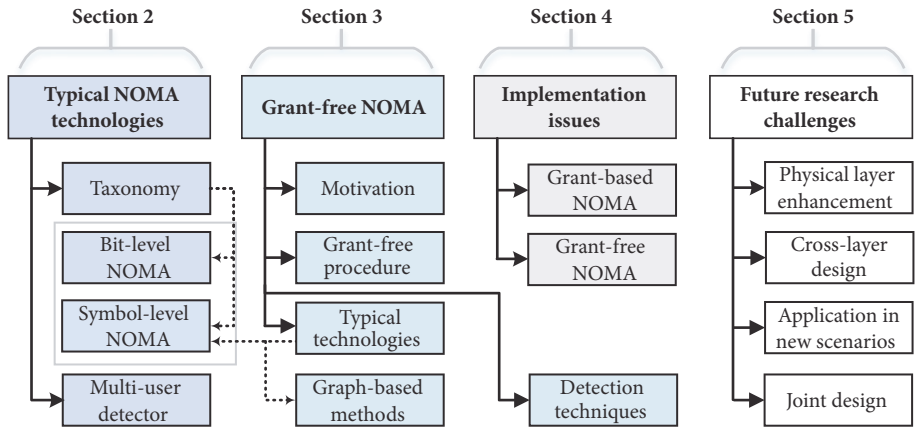


Figure 1: The outline of this paper.

by multiuser sharing technology (MUST), are specified in LTE Release-14 in 2017. Toward the evolution of 5G, the industrial community has proposed dozens of NOMA schemes as candidate multiple access technologies. In the meantime, a study item of NOMA, actuated by its potential advantages, has been approved by 3GPP RAN plenary [7] in March 2017, which promotes the standardization of NOMA in 5G.

The core idea of NOMA is to multiplex different data streams over the same radio resources and employ multiuser detection algorithm at the receiver to recover multiple users' signal streams. The major design target of NOMA is to introduce controllable mutual interference among users to achieve a fine tradeoff between multiplexing gain and detection reliability. According to both theoretical and numerical analysis, NOMA outperforms OMA with respect to SE, EE, and connectivity [8–11]. To grasp the development of NOMA technologies, some published review papers have presented different aspects of NOMA. We summarize the main contributions of these articles in Table 1.

Different from the existing literature, this paper presents a comprehensive review about the recent progress of NOMA proposed in the standardization process of 3GPP toward 5G, including candidate NOMA schemes and multiuser receiving technologies. Meanwhile, we also survey the state-of-the-art grant-free NOMA schemes, which are expected to satisfy the massive connectivity and high EE requirements in massive machine-type communication (mMTC) scenario. Additionally, we discuss the implementation issues about NOMA. The contributions of this survey are summarized in the following four aspects:

(i) It is a comprehensive survey about the candidate NOMA schemes proposed in 3GPP, as well as the promising multiuser detection methods. NOMA schemes are categorized into bit-level and symbol-level schemes for illustrations, according to the agreements in 3GPP [12].

(ii) The motivation and main idea of grant-free NOMA are presented in this survey. In addition to the grant-free procedures, this survey also introduces the typical grant-free NOMA schemes, as well as the detection algorithms.

(iii) The implementation issues about NOMA, especially grant-free NOMA, are discussed, with respect to resource allocation, procedures, and physical layer signals.

(iv) The future research challenges related to NOMA are identified, including physical layer enhancement, cross layer design, applications of NOMA in new scenarios, and the joint design of NOMA with other technologies.

Figure 1 illustrates the broad outline of this review. The rest of this review is organized as follows. Section 2 introduces the typical transmission and reception technologies of

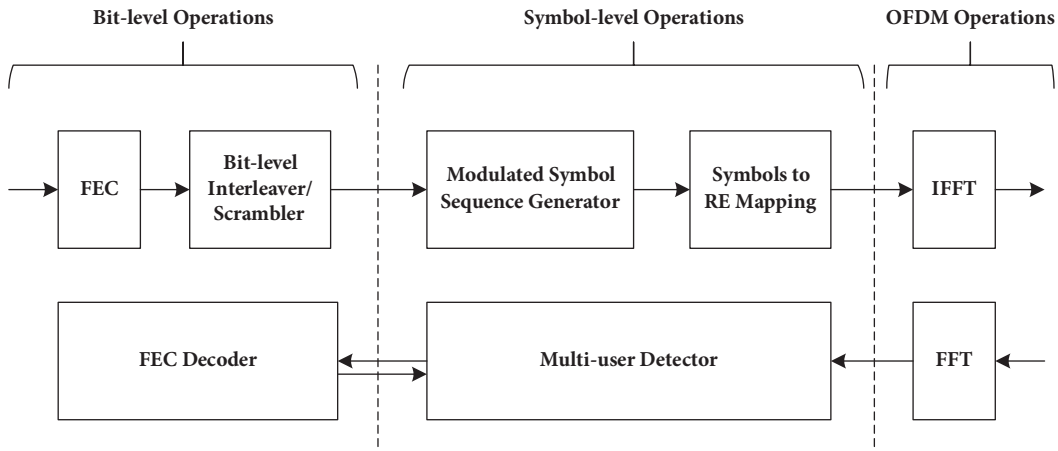


Figure 2: A unified structure of NOMA technologies.

NOMA. Section 3 analyzes the grant-free NOMA, including the motivations, procedures, and typical transceiving schemes. In Section 4, we discuss the implementation issues of NOMA toward 5G. The future research challenges about NOMA are highlighted in Section 5, and Section 6 concludes this paper.

2. Typical NOMA Technologies

The industrial community has proposed plenty of NOMA schemes to meet the diversified requirements toward 5G. Until 3GPP TSG-RAN WG1 (RAN-1) #86, at least 15 candidate NOMA schemes have been proposed for 5G new radio (NR) [19–32]. On RAN-1 #86b, a general framework of NOMA schemes is agreed upon [12], which helps to categorize existing operations in NOMA schemes into bit-level operations and symbol-level operations, as shown in Figure 2. Note that several component operations may be simultaneously adopted in the future 3GPP Release-15 to satisfy the requirements of different application scenarios.

Correspondingly, the proposed NOMA schemes can also be categorized into two classes, namely, bit-level NOMA and symbol-level NOMA, where the bit-level NOMA focuses on the design related to channel coding and bit-level interleaving, while the symbol-level NOMA mainly lays emphasis on symbol spreading and mapping. According to the above classifications, we summarize the state-of-the-art NOMA schemes in Table 2 and then give a comprehensive analysis. Meanwhile, the detection algorithms designed for these NOMA schemes are analyzed afterwards.

2.1. Bit-Level NOMA. Bit-level NOMA schemes exploit the low-rate forward error-correction (FEC) codes to enhance the detection accuracy, and/or take the advantage of user-specific interleaving to whiten the multiuser interference (MUI). In the following, we analyze several typical bit-level NOMA schemes including power-domain NOMA (PD-NOMA), low coding rate spreading (LCRS), low code rate and signature based shared access (LSSA), interleave-division multiple access (IDMA), and interleave-grid multiple access (IGMA).

2.1.1. PD-NOMA. PD-NOMA [27] multiplexes the users in power domain and applies the iteration-based SIC receiver to detect multiple signal streams at the receiver [33, 34]. In each iteration of SIC receiving, the MUI is regarded as thermal noise, which suggests that the user demultiplexing could be implemented by generating a large power difference among the multiplexed users. According to the simulation results, PD-NOMA can improve the resource utilization efficiency in both uplink and downlink [34]. Meanwhile, PD-NOMA can maintain low peak to average power ratio (PAPR) if single-carrier property is kept [35]. In addition, PD-NOMA does not depend on the information of instantaneous channel state information (CSI) of frequency-selective fading. Therefore, no matter the user mobility or CSI feedback latency, a robust performance gain in practical wide area deployments can be expected.

The major design aspect related to PD-NOMA is the resource allocation, including user association, radio resources assignment, and power allocation [36]. However, solving the resource allocation problem in one shot would be nontrivial. Therefore, this problem is usually decoupled into two subproblems, i.e., user scheduling and power allocation, respectively. In PD-NOMA, the users with large channel gain difference (e.g., large path-loss difference) are normally paired to enhance SE performance [37]. However, this simple criterion may cause unfairness in system-level deployment. Proportional fairness (PF) based scheduling [38], which simultaneously optimizes the user fairness and system throughput, is a practical user scheduling technology for PD-NOMA. The PF metric, calculated by dividing the instantaneous signal to interference and noise ratio (SINR) with the average data rate over the past period, is maximized during the user scheduling stage [39]. In uplink, user scheduling should consider the single-carrier frequency-division multiple access (SC-FDMA) where the subcarriers are distributed continuously to overcome the PAPR problem. One low complexity heuristic method based on greedy sub-band widening is proposed in [40] for practical deployment.

In the meantime, there have been abundant literature sources which address the power allocation problem of PD-NOMA [13]. Due to the nonconvexity of the power allocation

Table 2: Summarization of NOMA schemes toward 5G standardization.

NOMA scheme	Key technical point			Main advantage
IDMA	Low coding rate	Low rate FEC code	Bit-level Interleaving	Randomized the mutual interference
IGMA		Low rate FEC code or moderate one with repetition	Bit-level Interleaving (permutation matrix)	Sparse grid Mapping
LSSA		Low rate FEC code or moderate one with repetition	User-specific bit-level interleaving/permutation pattern	Large number of signatures
LCRS		Low rate FEC code and repetition	Bit-level spreading	Large coding gain
SCMA	Short low density spreading	Multidimensional modulation		Signal space diversity gain
PDMA		Irregular LDS		Irregular protection
LDS-SVE		LDS & User signature vector extension (SVE)		Higher diversity
MUSA	Short dense spreading (low cross-correlation sequence)	Short complex spreading sequence		Easy to generate & Large number
NCMA		NCC obtained by Grassmannian line packing problem		Optimal nonorthogonal sequence
NOCA		Zadoff-Chu sequence		Easy to generate, low PAPR
SSMA		Orthogonal or quasi-orthogonal codes		
GOCA	Long spreading/scrambling sequence	Group-based orthogonal/nonorthogonal sequences		Inter-group orthogonality
RDMA		Cyclic shift based time-frequency repetition		Easy implementation
RSMA		Low cross-correlation Sequence scrambling		Fit for asynchronized scenario
RSMA(single tone)	Single carrier (similar to CDMA), low PAPR modulation			Extended coverage and low PAPR for uplink

problem, advanced optimization techniques are usually employed to optimize the system throughput, reliability, and/or connectivity. In [41], the maximization of PF metric is presented. At first, the optimal power allocation of MAC is calculated iteratively. Then the optimization results can be converted to BC based on uplink-downlink duality. Several water-filling based methods are summarized and further studied in [42], where a weighted water-filling method is proposed in presence of user priority. In [43], an iterative suboptimal power allocation algorithm based on difference of convex (DC) programming is presented. The readers may refer to [13] for an extensive review about resource allocation algorithms in PD-NOMA.

Nevertheless, the existing methods may still be complex for system-level deployment. Hence, several power allocation methods are proposed by industrial community to enable efficient and practical applications. When the users have been paired into groups, one option is to apply the predefined power allocation ratios to different users as done in [44]. An alternative method is to choose one option that can maximize the PF metric out of several options; e.g., for two-user NOMA, the options of the power ratio can be $[0.2, 0.8]$ and $[0.3, 0.7]$, which is also termed fixed transmission power allocation (TPA). Since the indexes of TPA can be predefined, TPA can effectively decrease the amount of downlink signaling related to PD-NOMA. Another commonly used method is the fractional transmit power allocation (FTPA) inspired from the transmission power control used in the LTE uplink [39]. In FTPA, the users with poorer channel conditions are allocated with more power to partially compensate the channel loss. In the above FTPA, the related parameters can be optimized via system-level simulation. After the resource allocation stage, a sophisticated design of the constellations, e.g., constellation rotation [45], may provide additional gain in enhancing the detection accuracy.

When multicell or dense-network scenario is considered [46], the uplink PD-NOMA would increase the intercell interference (ICI) because multiple users are allowed to transmit on the shared carriers. Therefore, user association and ICI-aware power allocation should also be studied to control the transmission power and avoid causing severe ICI to the neighboring cells [47, 47].

2.1.2. IDMA. In addition to the power-domain multiplexing, the users can also be distinguished if they have different interleaving patterns, which is exploited in interleave-division multiple access (IDMA). IDMA is initially proposed by P. Li et al. [48] to enhance the performance of asynchronous code division multiple access (CDMA). It has the benefits of preventing the effect of fading and mitigating the ICI as in CDMA [48]. Researchers in [49] expound that IDMA can exhibit some other attractive characteristics such as flexible

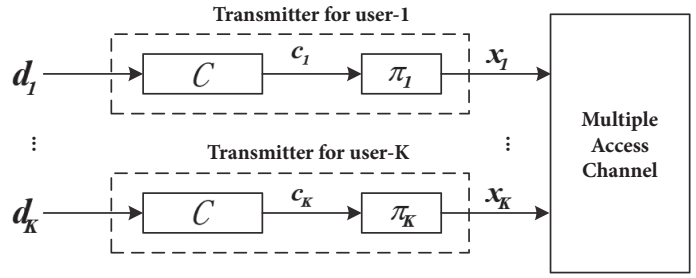


Figure 3: The transmission structure of IDMA with K multiplexed users.

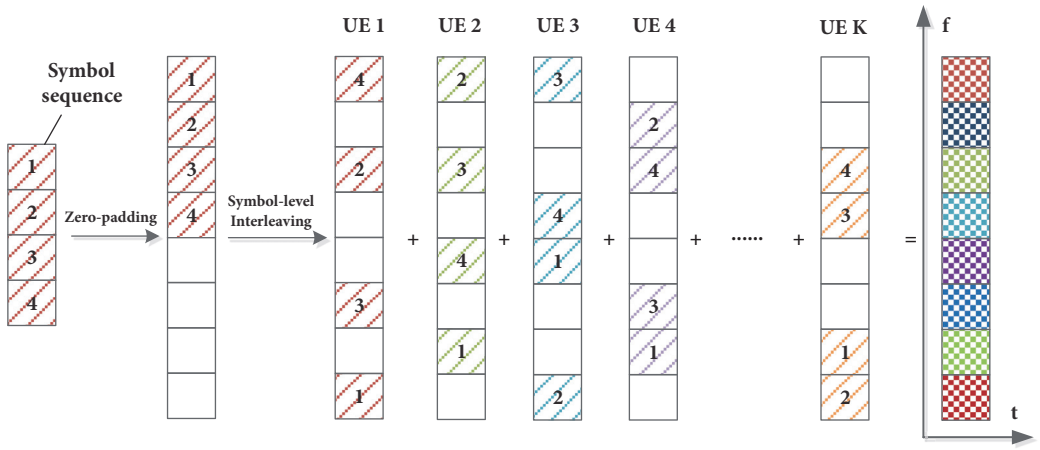


Figure 4: An illustration of the grid mapping procedure of IGMA.

rate adaptation, frequency diversity, and power efficiency. Besides, the theoretical study of IDMA also shows that the interleaved low-rate codes with a simple chip-by-chip iterative decoding strategy could achieve the capacity of a Gaussian MAC [50].

The transmission structure of IDMA is illustrated in Figure 3. The low-rate FEC encoder C is applied to encode the user-k's data bits $\mathbf{d}_k$. The output is referred as coded bits $\mathbf{c}_k$. The coded bits $\mathbf{c}_k$ pass through the interleaver π_k, after which multiple users' signals are multiplexed in the air. The interleaving patterns are generated independently and randomly and vary from each other in order to distinguish the users. Therefore, the design of reasonable interleavers is rather essential. A user-specific interleaver design method is proposed in [51], which can resolve the memory cost problem and reduce the signaling exchanging between the gNB and the users. Besides, to accommodate IDMA in multicarrier transmission, e.g., in OFDM, a multicarrier interleave-division-multiplexing-aided IDMA (MC-IDM-IDMA) is presented in [52].

IDMA has been wildly studied because of its robustness and user overload tolerance [19]. The structure of IDMA in single-path and multipath environments is elaborated in [53]. Besides, a power allocation method is introduced to enhance the performance of IDMA by taking the advantage of the semianalytical technique [48].

2.1.3. IGMA. Interleave-grid multiple access (IGMA) goes one step further than IDMA by introducing the grid mapping patterns [20], which can cooperate with the interleaving patterns to distinguish the signal streams from different users.

The flexibility to choose bit-level interleavers and/or grid mapping pattern for distinguishing the users could be easily supported in IGMA. Meanwhile, the scalability supporting different connection densities would be achieved with the abundant signatures generated by bit-level interleavers and grid mapping patterns.

Hereinafter, we briefly explain the general procedure of IGMA. Firstly, the user's data bits are encoded by the channel encoder to generate the coded bit sequence. The sequence is then interleaved to randomize the order of coded bits based on a preconfigured interleaver. The interleaved bit sequence is then modulated into the symbol sequence. Finally, the grid mapping process is conducted to interleave the symbol sequence as shown in Figure 4. The whole procedure of IGMA can further help in combating frequency selectivity and ICI due to the randomization.

2.1.4. LSSA. The low code rate and signature based shared access (LSSA) scheme is proposed to support asynchronous massive transmission in uplink [23]. LSSA randomizes the MUI among the users by multiplexing the users' data streams with user-specific signature patterns at bit-level, where the signature patterns are usually unknown to others. Besides,

Table 3: Distinction between long and short sequences.

	Long sequence	Short sequence
Level of operation	Bit level/symbol level	Usually symbol level
Generation of sequence	Randomly	Carefully designed
Usage	Disperses the encoded bit sequences so that the adjacent bits are approximately uncorrelated	To facilitate MUD
Receiving technique	Requires iterative detection between symbol-level and bit-level, e.g., ESE-SIC	Symbol level detection, e.g., MPA/EPA
Synchronization requirement	Supports asynchronous transmission when combined with single carrier waveform, e.g., RSMA	Synchronization is usually required, e.g. SCMA
Blind detection	Does not support	Support

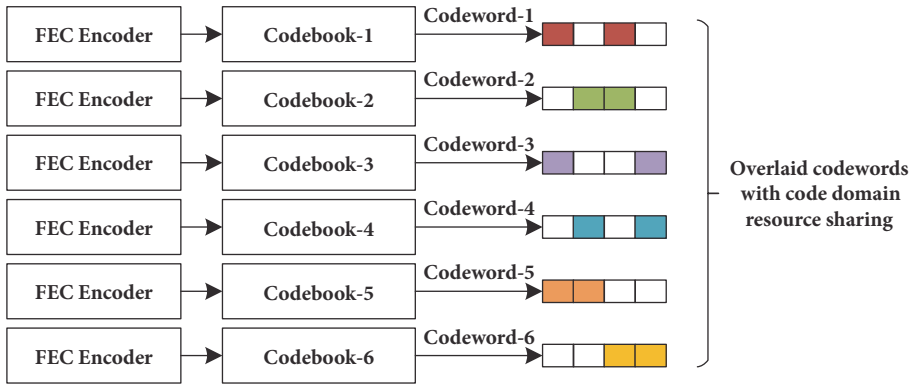


Figure 5: A typical transmission structure of SCMA with six users and four subcarriers.

the channel coding scheme which has very low code rate is adopted to encode each user's information bits in LSSA, which helps to mitigate the effect of the MUI. The low-rate FEC code can also be replaced by employing higher rate FEC code along with spreading. After the channel coding, bit-level multiplexing with user-specific signature would be used. The user-specific signature may relate to the reference signal, the complex/binary sequence, and the permutation pattern of a short length vector. The length of orthogonal spreading codes is a factor that influences the number of simultaneous transmissions. Fortunately, the receiver in LSSA does not depend on orthogonal multiplexing codes to distinguish the target users' signals. Instead, the interference cancelation is exploited, so that high user overloading is well supported. The signature of LSSA can be chosen randomly at the user side or assigned to the user by the gNB. Furthermore, LSSA can also be optionally modified to have a multicarrier variant in order to exploit frequency diversity provided by wider bandwidth and to achieve lower latency.

2.1.5. LCRS. Low code rate spreading (LCRS) is another NOMA scheme which utilizes the bit-level repetition and low-rate coding to spread information bits over the total nonorthogonal transmission area [21]. Therefore, LCRS can achieve the maximum coding gain by combining channel coding and spreading through low-rate codes. Under this circumstance, a user-specific channel interleaver [48] can be further exploited to aid the multiuser signal separation at the receiver.

2.2. Symbol-Level NOMA. Different from bit-level NOMA schemes which focus on the *bits*, symbol-level NOMA schemes play with *symbols* and mainly lay emphasis on the bit-to-symbol mapping. As illustrated in Table 2, a large portion of symbol-level NOMA schemes utilize the short sequence-based spreading to enhance the connectivity. These schemes can be further divided into two subcategories according to the densities of the spreading sequences. Some other symbol-level NOMA schemes make use of long sequence-based scrambling/spreading/permutation, where the receiver exploits the difference between these sequences. Table 3 compares the pros and cons of applying long or short sequences in symbol-level NOMA, which are further illustrated in the following subsections.

Short Sparse Spreading NOMA

(1) SCMA. Sparse code multiple access (SCMA) is a low density spreading-based NOMA scheme, which can achieve high overloading while maintaining high reliability [32, 54, 55]. The core idea of SCMA is to directly map the coded bits to the multidimensional modulation symbols, according to a predefined sparse codebook, instead of sequentially conducting modulation and low density spreading. Therefore, both the resource element mapping and the multidimensional constellation are essential designs in SCMA [56]. The transmission process of SCMA is illustrated in Figure 5, where multiple signal layers are multiplexed on the same radio resources. One major design aspect of SCMA is the sparse

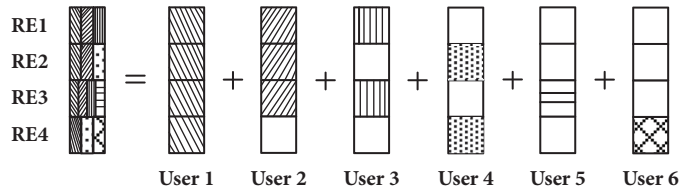


Figure 6: Resource mapping of PDMA, with six users and four subcarriers.

multidimensional codebook [57]. In [57], a SCMA codebook design method based on rotation, shuffling, and permutation is proposed, along with an example SCMA codebook which brings large shaping gain. SCMA can also achieve the signal space diversity by permuting the signal components to paired symbols located in multiple radio resources. Besides, with the sparse structure of SCMA, iterative multiuser detection algorithms, e.g., message passing algorithm (MPA), can be applied to simultaneously detect multiple data streams in symbol-level.

However, one concern of SCMA is that the sparse structure may be violated when single carrier is performed [23]. And MPA receiver may cause large computational burden and processing delay when the number of multiplexed users is large. Hence, a good tradeoff ought to be achieved between complexity and performance in the design of SCMA.

(2) PDMA. Inspired by unequal transmission diversity and sparse coding, pattern division multiple access (PDMA) is proposed as a novel NOMA scheme to enhance the performance of multiuser communication system [28]. Different from SCMA which utilizes regular spreading signatures, PDMA usually employs irregular sparse signatures to facilitate the SIC receiving [58]. Besides, with the irregular sparse spreading signatures, PDMA can have a total number of 2^N signatures where N is the length of spreading.

An example of the code domain pattern matrix of PDMA is shown in (1), which involves six users and four subcarriers. A "1" means that the subcarrier is occupied by a user. According to the spreading patterns, the signals of the six users are illustrated in Figure 6. However, we also see that one drawback of PDMA is that it cannot guarantee to accommodate the strict sparsity constraints as in SCMA; i.e., four users multiplex on the 3rd RE.

$$G_{\text{PDMA}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

To further enhance the ability of distinguishing the multiplexed users, PDMA allows utilizing multiple domains in the design of the signature matrix, including temporal, spatial, code, power, and interleave domains [59, 60]. For example, PDMA with large-scale antenna array (LSA-PDMA) is proposed where the spreading signatures are designed jointly in beam and power domain to improve the system sum rate and access connectivity, respectively [61]. In addition, an interleaver-based PDMA (IPDMA) scheme is proposed, where the signal separation can be done according to different bit-level interleavers and/or characteristic patterns [62]. With the joint design at the transmitter and the receiver, PDMA can meet the need of higher spectral efficiency in 5G, while ensuring a reasonable receiver complexity.

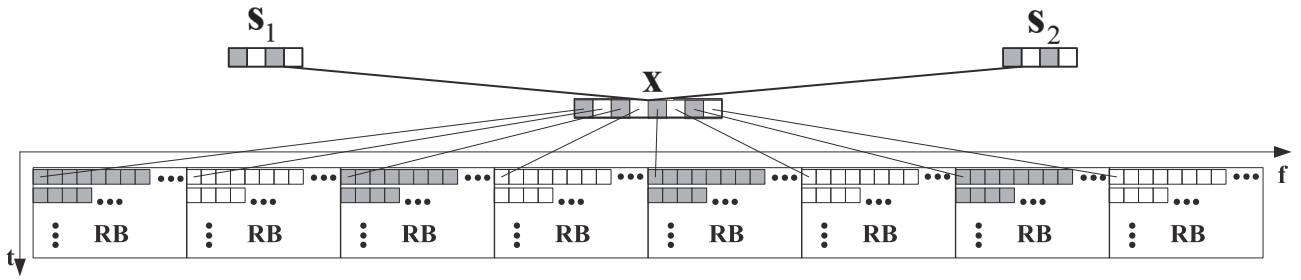
(3) LDS-SVE. Low density signature-signature vector extension (LDS-SVE) is another LDS-based NOMA scheme [22]. The major difference between LDS-SVE and the other LDS-based NOMAs, i.e., SCMA and PDMA, is that the former introduces user-specific signature vector extension, which is performed by transforming and concatenating two element signature vectors into a larger signature vector.

In the following, we show an example of LDS-SVE in Figure 7. The modulated symbols are first divided into two vectors, i.e., $\mathbf{s}_i$, $i = 1, 2$, according to a serial-to-parallel transformation. Define $\mathbf{s}_R$ as a real vector obtained by stacking the real and imaginary parts of signature $\mathbf{s}_1$ and $\mathbf{s}_2$, as shown in Figure 7. The SVE output is defined as a real vector $\mathbf{x}_R$, which is achieved by multiplying $\mathbf{s}_R$ with a transformation matrix $\mathbf{U}$. At last, transmission complex signal $\mathbf{x}$ can be recovered from $\mathbf{x}_R$, i.e., by reconstructing the complex symbols from the real vector $\mathbf{x}_R$. The main advantage of LDS-SVE is that, by multiplying $\mathbf{U}$, the original modulation symbols are spread on more REs, which brings higher order of diversity.

Short Dense Spreading NOMA

(4) MUSA. Multiuser sharing access (MUSA) is a NOMA scheme based on short complex spreading sequence and SIC receiver [24]. In general, the spreading sequences in MUSA do not have sparsity as in SCMA, PDMA, and LDS-SVE [15]. We illustrate the transmission procedure of MUSA in Figure 8. After channel encoding and modulation, as shown in Figure 8, each user's data symbols are spread by a complex sequence, whose elements take values in complex field. Then the spread symbols of each user are transmitted on the shared radio resources. At the receiver, the well-designed spreading sequences are exploited by the multiuser detectors to distinguish different users' data streams. It is worth mentioning that different symbols of the same user may use different spreading sequences, which can average the MUI and improve the system-level performance.

Short sequence-based spreading is the major operation in the MUSA transmitter. Each user can randomly pick one spreading sequence from a sequence pool consisting of multiple spreading sequences. The spreading sequence design



$$\mathbf{s}_R = \left[\text{real}(\mathbf{s}_1) \quad \text{real}(\mathbf{s}_2) \quad \text{imag}(\mathbf{s}_1) \quad \text{imag}(\mathbf{s}_2)\right]^T$$
$$\mathbf{x}_R = \left[\text{real}(\mathbf{x}) \quad \text{imag}(\mathbf{x})\right]^T \Bigg\} \quad \mathbf{x}_R = \mathbf{U}\mathbf{s}_R$$

Figure 7: An illustration of LDS-SVE.

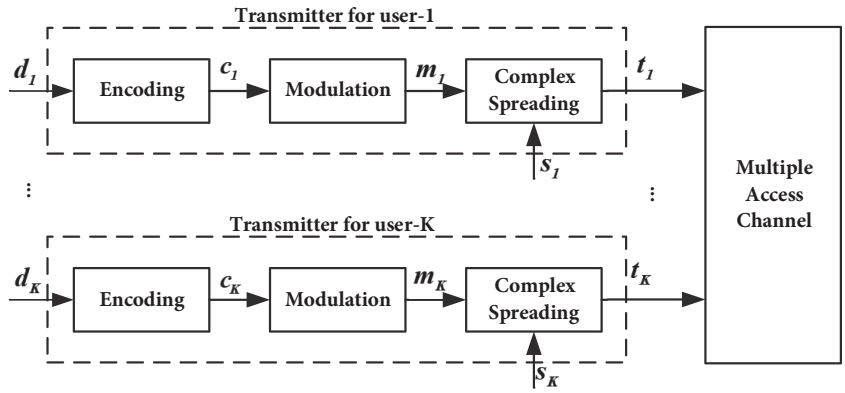


Figure 8: The transmission structure of MUSA with K simultaneous users.

in MUSA follows the guidelines of low cross-correlation, where each element of the sequence is chosen out of a complex scalar set, e.g., $\{\pm 1, 0, \pm i\}$. Due to the utilization of the imaginary part, complex spreading sequences could perform the lower cross-correlation compared to pseudorandom noise (PN), even with a short-spreading length [63]. In addition, with arbitrarily selected complex elements, the pool of the spreading sequences in MUSA can be very large.

(5) NCMA. Similar to MUSA, nonorthogonal coded multiple access (NCMA) also uses nonorthogonal dense spreading sequence to minimize MUI and support high overloading capability [25]. The spreading sequences of NCMA, also named as nonorthogonal cover codes (NCC), are obtained by solving the Grassmannian line packing problem, where the solutions of the problem guarantee the optima nonorthogonal sequences [64]. Due to the design of NCC, the interference level between two users is predictable.

The transmission structure of NCMA is illustrated in Figure 9. In NCMA, each user's data symbol is spread with NCC, and an additional FFT operation can be implemented before IFFT to reduce the PAPR. At the receiver, a simple despreading and parallel interference cancelation (PIC) detector can be implemented to recover the multiplexed signals. We note that applying IFFT on spare spreading-based NOMA schemes, e.g., SCMA and PDMA, would destroy the sparse structure and lead to high computational complexity in detection. To further improve the connectivity and bring additional throughput gain under specific QoS constraints, multistage spreading based on NCC can be applied. However, the correlation properties of the multistage spreading sequences, which are composed by multiplying several NCCs, need to be clarified. Hence, a good tradeoff between the connection density and the decoding performance needs to be further evaluated [23].

(6) NOCA. Nonorthogonal coded access (NOCA) is another spreading-based NOMA scheme. Similar to other symbol-level NOMA schemes, the data symbols in NOCA are spread according to nonorthogonal sequences before transmission [26]. The spreading in NOCA is operated in both time and frequency domain. We demonstrate the transmission structure of NOCA in Figure 10. The serial modulated symbol sequence is first converted to P parallel subsequences by a

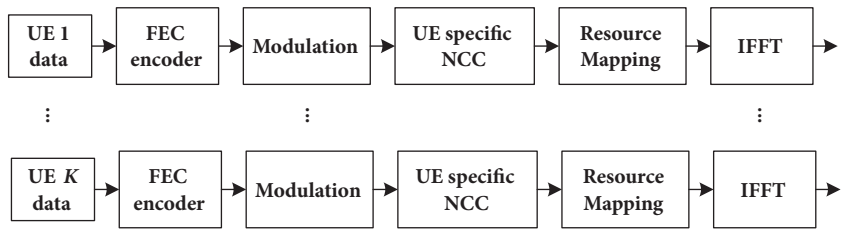


Figure 9: The transmission structure of NCMA with K simultaneous users.

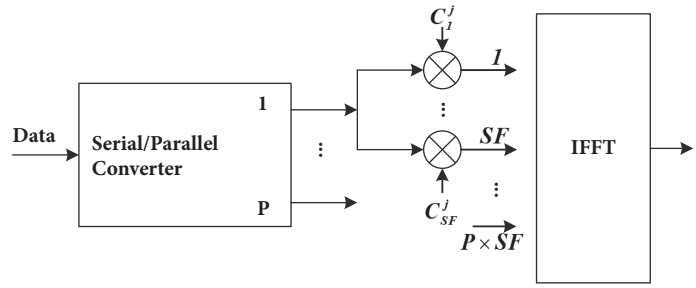


Figure 10: The transmission structure of NOCA.

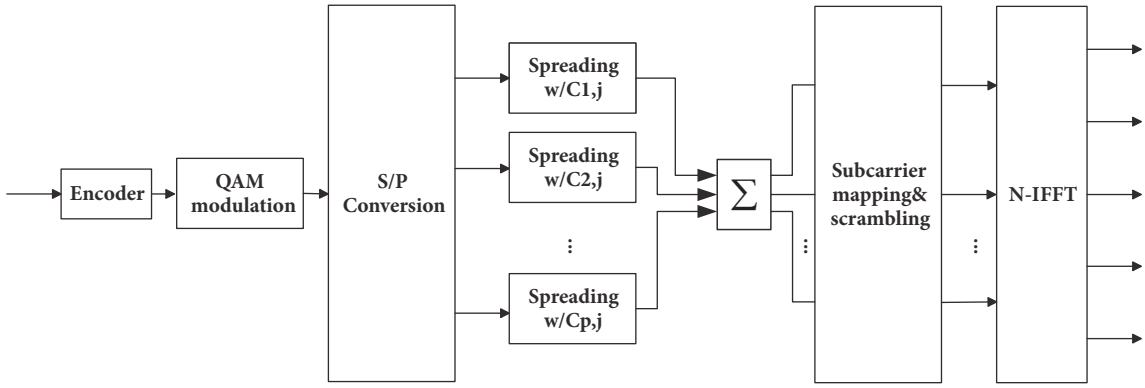


Figure 11: The transmission structures of SSMA.

S/P converter. We denote C_j as the nonorthogonal spreading sequence with length SF, where SF denotes the spreading factor. The jth subsequence is then spread on SF subcarriers according to C_j. Hence, a total number of $P \times SF$ subcarriers are required for NOCA. Besides, to accommodate the single-carrier transmission in uplink, FFT operation can also be applied before IFFT to reduce the PAPR.

To ensure high detection accuracy and high overloading, the spreading sequences used in NOCA should follow some properties, such as good autocorrelation, low cross-correlation, and low storage requirement. Meanwhile, the sequences should have constant modulus to ensure low cubic metric. Besides, multiple spreading factors might be supported for flexible adaptation.

(7) SSMA. Short sequence spreading-based multiple access (SSMA) is another spreading-based NOMA scheme [21], which directly spreads the modulation symbols with multiple orthogonal or quasi-orthogonal codes and transmits the spread symbols in time-frequency resources allocated for nonorthogonal transmission. The transmission structure of SSMA, as illustrated in Figure 11, is similar to NOCA, where user-specific scrambling is applied to average the MUI.

Long Sequence-Based NOMA

(8) RSMA. Resource spread multiple access (RSMA) is a novel NOMA scheme which applies long spreading or scrambling sequence to disperse the users signal over the entire radio resources. In RSMA, each user's codewords can be spread over all available time and frequency resources [24]. Therefore, RSMA can achieve full diversity compared to short-spreading-based NOMA schemes. At the receiver, different spreading/scrambling sequences can be exploited to distinguish different signal streams. Besides, low-rate FEC codes and advanced detection algorithms in RSMA can ensure high transmission reliability. The scrambler can also

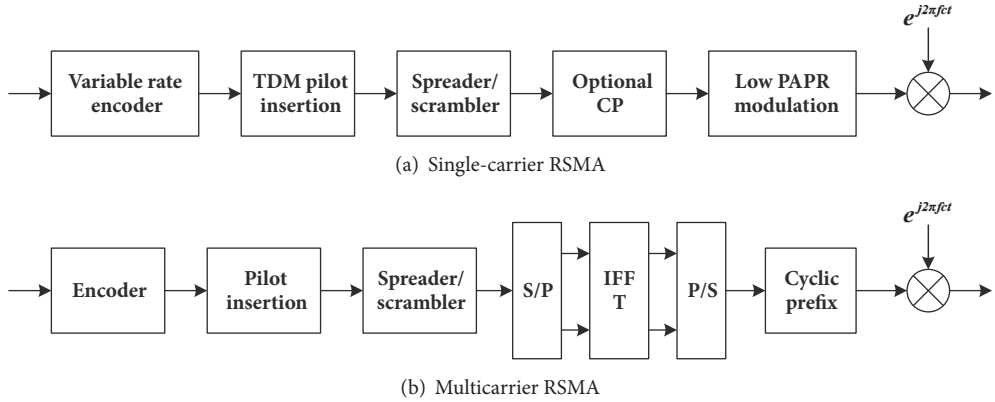


Figure 12: The transmission structures of single-carrier and multicarrier RSMAs.

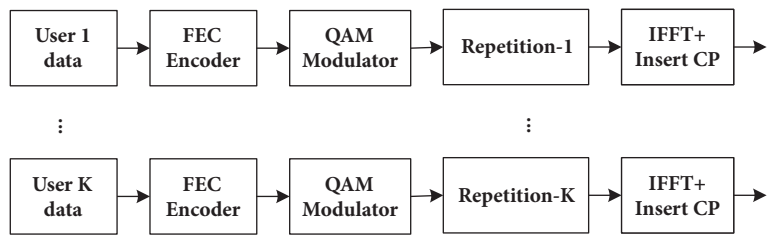


Figure 13: The transmission structure of RDMA with K simultaneous users.

be replaced by different interleavers for the sake of whitening the MUI.

According to different application scenarios, two kinds of RSMA schemes have been proposed, i.e., single-carrier RSMA and multicarrier RSMA [30], as shown in Figure 12. On the one hand, single-carrier RSMA employs the single-carrier waveforms and low PAPR modulations to enhance the performance of battery power consumption and coverage extension for small data transmission. Match-filter (MF) based receiver can be applied to distinguish different signals of single-carrier RSMA with low computational complexity. In addition, single-carrier RSMA does not rely on joint detection, which loses the synchronization requirement and makes it a good candidate for asynchronous access. On the other hand, multicarrier RSMA is studied to lower the latency and to promote the spectral efficiency for legacy users.

(9) RDMA. Repetition division multiple access (RDMA) can be regarded as an interleave-based NOMA scheme [29]. However, instead of deploying bit-level interleaving as in IDMA, RDMA focuses on the symbol-level interleaving which is designed based on simple cyclic-shift repetitions. In RDMA, each user's modulation symbol vector is repeatedly transmitted for several times, where different cyclic-shift indexes are assigned to the repetitions. Besides, different users would have different repetition and cyclic-shift patterns, which enables completely randomized MUI and achieves both time and frequency diversities.

The transmission structure of RDMA with K simultaneous users is illustrated in Figure 13. Compared with IDMA and RSMA, RDMA is simpler and may reduce the signaling overhead, since the user-specific scrambling and interleaving patterns are not needed. Meanwhile, SIC receiver is used in RDMA to provide good tradeoff between receiving complexity and detection performance.

(10) GOCA. Group orthogonal coded access (GOCA) is another long sequence-based NOMA scheme, which can be seen as an enhanced version of RDMA [29]. The major difference between GOCA and RDMA lies in the fact that the former employs the group orthogonal sequences to spread the modulation symbols into shared time and frequency resources after repetitions, as shown in Figure 14. Similar to RDMA, SIC receiver is expected to achieve good detection performance with moderate computational complexity.

The group orthogonal sequences have a two-stage structure, where orthogonal sequences and nonorthogonal sequences are used in first and second stage, respectively. Therefore, as shown in Figure 15, we can divide the GOCA sequences into several nonorthogonal groups according to the nonorthogonal sequences used in the second stage, while the sequences within a group are orthogonal to each other due to the design in the first stage.

2.3. Multiuser Detection Technologies. According to NOMA protocol, different users' signal streams are multiplexed on the same radio resources; therefore the multiuser detection (MUD) technologies are needed to distinguish independent signal streams. In the sequel, we analyze some essential MUD technologies, which are proposed to match the NOMA schemes in the above subsections.

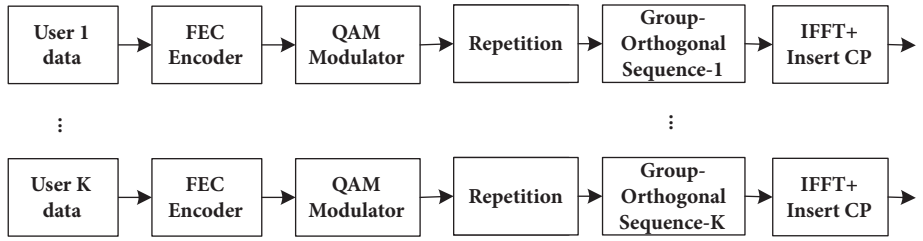


Figure 14: The transmission structure of GOCA with K simultaneous users.

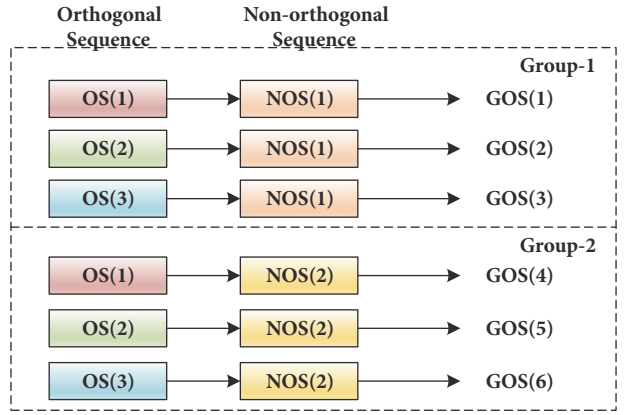


Figure 15: An example of group orthogonal sequences in GOCA, with two groups and a total of six sequences.

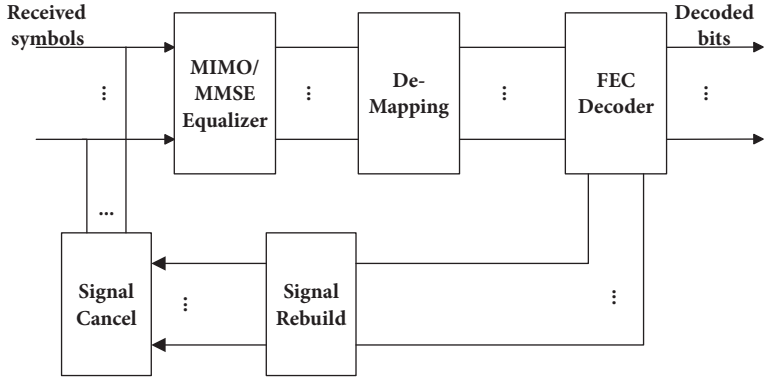


Figure 16: The structure of MMSE-SIC receiver.

2.3.1. MMSE-SIC. The minimum mean square error-successive interference cancelation (MMSE-SIC) receiver is a direct extension of MMSE receiver, as shown in Figure 16. In the first iteration of MMSE-SIC, the signal with largest received SINR is first detected by MMSE receiver by regarding the interference as noise, demapped, and then decoded to obtain the information bits. After that, the signal of this user is reconstructed and canceled from the received signal. The above procedure is repeated in the following iterations until no signal stream can be successfully recovered.

MMSE-SIC receiver suffers from error propagation problem, where the estimation errors in previous signal layers may propagate to the remaining layers. With the aim of mitigating the error propagation, the received SNRs of different data streams shall have large differences to ensure sufficient SINR in each iteration. Therefore, MMSE-SIC is especially suitable for PD-NOMA, as well as other NOMA schemes where users have diversified channel conditions.

2.3.2. MPA. MPA is an iteration-based nonlinear symbol detection algorithm, which can exploit the structure of sparse spreading sequences and achieve near maximum-likelihood (ML) performance. Different from ML receiving which estimates the entire spreading block with full search method, MPA only conducts localized optimal detection on each resource element to acquire the soft information about the transmitted symbols, and then delivers the information to the neighboring resource elements as the extrinsic information

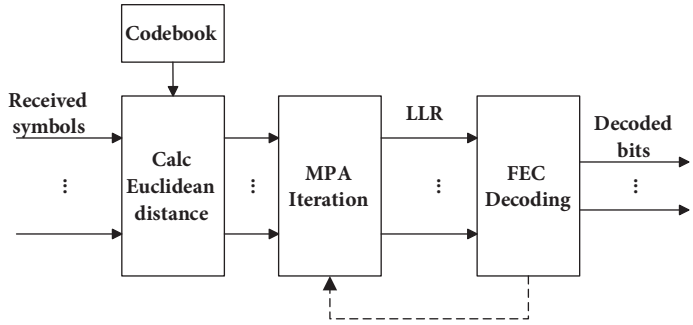


Figure 17: The structure of MPA receiver.

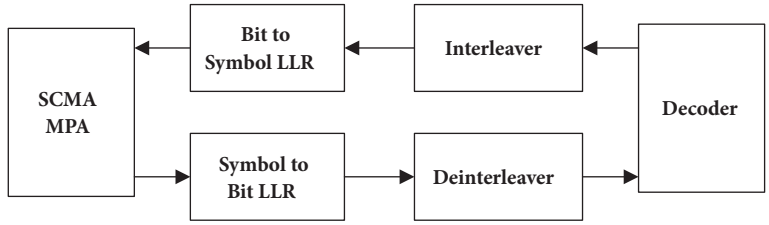


Figure 18: MPA-turbo.

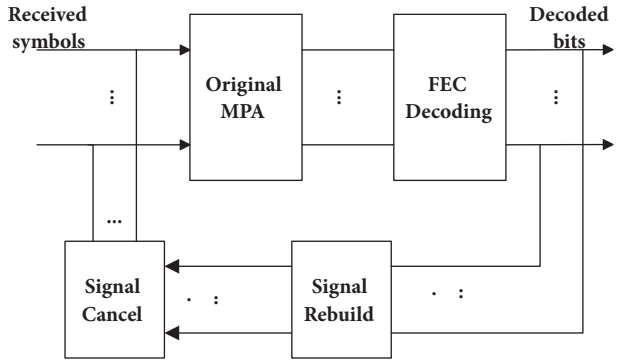


Figure 19: MPA-SIC.

of the next localized optimal detection. We show the MPA receiver in Figure 17.

The original MPA receiver focuses on symbol-level detection and does not exploit the error-correction ability of FECs to separate different signal streams. To resolve this problem, MPA-turbo and MPA-SIC are proposed as shown in Figures 18 and 19, respectively. As revealed in its name, MPA-turbo works just like the turbo decoding, where the FEC decoder processes the soft information provided by MPA and then gives feedback on the processed soft information to MPA module as extrinsic information. Different from MPA-turbo, MPA-SIC directly cancels the recovered signal streams from the received signal to mitigate the MUI.

2.3.3. EPA. Although MPA significantly reduces the computational complexity compared to ML, the complexity still grows exponentially with the number of multiplexed users on each radio resource. Estimation propagation algorithm (EPA) is another graphical-based multiuser detection algorithm, proposed for SCMA, to further reduce the computational complexity order from exponential to linear [65]. The idea of EPA originates from the variational approximate inference method, which is commonly applied in the machine learning era [45]. Different from MPA, EPA employs a Kullback-Leibler divergence based projection in the message update steps to align with the expectation propagation principle. We can directly replace the MPA module with the EPA module in the SCMA receiver, as shown in Figure 20, and generate new variants of EPA such as EPA-turbo and EPA-SIC using the similar approaches in MPA.

2.3.4. ESE-PIC. Elementary signal estimation-parallel interference cancelation (ESE-PIC) receiver is originally proposed in IDMA, which has shown robust performance even when a large number of users are multiplexed. As shown in Figure 21, ESE-PIC first detects transmitted symbols via ESE detection, a linear symbol detector. Then the detected signals are parallelly deinterleaved and decoded to acquire the coding

Table 4: Comparisons of NOMA receivers.

Receiver	Main character	Pros	Cons	Applications
MMSE-SIC	Reuses single-user receiver and SIC	Low complexity	Requires large SNR gaps among users	Almost all schemes, especially PD-NOMA
MPA/EPA	Symbol-level iterative detection	Near-ML symbol detection	Middle/high complexity	Sparse spreading NOMA
MPA-turbo/SIC	Iterative detection between symbol-level and bit-level	Better performance than MPA/EPA	High complexity	Sparse spreading NOMA
ESE-PIC	Iterative detection between symbol-level and bit-level	No requirement on sparsity	High complexity & hardware overhead	Especially bit-level NOMA

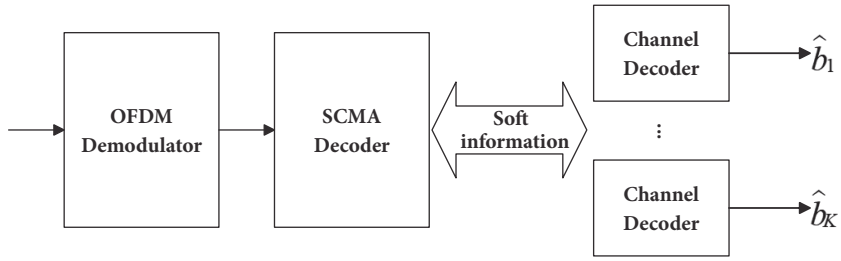


Figure 20: EPA.

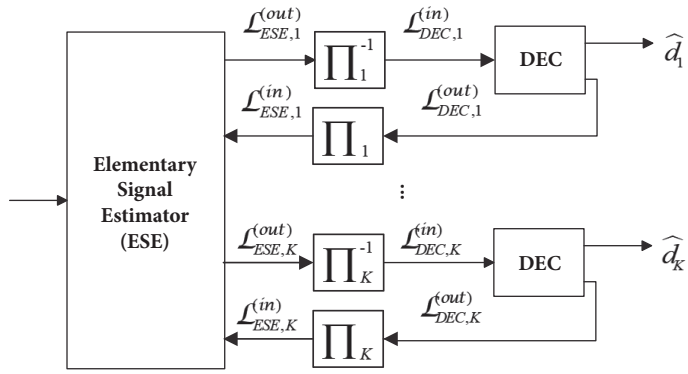


Figure 21: ESE-PIC.

gain. The output information from the decoder is then sent back to ESE module to aid symbol detection.

2.3.5. Comparisons on MUD Receivers. We now compare the pros and cons of the aforementioned receiving technologies, as well as their applicable NOMA schemes, as shown in Table 4. To sum up, the overall structure of MUD receivers consists of two parts, i.e., symbol detector and FEC decoder. Joint symbol detection, i.e., MPA and EPA, achieves better performance than single user detection, i.e., MMSE. However, they only work with short and sparse spreading sequences. Long sequence-based schemes require iterative detection, i.e., ESE-PIC; however, due to parallel message passing, several decoders may work at the same time, which leads to even larger hardware cost than MPA-turbo/SIC. To facilitate the implementation of NOMA and satisfy the diversified requirements of 5G, a good tradeoff between detection accuracy, computational complexity, latency, and hardware requirements should be achieved, which certainly requires further study.

3. Grant-Free NOMA for mMTC

The state-of-the-art NOMA schemes, mentioned in Section 2, are mainly based on centralized scheduling, where spreading sequences, interleaving patterns, and/or transmission powers of different users are scheduled by the gNB. However, the major drawback of the scheduling-based NOMA is that the signaling overhead occupies a large portion of radio resources, which makes the grant-free NOMA inevitable. In the following section, we analyze the motivation and the procedures of grant-free NOMA, as well as the typical transmission and reception technologies.

3.1. Motivation. The conventional human-type communications are normally optimized for mobile broadband (MBB) services [66], with small amount of users, high data rate,

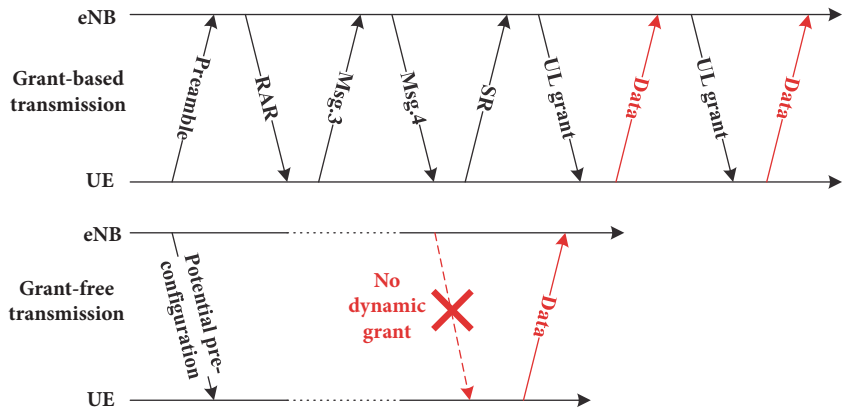


Figure 22: Comparison between the general procedures of grant-based and grant-free transmission.

and large packet size. Compared with the size of data packet in MBB, the control signaling is relatively few. Therefore, the human-type communications are not sensitive to the signaling overhead and usually involve frequent interactions between the gNB and the users to maintain high reliability and high data rate.

Despite the MBB services, 5G also aims at supporting mMTC, where massive connectivity and long battery life cycle are two key requirements. Different from in MBB scenario, the arrival of data packets in mMTC is sporadic and the packet sizes are rather small [67]. Based on [1], mMTC would support more than a million devices per square-kilometer, and this, together with the small packet sizes, makes the control signaling overhead rather significant. Therefore, the simplification on access procedure and the reduction on signaling overhead are both needed to satisfy the massive connectivity requirement of mMTC.

Grant-free NOMA, where multiple users conduct uplink instant transmissions without grant, can significantly reduce the signaling overhead. It is agreed that grant-free NOMA is more suitable for mMTC scenario due to the following concerns [66]:

(i) Energy saving: resource allocation has been well studied to extend the battery times of the devices, however, with a waste of the signaling overhead. Grant-free access can save the energy of the devices; i.e., the devices can decide to turn to active mode when small packets arrive or keep in sleep state if transmission is not needed.

(ii) Low cost devices: grant-free access can trade the computational complexity at the gNB with the hardware cost at the devices.

(iii) Latency and signaling overhead reduction: no additional latency or signaling overhead is induced by the signaling interactions.

Out of the above considerations, 3GPP has agreed that NR should target to support uplink autonomous/grant-free/contention-based transmission at least for mMTC scenario [68].

3.2. Grant-Free Procedure. In Figure 22, we compare the signaling procedures between the scheduling-based transmission in LTE and the grant-free transmission. In LTE, when a user becomes active, it would conduct random access procedure firstly, which includes at least 4 steps, i.e., preamble transmission, random access response (RAR), Message. 3, and Message. 4. After that, a scheduling request (SR) is transmitted to the gNB if the buffer is not empty, and the user does not transmit packets until it receives the uplink grant from the gNB. The above-mentioned procedures may take dozens of millisecond, which impose large signaling overhead on the network, consume more power for the signaling transmission/detection on the device side, and incur large latency for the data transmission. In contrast, grant-free transmission achieves autonomous transmission without explicit dynamic grant. Compared with the scheduling-based transmission, grant-free reduces signaling overhead, as well as control/user plane latency [69]. We note that, due to the decentralized uplink instant transmissions, the signals of users are multiplexed, which naturally leads to nonorthogonal transmissions.

We show the state graph of a grant-free user in Figure 23. If the user has no data in buffer, it stays in a sleep state; otherwise, it would wake up, synchronize according to reference signals, and acquire some necessary system broadcast information and some predefined uplink grant information. Before directly transmitting information block with the grant-free manner, preamble may be transmitted for the uplink synchronization for detection in the receiver side. Furthermore, some multiple access information could be implicitly indicated by the preamble, such as spreading signature, locations of radio resources, and the timing of retransmission. With this information, the collisions can be detected, and the blind detection complexity of the gNB can also be greatly reduced.

According to whether random access channel (RACH) is required, grant-free transmission can be classified into RACH-based grant-free and RACH-less grant-free.

3.2.1. RACH-Based Grant-Free. When all the users have performed RACH, grant-free transmission would occur in a more synchronized manner, i.e., the timing offsets among

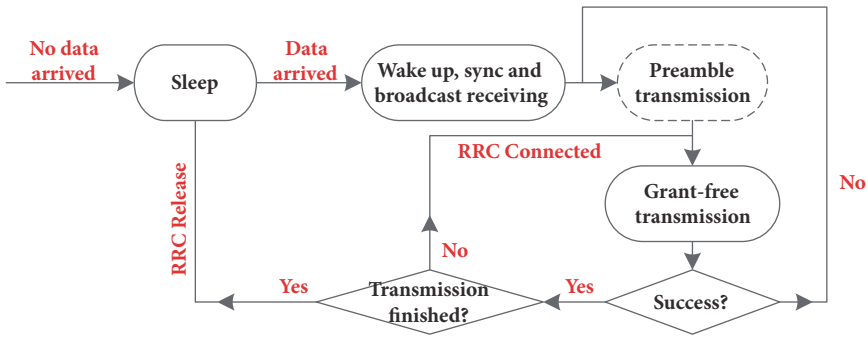


Figure 23: Grant-free uplink transmission illustrations.

users are mostly within the cyclic-prefix (CP) length. Therefore, this type of grant-free transmission is referred to as RACH-based grant-free [70] since RACH procedures have been done before data transmission. RACH-based grant-free could also reduce the overhead of SR and uplink grant, and, at the same time, it is beneficial for signal detection.

3.2.2. RACH-Less Grant-Free. In order to reduce the signaling overhead, the RACH procedure could also be canceled; i.e., data transmission phase starts whenever there are packets arriving. This method can be referred to as RACH-less grant-free [70, 71]. In this way, not only the RACH associated signaling, but also the battery energy can be saved, since the user can go to sleep if there is no data to transmit. However, the absence of RACH may result in the asynchronization among users, which may cause large detection complexity at the receiver.

3.3. Typical Grant-Free NOMA Technologies. In this subsection, we analyze the typical grant-free NOMA technologies, which are categorized into two classes, i.e., grant-free bit/symbol-level NOMA schemes and graph-based access, according to different design principles.

3.3.1. Grant-Free Bit/Symbol-Level NOMA. Grant-free bit/symbol-level NOMA can be obtained by directly incorporating grant-free access protocol with the existing NOMA schemes mentioned in Section 2, especially the short-spreading-based NOMA, e.g., SCMA, PDMA, and MUSA. To enable grant-free access in NOMA, a contention-based unit (CTU) is defined as the basic multiple access resource, as shown in Figure 24, where each CTU may consist of several fields including radio resources, reference signal, and spreading sequence [55]. One CTU may differ from the others in any fields, and these differences can be exploited by the receiver to distinguish different signal streams.

When a user has data in buffer, it randomly selects a CTU and then transmits its data packet accordingly, i.e., spreading the modulation symbols with the given spreading sequence over the given radio resources, as well as the given reference signals. When the number of active users is

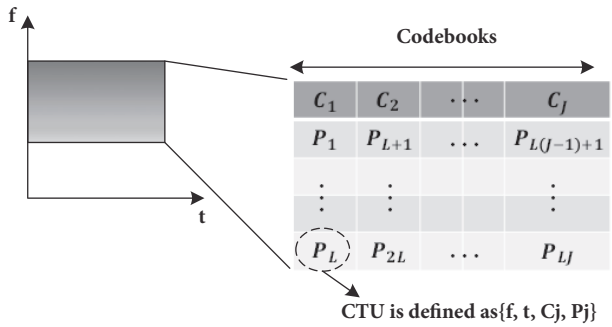


Figure 24: An illustration of CTU.

small, the users may choose different radio resources, and hence they are orthogonal. Otherwise, when the number of active users is large, the signals of the users are overlapped on the radio resources, and other fields in the CTU come into play in MUD. Therefore, grant-free NOMA can be regarded as a generalized orthogonal and nonorthogonal access.

The spreading sequences in CTUs can reuse the sequences designed for SCMA, PDMA, or MUSA. Spreading can also be replaced with sparse repetition, as proposed in [72], where simple inter- and intraslot SIC can be employed to recover multiple signal streams. In the meantime, with sparse spreading sequences, the MUI is mitigated and the receiving complexity also remains low. On the other hand, with dense spreading sequences, more diversity gain can be achieved which may combat the fading of wireless channel.

Due to the uncoordinated transmissions, the collision among users would be a severe problem in grant-free symbol-level NOMA. A hard collision happens when several users choose the same CTU. Under such circumstances, these users may be distinguished and detected only if they have distinctive channel gains. From this perspective, it is important to enlarge the pool of the spreading sequences, as done in MUSA [67]. However, enlarging the pool size may also increase the cross-correlations among the sequences, which may degrade the transmission reliability. Therefore, a good tradeoff between the pool size and the cross-correlations should be achieved to mitigate the collisions while maintaining high reliability.

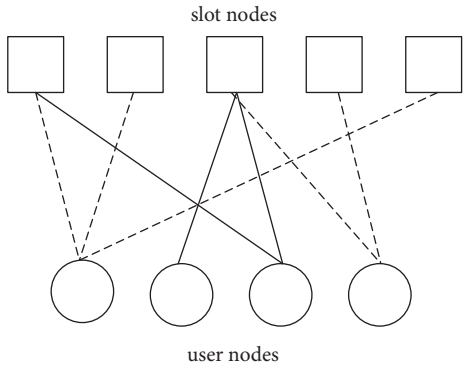


Figure 25: Bipartite graph representation of CSA.

3.3.2. Graph-Based Access Schemes. Slotted ALOHA (SA) is a conventional uncoordinated random access method which was proposed in 1970s. Recently, a class of graphical-based random access schemes has been proposed which introduces the theory of linear coding into ALOHA access [73–82]. The main idea of these schemes is to regard the random access as random coding and to optimize the access probability with coding theory. Interslot SIC receiving is applied to deal with the collisions. Besides, density evolution (DE) algorithm is usually applied to design or evaluate the transmission patterns of these schemes.

In [75], contention resolution diversity slotted ALOHA (CRDSA) is proposed which combines repetition codes with SA, where two replicas of each burst are transmitted randomly in two slots and the collided received bursts are divided by the SIC algorithm. An enhanced scheme of CRDSA, named irregular repetition slotted ALOHA (IRSA), is also introduced in [75] which optimizes the transmission method in CRDSA by bipartite graph and allows more feasible repetition pattern than CRDSA. In [73], a more generic scheme is proposed, named coded slotted ALOHA (CSA), which encodes the bursts via linear block code instead of the replicas and combines the iterative SIC with linear block code decoding to recover the source packets. A frameless ALOHA scheme based on rateless codes is provided in [83] where the transmissions of bursts act as the encoding process of rateless codes. Then, the receiver would send a feedback to the transmitter when its burst is recovered.

Hereinafter, we show a bipartite graph representation of the transmissions of CSA in Figure 25, where 4 users transmit bursts within 5 slots. Each burst node denotes the burst belonging to a user, each slot node denotes a slot, and each edge denotes that the replica of the corresponding burst is transmitted in the corresponding slot. Meanwhile, we also show the SIC process by a bipartite graph in Figure 26. In each iteration of SIC, the bursts occurred in the slots without collisions can be recovered immediately; thus the edges connected to these bursts can be removed. Then the next iteration starts and the iterations continue until no slots can be recovered. The nodes in green denote that the bursts of these users have been already recovered in the previous iterations.

3.4. Detection Techniques. In grant-free NOMA systems, the users randomly select the resources to transmit data without the dynamic scheduling. Therefore, the aforementioned MUD technologies in Section 2.3, where the identities of active users and their selected signatures are known to the receiver, are unreasonable in the practical grant-free NOMA system. Blind detection, where user activation, channel coefficients, and data packets are simultaneously detected, should be studied. Furthermore, since the transmission phase of grant-free NOMA is pretty simple, the complexity is transferred to the receiver side, which makes it rather important to design efficient blind detection algorithms.

We illustrate the general procedure at the grant-free NOMA receiver in Figure 27. The whole receiving process can be divided into two stages, i.e., user activity activation stage and data detection stage. In the first stage, active users are identified out of a potential user list. Then, in the second stage, the channel coefficients are estimated and the data packets are detected.

The idea of compressive sensing (CS) can be incorporated into the first stage due to the fact that the user activation is sparse. This sparsity is utilized in the CS-MPA detector which jointly uses CS and MPA to realize both stages simultaneously [84]. Compared with the conventional MPA without activity detection, it achieves better BLER and throughput. In addition to CS, the user activity detection could be realized by different algorithms and schemes. For example, focal underdetermined system solver (FOCUSS) and expectation maximization (EM) are proposed and analyzed for active pilot detection [85], and they can be combined with the blind data detection method, i.e., joint data and active codebook detection (JMPA), to recover the data in the spreading-based grant-free systems. It is seen that JMPA can achieve scarcely any performance degradation in decoding users' data without prior knowledge of active codebooks. Furthermore, to avoid the redundant pilot overhead, a novel sparsity-inspired sphere decoding (SI-SD) algorithm is proposed by introducing one additional all-zero codeword to achieve the maximum a posteriori (MAP) detection [86]. However, either CS or EM can only get the rough information about the active users. Detection-based group orthogonal matching pursuit (DGOMP) is a user activation detector which is promising to get a more accurate active user set [87]. Meanwhile, an enhanced version of JMPA is proposed in [87], which takes the channel gain and noise power into consideration when calculating the prior information of the zero codeword. The modified JMPA also helps to eliminate the false detection caused by noise, channel fading, and nonorthogonality of pilot sequences.

4. Implementation Issues

Nonorthogonal transmission is completely different from the orthogonal transmission which has been widely implemented in LTE. As a consequence, nonorthogonal transmission raises some implementation issues for practical deployment. In this section, we analyze some important implementation issues related to scheduling-based NOMA schemes and grant-free NOMA, respectively.

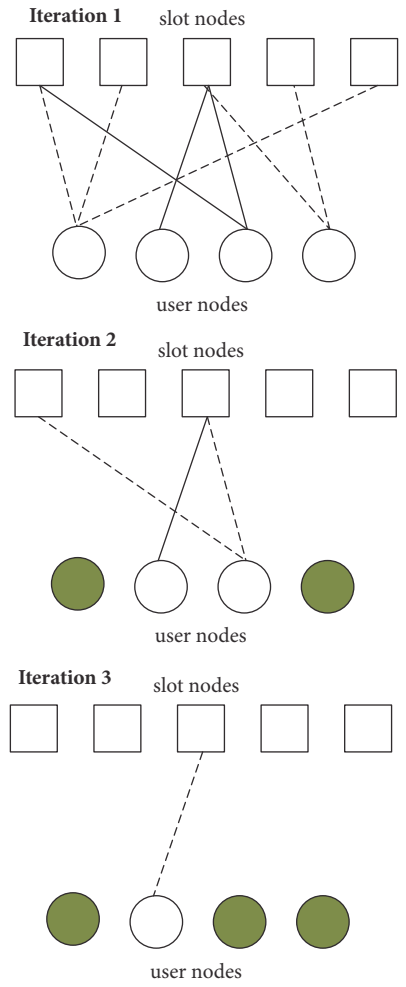


Figure 26: Bipartite graph representation of SIC process.

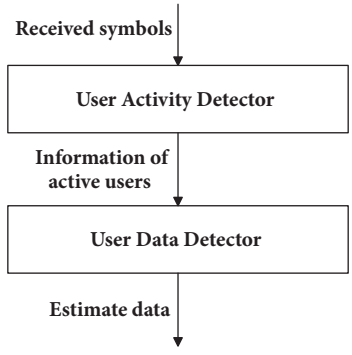


Figure 27: Grant-free NOMA receiver.

4.1. Scheduling-Based NOMA. Recall that the major difference between OMA and scheduling-based NOMA is that the latter allows multiple superimposed transmissions on the same radio resources, while the former only allows orthogonal transmissions. Hence, the resource allocation and demodulation reference signal (DM-RS) should be designed to facilitate scheduling-based NOMA.

4.1.1. Resource Allocation and Scheduling. Resource allocation, where the radio resources are assigned among users via centralized scheduling to meet certain optimization targets, has been extensively exploited in LTE to promote the system-level performance, including peak transmission data rate, average throughput, and user fairness. When multiple scheduling requests are transmitted from the users in LTE, the network orthogonally allocates the limited radio resources to a subset of the candidate users. However, the resource allocation in NOMA would be complex, since the resources in NOMA not only consist of radio resources, but also MA signature resources. Due to the fact that radio resources can be shared among users, the resource allocation problem would be even more complex. Besides, specific resource allocation methods should be designed for different NOMA schemes to match their unique characteristics. For example, PD-NOMA tends to multiplex the cell-center users and cell-edge users, while SCMA is more likely to superimpose the signals of colocated users.

The resource allocation in NOMA should also be designed to mitigate the effect of error propagation, if SIC-based receiver is employed. For example, the users with small

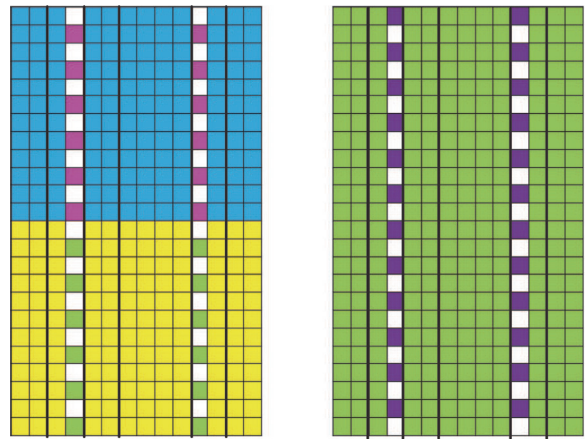


Figure 28: An example of the DM-RS structures with different combs.

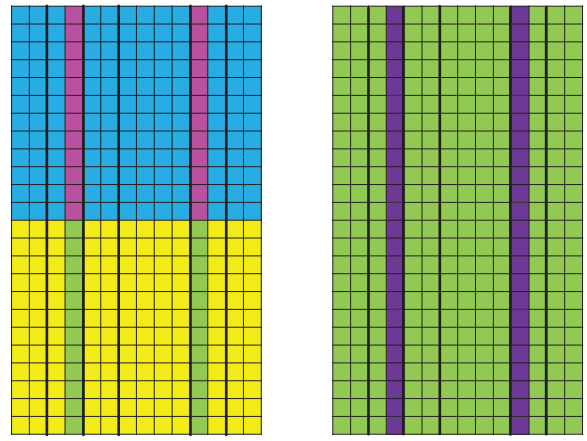


Figure 29: An example of the DM-RS structure with different OCC.

SIC order should be allocated with more radio resources to transmit signals with lower coding rates. When multicell system is considered, resource allocation should be designed to reduce the ICI. As an instance, very low density MA signatures may be allocated to the cell-edge users in uplink NOMA.

4.1.2. DM-RS. In LTE, DM-RSs with different cyclic shifts and orthogonal cover codes (OCCs) can be orthogonally multiplexed when the cyclic shift is longer than the channel delay spread. However, with the increasing demands for DM-RS ports in NOMA, it is unrealistic to add more cyclic shifts, and in the meantime, the OCC resources are also limited. Comb structures may be adopted in the design of DM-RS for NOMA. Figure 28 shows an example of the comb structure of DM-RS, which could increase the number of DM-RS resources without decreasing the accuracy in channel estimation. Compared with previous DM-RS schemes, as shown in Figure 29, the comb structure guarantees the orthogonal property of DM-RS via FDM [88, 89]. To support massive connectivity, nonorthogonal DM-RS may be further introduced to enlarge the number of DM-RS ports, where advanced channel estimation techniques should be exploited to mitigate the effect of nonorthogonality [90].

4.2. Grant-Free NOMA. As discussed in the previous section, data transmission in grant-free NOMA follows an arrive-and-go manner, which is very different from both OMA and scheduling-based NOMA. Therefore, implementing grant-free NOMA would require more efforts. This subsection presents several critical implementation issues related to grant-free NOMA, namely, resource allocation, hybrid automatic repeat request (HARQ), link adaptation, and physical signal design.

4.2.1. Resource Allocation. Similar to scheduling-based NOMA, the resources of grant-free NOMA also consist of radio resources and MA signatures. Two kinds of resource selection methods have been proposed in grant-free NOMA, i.e., random resource selection method and preconfigured method [91].

In random resource selection method, users randomly select the radio resources and the MA signatures and then transmit signals accordingly. In this case, the user activities

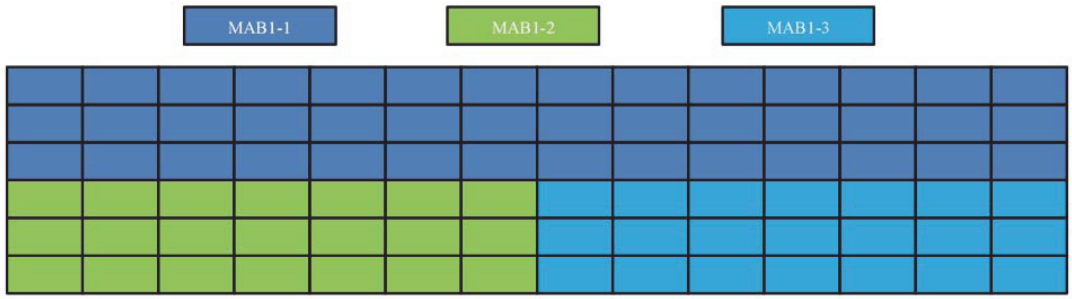


Figure 30: An example configuration of MAB.

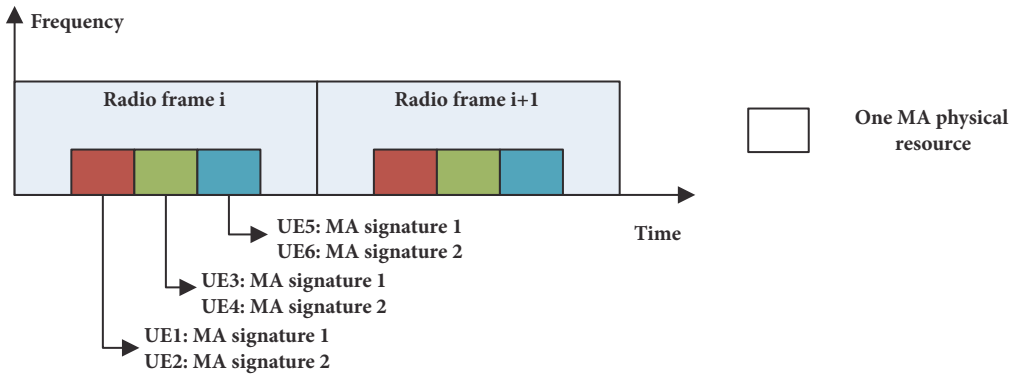


Figure 31: An illustration of predetermined MA resources for grant-free transmission.

are not available at the gNB [92], which may cause the ambiguity. In order to resolve the above problem, the radio resources should be divided into orthogonal multiple access blocks (MABs) in the time and frequency domains, as shown in Figure 30. Different MABs occupy different radio resources and may adopt different transmission settings, such as transmission block sizes (TBSs), modulation and coding schemes (MCSs), and transmission modes (TMs). The configurations of the MABs in a cell can be broadcasted by the gNB as system information. During data transmission phase, each active user first selects one MAB and then selects an MA signature. At the receiver, multiuser detection can be parallelly performed on each MAB. In addition, each MAB may be assigned with a limited number of MA signatures, which can reduce the computational complexity of blind detection at the receiver.

In the preconfigured resource allocation scheme, several users may be allocated the same radio resources along with unique MA signatures [93]. The MA signatures can be used for active user identification and collision reduction. We show an example of preconfigured scheme in Figure 31, where six users are scheduled with three distinctive pieces of MA radio resources, and the multiplexed users are allocated with different MA signatures.

4.2.2. HARQ. When the initial grant-free NOMA transmission is not successfully recovered, there is a need to retransmit the data for one or more times. HARQ is a profitable retransmission scheme which can merge the information of new transmission with previous transmissions in an effective way [94]. Due to the absence of uplink grant, one significant issue of supporting HARQ in uplink grant-free transmission is how gNB identifies the first transmission and the retransmissions for a HARQ process. One potential method is that the gNB can explicitly schedule retransmissions via downlink control signaling. Another method is to divide the MABs into several groups according to the maximum allowed number of retransmissions [95], where different users may select different MAB since they have different retransmission numbers. Another key issue in HARQ is the ACK/NACK indication. As discussed in [95], when collisions happen, gNB can utilize the RAR-style feedback, normally consisting of HARQ-ACK as well as user identification information, from which the collided users may identify whether or not their data are successfully decoded.

4.2.3. Link Adaptation. The link adaptation has been introduced in LTE to adapt to the instantaneous channel condition by adjusting the transmission parameters. Properly design link adaptation not only results in low BLER, but also reduces the retransmission number and collision probability [96]. In addition, the suitable link adaptation could achieve lower latency, which is an important target in NR application scenarios [97].

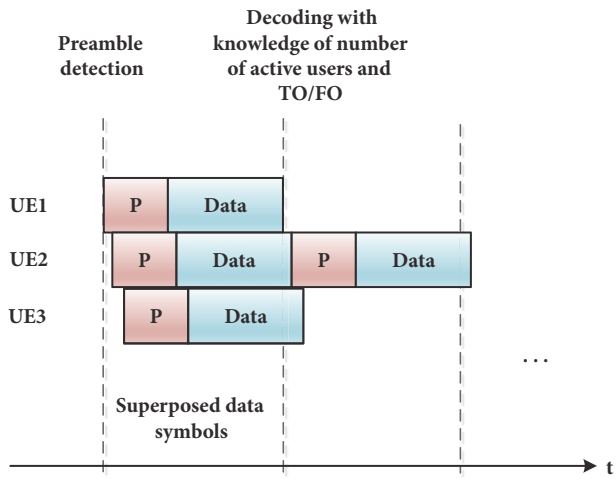


Figure 32: An illustration of RACH-less grant-free NOMA transmissions, where the data symbols are transmitted immediately after preambles.

In general, the link adaptation is realized by obtaining channel state information. However, the intermittent transmissions in uplink grant-free NOMA lead to the fact that the users might not be able to get accurate uplink channel status [98]. One solution is to use the measurements of downlink reference signals to determine the link adaptation parameters for uplink transmission. The link adaptation parameters may include MCS, number of repetitions, size of MA radio resources, and the transmission power during subsequent retransmissions [99].

4.2.4. Physical Signal Design. Physical signals, including preamble and DM-RS, are another important design aspect in grant-free NOMA. Preamble has been used in LTE for random access request [100]. However, with the aim of reducing signaling overhead, the complete random access procedure may be omitted (e.g., RACH-less grant-free). Instead, the preambles are usually directly followed by data symbols, as shown in Figure 32.

In RACH-less grant-free NOMA transmission, the users autonomously choose MA signatures as well as time instant for initial transmission, which are not known by gNB and may lead to asynchronization among received signals. In this case, well-designed preambles could assist the active user identification, MCS indication, timing offset (TO)/frequency offset (FO) estimation, and channel estimation.

Similar to the preambles, DM-RS can be used for channel estimation and user identification in grant-free NOMA [89]. However, due to the uncoordinated transmissions, different users may choose the same DM-RS, which greatly degrades the accuracy of channel estimation. To guarantee low collision probability on DM-RS, sufficient number of orthogonal/semiorthogonal DM-RSs should be provided [88]. Besides, advanced multiuser detection algorithms, such as SIC, could also help to increase the quality of channel estimation [27].

5. Future Research Challenges

Existing NOMA schemes have fully utilized the time, frequency, power, spatial, code, and interleave domains to enhance the connectivity and spectral efficiency. However, NOMA must be further studied and enhanced to satisfy the potential needs beyond NR. In this section, we highlight the future research directions of NOMA, as shown in Figure 33, including physical layer enhancement, cross layer design, joint design with other technologies, and applications of NOMA in new scenarios.

5.1. Physical Layer Enhancement. The existing NOMA schemes focus on either bit-level operations or symbol-level operations, which cannot achieve the global optimal designs. A straightforward idea is to conduct joint design of bit-level and symbol-level operations, e.g., joint design of channel encoding and symbol spreading, where the coding structure is optimized according to NOMA transmission [101, 102]. However, these designs are much too sensitive to certain channel conditions and require further enhancement for practical implementation. Despite the design at the transmitter side, signal detection at the receiver side is another physical layer technology which can be enhanced. To exploit the coding structure, several joint detection methods have been proposed in [103, 104]. However, the existing methods usually require many iterations between symbol detection and channel encoding, which leads to large detection latency. Furthermore, the high complexity and latency of blind detection still constitute an obstacle to the deployment of grant-free transmission. Therefore, simple and uniform design is required to reduce the computational complexity, as well as to maintain high reliability.

5.2. Cross Layer Design. Except for the physical layer enhancement, the cross layer design may also play an important role in the future development of NOMA [105]. For example,

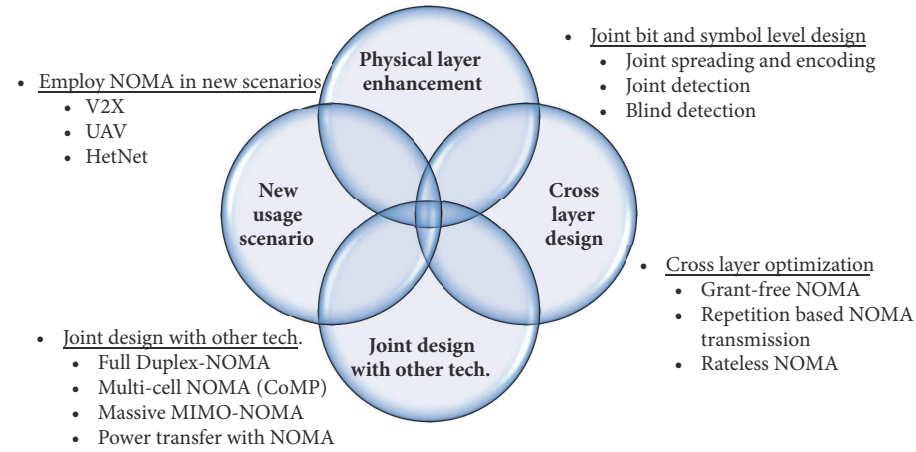


Figure 33: Future research aspects of NOMA.

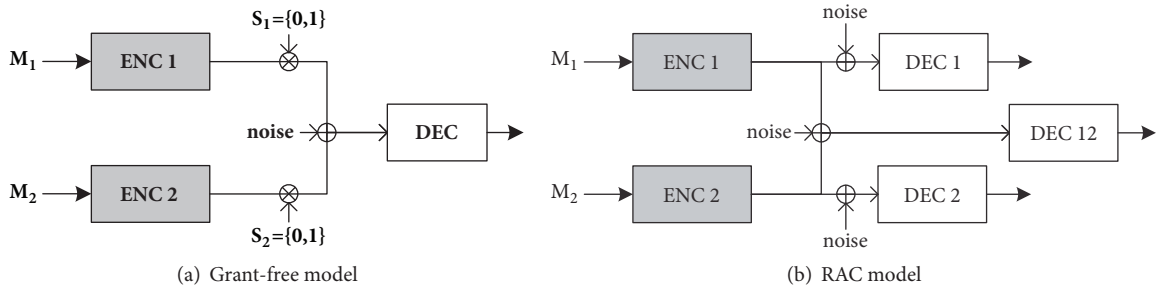


Figure 34: Illustrations of grant-free transmission and RAC models. M_1 and M_2 are the messages to be transmitted.

grant-free NOMA, which integrates the access layer protocol, i.e., grant-free protocol, into NOMA transmission, has been a promising technology for mMTC, as mentioned earlier in this paper. Besides, it is also promising to combine NOMA with other access layer techniques, such as repetition technique or rateless coding, and optimize the repetition number or code structure, respectively.

Unlike the physical layer technology which is always directed by the Shannon information theory, it is nontrivial to derive the potential limits of the channel when access layer is involved; e.g., the achievable channel capacity of grant-free NOMA is still an open problem. One possible solution to analyze the theoretical potential of NOMA with cross layer design is to formulate the Shannon information-theoretic channel model. For example, grant-free NOMA can be formulated as the random access channel (RAC) as shown in Figure 34 [106], which uses the auxiliary receivers to represent different states of user activation in grant-free access. We note that this channel model is similar to the interference channel model, where applying rate splitting can achieve a good capacity region. With this insight, one may naturally consider the deployment of rate splitting into grant-free NOMA. However, elaborate design and optimization are required to enhance the grant-free NOMA with rate splitting, for example, the coding rate and the power allocation coefficient for each splitting layer.

5.3. Joint Design with Other Technologies. Although NOMA, on its own, has met its bottleneck, we can always incorporate NOMA into other cutting edge technologies to see if there are additional advantages. For example, full duplex (FD) technology [107, 108], which is expected to increase the spectrum efficiency by a factor of two, can be jointly designed with NOMA. The conventional FD, which considers a point-to-point communication channel, is modeled by two-way channel (TWC) model. Consider a case where multiple users exist in a cell, and both gNB and the users have FD ability. Consequently, gNB and the users are simultaneously transmitting and receiving, so that the entire system can be regarded as a MAC in uplink, as well as a BC in downlink. Therefore, we may name the channel here as TW-MAC/BC model, which is illustrated in Figure 35. Obviously, NOMA technologies can be directly employed to enhance the throughput in either uplink or downlink, as it does in the conventional MAC and BC. However, there are two major differences between TW-MAC/BC and conventional MAC/BC: the first one is that, in the former case, the uplink and downlink transmissions may interfere each other due to nonideal interference cancelation, which means a good trade-off between mutual interference and sum capacity should be achieved; and the second one is that each user or gNB knows what they are receiving when they are transmitting, which means that the received signal may be exploited to derive

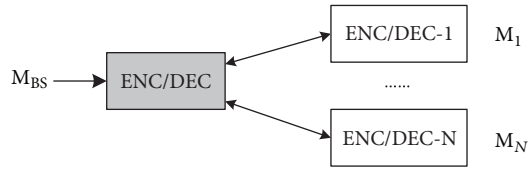


Figure 35: TW-MAC/BC. M_{gNB}, M_1, and M_N are the messages to be transmitted by the gNB, user-1, and user-N.

the hidden information about the channel condition and to enhance the transmission reliability.

Furthermore, NOMA can be jointly designed with cooperative multiple point (CoMP) [9, 109, 110] or multiple-antenna technology [41, 111–113]. Also, there have been some initial studies about employing NOMA in wireless power transfer network to increase the access opportunities [114–116]. Although jointly designing NOMA with the other technologies seems straightforward, whether the joint design can produce "a whole greater than the sum of its parts" is still an open question.

5.4. New Usage Scenarios. Despite the application of NOMA in the cellular networks, NOMA is also a promising technology in other new usage scenarios, e.g., vehicle to X (V2X) [117–119], unmanned aerial vehicle (UAV) [120, 121], and heterogeneous network (HetNet) [122], due to its superior performance. Although NOMA can be directly employed into these scenarios, elaborate design is still required to accommodate the channel characteristics of these scenarios and satisfy the diversified performance requirements. For example, due to the mobility of the vehicles in V2X and UAV, the handover is frequent and the cochannel interference is severe, which may degrade the reliability of existing NOMA technologies. Therefore, NOMA should be designed to maintain high reliability, as well as high throughput. As another example, in the HetNet, multiple kinds of cells, which have different transmission SNRs, may colocate together. NOMA should be designed to utilize this effect by paring the signals from different cells.

6. Conclusion

NOMA has been recognized as one of the key enabling technologies to accomplish the diversified requirements of 5G. By enabling multiple users to share the same radio resources and exploiting the advanced MUD algorithms, NOMA exhibits better performance than OMA, especially in SE and connectivity. As demonstrated in this review, the idea of superimposing the users has been carried forward into multiple domains, including power, code, interleave, and scramble, which have motivated many NOMA schemes. Meanwhile, various multiuser receiving technologies also facilitate the application of NOMA in different scenarios. Besides, we also look into grant-free NOMA, which aims to reduce the signaling overhead and increase the access probability for mMTC. Subsequently, the implementation issues and future scope of NOMA are analyzed. We hope that our survey would shed a light on the deployment and development of NOMA technologies.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61620106001.

References

[1] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies".

[2] A. E. Gamal and Y. H. Kim, "Lecture notes on network information theory," *Mathematics*, 2010.

[3] S. Shimamoto, Y. Onozato, and Y. Teshigawara, "Performance evaluation of power level division multiple access (PDMA) scheme," in *Proceedings of the [Conference Record] SUPERCOMM/ICC '92 Discovering a New World of Communications*, pp. 1333–1337, Chicago, IL, USA.

[4] K. Pedersen, T. Kolding, I. Seskar, and J. Holtzman, "Practical implementation of successive interference cancellation in DS/CDMA systems," in *Proceedings of the ICUPC - 5th International Conference on Universal Personal Communications*, pp. 321–325, Cambridge, MA, USA.

[5] G. Mazzini, "Power division multiple access," in *Proceedings of the ICUPC '98. IEEE 1998 International Conference on Universal Personal Communications. Conference Proceedings*, pp. 543–546, Florence, Italy.

[6] Y. Yuan, L. Anxin, and H. Kayama, "Superimposed radio resource sharing for improving uplink spectrum efficiency," in *Proceedings of the 2008 14th Asia-Pacific Conference on Communications, APCC 2008*, usa, October 2008.

[7] 3GPP TR 38.812, "Study on non-orthogonal multiple access (NOMA) for NR".

[8] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proceedings of the IEEE Conference on Standards for Communications and Networking, CSCN 2015*, pp. 18–23, jpn, October 2015.

[9] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.

[10] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[11] Z. Ding, Y. Liu, J. Choi et al., "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.

[12] 3GPP R1-165021, "WF on common features and general framework of MA schemes".

[13] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

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

[15] Y. Tao, L. Liu, S. Liu, and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Communications*, vol. 12, no. 10, pp. 1–15, 2015.

[16] Y. Wang, B. Ren, S. Sun, S. Kang, and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Communications*, vol. 13, pp. 52–66, 2016.

[17] Z. DIng, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[18] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Communications*, vol. 14, no. 4, pp. 17–25, 2016.

[19] 3GPP R1-165021, "Performance of interleave division multiple access (IDMA) in combination with OFDM family waveforms".

[20] 3GPP TSG-RAN WG1-163992, "Non-orthogonal multiple access candidate for NR".

[21] 3GPP R1-162385, "Multiple access schemes for new radio interface".

[22] 3GPP R1-164329, "Initial LLS results for UL non-orthogonal multiple access".

[23] 3GPP R1-164869, "Low code rate and signature based multiple access scheme for NR".

[24] 3GPP R1-162226, "Discussion on multiple access for new radio interface".

[25] 3GPP R1-162517, "Considerations on DL/UL multiple access for NR".

[26] 3GPP R1-165019, "Non-orthogonal multiple access for NR".

[27] 3GPP R1-163111, "Initial views and evaluation results on non-orthogonal multiple access for NR uplink".

[28] 3GPP R1-163383, "Candidate solution for new multiple access".

[29] 3GPP R1-167535, "New uplink non-orthogonal multiple access schemes for NR".

[30] 3GPP R1-163510, "Candidate NR multiple access schemes".

[31] 3GPP R1-164346, "MA for eMBB in mmWave spectrum".

[32] 3GPP R1-162153, "Overview of non-orthogonal multiple access for 5G".

[33] 3GPP RWS-150051, "5G vision for 2020 and beyond".

[34] K. Higuchi and A. Benjebbour, "Non-Orthogonal Multiple Access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98B, no. 3, pp. 403–414, 2015.

[35] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Uplink Non-Orthogonal Multiple Access (NOMA) with Single-Carrier Frequency Division Multiple Access (SC-FDMA) for 5G systems," *IEICE Transactions on Communications*, vol. E98B, no. 8, pp. 1426–1435, 2015.

[36] Z. Wei, D. W. K. Ng, J. Yuan, and H.-M. Wang, "Optimal Resource Allocation for Power-Efficient MC-NOMA with Imperfect Channel State Information," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3944–3961, 2017.

[37] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC*, pp. 611–615, September 2013.

[38] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proceedings of the 51st Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, IEEE, Tokyo, Japan, May 2000.

[39] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *Proceedings of the 2012 9th International Symposium on Wireless Communication Systems, ISWCS 2012*, pp. 476–480, August 2012.

[40] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *Proceedings of the 2014 79th IEEE Vehicular Technology Conference, VTC 2014-Spring*, kor, May 2014.

[41] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1640–1646, 2006.

[42] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access," *Wireless Personal Communications*, vol. 87, no. 3, pp. 837–867, 2016.

[43] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," in *Proceedings of the 2014 IEEE Globecom Workshops, GC Wkshps 2014*, pp. 1026–1031, usa, December 2014.

[44] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *Proceedings of the IEEE Globecom Workshops (GC '13)*, pp. 66–70, IEEE, Atlanta, Ga, USA, December 2013.

[45] N. Ye, A. Wang, X. Li, W. Liu, X. Hou, and H. Yu, "On Constellation Rotation of NOMA With SIC Receiver," *IEEE Communications Letters*, vol. 22, no. 3, pp. 514–517, 2018.

[46] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving Sustainable Ultra-Dense Heterogeneous Networks for 5G," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[47] Y. Fu, Y. Chen, and C. W. Sung, "Distributed Power Control for the Downlink of Multi-Cell NOMA Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6207–6220, 2017.

[48] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave-division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, 2006.

[49] L. Ping, Q. Guo, and J. Tong, "The OFDM-IDMA approach to wireless communication systems," *IEEE Wireless Communications Magazine*, vol. 14, no. 3, pp. 18–24, 2007.

[50] L. Ping, L. Liu, K. Y. Wu, and W. K. Leung, "Approaching the Capacity of Multiple Access Channels Using Interleaved Low-Rate Codes," *IEEE Communications Letters*, vol. 8, no. 1, pp. 4–6, 2004.

[51] H. Wu, L. Ping, and A. Perotti, "User-specific chip-level interleaver design for IDMA systems," *IEEE Electronics Letters*, vol. 42, no. 4, pp. 233-234, 2006.

[52] R. Zhang and L. Hanzo, "Three design aspects of multicarrier interleave division multiple access," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3607–3617, 2008.

[53] . Li Ping, . Lihai Liu, K. Wu, and . Leung WK, "On interleave-division multiple-access," in *Proceedings of the 2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577)*, pp. 2869–2873 Vol.5, Paris, France, June 2004.

[54] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 332–336, IEEE, London, UK, September 2013.

[55] K. Au, L. Zhang, H. Nikopour et al., "Uplink contention based SCMA for 5G radio access," in *Proceedings of the 2014 IEEE Globecom Workshops, GC Wkshps 2014*, pp. 900–905, usa, December 2014.

[56] H. Yu, Z. Fei, N. Yang, and N. Ye, "Optimal design of resource element mapping for sparse spreading non-orthogonal multiple access," *IEEE Wireless Communications Letters*, vol. PP, no. 99, p. 1, 2018.

[57] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proceedings of the 80th IEEE Vehicular Technology Conference, VTC 2014-Fall*, Canada, September 2014.

[58] B. Ren, Y. Wang, X. Dai, K. Niu, and W. Tang, "Pattern matrix design of PDMA for 5G UL applications," *China Communications*, vol. 13, pp. 159–173, 2016.

[59] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2015*, chn, October 2015.

[60] 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 Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3185–3196, 2017.

[61] P. Li, Y. Jiang, S. Kang, F. Zheng, and X. You, *Pattern division multiple access with large-scale antenna array*, 2017.

[62] J. Zeng, B. Liu, and X. Su, "Interleaver-Based Pattern Division Multiple Access with Iterative Decoding and Detection," in *Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Sydney, NSW, June 2017.

[63] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-User Shared Access for Internet of Things," in *Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, Nanjing, China, May 2016.

[64] H. Hu and J. Wu, "New constructions of codebooks nearly meeting the Welch bound with equality," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 60, no. 2, pp. 1348–1355, 2014.

[65] X. Meng, Y. Wu, Y. Chen, and M. Cheng, "Low complexity receiver for uplink SCMA system via expectation propagation," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference, WCNC 2017*, usa, March 2017.

[66] 3GPP R1-164268, "GB and GF MA for mMTC".

[67] 3GPP R1-166403, "Grant-free Multiple Access Schemes for mMTC".

[68] 3GPP R1-165021, "WF on clarification of grant-free transmission for mMTC".

[69] 3GPP R1-1609398, "Uplink grant-free access for 5G mMTC".

[70] 3GPP R1-167392, "Discussion on multiple access for UL mMTC".

[71] 3GPP R1-166405, "Discussion on grant-free concept for UL mMTC".

[72] N. Ye, A. Wang, X. Li, H. Yu, A. Li, and H. Jiang, "A Random Non-Orthogonal Multiple Access Scheme for mMTC," in *Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, Sydney, NSW, June 2017.

[73] Z. Sun, Y. Xie, J. Yuan, and T. Yang, "Coded Slotted ALOHA for Erasure Channels: Design and Throughput Analysis," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4817–4830, 2017.

[74] E. Paolini, Č. Stefanović, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 144–150, 2015.

[75] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.

[76] L. Toni and P. Frossard, "Prioritized Random MAC Optimization Via Graph-Based Analysis," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 5002–5013, 2015.

[77] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-coupled random access on graphs," in *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012*, pp. 478–482, usa, July 2012.

[78] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 57, no. 2, pp. 803–834, 2011.

[79] M. Ivanov, F. Brännström, A. GraellAmat, and G. Liva, "Unequal Error Protection in Coded Slotted ALOHA," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 536–539, 2016.

[80] D. Jia, Z. Fei, H. Lin, J. Yuan, and J. Kuang, "Distributed Decoding for Coded Slotted ALOHA," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1715–1718, 2017.

[81] C. Stefanovic and P. Popovski, "Coded slotted ALOHA with varying packet loss rate across users," in *Proceedings of the 2013 1st IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013*, pp. 787–790, usa, December 2013.

[82] Z. Sun, L. Yang, J. Yuan, and M. Chiani, "A novel detection algorithm for random multiple access based on physical-layer network coding," in *Proceedings of the 2016 IEEE International Conference on Communications Workshops, ICC 2016*, pp. 608–613, mys, May 2016.

[83] Č. Stefanović and P. Popovski, "ALOHA random access that operates as a rateless code," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4653–4662, 2013.

[84] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proceedings of the 82nd IEEE*

Vehicular Technology Conference, VTC Fall 2015, usa, September 2015.

[85] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proceedings of the 2014 11th International Symposium on Wireless Communications Systems, ISWCS 2014*, pp. 853–857, esp, August 2014.

[86] G. Chen, J. Dai, K. Niu, and C. Dong, "Sparsity-Inspired Sphere Decoding (SI-SD): A Novel Blind Detection Algorithm for Uplink Grant-Free Sparse Code Multiple Access," *IEEE Access*, vol. 5, pp. 19983–19993, 2017.

[87] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "Blind detection of uplink grant-free SCMA with unknown user sparsity," in *Proceedings of the 2017 IEEE International Conference on Communications, ICC 2017*, fra, May 2017.

[88] 3GPP R1-1612573, "Collision analysis of grant-free based multiple access".

[89] 3GPP R1-1608919, "Considerations on pre-configured resource for grant-free based UL non-orthogonal MA".

[90] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Transactions on Communications*, vol. PP, no. 99, p. 1, 2017.

[91] 3GPP R1-1609227, "On MA resource and MA signature configurations".

[92] 3GPP R1-1608917, "Considerations on random resource selection".

[93] 3GPP R1-1609647, "On MA resources for grant-free transmission".

[94] 3GPP R1-1608859, "The retransmission and HARQ schemes for grant-free".

[95] 3GPP R1-1609039, "HARQ operation for grant-free based multiple access".

[96] 3GPP R1-1609649, "Grant-free retransmission with diversity and combining for NR".

[97] 3GPP R1-1609648, "Collision handling for grant-free".

[98] 3GPP R1-1609654, "Link adaptation for grant-free transmissions".

[99] 3GPP R1-1610374, "Support of link adaptation for UL grant-free NOMA schemes".

[100] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: Form Theory to Practice*, Wiley, 2011.

[101] J. Dai, K. Niu, Z. Si, C. Dong, and J. Lin, "Polar-coded non-orthogonal multiple access," *IEEE Transactions on Signal Processing*, no. 99, p. 1, 2017.

[102] M. Qiu, Y. Huang, S. Shieh, and J. Yuan, "A Lattice-Partition Framework of Downlink Non-Orthogonal Multiple Access without SIC," in *Proceedings of the 2017 IEEE Global Communications Conference (GLOBECOM 2017)*, pp. 1–6, Singapore, December 2017.

[103] S. Chen, K. Peng, Y. Zhang, and J. Song, "Near capacity LDPC coded MU-BICM-ID for 5G," in *Proceedings of the 11th International Wireless Communications and Mobile Computing Conference, IWCMC 2015*, pp. 1418–1423, hrv, August 2015.

[104] L. Wen, R. Razavi, M. A. Imran, and P. Xiao, "Design of Joint Sparse Graph for OFDM System," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1823–1836, 2015.

[105] G. Liu, Z. Ma, X. Chen, Z. Ding, F. R. Yu, and P. Fan, "Cross-layer power allocation in non-orthogonal multiple access systems for statistical QoS provisioning," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11–388, Dec 2017.

[106] P. Minero, M. Franceschetti, and D. N. Tse, "Random access: an information-theoretic perspective," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 58, no. 2, pp. 909–930, 2012.

[107] M. S. Sim, M. Chung, D. Kim, J. Chung, D. K. Kim, and C.-B. Chae, "Nonlinear Self-Interference Cancellation for Full-Duplex Radios: From Link-Level and System-Level Performance Perspectives," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 158–167, 2017.

[108] A. Kord, D. L. Sounas, and A. Alu, "Achieving Full-Duplex Communication: Magnetless Parametric Circulators for Full-Duplex Communication Systems," *IEEE Microwave Magazine*, vol. 19, no. 1, pp. 84–90, 2018.

[109] Y. Tian, A. R. Nix, and M. Beach, "On the Performance of Opportunistic NOMA in Downlink CoMP Networks," *IEEE Communications Letters*, vol. 20, no. 5, pp. 998–1001, 2016.

[110] Z. Liu, G. Kang, L. Lei, N. Zhang, and S. Zhang, "Power Allocation for Energy Efficiency Maximization in Downlink CoMP Systems with NOMA," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, San Francisco, CA, USA, March 2017.

[111] X. Liu, Y. Liu, X. Wang, and H. Lin, "Highly Efficient 3-D Resource Allocation Techniques in 5G for NOMA-Enabled Massive MIMO and Relaying Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2785–2797, 2017.

[112] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low Complexity Beamforming and User Selection Schemes for 5G MIMO-NOMA Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2708–2722, 2017.

[113] V. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O. Shin, "Precoder Design for Signal Superposition in MIMO-NOMA Multicell Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2681–2695, 2017.

[114] Y. Xu, C. Shen, Z. Ding et al., "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4874–4886, 2017.

[115] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "The Impact of Power Allocation on Cooperative Non-orthogonal Multiple Access Networks with SWIPT," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4332–4343, 2017.

[116] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative Non-orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938–953, 2016.

[117] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance Analysis of NOMA-SM in Vehicle-to-Vehicle Massive MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2653–2666, 2017.

[118] B. Di, L. Song, Y. Li, and G. Y. Li, "NOMA-Based Low-Latency and High-Reliable Broadcast Communications for 5G V2X Services," in *Proceedings of the GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, December 2017.

[119] J. Dai, K. Niu, Z. Si, C. Dong, and J. Lin, "Polar-coded non-orthogonal multiple access," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1374–1389, 2018.

[120] W. Fawaz, C. Abou-Rjeily, and C. Assi, "UAV-Aided Cooperation for FSO Communication Systems," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 70–75, 2018.

[121] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV-Based IoT Platform: A Crowd Surveillance Use Case," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 128–134, 2017.

[122] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Resource allocation for non-orthogonal multiple access in heterogeneous networks," in *Proceedings of the ICC 2017 - 2017 IEEE International Conference on Communications*, pp. 1–6, Paris, France, May 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 1059231, 11 pages
https://doi.org/10.1155/2018/1059231

Research Article

A Novel Query Method for Spatial Data in Mobile Cloud Computing Environment

Guangsheng Chen,[1,2] Pei Nie,[1,2] and Weipeng Jing[1,2]

[1]*College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China*
[2]*Heilongjiang Province Engineering Technology Research Center for Forestry Ecological Big Data Storage and High Performance (Cloud) Computing, Harbin 150040, China*

Correspondence should be addressed to Pei Nie; 15546012870@163.com

Received 25 January 2018; Revised 5 April 2018; Accepted 17 April 2018; Published 17 May 2018

Academic Editor: Kai Yang

With the development of network communication, a 1000-fold increase in traffic demand from 4G to 5G, it is critical to provide efficient and fast spatial data access interface for applications in mobile environment. In view of the low I/O efficiency and high latency of existing methods, this paper presents a memory-based spatial data query method that uses the distributed memory file system Alluxio to store data and build a two-level index based on the Alluxio key-value structure; moreover, it aims to solve the problem of low efficiency of traditional method; according to the characteristics of Spark computing framework, a data input format for spatial data query is proposed, which can selectively read the file data and reduce the data I/O. The comparative experiments show that the memory-based file system Alluxio has better I/O performance than the disk file system; compared with the traditional distributed query method, the method we proposed reduces the retrieval time greatly.

1. Introduction

Under the background of the explosive growth of mobile data traffic and the emergence of various new business scenarios, the fifth-generation (5G) mobile communication network is proposed and becoming a hot topic in academical and industrial field. As a new generation of wireless mobile communication network, 5G is mainly used to meet the demand of mobile communication after 2020; driven by the rapid development of mobile Internet and growing demand for Internet of Things (IoT) services, 5G is required to have the features of low cost, low power consumption, and being safe and reliable [1, 2]; 5G will enable information and communication to exceed the time and space constraints, greatly shorten the distance between people and things, and quickly realize the interoperability of human and all things [3].

At present, the key technology of 5G network is still in the research phase; in addition to network architecture and transmission theory, mobile cloud computing is also one of the focuses of the future 5G network research [4]. Mobile cloud computing is a new model of delivery and usage of IT resources or information services; it is a product of cloud computing in the mobile Internet; mobile smart terminals in mobile networks are connected to remote service providers in an on-demand and scalable way to obtain the necessary resources, mainly including infrastructure, computing, storage capacity, and application resources [5, 6]. With the constant penetration of information technology into society, location-based services have been widely used in many fields such as military and transportation, for example, patient care in smart home and navigation service for mobile transportation; on the one hand, the positioning technology continues to evolve, providing increasingly accurate location information for mobile applications; on the other hand, with the increase in the number of users and the large increase in the number of mobile applications, the explosion of spatial data has brought tremendous pressure to upper-layer applications, especially for applications in mobile networks with huge traffic. Therefore, more and more researches combine mobile cloud computing with spatial data processing and use cloud platforms for data storage, calculation, indexing, and

querying to speed up processing and reduce response delay [7].

In this paper, we aim to provide a fast spatial data query interface in mobile cloud computing environment; for the reliable storage of massive spatial data, we first mesh the spatial data, fill the entire grid space by Hilbert curve, and then organize the data block using distributed memory file system Alluxio's KeyValueStore file and build a two-level index based on the file's internal index and file name. In order to provide real-time query, we use Spark, a distributed memory computing framework, to query for distributed data; in the meantime, we propose a Spark data input format based on the characteristic of Spark. The method eliminates disk I/O as much as possible and the entire query process from memory to memory and filtering through two layers. Experiments show that this method can well organize massive spatial data and has higher performance than traditional query methods.

The rest paper is organized as follows: Section 2 reviews the related work and in Section 3 we provide a background on spatial data queries and overview of Alluxio. Section 4 describes our proposed data indexing algorithm, storage structure, and parallel distributed retrieval algorithm. We discuss the experiment results in Section 5 and conclude in Section 6.

2. Related Work

Spatial data is a quantitative description of the geographical location of the world; based on computer technology, efficient use of spatial data in people's lives is of great significance. Spatial database emerged in the background of the rapid development of database and the rapid development of space application demand; it provides spatial data type definition interface and query language, which has very important pioneering significance in the early stage of GIS [8, 9]. However, with the rapid growth of spatial data, the traditional spatial database has been unable to meet the needs of real-time retrieval. Some works [10, 11] aimed at large-scale data query of spatial data and propose a parallel spatial database solution that distributes the data load and retrieval pressure of single computers to multiple servers. However, this method requires very expensive software licenses and dedicated hardware and requires complicated debugging and maintenance work [12].

As cloud computing has become a cost-effective and promising solution to computational and data-intensive issues, it is quite natural to integrate cloud computing into spatial data storage and processing. Wei et al. [13] applied the distributed NoSql database to the storage of spatial data and constructed efficient index to quickly retrieve spatial data. As MapReduce has become the standard solution for massively parallel data processing, more and more researchers apply Hadoop to GIS. Puri et al. [14] propose a MapReduce algorithm for distributed polygon retrieval based on Hadoop. Ji et al. [15] present a MapReduce-based method that constructs inverted grid index and processes k-NN query over massive spatial data sets. Hadoop-GIS [16] and Spatial-Hadoop [17] are two scalable and high-performance spatial data processing systems for running large-scale spatial queries in

Figure 1: Range query.

Hadoop. However, due to the limitations of MapReduce's own design, some researchers began to migrate spatial data processing to Spark. Wang et al. [18] use Spark for spatial range query; all the spatial data are stored in HDFS and propose a grid indexing method called Spark-Fat-Thin-Grid-Index. Cahsai et al. [19] propose a novel approach to process k-NN queries; this approach is based on a coordinator-based distributed query processing algorithm and uses Spark for data-parallel processing. At the system level, Yu et al. [20] introduce an in-memory cluster computing framework for processing large-scale spatial data, which efficiently executes spatial query processing algorithms and provides geometrical operations library that accesses Spatial RDDs to perform basic geometrical operations.

The above researches have eased the contradiction between the current rapidly increasing data set and real-time retrieval. However, there are still some performance bottlenecks in current approaches. First of all, the data is stored on disk, and the query is based on memory; the mismatch between memory processing speed and I/O speed restricts the performance. Secondly, the existing distributed search algorithm does not distinguish data strictly while readings, so many data that are not related to the query conditions are also read into the memory, and the query load of the calculation layer is increased. Therefore, in this paper, we store the data based on the memory file system and build a two-layer index structure to accelerate random access; in view of the lack of current work, Spark is used for parallel data processing and a data input format for spatial query is proposed, which filters out the irrelevant data with the query conditions based on the index structure.

3. Background

In this section, we provide the required background and preliminaries about the spatial data queries and a brief overview of distributed memory filesystem Alluxio.

3.1. Spatial Data Queries. For most applications, there are two common approaches to geospatial query. As shown in Figures 1 and 2, Figure 1 is range query [21]; given a set of data points P and a spatial range R, query aims to retrieve all spatial points within the given R boundary; the result of the range query is $\{P1, P2, P3, P4\}$. Figure 2 is K-NN query [22]; the nearest

Figure 2: K-NN query.

neighbor query is the most common query in geography; it is another kind of query method different from range query. It is used to find out the nearest neighbor in space from a given point; the nearest neighbor can be one or more; as shown in Figure 2, the given point is P marked by the red color and $k = 4$, 4-NN query here is to search for four points nearest to P, and the search result of this query is $\{P5, P6, P7, P8\}$.

3.2. Alluxio Overview. Alluxio (former Tachyon) is a new project that was unveiled by UC Berkeley AMPLab Labs in 2013 as the world's first memory-centric virtual distributed storage system that unifies data access and becomes the key to connecting computing infrastructure and underlying storage; Alluxio's memory-centric architecture enables data access faster than existing solutions. In essence, Alluxio is a distributed memory file system deployed on computing platforms such as Spark or MapReduce and storage platforms such as HDFS or S3 and by globally isolating computing platforms and storage platforms from alleviating the memory pressure in the computing framework, also given the ability to quickly read and write large amounts of data memory framework computing. Alluxio separates the power of memory storage from Spark/MapReduce, allowing Spark/MapReduce to focus more on computing itself for greater execution efficiency through finer partitioning [23, 24].

Alluxio architecture shown in Figure 3, which uses a standard Master-Worker mode, running Alluxio system, consists of a Master and multiple Workers and Alluxio Master support ZooKeeper for fault tolerance, used to manage the metadata of all files; it is also responsible for monitoring the status of individual Alluxio Workers. Each Alluxio Worker starts a daemon and manages a local Ramdisk, and Ramdisk stores specific file data. So far, Alluxio has been updated to 1.7.1.

4. Parallel Spatial Data Retrieval Based on Memory

In this section, because the spatial data indexing and storage are closely related to retrieval, we first introduce the spatial indexing algorithm in this paper and illustrate the data storage method and a two-level index structure in Alluxio. After that, we introduce the specific spatial data-parallel retrieval algorithm.

4.1. Indexing Spatial Data. Geographic space usually includes three kinds of vector data, namely, points, lines, and polygons, as shown in Figure 4. Faced with massive spatial data, users often care about some local information; therefore, how to index spatial data and respond quickly to user's requests is a key issue when storing. Common spatial indexing methods include gridding, KD Tree, R-tree, quad-tree index, etc. [25–28]; among them, grid index has the characteristics of simplicity and convenience; compared with other methods, the construction of grid index in Spark/MapReduce parallel system is less complicated. So in this paper we build the grid index for spatial data.

In order to construct the grid index, the geographic space needs to be divided into different grids; different data blocks contain the data intersecting with the geographic location of this grid. For uniquely identifying each data block and taking into account the spatial proximity of the data, the data block is coded by using the Hilbert curve [29, 30]; the original space area is block-coded as shown in Figure 5.

After meshing the geospatial space, we need to locate the vector data to a grid block and give its grid ID. As shown in Figure 5, the coordinates of lower-left and top-right corner of the space plane are (Lon0, Lat0) and (Lon1, Lat1), respectively; as parameters, the width and height of a grid are w and h, respectively, so the number of rows and the number of columns can be calculated as follows:

$$\begin{aligned} \text{cols} &= \frac{\text{Lon1} - \text{Lon0}}{w} \\ \text{rows} &= \frac{\text{Lat1} - \text{Lat0}}{h}. \end{aligned} \tag{1}$$

For each spatial data, we need to extract its location information, locate it to a specific grid, and then organize the data within the grid together to speed up the query based on the index. This paper abstracts all the spatial data to a point, and for polygon we have its inscribed center point as its position, and for the line we have its midpoint as its position; for vector point, its position point is itself. The point (x, y) is the spatial location of the data; according to the location points and cols and rows, we can get the col and row for any spatial data; finally use the Hilbert curve algorithm to get the grid ID.

$$\begin{aligned} \text{row} &= \frac{(y - \text{lat0})}{h} \quad 0 \le \text{row} \le \text{rows} \\ \text{col} &= \frac{(x - \text{lon0})}{w} \quad 0 \le \text{col} \le \text{cols} \\ \text{gridID} &= \text{HilbertCurve}\,(\text{row}, \text{col}) \\ &= \text{HilbertCurve}\left(\frac{y - \text{lat0}}{h}, \frac{x - \text{lon0}}{w}\right). \end{aligned} \tag{2}$$

After the above process, each spatial data has its grid ID; the data with the same ID are organized together to form a data table as shown in Table 1.

4.2. Data Storage Structure. Alluxio is a memory-based distributed file system that has many similarities with HDFS [31].

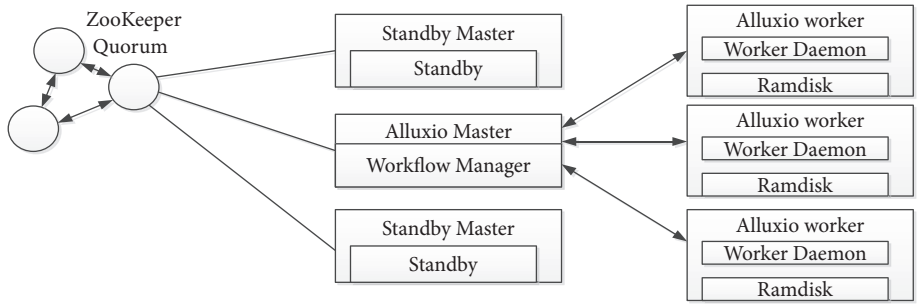


Figure 3: Alluxio architecture.

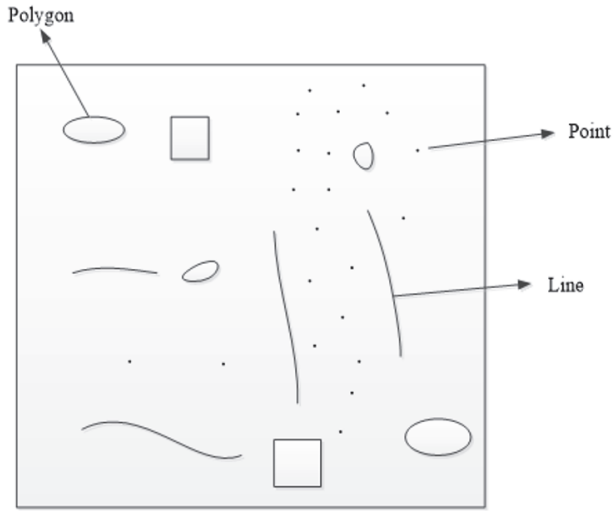


Figure 4: Space area.

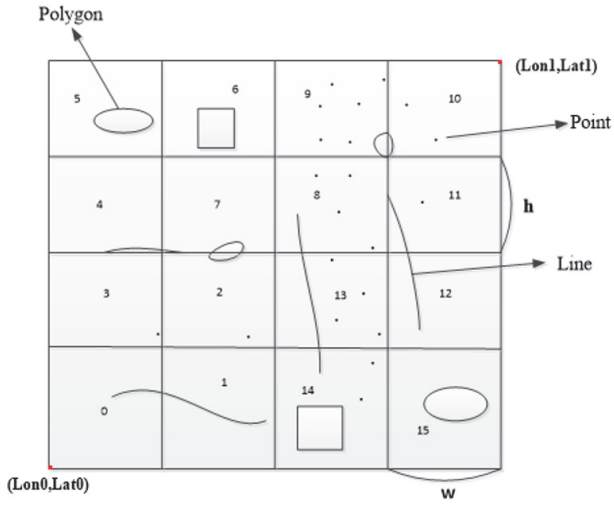


Figure 5: Grid index partitioning and coding.

Usually, data block tends to be smaller than file system block; if data blocks are directly stored on the file system, with the increase of data blocks, this will bring huge metadata storage pressure to the master node and small file problem [32]. Based on this problem, some researches use the key-value pair to structure data block together, solving the problem of small files [33, 34]. Therefore, this paper selected Alluxio built-in file structure KeyValueStore for data storage, the structure of the file in the form of key-value pairs, with the grid's Hilbert curve coding as the key; the corresponding data is stored as the value; KeyValueStore forces the key-value pairs to be written in ascending order because the internal index is built based on the key as the data is written.

Table 1: Data table.

Grid ID	Data List
1	Polygon 1, Line 2, Point 3, ...
2	Polygon 5, Polygon 7, ...
...	...
n	Polygon N, Line M, Point I, ...

This paper sets the size of each KeyValueStore file equal to the Alluxio block size, which is designed to optimize Alluxio storage and to avoid the problems of small files while improving the performance of file retrieval. Each filename contains the data range in the file, such as a file containing grid data with Hilbert code (0, 1, 2, 3); its filename is "0-3.kv," so that all the filename constitutes a global index. The internal index of all the KeyValueStore files and global index form a two-level index, as shown in Figure 6.

The shape of the vector data is irregular and there are many vector data in a single grid space. Therefore, it is very complex to determine the relationship between query conditions and data. So in this paper we introduce the thin-MBR and fat-MBR for the vector data in a single grid space [11]. Taking the data block with grid ID 14 as an example, the thin-MBR establishment process is as follows.

Step 1. Extract the center points of the vectors intersecting the space of the grid block; for the polygons, the center point is the incircle's center point; the line is its middle point, and the point is itself.

Step 2. Make the smallest circumscribed rectangle of all the vector center points in this grid area.

Step 3. The smallest circumscribed rectangle is the thin-MBR of vectors for the grid.

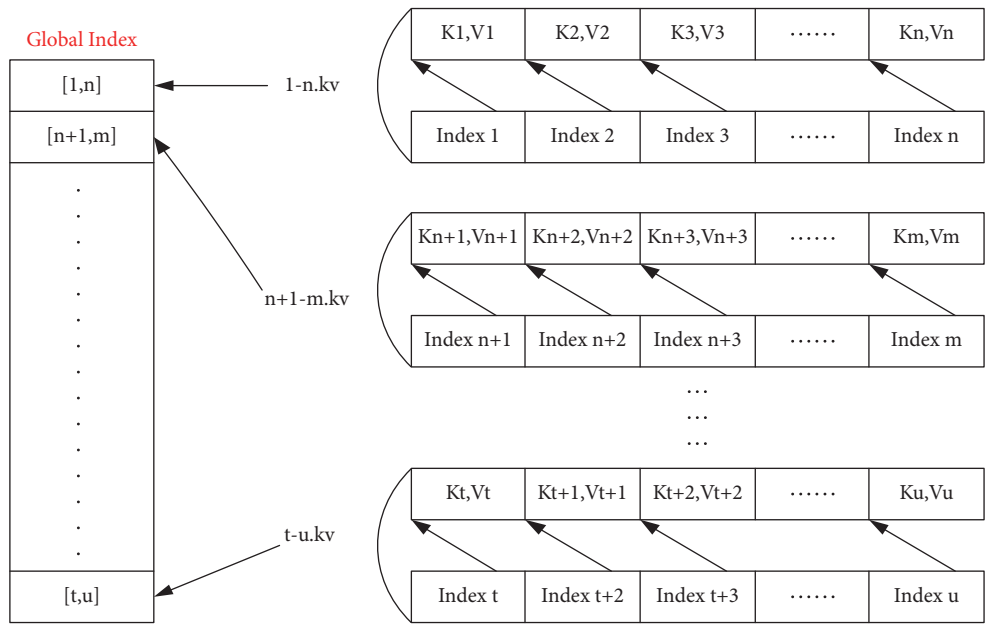


Figure 6: Data structure and two-level index.

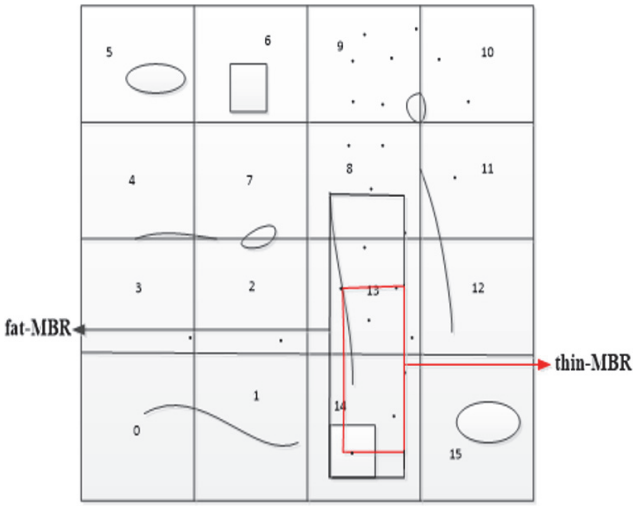


Figure 7: Fat-MBR and thin-MBR representation.

The establishment of fat-MBR is to traverse all the vectors in the space of this block; the minimum enclosing rectangle of all vectors is created, which is the fat-MBR of the space vector data of the grid block.

The thin-MBR and fat-MBR of grid 14 are shown in Figure 7; therefore, key-value pairs should contain more information, the grid code is used as the key, and the value includes the spatial data in the grid and the thin-MBR and the fat-MBR.

4.3. Spatial Data Retrieval. Based on the data-reading characteristics of Spark computing framework, this paper presents an input format called query-KVInputFormat for spatial data query, which filters the data when reading the memory filesystem. After the first layer of filtered data, each Spark task performs data query in memory, which is also called second-layer filtering; two layers of filtering are distributed; the first is to read distributed filesystem data, the second is that the parallelism tasks are handled distributedly.

4.3.1. The First Layer of Filtering. We cover two common distributed queries in Section 3, which can be retrieved after geospatial meshing and coding. For range queries, the range can be transformed into a set of grid numbers spanned by the range. For K-NN queries, record the grid number where the current position is located, such as y, and then record all the grids adjacent to y to form a set of grid numbers. As shown in Figure 8, the grid blocks intersecting the query range are (2, 7, 8, 13), current position is $P2$, and the K-NN query conditions are transformed into grid block group (6, 7, 8, 9, 10, 11). It can be seen that we have narrowed down the data range related to the spatial data query to a set of grid blocks so that we only need to read this set of grid blocks into memory for distributed retrieval and speed up data retrieval.

When interacting with MapReduce/Spark, Alluxio complies with the input and output formats of computational models. However, input formats of big data computing models are designed for batch processing and are read sequentially one by one; each task cannot be randomly read. Therefore, this paper designs a spatial-oriented query input format: query-KVInputFormat—this input format is based on KeyValueInputFormat; the implementation details are as follows.

Overrides protected List ⟨FileStatus⟩ listStatus (JobContext job): this method is used to filter out key-value files that do not contain data to be queried by using grid block groups, such as range queries that transform into grid block group (x, y), where $1 < x < n, n+1 < y < m$, building a global index in memory by traversing the filename, as shown in Figure 6 and

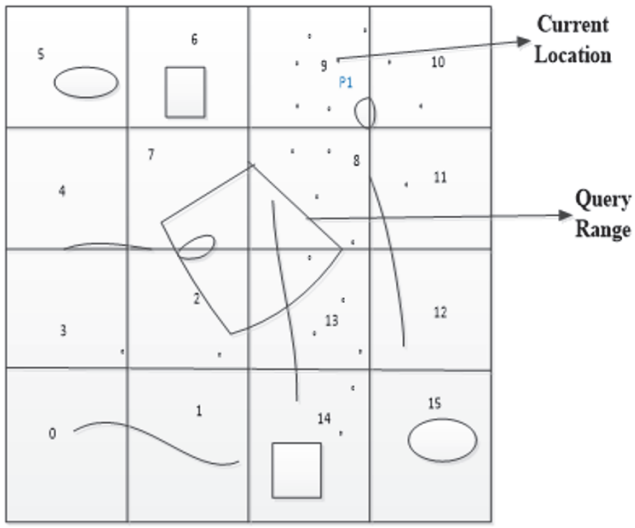


Figure 8: The grid block group representation.

retrieving the eligible files based on the global index, which returns $[1 - n.\text{kv}, n + 1 - m.\text{kv}]$.

Defined class query-KVRecordReader extends RecordReader $\langle K, V \rangle$; this class is internal class; the class defines the way of task read, in the traditional RecordReader class and its subclasses, and the order of each task to read records sequentially. But in the query we need to filter out some records; the sequential read method is no longer applicable, so rewrite the class; each task uses the block group as the query conditions to read assigned file (the file here is filtered after listStatus). Firstly, load the internal index of KeyValuetore into memory and use the query conditions for inquiries. If the record in the query condition exists in the file, the corresponding data block will be read out, if not, to the next record, the entire grid block group will be traversed. This process is concurrency on each task.

4.3.2. The Second Layer of Filtering

(1) Range Query. After the first level of retrieval, the data that meets the query is loaded into memory, forming RDDs, assigned to each task. For the range query, the data obtained through the first level query may be redundant; as shown in Figure 8, the query range is transformed into a grid block group, which reduces the complexity of the query but increases the scope of the query; data not in the range is also loaded into the memory. So in this layer we use Spark for further retrieval.

In a grid area, there are various types and styles of vector data; for the range to be queried, according to the conventional search method, traversing the vector data one by one in memory to determine whether there is intersection with the range, this method is quite consuming and unfavorable to extend. In this paper, data is filtered based on thin-MBR and fat-MBR; at first, whether the range contains the thin-MBR of the current grid region is judged; if yes, keep all the data in the area and jump to the next grid area; if no, judge the relationship between the range and the fat-MBR; if they do not intersect, discard the data in the area; otherwise, traverse the vector data in this area to determine the relationship between the vector and the range; if there is intersection, then retain this vector data; otherwise discard the vector. Combined with two filters, range query algorithm is shown in Algorithm 1.

```
Input: query range
Output: all data in the range
(0) Result set S = φ
(1) Calculate the grid block group corresponding to
    the query range
(2) for each task of spark do
(3)     Reading KeyValueStore to form RDDs based
        on the grid block group
(4) end for
(5) for each Partition of RDDs do
(6)     Determine the relationship between the
        query range and thin-MBR/fat-MBR
(7)     if the range contains the thin-MBR then
(8)         Add all data of this grid to S
(9)     else
(10)        if the range don't intersect with fat-MBR
then
(11)                Discard this grid
(12)        else
(13)                Traverse the data of this grid and
                    add eligible data to S
(14)        end if
(15)    end if
(16) end for
(17) return S
```

Algorithm 1: Range query.

(2) K-NN Query. At the first layer of filtering, the grid where query point lie and its bordering data block composed of the grid block group, each task reads the data into memory based on the grid block group, forming RDDs. For each task, we maintain a local queue, calculate the distance between each point of RDD-partitions and the query point, and add the nearest K points to the queue. After the task completes the calculation, the local queue is collected to the master node, and the master node recalculates and selects the nearest K vector points as the query results. K-NN query algorithm is shown in Algorithm 2.

5. Experimental Evaluation

In this section, we present the experimental evaluation of our retrieval algorithm. We divide this section into three parts; firstly, we introduce the dataset and the computing environment used in our experimental study; secondly, we evaluate and compare with existing solutions to see how the time is consumed by the query as the query area grows; after that, we compared Alluxio with HDFS. Finally, we evaluate our algorithms on various sizes of Spark cluster to measure the efficacy of our approach across various cluster sizes.

Input: query point
Output: the nearest K points
(0) Result set $\mathbf{S} = \phi$, local queue $\mathbf{Q} = \phi$
(1) Calculate ***the grid block group*** corresponding to the query point
(2) **for** *each Task* **do**
(3) Reading KeyValueStore to form RDDs based on ***the grid block group***
(4) **end for**
(5) **for** each *Partition of RDDs* **do**
(6) Calculate the distance between each point and the query point
(7) Add the nearest K points to **Q**
(8) **end for**
(9) **for** *each* **Q do**
(10) Collected to the master node
(11) **end for**
(12) Pick out the nearest K points add to **S** on master node
(13) **return S**

Algorithm 2: K-NN query.

5.1. Datasets and Experiment Environment. For the experimental dataset, we selected 43200 × 20880 global high-resolution images as the base map and a total of 11 layers' vector data of national borders, coastlines, ports, provinces, lakes, rivers, roads, and airports of all countries in the world as experimental data; the size of a single data grid is 512 × 512, encoded using Hilbert curve; the Alluxio block is set to 64 MB and the single KeyValueStore file is also 64 MB. We conducted our experiment on a cluster of 5 Inspur Yingxin I8000 blade servers, one of which served as the master node and the other four served as compute nodes. Each node was configured as a Xeon E5-2620 v2 6-core 2.10 GHz processor with 32 GB of memory and a 200 GB hard disk drive using Tenda TEG1024G Gigabit Ethernet switch, with Red-Hat 6.2 installed on each node and 2.6.32 Linux kernel, running Spark-1.5.0, Hadoop-2.5.2, and Alluxio-1.6.0.

5.2. Time Cost versus Query Size. We first demonstrate how the time costs grow as the size of the query area changes. In this experiment, we use all datasets as input, the cluster uses 4 compute nodes with 24 cores and 4 cores. For range queries, we randomly generated a polygon region for the query. For K-NN queries, we generate a pair of latitude and longitude coordinates using a random function to locate a point on the map as a query point. In order to compare our results with existing distributed techniques, we chose GeoSpark [17] as the benchmark. As shown in the previous section, we store the data in the Alluxio memory file system and use Spark to read the data for distributed data query.

Figures 9 and 10 show the relationship between the query time cost and the query size in 4 cores' environment, and Figures 11 and 12 show that in 24 cores' environment. For range queries, the size of the query range is measured in terms of acreage, and for K-NN queries, the number of query points is used for measuring. From the figure, we note that as the query size grows larger the time cost generally increases due to more data being processed. From the results, we also deduce that our algorithm is on average 1%–50% faster than the current technology in GeoSpark. As we discussed earlier, in the traditional approach data is stored in HDFS; a large amount of spatial data is directly loaded into the memory,

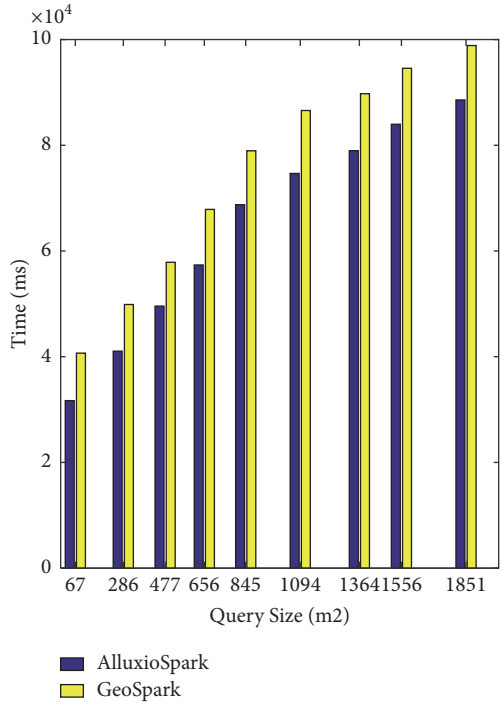


Figure 9: Range query execution time with 4 cores.

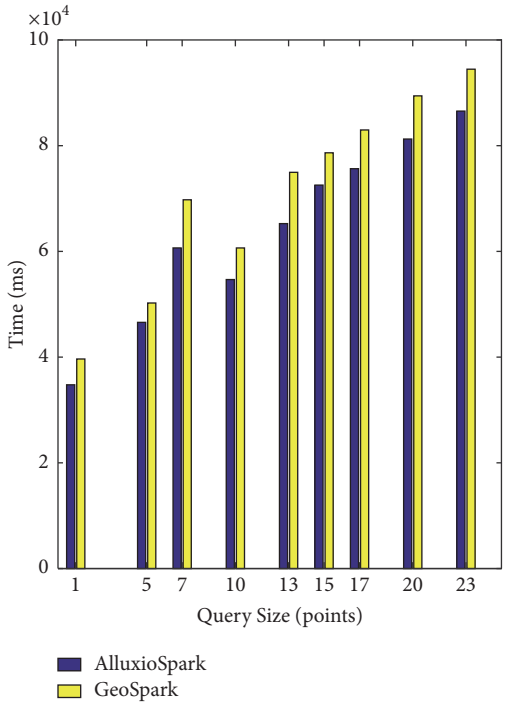


Figure 10: K-NN query execution time with 4 cores.

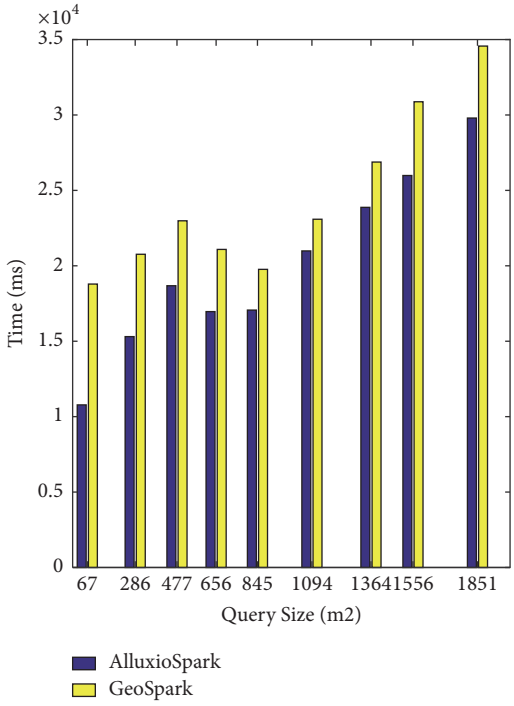


Figure 11: Range query execution time with 24 cores.

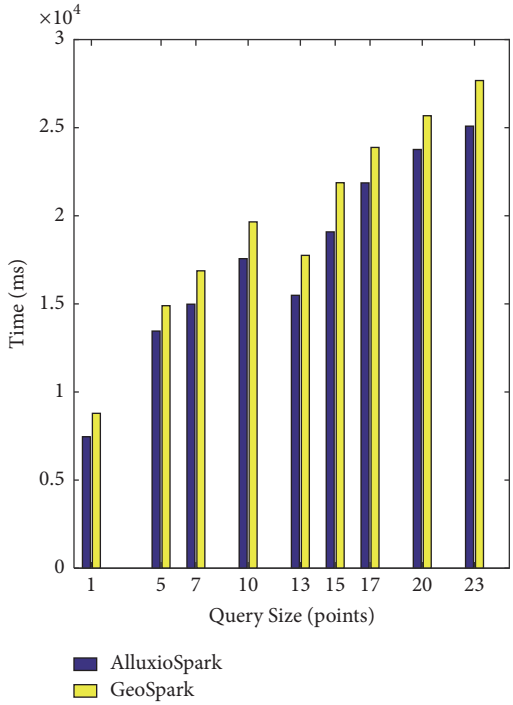


Figure 12: *K*-NN query execution time with 24 cores.

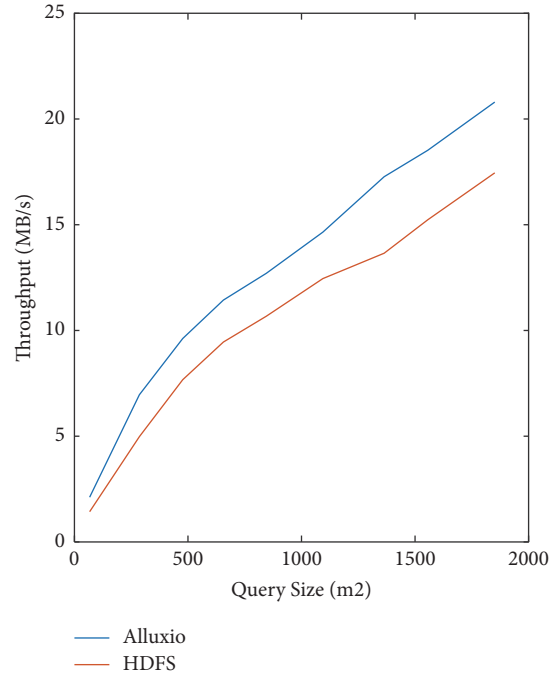


Figure 13: Throughput of range query with 4 cores.

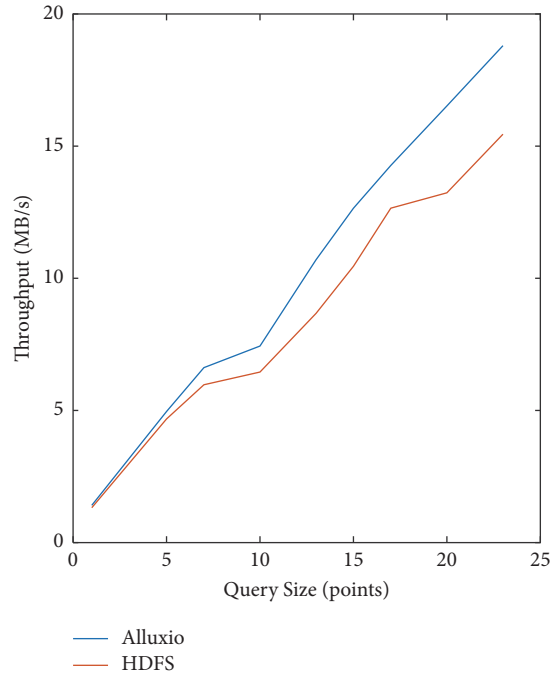


Figure 14: Throughput of *k*-NN query with 4 cores.

resulting in a large amount of data I/O and CPU load, but in the method proposed in this paper data is stored in memory; the entire retrieval process data flows from memory to memory, filtered while reading, making the irrelevant data read as little as possible into memory, reducing the CPU load and speeding up the query.

5.3. Alluxio versus HDFS. The underlying file system used by GeoSpark is HDFS, and Alluxio is used in this paper. We recorded the throughput of file system in experiment 5.2 to evaluate Alluxio's performance. Figures 13 and 14 show the throughput of Alluxio and HDFS in 4 cores' environment, and Figures 15 and 16 show that in 24 cores' environment. From these four figures, we can see that Alluxio has higher

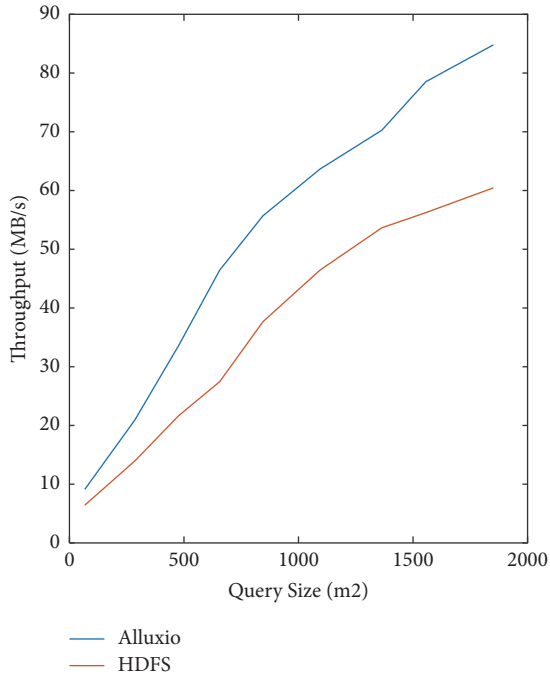


Figure 15: Throughput of range query with 24 cores.

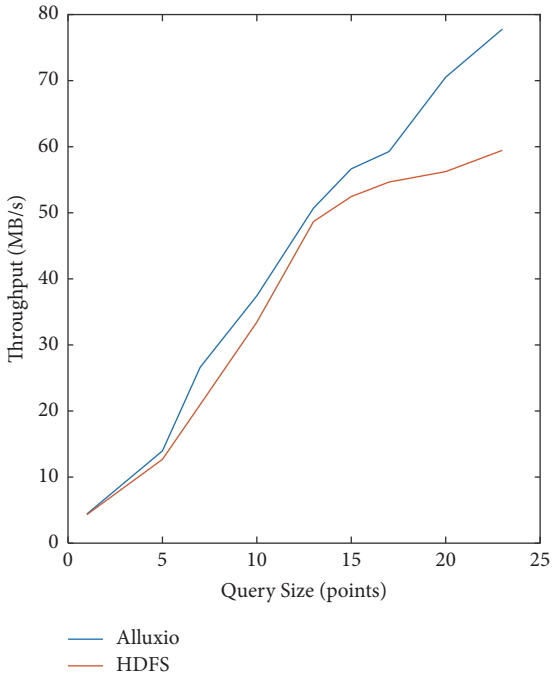


Figure 16: Throughput of k-NN query with 24 cores.

throughput than HDFS in the same experimental conditions, especially when it comes to range query; the main reason for this is that when executing a K-NN query, only the vector point is read into the memory, and the volume of vector point is small, so the difference between Alluxio and HDFS will not be obvious. And through this experimental result, we can see

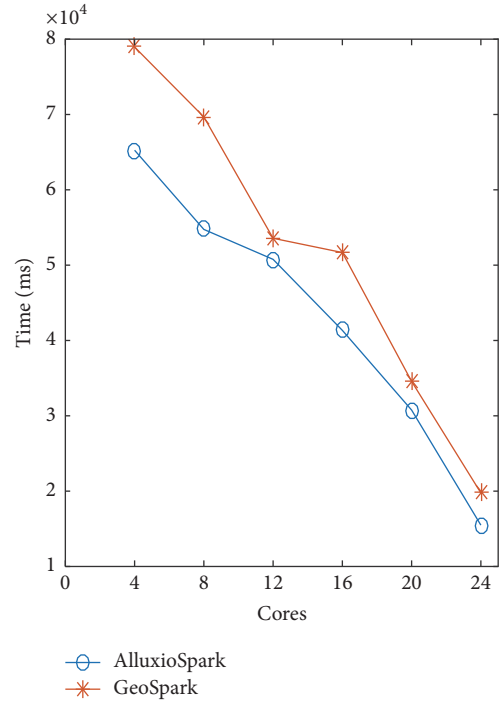


Figure 17: Impact of number of cores for range query.

that Alluxio has higher I/O rate than HDFS, which improves the overall job execution speed.

5.4. Time Cost versus Cluster Size. We next evaluate the effectiveness of our retrieval algorithm by varying the size of the Spark cluster in terms of the number of cores. For this experiment, we generated range query and K-NN query and used them to run queries on different cluster sizes. Figures 17 and 18 show the time cost on various cluster sizes when the range query size is 845 square meters and with K-NN query as 13-NN query. We infer from Figures 17 and 18 that the execution time decreases gradually as the cluster size becomes larger. Overall, we find that the proposed technique scales well with the number of nodes in the Spark cluster, showing a significant reduction in job execution time with increase in cluster size.

6. Conclusion and Future Work

With the development of mobile networks and the rapid growth of spatial data, traditional data storage and query models seem to be inadequate when dealing with massive spatial data. In this paper, the distributed memory file system Alluxio is used for data storage and indexing; at the same time, a big data input format for query of spatial data is proposed based on Spark computing framework, the entire retrieval process from memory to memory and read data selectively, reducing I/O load and CPU load. Through comparative experiments, the distributed retrieval method proposed in this paper has better query performance than the traditional method on the premise of efficient data

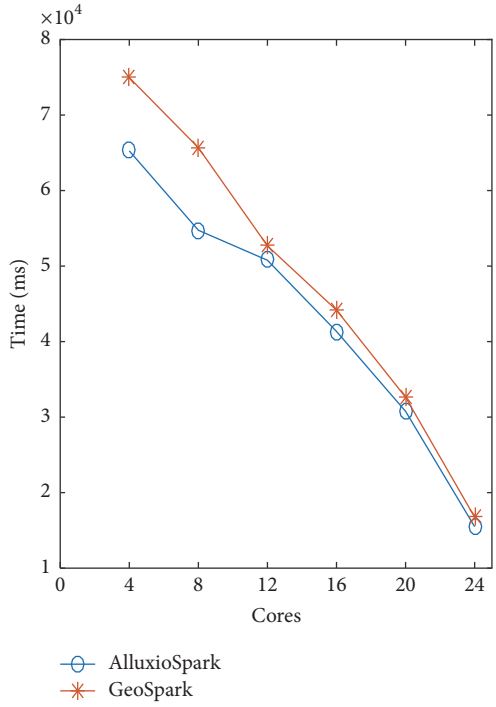


Figure 18: Impact of number of cores for k-NN query.

organization. The next step is to build more efficient spatial index and optimize Spark processing details to improve distributed query efficiency.

Data Availability

The spatial data (including global high-resolution remote sensing images and vector data of various layers) used to support the findings of this study comes from ENVI's own data set. Anyone can install ENVI and find experimental data in the "/data" folder.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported in part by National Natural Science Foundation of China (no. 31770768), Harbin Science and Technology Innovation Talent Research Project (no. 2014RFQXJ132), and China Forestry Nonprofit Industry Research Project (no. 201504307).

References

[1] Z. Wang, Z. Luo, and K. Wei, "5G Service Requirements and Progress on Technical Standards," *Zte Technology Journal*, vol. 20, no. 2, pp. 2–4, 25, 2014.

[2] 4G AMERICAS, 4G Americas' Recommendations on 5G Requirements and Solutions, white paper (EB/OL) 2014, http://www.4gamericas.org.

[3] IMT- 2020(5G) Promotion Group, 5G Vision and Requirements, white paper [EB/OL], 2014, http://www.IMT-2020.cn.

[4] X. Wang, G. Han, X. Du, and J. J. P. C. Rodrigues, "Mobile cloud computing in 5G: Emerging trends, issues, and challenges [Guest Editorial]," *IEEE Network*, vol. 29, no. 2, pp. 4-5, 2015.

[5] R. Kumar and S. Rajalakshmi, "Mobile cloud computing: Standard approach to protecting and securing of mobile cloud ecosystems," in *Proceedings of the 2013 International Conference on Computer Sciences and Applications, CSA 2013*, pp. 663–669, December 2013.

[6] N. Gupta and A. Agarwal, "Context aware mobile cloud computing: Review," in *Proceedings of the 2nd International Conference on Computing for Sustainable Global Development, INDIACom 2015*, pp. 1061–1065, March 2015.

[7] Y. Cui, J. Song, C.-C. Miao, and J. Tang, "Mobile Cloud Computing Research Progress and Trends," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 40, no. 2, pp. 273–295, 2017.

[8] R. H. Güting, "An introduction to spatial database systems," *The VLDB Journal*, vol. 3, no. 4, pp. 357–399, 1994.

[9] T. Devogele, C. Parent, and S. Spaccapietra, "On spatial database integration," *International Journal of Geographical Information Science*, vol. 12, no. 4, pp. 335–352, 1998.

[10] F. Wang, J. Kong, L. Cooper et al., "A data model and database for high-resolution pathology analytical image informatics," *Journal of Pathology Informatics*, vol. 2, no. 1, article 32, 2011.

[11] F. Wang, J. Kong, J. Gao et al., "A high-performance spatial database based approach for pathology imaging algorithm evaluation," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 5, 2013.

[12] A. Pavlo, E. Paulson, A. Rasin et al., "A comparison of approaches to large-scale data analysis," in *Proceedings of the International Conference on Management of Data and 28th Symposium on Principles of Database Systems, SIGMOD-PODS'09*, pp. 165–178, July 2009.

[13] L. Y. Wei, Y. T. Hsu, W. C. Peng et al., "Indexing spatial data in cloud data managements," *Pervasive & Mobile Computing*, vol. 15(C), pp. 48–61, 2014.

[14] S. Puri, D. Agarwal, X. He et al., *MapReduce Algorithms for GIS Polygonal Overlay Processing*, 2013.

[15] C. Ji, T. Dong, Y. Li et al., "Inverted Grid-Based kNN Query Processing with MapReduce," in *Proceedings of the 2012 Seventh ChinaGrid Annual Conference (ChinaGrid)*, pp. 25–32, Beijing, China, September 2012.

[16] A. Eldawy and M. F. Mokbel, "A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data," *VLDB Endowment*, pp. 1230–1233, 2013.

[17] A. Aji, F. Wang, H. Vo et al., "Hadoop gis: a high performance spatial data warehousing system over mapreduce," in *Proceedings of the VLDB Endowment*, pp. 1009–1020, 2013.

[18] F. Wang, X. Wang, W. Cui, X. Xiao, Y. Zhou, and J. Li, "Distributed retrieval for massive remote sensing image metadata on spark," in *Proceedings of the 36th IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016*, pp. 5909–5912, July 2016.

[19] A. Cahsai, N. Ntarmos, C. Anagnostopoulos, and P. Triantafillou, "Scaling k-Nearest Neighbours Queries (The Right Way)," in *Proceedings of the 37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017*, pp. 1419–1430, June 2017.

[20] J. Yu, J. Wu, and M. Sarwat, "GeoSpark: a cluster computing framework for processing large-scale spatial data," in *Proceedings of the Sigspatial International Conference on Advances in Geographic Information Systems*, vol. 70, ACM, 2015.

[21] L. Chen, Y. Tang, M. Lv, and G. Chen, "Partition-based range query for uncertain trajectories in road networks," *GeoInformatica*, vol. 19, no. 1, pp. 61–84, 2014.

[22] H. D. Chon, D. Agrawal, and A. E. Abbadi, "Range and k NN query processing for moving objects in grid model," *Mobile Networks & Applications*, vol. 8, no. 4, pp. 401–412, 2003.

[23] H. Li, A. Ghodsi, M. Zaharia, S. Shenker, and I. Stoica, "Tachyon: Reliable, memory speed storage for cluster computing frameworks," in *Proceedings of the 5th ACM Symposium on Cloud Computing, SOCC 2014*, November 2014.

[24] Z. Li, Y. Yan, J. Mo, Z. Wen, and J. Wu, "Performance Optimization of In-Memory File System in Distributed Storage System," in *Proceedings of the 2017 International Conference on Networking, Architecture, and Storage (NAS)*, Shenzhen, China, August 2017.

[25] D. Šidlauskas, S. Šaltenis, C. W. Christiansen, J. M. Johansen, and D. Šaulys, "Trees or grids? Indexing moving objects in main memory," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2009*, pp. 236–245, November 2009.

[26] A. Guttman, "R-trees: a dynamic index structure for spatial searching," *ACM SIGMOD Record*, vol. 14, no. 2, pp. 47–57, 1984.

[27] Z. Liu, "A k-d tree-based algorithm to parallelize Kriging interpolation of big spatial data," *Giscience & Remote Sensing*, vol. 52, no. 1, pp. 40–57, 2015.

[28] M. Demirbas and X. Lu, "Distributed Quad-Tree for Spatial Querying in Wireless Sensor Networks," in *Proceedings of the IEEE International Conference on Communications*, pp. 3325–3332, IEEE, 2007.

[29] J. McVay, N. Engheta, and A. Hoorfar, "High Impedance Metamaterial Surfaces Using Hilbert-Curve Inclusions," *IEEE Microwave and Wireless Components Letters*, vol. 14, no. 3, pp. 130–132, 2004.

[30] T. Su, W. Wang, Z. Lv, W. Wu, and X. Li, "Rapid Delaunay triangulation for randomly distributed point cloud data using adaptive Hilbert curve," *Computers and Graphics*, vol. 54, article no. 2631, pp. 65–74, 2016.

[31] F. Azzedin, "Towards a scalable HDFS architecture," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, pp. 155–161, May 2013.

[32] X. Li, B. Dong, L. Xiao, L. Ruan, and Y. Ding, "Small files problem in parallel file system," in *Proceedings of the 2011 International Conference on Network Computing and Information Security, NCIS 2011*, pp. 227–232, May 2011.

[33] B. Dong, Q. Zheng, F. Tian, K.-M. Chao, R. Ma, and R. Anane, "An optimized approach for storing and accessing small files on cloud storage," *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1847–1862, 2012.

[34] Z. Chi, F. Zhang, Z. Du, and R. Liu, "A distributed storage method of remote sensing data based on image blocks organization," *Journal of Zhejiang University, Science Edition*, vol. 41, no. 1, pp. 95–112, 2014.