SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

August 2022, Vol. 65 182301:1–182301:18
https://doi.org/10.1007/s11432-021-3328-y

Active device detection and performance analysis of massive non-orthogonal transmissions in cellular Internet of Things

Donghong CAI[1,2], Pingzhi FAN[2], Qiuyun ZOU[3*], Yanqing XU[4,5], Zhiguo DING[6] & Zhiquan LIU[1]

[1]*College of Cyber Security, Jinan University, Guangzhou* 510632*, China;*
[2]*Institute of Mobile Communications, Southwest Jiaotong University, Chengdu* 610031*, China;*
[3]*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;*
[4]*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen* 518172*, China;*
[5]*The University of Science and Technology of China, Hefei* 230026*, China;*
[6]*School of Electrical and Electronic Engineering, The University of Manchester, Manchester* M13 9PL*, UK*

Received 6 March 2021/Revised 14 June 2021/Accepted 23 August 2021/Published online 25 July 2022

Abstract This paper investigates multiple access schemes for uplink and downlink transmissions in cellular networks with massive Internet of Things (IoT) devices. Recall that single-carrier frequency division multiple access and orthogonal frequency division multiple access, which are orthogonal multiple access (OMA) schemes, have been conventionally adopted for uplink and downlink transmissions in narrow-band IoT, respectively. Unlike these OMA schemes, we propose two non-orthogonal multiple access (NOMA) schemes for cellular IoT with short-packet transmissions. Especially, a generalized expectation consistent signal recovery-based algorithm is proposed to estimate active devices, channel state information and data in uplink transmission, where all of the active devices are allowed to transmit their pilots and data through the same resource block without authorization. On the other hand, the active devices estimated during uplink transmission are grouped for downlink transmission with a trade-off between performance and detection complexity. Additionally, the data error rates are analysed for both uplink and downlink transmissions with low-resolution analog-to-digital converters (ADCs), where the effects of critical parameters such as the estimation error, ADC bits, packet length, and message bits are revealed. Both simulation and analytical results are provided to demonstrate the excellent performance of the proposed NOMA schemes and algorithms, especially for active device, channel, and data estimations. More importantly, the obtained results show that the data error rate performance of downlink NOMA is superior to that of OMA when the message bits of devices in one group are selected following the proposed strategy.

Keywords massive connections, non-orthogonal multiple access, active device detection, short-packet.

Citation Cai D H, Fan P Z, Zou Q Y, et al. Active device detection and performance analysis of massive non-orthogonal transmissions in cellular Internet of Things. Sci China Inf Sci, 2022, 65(8): 182301, https://doi.org/10.1007/s11432-021-3328-y

1 Introduction

Machine-type communication (mMTC) has been considered a representative service category in 5G networks [1–3] because of its wide applications of the Internet of Things (IoT) such as smart city, smart health care, factory automation, and autonomous driving [1, 4]. Notably, the number of IoT devices is growing exponentially and will reach hundreds of billions in 2030 [5]. A ket to enhance connection density is to provide device access over a large range. The cellular technique is one of the main access techniques for IoT [5, 6]. In narrow-band IoT, single-carrier frequency division multiple access (SC-FDMA) and orthogonal frequency division multiple access (OFDMA) have been adopted in uplink and downlink transmissions, respectively, which are based on the conventional granted orthogonal multiple

* Corresponding author (email: qiuyun.zou@bupt.edu.cn)

access (OMA) scheme [7], and orthogonal time/frequency resources are allocated to different devices. Therefore, the conventional scheme cannot support the massive connections required of mMTC networks owing to the high probability of collisions among devices. In addition, the significant signalling overhead and excessive latency caused by complicated scheduling procedures are inefficient to send sporadic short packets of IoT devices.

Recently, grant-free non-orthogonal multiple access (NOMA) schemes have been considered a compelling alternative [8–12]. In grant-free NOMA schemes, devices transmit short data packets through the same time/frequency resource without a granting procedure. Consequently, grant-free NOMA has an excellent performance in terms of resource utilisation and latency/signalling overhead. Especially, the active device and channel estimations in the uplink transmission are formulated as a sparse signal recovery problem for grant-free NOMA schemes. Various compressed sensing-based algorithms have been proposed to solve this problem [8,10,13–18]. In [18], approximate message passing is used to detect active devices and estimate their channel state information (CSI) based on Gaussian pilot sequences. To maintain orthogonality and mitigate the inter-user interference, Ref. [19] proposed an active device detection algorithm with Zadoff-Chu (ZC) sequences. To further improve active device detection and channel estimation performance over existing algorithms, the authors proposed an expectation propagation-based technique in [8]. However, these existing algorithms are only suitable for linear measurements corrupted by additive white Gaussian noise (AWGN) [20]. Therefore, Refs. [21,22] proposed message passing based active device detection algorithms for multiple antennas case, which allow the measurement at the receiver in arbitrary form and shows that increasing the number of receiving antennas reduces the pilot cost.

Although grant-free NOMA provides massive low latency connections and multiple receiving antennas reduce the pilot length, there still exist different time offsets since signals from different devices arrive at the receiver asynchronously [23]. This is because devices are geographically distributed and signals from different devices begin to propagate at any time. Hence, we have to develop an effective grant-free NOMA scheme with time offsets and improve the algorithms for activity, CSI, and data estimations. On the other hand, to achieve the prominent spectral performance of multiple access, short data packet transmission, and detection have been designed in [8, 10, 24–27]. For example, in [10], the authors proposed a greedy algorithm to perform active device detection, channel estimation, and data decoding jointly for uplink transmission. To reduce the system cost, a joint channel-and-data estimation method based on Bayes-optimal inference has been proposed for the quantized uplink systems [24]. However, existing research on cellular IoT mainly focuses on active device detection and channel estimation for uplink transmission. One of the primary issues is to reveal impacts of the packet length, error propagation, and low-resolution analog-to-digital converters (ADCs) on system performance.

In addition, the downlink NOMA scheme design for cellular IoT is another important issue due to the limited computing capacity of IoT devices. In [28], a backscatter-NOMA is proposed for downlink transmission of cellular IoT. The outage probabilities and ergodic rates are analysed. Furthermore, the authors in [25] designed a three-phase transmission protocol operated in time division duplex (TDD) mode, including the training phase, uplink data transmission phase, and downlink data transmission phase. The closed-form expressions of uplink and downlink individual achievable rates are derived, and the pilot length and data packet length are optimized under the rate constraint. However, imperfect CSI might lead to a significant performance degradation of NOMA. It is impractical to assume that the active devices and the CSI are estimated perfectly and the achievable rate of a short data packet is formulated using the Shannon formula.

In this paper, we design the NOMA schemes for uplink and downlink cellular IoT with short-packet transmissions and low-resolution ADC receivers. In uplink transmission, a grant-free access strategy is adopted. Different from [8, 25], a generalized expectation consistent signal recovery (GEC-SR) based algorithm is proposed to detect active devices, estimate their CSI, and detect their signals from the received quantified signal. On the other hand, we propose a hybrid NOMA scheme for downlink transmission. Especially, the active devices detected in uplink transmission are first grouped for the performance and complexity trade-off. The downlink CSI is then estimated in each group by allocated orthogonal time slots to the devices in the training phase. In the data transmission phase, power domain NOMA is used within one group, and orthogonal resources are allocated to different groups. In order to further reveal the system performance with short packet transmission, we use the finite block-length coding (FBC) theory [29,30] to formulate the achievable rate of each device. The contributions of this paper are summarised as follows.

• Two short-packet non-orthogonal transmission schemes are proposed for uplink and downlink transmissions in cellular IoT. Especially, the potential active devices access to the network without authorisation and the message bits of active devices are modulated as short packets. Low-cost algorithms based on the GEC-SR algorithm are proposed to estimate active devices, channels, and data from uplink receiving ADC quantified signals. While the detected active devices are grouped for downlink transmission based on uplink estimations, where a trade-off between performance and complexity is revealed.

• With the imperfect estimated CSI, the bit error rate (BER) performance of uplink non-orthogonal short-packet transmission is obtained using a GEC-SR based linear minimum mean square error (LMMSE) detector, which shows the impact of error propagation caused by the active device detection on uplink data detection. In addition, the average block error rate (BLER) of downlink NOMA with low-resolution ADCs is derived in approximated closed-form, which quantifies the effect of channel estimation errors on system performance.

• By investigating the impact of message bits of downlink transmission based on the analyzed average BLER for a given block-length, we obtain a trade-off between the reliability and the effectiveness. Particularly, a device pairing strategy based on message bits is proposed to guarantee NOMA performance compared with OMA. Simulation results demonstrate the accuracy of the active device detection, channel estimation, and obtained analytical results. More importantly, the obtained results show that the BLER performance of downlink NOMA is superior to that of OMA when message bits are selected according to the proposed strategy.

The remainder of this paper is organized as follows. Section 2 describes the uplink and downlink short-packet non-orthogonal transmissions in large scale cellular IoT. A low-complexity active device detection based on GEC-SR and the BER of uplink data with the LMMSE detector are presented in Section 3, whereas the downlink channel estimation and the performance analysis of short-packet transmission are investigated in Section 4. In Section 5, numerical results and simulations are applied to verify the performance of proposed algorithms and developed analysis. Finally, Section 6 concludes the paper.

Notations. The identity matrix, the all-one vector, and the all-zero vector of size M are denoted as $\boldsymbol{I}_M$, $\mathbf{1}_M$, and $\mathbf{0}_M$, respectively. The distribution of a circularly symmetric complex (or real) Gaussian random vector $\boldsymbol{x}$ with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V}$ is denoted as $\mathcal{N}_c(\boldsymbol{x};\boldsymbol{m},\boldsymbol{V})$ (or $\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{V})$). $\odot$ denotes componentwise multiply and $\oslash$ denotes componentwise divide.

2 System model

We consider a single-cellular mMTC network, in which one single-antenna central base-station (BS) and N single-antenna IoT devices located in the cellular. A block fading channel model with L symbol durations is considered. To meet the massive connections of devices, different NOMA schemes are designed for uplink and downlink transmission in cellular networks. On one hand, a grant-free NOMA scheme is employed for uplink transmission, where unknown active devices are allowed to access the network without authorization or scheduling. On the other hand, the detected active devices are paired/grouped for downlink transmission due to the limited capacity of the IoT devices. A hybrid NOMA scheme is proposed to serve each group of active users. Specially, the detected active devices are grouped based on uplink estimations. Orthogonal time slots are allocated to all active devices for channel estimation in the downlink training phase. The power domain NOMA scheme is used within each group for information transmission and orthogonal bandwidth resources are employed among different devices groups [31]. The details of theses two NOMA schemes for uplink and downlink of the cellular network are shown as Figure 1 and described in the following two subsections, respectively.

2.1 Uplink NOMA scheme

In uplink grant-free NOMA scheme, as shown in Figure 2, the transmission occurs in two phases and it shows the time-domain structure of the received signal from multiple asynchronous devices. This asynchronous communication may occur since the signals from different users arrive at the receiver asynchronously. Without loss of generality, it is assumed that the signal transmitted by device 1 arrives at the BS prior to that transmitted by device $\bar{n}$ ($\bar{n} = 2, \ldots, N$) by a time offset $\Delta_{\bar{n}}$, $\Delta_{\bar{n}} > 0$. Note that only a small fraction of potential devices are active and send their small packets sporadically in IoT [18]. In order to mitigate the asynchronous effect, a guard space is used to prevent the interference between pilot and data symbols [23]. For example, the cyclic prefix (CP) can be utilized as the protection prefix

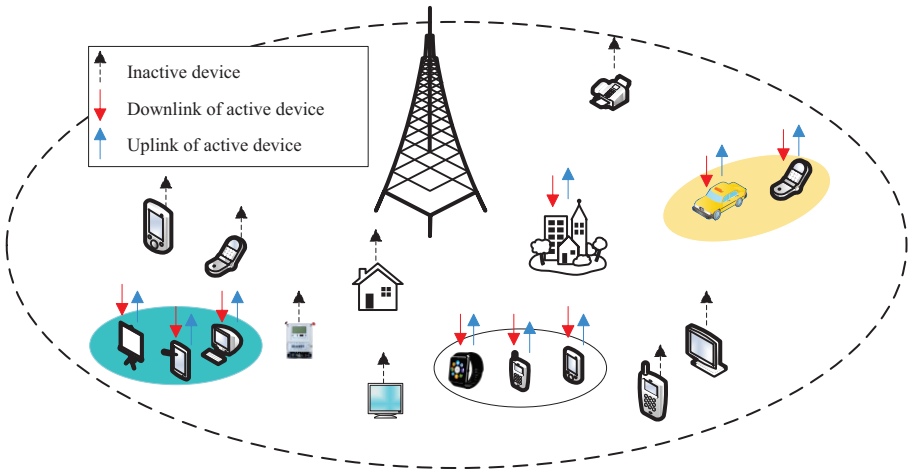

Figure 1 (Color online) Sporadic traffic scenario in cellular IoT.

if the non-orthogonal transmission is implemented on an orthogonal OFDMA tones [32]. We first detect active devices and estimate their CSI in the training phase. The BERs of active devices' data packets are then obtained in the uplink data transmission phase.

2.1.1 *Uplink channel training phase*

The BS assigns a pilot sequence $\boldsymbol{a}_n = (a_{n,1}, \ldots, a_{n,L^p})^{\mathrm{T}} \in \mathbb{C}^{L^p}$ to device n $(n = 1, 2, \ldots, N)$ in advance, where $L^p < N$ and $a_{n,l}$ is independently chosen from $\{-1, 1\}$ with equal probability. It is assumed that only a small portion of devices are active, and we define the device activity indicator as

$$\varpi_n = \begin{cases} 1, & \text{device } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

for $n \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$. Each device decides whether or not to access the channel with probability ϵ in an independent manner [18]. Then, ϖ_n can be modeled as a Bernoulli random variable such that $\Pr(\varpi_n = 1) = \epsilon$, $\Pr(\varpi_n = 0) = 1 - \epsilon$, $\forall n$. Let T_p, T_d, and T_g represent the pilot transmission duration, data transmission duration, and the guard interval duration, respectively. The receiver removes the data of the guard position, and selects the remaining L^p as the received signals in the training phase. So there are two types of observation windows. As shown in Figure 2, the part intercepted by the receiver in type I observation window can not completely contain the pilot sequences of active devices. Then the inter-symbol interference (ISI) will be introduced. Instead, type II completely contains the pilot sequences of active devices and will not effected by data symbols. Therefore, the ISI can be perfectly eliminated if the length of guard position is long enough for asynchronous time offset. The received signal at the BS for active device and channel estimations is

$$\boldsymbol{y}^{\mathrm{u},p} = \sum_{n \in \mathcal{N}} \varpi_n \boldsymbol{a}_n h_n^{\mathrm{u}} + \boldsymbol{w}^p \triangleq \boldsymbol{A}\boldsymbol{x}^p + \boldsymbol{w}^{\mathrm{u},p}, \tag{2}$$

where $\boldsymbol{y}^{\mathrm{u},p} = (y_1^{\mathrm{u},p}, \ldots, y_{L^p}^{\mathrm{u},p})^{\mathrm{T}} \in \mathbb{C}^{L^p}$, $\boldsymbol{w}^{\mathrm{u},p} = (w_1^{\mathrm{u},p}, \ldots, w_{L^p}^{\mathrm{u},p}) \in \mathbb{C}^{L^p}$ with $w_l^{\mathrm{u},p}$ $(l = 1, 2, \ldots, L^p)$, is the independent AWGN following zero mean and σ_p^2 variance complex Gaussian distribution[1)], $\boldsymbol{A} \triangleq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N] \in \mathbb{C}^{L^p \times N}$ with $|\boldsymbol{a}_n|^2 = 1$ is the collection of pilot sequences of all devices, h_n^{u} is the channel between device n and the BS, and $\boldsymbol{x}^p = (x_1^p, \ldots, x_N^p)^{\mathrm{T}} \in \mathbb{C}^N$ with $x_n^p \triangleq \varpi_n h_n^{\mathrm{u}}$. We define h_n^{u} as $h_n^{\mathrm{u}} = \frac{g_n^{\mathrm{u}}}{\sqrt{1+d_n^{\zeta}}}$, where $g_n^{\mathrm{u}} \sim \mathcal{N}_c(0, 1)$, d_n is the distance and ζ denotes the path loss factor. Then we have $h_n^{\mathrm{u}} \sim \mathcal{N}_c(0, \lambda_n)$ with $\lambda_n = \frac{1}{1+d_n^{\zeta}}$.

1) The variances of zero mean AWGNs of pilot and data transmission phases in this paper are assumed to be σ_p^2 and σ_0^2, respectively.

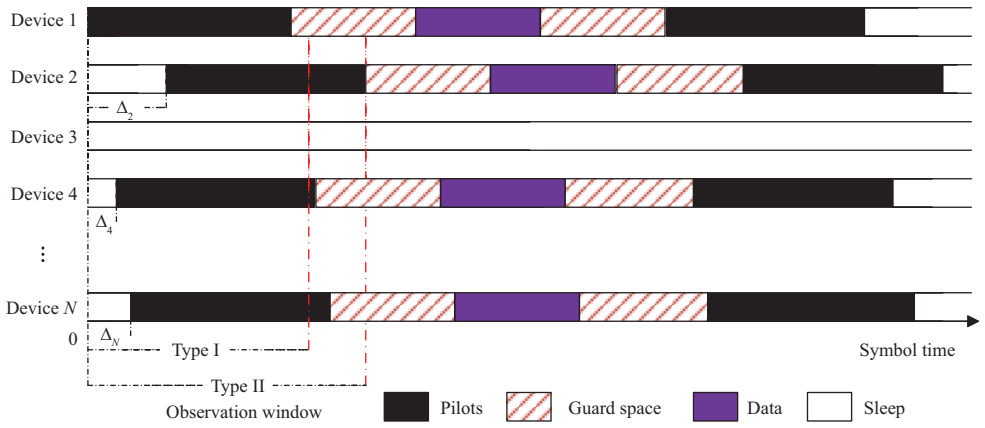

Figure 2 (Color online) Time-domain structure of the received signals for uplink NOMA scheme.

2.1.2 *Uplink data non-orthogonal transmission phase*

Device n's uplink message bits are modulated as x_n^{u} chosen from the M-order constellation $\mathcal{A}$ when it is active; otherwise, we use $x_n^{\mathrm{u}} = 0$ to represent that the device n is inactive. With spreading sequence, $\boldsymbol{s}_n = (s_{n,1}, \ldots, s_{n,L^{\mathrm{u}}})^{\mathrm{T}}, s_{l^{\mathrm{u}}} \sim \mathcal{N}_c(0, 1/L^{\mathrm{u}})$, $L^{\mathrm{u}} = L - L^p$, the received uplink data at the BS is

$$\boldsymbol{y}^{\mathrm{u}} = \boldsymbol{S}\mathrm{Diag}(\hat{\boldsymbol{h}}^{\mathrm{u}})\boldsymbol{x}^{\mathrm{u}} + \boldsymbol{w}^{\mathrm{u}}, \tag{3}$$

where $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{C}^{L^{\mathrm{u}} \times N}$, $\hat{\boldsymbol{h}}^{\mathrm{u}} = (\hat{h}_1^{\mathrm{u}}, \ldots, \hat{h}_N^{\mathrm{u}})^{\mathrm{T}}$, $\boldsymbol{x}^{\mathrm{u}} = (x_1^{\mathrm{u}}, \ldots, x_N^{\mathrm{u}})^{\mathrm{T}}$, and $\boldsymbol{w}^{\mathrm{u}} = (w_1^{\mathrm{u}}, \ldots, w_{L^{\mathrm{u}}}^{\mathrm{u}})^{\mathrm{T}} \in \mathbb{C}^{L^{\mathrm{u}}}$, $w_l^{\mathrm{u}} \sim \mathcal{N}_c(0, \sigma_0^2)$ is an AWGN vector in uplink data transmission phase. In order to reduce the cost, the received signal in (2) is quantized by a uniform complex-valued ADC quantization Φ_c [33] with B bits and step size Δ_τ, therefore the quantized signal is

$$\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p} \triangleq \Phi_c(\boldsymbol{y}^{\mathrm{u},p}), \tag{4}$$

where $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}^p$. The B-bit uniform ADC quantizer with 2^B bins is characterized by a set of $2^B - 1$ thresholds $\Pi := [\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{2^B-1}] \in \mathbb{R}^{2^B-1}$ with $\hat{\tau}_{b_0} = (-2^{B-1} + b_0)\Delta_\tau$, $b_0 = 1, 2, \ldots, 2^B - 1$, such that $-\infty \triangleq \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_{2^B-1} < \hat{\tau}_{2^B} \triangleq \infty$. The quantized output is $\Delta_\tau(-2^{B-1} + b - \frac{1}{2})$, $b = 1, 2, \ldots, 2^B$, defined by the interval $(\hat{\tau}_{b-1}, \hat{\tau}_b]$ if the input falls in the b-th bin.

Similarly, the quantized output signal of (3) is

$$\boldsymbol{y}_{\Phi_c}^{\mathrm{u}} \triangleq \Phi_c(\boldsymbol{y}^{\mathrm{u}}) = \Phi_c(\boldsymbol{S}\mathrm{Diag}(\hat{\boldsymbol{h}}^{\mathrm{u}})\boldsymbol{x}^{\mathrm{u}} + \boldsymbol{w}^{\mathrm{u}}). \tag{5}$$

2.2 Downlink NOMA scheme

For downlink transmission, we propose a hybrid NOMA scheme to serve the active devices detected in the uplink, where the active devices are grouped due to the limited capacity of the IoT devices as well as the trade-off between the performance and the complexity. Here, we first focus on two-device case and the results can be easily extended to a general case in Subsection 4.2. Different from the uplink NOMA scheme, a hybrid NOMA scheme is considered in downlink transmission [31]. Especially, the power domain NOMA scheme is used within each group for data transmission and orthogonal bandwidth resources are employed among different groups. Besides, orthogonal time slots are allocated to the active devices of one group for downlink channel estimations before non-orthogonal data transmission.

2.2.1 *Downlink channel training phase*

In the downlink NOMA scheme, each active device i $(i = u, v)$ in one group should estimate its channel response before signal detection. During the channel estimation phase, the BS sends a special pilot sequence[2)], $\boldsymbol{\phi}_i \in \mathbb{C}^{L^q}$ $(2L^q < L)$, to device i. The received training signal at the i-th device is

$$\boldsymbol{y}_i^{\mathrm{d},p} = \boldsymbol{\phi}_i h_i^{\mathrm{d}} + \boldsymbol{w}_i^{\mathrm{d},p}, \tag{6}$$

2) For the general case, the sum of the length of the pilot sequence of all devices in one group is not large than the coherence symbol durations.

where $\boldsymbol{w}_i^{\mathrm{d},p} \in \mathbb{C}^{L^q}$ denotes the AWGN vector at the i-th device during the downlink channel estimation phase. $\boldsymbol{y}_i^{\mathrm{d},p} \in \mathbb{C}^{L^q}$, and h_i^{d} is the channel between device i and the BS. Due to the ADC at receiver, the quantized $\boldsymbol{y}_i^{\mathrm{d},p}$ is

$$\boldsymbol{y}_{i,\Phi_c}^{\mathrm{d},p} = \Phi_c(\boldsymbol{y}_i^{\mathrm{d},p}) = \Phi_c(\boldsymbol{\phi}_i h_i^{\mathrm{d}} + \boldsymbol{w}_i^{\mathrm{d},p}). \tag{7}$$

2.2.2 *Downlink data transmission phase*

The message bits B_i of device i are encoded as a unit-power codeword, x_i^{d}, with block-length $L_i^{\mathrm{d}} = L - 2L^q$. For the fairness and the superposition coding with successive interference cancellation (SIC) in downlink NOMA, we assume that the transmit powers at the BS are sorted as $p_v \leqslant p_u$ based on the uplink estimated channel gain[3] $|\hat{h}_v^{\mathrm{u}}|^2 \geqslant |\hat{h}_u^{\mathrm{u}}|^2$ and $L_i^{\mathrm{d}} = L^{\mathrm{d}}$. Then the superposition codeword of one pair of devices at the BS is $x_s^{\mathrm{d}} = \sqrt{p_u}x_u^{\mathrm{d}} + \sqrt{p_v}x_v^{\mathrm{d}}$, where p_j is the transmit power of device i, x_i^{d} is the signal intended for the i-th user satisfying $\mathrm{E}\{|x_i^{\mathrm{d}}|^2\} = 1$. The received signal at the i-th user is given by

$$y_i^{\mathrm{d}} = h_i^{\mathrm{d}}(\sqrt{p_u}x_u^{\mathrm{d}} + \sqrt{p_v}x_v^{\mathrm{d}}) + w_i^{\mathrm{d}}, \tag{8}$$

where w_i^{d} denotes the AWGN at the i-th user. Then the output signal of ADC is

$$y_{i,\Phi_c}^{\mathrm{d}} = \Phi_c(y_i^{\mathrm{d}}) = \Phi_c(h_i^{\mathrm{d}}(\sqrt{p_u}x_u^{\mathrm{d}} + \sqrt{p_v}x_v^{\mathrm{d}}) + w_i^{\mathrm{d}}). \tag{9}$$

At device u, the signal of device v, x_v^{d}, is always treated as interference. Then the received signal-to-interference-and-noise ratio (SINR) for decoding its own signal is $\gamma_{u\to u}$. The instantaneous BLER of device u is approximated as

$$\mathcal{E}_u \approx Q\left(\frac{\mathcal{C}(\gamma_{u\to u}) - \frac{B_u}{L^{\mathrm{d}}}}{\sqrt{V(\gamma_{u\to u})/L^{\mathrm{d}}}}\right), \tag{10}$$

where $\mathcal{C}(\gamma_{u\to u}) = \log_2(1+\gamma_{u\to u})$ is Shannon capacity, $V(\gamma_{u\to u}) = (1 - \frac{1}{(1+\gamma_{u\to u})^2})(\log_2 e)^2$ is the channel dispersion, and $Q^{-1}(\cdot)$ is the inverse of Q-function, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$.

In contrast, SIC is performed at device v. In particular, device v first decodes x_u^{d} by treating x_v^{d} as interference. The received SINRs for decoding x_u^{d} and x_v^{d} at device v are $\gamma_{v\to u}$ and $\gamma_{v\to v}$, respectively. Similar to (10), we have the instantaneous BLERs $\mathcal{E}_{v\to u}$ and $\mathcal{E}_{v\to v}$. Then the instantaneous BLER of device v is approximated as

$$\mathcal{E}_v = \mathcal{E}_{v\to u} + (1 - \mathcal{E}_{v\to u})\mathcal{E}_{v\to v} \overset{\text{(a)}}{\approx} \mathcal{E}_{v\to u} + \mathcal{E}_{v\to v}, \tag{11}$$

where step (a) holds for the case that the signal of device u can be decoded successfully with high probability. Moreover, we will design a coding strategy for each device pair to guarantee transmission reliability later.

3 Detection and performance analysis for uplink transmission

In this section, we propose GEC-SR based detection methods for active device detection, channel estimation, and data detection in the uplink transmission phase. In particular, the updated messages of the proposed iteration algorithms are approximated by the complex Gaussian distribution with the projection operations and a joint active device detection and channel estimation method is presented. With the estimated channel, a signal decoding method of active device is then proposed for the uplink data non-orthogonal transmission.

3.1 Active device detection and channel estimation

Note that the MMSE estimator of $\boldsymbol{x}^p$ in (4) is given by [34]

$$\hat{\boldsymbol{x}}^p = \mathrm{E}\big[\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}\big] = \int \boldsymbol{x}^p p(\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p})\mathrm{d}\boldsymbol{x}^p, \tag{12}$$

3) For average power allocation, we allocate the transmit powers based on $\mathrm{E}\{|\hat{h}_v^{\mathrm{u}}|^2\} \geqslant \mathrm{E}\{|\hat{h}_u^{\mathrm{u}}|^2\}$, i.e., $\lambda_u \geqslant \lambda_v$.

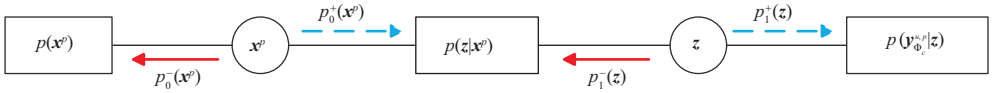

Figure 3 (Color online) The factor graph, where the circle refers to variable node while the square denotes factor node. In addition, the message update rules are shown in [20].

where the expectation is taken over the posterior distribution $p(\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p})$ denoted as

$$p(\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}) = \frac{p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{x}^p)p(\boldsymbol{x}^p)}{p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p})} \propto p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{x}^p)p(\boldsymbol{x}^p) = \int p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{z})\delta(\boldsymbol{z}-\boldsymbol{A}\boldsymbol{x}^p)p(\boldsymbol{x}^p)\mathrm{d}\boldsymbol{z}. \tag{13}$$

The posterior probability in (13) is computationally intractable due to the discrete nature of the active pattern. Thus, we aim at constructing a multi-variate Gaussian approximation of $p(\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p})$ and finding the corresponding mean and variance that are close to that of $p(\boldsymbol{x}^p|\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p})$ in iterative fashion.

As shown in Figure 3, we initialize the forward messages, $p_0^{1,+}(\boldsymbol{x}^p)$, $p_1^{1,+}(\boldsymbol{z})$, as complex Gaussian distributions, i.e., $p_0^{1,+}(\boldsymbol{x}^p) \propto \mathcal{N}_c(\boldsymbol{x}^p; \boldsymbol{m}_0^{1,+}, \mathrm{Diag}(\boldsymbol{v}_0^{1,+}))$ and $p_1^{1,+}(\boldsymbol{z}) \propto \mathcal{N}_c(\boldsymbol{z}; \boldsymbol{m}_1^{1,+}, \mathrm{Diag}(\boldsymbol{v}_1^{1,+}))$. Then the messages, $p_1^{t+1,-}(\boldsymbol{z}), p_0^{t+1,-}(\boldsymbol{x}^p)$, at the back direction (red line) are updated firstly for the t-th iteration due to the observed signals are involved and the messages in the forward direction (cyan line) are then updated. Note that the passing message in each factor node is approximated by the projection operation [20], which is defined as follows:

$$\mathrm{Proj}_x[p(x)] = \arg\min_{q(x)\in\Omega(x)} \mathcal{D}_{\mathrm{KL}}[p(x)||q(x)] = \mathcal{N}_c(x; m, v), \tag{14}$$

where $\mathcal{D}_{\mathrm{KL}}$ is Kullback-Liebler (KL)-divergence. $\Omega(x)$ is a Gaussian family distribution and

$$m = \int xq(x)\mathrm{d}x, \quad v = \int |x-m|^2 q(x)\mathrm{d}x. \tag{15}$$

With the factor graph shown in Figure 3 and the message update rules found in [20], a joint active device detection and channel estimation is presented in Algorithm 1, and we provide some intuition to understand the algorithm. For the second layer in the back direction, the projection of $p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{z})$ is calculated, and the mean vector and variance matrix are expressed as (A1) and (A2), where

$$\zeta^t \triangleq \frac{p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{z})\mathcal{N}_c(\boldsymbol{z}; \boldsymbol{m}_1^{t,+}, \boldsymbol{v}_1^{t,+})}{\int p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{z})\mathcal{N}_c(\boldsymbol{z}; \boldsymbol{m}_1^{t,+}, \boldsymbol{v}_1^{t,+})\mathrm{d}\boldsymbol{z}}. \tag{16}$$

Then the extrinsic information of $\boldsymbol{z}$ is calculated by (A3) and (A4). For the first and the second layers in the back direction, the passing mean vector and variance matrix can be obtained by (A5)–(A11). On the other hand, the projection of the prior probability of $\boldsymbol{x}^p$, i.e., $p(\boldsymbol{x}^p)$, of the first layer in the forward direction is evaluated, and the mean vector and variance matrix are expressed as (A12) and (A13), where

$$\chi^{t+1} \triangleq \frac{p(\boldsymbol{x}^p)\mathcal{N}_c(\boldsymbol{x}^p; \boldsymbol{m}_0^{t+1,-}, \boldsymbol{v}_0^{t+1,-})}{\int p(\boldsymbol{x}^p)\mathcal{N}_c(\boldsymbol{x}^p; \boldsymbol{m}_0^{t+1,-}, \boldsymbol{v}_0^{t+1,-})\mathrm{d}\boldsymbol{x}^p}. \tag{17}$$

The extrinsic information of $\boldsymbol{x}^p$ is calculated by (A14)–(A17). The first and the second layers in the forward direction, the passing mean vector, and variance matrix are obtained by (A18) and (A19).

Note that the received signal is quantized by the ADCs, then the closed-form expressions of $\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}$ and $\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}$ in (A1) and (A2) can be derived based on (16). In addition, the entries of the sparse signal $\boldsymbol{x}^p$ can be formulated as independent and identically distributed (i.i.d.) complex Bernoulli-Gaussian distribution with the probability distribution function (PDF):

$$p(x_n^p) = (1-\epsilon)\delta(x_n^p) + \epsilon\mathcal{N}_c\left(x_n^p; 0, \lambda_n\right). \tag{18}$$

Thus, the closed-form expressions of (A10) and (A11) are derived according to (17) and (18). In the following two propositions, the detail expressions of $\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}$, $\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}$, $\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,+}$, and $\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,+}$ are presented.

Algorithm 1 GEC-SR

1. Initialization: $t=1$, $\boldsymbol{m}_1^{1,+}=\mathbf{0}$, $\boldsymbol{v}_1^{1,+}=\mathbf{1}$, $\boldsymbol{m}_0^{1,+}=\mathbf{0}$, $\boldsymbol{v}_0^{1,+}=\mathbf{1}$.

2. While $t\leqslant T_{\max}$ **do**

$$\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}=\mathrm{E}_{\zeta^t}[\zeta^t], \tag{A1}$$
$$\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}=\mathrm{Var}_{\zeta^t}[\zeta^t], \tag{A2}$$
$$\boldsymbol{v}_1^{t+1,-}=\mathbf{1}\oslash(\mathbf{1}\oslash\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}-\mathbf{1}\oslash\boldsymbol{v}_1^{t,+}), \tag{A3}$$
$$\boldsymbol{m}_1^{t+1,-}=\boldsymbol{v}_1^{t+1,-}\odot(\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}\oslash\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}-\boldsymbol{m}_1^{t,+}\oslash\boldsymbol{v}_1^{t,+}), \tag{A4}$$
$$\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,-}=(\boldsymbol{A}^{\mathrm{H}}(\mathbf{1}\oslash\boldsymbol{v}_1^{t+1,-})\boldsymbol{A}+(\mathbf{1}\oslash\boldsymbol{v}_0^{t,+}))^{-1}, \tag{A5}$$
$$\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,-}=\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,-}(\boldsymbol{A}^{\mathrm{H}}(\mathbf{1}\oslash\boldsymbol{v}_1^{t+1,-})\boldsymbol{m}_1^{t+1,-}+\boldsymbol{m}_0^{t,+}\oslash\boldsymbol{v}_0^{t,+}), \tag{A6}$$
$$\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,-}=\mathrm{diag}(\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,-}), \tag{A7}$$
$$\boldsymbol{v}_0^{t+1,-}=\mathbf{1}\oslash(\mathbf{1}\oslash\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,-}-\mathbf{1}\oslash\boldsymbol{v}_0^{t,+}), \tag{A8}$$
$$\boldsymbol{m}_0^{t+1,-}=\boldsymbol{v}_0^{t+1,-}\odot(\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,-}\oslash\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,-}-\boldsymbol{m}_0^{t,+}\oslash\boldsymbol{v}_0^{t,+}), \tag{A9}$$
$$\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,+}=\mathrm{E}_{\chi^{t+1}}[\chi^{t+1}], \tag{A10}$$
$$\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,+}=\mathrm{Var}_{\chi^{t+1}}[\chi^{t+1}], \tag{A11}$$
$$\boldsymbol{v}_0^{t+1,+}=\mathbf{1}\oslash(\mathbf{1}\oslash\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,+}-\mathbf{1}\oslash\boldsymbol{v}_0^{t+1,-}), \tag{A12}$$
$$\boldsymbol{m}_0^{t+1,+}=\boldsymbol{v}_0^{t+1,+}\odot(\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,+}\oslash\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,+}-\boldsymbol{m}_0^{t+1,-}\oslash\boldsymbol{v}_0^{t+1,-}), \tag{A13}$$
$$\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,+}=(\boldsymbol{A}^{\mathrm{H}}(\mathbf{1}\oslash\boldsymbol{v}_1^{t+1,-})\boldsymbol{A}+(\mathbf{1}\oslash\boldsymbol{v}_0^{t+1,+}))^{-1}, \tag{A14}$$
$$\hat{\boldsymbol{m}}_{\boldsymbol{x}^p}^{t+1,+}=\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,+}(\boldsymbol{A}^{\mathrm{H}}(\mathbf{1}\oslash\boldsymbol{v}_1^{t+1,-})\boldsymbol{m}_1^{t+1,-}+\boldsymbol{m}_0^{t+1,+}\oslash\boldsymbol{v}_0^{t+1,+}), \tag{A15}$$
$$\boldsymbol{m}_{\boldsymbol{z}}^{t+1,+}=\boldsymbol{A}\hat{\boldsymbol{m}}_{\boldsymbol{x}^p}^{t+1,+}, \tag{A16}$$
$$\boldsymbol{v}_{\boldsymbol{z}}^{t+1,+}=\mathrm{diag}(\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{x}^p}^{t+1,+}\boldsymbol{A}^{\mathrm{H}}), \tag{A17}$$
$$\boldsymbol{v}_1^{t+1,+}=\mathbf{1}\oslash(\mathbf{1}\oslash\boldsymbol{v}_{\boldsymbol{z}}^{t+1,+}-\mathbf{1}\oslash\boldsymbol{v}_1^{t+1,-}), \tag{A18}$$
$$\boldsymbol{m}_1^{t+1,+}=\boldsymbol{v}_1^{t+1,+}\odot(\boldsymbol{m}_{\boldsymbol{z}}^{t+1,+}\oslash\boldsymbol{v}_{\boldsymbol{z}}^{t+1,+}-\boldsymbol{m}_1^{t+1,-}\oslash\boldsymbol{v}_1^{t+1,-}), \tag{A19}$$

3. End while

4. Output: $\boldsymbol{m}_{\boldsymbol{x}^p}^{T_{\max}}, \boldsymbol{v}_{\boldsymbol{x}^p}^{T_{\max},+}$.

Proposition 1. Let $y_{\Phi_c}^{\mathrm{u},p}$, $m_z^{t+1,-}$, $v_z^{t+1,-}$, $m_1^{t,+}$, and $v_1^{t,+}$ be the real part or image part of any one element in $\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}$, $\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}$, $\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}$, $\boldsymbol{m}_1^{t,+}$, and $\boldsymbol{v}_1^{t,+}$, the expressions of $m_z^{t+1,-}$ and $v_z^{t+1,-}$ in closed-form are given by

$$m_z^{t+1,-}=m_1^{t,+}-\frac{v_1^{t,+}}{\sqrt{2(\sigma_p^2+v_1^{t,+})}}\frac{\Theta(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\Theta(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}{\Psi(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\Psi(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}, \tag{19}$$

$$v_z^{t+1,-}=\frac{v_1^{t,+}}{2}-\frac{(v_1^{t,+})^2}{2(\sigma_p^2+v_1^{t,+})}\left[\frac{\eta_1(y_{\Phi_c}^{\mathrm{u},p})\Theta(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\eta_2(y_{\Phi_c}^{\mathrm{u},p})\Theta(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}{\Psi(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\Psi(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}\right.$$
$$\left.+\left(\frac{\Theta(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\Theta(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}{\Psi(\eta_1(y_{\Phi_c}^{\mathrm{u},p}))-\Psi(\eta_2(y_{\Phi_c}^{\mathrm{u},p}))}\right)^2\right], \tag{20}$$

where $\Psi(x)=\int_{-\infty}^{x}\mathcal{N}(t;0,1)\mathrm{d}t$, $\Theta(x)=\mathcal{N}(x;0,1)$, the expectation and variance are taken over

$$\frac{p(y_{\Phi_c}^{\mathrm{u},p}|z)\mathcal{N}_c(z;m_1^{t,+},v_1^{t,+}/2)}{\int p(y_{\Phi_c}^{\mathrm{u},p}|z)\mathcal{N}_c(z;m_1^{t,+},v_1^{t,+}/2)\mathrm{d}z},$$

and

$$\eta_1(y^{\mathrm{u},p})\triangleq\frac{y^{\mathrm{up}}-m_1^{t,+}}{\sqrt{(\sigma_p^2+v_1^{t,+})/2}},\quad \eta_2(y^{\mathrm{u},p})\triangleq\frac{y^{\mathrm{low}}-m_1^{t,+}}{\sqrt{(\sigma_p^2+v_1^{t,+})/2}}, \tag{21}$$

where y^{up}, y^{low} are defined in [33].

Proof. See Appendix A.

Remark 1. When the bit of ADCs is infinite, we have $p(\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}|\boldsymbol{z})=\mathcal{N}_c(\boldsymbol{z};\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p},\sigma_p^2\boldsymbol{I})$ and $\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}=\mathbf{1}\oslash(\mathbf{1}\oslash\boldsymbol{v}_1^{t,+}+\sigma_0^{-2}\boldsymbol{I})$, $\boldsymbol{m}_{\boldsymbol{z}}^{t+1,-}=\boldsymbol{v}_{\boldsymbol{z}}^{t+1,-}\odot(\sigma_p^{-2}\boldsymbol{y}_{\Phi_c}^{\mathrm{u},p}+\boldsymbol{m}_1^{t,+}\oslash\boldsymbol{v}_1^{t,+})$.

Proposition 2. The closed-form expressions of elements of $\boldsymbol{m}_{\boldsymbol{x}^p}^{t+1,+}$, $\boldsymbol{v}_{\boldsymbol{x}^p}^{t+1,+}$ in (A10) and (A11) are given by (22) and (23) at the top of next page.

$$\boldsymbol{m}_{\boldsymbol{x}^p,n}^{t+1,+}=\frac{\epsilon\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-}+\lambda_n)m_{0,n}^{t+1,-}\lambda_n}{[(1-\epsilon)\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-})+\epsilon\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-}+\lambda_n)](\lambda_n+v_{0,n}^{t+1,-})}, \tag{22}$$

$$\boldsymbol{v}_{\boldsymbol{x}^p,n}^{t+1,+}=\frac{\epsilon\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-}+\lambda_n)}{(1-\epsilon)\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-})+\epsilon\mathcal{N}_c(0;m_{0,n}^{t+1,-},v_{0,n}^{t+1,-}+\lambda_n)}\left(\frac{\lambda_n v_{0,n}^{t+1,-}}{\lambda_n+v_{0,n}^{t+1,-}}+\left|\frac{m_{0,n}^{t+1,-}\lambda_n}{\lambda_n+v_{0,n}^{t+1,-}}\right|^2\right)-|\boldsymbol{m}_{\boldsymbol{x}^p,n}^{t+1,+}|^2. \tag{23}$$

Proof. See Appendix B.

Therefore, the estimated signal in (13) is $\hat{\boldsymbol{x}}^p=\boldsymbol{m}_{\boldsymbol{x}^p}^{T_{\max},+}$ with the variance matrix $\boldsymbol{v}_{\boldsymbol{x}^p}^{T_{\max},+}$. Then the log-likelihood ratio (LLR) test for active pattern estimation is given by

$$\mathrm{LLR}\left(\hat{x}_n^p\right)=\log\left(\frac{p_{\hat{x}_n^p|\varpi_n}\left(\hat{x}_n^p|\varpi_n=1\right)}{p_{\hat{x}_n^p|\varpi_n}\left(\hat{x}_n^p|\varpi_n=0\right)}\right). \tag{24}$$

With the estimated $\hat{\boldsymbol{x}}^p$, the conditional probability of $\hat{x}_n^p$ on ϖ_n is given by

$$p_{\hat{x}_n^p|\varpi_n}(\hat{x}_n^p|\varpi_n)=\begin{cases}\dfrac{1}{\pi(\lambda_n+v_n)}\exp\left(-\dfrac{|\hat{x}_n^p|^2}{\lambda_n+v_n}\right), & \text{if } \varpi_n=1,\\ \dfrac{1}{\pi v_n}\exp\left(-\dfrac{|\hat{x}_n^p|^2}{v_n}\right), & \text{if } \varpi_n=0,\end{cases} \tag{25}$$

where v_n is the n-th component of $\boldsymbol{v}_{\boldsymbol{x}^p}^{T_{\max},+}$. Then the $\mathrm{LLR}(\hat{x}_n^p)$ in (24) can be calculated as

$$\mathrm{LLR}(\hat{x}_n^p)=\log\left(\frac{v_n}{\lambda_n+v_n}\exp\left(|\hat{x}_n^p|^2\left(\frac{1}{v_n}-\frac{1}{\lambda_n+v_n}\right)\right)\right). \tag{26}$$

Then, we have

$$\log\left(\frac{v_n}{\lambda_n+v_n}\exp\left(|\hat{x}_n^p|^2\left(\frac{1}{v_n}-\frac{1}{\lambda_n+v_n}\right)\right)\right)\gtrless 0\Leftrightarrow|\hat{x}_n^p|^2\gtrless\Pi_n, \tag{27}$$

where the threshold is

$$\Pi_n=\frac{\log(1+\frac{\lambda_n}{v_n})}{\frac{1}{v_n}-\frac{1}{\lambda_n+v_n}}. \tag{28}$$

Therefore, the device activity indicator function of device n is obtained as

$$\hat{\varpi}_n=\begin{cases}1, & \text{if } |\hat{x}_n^p|^2\geqslant\Pi_n,\\ 0, & \text{if } |\hat{x}_n^p|^2<\Pi_n,\end{cases} \tag{29}$$

and the estimated CSI of device n is $\hat{x}_n^p$ if $\hat{\varpi}_n=1$.

The computational cost of each iteration of GEC-SR based algorithm can be divided into two parts: linear operations and nonlinear operation (the rest of equations). The nonlinear operations refer to (A1), (A2), (A10), and (A11) which do not change with dimensions. The complexity of them is $\mathcal{O}(N^2)$. The linear operations refer to the rest equations i.e., (A3)–(A9) and (A12)–(A19). The complexity of them is dominated by matrix inverse referring to (A5) which is the cost of $\mathcal{O}(N^3)$. Hence, the complexity of the GEC-SR based method is $\mathcal{O}(N^3T)$.

3.2 Signal detection for uplink data transmission

Define $\hat{N}=\{n|\hat{\varpi}_n=1,n=1,2,\ldots,N\}$, $\hat{h}_n=\hat{x}_n^p$, $\forall\hat{n}\in\hat{N}$, the received signal of uplink transmission data in (3) based on detected active devices can be written as

$$\boldsymbol{y}^{\mathrm{u}}=\Phi_c\big(\hat{\boldsymbol{S}}\mathrm{Diag}(\hat{\boldsymbol{h}}_{\hat{N}}^{\mathrm{u}})\boldsymbol{x}^{\mathrm{u}}+\boldsymbol{w}^{\mathrm{u}}\big), \tag{30}$$

where $\hat{\boldsymbol{h}}_{\hat{N}}^{\mathrm{u}}=(\hat{h}_1^{\mathrm{u}},\ldots,\hat{h}_{|\hat{N}|}^{\mathrm{u}})^{\mathrm{T}}$, $\hat{\boldsymbol{S}}=[\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{|\hat{N}|}]\in\mathbb{C}^{L^{\mathrm{u}}\times\hat{N}}$, and $\boldsymbol{x}^{\mathrm{u}}=(x_1^{\mathrm{u}},\ldots,x_{\hat{N}}^{\mathrm{u}})^{\mathrm{T}}$.

If the linear MMSE (LMMSE)[4)] is employed at the BS, the detected data symbols are [8]

$$\hat{\boldsymbol{x}}^{\mathrm{u}}=\Phi_{\mathcal{A}}\big((\sigma_0^2\boldsymbol{I}+\tilde{\boldsymbol{L}}_{\hat{N}}^{\mathrm{H}}\tilde{\boldsymbol{L}}_{\hat{N}})^{-1}\tilde{\boldsymbol{L}}_{\hat{N}}^{\mathrm{H}}\boldsymbol{y}^{\mathrm{u}}\big), \tag{31}$$

4) We consider a linear model $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{w}$, where $\mathrm{E}[\boldsymbol{w}]=\boldsymbol{0}$, $\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^{\mathrm{H}}]=\sigma_w^2\boldsymbol{I}_m$, $\mathrm{E}[\boldsymbol{x}]=\boldsymbol{0}$, $\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}]=\sigma_x^2\boldsymbol{I}_n$. By the orthogonality condition ($\mathrm{E}[(\boldsymbol{x}-\hat{\boldsymbol{x}})\boldsymbol{y}^{\mathrm{H}}]=\boldsymbol{0},\hat{\boldsymbol{x}}=\boldsymbol{B}\boldsymbol{y}$), we have the LMMSE estimator $\hat{\boldsymbol{x}}=\Sigma_{xy}\Sigma_{yy}^{-1}\boldsymbol{y}$, where $\Sigma_{xy}=\mathrm{E}[\boldsymbol{x}\boldsymbol{y}^{\mathrm{H}}]=\mathrm{E}[\boldsymbol{x}(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{w})^{\mathrm{H}}]=\sigma_x^2\boldsymbol{H}^{\mathrm{H}}$, $\Sigma_{yy}=\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}]=\sigma_x^2\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}+\sigma_w^2$.

where $\tilde{\boldsymbol{L}}_{\hat{N}} = \hat{\boldsymbol{S}}\text{Diag}(\hat{\boldsymbol{h}}^{\text{u}}_{\hat{N}})$, $\Phi_{\mathcal{A}}$ is the quantization operator on the constellation $\mathcal{A}$. An LMMSE detector-based GEC-SR is proposed to detect the data packet, which is summarized in Algorithm 2. Note that the error propagation caused by the active device detection is considered in this symbol detector.

Algorithm 2 GEC-SR-based detection for uplink data transmission

1: **Input:** The pilot sequences $\{\boldsymbol{s}_n\}$, the quantized signal $\tilde{\boldsymbol{y}}$, the activity probability ϵ, the channel variances $\{\lambda_n\}$, the constellation $\mathcal{A}$.
2: **Step 1:** Solve the problem: $\Phi_c(\boldsymbol{A}\boldsymbol{x}^p + \boldsymbol{w}^p)$, by using the GEC-SR, the estimated signal $\hat{\boldsymbol{x}}^p = \boldsymbol{m}_{\boldsymbol{x}^p}^{T_{\max},+}$ with the variance matrix $\boldsymbol{v}_{\boldsymbol{x}^p}^{T_{\max},+}$, where (A10) and (A11) are derived based on the PDF in (17);
3: **Step 2:** Calculate the threshold $\prod_n$ in (28), the active devices and the CSIs are estimated as $\hat{N} = \{n||\hat{x}_n^p|^2 \geqslant \prod_n\}$ and $\hat{h}_n = \hat{x}_n^p$, $\forall n \in \hat{N}$;
4: **Step 3:** The received uplink transmission data is formulated as (30) and solved by the LMMSE detector defined in (31).
5: **Output:** Decoded symbols $\hat{\boldsymbol{x}}^u$.

4 Design and performance analysis for downlink transmission

In this section, we focus on the design and analysis of the downlink hybrid NOMA scheme. Specially, the downlink channel information is first estimated by the devices with low-resolution ADCs. Then, we drive the BLER of short-packet transmission for one pair of devices performed NOMA. In order to obtain insight, we further approximate the analytical BLERs. Moreover, we reveal the impact of message bits with a fixed block-length, which can be used to guide the devices pairing/grouping. Finally, the analytical results of two-device case can be extended to the general case that consists of more than two devices in one group.

4.1 Downlink channel estimation

With the aid of additive quantization noise model (AQNM) [35–37], the quantized vector $\boldsymbol{y}_{i,\Phi_c}^{\text{d},p}$ is decomposed as

$$\boldsymbol{y}_{i,\Phi_c}^{\text{d},p} = \Phi_c(\boldsymbol{y}_i^{\text{d},p}) \approx \alpha(\boldsymbol{\phi}_i h_i^{\text{d}} + \boldsymbol{w}_i^{\text{d},p}) + \boldsymbol{w}_q^{\text{d},p}, \tag{32}$$

where $\boldsymbol{w}_q^{\text{d},p}$ denotes the additive Gaussian quantization noise vector which is uncorrelated with $\boldsymbol{y}_i^{\text{d},p}$, $\boldsymbol{w}_q^{\text{d},p} \sim \mathcal{N}_c(\boldsymbol{0}_{L^q}, \sigma_{q1}^2 \boldsymbol{I}_{L^q})$, $\sigma_{q1}^2 = \alpha\varsigma\text{E}\{|\boldsymbol{y}_i^{\text{d},p}|^2\} = \alpha\varsigma(\lambda_i + \sigma_p^2)$, and $\alpha = 1 - \varsigma$ with [35]

$$\varsigma = \frac{\text{E}\{\|\boldsymbol{y}_{i,\Phi_c}^{\text{d},p} - \boldsymbol{y}_i^{\text{d},p}\|^2\}}{\text{E}\{\|\boldsymbol{y}_i^{\text{d},p}\|^2\}} \tag{33}$$

denoting the distortion factor of the low-resolution ADC [37]. When the ADC resolution B is large ($B \geqslant 5$), the distortion factor ς can be approximated as [38] $\varsigma \approx \frac{\pi\sqrt{3}}{2} 2^{-2B}$.

Lemma 1. The LS estimator of h_i^{d} is

$$\hat{h}_i^{\text{d}} = \frac{1}{\alpha} \boldsymbol{\phi}_i^{\text{H}} \boldsymbol{y}_{i,\Phi_c}^{\text{d},p} \tag{34}$$

and its distribution is

$$\hat{h}_i^{\text{d}} \sim \mathcal{N}_c\left(0, \lambda_i + \sigma_p^2 + \frac{\varsigma}{\alpha}\left(\lambda_i + \sigma_p^2\right)\right). \tag{35}$$

In addition, the estimation error is $e_i = h_i^{\text{d}} - \hat{h}_i^{\text{d}}$, which is independent with $\hat{h}_i^{\text{d}}$, and the correlation coefficient is $\sigma_{ei}^2 = \text{E}\{e_i e_i^{\text{H}}\} = \sigma_p^2 + \frac{\varsigma}{\alpha}(\lambda_i + \sigma_p^2)$, and the relationship between $\hat{h}_i^{\text{d}}$ and h_i^{d} is

$$\hat{h}_i^{\text{d}} = h_i^{\text{d}} + e_i. \tag{36}$$

Proof. See Appendix C.

The transmit SNR in the downlink channel training phase is defined as $\varrho \triangleq \frac{|\boldsymbol{\phi}_i|^2}{L^q \sigma_p^2} = \frac{1}{L^q \sigma_p^2}$. As shown in Figure 4 with $\lambda_i = 0.5$, the distribution of estimated CSI matches well with the simulation.

Remark 2. Lemma 1 reveals the impact of the parameters on downlink channel estimation. From (75) in the proof of Lemma 1, one can see that the variance of estimation error is decided by the quantify precision of ADC and the noise power with a given length of the pilot sequence. Therefore, the channel estimation accuracy can be controlled by adjusting the parameters according to the relationship shown in Lemma 1.

4.2 Average BLER analysis

From (32) and (36), the output signal of ADC in (9) is approximated as

$$y^{\mathrm{d}}_{i,\Phi_c} \approx \alpha(h^{\mathrm{d}}_i(\sqrt{p_u}x^{\mathrm{d}}_u + \sqrt{p_v}x^{\mathrm{d}}_v) + w^{\mathrm{d}}_i) + w^{\mathrm{d}}_q = \alpha\hat{h}^{\mathrm{d}}_i(\sqrt{p_u}x^{\mathrm{d}}_u + \sqrt{p_v}x^{\mathrm{d}}_v) + \Delta_{ei,w}, \tag{37}$$

where w^{d}_q follows Gaussian distribution with zero mean and variance $\sigma^2_{q_2} = \alpha\varsigma\mathrm{E}\{|y^{\mathrm{d}}_i|^2\} = \alpha\varsigma[(p_u+p_v)\lambda_i + \sigma^2_0]$, and $\Delta_{ei,w} \triangleq -\alpha e_i(\sqrt{p_u}x^{\mathrm{d}}_u + \sqrt{p_v}x^{\mathrm{d}}_v) + \alpha w^{\mathrm{d}}_i + w^{\mathrm{d}}_q$. Note that $\Delta_{ei,w}$ is also a Gaussian distribution with zero mean and variance $\sigma^2_{\Delta_{ei,w}} = \alpha^2\sigma^2_{ei}(p_u+p_v) + \alpha^2\sigma^2_0 + \sigma^2_{q_2}$. Then the SINRs and SNR are

$$\gamma_{i\to u} = \frac{p_u|\hat{h}_i|^2}{p_v|\hat{h}_i|^2 + \tilde{\sigma}^2_{ei} + \sigma^2_0}, \quad \gamma_{v\to v} = \frac{p_v|\hat{h}_v|^2}{\tilde{\sigma}^2_{ev} + \sigma^2_0}, \tag{38}$$

where $\tilde{\sigma}^2_{ei} = \sigma^2_{ei}(p_u+p_v) + \sigma^2_{q_2}/\alpha^2$. The average BLER of device j at device i is

$$\bar{\mathcal{E}}_{i\to j} \approx \int_0^\infty Q\left(\frac{\mathcal{C}(\gamma_{i\to j}) - \frac{B_j}{L^{\mathrm{d}}}}{\sqrt{V(\gamma_{i\to j})/L^{\mathrm{d}}}}\right) f_{\gamma_{i\to j}}(x)\mathrm{d}x, \tag{39}$$

where $f_{\gamma_{i\to u}}(x)$ is the PDF of $\gamma_{i\to u}$. Then we derive the average BLER of NOMA with imperfect CSI and finite ADC bits in Theorem 1.

Theorem 1. The average BLERs of device u and device v in NOMA with imperfect CSI are expressed as (40) and (41), respectively, shown at the top of the next page,

$$\begin{aligned}\bar{\mathcal{E}}_u \approx & 1 - \frac{\alpha_{u,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}\alpha_u(\tilde{\sigma}^2_{eu}+1/\rho)\exp(\frac{\tilde{\sigma}^2_{eu}+1/\rho}{\alpha_v\tilde{\lambda}_u})}{\alpha^2_v\tilde{\lambda}_u} \\ & \cdot\left[E_1\left(-\frac{\alpha_u(\tilde{\sigma}^2_{eu}+1/\rho)}{(\alpha_u-\alpha_v\mu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_u}\right) - E_1\left(-\frac{\alpha_u(\tilde{\sigma}^2_{eu}+1/\rho)}{(\alpha_u-\alpha_v\nu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_u}\right)\right] \\ & + \frac{\alpha_{u,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}\exp(\frac{\tilde{\sigma}^2_{eu}+1/\rho}{\alpha_v\tilde{\lambda}_u})}{\alpha_v}\cdot(\alpha_u-\alpha_v\nu_{u,L^{\mathrm{d}}}) \\ & \cdot\left[\exp\left(-\frac{\alpha_u(\tilde{\sigma}^2_{eu}+1/\rho)}{(\alpha_u-\alpha_v\nu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_u}\right) - \exp\left(-\frac{\alpha_u(\tilde{\sigma}^2_{eu}+1/\rho)}{(\alpha_u-\alpha_v\mu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_u}\right)\right],\end{aligned} \tag{40}$$

and

$$\begin{aligned}\bar{\mathcal{E}}_v \approx & 2 - \frac{\alpha_{v,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}\alpha_v(\tilde{\sigma}^2_{ev}+1/\rho)\exp(\frac{\tilde{\sigma}^2_{ev}+1/\rho}{\alpha_v\tilde{\lambda}_v})}{\alpha^2_v\tilde{\lambda}_v} \\ & \cdot\left[E_1\left(\frac{\alpha_u(\tilde{\sigma}^2_{ev}+1/\rho)}{(\alpha_u-\alpha_v\mu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_v}\right) - E_1\left(\frac{\alpha_u(\tilde{\sigma}^2_{ev}+1/\rho)}{(\alpha_u-\alpha_v\nu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_v}\right)\right] \\ & + \frac{\alpha_{u,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}\exp(\frac{\tilde{\sigma}^2_{\varepsilon}+1/\rho}{\alpha_v\tilde{\lambda}_v})}{\alpha_v}\cdot(\alpha_u-\alpha_v\nu_{u,L_u}) \\ & \cdot\left[\exp\left(-\frac{\alpha_u(\tilde{\sigma}^2_{ev}+1/\rho)}{(\alpha_u-\alpha_v\nu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_v}\right) - \exp\left(-\frac{\alpha_u(\tilde{\sigma}^2_{ev}+1/\rho)}{(\alpha_u-\alpha_v\mu_{u,L^{\mathrm{d}}})\alpha_v\tilde{\lambda}_v}\right)\right] \\ & - \frac{\alpha_{v,L_v}\sqrt{L_v}\alpha_v\tilde{\lambda}_v}{\tilde{\sigma}^2_{ev}+1/\rho}\left[\exp\left(-\frac{\mu_{v,L^{\mathrm{d}}}(\tilde{\sigma}^2_{ev}+1/\rho)}{\alpha_v\tilde{\lambda}_v}\right) - \exp\left(-\frac{\nu_{v,L^{\mathrm{d}}}(\tilde{\sigma}^2_{ev}+1/\rho)}{\alpha_v\tilde{\lambda}_v}\right)\right],\end{aligned} \tag{41}$$

where $\alpha_{u,L^{\mathrm{d}}}=\frac{1}{\sqrt{2\pi(2^{2B_u/L^{\mathrm{d}}}-1)}}$, $\beta_{u,L^{\mathrm{d}}}=2^{\frac{B_u}{L^{\mathrm{d}}}}-1$, $\mu_{u,L^{\mathrm{d}}}=\beta_{u,L^{\mathrm{d}}}-\frac{1}{2\alpha_{u,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}}$, $\nu_{u,L^{\mathrm{d}}}=\beta_{u,L^{\mathrm{d}}}+\frac{1}{2\alpha_{u,L^{\mathrm{d}}}\sqrt{L^{\mathrm{d}}}}$, $\tilde{\lambda}_i=\lambda_i+\sigma_{ei}^2$, $\alpha_u=p_u/(p_u+p_v)$, $p_v=1-p_u$, $\rho=(p_u+p_v)/\sigma_0^2$, and the exponential integral E_1 is defined as $E_1(\mu)=\int_{\mu}^{\infty}\frac{\exp(-x)}{x}\mathrm{d}x$, $\mu>0$.

Proof. See Appendix D.

Similarly, the individual average BLER of two-device OMA with FBC is approximated as

$$\hat{\mathcal{E}}_i\approx 1-\frac{\hat{\alpha}_{i,L^{\mathrm{d}}}\sqrt{0.5L^{\mathrm{d}}}(\tilde{\lambda}_i+\sigma_{ei}^2)}{\tilde{\sigma}_{ei}^2+1/\rho}\left(\exp\left(-\frac{\hat{\mu}_{i,L^{\mathrm{d}}}(\tilde{\sigma}_{ei}^2+1/\rho)}{\tilde{\lambda}_i+\tilde{\sigma}_{ei}^2}\right)-\exp\left(-\frac{\hat{\nu}_{i,L^{\mathrm{d}}}(\tilde{\sigma}_{ei}^2+1/\rho)}{\tilde{\lambda}_i+\tilde{\sigma}_{ei}^2}\right)\right),\tag{42}$$

where $\hat{\alpha}_{i,L^{\mathrm{d}}}=\frac{1}{\sqrt{2\pi(2^{4B_i/L^{\mathrm{d}}}-1)}}$, $\hat{\mu}_{i,L^{\mathrm{d}}}=2^{\frac{2B_i}{L^{\mathrm{d}}}}-1-\frac{1}{2\hat{\alpha}_{i,L^{\mathrm{d}}}\sqrt{0.5L^{\mathrm{d}}}}$, $\hat{\nu}_{i,L^{\mathrm{d}}}=2^{\frac{2B_i}{L^{\mathrm{d}}}}-1+\frac{1}{2\hat{\alpha}_{i,L^{\mathrm{d}}}\sqrt{0.5L^{\mathrm{d}}}}$.

The developed approximated closed-form expressions of average BLER in Theorem 1 can be easily used to evaluate the performance of NOMA with FBC.

4.3 Discussion on device grouping

From (10), we have the maximal achievable rate $\frac{B_j}{L^{\mathrm{d}}}\approx\mathcal{C}(\gamma_j)-Q^{-1}(\mathcal{E})\sqrt{\frac{V(\gamma_j)}{L^{\mathrm{d}}}}$. Note that $Q^{-1}(0.5)=0$, which means that if $\mathcal{E}=0.5$ the maximal achievable rate $\frac{B_j}{L^{\mathrm{d}}}$ is equal to the capacity $\mathcal{C}(\gamma_j)$. In order to reveal the impact of devices grouping, we define the outage probability as $P_j^{\mathrm{out}}=\Pr(\mathcal{C}(\gamma_j)<\hat{R}_j)$, where $\hat{R}_j=\frac{B_j}{L^{\mathrm{d}}}$ is target rate. We assume that there are J devices in one group performed NOMA with power allocation coefficients $\alpha_1>\cdots>\alpha_J$, $\sum_{j=1}^{J}\alpha_j=1$. Then the outage probability of device j $(1\leqslant j<J)$ is expressed as

$$\begin{aligned}P_j^{\mathrm{out}}&=1-\Pr\left(\frac{\alpha_1|\hat{h}_j^{\mathrm{d}}|^2}{\sum_{i=2}^{J}\alpha_i|\hat{h}_j^{\mathrm{d}}|^2+\tilde{\sigma}_{ej}^2+1/\rho}>r_1,\ldots,\frac{\alpha_j|\hat{h}_j^{\mathrm{d}}|^2}{\sum_{i,i>j}\alpha_i|\hat{h}_j^{\mathrm{d}}|^2+\tilde{\sigma}_{ej}^2+1/\rho}>r_j\right)\\&=1-\Pr\left(|\hat{h}_j^{\mathrm{d}}|^2>\frac{(\tilde{\sigma}_{ej}^2+1/\rho)r_1}{\alpha_1-r_1\sum_{i=2}^{J}\alpha_i},\ldots,|\hat{h}_j^{\mathrm{d}}|^2>\frac{(\tilde{\sigma}_{ej}^2+1/\rho)r_j}{\alpha_j-r_j\sum_{i,i>j}\alpha_i}\right),\end{aligned}\tag{43}$$

where $r_j=2^{\hat{R}_j}-1$. Note that the outage probability in (43) is always one when $\alpha_j-r_j\sum_{i,i>j}\alpha_i\leqslant 0$. Therefore, the power allocation coefficients and the target rate of device j should satisfy the following condition:

$$\frac{\alpha_j}{\sum_{i,i>j}\alpha_i}>r_j\Leftrightarrow\frac{\alpha_j}{\sum_{i,i>j}\alpha_i}>2^{\hat{R}_j}-1.\tag{44}$$

By defining $\varphi=\max\{\frac{r_1}{\alpha_1-r_1\sum_{i=2}^{J}\alpha_i},\ldots,\frac{r_j}{\alpha_j-r_j\sum_{i,i>j}\alpha_i}\}$, the outage probability in (43) can be calculated by

$$P_j^{\mathrm{out}}=1-\exp\left(-\frac{(\tilde{\sigma}_{ej}^2+1/\rho)\varphi}{\tilde{\lambda}_j}\right).\tag{45}$$

Similarly, the outage probability of device j in OMA is expressed as

$$P_j^{\mathrm{out}}=\Pr\left(\frac{1}{J}\log_2\left(1+\frac{|\hat{h}_j^{\mathrm{d}}|^2}{\tilde{\sigma}_{ej}^2+1/\rho}\right)<\hat{R}_j\right)=1-\exp\left(-\frac{(2^{J\hat{R}_j}-1)(\tilde{\sigma}_{ej}^2+1/\rho)}{\tilde{\lambda}_j}\right).\tag{46}$$

Based on (45) and (46), the condition on the superiority of NOMA to OMA for $1\leqslant j\leqslant J-1$ is

$$\varphi<2^{J\hat{R}_j}-1\Leftrightarrow\max\left\{\frac{2^{\hat{R}_1}-1}{\alpha_1-(2^{\hat{R}_1}-1)\sum_{i=2}^{J}\alpha_i},\ldots,\frac{2^{\hat{R}_j}-1}{\alpha_j-(2^{\hat{R}_j}-1)\sum_{i,i>j}\alpha_i}\right\}<2^{J\hat{R}_j}-1.\tag{47}$$

Similarly, the condition on the superiority of NOMA to OMA for device J can be expressed as

$$\max\left\{\frac{2^{\hat{R}_1}-1}{\alpha_1-(2^{\hat{R}_1}-1)\sum_{i=2}^{J}\alpha_i},\ldots,\frac{2^{\hat{R}_j}-1}{\alpha_j-(2^{\hat{R}_j}-1)\sum_{i,i>j}\alpha_i},\ldots,\frac{2^{\hat{R}_J}-1}{\alpha_J}\right\}<2^{J\hat{R}_J}-1.\tag{48}$$

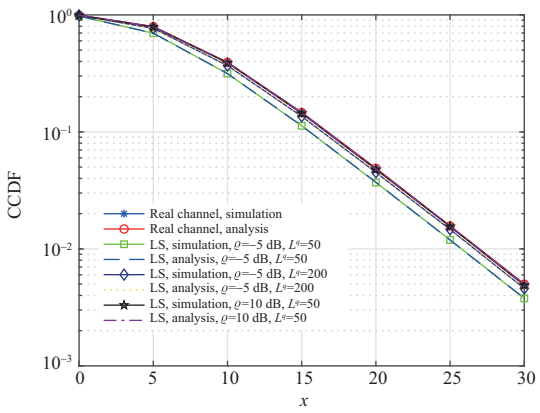

Figure 4 (Color online) The complementary CDF (CCDF) of the estimated channel in Lemma 1.

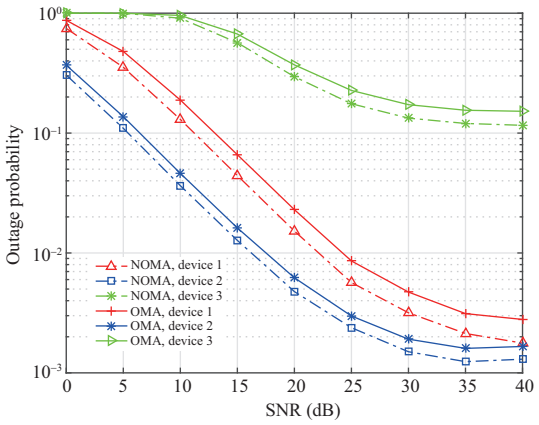

Figure 5 (Color online) Outage probability of three-device NOMA with ADCs.

From the analytical results in (44), (47) and (48), we can make the following conclusion:

(1) With the increase of the number of devices, to guarantee the reliability of NOMA, the target rate of each device in the group should be smaller.

(2) The power allocation and the target rates should satisfy the conditions in (47) and (48).

(3) The power allocation and the target rate selection are very complicated for the case that more than two devices in one group.

For example, we consider the case that $J = 3$, then the power allocation coefficients and the target rates should satisfy the following conditions.

(1) For device 1, the following conditions should be satisfied:

$$\frac{\alpha_1}{\alpha_2+\alpha_3} > 2^{\hat{R}_1} - 1, \quad \frac{2^{\hat{R}_1}-1}{\alpha_1 - (2^{\hat{R}_1}-1)(\alpha_2+\alpha_3)} < 2^{3\hat{R}_1} - 1. \tag{49}$$

(2) For device 2, the following conditions should be satisfied:

$$\frac{\alpha_2}{\alpha_3} > 2^{\hat{R}_2} - 1, \quad \max\left\{\frac{2^{\hat{R}_1}-1}{\alpha_1 - (2^{\hat{R}_1}-1)(\alpha_2+\alpha_3)}, \frac{2^{\hat{R}_2}-1}{\alpha_2 - (2^{\hat{R}_2}-1)\alpha_3}\right\} < 2^{3\hat{R}_2} - 1. \tag{50}$$

(3) For device 3, the following conditions should be satisfied:

$$\alpha_3 > 2^{\hat{R}_3} - 1, \quad \max\left\{\frac{2^{\hat{R}_1}-1}{\alpha_1 - (2^{\hat{R}_1}-1)(\alpha_2+\alpha_3)}, \frac{2^{\hat{R}_2}-1}{\alpha_2 - (2^{\hat{R}_2}-1)\alpha_3}, \frac{2^{\hat{R}_3}-1}{\alpha_3}\right\} < 2^{3\hat{R}_3} - 1. \tag{51}$$

As shown in Figure 5 with $\varrho = 15$ dB, $L^q = 50$, $B = 5$ bits, the individual outage probability of NOMA outperforms that of OMA when $\alpha_1 = 0.5$, $\alpha_2 = 0.4$, $\alpha_3 = 0.1$, $\zeta = 3$, $d_1 = 2$, $d_2 = 1$, $d_3 = 0.5$, $\hat{R}_1 = \hat{R}_2 = 0.1$ bps/Hz, $\hat{R}_3 = 1.6$ bps/Hz. From (49)–(51), we can see that the power allocation and the target rate selection are complicated. Even if the power coefficient is given, the target rate selection is still complicated.

5 Simulations and numerical results

In this section, we provide numerical examples to verify the proposed algorithms and the analytical results for both uplink and downlink NOMA schemes with short-packet transmissions.

5.1 Uplink short-packet transmission

The GAMP [20,39] is shown as the benchmark for the performance comparison for the uplink simulations due to the quantized system with ADC. The performance comparisons between the GAMP and GEC-SR are shown in Figures 6(a) and (b) for channel and active device estimation, where the normalized mean

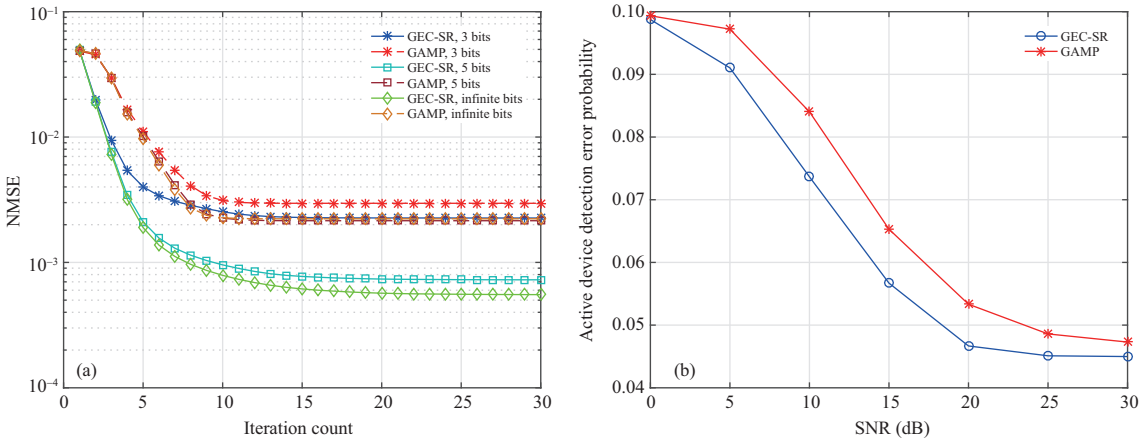

Figure 6 (Color online) (a) Iteration count vs. NMSE where $(N, L^p, \epsilon) = (1000, 200, 0.05)$ and SNR = 25 dB; (b) SNR vs. active device detection error probability, where $(N, L^p, \epsilon) = (1000, 250, 0.1)$ and $B = 3$ bits.

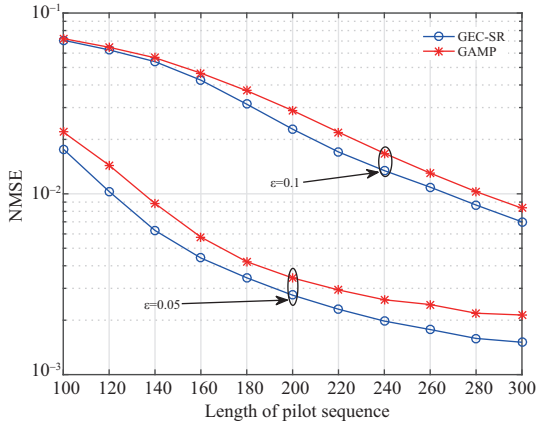

Figure 7 (Color online) The impact of the length of the pilot sequence.

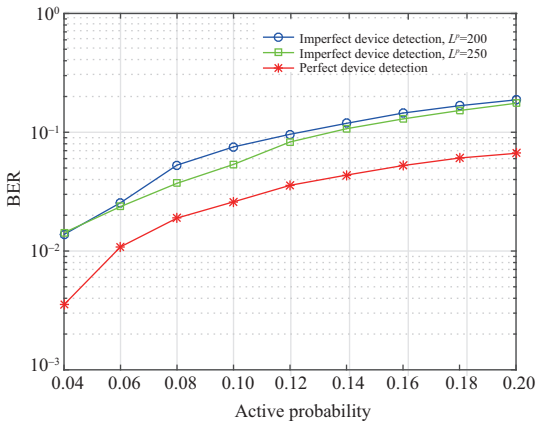

Figure 8 (Color online) BER of uplink data transmission with error propagation.

squared error (NMSE) is defined as the MSE between the estimated $\hat{\boldsymbol{x}}^p$ and actual $\boldsymbol{x}^p$. In Figure 6(a), we can observe that the NMSE converges for both GEC-SR and GAMP with different ADC bits at 25 dB. Moreover, compared to the GAMP algorithm, the convergency value NMSE is much smaller for the GEC-SR based algorithm. It is also observed that the performance of GAMP and GEC-SR with 5 ADC bits is close to the case that with infinite ADC bits. In Figure 6(b), the active device detection error probabilities for GAMP and GEC-SR are shown as the functions of SNR, respectively. One can observe that the active device detection error probability of GEC-SR is smaller than that of GAMP from Figure 6(b).

Figure 7 shows the NMSE performance of GEC-SR as a function of the length of the pilot sequence by using GAMP as a benchmark algorithm.

It can be seen that GEC-SR has a better performance than GAMP. The NMSE decreases as the length of the pilot sequence increases and it will be flat when the length of the pilot sequence is large enough.

It can be also seen that the shorter length of the pilot sequence can satisfy the same NMSE requirements of different device active probabilities.

The impact of error propagation caused by active detection estimation errors on the BER performance of uplink data is shown in Figure 8, where the BER is defined as

$$\text{BER} = \frac{L_{\text{num}} + d_{\text{num}} \log_2 M}{N \log_2 M}, \tag{52}$$

where L_{num} denotes the number of data detection error bit and d_{num} is the active device detection error. And $d_{\text{num}} = 0$ for the case that the active device detection is perfect. As shown in Figure 8, the BER

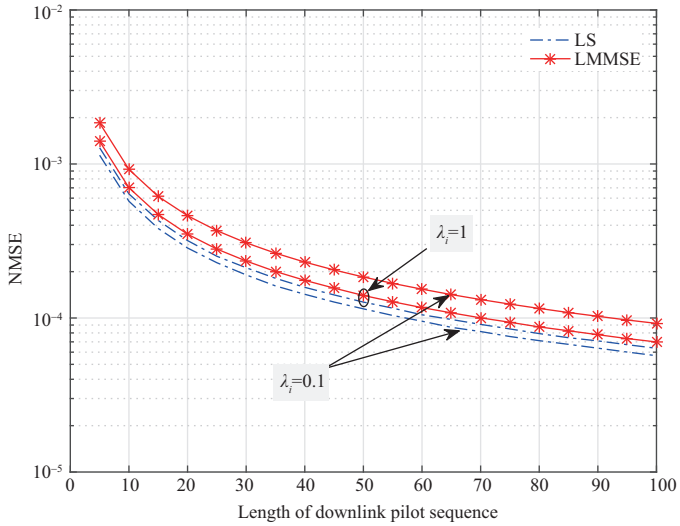

Figure 9 (Color online) NMSE vs. length of downlink pilot sequence.

gap between imperfect active device detection and perfect device detection is increased when the active probability is large due to the error propagation of active device detection.

5.2 Downlink short-packet transmission

The NMSE of downlink channel estimation is shown in Figure 9, where includes LS and LMMS estimations at 20 dB. It shows that the NMSE performance of LS is better than that of LMMSE. One can observe that the gap between LS and LMMSE becomes larger when the parameter λ_i is smaller. It is worthy to note that the NMSE approximates 10^{-4} when the length of the pilot sequence is 60. Thus, we can control the variance, σ_{ei}^2, of estimation error by adjusting the parameters including the length of the pilot sequence, SNR and ADC bits according to the results in Lemma 1.

In Figure 10, we set $\varrho = 15$ dB, $\lambda_i = 1$, $L^q = 50$, $B = 5$ bits for channel estimations and $\alpha_u = 0.7$, $\alpha_v = 0.3$, $L^{\mathrm{d}} = 168$ for data transmission. The ADCs at the training and data phases have the same parameters.

In Figure 10(a), the analytical results of average BLERs in (40) and (41) are presented for downlink NOMA with FBC, where NOMA with imperfect CSI and perfect CSI is considered. The two curves show that the analytical results match well with the computer simulation results for the whole range of the SNR, which confirms the accuracy of the derived expressions. Furthermore, it can be observed from Figure 10(a) that the average BLER of NOMA with imperfect CSI can result in an error floor at the high SNR region.

In Figures 10(b) and (c), the average BLERs of devices in one group are presented as a function of message bits. As shown in Figure 10(b), the average BLER of device u in one group increases as the message bits increase with a fixed block-length. It is important to note that the average BLER performance of device u in NOMA is superior to that in OMA before the threshold. While the average BLER of device u in NOMA becomes worse when the message bits are greater than the threshold. More importantly, the average BLER of device u in NOMA is always one if the message bits are sufficiently large. On the other hand, the impact of message bits on the average BLER of devices v with a fixed block-length is shown in Figure 10(c). It can be observed that the average BLER performance of device v is better than that in OMA if the message bits are greater than the threshold for the case in which device v selects proper message bits. Otherwise, the OMA transmission scheme should be selected due to its excellent performance. Moreover, the performance of device v in NOMA is always inferior to that in OMA if device u selects its message bits irrelevantly. This is because the error propagation becomes more severe. Finally, Figure 10(d) verifies the proposed message bit selection strategy analysis results and simulations of NOMA and OMA with low-resolution ADCs and imperfect CSI.

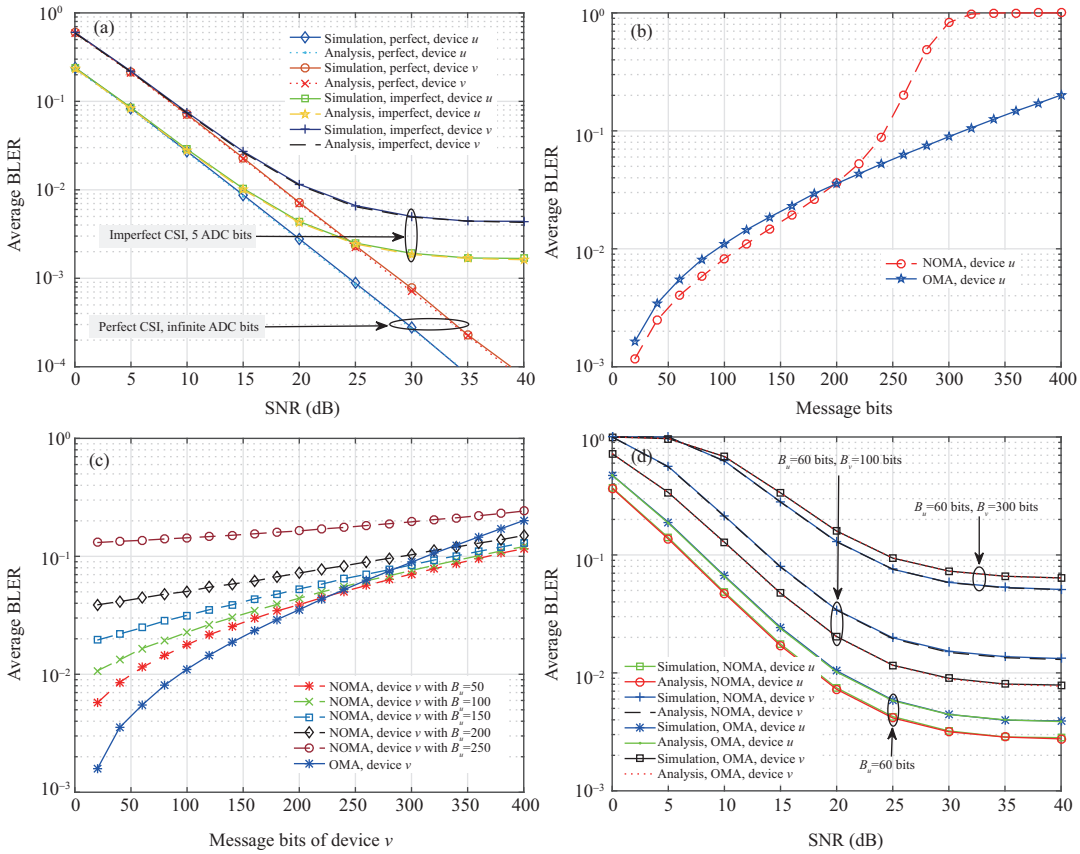

Figure 10 (Color online) (a) Average BLER for downlink NOMA; (b) message bits vs. average BLER for device u; (c) the impact of message bits for device v; (d) BLER performance comparison of NOMA and OMA with different message bits.

6 Conclusion

In this paper, we have proposed two NOMA schemes for uplink and downlink transmissions in cellular Internet of Things with short-packet transmission. Particularly, a low-complexity algorithm based on GEC-SR has been proposed to detect active devices and estimate their CSI for uplink grant-free NOMA. Furthermore, we have obtained the BER of uplink data transmission with imperfect estimated CSI and discussed the impact of error propagation caused by the device detection. On the other hand, a hybrid NOMA scheme is proposed for downlink transmission. The CSI is estimated and the BLERs of each pair of active devices with finite block-length coding are derived in closed-form. With the analytical results, a message bit selection strategy in each pair of devices has been proposed to ensure better NOMA performance than OMA, and an extended strategy for devices grouping has been proposed. Finally, we presented simulation results to demonstrate the accuracy of the proposed algorithms and the obtained analytical results. More importantly, the obtained results show that the performance of NOMA is superior to OMA when the message bits are selected according to the proposed strategy, which can be used to guide devices grouping in NOMA.

Acknowledgements This work of Donghong CAI was supported by National Natural Science Foundation of China (Grant No. 62001190) and China Postdoctoral Science Foundation (Grant No. 2021M691249). This work of Pingzhi FAN was supported by National Natural Science Foundation of China (Grant No. 62020106001) and 111 Project (Grant No. 111-2-14). This work of Yanqing XU was supported by China Postdoctoral Science Foundation (Grant No. 2021M693100). This work of Zhiquan LIU was supported by National Natural Science Foundation of China (Grant No. 61802146) and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011017).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Bockelmann C, Pratas N, Nikopour H, et al. Massive machine-type communications in 5G: physical and MAC-layer solutions. IEEE Commun Mag, 2016, 54: 59–65

2 Jiao J, Liao S Y, Sun Y Y, et al. Fairness-improved and QoS-guaranteed resource allocation for NOMA-based S-IoT network. Sci China Inf Sci, 2021, 64: 169306

3 Gao N, Ni Q, Feng D Q, et al. Physical layer authentication under intelligent spoofing in wireless sensor networks. Signal Process, 2020, 166: 107272

4 Ye N, Li X M, Yu H X, et al. Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things. IEEE Trans Ind Inf, 2019, 15: 2995–3005

5 Jia R D, Chen X M, Qi Q, et al. Massive beam-division multiple access for B5G cellular Internet of Things. IEEE Int Things J, 2020, 7: 2386–2396

6 New W K, Leow C Y, Navaie K, et al. Application of NOMA for cellular-connected UAVs: opportunities and challenges. Sci China Inf Sci, 2021, 64: 140302

7 Ma Z, Zhang Z Q, Ding Z G, et al. Key techniques for 5G wireless communications: network architecture, physical layer, and MAC layer perspectives. Sci China Inf Sci, 2015, 58: 041301

8 Ahn J, Shim B, Lee K B. EP-based joint active user detection and channel estimation for massive machine-type communications. IEEE Trans Commun, 2019, 67: 5178–5189

9 Shahab M B, Abbas R, Shirvanimoghaddam M, et al. Grant-free non-orthogonal multiple access for IoT: a survey. IEEE Commun Surv Tut, 2020, 22: 1805–1838

10 Irtaza S A, Lim S H, Choi J W. Greedy data-aided active user detection for massive machine type communications. IEEE Wirel Commun Lett, 2019, 8: 1224–1227

11 Yu H X, Fei Z S, Zheng Z, et al. Finite-alphabet signature design for grant-free NOMA: a quantized deep learning approach. IEEE Trans Veh Technol, 2020, 69: 10975–10987

12 Jiang S C, Yuan X J, Wang X, et al. Joint user identification, channel estimation, and signal detection for grant-free NOMA. IEEE Trans Wirel Commun, 2020, 19: 6960–6976

13 Choi J. NOMA-based compressive random access using gaussian spreading. IEEE Trans Commun, 2019, 67: 5167–5177

14 Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans Inform Theory, 2007, 53: 4655–4666

15 Schepker H F, Bockelmann C, Dekorsy A. Exploiting sparsity in channel and data estimation for sporadic multi-user communication. In: Proceedings of International Symposium on Wireless Communication Systems, Ilmenau, 2013. 1–5

16 Majumdar A, Ward R K. Fast group sparse classification. In: Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 2009. 11–16

17 Donoho D L, Maleki A, Montanari A. Message-passing algorithms for compressed sensing. Proc Natl Acad Sci, 2009, 106: 18914–18919

18 Liu L, Yu W. Massive connectivity with massive MIMO-part I: device activity detection and channel estimation. IEEE Trans Signal Process, 2018, 66: 2933–2946

19 Huang N H, Chiueh T D. Sequence design and user activity detection for uplink grant-free NOMA in mMTC networks. IEEE Open J Commun Soc, 2021, 2: 384–395

20 Zou Q Y, Zhang H C, Wen C K, et al. Concise derivation for generalized approximate message passing using expectation propagation. IEEE Signal Process Lett, 2018, 25: 1835–1839

21 Zou Q Y, Zhang H C, Cai D H, et al. Message passing based joint channel and user activity estimation for uplink grant-free massive MIMO systems with low-precision ADCs. IEEE Signal Process Lett, 2020, 27: 506–510

22 Zou Q Y, Zhang H C, Cai D H, et al. A low-complexity joint user activity, channel and data estimation for grant-free massive MIMO systems. IEEE Signal Process Lett, 2020, 27: 1290–1294

23 Fu J W, Wu G, Zhang Y Z, et al. Active user identification based on asynchronous sparse bayesian learning with SVM. IEEE Access, 2019, 7: 108116

24 Wen C K, Wang C J, Jin S, et al. Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs. IEEE Trans Signal Process, 2016, 64: 2541–2556

25 Shao X D, Chen X M, Zhong C J, et al. A unified design of massive access for cellular Internet of Things. IEEE Internet Things J, 2019, 6: 3934–3947

26 Ren H, Pan C H, Deng Y S, et al. Joint power and blocklength optimization for URLLC in a factory automation scenario. IEEE Trans Wirel Commun, 2020, 19: 1786–1801

27 Hu Y L, Gursoy M C, Schmeink A. Efficient transmission schemes for low-latency networks: NOMA vs. relaying. In: Proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, Montreal, 2017. 1–6

28 Zhang Q Q, Zhang L, Liang Y C, et al. Backscatter-NOMA: a symbiotic system of cellular and Internet-of-Things networks. IEEE Access, 2019, 7: 20000–20013

29 Ma Z, Xiao M, Xiao Y, et al. High-reliability and low-latency wireless communication for Internet of Things: challenges, fundamentals, and enabling technologies. IEEE Int Things J, 2019, 6: 7946–7970

30 Yu Y H, Chen H, Li Y H, et al. On the performance of non-orthogonal multiple access in short-packet communications. IEEE Commun Lett, 2018, 22: 590–593

31 Ding Z G, Fan P Z, Poor H V. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. IEEE Trans Veh Technol, 2016, 65: 6010–6023

32 Nikopour H, Yi E, Bayesteh A, et al. SCMA for downlink multiple access of 5G wireless networks. In: Proccessing of IEEE Global Communications Conference, Austin, 2014. 3940–3945

33 He H T, Wen C K, Jin S. Bayesian optimal data detector for hybrid mmWave MIMO-OFDM systems with low-resolution ADCs. IEEE J Sel Top Signal Process, 2018, 12: 469–483

34 Rangan S, Schniter P, Fletcher A K, et al. On the convergence of approximate message passing with arbitrary matrices. IEEE Trans Inform Theory, 2019, 65: 5339–5351

35 Fletcher A K, Rangan S, Goyal V K, et al. Robust predictive quantization: analysis and design via convex optimization. IEEE J Sel Top Signal Process, 2007, 1: 618–632

36 Jacobsson S, Durisi G, Coldrey M, et al. Throughput analysis of massive MIMO uplink with low-resolution ADCs. IEEE Trans Wirel Commun, 2017, 16: 4038–4051

37 Zhang J Y, Dai L L, He Z Y, et al. Performance analysis of mixed-ADC massive MIMO systems over rician fading channels. IEEE J Sel Areas Commun, 2017, 35: 1327–1338

38 Mo J, Alkhateeb A, Abu-Surra S, et al. Hybrid architectures with few-bit ADC receivers: achievable rates and energy-rate tradeoffs. IEEE Trans Wirel Commun, 2017, 16: 2274–2287

39 Rangan S. Generalized approximate message passing for estimation with random linear mixing. In: Proceedings of IEEE International Symposium on Information Theory Proceedings, Petersburg, 2011. 2168–2172