# Next-Generation URLLC with Massive Devices: A Unified Semi-Blind Detection Framework for Sourced and Unsourced Random Access

Malong Ke, Zhen Gao, *Member, IEEE,* Mingyu Zhou, Dezhi Zheng,

Derrick Wing Kwan Ng, *Fellow, IEEE,* and H. Vincent Poor, *Life Fellow, IEEE*

## Abstract

This paper proposes a unified semi-blind detection framework for sourced and unsourced random access (RA), which enables next-generation ultra-reliable low-latency communications (URLLC) with massive devices. Specifically, the active devices transmit their uplink access signals in a grant-free manner to realize ultra-low access latency. Meanwhile, the base station aims to achieve ultra-reliable data detection under severe inter-device interference without exploiting explicit channel state information (CSI). We first propose an efficient transmitter design, where a small amount of reference information (RI) is embedded in the access signal to resolve the inherent ambiguities incurred by the unknown CSI. At the receiver, we further develop a successive interference cancellation-based semi-blind detection scheme, where a bilinear generalized approximate message passing algorithm is utilized for joint channel and signal estimation (JCSE), while the embedded RI is exploited for ambiguity elimination. Particularly, a rank selection approach and a RI-aided initialization strategy are incorporated to reduce the algorithmic computational complexity and to enhance the JCSE reliability, respectively. Besides, four enabling techniques are integrated to satisfy the stringent latency and reliability requirements of massive URLLC. Numerical results demonstrate that the proposed semi-blind detection framework offers a better scalability-latency-reliability tradeoff than the state-of-the-art detection schemes dedicated to sourced or unsourced RA.

M. Ke, Z. Gao, and D. Zheng are with the School of Information and Electronics, Beijing Institute of Technology, 100081 Beijing, China (e-mails: kemalong@bit.edu.cn; gaozhen16@bit.edu.cn; zhengdezhi@bit.edu.cn). M. Zhou is with Baicells Technologies Co. Ltd., Beijing 100089, China (e-mail: zhoumingyu@baicells.com). D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, 2052 Sydney, Australia (e-mail: w.k.ng@unsw.edu.au). H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, NJ 08542 Princeton, USA (e-mail: poor@princeton.edu).

**Index Terms**

Massive URLLC, grant-free, sourced/unsourced random access, semi-blind detection, approximate message passing.

## I. INTRODUCTION

### A. Background and Motivation

The emerging Internet-of-Things (IoT) applications in various vertical sectors have driven the massive machine-type communication (mMTC) and ultra-reliable low-latency communication (URLLC) services in the fifth-generation (5G) cellular systems, which pursue scalability and reliability with low user plane latency, respectively [1]–[3]. Motivated by the grander Internet-of-Everything (IoE) that is envisioned to connect millions of people and billions of machines, the next-generation, i.e., Beyond 5G or sixth-generation (6G), cellular systems must further scale the classical URLLC across the device dimension, leading to a new massive URLLC service that merges legacy mMTC and URLLC [4]. The application scenarios range from extended reality (XR) services to flying vehicles, brain-computer interfaces, and connected autonomous systems. Although the conventional network slicing is effective in supporting a simple mixture of mMTC and URLLC, it is still very challenging to simultaneously satisfy the stringent scalability, latency, and reliability requirements (e.g., $10^6$ devices/km$^2$, 1 ms user plane latency, and 99.99999% reliability) of massive URLLC [5], [6].

More specifically, the unprecedentedly high density of wireless devices has already posed great challenges in random access (RA), which is essential for ensuring ubiquitous IoE connectivity [7]–[9]. In legacy cellular systems, the widely adopted grant-based RA protocol requires multiple signaling interactions to facilitate the scheduling of interference-free transmissions [8]. Despite its simplicity and reliability, this protocol would become inefficient or even impractical in the context of massive URLLC due to its extremely high access latency resulting from severe access collisions among the massive devices [9]. To tackle this issue, the promising grant-free RA protocol has been recently proposed as a key enabler to achieve ultra-low access latency, where the active devices directly transmit their access signals to the base station (BS) without any scheduling in advance [10]. However, since the signals of all the active devices are transmitted via the same physical resources, the inter-device interference becomes a severely limiting factor for realizing ultra-reliable data detection. Therefore, the key challenge of massive URLLC lies in the improvement of data detection reliability for grant-free massive RA (MRA) [11].

In general, grant-free MRA can be classified into two paradigms, i.e., sourced and unsourced RA, which focus on two practical RA scenarios having different access requirements [12]. For sourced RA, the BS is interested in both the transmitted messages and the identities (IDs) of the devices that generated them. Hence, some reference information (RI), such as pilot sequence, should be transmitted along with the payload data for device identification. While for unsourced RA, the BS is solely interested in estimating a list of sent messages, without any interest in the identities of the transmitters. Therefore, the payload efficiency can be improved by omitting the device ID information in the transmission. Considering their different access requirements, the research community has developed two independent lines of research to study the reliable data detection for grant-free sourced and unsourced RA, respectively. However, designing a unified data detection framework for incorporating both RA paradigms is still an open issue, which is indispensable to satisfy the heterogeneous access requirements of future IoE applications [12]. Meanwhile, the previous works generally focus on the traditional mMTC and fail to support the emerging massive URLLC that simultaneously pursues the stringent scalability, latency, and reliability requirements [5], [6].

## B. Related Work

Grant-free sourced RA has been intensively investigated in the literature, e.g., [11]–[17], where the non-orthogonal pilot-based coherent detection framework is generally considered. Specifically, each active device transmits a non-orthogonal pilot sequence along with its payload data to the BS in a grant-free manner. Meanwhile, the BS first performs active device detection (ADD) and channel estimation (CE) based on the received pilot signal, then the acquired results are adopted for the subsequent coherent data detection [11]. A key feature of massive URLLC is the sporadic uplink traffic, i.e., for any given time interval, only a small number of devices are activated by external events and desire to access the network [12], [13]. By leveraging the sparse device activity, the authors in [14] formulated the joint ADD and CE design as a compressive sensing (CS) problem and an orthogonal matching pursuit-based algorithm was developed for the related sparse signal recovery. However, this work assumes only a single-antenna receiver at the BS and the solution is not applicable to multi-antenna systems. Also, the work in [15] revealed that the detection error probability of ADD can be driven to zero as the number of BS antennas is sufficiently large. On the other hand, to reduce the computational complexity in the case of large numbers of devices and BS antennas, a dimension reduction-based joint ADD and CE approach

was further proposed in [16]. Particularly, the massive multiple-input multiple-output (MIMO) channels between the devices and the BS usually exhibit clustered sparsity in the virtual angular domain [18]. In this context, the authors in [17] developed an approximate message passing (AMP)-based ADD and CE scheme to leverage the angular-domain clustered sparsity for further enhanced MRA performance. Overall, the previous works on grant-free sourced RA generally focus on the scalability of the traditional mMTC service, where the transmission latency (or pilot length) must increase linearly with the number of active devices to guarantee the reliable data detection [17]. Therefore, it is challenging for them to simultaneously satisfy the stringent latency and reliability requirements of massive URLLC.

Recent studies on grant-free unsourced RA mainly rely on the common codebook-based non-coherent detection framework introduced in [19]. Specifically, according to the payload data bits to be transmitted, each active device sends a codeword selected from a common codebook. Unlike the sourced RA counterpart, the BS in this case is solely interested in estimating a list of sent messages without any interests in the identities of the transmitters, i.e., the estimated messages have an unknown permutation. The main obstacle of realizing the scheme stems from the extremely large size of the codebook, i.e., the number of codewords, which grows exponentially with respect to the payload data length and causes prohibitive computational complexity [19]. To overcome this limitation, the first low-complexity coding scheme for unsourced RA was proposed in [20], where the transmission period was divided into multiple small sub-blocks and each active device randomly chose a sub-block to transmit its codeword. Relying on a similar transmission structure, the subsequent work in [21] further proposed a close-to-optimal coding strategy, where user-independent successive interference cancellation (SIC) was applied for improved decoding performance. Subsequently, the authors in [22] proposed another efficient approach, which leveraged recent advances in the CS field to further reduce the decoding complexity. For this scheme, the message of each active device is split into several sub-messages and the coding scheme is divided into two parts, i.e., inner and outer encoder/decoder. Here, a CS-based inner encoder/decoder is adopted to map a sub-message into a codeword at the devices and estimate the transmitted sub-messages at the BS, as in [19]. Meanwhile, a tree-based outer decoder is employed to acquire the original messages by stitching the estimated sub-messages together. The works in [19]–[22] consider a Gaussian multiple access channel model, where the BS is equipped with a single-antenna and the channel gains between the devices and the BS are assumed to be unity. Although this assumption facilitates the performance analysis of the

proposed coding scheme, it hinders the practical application of the results. Moreover, the authors in [23], [24] revealed that the required transmit power-per-bit can be driven to an arbitrarily small value as the number of BS antennas grows sufficiently large. Considering the emerging massive MIMO systems, an uncoupled CS-based unsourced RA solution was proposed, which exploited the rich spatial dimensionality offered by the large-scale antenna array to enhance the decoding performance [25]. The strong common characteristic of the aforementioned works lies in the employment of the coding scheme based on a common codebook. It is also challenging for them to simultaneously satisfy the stringent latency and reliability requirements of massive URLLC due to the low payload efficiency or the high computational complexity resulting from the employment of the common codebook-based coding scheme [6], [25].

In previous works, the traditional sourced and unsourced RA paradigms generally adopt their dedicated data detection frameworks, i.e., coherent and non-coherent detection, respectively, which rely on different transceiver designs, cf. [17], [22]. The authors in [12] have tried to support both sourced and unsourced RA services in the same IoE system. However, the two RA paradigms still adopt their dedicated data detection frameworks, which rely on different transmission schemes, signal models, and data detection schemes. Here, only the related activity detection algorithm is unified. In this context, we have to integrate two different transceivers into the same system, allowing the network to switch between sourced and unsourced RA modes according to practical access requirements. This solution is unattractive in terms of device size, hardware complexity, and overall cost [12]. Therefore, a more beneficial unified detection framework is needed, where both RA paradigms can share almost the same RA procedure, transceiver hardware design, and receive algorithm.

*C. Main Contributions*

In this paper, we design a unified semi-blind detection framework for grant-free sourced and unsourced RA, which pursues the ultra-reliable and low-latency requirements of massive URLLC. Specifically, the active devices directly transmit their uplink access signals exploiting the same physical resources, where a small amount of RI is embedded in the access signals. Based on the overlapped received signal, the BS jointly estimates the channels and detects the signals of the active devices, then the embedded RI is exploited to eliminate the inherent ambiguities. For sourced RA, the RI contains device ID bits, cyclic redundancy check (CRC) bits, and a scalar pilot symbol, which are adopted for eliminating the phase and permutation ambiguities.

While for unsourced RA, only CRC bits and a scalar pilot symbol are transmitted for phase ambiguity elimination, and thus higher payload efficiency can be achieved. In summary, our main contributions are listed as follows:

- We propose a unified semi-blind detection framework for enabling grant-free sourced and unsourced RA, under which both RA paradigms share almost the same RA procedure, transceiver hardware design, and receive algorithm. Moreover, in contrast to the existing non-orthogonal pilot-based coherent detection for sourced RA [11]–[17], the proposed detection framework results in a significant transmission latency reduction when the same detection reliability is considered. Furthermore, compared to the common codebook-based non-coherent detection for unsourced RA [19]–[25], the proposed detection framework dramatically reduces the processing latency by circumventing the common codebook-based coding scheme. Due to the reduced transmission and processing latencies, the proposed detection framework achieves a much lower user plane latency than its counterparts [13].

- We propose an SIC-based semi-blind detection scheme at the BS, which mitigates the inter-device interference iteratively. In each SIC iteration, the channels and the signals of the active devices are jointly inferred from the overlapped received signal, while the embedded RI is exploited for ambiguity elimination. Moreover, the signal components of reliably detected active devices are removed from the received signal to alleviate the inter-device interference in the following iterations.

- We propose a bilinear generalized AMP (BiG-AMP)-based joint channel and signal estimation (JCSE) algorithm, where the JCSE is formulated as a matrix factorization problem based on the Bayesian theory and the advanced BiG-AMP algorithm is employed to obtain a low-complexity approximate solution. Particularly, we develop a rank selection approach to estimate the unknown number of active devices, which facilitates the computational complexity reduction of the BiG-AMP algorithm. Moreover, a RI-aided initialization strategy is further incorporated for improved JCSE reliability. The proposed algorithm significantly outperforms the classic BiG-AMP algorithm adopting the random initialization strategy [26].

- We introduce four enabling techniques that can be flexibly integrated into the proposed semi-blind detection framework to further reduce the user plane latency and enhance the detection reliability. The obtained URLLC-enhanced version of the proposed detection framework is capable of simultaneously satisfying the stringent scalability, latency, and

Fig. 1. An illustration of future massive URLLC scenarios with sparse device activity. A one-ring channel model is considered between the devices and the massive MIMO BS.

reliability requirements of massive URLLC.

*D. Notations*

We adopt normal-face letters to denote scalars and lowercase (uppercase) boldface letters to denote column vectors (matrices). The $(n, k)$th element, the $n$th row vector, and the $k$th column vector of the matrix $\mathbf{G} \in \mathbb{C}^{N \times K}$ are denoted as $g_{n,k}$, $[\mathbf{G}]_{n,:}$, and $[\mathbf{G}]_{:,k}$, respectively, where $\mathbb{C}$ is the set of complex numbers. $\mathbb{B}$ is the set of binary numbers and $\mathbf{0}_{N \times K}$ is the zero matrix of size $N \times K$. The superscripts $(\cdot)^{\mathrm{T}}$, $(\cdot)^{*}$, $(\cdot)^{\mathrm{H}}$ and $(\cdot)^{\dagger}$ represent the transpose, complex conjugate, conjugate transpose, and pseudo-inverse operators, respectively. $[K]$ denotes the set of integers $\{1, 2, \cdots, K\}$, $|\mathcal{A}|_c$ is the cardinal number of the set $\mathcal{A}$, $\emptyset$ is an empty set, and $\mathrm{supp}\{\cdot\}$ denotes the support set of a sparse vector or matrix. $\|\mathbf{G}\|_{\mathrm{F}}$ denotes the Frobenius-norm of the matrix $\mathbf{G}$ and $\|\mathbf{G}\|_{0}$ denotes the zero-norm of $\mathbf{G}$, i.e., the number of non-zero elements in $\mathbf{G}$. $[\mathbf{G}]_{:,\mathcal{A}}$ represents the matrix that stacks the columns of $\mathbf{G}$ indexed by the set $\mathcal{A}$, while $[\mathbf{G}]_{\mathcal{A},:}$ is the matrix that stacks the rows of $\mathbf{G}$ indexed by the set $\mathcal{A}$. $\mathcal{R}(\cdot)$ is the real part of a complex number. $\lceil b \rceil$ rounds $b$ to the nearest integer greater than or equal to $b$. $\mathcal{U}(x; a, b)$ denotes that the variable $x$ follows the uniform distribution between $a$ and $b$. Finally, $\mathcal{CN}(x; \mu, v)$ denotes the complex Gaussian distribution of a random variable $x$ with mean $\mu$ and variance $v$. $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote statistical expectation and variance operators, respectively.

## II. System Model

Consider the uplink of a typical massive URLLC scenario in massive MIMO systems, as depicted in Fig. 1. Here, we employ a BS equipped with an $N$-antenna uniform linear array

(ULA) to provide access service for $K$ synchronized single-antenna devices. Due to the sporadic uplink traffic of IoE, it is assumed that only $K_a$ ($K_a \ll K$) out of the total $K$ devices are activated by external events and desire to access the network [17]. To avoid the complicated access scheduling for ultra-low access latency, the promising grant-free RA protocol is adopted for uplink transmission, where the active devices directly transmit their access signals to the BS via the same time-frequency resources. At the BS, the signal $\mathbf{r}_l \in \mathbb{C}^{N \times 1}$ received in the $l$th symbol duration is expressed as

$$\mathbf{r}_l = \sum_{k=1}^{K} \mathbf{g}_k \alpha_k x_{k,l} + \mathbf{w}_l = \mathbf{G}\mathbf{x}_l + \mathbf{w}_l, \tag{1}$$

where $\mathbf{g}_k \in \mathbb{C}^{N \times 1}$ denotes the uplink channel between the $k$th device and the BS, the binary variable $\alpha_k$ indicates the device activity, i.e., $\alpha_k = 1$ for active and 0 otherwise, $x_{k,l} \in \mathbb{C}$ is the transmitted signal (i.e., modulated symbol) of the $k$th device in the $l$th symbol duration, $\mathbf{w}_l \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$ is the additive white Gaussian noise (AWGN), and $\sigma^2$ is the noise variance. Incorporating both channel response and device activity, $\mathbf{G} = [\alpha_1 \mathbf{g}_1, \alpha_2 \mathbf{g}_2, \cdots, \alpha_K \mathbf{g}_K] \in \mathbb{C}^{N \times K}$ is referred to as the MRA channel matrix and $\mathbf{x}_l = [x_{1,l}, x_{2,l}, \cdots, x_{K,l}]^{\mathrm{T}} \in \mathbb{C}^{K \times 1}$. Further focusing on small data packets, the length of the symbol frame $L$ is usually far smaller than the channel coherence time. Meanwhile, the device activity remains constant during the frame. In this context, the number of active devices is fixed within each frame but may change across different frames. For a specific frame, the received signal over $L$ successive symbol durations is given as

$$\mathbf{R} = \mathbf{G}\mathbf{X} + \mathbf{W}, \tag{2}$$

where $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_L] \in \mathbb{C}^{N \times L}$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L] \in \mathbb{C}^{K \times L}$, and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_L]$.

Considering the widely studied spatial channel model [17], the channel between the $k$th device and the BS is modeled as

$$\mathbf{g}_k = \rho_k \sum_{p=1}^{P} \beta_{k,p} \mathbf{a}_R(\phi_{k,p}), \tag{3}$$

where $\rho_k$ is the large-scale fading parameter, $P$ is the number of multi-path components (MPCs), $\beta_{k,p}$ denotes the complex gain of the MPC, and $\mathbf{a}_R(\phi_{k,p}) = [1, e^{-j2\pi\phi_{k,p}}, \cdots, e^{-j2\pi(N-1)\phi_{k,p}}]/\sqrt{N}$ is the array response vector at the BS. Here, $\phi_{k,p} = \frac{d}{\lambda}\sin(\psi_{k,p})$, where $\psi_{k,p}$ is the physical angle-of-arrival (AoA) associated with the $k$th device and the $p$th MPC, $d$ is the antenna spacing, and $\lambda$ is the wavelength.

Fig. 2. The clustered sparsity of the angular-domain MRA channel matrix $\mathbf{H}$, where $K = 30$ devices, $K_a = 10$ active devices, and $N = 500$ BS antennas are considered.

For a typical network deployment, the spatial propagation characteristics of the channels between the devices and the BS can be modeled as an one-ring channel model, see Fig. 1 [17]. Here, the MPCs only can be observed within a small angular window at the BS, i.e., $\psi_{k,p} \in [\psi_0 - \Delta, \psi_0 + \Delta]$, where $\psi_0$ is the central AoA and $2\Delta \ll 180°$ is the angular spread. Define $\mathbf{H} = \mathbf{A}_R \mathbf{G}$ as the angular-domain representation of the MRA channel matrix $\mathbf{G}$, where $\mathbf{A}_R$ denotes the transformation matrix and becomes a discrete Fourier transform matrix for ULA with $d = \lambda/2$. The limited AoA spread leads to the clustered angular-domain sparsity of massive MIMO channels, i.e.,

$$1 < \left|\mathrm{supp}\left\{[\mathbf{H}]_{:,k}\right\}\right|_c \ll N, \forall k \in [K].\tag{4}$$

Moreover, considering the sparse device activity, we further have

$$\left|\mathrm{supp}\left\{[\mathbf{H}]_{n,:}\right\}\right|_c \ll K_a, \forall n \in [N].\tag{5}$$

By combining the sparsity features presented in (4) and (5), the clustered sparsity of the angular-domain MRA channel matrix $\mathbf{H}$ is illustrated in Fig. 2, which will be exploited to facilitate the development of a semi-blind detection scheme at the BS.

*Remark 1:* It should be noted that the received signal model in (2) is identical for both the coherent detection framework dedicated to sourced RA and the non-coherent detection framework dedicated to unsourced RA. The major differences between two detection frameworks lie in the transmitted signal $\mathbf{X}$ and the receive algorithm, which will be detailed in *Section III*.

## III. TRADITIONAL DETECTION FRAMEWORKS FOR SOURCED AND UNSOURCED RA

As described in *Section II*, the key idea of grant-free RA protocol is to avoid complicated signaling interactions between the devices and the BS, thus achieving the ultra-low access latency, but at the expense of severe inter-device interference. Without access scheduling in advance, the uplink signals of all the active devices are overlapped on the same time-frequency resources, which makes reliable data detection at the BS a challenging problem. In this section, we first introduce two state-of-the-art detection frameworks for grant-free sourced and unsourced RA, respectively, which focus on different access requirements. Moreover, the related merits and faults are discussed.

### A. Non-Orthogonal Pilot-Based Coherent Detection for Sourced RA

The non-orthogonal pilot-based coherent detection framework for sourced RA adopts a two-phase transmission scheme [11]–[17], where each frame is divided into the pilot and payload data phases, i.e., $\mathbf{X} = [\mathbf{X}_p, \mathbf{X}_d]$ with $\mathbf{X}_p \in \mathbb{C}^{K \times L_p}$ and $\mathbf{X}_d \in \mathbb{C}^{K \times L_d}$, respectively. Here, the first $L_p$ symbol durations are used to transmit the non-orthogonal pilot sequences of active devices and the remaining $L_d = L - L_p$ symbol durations are reserved for payload data transmission. Similarly, the received signal can be expressed as $\mathbf{R} = [\mathbf{R}_p, \mathbf{R}_d]$, where $\mathbf{R}_p \in \mathbb{C}^{N \times L_p}$ and $\mathbf{R}_d \in \mathbb{C}^{N \times L_d}$ correspond to the received pilot and data signals, respectively. At the receiver, the BS first performs joint ADD and CE based on the received pilot signal $\mathbf{R}_p = \mathbf{G}\mathbf{X}_p + \mathbf{W}_p$, which is equivalent to estimating $\mathbf{G}$ based on the known $\mathbf{X}_p$ and $\mathbf{R}_p$. By leveraging the sparse device activity, the problem can be formulated as a CS problem and the advanced AMP algorithm in [17] can be employed to acquire the solution. With the estimated active device set $\widehat{\mathcal{A}}$ and channel matrix $\widehat{\mathbf{G}}$, the coherent data detection is then achieved as

$$\widehat{\mathbf{X}}_d = \left[ \widehat{\mathbf{G}} \right]_{:,\widehat{\mathcal{A}}}^{\dagger} \mathbf{R}_d, \tag{6}$$

where $\mathbf{R}_d = \mathbf{G}\mathbf{X}_d + \mathbf{W}_d$. At this point, the inter-device interference can be effectively resolved as long as the reliable estimates of the active device set and the MRA channel matrix, i.e., $\widehat{\mathcal{A}}$ and $\widehat{\mathbf{G}}$, respectively, are obtained. However, according to the CS theory, the pilot length $L_p \geq K_a \log_2(K)$ is required to obtain the satisfactory ADD and CE performance, which significantly degrades the payload efficiency, especially in the scenarios of massive URLLC conveying small data packets [17]. By further utilizing the angular-domain sparsity of massive MIMO channels, i.e.,

$$\mathbf{Y} = [\mathbf{Y}_p, \mathbf{Y}_d] = \mathbf{A}_R \mathbf{R} = \mathbf{H}[\mathbf{X}_p, \mathbf{X}_d] + \mathbf{N}, \tag{7}$$

with $\mathbf{N} = \mathbf{A}_R \mathbf{W}$ denoting the noise matrix, the authors in [17] revealed that the minimum pilot overhead can be reduced to $S_a \log_2(K)$ with $S_a = \max\{\text{supp}\{\mathbf{H}_{n,:}\}, \forall n \in [N]\}$ and $S_a \ll K_a$. Yet, the payload efficiency is still limited when $K$ is extremely large. Note that given the fixed payload data length, a lower payload efficiency indicates a higher transmission latency.

### B. Common Codebook-Based Non-Coherent Detection for Unsourced RA

The common codebook-based non-coherent detection framework is dedicated to unsourced RA, where each active device delivers $B$-bit information using a common codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{2^B}\} \subset \mathbb{C}^{L \times 1}$. Specifically, the $B$-bit information $\mathbf{b}_k \in \mathbb{B}^{B \times 1}$ produced by the active device $k$ is mapped to an integer $b_k \in \{1, 2, \cdots, 2^B\}$. Then, the active device simply sends the $b_k$th codeword of the common codebook, i.e., $\mathbf{c}_{b_k}$, to the BS. We can model the codeword selection by a set of $2^B K$ Bernoulli random variables $\delta_{b,k}$, $\forall b \in [2^B]$ and $\forall k \in [K]$. Here, $\delta_{b,k} = 1$ if the $k$th device is active and transmits the code $\mathbf{c}_b$, and $\delta_{b,k} = 0$ otherwise. On this basis, the transmitted signal of the $k$th device can be expressed as $[\mathbf{X}]_{k,:} = \sum_{b=1}^{2^B} \delta_{b,k} \mathbf{c}_b^{\mathrm{T}}$, and the signal model in (2) can be re-formulated as

$$\mathbf{R} = \mathbf{G}[\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_K]^{\mathrm{T}} \mathbf{C} + \mathbf{W}$$
$$= \mathbf{G}\boldsymbol{\Delta}\mathbf{C} + \mathbf{W} = \widetilde{\mathbf{G}}\mathbf{C} + \mathbf{W}, \tag{8}$$

where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{2^B}]^{\mathrm{T}} \in \mathbb{C}^{2^B \times L}$ is the common codebook, $\boldsymbol{\delta}_k = [\delta_{1,k}, \delta_{2,k}, \cdots, \delta_{2^B,k}]^{\mathrm{T}} \in \mathbb{B}^{2^B \times 1}$, and $\widetilde{\mathbf{G}}$ is the matrix combining the spatial-domain MRA channel matrix $\mathbf{G}$ and the codeword selection matrix $\boldsymbol{\Delta} \in \mathbb{B}^{K \times 2^B}$. The matrix $\boldsymbol{\Delta}$ contains only $K_a$ non-zero rows, each of which has a single non-zero entry. With this formulation, each active device contributes a single non-zero coefficient in $[\widetilde{\mathbf{G}}]_{n,:}$, thereby resulting in a $K_a$-sparse $2^B$-dimensional vector. Considering the BS with $N$ receive antennas, the problem can be formulated as a multiple measurement vectors (MMV) support detection problem, where the different rows of $\widetilde{\mathbf{G}}$ have a common sparsity pattern. The problem can be effectively addressed by the CS recovery algorithm such as AMP [17], but the computational complexity scales exponentially with $B$, which is prohibitive even for short packets with dozens of bits. The prohibitive computational complexity leads to an extremely high processing latency at the BS. Although several low-complexity solutions have been proposed [21], [22], the payload efficiency is dramatically degraded due to the introduced redundant coding.

The user plane latency accounts for the one-way latency from the beginning of the packet processing at the transmitter to the successful detection at the receiver. In grant-free MRA, the transmission and receive processing latencies are the two most dominant components contributing to the user plane latency [27]. Therefore, it is generally challenging for the traditional detection frameworks to satisfy the ultra-low latency requirement of massive URLLC due to the low payload efficiency or the high data detection complexity. Moreover, their applications are limited to either sourced or unsourced RA, which is not conductive to accommodating future massive URLLC with heterogeneous access requirements.

## IV. PROPOSED UNIFIED SEMI-BLIND DETECTION FRAMEWORK: TRANSMITTER DESIGN

To overcome the limitations of conventional coherent and non-coherent detection frameworks, this paper develops a unified semi-blind detection framework for supporting both sourced and unsourced RA. Particularly, our goal is to jointly infer the sparse MRA channel matrix $\mathbf{H}$ and the signal matrix $\mathbf{X}$ from the received signal $\mathbf{Y}$ in (7), based on which the payload data of active devices can be further detected. By avoiding the pilot phase, an extremely high payload efficiency can be achieved, which leads to an ultra-low transmission latency. However, the JCSE problem suffers from the inherent phase and permutation ambiguities. Specifically, define $\mathbf{\Sigma}$ and $\mathbf{\Pi}$ as a diagonal matrix with phase shifts in the diagonal and a permutation matrix, respectively. The ambiguities are caused by the fact that if $\left(\widehat{\mathbf{H}}, \widehat{\mathbf{X}}\right)$ is a solution to the JCSE problem based on (7), then $\left(\widehat{\mathbf{H}}\mathbf{\Sigma}^{-1}\mathbf{\Pi}^{-1}, \mathbf{\Pi}\mathbf{\Sigma}\widehat{\mathbf{X}}\right)$ is also a valid solution. In fact, the cost function $\left\|\mathbf{Y} - \widehat{\mathbf{H}}\widehat{\mathbf{X}}\right\|_{\mathrm{F}}^{2}$ is invariant to any phase shifts and permutations of the rows of $\mathbf{X}$. The phase shift will lead to the demodulation error of estimated signals, while the row permutation will lead to the identification error of active devices. To tackle this issue, we propose to insert a small amount of RI in the access signal $\mathbf{X}$ to eliminate the ambiguities.

The proposed detection framework involves the transmitter design at the devices and the SIC-based semi-blind detection scheme at the BS. This section first introduces a unified transmitter design for sourced and unsourced RA, where the required modules are almost identical for both RA paradigms, as illustrated in Fig. 3. Therefore, our explanation mainly focuses on the sourced RA and the major differences between the two RA paradigms will be further clarified.

Fig. 3. The proposed unified transmitter design for sourced and unsourced RA: (a) Block diagram; (b) Data packet structure and frame structure.

## A. Transmitter Design for Sourced RA

For arbitrary active device with index $k$, its uplink access signal is generated based on the following key steps.

- **Step 1:** To eliminate the permutation ambiguity, a binary device ID sequence of $B_i = \lceil \log_2(K) \rceil$ bits is inserted at the head of the payload data packet to identify the $K$ devices. For the $k$th device, its ID sequence is provided as $\mathbf{b}_k^i = \text{dec2bin}(k)$, where the operator $\text{dec2bin}(\cdot)$ converts a decimal integer to its binary representation.

- **Step 2:** To verify the correctness of the detected ID bits, a $B_c$-bit CRC code $\mathbf{b}_k^c$ is added to the end of the device ID sequence, as $\mathbf{b}_k = \left[\mathbf{b}_k^i; \mathbf{b}_k^c; \mathbf{b}_k^d\right] \in \mathbb{C}^{B \times 1}$, where $\mathbf{b}_k^d \in \mathbb{C}^{B_d \times 1}$ is the payload data packet and $B = B_i + B_c + B_d$. The CRC code is generated as

$$\mathbf{b}_k^c = f\left(\left[\mathbf{b}_k^i; \mathbf{0}_{B_c \times 1}\right] \div \mathbf{p}_c\right), \tag{9}$$

where $\div$ denotes the binary (modulo-2) division, $\mathbf{p}_c$ is the generator polynomial of CRC, and $f(\cdot)$ is the function to compute the remainder of the binary division.

- **Step 3:** The overall data packet $\mathbf{b}_k$ is modulated by an $M$-order phase shift keying (PSK) modulator, where the modulated symbol sequence is defined as $\mathbf{x}_k^d \in \mathbb{C}^{(L-1) \times 1}$ with $L = \lceil B/\log_2(M) \rceil + 1$.

- **Step 4:** To eliminate the phase ambiguity, a known scalar pilot symbol $x_p$ is inserted at the head of the modulated symbol sequence, i.e., $\mathbf{x}_k = [x_p; \mathbf{x}_k^d] \in \mathbb{C}^{L \times 1}$, where $\mathbf{x}_k$ is the uplink access signal of the $k$th device to be transmitted. Here, $x_p$ is drawn from the constellation set of the adopted modulation scheme and is identical for all active devices. Note that since $\boldsymbol{\Sigma}$ is a diagonal matrix, the phase shifts of phase ambiguity are identical for all the transmitted symbols of a specific active device, but different for the symbol frames of different active devices. In this case, only one pilot symbol in each $\mathbf{x}_k$ is sufficient to estimate the phase shift matrix $\boldsymbol{\Sigma}$.

### B. Extension to Unsourced RA

The aforementioned transmitter design for sourced RA can be further extended to the unsourced RA, where the major difference lies in the structure of the data packet, see Fig. 3. For unsourced RA, the BS is solely interested in the list of the sent messages, without regard for the identities of individual sources, i.e., the permutation ambiguity could be ignored. Therefore, the device ID sequence is removed from the data packet for improved payload efficiency. Meanwhile, the CRC code is attached to the end of the payload data packet, as $\mathbf{b}_k = \left[ \mathbf{b}_k^d; \mathbf{b}_k^c \right] \in \mathbb{C}^{B \times 1}$ with $B = B_c + B_d$, and the generation of the CRC code is modified to

$$\mathbf{b}_k^c = f\left( [\mathbf{b}_k^d; \mathbf{0}_{B_c \times 1}] \div \mathbf{p}_c \right). \tag{10}$$

Different from unsourced RA, the CRC code in sourced RA is mainly used for evaluating the reliability of the detected device ID bits, which effectively avoids the whole packet loss due to the detection error of few payload data bits, thus dramatically reducing the probability of miss detection. Based on the proposed transmitter design, both sourced and unsourced RA could share the same hardware modules and only a software-defined switch is required to determine which data packet structure is adopted. Compared to the traditional detection frameworks detailed in *Section III*, the proposed unified transmitter design is more beneficial to satisfying the ultra-low latency requirement of massive URLLC due to the significantly improved payload efficiency.

## V. PROPOSED UNIFIED SEMI-BLIND DETECTION FRAMEWORK: RECEIVER DESIGN

Adopting the transmitter design proposed in *Section IV*, the inserted RI is insufficient to achieve reliable ADD and CE, which significantly degrades the performance of traditional coherent detection. In this section, we develop an SIC-based semi-blind detection scheme at the BS, where

the payload data of active devices is directly detected from the overlapped received signal without exploiting explicit channel state information. Specifically, we first propose a BiG-AMP-based JCSE algorithm, where the channel and signal matrices are jointly estimated by factorizing the noisy received signal, without regard for the phase and permutation ambiguities. In particular, a singular value decomposition (SVD)-based rank selection approach and a RI-aided initialization strategy are incorporated to reduce the computational complexity and to enhance the JCSE reliability, respectively, for the conventional BiG-AMP algorithm. Finally, the SIC-based semi-blind data detection scheme is developed, where the inserted RI is exploited to resolve the ambiguities and the SIC technique is utilized to mitigate the inter-device interference iteratively.

### A. SVD-Based Rank Selection

As clarified in *Section II*, only $K_a$ ($K_a \ll K$) active devices contribute to the received signal $\mathbf{Y}$, thus the signal model in (7) can be re-expressed as

$$\mathbf{Y} = [\mathbf{H}]_{:,\mathcal{A}} [\mathbf{X}]_{\mathcal{A},:} + \mathbf{N} = \mathbf{H}_{\text{act}} \mathbf{X}_{\text{act}} + \mathbf{N}. \tag{11}$$

Here, $\mathcal{A}$ is the active device set, $\mathbf{H}_{\text{act}} = [\mathbf{H}]_{:,\mathcal{A}} \in \mathbb{C}^{N \times K_a}$ and $\mathbf{X}_{\text{act}} = [\mathbf{X}]_{\mathcal{A},:} \in \mathbb{C}^{K_a \times L}$ represent the MRA channel matrix and the transmitted signal matrix associated with the active devices, respectively. For JCSE, our goal is to jointly infer the channel matrix $\mathbf{H}_{\text{act}}$ and the signal matrix $\mathbf{X}_{\text{act}}$ based on $\mathbf{Y}$. By exploiting the angular-domain sparsity of massive MIMO channels, as well as the statistical information of $\mathbf{H}_{\text{act}}$ and $\mathbf{X}_{\text{act}}$, the efficient BiG-AMP algorithm derived in [26] can be employed to achieve the goal, where the concerned problem is formulated as a matrix factorization problem. In practice, since the number of active devices $K_a$ is generally unknown in advance, a straightforward solution is to apply the BiG-AMP algorithm to the model (7), where $\mathbf{H}$ and $\mathbf{X}$ can be jointly estimated. Then, the estimates of $\mathbf{H}_{\text{act}}$ and $\mathbf{X}_{\text{act}}$ are obtained by removing the channels and the signals of the devices whose channel gains are smaller than a predefined threshold. However, this solution poses stringent requirements on the number of BS antennas and the length of uplink access signal, i.e., $N > K$ and $L > K$, which is impractical in massive URLLC with small data packets [28]. Meanwhile, the resulting computational complexity at each BiG-AMP iteration scales with the number of the total devices, i.e., $\mathcal{O}(NK + KL + NL)$ [29].

*Proposition 1:* When $N > K_a$ and $L > K_a$, the rank of the noiseless received signal $\mathbf{Z} = \mathbf{H}_{\text{act}} \mathbf{X}_{\text{act}}$ is $K_a$.

Fig. 4. The noisy received signal $\mathbf{Y}$ has a prominent peak in the pairwise ratios of its adjacent descending singular values, where $N = 512$ and $L = 274$ are considered.

*Proof:* Due to $\mathbf{Z} = \mathbf{H}_{\text{act}}\mathbf{X}_{\text{act}}$, the rank of $\mathbf{Z}$ satisfies the following inequalities, as

$$\text{rank}\left(\mathbf{Z}\right) \leq \min\left\{\text{rank}\left(\mathbf{H}_{\text{act}}\right), \text{rank}\left(\mathbf{X}_{\text{act}}\right)\right\} \tag{12}$$

and

$$\text{rank}\left(\mathbf{H}_{\text{act}}\right) + \text{rank}\left(\mathbf{X}_{\text{act}}\right) - K_a \leq \text{rank}\left(\mathbf{Z}\right), \tag{13}$$

where $\text{rank}(\cdot)$ denotes the rank of a matrix. On the one hand, the assumptions $N > K_a$ and $L > K_a$ lead to $\min\left\{\text{rank}\left(\mathbf{H}_{\text{act}}\right), \text{rank}\left(\mathbf{X}_{\text{act}}\right)\right\} \leq K_a$. Thus, the inequality (12) can be re-expressed as $\text{rank}\left(\mathbf{Z}\right) \leq K_a$. On the other hand, since the access signals of different active devices are generated independently, we have $\text{rank}\left(\mathbf{X}_{\text{act}}\right) = K_a$. Meanwhile, since the active devices are independently distributed in the BS coverage, their channels are linearly independent, which results in $\text{rank}\left(\mathbf{H}_{\text{act}}\right) = K_a$. Therefore, the inequality (13) can be re-expressed as $K_a \leq \text{rank}\left(\mathbf{Z}\right)$. At the point, the rank of $\mathbf{Z}$ is proofed to be $\text{rank}\left(\mathbf{Z}\right) = K_a$ by combining the inequalities in (12) and (13).

With $\text{rank}\left(\mathbf{Z}\right) = K_a$, the authors in [16] have revealed that the space of $\mathbf{Y} = \mathbf{Z} + \mathbf{N}$ can be divided into a noisy signal subspace and a pure noise subspace in high signal-to-noise ratio (SNR) cases. Specifically, by exploiting SVD, the noisy received signal is re-expressed as $\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\text{H}}$, where $\mathbf{U} \in \mathbb{C}^{N \times N}$ and $\mathbf{V} \in \mathbb{C}^{L \times L}$ are unitary matrices, $\boldsymbol{\Sigma} \in \mathbb{C}^{N \times L}$ is a diagonal matrix with $K_{\max}$ non-zero real numbers, i.e., the singular values of $\mathbf{Y}$, on the diagonal, and $K_{\max} = \min\left(N, L\right) > K_a$. Then, the signal subspace is constructed as $\mathbf{Y}_s = [\mathbf{U}]_{:,\mathcal{K}_s} [\boldsymbol{\Sigma}]_{\mathcal{K}_s,\mathcal{K}_s} [\mathbf{V}]_{\mathcal{K}_s,:}^{\text{H}} =$

$\mathbf{Z} + \mathbf{N}_s$, with $\mathcal{K}_s = \{1, 2, \cdots, K_a\}$ and $\mathbf{N}_s$ denoting the noise incorporated in the signal space. Meanwhile, the noise subspace is constructed as $\mathbf{Y}_n = [\mathbf{U}]_{:,\mathcal{K}_n} [\mathbf{\Sigma}]_{\mathcal{K}_n, \mathcal{K}_n} [\mathbf{V}]^{\mathrm{H}}_{\mathcal{K}_n,:} = \mathbf{N}_n$ with $\mathcal{K}_n = \{K_a + 1, \cdots, K_{\max}\}$ and $\mathbf{N}_n$ denoting the noise incorporated in the noise space. Particularly, the singular values of $\mathbf{Y}_s$ are considerably larger than those of $\mathbf{Y}_n$ as a relatively high SNR is considered. Therefore, the received signal $\mathbf{Y}$ has a prominent peak in the pairwise ratios of its adjacent descending singular values, i.e., $\left\{ [\mathbf{\Sigma}]_{k,k} / [\mathbf{\Sigma}]_{k+1,k+1} | \forall k \in [K_{\max} - 1] \right\}$, as illustrated in Fig. 4. Moreover, the singular value index corresponding to the maximum ratio is exactly $K_a$. Based on this remarkable characteristic, the number of active devices $K_a$ can be estimated via the following rank selection procedure,

$$\widehat{K}_a = \underset{k \in [K_{\max} - 1]}{\arg\max} \, [\mathbf{\Sigma}]_{k,k} / [\mathbf{\Sigma}]_{k+1,k+1} . \tag{14}$$

In this context, we can apply the BiG-AMP algorithm to model (11) to jointly estimate $\mathbf{H}_{\mathrm{act}}$ and $\mathbf{X}_{\mathrm{act}}$, where the dimension constraint relaxes to $N > K_a$ and $L > K_a$ with $K_a \ll K$, i.e., the considered problem is independent of the number of potential devices. Meanwhile, the computational complexity of each iteration of the BiG-AMP algorithm reduces to $\mathcal{O}\left(N K_a + K_a L + N L\right)$. This leads to the dramatically reduced processing latency, which is another key to guarantee the ultra-low user plane latency of massive URLLC.

*Remark 2:* In this paper, the considered JCSE problem is formulated based on the angular-domain signal model (7), rather than the spatial-domain model (2). Compared with the spatial-domain channel matrix $\mathbf{G}$, the angular-domain channel matrix $\mathbf{H}$ exhibits an enhanced sparsity, which dramatically reduces the number of unknown channel coefficients to be estimated. In this case, for a given number of measurements, the JCSE performance can be significantly improved by further leveraging the angular-domain sparsity of massive MIMO channels. The authors in [40] have revealed that the performance can be very close to the ideal case with perfect CSI as long as the channel matrix is sufficiently sparse.

*Remark 3:* For the cases with an extremely low SNR (e.g., $\mathrm{SNR} < 0$ dB) or an extremely large number of active devices (e.g., $K_a > 500$), the singular values of $\mathbf{Y}$ will decay smoothly, which makes the signal and noise subspaces indistinguishable. In this context, the proposed SVD-based rank selection approach fails to work. However, due to the sporadic uplink traffic of massive URLLC and the adaptive transmit power control, such extreme cases are rare to occur in practice.

## B. BiG-AMP-Based JCSE Algorithm

Next, we utilize the BiG-AMP algorithm to address the aforementioned matrix factorization problem, where the expectation maximization (EM) algorithm is incorporated to learn the unknown hyper-parameters and a RI-based initialization strategy is proposed to improve the estimation accuracy. Under the Bayesian inference framework, the detailed description of the BiG-AMP algorithm begins with the probabilistic model of the problem. Specifically, the minimum mean-square-error (MMSE) estimates of $\mathbf{H}_{\mathrm{act}}$ and $\mathbf{X}_{\mathrm{act}}$, denoted by $\widehat{\mathbf{H}}_{\mathrm{act}}$ and $\widehat{\mathbf{X}}_{\mathrm{act}}$, respectively, are expressed as

$$\left(\widehat{\mathbf{H}}_{\mathrm{act}}, \widehat{\mathbf{X}}_{\mathrm{act}}\right) = \mathbb{E}\left[\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}|\mathbf{Y}\right] = \iint p\left(\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}|\mathbf{Y}\right) d\mathbf{H}_{\mathrm{act}} d\mathbf{X}_{\mathrm{act}}, \tag{15}$$

where the joint posterior distribution is given as

$$\begin{aligned} p\left(\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}|\mathbf{Y}\right) &= \frac{p\left(\mathbf{Y}|\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}\right) p\left(\mathbf{H}_{\mathrm{act}}\right) p\left(\mathbf{X}_{\mathrm{act}}\right)}{p\left(\mathbf{Y}\right)} \\ &\propto p\left(\mathbf{Y}|\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}\right) p\left(\mathbf{H}_{\mathrm{act}}\right) p\left(\mathbf{X}_{\mathrm{act}}\right), \end{aligned} \tag{16}$$

with the notation $\propto$ denoting an equality up to a constant scaling factor. It is assumed that the elements of the noise matrix $\mathbf{N}$ are independently drawn from $\mathcal{CN}\left(0, \sigma^2\right)$. Hence, given $\mathbf{H}_{\mathrm{act}}$ and $\mathbf{X}_{\mathrm{act}}$, the likelihood function can be factorized into

$$\begin{aligned} p\left(\mathbf{Y}|\mathbf{H}_{\mathrm{act}}, \mathbf{X}_{\mathrm{act}}\right) &= \prod_{n=1}^{N} \prod_{l=1}^{L} p\left(y_{n,l}|z_{n,l} = \sum_{k=1}^{K_a} h_{n,k} x_{k,l}\right) \\ &= \prod_{n=1}^{N} \prod_{l=1}^{L} \frac{1}{\pi\sigma^2} \exp\left(-\frac{1}{\sigma^2}\left|y_{n,l} - z_{n,l}\right|^2\right), \end{aligned} \tag{17}$$

where the subscript "act" is omitted in $h_{n,k}$ and $x_{k,l}$ for notational simplicity. Meanwhile, we adopt the well-studied spike-and-slab *a priori* distribution to capture the sparse feature of the angular-domain channel matrix $\mathbf{H}_{\mathrm{act}}$, i.e.,

$$p\left(\mathbf{H}_{\mathrm{act}}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K_a} p\left(h_{n,k}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K_a} \left[\left(1 - \gamma_{n,k}\right) \delta\left(h_{n,k}\right) + \gamma_{n,k} \tilde{f}\left(h_{n,k}\right)\right], \tag{18}$$

where $0 \leq \gamma_{n,k} \leq 1$ denotes the sparsity ratio, i.e., the probability of $h_{n,k}$ being non-zero, $\delta\left(\cdot\right)$ is the Dirac delta function, $\tilde{f}\left(h_{n,k}\right) = \mathcal{CN}\left(h_{n,k}; \mu_{n,k}, \tau_{n,k}\right)$ is the *a priori* distribution of non-zero channel coefficients. This distribution has been widely applied in the literature for AMP-based MIMO channel estimation [17], which shows its effectiveness in modeling the *a priori* distribution of real-world MIMO channels Here, the channel coefficients associated with different BS antennas are assumed to be mutually independent. This assumption simplifies the

considered problem and facilitates the application of the efficient AMP inference framework with acceptable performance loss, as discussed in [17]. Note that although taking into account the correlation of different antennas may further enhance the performance, the corresponding algorithm would be much more involved. In addition, since the transmitted signals are randomly drawn from a finite constellation set $\mathcal{S}$, the *a priori* distribution of $\mathbf{X}_{\mathrm{act}}$ is provided as

$$p\left(\mathbf{X}_{\mathrm{act}}\right) = \prod_{k=1}^{K_a}\prod_{l=1}^{L} p\left(x_{k,l}\right) = \prod_{k=1}^{K_a}\prod_{l=1}^{L}\frac{1}{M}\sum_{m=1}^{M}\delta\left(x_{k,l}-s_m\right), \tag{19}$$

where $s_m \in \mathcal{S}, \forall m \in [M]$ are the constellation symbols. Benefitting from the factorizability of the likelihood function and *a priori* distributions, as in (17)-(19), the joint posterior distribution in (16) can be represented by a factor graph. In this context, the standard sum-product algorithm can operate to compute the means of the marginal posterior distributions $p\left(h_{n,k}|\mathbf{Y}\right)$ and $p\left(x_{k,l}|\mathbf{Y}\right)$ for all pairs $(n,k)$ and $(k,l)$, i.e., the solution of the problem in (15) [31]. However, for massive URLLC in massive MIMO systems, the exact implementation of the sum-product algorithm is impractical, as the large numbers of BS antennas and active devices make the related computational complexity prohibitive. To overcome this obstacle, the key idea of the BiG-AMP algorithm is to provide a low-complexity approximation of the sum-product algorithm by applying the central-limit theorem and Taylor-series approximations in the large system limits [26]. Intuitively, with the approximations, the matrix estimation problem in (15) can be decoupled into multiple independent scalar estimation problems, which avoids high-dimensional integrals and facilitates the practical implementation of the algorithm.

---

**Algorithm 1** BiG-AMP-Based JCSE Algorithm

---

**Input:** Angular-domain received signal $\mathbf{Y}$, the maximum number of iterations $U$, and termination threshold $\epsilon_{\mathrm{amp}}$.

**Output:** The estimates of the channel matrix $\mathbf{H}_{\mathrm{act}}$ and transmitted signal matrix $\mathbf{X}_{\mathrm{act}}$ associated with the active devices.

1: Determine the problem dimensions, i.e., $[N,L] = \mathrm{size}\left(\mathbf{Y}\right)$, and $K_a$ is estimated by (14).

2: $\widehat{\mathbf{H}}_{\mathrm{act}} = \mathbf{0}_{N\times\widehat{K}_a}$, $\widehat{\mathbf{X}}_{\mathrm{act}} = \mathbf{0}_{\widehat{K}_a\times L}$.

3: $\forall n,k$: Initialize the hyper-parameters $\sigma^2\left(1\right)$, $\mu_{n,k}\left(1\right)$, $\tau_{n,k}\left(1\right)$, and $\gamma_{n,k}\left(1\right)$.

4: $\forall n,k,l$: Initialize the marginal posterior means and variances of $\mathbf{H}_{\mathrm{act}}$ and $\mathbf{X}_{\mathrm{act}}$, i.e., $\widehat{h}_{n,k}\left(1\right)$, $v_{n,k}^h\left(1\right)$, $\widehat{x}_{k,l}\left(1\right)$, and $v_{k,l}^x\left(1\right)$.

5: **for** $u = 1,\cdots,U$ **do**

6:    // BiG-AMP Updates:

7:      $\forall n, l: \overline{v}_{n,l}^{p}\left(u\right) = \sum_{k=1}^{K_a} \left|\widehat{h}_{n,k}\left(u\right)\right|^2 v_{k,l}^{x}\left(u\right) + v_{n,k}^{h}\left(u\right)\left|\widehat{x}_{k,l}\left(u\right)\right|^2$

8:      $\forall n, l: \overline{p}_{n,l}\left(u\right) = \sum_{k=1}^{K_a} \widehat{h}_{n,k}\left(u\right) \widehat{x}_{k,l}\left(u\right)$

9:      $\forall n, l: v_{n,l}^{p}\left(u\right) = \overline{v}_{n,l}^{p}\left(u\right) + \sum_{k=1}^{K_a} v_{n,k}^{h}\left(u\right) v_{k,l}^{x}\left(u\right)$

10:      $\forall n, l: \widehat{p}_{n,l}\left(u\right) = \overline{p}_{n,l}\left(u\right) - \widehat{g}_{n,l}\left(u - 1\right)\overline{p}_{n,l}\left(u\right)$

11:      $\forall n, l: v_{n,l}^{z}\left(u\right) = \mathbb{V}\left[z_{n,l}|\widehat{p}_{n,l}\left(u\right), v_{n,l}^{p}\left(u\right)\right]$

12:      $\forall n, l: \widehat{z}_{n,l}\left(u\right) = \mathbb{E}\left[z_{n,l}|\widehat{p}_{n,l}\left(u\right), v_{n,l}^{p}\left(u\right)\right]$

13:      $\forall n, l: v_{n,l}^{g}\left(u\right) = \left(1 - v_{n,l}^{z}\left(u\right)/v_{n,l}^{p}\left(u\right)\right)/v_{n,l}^{p}\left(u\right)$

14:      $\forall n, l: \widehat{g}_{n,l}\left(u\right) = \left(\widehat{z}_{n,l}\left(u\right) - \widehat{p}_{n,l}\left(u\right)\right)/v_{n,l}^{p}\left(u\right)$

15:      $\forall n, k: v_{n,k}^{q}\left(u\right) = \left(\sum_{l=1}^{L}\left|\widehat{x}_{k,l}\left(u\right)\right|^2 v_{n,l}^{g}\left(u\right)\right)^{-1}$

16:      $\forall n, k: \widehat{q}_{n,k}\left(u\right) = v_{n,k}^{q}\left(u\right)\sum_{l=1}^{L}\widehat{x}_{k,l}^{*}\left(u\right)\widehat{g}_{n,l}\left(u\right) + \widehat{h}_{n,k}\left(u\right)\left(1 - v_{n,k}^{q}\left(u\right)\sum_{l=1}^{L} v_{k,l}^{x}\left(u\right) v_{n,l}^{g}\left(u\right)\right)$

17:      $\forall n, k: v_{n,k}^{h}\left(u + 1\right) = \mathbb{V}\left[h_{n,k}|\widehat{q}_{n,k}\left(u\right), v_{n,k}^{q}\left(u\right)\right]$

18:      $\forall n, k: \widehat{h}_{n,k}\left(u + 1\right) = \mathbb{E}\left[h_{n,k}|\widehat{q}_{n,k}\left(u\right), v_{n,k}^{q}\left(u\right)\right]$

19:      $\forall k, l: v_{k,l}^{r}\left(u\right) = \left(\sum_{n=1}^{N}\left|\widehat{h}_{n,k}\left(u\right)\right|^2 v_{n,l}^{g}\left(u\right)\right)^{-1}$

20:      $\forall k, l: \widehat{r}_{k,l}\left(u\right) = v_{k,l}^{r}\left(u\right)\sum_{n=1}^{N}\widehat{h}_{n,k}^{*}\left(u\right)\widehat{g}_{k,l}\left(u\right) + \widehat{x}_{k,l}\left(u\right)\left(1 - v_{k,l}^{r}\left(u\right)\sum_{n=1}^{N} v_{n,k}^{h}\left(u\right) v_{k,l}^{g}\left(u\right)\right)$

21:      $\forall k, l: v_{k,l}^{x}\left(u + 1\right) = \mathbb{V}\left[x_{k,l}|\widehat{r}_{k,l}\left(u\right); v_{k,l}^{r}\left(u\right)\right]$

22:      $\forall k, l: \widehat{x}_{k,l}\left(u + 1\right) = \mathbb{E}\left[x_{k,l}|\widehat{r}_{k,l}\left(u\right); v_{k,l}^{r}\left(u\right)\right]$

23:      // EM Updates:

24:      $\sigma^2\left(u + 1\right) = \frac{1}{NL}\sum_{n=1}^{N}\sum_{l=1}^{L}\left[\frac{\left|y_{n,l} - \widehat{p}_{n,l}\left(u\right)\right|^2}{1 + v_{n,l}^{p}\left(u\right)/\sigma^2\left(u\right)} + \frac{\sigma^2\left(u\right) v_{n,l}^{p}\left(u\right)}{\sigma^2\left(u\right) + v_{n,l}^{p}\left(u\right)}\right]$

25:      $\tau_{n,k}\left(u + 1\right) = \frac{\pi_{n,k}\left(u\right)\left[\left|\mu_{n,k}\left(u\right) - \widehat{d}_{n,k}\left(u\right)\right|^2 + v_{n,k}^{d}\left(u\right)\right]}{\pi_{n,k}\left(u\right)}$

26:      $\mu_{n,k}\left(u + 1\right) = \frac{\pi_{n,k}\left(u\right)\widehat{d}_{n,k}\left(u\right)}{\pi_{n,k}\left(u\right)}$

27:      **if** $\frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\left|\widehat{p}_{n,l}\left(u\right) - \widehat{p}_{n,l}\left(u - 1\right)\right|^2}{\sum_{n=1}^{N}\sum_{l=1}^{L}\left|\widehat{p}_{n,l}\left(u\right)\right|^2} \leq \epsilon_{\mathrm{amp}}$ **then**

28:        break;

29:      **end if**

30: **end for**

31: **return** $\widehat{\mathbf{H}}_{\mathrm{act}}, \widehat{\mathbf{X}}_{\mathrm{act}}$

---

The overall steps of the BiG-AMP-based JCSE algorithm are summarized in *Algorithm 1*. For completeness, we provide more detailed descriptions as follows. *Lines 7* and *8* acquire a

plug-in[1] estimate of the noiseless received signal $\mathbf{Z} = \mathbf{H}_{\text{act}}\mathbf{X}_{\text{act}}$, where the corresponding means $\{\overline{p}_{n,l}\}$ and variances $\{\overline{v}^{p}_{n,l}\}$ are computed in element-wise. *Lines 9* and *10* introduce the so called Onsager reaction term[2] (i.e., the last term on the right-hand side of the equation) to correct the the rough plug-in estimates, which further improves the estimation accuracy. With the obtained quantities $\{\widehat{p}_{n,l}\}$ and $\{v^{p}_{n,l}\}$, *lines 11* and *12* compute the marginal posterior means $\{\widehat{z}_{n,l}\}$ and variances $\{v^{z}_{n,l}\}$ of $\mathbf{Z}$. Specifically, the MMSE estimation of $\mathbf{Z}$ is decoupled into $NL$ independent scalar inference problems, i.e., $\widehat{p}_{n,l} = z_{n,l} + w^{z}_{n,l}, \forall n, l$, with $z_{n,l} \sim \mathcal{CN}\left(z_{n,l}; y_{n,l}, \sigma^{2}\right)$ and $w^{z}_{n,l} \sim \mathcal{CN}\left(w^{z}_{n,l}; 0, v^{p}_{n,l}\right)$. Therefore, *lines 11* and *12* are explicitly computed as

$$v^{z}_{n,l}(u) = \frac{\sigma^{2}(u)\, v^{p}_{n,l}(u)}{\sigma^{2}(u) + v^{p}_{n,l}(u)}, \tag{20}$$

$$\widehat{z}_{n,l}(u) = \frac{y_{n,l} v^{p}_{n,l}(u) + \sigma^{2}(u)\, \widehat{p}_{n,l}(u)}{\sigma^{2}(u) + v^{p}_{n,l}(u)}, \tag{21}$$

respectively. Subsequently, *lines 13* and *14* use the related posterior moments to compute the scaled residual $\{\widehat{g}_{n,l}\}$ and its inverse variances $\{v^{g}_{n,l}\}$. Finally, *lines 15* and *16* obtain an equivalent AWGN corrupted observation of the true $h_{n,k}$, i.e., $\widehat{q}_{n,k} = h_{n,k} + w^{h}_{n,k}, \forall n, k$, with $w^{q}_{n,k} \sim \mathcal{CN}\left(w^{q}_{n,k}; 0, v^{q}_{n,k}\right)$. Adopting the *a priori* distribution $p(h_{n,k})$ given in (18), the posterior distribution of $h_{n,k}$ is computed as

$$p\left(h_{n,k} | \widehat{q}_{n,k}(u), v^{q}_{n,k}(u)\right) = (1 - \pi_{n,k}(u))\, \delta(h_{n,k}) + \pi_{n,k}(u)\, \mathcal{CN}\left(h_{n,k}; \widehat{d}_{n,k}(u), v^{d}_{n,k}(u)\right), \tag{22}$$

where

$$v^{d}_{n,k}(u) = \frac{\tau_{n,k}(u)\, v^{q}_{n,k}(u)}{\tau_{n,k}(u) + v^{q}_{n,k}(u)}, \tag{23}$$

$$\widehat{d}_{n,k}(u) = \frac{\mu_{n,k}(u)\, v^{q}_{n,k}(u) + \tau_{n,k}(u)\, \widehat{q}_{n,k}(u)}{\tau_{n,k}(u) + v^{q}_{n,k}(u)}, \tag{24}$$

$$\mathcal{L} = \ln \frac{v^{q}_{n,k}(u)}{\tau_{n,k}(u) + v^{q}_{n,k}(u)} + \frac{|\widehat{q}_{n,k}(u)|^{2}}{v^{q}_{n,k}(u)} + \frac{|\widehat{q}_{n,k}(u) - \mu(u)|^{2}}{\tau_{n,k}(u) + v^{q}_{n,k}(u)}, \tag{25}$$

$$\pi_{n,k}(u) = \frac{\gamma_{n,k}(u)}{\gamma_{n,k}(u) + (1 - \gamma_{n,k}(u))\exp(-\mathcal{L})}. \tag{26}$$

---

[1]The plug-in principle is a technique used in the probability theory and statistics to approximately estimate a feature of a distribution (e.g., the expected value and the variance) that cannot be computed exactly. It is widely used in the theories of Monte Carlo simulation and bootstrapping [32].

[2]The Onsager reaction term has been extensively discussed in the context of AMP. For more details, please refer to reference [33].

Then, the posterior mean $\widehat{h}_{n,k}$ and variance $v_{n,k}^h$ of $h_{n,k}$ in *lines 17* and *18*, respectively, are explicitly given as

$$\widehat{h}_{n,k}\left(u+1\right) = \pi_{n,k}\left(u\right)\widehat{d}_{n,k}\left(u\right), \tag{27}$$

$$v_{n,k}^h\left(u+1\right) = \pi_{n,k}\left(u\right)\left[\left|\widehat{d}_{n,k}\left(u\right)\right|^2 + v_{n,k}^d\left(u\right)\right] - \left|\widehat{h}_{n,k}\left(u+1\right)\right|^2, \tag{28}$$

respectively. Similarly, the AWGN corrupted observation of the true $x_{k,l}$ and its variance are computed in *lines 19* and *20*. Based on $\widehat{r}_{k,l} = x_{k,l} + w_{k,l}^x, \forall k, l$, with $w_{k,l}^x \sim \mathcal{CN}\left(w_{k,l}^x; 0, v_{k,l}^r\right)$ and $p\left(x_{k,l}\right)$ in (19), the posterior mean $\widehat{x}_{k,l}$ and variance $v_{k,l}^x$ of $x_{k,l}$ in *lines 21* and *22*, respectively, are explicitly computed as

$$\widehat{x}_{k,l}\left(u+1\right) = \frac{\sum\limits_{m=1}^{M} s_m \exp\left[-\left|s_m\right|^2 - \frac{2\mathcal{R}\left(\widehat{r}_{k,l}^*(u)s_m\right)}{v_{k,l}^r(u)}\right]}{\sum\limits_{m=1}^{M} \exp\left[-\left|s_m\right|^2 - \frac{2\mathcal{R}\left(\widehat{r}_{k,l}^*(u)s_m\right)}{v_{k,l}^r(u)}\right]}, \tag{29}$$

$$v_{k,l}^x\left(u+1\right) = \frac{\sum\limits_{m=1}^{M} \left|s_m\right|^2 \exp\left[-\left|s_m\right|^2 - \frac{2\mathcal{R}\left(\widehat{r}_{k,l}^*(u)s_m\right)}{v_{k,l}^r(u)}\right]}{\sum\limits_{m=1}^{M} \exp\left[-\left|s_m\right|^2 - \frac{2\mathcal{R}\left(\widehat{r}_{k,l}^*(u)s_m\right)}{v_{k,l}^r(u)}\right]}, \tag{30}$$

respectively. Note that *lines 7-22* of *Algorithm 1* constitute the basic version of the BiG-AMP algorithm developed in [26]. Here, based on the likelihood function and the *a priori* distributions provided in (17)-(19), we re-derived the explicit expressions of the MMSE estimates of $\mathbf{H}_{\text{act}}$ and $\mathbf{X}_{\text{act}}$, i.e., (20)-(30). In this paper, we further introduce the following two mechanisms to improve the realizability and the estimation reliability of the algorithm.

*1) EM-Based Hyper-Parameter Learning:* The implementation of the BiG-AMP algorithm requires the full knowledge of the likelihood function $p\left(\mathbf{Y}|\mathbf{H}_{\text{act}}, \mathbf{X}_{\text{act}}\right)$ and the *a priori* distributions $p\left(\mathbf{H}_{\text{act}}\right)$ and $p\left(\mathbf{X}_{\text{act}}\right)$. In practice, only the families of these distributions are known in advance and the governing hyper-parameters $\boldsymbol{\xi} = \left\{\sigma^2, \mu_{n,k}, \tau_{n,k}, \gamma_{n,k}, \forall n, k\right\}$ are generally unknown to the BS. Therefore, the EM algorithm proposed in [34] is incorporated to iteratively learn the unknown hyper-parameters. Intuitively, each iteration of the EM algorithm consists of two steps: E-step computes the joint distribution of all involved variables given the current estimate of the hyper-parameters $\widehat{\boldsymbol{\xi}}\left(u\right)$; M-step re-estimates the hyper-parameters with the goal of maximizing the likelihood, as

$$\widehat{\boldsymbol{\xi}}(u+1) = \arg\max_{\boldsymbol{\xi}} \mathbb{E}\left[\ln p\left(\mathbf{H}_{\text{act}}, \mathbf{X}_{\text{act}}, \mathbf{Z}, \mathbf{Y}; \boldsymbol{\xi}\right) | \mathbf{Y}; \widehat{\boldsymbol{\xi}}(u)\right]$$

$$= \arg\max_{\boldsymbol{\xi}} \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K_a} \mathbb{E}\left[\ln p\left(h_{n,k}; \boldsymbol{\xi}\right) | \mathbf{Y}; \widehat{\boldsymbol{\xi}}(u)\right] \right. \tag{31}$$

$$+ \sum_{k=1}^{K_a} \sum_{l=1}^{L} \mathbb{E}\left[\ln p\left(x_{k,l}; \boldsymbol{\xi}\right) | \mathbf{Y}; \widehat{\boldsymbol{\xi}}(u)\right]$$

$$\left. + \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}\left[\ln p\left(y_{n,l}|z_{n,l}; \boldsymbol{\xi}\right) | \mathbf{Y}; \widehat{\boldsymbol{\xi}}(u)\right] \right\}.$$

Here, the factorizability of $p\left(\mathbf{H}_{\text{act}}\right)$, $p\left(\mathbf{X}_{\text{act}}\right)$, and $p\left(\mathbf{Y}|\mathbf{Z}\right)$ simplifies the computation of the joint distribution in (31). Moreover, instead of jointly optimizing all parameters in $\boldsymbol{\xi}$, we adopt the incremental update strategy from [35], where $\boldsymbol{\xi}$ is updated one element at a time and the other parameters are held constant. By setting the derivative of (31) with respect to one element of $\boldsymbol{\xi}$ to zero, the estimates of the hyper-parameters are provided in *lines 24-26* of *Algorithm 1*.

*2) RI-Aided Initialization Strategy:* For *lines 3* and *4* of *Algorithm 1*, the traditional random initialization strategy may lead the algorithm to converge to a local extremum of the mean-square-error function [17]. To avoid this situation, the authors in [29] proposed to initialize the algorithm multiple times and select the optimal pair of solutions as the final estimates, which improves the estimation accuracy but significantly increases the computational complexity. In this paper, we propose a more efficient RI-aided initialization strategy, where the transmitted reference signal for eliminating phase and permutation ambiguities also serves as a short pilot sequence to acquire an initial estimate of the MRA channel matrix. Specifically, the reference signal is composed of the modulated symbols of device ID bits and CRC bits, as well as the scalar pilot symbol. By stacking all devices' reference signals in rows as the pilot matrix, the angular-domain joint ADD and CE scheme proposed in [17] is employed to acquire the coarse estimates of the active device set and MRA channel matrix. On this basis, the transmitted signal matrix of active devices can be further estimated by the least squares (LS) method. The aforementioned processing has been detailed in *Section III-A*, i.e., non-orthogonal pilot-based coherent detection. Note that the hyper-parameters can be simultaneously estimated with the incorporated EM algorithm, as in [17] . In this context, the estimated channel matrix, signal matrix, and hyper-parameters are exploited as the initial estimates of the proposed BiG-GAMP-based JCSE algorithm.

## C. SIC-Based Semi-Blind Detection Scheme

With the BiG-AMP-based JCSE algorithm developed in *Section V-B*, we further propose an SIC-based semi-blind detection scheme, where the embedded RI is utilized for ambiguity elimination. Meanwhile, the SIC technique is incorporated to mitigate the inter-device interference iteratively, as summarized in *Algorithm 2*. In the $j$th SIC iteration, *line 6* computes the residual received signal $\widetilde{\mathbf{Y}}^j$ and the residual number of active devices $\widehat{K}_a^j$, where $\widehat{\mathcal{A}}^{j-1}$, $\widehat{\mathbf{H}}^{j-1}$, and $\widehat{\mathbf{X}}^{j-1}$ are the estimated active device set, channel matrix, and signal matrix in the last iteration, respectively. If all the active devices have been detected or the power of the residual received signal is small enough, the processing is terminated to avoid unnecessary iterations, see *lines 7-9*. In *line 10*, without regard for the phase and permutation ambiguities, we employ the BiG-AMP-based JCSE algorithm to jointly infer the residual channel and signal matrices, i.e., $[\mathbf{H}]_{:,\mathcal{A}^j}$ and $[\mathbf{X}]_{\mathcal{A}^j,:}$, respectively, based on the following model

$$\widetilde{\mathbf{Y}}^j = \mathbf{Y} - \widehat{\mathbf{H}}^{j-1}\widehat{\mathbf{X}}^{j-1} = [\mathbf{H}]_{:,\mathcal{A}^j}[\mathbf{X}]_{\mathcal{A}^j,:} + \widehat{\mathbf{N}} + \mathbf{N}, \tag{32}$$

with the estimation error of the last iteration given as

$$\widehat{\mathbf{N}} = \left([\mathbf{H}]_{:,\widehat{\mathcal{A}}^{j-1}} - \left[\widehat{\mathbf{H}}^{j-1}\right]_{:,\widehat{\mathcal{A}}^{j-1}}\right)\left([\mathbf{X}]_{\widehat{\mathcal{A}}^{j-1},:} - \left[\widehat{\mathbf{X}}^{j-1}\right]_{\widehat{\mathcal{A}}^{j-1},:}\right). \tag{33}$$

Here, $\mathcal{A}^j = \mathcal{A} - \widehat{\mathcal{A}}^{j-1}$ denotes the residual active devices to be detected.

---

**Algorithm 2** SIC-Based Semi-Blind Detection Scheme

---

**Input:** Angular-domain received signal $\mathbf{Y}$, the maximum number of SIC iterations $J$, and termination threshold $\epsilon_{\text{sic}}$.

**Output:** The estimated active device set $\widehat{\mathcal{A}}$, channel matrix $\widehat{\mathbf{H}}$, signal matrix $\widehat{\mathbf{X}}$, and binary data matrix $\widehat{\mathbf{B}}$.

1: Estimate the number of active devices based on (14).

2: // SIC Initializations:

3: $\widehat{\mathcal{A}}^0 = \emptyset$, $\widehat{\mathbf{H}}^0 = \mathbf{0}_{N \times K}$, $\widehat{\mathbf{X}}^0 = \mathbf{0}_{K \times L}$, $\widehat{\mathbf{B}}^0 = \mathbf{0}_{K \times B}$.

4: // SIC Loops:

5: **for** $j = 1, \cdots, J$ **do**

6: $\quad \widetilde{\mathbf{Y}}^j = \mathbf{Y} - \widehat{\mathbf{H}}^{j-1}\widehat{\mathbf{X}}^{j-1}$, $\widehat{K}_a^j = \widehat{K}_a - \left|\widehat{\mathcal{A}}^{j-1}\right|_c$

7: $\quad$ **if** $\widehat{K}_a^j \leq 0 \mathbin{||} \left\|\widetilde{\mathbf{Y}}^j\right\|_{\text{F}}^2 / \|\mathbf{Y}\|_{\text{F}}^2 < \epsilon_{\text{sic}}$ **then**

8: $\quad\quad$ break;

9: $\quad$ **end if**

10:     Employ the BiG-AMP-based JCSE algorithm to acquire the estimates of residual channel and residual signal matrices, i.e., $\widehat{\mathbf{H}}^j_{\text{act}}$ and $\widehat{\mathbf{X}}^j_{\text{act}}$, respectively, based on $\widetilde{\mathbf{Y}}^j$, where the phase and permutation ambiguities are ignored, as in *Algorithm 1*.

11:     // Phase Ambiguity Elimination:

12:     $\widehat{\Sigma} = \text{diag}\left( x_p / \left[ \widehat{\mathbf{X}}^j_{\text{act}} \right]_{:,1} \right),\ \widehat{\mathbf{H}}^j_{\text{act}} = \widehat{\mathbf{H}}^j_{\text{act}} \widehat{\Sigma}^{-1},\ \widehat{\mathbf{X}}^j_{\text{act}} = \widehat{\Sigma} \widehat{\mathbf{X}}^j_{\text{act}}$

13:     With $\widehat{\mathbf{X}}^j_{\text{act}}$, perform $M$-PSK demodulation to obtain the estimated binary data matrix $\widehat{\mathbf{B}}^j_{\text{act}}$.

14:     With $\widehat{\mathbf{B}}^j_{\text{act}}$, perform CRC to acquire the corresponding checking result $\mathbf{c}^j \in \mathbb{C}^{\widehat{K}^j_a \times 1}$.

15:     // Permutation Ambiguity Elimination:

16:     **for** $k = 1, \cdots, \widehat{K}^j_a$ **do**

17:         $\widehat{k} = \text{bi2dec}\left( \left[ \widehat{\mathbf{B}}^j_{\text{act}} \right]_{k, \mathcal{B}_i} \right),\ \mathcal{B}_i = \{1, \cdots, B_i\}$

18:         **if** $0 \leq \widehat{k} \leq K$ && $\mathbf{c}^j(k) == 0$ **then**

19:             $\widehat{\mathcal{A}}^j = \widehat{\mathcal{A}}^{j-1} \cup \widehat{k},\ \left[ \widehat{\mathbf{B}}^j \right]_{\widehat{k},:} = \left[ \widehat{\mathbf{B}}^j_{\text{act}} \right]_{k,:}$

20:             $\left[ \widehat{\mathbf{H}}^j \right]_{:,\widehat{k}} = \left[ \widehat{\mathbf{H}}^j_{\text{act}} \right]_{:,k},\ \left[ \widehat{\mathbf{X}}^j \right]_{\widehat{k},:} = \left[ \widehat{\mathbf{X}}^j_{\text{act}} \right]_{k,:}$

21:         **end if**

22:     **end for**

23: **end for**

24: **return** $\widehat{\mathcal{A}}^j,\ \widehat{\mathbf{H}}^j,\ \widehat{\mathbf{X}}^j,\ \widehat{\mathbf{B}}^j$

For a specific active device, it has been revealed in [36] that the phase shifts of phase ambiguity are identical for all transmitted symbols. Therefore, the phase ambiguity can be eliminated by computing the corresponding phase shift as in *line 12* of *Algorithm 2*, where $x_p$ is the common scalar pilot symbol inserted in the access signals for all the active devices. Given the estimated signal matrix with phase correction, i.e., $\widehat{\mathbf{X}}^j_{\text{act}}$, *line 13* further executes $M$-PSK demodulation to obtain the estimated binary data matrix $\widehat{\mathbf{B}}^j_{\text{act}}$. Moreover, leveraging the validity checking procedure and the inserted device ID bits, the permutation ambiguity is further resolved, as in *lines 16-21* of *Algorithm 2*. Specifically, the CRC is firstly adopted to validate the correctness of the detected ID bits, as

$$\mathbf{c}^j(k) = \hat{f}\left( \left[ \widehat{\mathbf{B}}^j_{\text{act}} \right]_{k, \mathcal{B}_i} \div \mathbf{p}_c \right), \forall k \in \left[ \widehat{K}^j_a \right], \tag{34}$$

where $\mathcal{B}_i$ is the index set of device ID bits, $\hat{f}(\cdot) = 1$ if the remainder of the binary division is non-zero and $\hat{f}(\cdot) = 0$ otherwise. For the $k$th detected active device, if all the device ID bits are correctly recovered, i.e., $\mathbf{c}^j(k) = 0$, the estimated device ID can be used to identify

Fig. 5. Block diagram of the proposed unified SIC-based semi-blind detection scheme at the BS.

the corresponding active device, which is added to the estimated active device set $\widehat{\mathcal{A}}^j$, and the corresponding estimates are updated, as in *lines 19 and 20*.

Obviously, the phase and permutation ambiguities can be effectively resolved based on the embedded RI, i.e., the device ID bits, the CRC bits, and the scalar pilot symbol. Meanwhile, the amount of RI scales logarithmically with the number of devices, as detailed in *Section IV*, which leads to a very small time resource consumption. Moreover, due to the fact that the reliable BiG-AMP-based JCSE makes the error propagation controllable, the proposed SIC-based semi-blind detection scheme can further enhance the detection reliability by mitigating the inter-device interference iteratively.

### D. Extension to Unsourced RA

As analyzed in *Section IV*, the major difference between sourced and unsourced RA lies in that the permutation ambiguity in the unsourced RA does not have to be resolved. Therefore, the proposed SIC-based semi-blind scheme can be directly applied to the unsourced RA by just making some minor modifications at the software level. Specifically, since the CRC is utilized to validate the correctness of the detected payload data bits rather than the detected ID bits, the validity checking procedure in *lines 14* is modified to

$$\mathbf{c}^j\left(k\right) = \hat{f}\left(\left[\widehat{\mathbf{B}}^j_{\text{act}}\right]_{k,\mathcal{B}_d} \div \mathbf{p}_c\right), \forall k \in \left[\widehat{K}^j_a\right], \tag{35}$$

where $\mathcal{B}_d$ is the index set of payload data bits. Moreover, the step for device identification, i.e., *line 17*, is removed. Finally, the update rules in *lines 19* and *20* are modified as

$$\widehat{\mathcal{A}}^j = \widehat{\mathcal{A}}^{j-1} \cup \left( \widehat{K}_a^{j-1} + k \right), \text{if } \mathbf{c}^j\left(k\right) == 0, \tag{36}$$

$$\left[\widehat{\mathbf{H}}^j\right]_{:,\widetilde{\mathcal{A}}^j} = \widehat{\mathbf{H}}^j_{\mathrm{act}}, \left[\widehat{\mathbf{X}}^j\right]_{:,\widetilde{\mathcal{A}}^j} = \widehat{\mathbf{X}}^j_{\mathrm{act}}, \left[\widehat{\mathbf{B}}^j\right]_{:,\widetilde{\mathcal{A}}^j} = \widehat{\mathbf{B}}^j_{\mathrm{act}}, \tag{37}$$

where $\widehat{\mathcal{A}}^j$ represents the active device set with correctly detected data bits and unknown identity, $\widehat{K}_a^{j-1}$ denotes the number of devices that have been detected in the $(j-1)$th SIC iteration, and $\widetilde{\mathcal{A}}^j = \left\{ k | \widehat{K}_a^{j-1} + 1 \leq k \leq \widehat{K}_a^{j-1} + \widehat{K}_a^j \right\}$. In this context, we propose a unified semi-blind data detection scheme at the BS, as shown in Fig. 5. Here, both RA paradigms share the same hardware modules and a software-defined switch is utilized to determine which RA mode is enabled. This facilitates more flexible network deployment and reduces the cost of network re-configuration.

### E. Computational Complexity Analysis

For the practical implementation, the processing latency mainly depends on the computational complexity of the adopted receive algorithm. In the non-orthogonal pilot-based coherent detection for sourced RA, the complexity of AMP-based joint ADD and CE is calculated by the function as

$$C_{\mathrm{amp}}\left(N, K, L_p, U_a\right) = U_a \left( 4NKL_p + 3KL_p + 16NL_p + 20NK \right), \tag{38}$$

and the complexity of LS-based coherent data detection is calculated as

$$C_{\mathrm{ls}}\left(N, K_a, L_d\right) = K_a^3 + NK_a^2 + NK_aL_d, \tag{39}$$

where $U_a$ is the number of AMP iterations. Therefore, the overall computational complexity is in the order of $\mathcal{O}\left[C_{\mathrm{amp}}\left(N, K, L_p, U_a\right) + C_{\mathrm{ls}}\left(N, K_a, L_d\right)\right]$. While in the common codebook-based non-coherent detection for unsourced RA, the computational complexity mainly stems from the AMP-based codeword detection, which is in the order of $\mathcal{O}\left[C_{\mathrm{amp}}\left(N, 2^B, L, U_a\right)\right]$.

The computational complexity of the proposed unified semi-blind detection framework is mainly composed of three parts. Specifically, the complexity of SVD-based rank selection is calculated as

$$C_{\mathrm{svd}}\left(N, L\right) = 2NL^2 + L^3 + L + NL, \tag{40}$$

the complexity of RI-aided initialization is calculated as

$$C_{\text{init}} \left( N, K, K_a, L, L_r, U_a \right) = C_{\text{amp}} \left( N, K, L_r, U_a \right) + C_{\text{ls}} \left( N, K_a, L - L_r \right), \tag{41}$$

and the complexity of BiG-AMP-based JCSE is calculated as

$$C_{\text{jcse}} \left( N, K_a, L, U \right) = U \left( N K_a + K_a L + N L \right). \tag{42}$$

Here, $L_r$ is the number of consumed symbol durations for transmitting the RI, which is expressed as

$$L_r = \left\lceil \frac{B_i + B_c}{\log_2 (M)} \right\rceil + 1. \tag{43}$$

Further considering the SIC procedure, the overall computational complexity of the proposed semi-blind detection framework is in the order of $\mathcal{O} \left[ C_{\text{sic}} \left( N, K, K_a, L, L_r, U_a \right) \right]$ with

$$\begin{aligned}
C_{\text{sic}} \left( N, K, K_a, L, L_r, U_a, U \right) &= \sum_{j=1}^{J} C_{\text{svd}} \left( N, L \right) + C_{\text{init}} \left( N, K, \widehat{K}_a^j, L, L_r, U_a \right) \\
&\quad + C_{\text{jcse}} \left( N, \widehat{K}_a^j, L, U \right) + C_{\text{res}} \left( N, \left| \widehat{\mathcal{A}}^{j-1} \right|_c, L \right),
\end{aligned} \tag{44}$$

where $C_{\text{res}}(N, |\widehat{\mathcal{A}}^{j-1}|_c, L) = N|\widehat{\mathcal{A}}^{j-1}|_c L$ is the complexity for computing the residual received signal, $\widehat{K}_a^j$ is the number of active devices to be estimated in the $j$th SIC iteration, $|\widehat{\mathcal{A}}^{j-1}|_c$ is the detected active devices in the $(j-1)$th SIC iteration, and $J$ is the number of SIC iterations. Since the RI-aided initialization strategy is only applicable for sourced RA, the complexity of RI-aided initialization, i.e., $C_{\text{init}} \left( N, K, K_a, L, L_r, U_a \right)$, should be removed from (44) when unsourced RA is considered.

## VI. SOURCED RA AND UNSOURCED RA COEXISTENCE SCHEMES

For simplicity, the previous descriptions on the proposed unified transceiver design assume that the BS provides only one of the sourced and unsourced RA services during a given time interval, and the RA mode may switch between sourced and unsourced RA in different time intervals. Meanwhile, the enabled RA mode is assumed to be known in advance at both the devices and the BS. These assumptions fail to consider the more general sourced RA and unsourced RA (SRA-URA) coexistence scenarios having unknown device access requirements. To this end, we further develop two SRA-URA coexistence schemes, where the aforementioned unified transceiver design can be directly applied by making minor software-level updates.

## A. Orthogonal SRA-URA Coexistence Scheme

Considering the devices with periodic uplink traffic and predictable access requirements, such as the sensors that need to report their data periodically, we first propose an orthogonal SRA-URA coexistence scheme. Specifically, for a specific time interval, the potential devices are divided into two groups, i.e., sourced and unsourced RA device groups, according to their practical access requirements. Meanwhile, the time-frequency resources reserved for grant-free RA are divided into multiple orthogonal resource blocks (RBs), which are then allocated to the two device groups for avoiding inter-group interferences. Here, due to the predictable uplink traffic and access requirements, the associations between the device groups and the orthogonal RBs are pre-configured. In this context, the received signals of sourced and unsourced RA are distinguishable at the BS. Therefore, the proposed unified transceiver design in *Sections IV* and *V* can be independently applied to all RBs for semi-blind data detection, and the RA mode can flexibly switch between sourced and unsourced RA according to the served device type in different RBs. The proposed orthogonal SRA-URA coexistence scheme is inspired by the traditional orthogonal frequency division multiple access (OFDMA) developed in the fourth-generation (4G) Long-Term Evolution (LTE), where the interferences among all the active devices are avoided through orthogonal resource allocation. The key difference lies in that only the access signals from different device groups are orthogonal, while the signals from the same device group are still overlapped on the same RB.

## B. Non-Orthogonal SRA-URA Coexistence Scheme

In practice, since a considerably number of devices may randomly access the network and change their access requirements, the application scenario of the aforementioned orthogonal SRA-URA coexistence scheme is still very limited. To overcome this limitation, we further propose a non-orthogonal SRA-URA coexistence scheme, where the access signals of both types of devices, i.e., sourced and unsourced RA devices, are directly transmitted exploiting the same time-frequency resources without uplink resource pre-allocation or scheduling. In this case, the received signals of sourced and unsourced RA are overlapped at the BS and unable to be separated. By exploiting the common receive modules of sourced and unsourced RA paradigms, i.e., *lines 1-14* of *Algorithm 2*, the overlapped received signals can be jointly processed to obtain the estimated data packets of active devices, while their adopted RA modes are still unavailable. To tackle this issue, we propose to add a one-bit mode indicator at the beginning of the data

packet, which serves as the reference information and indicates the adopted RA mode. Here, the mode indicator takes 1 for sourced RA and 0 for unsourced RA. Since there are two types of RA modes, only one-bit reference information is sufficient to identify which RA mode is adopted by the active devices. At this point, with the estimated data packets and corresponding mode indicators, the remaining steps of the proposed semi-blind detection scheme can be executed to acquire the final estimates of the payload data bits. Specifically, for a specific detected active device, if its RA mode is judged to be sourced RA, the corresponding estimated data packet is processed by *lines 16-22* of *Algorithm 2* for permutation ambiguity elimination and estimates update; otherwise, the estimated data packet is processed by (36) and (37) for estimates update, as detailed in *Section V* and illustrated in Fig. 5. It is clear that, based on the proposed unified transceiver design presented in *Sections IV* and *V*, the aforementioned non-orthogonal SRA-URA coexistence scheme can be realized by making minor software-level updates.

## VII. URLLC ENHANCEMENTS

According to previous discussions, the proposed unified semi-blind detection framework facilitates massive URLLC via simplifying the access scheduling, improving the payload efficiency, reducing the computational complexity, and enhancing the JCSE reliability. However, these are still not enough to satisfy the stringent latency and reliability requirements in the context of massive devices, e.g., over $99.99999\%$ reliability within 1 ms user plane latency for 32 bytes [27]. Indeed, it is generally difficult for a single grant-free MRA technique to satisfy these requirements and several key enabling techniques should be further integrated to achieve the goal.

### A. Multi-Carrier Deployment

The basic version of the proposed semi-blind detection framework considers the single-carrier transmission. In practical orthogonal frequency division multiplexing (OFDM) systems, it can be directly deployed by selecting one of the subchannels for grant-free MRA, while the remaining subchannels are reserved for other purposes, such as control signaling exchanges. Meanwhile, it can be easily extended to multi-carrier deployment, where the payload data bits are delivered in parallel at multiple subcarriers for further reduced transmission latency. Specifically, for each active device, its data packet is uniformly split into $N_{sc}$ sub-blocks, where $N_{sc}$ is the number of occupied subcarriers. Then, the device ID bits, the CRC bits, and the scalar pilot symbol are inserted in each sub-blocks to eliminate the ambiguities, as detailed in *Section IV*. On this basis,

Fig. 6. Multi-carrier deployment of the proposed semi-blind detection framework.

the single-carrier version of the proposed SIC-based semi-blind detection scheme is employed to detect the sub-blocks carried by different subcarriers. Finally, the original data packet is acquired by stitching the detected sub-blocks together according to the device ID. However, it is not efficient to insert the device ID bits in all the sub-blocks, which significantly degrades the payload efficiency. The authors in [18] have revealed that the subchannels across different subcarriers generally have a common sparsity pattern in the angular domain. Meanwhile, the AoAs of different devices are usually distinguishable. Considering this characteristic, we propose a more efficient deployment for the improved payload efficiency, where only one sub-block is selected to carry the device ID bits to eliminate the permutation ambiguity of the corresponding subchannels, as depicted in Fig. 6. While the permutation ambiguities of the remaining subchannels are resolved by leveraging the fact that the subchannels having a common angular-domain sparsity pattern belong to the same active device. This can be realized by various clustering algorithms, such as $K$-means algorithm in [37]. Based on the above descriptions, the complexity of the multi-carrier deployment of the proposed detection framework is calculated as

$$
\begin{aligned}
C_{\mathrm{mc}}\left(N, K, K_a, L, L_r, U_a, U, N_{\mathrm{sc}}\right) &= C_{\mathrm{sic}}\left(N, K, K_a, L_1, L_r, U_a, U\right) \\
&+ \left(N_{\mathrm{sc}} - 1\right)\left[C_{\mathrm{svd}}\left(N, L_2\right) + C_{\mathrm{jcse}}\left(N, K_a, L_2, U\right)\right],
\end{aligned}
\tag{45}
$$

where $L_{\mathrm{sc}} = (L - L_r)/N_{\mathrm{sc}}$ is the payload signal length at each subcarrier, $L_1 = L_{\mathrm{sc}} + L_r$ is the overall signal length of the first subcarrier, $L_2 = L_{\mathrm{sc}} + L_r - L_i$ is the overall signal length of the remaining subcarriers, and $L_i$ is the signal length for transmitting the device ID bits.

Fig. 7. The flexible frame structure introduced in 5G NR, where different SCSs affects the slot duration and TTI.

## B. Flexible Frame Structure

In addition to payload efficiency, the transmission time interval (TTI) also plays an important role in contributing to the transmission latency [27]. Hence, reducing TTI is another key to satisfying the ultra-low user plane latency of massive URLLC. In the 4G LTE, the subcarrier spacing (SCS) is fixed at 15 kHz and each TTI contains 14 OFDM symbols, leading to a TTI (equals to two slots) of 1ms. This is only the transmission time on the air interface. The overall user plane latency would be much larger than 1 ms by further considering other delay components. Therefore, the 5G New Radio (NR) has introduced a more flexible frame structure, where the TTI can be shortened by using the scalable SCS [38]. Specifically, each frame with 10 ms consists of 10 subframes and the number of slots within a certain subframe depends on the SCS, as illustrated in Fig. 7. Furthermore, each slot is composed of 14 OFDM symbols. By using different SCSs in 5G NR, different slot durations and TTIs are configurable. For example, 15 kHz SCS with 14 symbols spanning the entire 1 ms subframe corresponds to the LTE's configuration. While at 240 kHz SCS, 14 symbols are squeezed into a mini-slot with 62.5 $\mu$s, thus significantly reducing the transmission latency.

## C. Concurrent Access Mechanism

By reducing the transmission latency, the aforementioned multi-carrier deployment and flexible frame structure also create more retransmission opportunities within a target latency. As a result, various hybrid automatic repeat request (HARQ) transmission schemes, including re-active HARQ, $K$-repetition HARQ, and proactive HARQ, can be incorporated into grant-free MRA for further enhanced reliability [27]. However, the number of retransmission times is still very limited due to the stringent latency requirement. To overcome this limitation, we

further propose a concurrent access mechanism in this paper, which resorts to the relatively richer frequency resource for diversity gain. Specifically, the whole bandwidth is divided into multiple independent sub-bands. Moreover, the same payload data is repeatedly delivered in these sub-bands, where the proposed semi-blind detection framework is employed for each sub-band. In this context, the detection reliability could be dramatically improved due to the frequency diversity. More specifically, although an active device may be missed in a specific sub-band, it can be successfully detected in other sub-bands with a high probability. The overall computational complexity of the URLLC-enhanced semi-blind detection framework is in the order of $\mathcal{O}\left[N_b C_{\mathrm{mc}}\left(N, K, K_a, L, L_r, U_a, U, N_{\mathrm{sc}}\right)\right]$, where $N_b$ is the number of sub-bands.

### D. Adaptive Transmit Power Control

In the proposed semi-blind detection framework, the corresponding whole payload data packet would be lost if an active device is not successfully detected, i.e., miss detection. On the other hand, all the active devices have an identical transmit power. Due to the severe path loss, the active devices located in the cell edge generally suffer from a far smaller received SNR than those in the cell center. This leads to a high miss detection probability of active devices in the cell edge and becomes a major limiting factor for improving the data detection reliability [13]. To tackle this issue, we propose an adaptive transmit power control (ATPC), where the transmit power of the $k$th device is given as $P_k = P_{\max} d_k^{\tilde{\alpha}} / d_{\max}^{\tilde{\alpha}}$. Here, $P_{\max}$ is the maximum transmit power, $d_k$ is the distance between the $k$th device and the BS, $\tilde{\alpha}$ is the path loss decay exponent, and $d_{\max}$ is the cell radius. In this context, all the active devices will have a similar received SNR at the BS, which significantly improves the data detection reliability by reducing the miss detection probability.

## VIII. NUMERICAL RESULTS

This section conducts exhaustive Monte-Carlo simulations to assess the performance of the proposed unified semi-blind detection framework. We consider a grant-free massive URLLC scenario in massive MIMO systems, where a BS equipped with an $N$-antenna ULA is employed to serve $K$ single-antenna devices. The devices are uniformly distributed in the BS's coverage and only $K_a$ out of the total $K$ devices are active within any given time interval. The massive MIMO channels are generated as in (3). Moreover, considering the perfect synchronization between different devices, we further assume the device activity and the massive MIMO channels remain

TABLE I: Simulation Parameters [17]

| Parameter | Value |
|---|---|
| Number of potential devices $K$ | 500 |
| Number of BS antennas $N$ | 512 |
| Number of payload data bits $B_d$ | 100 |
| Number of CRC bits $B_c$ | 8 |
| Generator polynomial of CRC $\mathbf{p}_c$ | $x^8 + x^7 + x^6 + x^4 + x^2 + 1$ |
| Modulation order $M$ | 2 |
| Carrier frequency | 3.9 GHz |
| System bandwidth | 400 MHz |
| Number of MPCs $P$ | $\mathcal{U}(P; 31, 61)$ |
| Angular spread in degree | $10°$ |
| Complex gain of the MPCs $\beta_{k,p}$ | $\mathcal{CN}(\beta_{k,p}; 0, 1)$ |
| Maximum transmit power $P_{\max}$ | 35 dBm |
| Background noise power | -174 dBm/Hz |
| Number of SIC iterations $J$ | 3 |
| Number of BiG-AMP iterations $U$ | 500 |
| Termination threshold $\epsilon_{\mathrm{amp}}$ | $10^{-5}$ |
| Termination threshold $\epsilon_{\mathrm{sic}}$ | $10^{-5}$ |
| Device-to-BS distance $d_k$ in km | $\mathcal{U}(d_k; 0.1, 1)$ |
| Path loss in dB at the distance $d_k$ | $128.1 + 37.6\log_{10}(d_k)$ |

unchanged during the considered transmission duration. The assumed simulation parameters are provided in *Table I* unless otherwise specified. According to the practical access requirements, the system can flexibly switch to either sourced or unsourced RA mode, where the proposed unified transceiver design detailed in *Sections IV and V* is adopted. Here, we first focus on the basic version of the proposed semi-blind detection framework, then the effectiveness of the URLLC-enhanced version is further verified. All simulation results are obtained by averaging over 10000 independent channel realizations.

## A. Performance of Sourced RA

For sourced RA, the state-of-the-art non-orthogonal pilot-based coherent detection framework detailed in *Section III-A* is compared as the benchmark. Particularly, based on the angular-domain received pilot signal, the advanced generalized MMV-AMP algorithm is employed for joint ADD and CE, as in [17]. The length of non-orthogonal pilot sequence in the baseline scheme is set

(a)                                                      (b)                                                      (c)

Fig. 8. Activity detection performance comparison of the traditional non-orthogonal pilot-based coherent detection framework and the proposed semi-blind detection framework under different numbers of BS antennas $N$: (a) AER performance; (b) Miss detection probability; (c) False alarm probability.

to $L_p = L_r$ for comparison fairness. For performance evaluation, we consider the activity error rate (AER) of ADD and the bit error rate (BER) of data detection, which are defined as

$$\text{AER} = \frac{\left|\mathcal{A} - \widehat{\mathcal{A}}\right|_c + \left|\widehat{\mathcal{A}} - \mathcal{A}\right|_c}{K}, \tag{46}$$

$$\text{BER} = \frac{\left\|[\mathbf{B}]_{\mathcal{A} \cap \widehat{\mathcal{A}}, \mathcal{B}_d} - [\widehat{\mathbf{B}}]_{\mathcal{A} \cap \widehat{\mathcal{A}}, \mathcal{B}_d}\right\|_0 + B_d \left|\mathcal{A} - \widehat{\mathcal{A}}\right|_c}{K_a B_d}, \tag{47}$$

respectively. Here, the set $\mathcal{B}_d = \{B_i + B_c + 1, \cdots, B\}$ with $|\mathcal{B}_d|_c = B_d$ denotes the column indexes corresponding to the payload data bits in the binary data matrix $\mathbf{B}$. The AER takes both miss detection and false alarm into account, as expressed in the numerator of (46). The BER also consists of two parts: the number of error bits due to the failure of symbol detection and the number of bits that are lost due to the miss detection, cf. the numerator of (47).

In Figs. 8 and 9, we first compare the sourced RA performances of the conventional non-orthogonal pilot-based coherent detection framework and the proposed semi-blind detection framework. To validate the most fundamental superiority of the proposed detection framework, the SIC procedure is disabled to exclude the performance gain provided by SIC. Meanwhile, both detection frameworks occupy the same number of time-frequency resources for comparison fairness. Meanwhile, both detection frameworks occupy the same number of time-frequency resources for comparison fairness. As can be observed, the AER and BER performances of all the considered schemes degrade as the number of active devices increases. This is because a larger number of active devices indicates severer inter-device interferences and a larger number of unknown variables to be estimated. Note that only the devices whose estimated data packet passes

Fig. 9. Data detection performance comparison of the traditional non-orthogonal pilot-based coherent detection framework and the proposed semi-blind detection framework under different numbers of BS antennas $N$: (a) Overall BER; (b) Symbol detection error rate.

the CRC will be identified as the active devices. Therefore, in the proposed semi-blind detection framework, the miss detection probability is generally larger than the false alarm probability. Meanwhile, the proposed semi-blind detection framework achieves much better AER and BER performance than the baseline scheme. This verifies the superiority of the proposed detection framework in combating inter-device interferences when the same number of physical resources is consumed for grant-free MRA. As for the baseline scheme, the length of non-orthogonal pilot sequence is too short to realize satisfactory ADD and CE performance, which leads to an inaccurate signal matrix estimate in (6) and becomes the major limiting factor of the reliable data detection. The performance of the baseline scheme can be improved by increasing the pilot length, but the payload efficiency would be significantly degraded, especially for massive URLLC with short data packets. While for the proposed semi-blind detection framework, the channel and signal matrices are jointly estimated via the advanced BiG-AMP algorithm, which does not rely heavily on the length of pilot sequence, thus reaps a better performance. Besides, the sourced RA performance of both detection frameworks becomes better as the number of BS antennas increases. This is because a larger number of BS antennas indicates the enhanced angular-domain sparsity of the MRA channel matrix, which leads to the improved ADD and CE performance in the pilot phase. In this context, a more accurate initialization for JCSE is available and the BiG-AMP algorithm will converge to the global optimum with a higher probability and

Fig. 10. Sourced RA performance of the proposed semi-blind detection framework under different numbers of SIC iterations $J$ and initialization strategies.: (a) AER performance; (b) BER performance.



Fig. 11. Convergence of the proposed BiG-AMP-based JCSE algorithm and the proposed SIC-based semi-blind detection scheme, where $K_a = 80$ is considered: (a) BiG-AMP iteration; (b) SIC iteration.

a faster speed. On the other hand, more spatial observations are available by equipping more antennas at the BS, which further improves the JCSE performance. Indeed, the BS must have a sufficient number of measurement samples, i.e., $N \gg K_a$, to avoid over-fitting, thus allowing more flexible and richer channel matrix estimates [39], [40]. Therefore, we conclude that the massive MIMO shows great benefits in grant-free MRA and makes the semi-blind detection framework practical.

Fig. 12. Sourced RA performance comparison of the traditional coherent detection framework and the proposed semi-blind detection framework under different modulation orders $M$: (a) AER performance; (b) BER performance.

In Fig. 10, we further investigate the effectiveness of the proposed SIC-based semi-blind detection scheme and the RI-aided initialization strategy. Note that the SIC procedure is disabled when the number of SIC iterations is set to $J = 1$. It is clear that the proposed semi-blind detection framework reaps better AER and BER performance as the number of SIC iterations increases, and only $J = 3$ iterations are sufficient to converge. This is due to the fact that the reliable BiG-AMP-based JCSE makes the error propagation of SIC controllable. Meanwhile, as the SIC iterations proceed, the signal components associated with the reliably detected active devices, i.e., whose detected device ID bits passes the CRC, are removed from the received signal, which mitigates the inter-device interference in the following SIC iterations. On this basis, the activity and the payload data of residual active devices can be detected with improved reliability. Moreover, the proposed RI-aided initialization strategy outperforms the traditional random initialization strategy. This demonstrates that the traditional BiG-AMP algorithm using the traditional random initialization is easy to stuck in the local extremum, while the proposed RI-aided initialization can guarantee the algorithm to converge to the global optimum.

Fig. 11 validates the convergence of the proposed BiG-AMP-based JCSE algorithm and the proposed SIC-based semi-blind detection scheme, where only $U = 100$ and $J = 3$ are sufficient to converge. Fig. 12 compares the sourced RA performances of the traditional non-orthogonal pilot-based coherent detection framework and the proposed semi-blind detection framework, where $N = 128$ and different modulation orders are considered. It is clear that the AER and

Fig. 13. Sourced RA performance of the proposed semi-blind detection framework for different numbers of CRC bits $B_c$: (a) AER performance; (b) BER performance.

BER performances of both detection frameworks degrade when a higher modulation order is adopted. This is because a higher modulation order leads to smaller Euclidean distances among the constellations and a worse RI-aided initialization. However, the proposed detection framework still achieves a much better performance than the traditional coherent detection framework dedicated to sourced RA.

In Fig. 13, the AER and BER performances of the proposed semi-blind detection framework for different numbers of CRC bits $B_c$ are also studied. In the simulations, $B_c \in [4, 8, 16]$ are investigated, where the corresponding $L_p = L_r$ with $L_r \in [14, 18]$ are considered in the baseline scheme for comparison fairness. The relationship between $B_c$ and $L_r$ is provided in (43). The generator polynomials of 4-bit and 16-bit CRC codes are given as $x^4 + x^3 + x^2 + x + 1$ and $x^{16} + x^{15} + x^2 + 1$, respectively. Meanwhile, the payload efficiency of the proposed semi-blind detection framework is defined as

$$\mathrm{PE}_{\mathrm{sbd}} = \frac{B_d}{\lceil B/\log_2(M) \rceil + 1}, \tag{48}$$

that is, the number of payload data bits to the number of consumed symbol durations for transmission. As shown in the Fig. 13 and (48), a larger $B_c$ improves both AER and BER performance, but at the expense of payload efficiency. Meanwhile, the performance gain is limited when $B_c > 8$. With the limited number of CRC bits, the generator polynomial can be optimized to fulfill a specified error detection performance [41].

(a)                                                                (b)

Fig. 14. The user-plane latency comparison of the traditional coherent detection framework, the basic semi-blind detection framework, and the URLLC-enhanced semi-blind detection framework, where $N = 512$, $K_a = 50$, and $B_d = 256$ are considered: (a) Transmission latency; (b) Processing latency.



(a)                                                                (b)

Fig. 15. Data detection performance comparison of the basic version and the URLLC-enhanced version of the proposed semi-blind detection framework for $N = 512$ and $B_d = 256$: (a) BER under different numbers of active devices; (b) BER under different maximum transmit powers with $K_a = 50$.

The aforementioned simulation results focus the basic version of the proposed semi-blind detection framework, which is still not enough to satisfy the stringent latency and reliability requirements of massive URLLC. Therefore, the enabling techniques introduced in *Section VII* should be further integrated to obtain a URLLC-enhanced version of the proposed detection framework. Fig. 14 compares the user-plane latency of the traditional non-orthogonal pilot-

based coherent detection for sourced RA, the basic semi-blind detection framework, and the URLLC-enhanced semi-blind detection framework. The user plane latency mainly consists of the transmission latency, propagation latency, and receive processing latency. The propagation latency only depends on the distance between the device and the BS, which is at a maximum of 3.3 $\mu$s for a cell radius of 1 km. Therefore, the propagation latency is identical for all the considered schemes. The processing latency generally depends on the computational complexity of the receive algorithm and the computing power of the processing unit. Since the available computing resources are identical for all the considered schemes, we mainly analyze the processing latency in terms of the computational complexity, which is quantified by the number of required complex multiplications. As can be observed, by exploiting the multi-carrier deployment, the proposed URLLC-enhanced semi-blind detection framework achieves a significantly reduced transmission latency than its counterparts. Moreover, the basic semi-blind detection framework has a slightly higher computational complexity than the non-orthogonal pilot-based coherent detection. Meanwhile, for the URLLC-enhanced version, the complexity will further increase by incorporating the concurrent access mechanism. However, it should be noted that the complexities of the all considered schemes are in the same order of magnitudes. Considering the rapid development of high-performance processing units and the rich computing resources at the BS, the increased processing latency is expected to be very minor. Since the traditional coherent detection framework and the proposed basic semi-blind detection framework occupy the same number of time-frequency resources, they have an identical transmission latency. Hence, the 1 ms user-plane latency can be satisfied as long as the processing latency is less than 0.4 ms. Fig. 15 verifies the effectiveness of the proposed concurrent access mechanism and ATPC in improving data detection reliability. It is observed that the URLLC-enhanced version effectively satisfy the 99.99999% reliability requirement when $K_a \leq 65$. Moreover, increasing the maximum transmit power further improves the detection reliability.

## B. Performance of Unsourced RA

For unsourced RA, we employ the most widely studied common codebook-based non-coherent detection framework as the benchmark, which was briefly introduced in *Section III-B* and detailed in [19]. Moreover, the coupled CS-based coding scheme is adopted for reducing the computational complexity, where the concatenated coding is utilized to couple an outer tree code and an inner CS code [22]. Specifically, for each active device, its payload data of $B = 180$ bits is

Fig. 16. PUPE performance comparison of the proposed semi-blind detection framework and the conventional common codebook-based non-coherent detection framework.



Fig. 17. The user-plane latency comparison of the traditional non-coherent detection framework, the basic semi-blind detection framework, and the URLLC-enhanced semi-blind detection framework: (a) Transmission latency; (b) Processing latency.

non-uniformly divided into $Q = 32$ fragments and the length of the $q$th fragment is $b_q$ satisfying $\sum_{q=1}^{Q} b_q = B$. To realize the coupling of different fragments, $a_q$ redundant parity check bits are added to the end of each fragment to form a sub-block with a fixed-length of $\widetilde{B} = b_q + a_q = 12$. Considering the typical simulation setup in [22], the parity profile $\mathbf{a} = [a_1, a_2, \cdots, a_Q]$ is set to $\mathbf{a} = [0, 6, \cdots, 6, 12, 12, 12]$. Subsequently, the inner encoder maps each sub-block to a codeword of a common codebook with size $L \times 2^{\widetilde{B}}$, which will be transmitted over $L$ successive symbol durations. Due to the absence of device identification, i.e., the permutation ambiguity is ignored,

the widely adopted RA performance metrics including AER and BER are not available. In this paper, the unsourced RA performance is evaluated in terms of per-user probability of error (PUPE), defined as the average fraction of transmitted messages not contained in the detected message list, i.e.,

$$\text{PUPE} = \frac{K_a - \left|\mathcal{M} - \widehat{\mathcal{M}}\right|_c}{K_a}, \tag{49}$$

where $\widehat{\mathcal{M}} = \left\{[\widehat{\mathbf{B}}]_{k,\mathcal{B}_d} | k \in [K_a]\right\}$ and $\mathcal{M} = \{[\mathbf{B}]_{k,\mathcal{B}_d} | k \in \mathcal{A}\}$ are the detected message list and the transmitted message list, respectively.

Fig. 16 examines the unsourced RA performance of the proposed semi-blind detection framework and the traditional common codebook-based non-coherent detection framework, where the codeword lengths $L \in [45, 50, 55]$ are considered and the corresponding payload efficiencies are indicated. The payload efficiency is defined as the number of transmitted payload data bits to the total number of consumed symbol durations, which turns out to be

$$\text{PE}_{\text{ncd}} = \frac{B_d}{QL}, \tag{50}$$

for the traditional common codebook-based non-coherent detection and

$$\text{PE}_{\text{sbd}} = \frac{B_d}{\lceil B/\log_2(M) \rceil + 1}, \tag{51}$$

for the proposed semi-blind detection. As shown in Fig. 16, the proposed semi-blind detection framework significantly outperforms the baseline scheme and offers a much better payload efficiency. This is because, in the baseline scheme, a large proportion of time resources are consumed to transmit the redundant parity check bits to guarantee the reliable stitching of the message fragments. Indeed, even under a very low payload efficiency, e.g., $\text{PE}_{\text{ncd}} = 10.23\%$, the corresponding number of measurements, i.e., the codeword length $L$, is too small to achieve the reliable CS-based inner decoding for the baseline scheme. Although the performance can be further improved by increasing the codeword length, the payload efficiency would continue to degrade, as in (50). While for the proposed detection framework, only a small proportion of time resources are consumed to transmit a $B_c$-bit CRC code and a scalar pilot symbol, leading to a payload efficiency over $90\%$. Therefore, compared to the baseline scheme, the proposed semi-blind detection framework has a better tradeoff between the transmission latency and the detection reliability. Fig. 17 compares the user-plane latencies of the traditional non-coherent detection framework, the basic semi-blind detection framework, and the URLLC-enhanced semi-blind

detection framework. Here, the transmission latency of the traditional non-coherent detection framework is computed as

$$d_{\mathrm{ncd}} = \frac{QL}{N_{\mathrm{scs}}},$$

(52)

where $N_{\mathrm{scs}}$ is the subcarrier spacing. While for the proposed unified detection framework, the transmission latency for unsourced RA is given as

$$d_{\mathrm{sbd}} = \frac{\lceil (B_d + B_c)/\log_2(M) \rceil + 1}{N_{\mathrm{scs}} N_{\mathrm{sc}}}.$$

(53)

The subcarrier spacing is set to 240 kHz for URLLC-enhanced scheme and 15 kHz for other schemes. Compared with the traditional non-coherent detection framework, the proposed unified semi-blind detection framework significantly reduces the transmission latency and computational complexity thus leading to a much smaller user-plane latency.

## IX. CONCLUSION

This paper has proposed a unified semi-blind detection framework for grant-free sourced and unsourced RA, which effectively facilitates the massive URLLC in massive MIMO systems. Under this framework, the system can flexibly switch to either sourced or unsourced RA mode according to the practical heterogeneous access requirements, making the network configuration more efficient and economical. By leveraging the large spatial degrees-of-freedom offered by the massive MIMO BS, we have developed a BiG-AMP-based JCSE algorithm to jointly infer the channel and signal matrices, where a rank selection approach and a RI-aided initialization strategy are incorporated for the reduction of computational complexity and the improvement of estimation reliability, respectively. Moreover, a small amount of RI is embedded in the access signal to eliminate the inherent phase and permutation ambiguities and the SIC technique has been introduced for further enhanced detection reliability. Besides, the four enabling techniques have also been integrated to satisfy the stringent latency and reliability requirements of massive URLLC. Numerical results have revealed that the proposed semi-blind detection framework offers a much better scalability-latency-reliability tradeoff than its counterparts dedicated to either sourced or unsourced RA, and thus it is more attractive for supporting massive URLLC.

## REFERENCES

[1] M. Ke, Z. Gao, S. Tan *et al.*, "Massive MIMO-enabled semi-blind detection for grant-free massive connectivity," in *Proc IEEE Int. Wireless Commun. Mobile Comput. (IWCMC)*, Dubrovnik, Croatia, July 2022, pp. 38-43.

[2] D. C. Nguyen, M. Ding, P. N. Pathirana *et al.*, "6G Internet-of-Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359-383, Jan. 2022.

[3] P. Popovski, C. Stefanovic, J. N. Jimmy *et al.*, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783-5801, Aug. 2019.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134-142, Oct. 2019.

[5] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765-55779, Sept. 2019.

[6] S. R. Pokhrel, J. Ding, J. Park, O.-S Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796-131813, Jul. 2020.

[7] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615-637, Mar. 2021.

[8] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148-156, Aug. 2020.

[9] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4-16, 1st Quart., 2018.

[10] Z. Zhang, X. Wang, Y. Zhang, and Y. Chen, "Grant-free rateless multiple access: A novel massive access scheme for Internet-of-Things," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2019-2022, Oct. 2016.

[11] X. Shao, X. Chen, C. Zhong, J. Zhao, and Z. Zhang, "A unified design of massive access for cellular Internet-of-Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3934-3947, Apr. 2019.

[12] X. Shao, X. Chen, D. W. K. Ng, C. Zhong, and Z. Zhang, "Cooperative activity detection: Sourced and unsourced massive random access paradigms," *IEEE Trans. Signal Process.*, vol. 68, pp. 6578-6593, Nov. 2020.

[13] M. Ke, Z. Gao, Y. Huang, G. Ding, D. W. K. Ng, Q. Wu, J. Zhang, "An edge computing paradigm for massive IoT connectivity over high-altitude platform networks," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 102-109, Oct. 2021.

[14] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972–974, July 2012.

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

[16] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420-435, Dec. 2020.

[17] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, Jan. 2020.

[18] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169-6183, Dec. 2015.

[19] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2523-2527.

[20] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access Gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2528-2532.

[21] A. Vem, K. R. Narayanan, J. F. Chamberland, and J. Cheng, "A user-independent successive interference cancellation based coding scheme for the unsourced random access Gaussian channel," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8258-8272, Dec. 2019.

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

[23] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, "Massive MIMO unsourced random access," [Online]. Available: https://arxiv. org/abs/1901.00828, Jan. 2019.

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

[25] V. Shyianov, F. Bellili, A. Mezghani, and E. Hossain, "Massive unsourced random access based on uncoupled compressive sensing: Another blessing of massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 820-834, Mar. 2021.

[26] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing-Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839-5853, Nov. 2014.

[27] J. Ding, M. Nemati, S. R. Pokhrel, O.-S. Park, J. Choi, and F. Adachi, "Enabling grant-free URLLC: An overview of principle and enhancements by massive MIMO," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 384-400, Jan. 2022.

[28] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing-Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854-5867, Nov. 2014.

[29] W. Yan and X. Yuan, "Semi-blind channel-and-signal estimation for uplink massive MIMO with channel sparsity," *IEEE Access*, vol. 7, pp. 95008-95020, July 2019.

[30] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Programm. Comput.*, vol. 4, pp. 333-361, July 2012. [Online]. Available: http://dx.doi.org/10.1007/s12532-012-0044-1.

[31] F. R. Kschischang, B. J. Frey, and H-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 498-519, Feb. 2001.

[32] A. W. Van der Vaart, "Asymptotic Statistics," *Cambridge University Press*, 1998.

[33] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Inf. Theory Workshop. (ITW)*, Jan. 2010, pp. 1-5.

[34] A. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1-17, 1977.

[35] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models.*, Springer, 1998, pp. 355-368.

[36] H. Q. Ngo and E. G. Larsson, "EVD-based channel estimation in multicell multiuser MIMO systems with very large antenna arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 3249-3252.

[37] X. Xie, Y. Wu, J. An *et al.* , "Massive unsourced random access: Exploiting angular domain sparsity," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2480-2498, Apr. 2022.

[38] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1-6.

[39] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Proc. JMLR: Workshop Conf. Proc. 25th Annu. Conf. Learn. Theory*, 2012, vol. 23, pp. 37.1-37.18.

[40] J. Zhang, X. Yuan, and Y. J. A. Zhang, "Blind signal detection in massive MIMO: Exploiting the channel sparsity," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 700-712, Feb. 2018.

[41] T. S. Baicheva, "Determination of the best CRC codes with up to 10-bit redundancy," *IEEE Trans. Commun.*, vol. 56, no. 8, pp. 1214-1220, Aug. 2008.