On Lattice Network Coding based Cell-free MIMO with Uncoordinated Base Stations

Tao Yang, *Member, IEEE*

***Abstract*—This paper studies a cell-free MIMO (cf-MIMO) system that consists of a central unit (CU), N distributed base stations (BSs) and K users. The BSs are connected to CU via N independent backhaul (BH) links of finite capacities. The BSs are uncoordinated, i.e. each one is ignorant of the others. We study a lattice network coding (LNC) scheme for cf-MIMO. In the uplink, the K users encode their messages with a practical 2^m-ary channel code mapped to 2^m-PAM, belonging to the ensemble of *lattice codes*, and transmit simultaneously. Each BS computes independent streams of integer-combinations of the users' messages, referred to as *network coding* (NC) streams. The NC streams are forwarded to the CU via BH. The CU aggregates all NC streams from the N BSs, and recovers K users' messages. We derive the achievable symmetric rate of the LNC scheme. Further, we solve a "*bounded independent vectors problem*" (BIVP) which identifies a near-optimal set of NC coefficient vectors. The solution maximizes the number of correctly computed NC streams at each BS. For practical implementation, we develop new soft detection algorithms for LNC decoding. The per-user complexity is proved to be no greater than $O(K)$, suitable for cf-MIMO with a large number of users. Our developed LNC based cf-MIMO exhibits superior frame error rate (FER) and spectral efficiency over existing non-LNC based schemes. Such advances are achieved with low-cost parallel processing and efficient usage of the BH.**

***Index Terms*—Cell-free, distributed MIMO, multi-user MIMO, massive access, coded modulation, lattice-codes, compute-forward, physical-layer network coding, soft detection, NOMA**

I. Introduction

In conventional cellular networks, inter-cell interference (ICI) has been deemed as a major limiting factor. Cell-free MIMO (cf-MIMO), also called distributed MIMO, has been proposed to overcome the ICI [1], [2]. A generic cf-MIMO system deploys a number of distributed base stations (BSs) which are connected to a central unit (CU) [3]. There is no cell boundaries in the system, where the CU and distributed BSs jointly serve a large number users [4], [5]. It was reported that cf-MIMO achieves a higher spectral efficiency than centralized massive MIMO and small-cell system [6]. Several studies on cell-free networks have investigated the per-user transmission rate [7], per-user packet delay [8], as well as implementation issues such as pilot contamination [5] and network backhauling [9]. Various beamforming techniques were studied for cell-free networks. [10] proposed the use of conjugate beamforming on the downlink and matched filtering on the uplink. [11] showed that partially or fully centralized processing at the CU can achieve higher spectral efficiency. [7] showed that a centralized implementation with optimal minimum mean square error (MMSE) processing not only maximizes the spectral efficiency but also largely reduces the backhaul signaling. Existing results in the literature have demonstrated the capabilities of cf-MIMO in meeting the demands of future wireless systems.

Albeit the promising potentials, the design of cf-MIMO is facing a set of new challenges, including high computation cost, high consumption of backhaul, high latency caused by BS coordination, and difficulty in synchronization [12], [13]. Considering the massive increase of the number of users per unit area in 6G network, the computation complexities at the distributed BSs and CU skyrocket [14]. Accordingly, the consumption of backhaul also explodes, while the coordination among BSs becomes more difficult. Henceforth, advanced architecture and processing for cf-MIMO are required.

A. Motivations

In essence, there are two inadequacies of existing techniques for cf-MIMO. First, from network information theory, a cf-MIMO is a *multi-point multi-hop relay network* [15], where the *network information flow* (NIF) problem arises [16] [17]. To achieve the network information capacity, each node of the multi-hop network needs to forward a function of the incoming messages, rather than just simply repeat them as in conventional routing. Yet, existing coding, multi-user detection, precoding methods are mostly developed without considering the NIF problem, which may not attain the full capability of cf-MIMO. Second, most of the conventional ways of treating the multiuser interference (MUI), such as those in decode-forward and compress-forward schemes coupled with MMSE filtering, insufficiently exploits the *interference structure* of the multi-point multi-hop cf-MIMO network.

From the literature, it is understood that the notion of *lattices and lattice codes* provide a viable approach that could address the above two issues. Base on the property that the integer sum of lattice codewords is still a valid codeword, integer-combinations of the users' messages (or network coding streams) can be efficiently computed and then forwarded, akin to the fashion of NIF [18]–[21]. Such notion was referred to as lattice network coding (LNC) [22], physical-layer network coding (PNC) [23], [24] or compute-and-forward (C&F) [18], [25]. For a multi-antenna environment, the interference structure in the spatial dimension is explicitly exploited via lattice reduction techniques [26] [27] [28], as in the LNC based integer-forcing (IF) detection and precoding [29] [30].

This work is supported by the National Key R&D Program of China (No. 2022YFB2902604 and No.2020YFB1807102), National Natural Science Foundation of China (No. 62371020) and Beijing Science Foundation (No. L232044). T. Yang is with Beihang University, Beijing, China. Email: tyang@buaa.edu.cn.

Recently, LNC and IF have been extended to address time-varying or frequency-selective fading channels [26], [31], as well as inter-symbol-interference equalization problem with the help of cyclic linear codes [32]. Zhu and Gastpar showed that any rate-tuple of the 2-user Gaussian multiple-access capacity region can be achieved using C&F [33], and we recently generalized this result to K-user fading multiple-access channel [24], [34].

For cf-MIMO specifically, [25] advocated using lattice codes and quantized C&F, where a larger capacity region was demonstrated. For the downlink, reverse C&F was studied by exploiting the uplink-downlink duality. For the case with multi-antenna users and multi-antenna BSs, signal-space alignment based lattice network coding methods are presented for both the uplink and downlink cf-MIMO [12], [13]. An advance in achievable d.o.f. was demonstrated therein. A compute-compress-forward scheme was studied in [35] [36], which exploited distributed source coding to further reduce the backhaul consumption.

This paper is set to exploit lattice codes and lattice network coding (LNC) to address the two issues raised above. Base on the algebraic property of lattice codes, integer linear combinations of the users' messages (or network coding streams) can be efficiently computed and then forwarded, which offers increased spectral efficiency and reduced backhaul consumption relative to existing non-LNC based schemes for cf-MIMO. Albeit the advances of LNC notion, some theoretical and practical aspects are still under research, including the attainable system load, FER, exact backhaul consumption and etc.. Moreover, for a massive number of users, there lacks algorithms for the LNC detection and identification of the network coding coefficient matrix with realistic costs.

B. Main Results

This paper studies an uplink cf-MIMO where the BSs are uncoordinated, i.e. each BS is ignorant of other BSs. We study a LNC based scheme. The users encode their messages with a practical lattice code, and transmit simultaneously with full bandwidth. Each BS computes a number of independent streams of integer-combinations of the users' messages, referred to as *network coding* (NC) streams. The correctly computed NC streams are forwarded to CU via the digital BH link, subject to the BH capacity constraint. The CU aggregates all NC streams from all BSs, and recovers all users' messages. This paper contributes to this subject in the following:

1) We consider a cf-MIMO system with moderate-to-large numbers of users and antennas at each BS. From a theoretical perspective, we derive the achievable rate of the LNC based cf-MIMO subject to a certain BH capacity. We also present the outage probability and ε-outage rate [37] of the scheme.

2) Base on the rate characterization, we establish and solve a new "*bounded independent vectors problem*" (BIVP) which identifies a near-optimal set of NC coefficient vectors. The solution maximizes the number of correctly computed NC streams at each BS, resulting in an advance of cf-MIMO.

3) From a practical perspective, we suggest a 2^m-ary LDPC code with 2^m-PAM as the underlying lattice code, to materialize the theoretical gains of LNC. For it, we develop new *soft detection algorithms* that calculate the a posteriori probability w.r.t. a NC stream over the lattice. We prove that our proposed soft detection algorithm has a per-user complexity no greater than $O(K)$. As such, the developed algorithms are suitable for massive access in cf-MIMO.

It is demonstrated that our developed LNC based cf-MIMO exhibits advanced frame error rate (FER) performance and increased spectral efficiency over existing non-LNC based cf-MIMO, such as decode-forward and compress-forward schemes. Such advances are achieved with just parallel processing and no more than NK single-user decoding operations. Moreover, the BH consumption is shown to have the same order as the capacity of the air-interface, offering an efficient usage of the BH.

C. Difference to Prior Works

First, the numbers of users and BS antennas under consideration are moderate-to-large, while the number of BSs is moderate. Such a setup is more relevant to existing network architecture featuring multi-antenna BSs and a small-to-moderate number of BSs. This is in contrast to [7], [10] which considered single-antenna BSs while the number of BSs is huge. Second, our work explicitly incorporates the BH capacity constraint, while [7], [10] assumed infinite BH capacity. Third, we put forth practical 2^m-ary LDPC codes for lattice coding, as well as soft detection algorithms for decoding, whose FER is shown to agree with the rate characterization. In contrast, [25], [35], [36], [38], [39] only studied the achievable rates. Our works involves MIMO processing for lattice decoding which is not studied in [25], [38], [39]. Lastly, we consider an open-loop system where the coordination of BSs are minimized, which is of particular importance for massive access.

This paper is different from [40] in the following aspects. First, the system model of this paper and that in [40] are different. [40] considered an uplink multiple access scenario. In contrast, this paper studies a cell-free MIMO network of N distributed BSs, with rate-constrained backhaul links. The rate analysis for the cell-free network with LNC (Theorem 2 on Page 8) is different from that in [40] for the multiple access. Second, the detection algorithm presented in [40] does not apply to the scenario with K being large, and is not suitable to the cell-free network . In this paper, we developed a new soft detection algorithm, whose complexity is proved to be no greater than $O(K)$. Third, the algorithms for identifying the LNC coefficient matric on Page 8 is different to that in [40]. To be specific, [40] only borrows the existing LLL algorithm, which is known to have significant loss for the case with a large K or an overloaded case where K is greater than the number of BS antennas. In contrast, we formulate a new bounded integer vector problem (BIVP) and solve it by a constraint sphere decoding (CSD) algorithm. Our proposed CSD algorithm yields significantly increased number of reliably computing integer combinations at the BSs, that contributes to the overall performance enhancement of the proposed LNC scheme for the cell-free MIMO.

II. System Model

Consider an uplink cf-MIMO system that consists of a CU, N distributed BSs, and K users. The number of users is moderate-to-large, e.g. K is tens to a hundred. The number of BSs is moderate, i.e. N less than ten. Each BS is connected to CU via a digital BH link. The BSs are not mutually connected and are uncoordinated, i.e. each BS is ignorant of other BSs. Following the convention in studying the uplink multi-user communication, we consider an *open-loop* system. There is no return links from a BS to a user and from CU to a BS, for delivering channel state information (CSI) or adaptive coding-modulation (ACM) information. Each user transmits at a target rate R_0. *Frame error rate* (FER) and *spectral efficiency* (SE) KR_0 are the performance indicators.

A. Air-interface

Consider that each user is equipped with single-antenna while each BS is equipped with n_R antennas. The extension to multi-antenna users is straightforward [37]. Let a row vector $\mathbf{x}_i^T$ denote the length-n coded-modulated signal sequence of user i, $i = 1, \cdots, K$. For a real-valued model, the baseband equivalent discrete signal at the receiver of BS j is

$$\mathbf{Y}_j = \sum_{i=1}^{K} \sqrt{\rho}\mathbf{h}_{j,i}\mathbf{x}_i^T + \mathbf{Z}_j = \sqrt{\rho}\mathbf{H}_j\mathbf{X} + \mathbf{Z}_j, j = 1, \cdots, N \tag{1}$$

where $\mathbf{Z}_j$ denotes the additive white Gaussian noise (AWGN) matrix whose entries are i.i.d with zero mean and unit variance; ρ denotes the symbol energy, which is equivalent to the per-user SNR. The column vector $\mathbf{h}_{j,i}$ represents the fading channel coefficients from user i to n_R antennas of BS j, and $\mathbf{H}_j$ is the channel coefficient matrix. A complex-valued model can be represented by a real-valued model of doubled dimension as in [18], [41]. For a better readability, this paper presents with a real-valued model.

BS j processes the received signal $\mathbf{Y}_j$:

$$\mathbf{Y}_j \overset{\text{BS processing}}{\rightarrow} \mathbf{U}_j. \tag{2}$$

The resultant signal $\mathbf{U}_j$ is then delivered to the CU via BH.

Remark 1: For the wireless channel model in (1), we assume no inter-symbol-interference. This paper focuses on a flat-fading model, where the channel coefficients remain unchanged for each coded block while differing over blocks. Such channel model applies to scenarios where the bandwidth of a user's signal is within the coherence bandwidth of the multi-path channel, e.g. in the case that each user only occupies a smaller number of subcarriers. We consider that the coherence time of channels is larger than a time block, i.e., the channels remain unchanged in a block. Our developed techniques can be extended to fast fading or frequency selective fading models by borrowing the notion of ring C&F or multi-mode integer-forcing as treated in [26], [31]. Moreover, in (1) we assumed that the users' signals are synchronized at the receiver of a BS[1]. This holds if the difference of arrival time of the users' signals are within the duration of the length of the cyclic prefix (CP).

[1]The asynchrony can be addressed by cyclicly coded LNC [32].

B. Backhaul (BH) and Central Unit (CU)

This paper considers wired digital BH links with finite capacities given by $C_1^{BH}, \cdots, C_N^{BH}$. This applies to existing network where distributed units are connected to the CU with optical fibre cables [7]. Our developed techniques also apply to the scenario with wireless BH links. From source coding theorem [15], $\mathbf{U}_j$ can be recovered by the CU if

$$\frac{1}{n} H(\mathbf{U}_j) < C_j^{BH}, j = 1, \cdots, N, \tag{3}$$

where $H(\cdot)$ denotes the entropy function and C_j^{BH} denotes the BH capacity per channel-use. The CU aggregates $\mathbf{U}_1, \cdots, \mathbf{U}_N$, and then attempts to recover all K users' messages. A *frame error* is declared if a recovered user's message sequence is not identical to the genuine one.

Problem statement: how to design transceiver architecture and BS processing algorithms, such that cf-MIMO achieves a low FER for a given SE, or a high SE for a target FER?

Remark 2: This paper considers that the capacities of the air-interface and the BH are comparable, i.e., they are of the same order of magnitude. The scenario where the air-interface capacity and BH capacity tremendously differ is of much less interests in both theory and practice. Since there is no coordination among the N BSs, distributed source-coding such as in Slepian-Wolf [15] cannot be implemented. The extension to BS coordination with distributed source coding [35], [36] is beyond the scope of this paper.

III. Lattice Network Coding based cf-MIMO

Fig. 1 shows the architecture of LNC based cf-MIMO.

A. Transmitters

Prior works on the theoretical aspects of LNC largely rely on the existence of good nested lattice codes of infinite block length. In this paper, we suggest a simple yet powerful lattice code, referred to as *ring-coded PAM* (RC-PAM). Consider an integer ring $\mathbb{Z}_q \triangleq \{0, \cdots, q-1\}$. For a prime q, the integer ring $\mathbb{Z}_q$ is a *field*. For a non-prime q, $\mathbb{Z}_q$ is not a field. To match the mainstream 2^m-PAM or 2^{2m}-QAM signaling, this paper focuses on $q = 2^m, m = 1, 2, \cdots$, where 2^m is not a prime number (except $m = 1$).

For user i, $i = 1, \cdots, K$, let $\mathbf{b}_i = [b[1], \cdots, b[k]] \in \mathbb{Z}_{2^m}^k$ denote its 2^m-ary message sequence of length k. Each entry of $\mathbf{b}_i$ is uniformly drawn from $\{0, \cdots, 2^m - 1\}$. A 2^m-ary *ring-code* with *generator matrix* $\mathbf{G}$ is adopted to encode $\mathbf{b}_i$. The encoded sequence is given by

$$\mathbf{c}_i = \text{mod}(\mathbf{G}\mathbf{b}_i, 2^m) = \mathbf{G} \otimes \mathbf{b}_i \tag{4}$$

where "$\otimes$" represents matrix multiplication modulo-2^m, and $\mathbf{c}_i \in \mathbb{Z}_{2^m}^n$. Let $\mathcal{C}^n$ denote the codebook that collects all 2^{mk} codewords w.r.t. (4). The codebook is revealed to all BSs.

Each entry of $\mathbf{c}_i$ is *one-to-one* mapped to a regular 2^m-PAM constellation symbol, given by

$$\mathbf{x}_i = \frac{1}{\gamma}\left(\mathbf{c}_i - \frac{2^m - 1}{2}\right) \in \frac{1}{\gamma}\left\{\frac{1 - 2^m}{2}, \cdots, \frac{2^m - 1}{2}\right\}^n. \tag{5}$$

Here γ normalizes the symbol energy. The rate is $R = \frac{km}{n}$ bits/symbol. The K users encode their messages with the

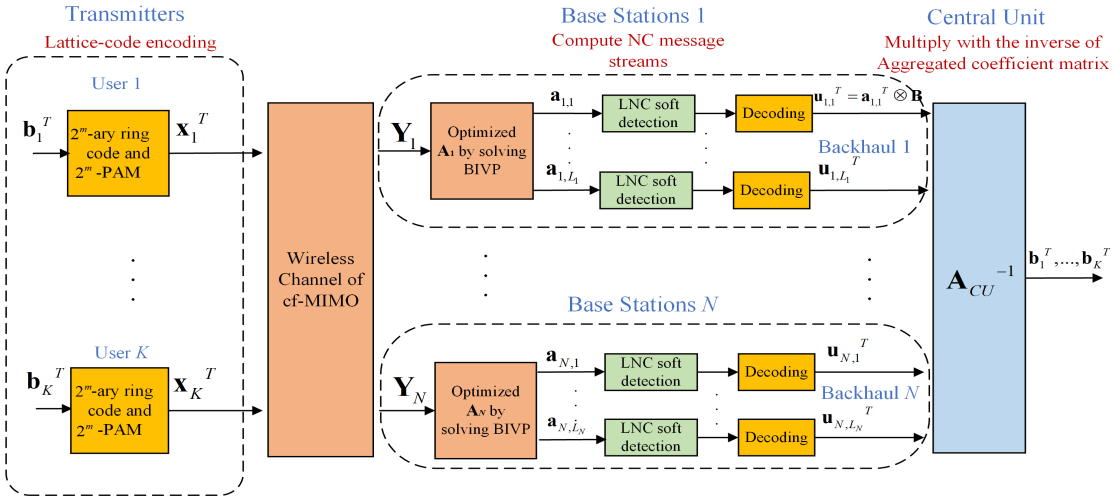

Fig. 1. Block diagram of a LNC based cf-MIMO system with K users and N BSs. All users utilize the same 2^m-ary code and PAM. For each BS, the optimized coefficient matrix is identified by solving the BIVP w.r.t. the channel state information.

same[2] RC-PAM, and transmit simultaneously. For a complex-valued model, two 2^m-level RC-PAM, one for the inphase and the other for the quadrature part, form 2^{2m}-level RC-QAM. Einstein integers are beyond the scope of this paper [22].

Remark 3: [*Difference to existing code-modulation*] The presented RC-PAM is a "*Construction-A*" type of lattice code [18] with a *one-dimension* shaping using modulo-2^m operation. It is much simpler than existing low-density lattice codes [42] that does not apply to the mainstream 2^m-PAM modulation schemes.

Remark 4: [*"Good" generator matrices* $\mathbf{G}$] For $m = 1$, RC-PAM reduces to binary channel coding with BPSK signaling. Any state-of-the-art binary channel codes, such as LDPC codes and polar codes in 5G NR standards, can be utilized to execute (4) and the associated decoding. For $m = 2, 3, \cdots$, LDPC and irregular repeat-accumulate (IRA) ring-codes developed in [43] are ready to use in cf-MIMO.

Remark 5: [*Necessity of ring codes*] Here $\mathbb{Z}_{2^m}$ is not identical to GF(2^m). GF(2^m) is an extension field of GF(2) of elements $\left\{0, 1, \beta, \cdots \beta^{2^m-2}\right\}$ [44]. The additive rule is based on the primitive element of the polynomials, which is different from the additive rule of integers $\mathbb{Z}_{2^m}$. To form a lattice code for 2^m-PAM, a ring code over $\mathbb{Z}_{2^m}$ is required.

In practice, the implementation of lattice codes can be done by using q-ary linear codes coupled with q-PAM modulation. Such coded-modulation belongs to the ensemble of lattice codes, and the required structural properties hold. For q being a prime number, a GF(q) code with q-PAM as in [45] can be used. Yet, it is well-known that the mainstream modulation schemes are 2^m-PAM or 2^m-QAM, where $q = 2^m$ is not a prime number. As such, a linear code over the integer ring $\mathbb{Z}_{2^m}$ is necessary in LNC for the mainstream 2^m -PAM or 2^m-QAM modulation.

[2]The extension to asymmetric rate is straightforward. A low rate user' message are zero-padded to form a length k message sequence. Then, the same channel code encoder can be utilize to encode all users' messages.

Albeit our developed techniques can be extended to asymmetric rate setup, to make this article highly focused on presenting the notion and mechanisms, this work studies the symmetric rate case, where each user transmits at the same target rate R_0 [37]. The symmetric rate is widely acknowledged as a key indicator of the performance of the uplink multi-user communication scheme for the open-loop model. We note that our developed techniques can be extended to the asymmetric rate case by introducing the nesting of linear codes as in [24]. A mother code of a high rate is used for a high rate user. Then, the code for a low rate user is given by strategically zero-padding the messages [24]. In such a way, the treatment of lattice network coding applies to the asymmetric rate setup.

B. LNC Processing at BSs

Assume local CSI is acquired at each BS. Recall the received signals at the BSs given in (1). BS j attempts to directly compute L_j, $L_j \leq K$, streams of *integer-combinations* of the K users' messages over $\mathbb{Z}_{2^m}$, written as

$$\mathbf{u}_{j,l}^T \triangleq \mathbf{a}_{j,l}^T \otimes \mathbf{B}, l = 1, \cdots, L_j, \tag{6}$$

where $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_K]^T$. We refer to $\mathbf{u}_{j,l}$ as the lth *network coding* (NC) *stream*, and $\mathbf{a}_{j,l} \in \mathbb{Z}^{K\times 1}$ as the *NC coefficient vector*. Note that $\mathbf{u}_{j,l}, l = 1, \cdots, L_j$ are *digital streams*.

Let $\mathbf{U}_j = \left[\mathbf{u}_{j,1}, \cdots, \mathbf{u}_{j,L_j}\right]^T$ denote the L_j NC streams obtained at BS j. Let $\mathbf{A}_j = \left[\mathbf{a}_{j,1}, \cdots, \mathbf{a}_{j,L_j}\right]^T$ denote the L_j NC coefficient vectors, referred to as the *NC coefficient matrix*. Then

$$\mathbf{U}_j = \mathbf{A}_j \otimes \mathbf{B}. \tag{7}$$

Remark 6: [*Mechanism of LNC*] In cf-MIMO, even if BS j is not able to decode all K users' messages $\mathbf{B}$, it would still be possible to compute $\mathbf{u}_{j,1}, \cdots, \mathbf{u}_{j,L_j}$ thanks to the structural property of the underlying lattice code [18]. Intuitively, LNC is set to remove a "partial entropy" $H(\mathbf{U}_j) < H(\mathbf{B})$ for $L_j < K$. This task is "easier" than removing the full entropy

$H(\mathbf{B})$, and is more likely to be accomplished. Section IV will develop practical algorithms for LNC in detail.

Remark 7: [*Choice of* $\mathbf{A}_j$] The choice of coefficient matrix $\mathbf{A}_j$ largely affects the performance. A well-chosen $\mathbf{A}_j$ *maximizes the number of correctly computed NC streams* L_j. This leads to the maximized number of aggregated NC streams $\sum_{j=1}^{N} L_j$ at the CU. Section V will develop efficient algorithms for acquiring the optimized $\mathbf{A}_j$ in detail.

The correctly computed $\mathbf{U}_j$ and the associated $\mathbf{A}_j$ are forwarded to the CU via its BH link, which are *digital streams*. The BH consumption per channel-use is

$$\frac{H(\mathbf{U}_j)}{n} = R_0 L_j \text{ bits}. \tag{8}$$

As $\mathbf{A}_j$ is chosen once per block, the BH consumption of delivering $\mathbf{A}_j$ is minor for large n.

C. CU's Processing

The CU collects all $L_{CU} \triangleq \sum_{j=1}^{N} L_j$ NC streams from the N BSs, forming the *aggregated NC streams* $\mathbf{U} = \left[\mathbf{U}_1^T, \cdots, \mathbf{U}_N^T\right]^T$. Also, the CU forms the *aggregated NC coefficient matrix* $\mathbf{A}_{CU} \triangleq \left[\mathbf{A}_1^T, \cdots, \mathbf{A}_N^T\right]^T$, which is of size L_{CU}-by-K. If $\mathbf{A}_{CU}$ is of full rank K in $\mathbb{Z}_{2^m}$, it has a unique inverse $\mathbf{A}_{CU}^{-1}$, i.e., $\mathbf{A}_{CU}^{-1} \otimes \mathbf{A}_{CU} = \mathbf{I}_K$. Then, CU can correctly recover all K users' messages by implementing

$$\mathbf{B} = \mathbf{A}_{CU}^{-1} \otimes \mathbf{U}. \tag{9}$$

If $L_{CU} < K$, or $L_{CU} \geq K$ but $\mathbf{A}_{CU}$ is not full rank, some users' messages cannot be recovered.

Example 1: Consider a system of $K = 4$ users and $N = 2$ BSs. Each BS has $n_R = 2$ antennas. Consider a 4-ary ring code and 4-PAM are utilized. The channel realization of BSs 1 and 2 are

$$\mathbf{H}_1 = \begin{bmatrix} 1.02 & 1.96 & 0.03 & 0.12 \\ 0.17 & 1.05 & 0.97 & 1.01 \end{bmatrix}, \tag{10}$$

$$\mathbf{H}_2 = \begin{bmatrix} 0.95 & 1.03 & 0.03 & 0.15 \\ 0.08 & 1.08 & 2.02 & 1.04 \end{bmatrix}. \tag{11}$$

BS 1 selects

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \tag{12}$$

and computes 2 NC streams

$$\mathbf{u}_{1,1}^T \triangleq \mathbf{a}_{\mathbf{1},\mathbf{1}}^{\mathbf{T}} \otimes \mathbf{B} = \mathbf{b}_1^T \oplus 2\mathbf{b}_2^T, \quad \mathbf{u}_{1,2}^T \triangleq \mathbf{a}_{1,2}^T \otimes \mathbf{B} = \mathbf{b}_2^T \oplus \mathbf{b}_3^T \oplus \mathbf{b}_4^T.$$

The decisions on $\mathbf{u}_{1,1}^T$,$\mathbf{u}_{1,2}^T$, together with $\mathbf{A}_1$, are forwarded to CU. Meanwhile, BS 2 selects

$$\mathbf{A}_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 \end{bmatrix},$$

and computes 2 NC streams

$$\mathbf{u}_{2,1}^T \triangleq \mathbf{b}_1^T \oplus \mathbf{b}_2^T, \quad \mathbf{u}_{2,2}^T \triangleq \mathbf{b}_2^T \oplus 2\mathbf{b}_3^T \oplus \mathbf{b}_4^T.$$

The CU aggregates 4 NC streams $\mathbf{u}_{1,1}^T, \mathbf{u}_{1,2}^T, \mathbf{u}_{2,1}^T$ and $\mathbf{u}_{2,2}^T$, as well as $\mathbf{A}_{CU} \triangleq \left[\mathbf{A}_1^T, \mathbf{A}_2^T\right]^T$. In this example, $\mathbf{A}_{CU}$ is of full rank 4, and has a unique inverse given by

$$\mathbf{A}_{CU}^{-1} = \begin{bmatrix} 3 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 0 & 3 & 0 & 1 \\ 3 & 2 & 1 & 3 \end{bmatrix}. \tag{13}$$

The CU recover's all 4 users' messages by implementing (9).

IV. Practical Soft Detection and Decoding for LNC

In this section, we develop new soft detection algorithms for LNC. This section considers that $\mathbf{A}_j$ is given. The optimized $\mathbf{A}_j$ will be presented in Section V.

A. Parallel Computation Rule of LNC

BS j aims to compute NC streams $\mathbf{U}_j = \left[\mathbf{u}_{j,1}, \cdots, \mathbf{u}_{j,L_j}\right]^T$. The optimal rule requires the jointly computing of $p\left(\mathbf{u}_{j,1}, \cdots, \mathbf{u}_{j,L_j} | \mathbf{Y}_j\right)$. This is well-known to be highly conceptual and is intangible to implement. This paper relies on a *parallel rule* [29]

$$p\left(\mathbf{u}_{j,l} | \mathbf{Y}_j\right), l = 1, \cdots, L_j. \tag{14}$$

Such parallel rule has low-cost implementation, low-latency, tractable analysis and competitive performance. The extension to successive computation can be done as in [24].

We next present how to implement (14). Let $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_K]^T$ stack up all users' coded sequences generated by the 2^m-ary ring-code in (4). Define

$$\mathbf{v}_{j,l}^T \triangleq \bmod\left(\sum_{i=1}^{K} a_{l,i}^j \mathbf{c}_i^T, 2^m\right) = \mathbf{a}_{j,l}^T \otimes \mathbf{C}, \tag{15}$$

referred to as a *codeword-level* NC stream. The following properties of RC-PAM will be utilized.

Property 1: For any K codewords $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K \in \mathcal{C}^n$, RC-PAM satisfies

$$\bmod\left(\sum_{i=1}^{K} a_i \mathbf{c}_i, 2^m\right) \in \mathcal{C}^n \tag{16}$$

for any integers $a_1, \cdots, a_K$, i.e. the integer-sum of K codewords modulo-2^m is a valid codeword.

Property 2: With the generator matrix $\mathbf{G}$ in (4), we have

$$\mathbf{v}_{j,l} = \mathbf{G} \otimes \mathbf{u}_{j,l}. \tag{17}$$

That is, the codeword-level NC stream and the message-level NC stream are also related by $\mathbf{G}$.

Proof. The codeword-level NC stream can be written as

$$\mathbf{v}_{jl}^T \triangleq \bmod(\sum_{i=1}^{K} a_{l,i}^j \mathbf{c}_i^T, 2^m) = \bmod\left(\sum_{i=1}^{K} a_{l,i}^j \mathbf{G} \otimes \mathbf{b}_i, 2^m\right) \tag{18}$$

$$\overset{(a)}{=} \bmod\left(\mathbf{G} \sum_{i=1}^{K} a_{l,i}^j \otimes \mathbf{b}_i, 2^m\right) = \mathbf{G} \otimes \mathbf{u}_{j,l}$$

where "$\overset{(a)}{=}$" utilizes the associative law for the summation and the modulo-2^m operation. ■

In conventional channel-code decoding, the a posteriori probabilities (APPs) of the coded-sequence $\mathbf{c}$, denoted by $p\left(c\left[t\right]|y\left[t\right]\right), t=1,\cdots,n$, is first calculated. The decoder takes in the APPs the exploits the structure of $\mathbf{G}$, and outputs the probabilities on the message sequence $\mathbf{b}$. Prop. 2 suggests that in LNC, the symbol-wise APPs of the codeword-level NC stream $\mathbf{v}_{j,l}$ is first calculated. The ring code decoder takes in such APPs and outputs the probabilities on the message-level NC stream $\mathbf{u}_{j,l}$. As such, the practical implementation of (14) is carried out via:

a) A *soft LNC detector* calculates the APPs w.r.t. the codeword-level NC stream $\mathbf{v}_{j,l}$, written as $p\left(v_{j,l}\left[t\right]|\mathbf{y}_j\left[t\right]\right), t=1,\cdots,n$. (Here $v_{j,l}\left[t\right]$ and $\mathbf{y}_j\left[t\right]$ denote the t-th column of $\mathbf{v}_{j,l}$ and $\mathbf{Y}_j$.)

b) A 2^m-ary *ring-code decoder* takes the APPs as input and outputs the probabilities of the message-level NC stream $\mathbf{u}_{j,l}$, written as $p(u_{j,l}[t]), t=1,\cdots,k$. The hard-decision is then made.

B. Soft LNC Detection

Since symbol-by-symbol detection is utilized, we omit the index "t" for simplicity. Moreover, since the processing of the N BSs follows the same procedures, we omit the BS index j. Recall the bijective mapping between x_i and c_i in (5), the received signal (1) can be written as

$$\mathbf{y}=\frac{\sqrt{\rho}}{\gamma}\sum_{i=1}^{K}\mathbf{h}_i c_i+\mathbf{z}-\varkappa=\frac{\sqrt{\rho}}{\gamma}\mathbf{H}\mathbf{c}+\mathbf{z}+\varkappa. \tag{19}$$

where $\varkappa=-\frac{\sqrt{\rho}}{\gamma}\sum_{i=1}^{K}\mathbf{h}_i\frac{2^m-1}{2}$ can be straightforwardly compensated. Recall that each user's coded symbol c_i belongs to $\mathbb{Z}_{2^m}=\{0,\cdots,2^m-1\}$. For BS's received signal, the superposition of the K users' symbols, denoted by $\mathbf{s}=\sum_{i=1}^{K}\mathbf{h}_i c_i=\mathbf{H}\mathbf{c}$, forms a "super-constellation" of 2^{mK} *candidates* in a n_R dimension space. Let all the candidate vectors $\mathbf{s}$ of the super-constellation be collected by the set

$$\mathcal{S}=\left\{\mathbf{s}=\mathbf{H}\mathbf{c},\forall\mathbf{c}\in\mathbb{Z}_{2^m}^{K}\right\}. \tag{20}$$

For a NC stream, the 2^{mK} candidates are partitioned in to 2^m "*bins*". Those candidates with identical (non-identical) NC value $v_l\triangleq\mathbf{a}_l^T\otimes\mathbf{c}=\theta$, belong to the same (different) bin. This is referred to as "*algebraic binning*" [18], $\mathbf{a}_l^T$ determines the rule of the partition.

The soft LNC detector computes the *APPs of the bin-indices* θ. Using the Baye's rule, for $\theta=0,\cdots,2^m-1$,

$$p\left(v_l=\theta|\mathbf{y}\right)=\frac{1}{\eta}\sum_{\mathbf{c}:\mathbf{a}_l^T\otimes\mathbf{c}=\theta}\exp\left(-\frac{\left\|\mathbf{y}-\frac{\sqrt{\rho}}{\gamma}\mathbf{H}\mathbf{c}-\varkappa\right\|^2}{2}\right). \tag{21}$$

The scalar η is to ensure the APPs w.r.t. $\theta=0,\cdots,2^m-1$ add up to 1. The APP calculation is performed in a n_R-dimension space, which is different from that in the IF methodology.

The complexity of directly executing (21) has order $O\left(2^{mK}\right)$. For a large K, alternative soft detection algorithms are required, which is studied below.

C. Efficient Linear Soft LNC Detection

A linear filtering is first utilized to transform the n_R-dimension received signal into L ($L\leq K$) streams of single-dimension signals. Then, each signal stream is used to compute one NC stream.

1) Derivation of APP: Denote by $\mathbf{W}$ a size L-by-n_R linear filtering matrix of real-valued entries. Let $\mathbf{w}_l^T$ denote the lth row of $\mathbf{W}$, $\|\mathbf{w}_l\|=1$. The lth filtered signal stream is

$$\overline{y}_l=\mathbf{w}_l^T\mathbf{y}=\frac{\sqrt{\rho}}{\gamma}\sum_{i=1}^{K}\psi_{l,i}c_i+\overline{z}_l+\varkappa_l \tag{22}$$

where $\psi_{l,i}=\mathbf{w}_l^T\mathbf{h}_i$ denotes the "effective gain" w.r.t. user i after the filtering, the noise $\overline{z}_l$ has a unit variance, and the term $\varkappa_l=\mathbf{w}_l^T\varkappa$ can be straightforwardly compensated.

Let $\mathcal{I}_l\triangleq\{i:a_{l,i}\neq 0\}$ collects the positions of non-zero entries of $\mathbf{a}_l$, and $\mathcal{I}_l^c$ be the complementary set. Let $\omega\left(\mathbf{a}_l\right)\triangleq|\mathcal{I}_l|$. Then, $\overline{y}_l$ is re-arranged into

$$\begin{aligned}\overline{y}_l&=\frac{\sqrt{\rho}}{\gamma}\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i+\frac{\sqrt{\rho}}{\gamma}\sum_{i\in\mathcal{I}_l^c}\psi_{l,i}c_i+\overline{z}_l+\varkappa_l\\&=\frac{\sqrt{\rho}}{\gamma}\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i+\xi_l+\varkappa_l'.\end{aligned} \tag{23}$$

The term $\sum_{i\in\mathcal{I}_l}\sqrt{\rho}\psi_{l,i}c_i$ is the superposition of the signals of the $\omega\left(\mathbf{a}_l\right)$ users whose NC coefficients are non-zero, which is the *useful signal* part. The term $\sum_{i\in\mathcal{I}_l^c}\psi_{l,i}c_i$ contains the signals of the remaining $K-\omega\left(\mathbf{a}_l\right)$ users whose NC coefficients are zero. This term is statistically independent from $\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i$, and is regarded as irrelevant w.r.t. the computation of the NC stream. The term $\xi_l=\sum_{i\in\mathcal{I}_l^c}\sqrt{\rho}\psi_{l,i}x_i+\overline{z}_l$ is the *effective noise*, which is uncorrelated with the useful signal part.

For a sufficiently large K, $|\mathcal{I}_l^c|$ is sufficiently large to apply Central Limit Theorem. As such, ξ_l follows a Gaussian distribution with zero mean and variance $\overline{\sigma}_l^2=\rho\sum_{i\in\mathcal{I}_l^c}\psi_{l,i}^2+1$.

Let $\overline{\mathbf{c}}_l$ consist of only the entries $\{c_i,i\in\mathcal{I}_l\}$. The APP w.r.t. the lth NC stream is given by

$$\begin{aligned}p\left(v_l=\theta|\overline{y}_l\right)&=\frac{1}{\eta}\sum_{\overline{\mathbf{c}}_l:\mathbf{a}_l^T\otimes\overline{\mathbf{c}}_l=\theta}p\left(\overline{y}_l|\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i\right)\\&=\frac{1}{\eta}\sum_{\overline{\mathbf{c}}_l:\mathbf{a}_l^T\otimes\overline{\mathbf{c}}_l=\theta}\exp\left(-\left|\overline{y}_l-\frac{\sqrt{\rho}}{\gamma}\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i-\varkappa_l'\right|^2/2\overline{\sigma}_l^2\right),\end{aligned} \tag{24}$$

where η is the normalization factor. The APP $p\left(v_l=\theta|\overline{y}_l\right)$ is equal to the sum of the likelihood functions of the $2^{m(\omega(\mathbf{a}_l)-1)}$ candidates whose underlying NC is equal to θ. Here we abused the notation of $\mathbf{a}_l^T\otimes\overline{\mathbf{c}}_l$ by considering the multiplication with the non-zeros entries of $\mathbf{a}_l^T$ only.

2) Gaussian Approximation: A direct execution of (24) needs to evaluate the Euclidean distances of $2^{m\omega(\mathbf{a}_l)}$ candidates. The following theorem shows that for a sufficiently large K, the APP (24) can be efficiently computed using a Gaussian approximation. Define $\omega_H(\mathbf{a}_l) \triangleq \sum_{i\in\mathcal{I}_l} |a_{l,i}|$ as the "*weight*" of $\mathbf{a}_l$.

Theorem 1: As $K \to \infty$, the APP in (24) can be computed with a complexity of order

$$O\left(\omega_H(\mathbf{a}_l)\left(2^m-1\right)+1\right) \tag{25}$$

Proof. For convenience, we consider that the term $\varkappa_l'$ is compensated. Note that there is a many-to-one mapping between $\mathbf{a}_l^T\mathbf{c}$ and $\mathbf{a}_l^T \otimes \mathbf{c}$. Specifically, all events $\left\{\mathbf{a}_l^T\mathbf{c} = \theta \pm \beta\cdot 2^m\right\}$ with various values $\overline{\theta} = \theta \pm \beta\cdot 2^m$ have an identical value of $\mathbf{a}_l^T\otimes\mathbf{c} = \theta$ after the modulo-2^m operation. As such, using the Total Probability Rule, the APP is written as

$$p\left(v_l=\theta|\overline{y}_l\right) = \frac{1}{\eta}\sum_{\overline{\theta}:\mathrm{mod}\left(\overline{\theta},2^m\right)=\theta} p\left(\overline{y}_l|\mathbf{a}_l^T\mathbf{c}=\overline{\theta}\right)p\left(\overline{\theta}\right). \tag{26}$$

Let $\Omega_l\left(\overline{\theta}\right) = \left\{\mathbf{c}:\mathbf{a}_l^T\mathbf{c}=\overline{\theta}\right\}$ collect the candidates $\mathbf{c}$ with $\mathbf{a}_l^T\mathbf{c}$ equal to $\overline{\theta}$. The conditional mean is

$$\mu_l\left(\overline{\theta}\right) = E_{\mathbf{c}}\left(\overline{y}_l|\mathbf{a}_l^T\mathbf{c}=\overline{\theta}\right) = E_{\mathbf{c}}\left(\frac{\sqrt{\rho}}{\gamma}\sum_{i\in\mathcal{I}_l}\psi_{l,i}c_i + \xi_l|\mathbf{a}_l^T\mathbf{c}=\overline{\theta}\right)$$
$$= \frac{1}{\left|\Omega_l\left(\overline{\theta}\right)\right|}\sum_{\mathbf{c}\in\Omega_l\left(\overline{\theta}\right)}\sum_{i\in\mathcal{I}_l}\frac{\sqrt{\rho}}{\gamma}\psi_{l,i}c_i. \tag{27}$$

The conditional variance is

$$\sigma_l^2\left(\overline{\theta}\right) = E_{\mathbf{c}}\left(\left|\sum_{i\in\mathcal{I}_l}\frac{\sqrt{\rho}}{\gamma}\psi_{l,i}c_i + \xi_l - \mu_l\left(\overline{\theta}\right)\right|^2\right)$$
$$= \frac{1}{\left|\Omega_l\left(\overline{\theta}\right)\right|}\sum_{\mathbf{c}\in\Omega_l\left(\overline{\theta}\right)}\left(\sum_{i\in\mathcal{I}_l}\frac{\sqrt{\rho}}{\gamma}\psi_{l,i}c_i\right)^2 - \mu_l^2\left(\overline{\theta}\right) + \overline{\sigma}_l^2. \tag{28}$$

As $K\to\infty$, $\omega_H(\mathbf{a}_l)$ also tends to be large. Then, $\overline{y}_l$ is of a conditional Gaussian distribution for all values of $\overline{\theta}$ in probability. The APP is then calculated as

$$p\left(v_l=\theta|\overline{y}_l\right) = \frac{1}{\eta}\sum_{\overline{\theta}:\mathrm{mod}\left(\overline{\theta},2^m\right)=\theta}\exp\left(-\frac{\left(\overline{y}_l-\mu_l\left(\overline{\theta}\right)\right)^2}{2\sigma_l^2\left(\overline{\theta}\right)}\right)p\left(\overline{\theta}\right). \tag{29}$$

Here, $\overline{\theta} \in \{\sum_{i:a_{l,i}<0} a_{l,i}\left(2^m-1\right), \cdots, \sum_{i:a_{l,i}>0} a_{l,i}\left(2^m-1\right)\}$. The cardinality of the set for $\overline{\theta}$ is $\omega_H(\mathbf{a}_l)\left(2^m-1\right)+1$. In other words, there are $\omega_H(\mathbf{a}_l)\left(2^m-1\right)+1$ Euclidean distances needs to be calculated in (29). This completes the proof. ■

Remark 8: Note that if K is not approaching infinity, the order of complexity of the proposed soft detection algorithm is still given by (25), which is no greater than $O(2^m K)$. However, when K is relatively small, i.e. the central limit theorem is not very effective, the APP computed via our proposed soft detection based on Gaussian approximation is not identical to that computed via the direct execution of (24). Empirically, the loss is about 0.2 dB in FER performance for K being less than 4. As K becomes sufficiently large, the loss becomes unnoticeable.

Remark 9: For a wide range of cf-MIMO configurations, $\omega_H(\mathbf{a}_l)$ is just a (small) fraction of K. In particular if the lattice basis vectors are already short and near-orthogonal, $\mathbf{a}_l$ tends to be sparse for K being large.

3) Details on the statistics: The computation of APP presented above requires a) the a priori probability $p\left(\overline{\theta}\right)$, b) the conditional mean $\mu_l\left(\overline{\theta}\right)$ and c) conditional variance $\sigma_l^2\left(\overline{\theta}\right)$ (29), to be detailed below. Since these statistics are required to be calculated once per-block, the cost is minor compared to that in (29) which are computed n times per-block. For notational simplify, the index l is omitted in this part.

a) Let $n_1\left[\overline{\theta}\right] = 1$ for $\overline{\theta} = 0, a_1, \cdots, (2^m-1)a_1$ if $a_1 > 0$, and $\overline{\theta} = (2^m-1)a_1, \cdots, 0$ if $a_1 < 0$. Let $n_1\left[\overline{\theta}\right] = 0$ for the rest values of $\overline{\theta}$. Then $p\left(\overline{\theta}\right)$ is obtained by sequentially implementing

$$n_k\left[\overline{\theta}\right] = \sum_{\tau=0,\cdots,2^m-1} n_{k-1}\left[\overline{\theta}-a_i\tau\right] \tag{30}$$

until layer $K' = \omega(\mathbf{a})$ is reached. This requires no more than

$$\sum_{k=1}^{K'}\omega_H\left(\left[a_1,\cdots,a_k\right]\right)\left(2^m-1\right)^2 \tag{31}$$

additions in total and does not involve multiplication.

b) The conditional means can be obtained by sequentially implementing

$$\widetilde{\mu}_k\left[\overline{\theta}\right] = \sum_{\tau=0,\cdots,2^m-1}\widetilde{\mu}_{k-1}\left[\overline{\theta}-a_i\tau\right] + \tau\sqrt{\rho}\psi_k. \tag{32}$$

When reaching layer $K' = \omega(\mathbf{a})$, the conditional mean is computed by $\mu\left(\overline{\theta}\right) = \widetilde{\mu}_{K'}\left[\overline{\theta}\right]/n_{K'}\left[\overline{\theta}\right]$.

c) The term $\sum_{\mathbf{c}\in\Omega\left(\overline{\theta}\right)}\left(\sum_{i\in\mathcal{I}}\sqrt{\rho}\psi_i c_i\right)^2$ is calculated by sequentially implementing

$$\vartheta_k\left[\overline{\theta}\right] =$$
$$\sum_{\tau}\left(\vartheta_{k-1}\left[\overline{\theta}-a_k\tau\right] + 2\tau\sqrt{\rho}\psi_i u_{k-1}\left[\overline{\theta}-a_k\tau\right] + \left(\tau\sqrt{\rho}\psi_i\right)^2\right).$$

When reaching layer K', the conditional variance is

$$\sigma^2\left(\overline{\theta}\right) = s_{K'}\left[\overline{\theta}\right]/n_{K'}\left[\overline{\theta}\right] - \mu^2\left(\overline{\theta}\right) + \gamma^2\overline{\sigma}^2. \tag{33}$$

D. Example with Integer-forcing (IF)

Our developed soft LNC detection applies to any $\mathbf{W}$, not only for IF. For illustration purpose, we briefly illustrate an example with regularized IF (RIF), whose filter is

$$\mathbf{W}_{RIF} = \mathbf{A}\mathbf{H}^T\left(\rho\mathbf{H}\mathbf{H}^T + \mathbf{I}_N\right)^{-1}. \tag{34}$$

The filtered signal is

$$\overline{y}_l = \mathbf{w}_l^T\mathbf{y} = \sum_{i\in\mathcal{I}_l}\sqrt{\rho}\psi_{l,i}c_i + \xi_l = \sum_{i\in\mathcal{I}_l}\sqrt{\rho}a_{l,i}c_i + e_l, \tag{35}$$

where the estimation error term is

$$e_l = \sum_{i\in\mathcal{I}_l}\sqrt{\rho}\left(\psi_{l,i}-a_{l,i}\right)c_i + \xi_l. \tag{36}$$

The error term e_l is correlated with the useful signal part $\sum_{i\in\mathcal{I}_l}\sqrt{\rho}a_{l,i}c_i$. This leads to $\mu_l\left(\overline{\theta}\right)\neq\overline{\theta}$ and $\sigma_l^2\left(\overline{\theta}\right)\neq\gamma^2\widetilde{\sigma}_l^2$, which need to be calculated as in (27) and (28), respectively.

For a sufficiently large K, the number of terms that adds up in (36) is sufficiently large to apply Central Limit Theorem for e_l. Hence, one may approximate e_l as a Gaussian random variable with variance $E\left(e_l^2\right)$. It can be easily shown that the MSE of e_l has a close-form representation

$$E\left(e_l^2\right)=\gamma^2\mathbf{a}_l^T\left(\rho\mathbf{H}^T\mathbf{H}+\mathbf{I}\right)^{-1}\mathbf{a}_l^T. \tag{37}$$

Further, by disregarding the bias in the estimation error term, the mean of e_l is approximated as zero. As such, the calculation of the APP in (29) may be further simplified into

$$p\left(v_l=\theta|\overline{y}_l\right)\approx\frac{1}{\eta}\sum_{\overline{\theta}:\mathbf{a}_l^T\otimes\mathbf{c}=\theta}\exp\left(-\frac{\left(\overline{y}_l-\overline{\theta}\right)^2}{2E\left(e_l^2\right)}\right)p\left(\overline{\theta}\right). \tag{38}$$

For a small K, the loss by using (30) could be considerable. For a large K, the loss shrinks.

E. Decoding

The symbol-wise APPs of each NC stream obtained from the aforementioned soft LNC detector is forwarded to a 2^m-ary ring code decoder. For LDPC and IRA ring codes, 2^m-ary message passing decoding is conducted. When the parity-check rule of the decoder's hard-decision output is satisfied, the message passing decoding is terminated, yielding the decision $\widehat{\mathbf{u}}_l$. The details on the ring-code decoder can be found in [43]. Algorithm 1 summarizes the procedures of BSs and CU of the LNC based cf-MIMO system. Note that the computations of the L_j NC streams $\mathbf{u}_{j,1},\cdots,\mathbf{u}_{j,K}$ at BS j are executed in parallel. The operations of the N BS are also carried out in parallel.

Algorithm 1 Summary of the procedures of LNC based cf-MIMO system

Step 1) At each BS j, calculate the filter matrix, e.g. $\mathbf{W}_{RIF}$ in (34). Perform the filtering (22).
Step 2) Calculate the a priori probability as in (30), the conditional mean as in (32), the conditional variance as in (33), for $l=1,\cdots,L_j$.
Step 3) Perform (29) in parallel to calculate the APPs $p\left(v_l=\theta|\overline{y}_l\right)$ for the L_j streams of codeword-level NC streams. Forward the L_j streams of APPs to the L_j ring-code decoders.
Step 4) Perform ring-code decoding for the L_j streams parallely, which yields the decisions $\widehat{\mathbf{u}}_1,..,\widehat{\mathbf{u}}_{L_j}$. Forward the L_j NC streams to the CU via the BH link.
Step 5) The CU collects all L_{CU} NC streams from the N BSs, forming the aggregated NC streams $\widehat{\mathbf{U}}=\left[\widehat{\mathbf{U}}_1^T,\cdots,\widehat{\mathbf{U}}_N^T\right]^T$ and $\mathbf{A}\triangleq\left[\mathbf{A}_1^T,\cdots,\mathbf{A}_N^T\right]^T$. Then, the CU implements (9).

The detection algorithm resented in [40] utilized an exhaustive search method to calculate the a posteriori probability (APP) w.r.t. to an integer combination. The complexity therein is exponential to the weight of the integer coefficient vector, which prevents its usage for the massive access scenario with K being large. In this paper, we develop a new soft detection algorithm. This algorithm calculates the a priori distribution of the integer sum of the extended lattice constellation, the conditional mean and conditional variance. Then, a Gaussian approximation technique is used to calculate the APP. We show that with this technique, the complexity is proved to be no greater than $O(2^mK)$, see Theorem 1. This new contribution is of pivotal importance to the massive access cell-free MIMO.

Here we note that in the proposed scheme, each BS just computes LNC streams locally, and forwards them to the central unit via its own BH link. Its processing operations are independent from those of other BSs. The synchronization procedure is exactly the same as in the conventional single-BS setup. The preamble sequence (e.g. using m-sequence, gold-sequence or zadoff-chu sequence) of each user is transmitted. The BS correlates with the preamble sequence locally and attains precise synchronization. This is performed independently at the N BSs. Thus, the proposed LNC impose no new challenges to the synchronization. In practice, the maximum relative delay among the K users at the BS is usually within the length of the cyclic prefix (CP) of the OFDM symbol. As such, the asynchrony among the users only introduces some phase shift. Our proposed LNC techniques, e.g. the soft detection and network coding matrix selection, will adjust accordingly w.r.t. the phase shifts.

V. On Optimized Design of LNC for cf-MIMO

In this section, we study the optimized NC coefficient matrices $\mathbf{A}_1,\cdots,\mathbf{A}_N$, based on the accurate local CSI at each BS. For a tractable analysis, we consider LNC with linear detection and rely on the existence of "good" lattice codes. This enables us to provide an efficient yet powerful suboptimal solution to $\mathbf{A}_1,\cdots,\mathbf{A}_N$.

A. Analysis of Achievable Symmetric Rate

Lemma 1: For BS j, $j=1,\cdots,N$, as $n\to\infty$, the probability of computation error of the lth NC stream $\Pr\left(\widehat{\mathbf{u}}_{j,l}\neq\mathbf{u}_{j,l}\right)<\varepsilon$ for any arbitrarily small ε if the rates of the K users satisfies

$$R_i<\frac{1}{2}\log_2^+\left(\frac{1}{\mathbf{a}_{j,l}^T\left(\rho\mathbf{H}_j^T\mathbf{H}_j+\mathbf{I}_K\right)^{-1}\mathbf{a}_{j,l}}\right),\forall i\in\mathcal{I}_l^j. \tag{39}$$

Proof. The mean square error (MSE) in the linear estimator of $\mathbf{a}_l^T\mathbf{x}$ with $\mathbf{w}_{j,l}$ is given by

$$E\left(\left|\mathbf{w}_{j,l}^T\mathbf{y}_j-\mathbf{a}_{j,l}^T\mathbf{x}\right|^2\right). \tag{40}$$

The derivative of the MSE w.r.t. $\mathbf{w}_{j,l}$ is

$$\frac{\partial E\left(\left|\mathbf{w}_{j,l}^T\mathbf{y}_j-\mathbf{a}_{j,l}^T\mathbf{x}\right|^2\right)}{\partial\mathbf{w}_{j,l}}. \tag{41}$$

By setting the derivative to zero, the minimum MSE (MMSE) of the linear estimator is

$$\min_{\mathbf{w}_{j,l}}E\left(\left|\mathbf{w}_{j,l}^T\mathbf{y}_j-\mathbf{a}_{j,l}^T\mathbf{x}\right|^2\right)=\mathbf{a}_{j,l}^T\left(\rho\mathbf{H}_j^T\mathbf{H}_j+\mathbf{I}_K\right)^{-1}\mathbf{a}_{j,l}. \tag{42}$$

As n tends to infinity, the effective noise sphere is given by this MMSE for computing the lth NC stream. Then, there exist a nested lattice code with simultaneous "Roger-goodness" and "Poltyrev-goodness", such that the rate

$$\frac{1}{2}\log_2^+\left(\frac{1}{\mathbf{a}_{j,l}^T\left(\rho\mathbf{H}_j^T\mathbf{H}_j+\mathbf{I}_K\right)^{-1}\mathbf{a}_{j,l}}\right) \tag{43}$$

w.r.t. the lth NC stream is achievable [18], [29]. ■

The symmetric rate is pertain to the open-loop uplink system [37]. For the LNC-based cf-MIMO, the symmetric rate is characterized in the following theorem. Recall that L_j denotes the number of rows of $\mathbf{A}_j$, and C_j^{BH} denote the finite BH capacity of BS j. For $\mathbf{H}_j$, let the MMSE matrix be denoted by $\mathbf{\Psi}_j=\left(\rho\mathbf{H}_j^T\mathbf{H}_j+\mathbf{I}_K\right)^{-1}$. Its eigen-decomposition is given by

$$\mathbf{\Psi}_j=\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^T. \tag{44}$$

Theorem 2: A symmetric rate R_0 is achievable if there exist integer valued coefficient matrices $\mathbf{A}_1,\cdots,\mathbf{A}_N$ at the N BSs, such that

$$\mathbf{D}_j^{\frac{1}{2}}\mathbf{V}_j^T\mathbf{a}_{j,l}<\sqrt{\frac{1}{2^{2R_0}}},\forall l=1,\cdots,L_j,j=1,\cdots,N, \tag{45}$$

$$\mathbf{A}=\left[\mathbf{A}_1^T,\cdots,\mathbf{A}_N^T\right]^T \text{ is of full rank } K \text{ in } \mathbb{Z}_{2^m}. \tag{46}$$

and

$$R_0L_j<C_j^{BH},j=1,\cdots,N. \tag{47}$$

Proof. From Lemma 1, the symmetric rate with which BS j can compute L_j NC streams is

$$\begin{aligned}R_0\le R_j&=\min_{l=1,\cdots,L_j}R_{j,l}^{comp}\\&=\min_{l=1,\cdots,L_j}\frac{1}{2}\log_2^+\left(\frac{1}{\mathbf{a}_{j,l}^T\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^T\mathbf{a}_{j,l}}\right).\end{aligned} \tag{48}$$

It follows directly that all L_j NC streams at BSs $1,\cdots,N$ can be reliably computed if

$$\min_{j=1,\cdots,N,l=1,\cdots,L_j}\frac{1}{2}\log_2^+\left(\frac{1}{\mathbf{a}_{j,l}^T\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^T\mathbf{a}_{j,l}}\right)>R_0. \tag{49}$$

This leads to (45). The entropy of the NC streams of BS j is precisely

$$H\left(\mathbf{U}_j\right)=R_0L_j \tag{50}$$

According to the source-coding theorem, $\mathbf{U}_j$ can be lossless recovered by the CU if

$$H\left(\mathbf{U}_j\right)=R_0L_j<C_j^{BH},j=1,\cdots,N, \tag{51}$$

which leads to (47). Given the full rank condition (46), all users' messages are recovered. ■

Remark 10:

For lattice-based processing with ring-coded PAM, the achievable mutual information that takes into account the 2^m-PAM modulation should be used for characterization. We have reported this in our previous work [24]. However, for the case with a large number of users and BS antennas, it is well-known that calculating the exact mutual information for 2^m-PAM requires a multi-dimension integration. This makes finding the optimized $\mathbf{A}$ intangible. To obtain a viable and pragmatic solution to $\mathbf{A}$, in this paper we resort to the succinct rate expression presented in Lemma 1, which is an upper bound for the achievable mutual information with ring-code PAM. Then, the BIVP w.r.t. Theorem 2 is formulated based on this rate upper bound. Solving that problem with our proposed constrained sphere decoding algorithm provides a competitive pragmatic solution, as demonstrated with our extensive numerical results. We note that the NC coefficient matrices are determined based on the instantaneous rate, which is calculated based on the receiver side CSI at each BS. The solution is found by solving the BIVP given in Theorem 2.

B. Optimized Coefficient Matrices $\mathbf{A}_1,\cdots,\mathbf{A}_N$

1) Optimized Design with BIVP: The central problem to be addressed is: based on the *local* CSI of $\mathbf{H}_j$, BS j finds all linearly independent lattice points, formed by the basis vectors $\mathbf{D}_j^{\frac{1}{2}}\mathbf{V}_j^T$, within the boundary of radius $\sqrt{\frac{1}{2^{2R_0}}}$ (45). We referred to this as a *bounded independent vectors problem* (BIVP).

2) Solution to BIVP: We suggest a *constrained sphere decoding* (CSD) which solves BIVP (45). The goal is to find all coefficient vectors that are within the boundary of radius $\sqrt{\frac{1}{2^{2R_0}}}$. In CSD, we set the search center to an all-zero vector, and set the radius to $\sqrt{\frac{1}{2^{2R_0}}}$. Choleski factorization

$$\mathbf{T}\mathbf{T}^T=\mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^T$$

is applied where $\mathbf{T}$ is a triangular matrix. Next, a tree search is implemented over the layers of $\mathbf{T}$ as in sphere decoding, which identifies all coefficient vectors within the radius. The rank of these coefficient vectors in $\mathbb{Z}_{2^m}$ are calculated, denoted by $\widetilde{L}_j$. Finally, we pick those $L_j\le\widetilde{L}_j$ linearly independent vectors with the smallest norms, with L_j being set to meet the BH constraint (47).

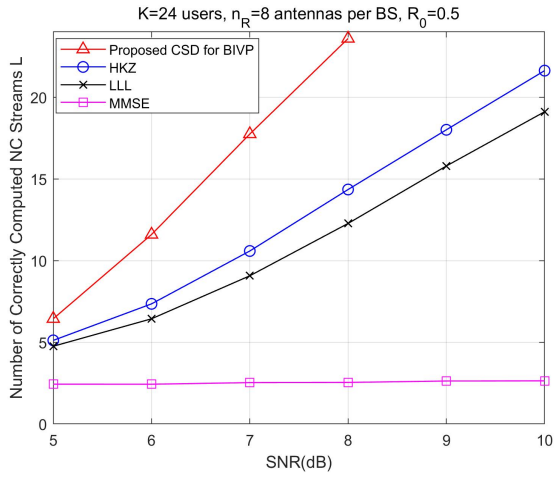

Fig. 2. Averaged number of correctly computed NC streams L at each BS, $n_R=8,K=24,R_0=0.5$.

3) Comparison: Recall that the objective of the design is to identify a good coefficient matrix at each BS, such that

it can correctly compute and forward as many NC streams as possible. Fig. 2 shows the averaged number of correctly compute NC messages L at each BS, where $n_R = 8, K = 24$. The greater the L, the higher the chance that CU can recover all users' messages. For the conventional non-lattice coded based scheme with MMSE detection where $\mathbf{A} = \mathbf{I}$, L is about 2.5 and barely increases with SNR. The conventional processing with MMSE is a user-by-user based processing method. A linear MMSE filter is employed to just suppress the multi-user interference, which does not sufficiently exploit the interference structure. In contrast, the proposed scheme is set to operate over integer linear combination (or network coding) streams. By adopting an optimized $\mathbf{A}$ matrix, the multi-user interference structure is much better harnessed in the LNC scheme. For the LNC based cf-MIMO scheme, L increases with SNR, suggesting the CU can recover all users' messages at a sufficiently high SNR. It is clear that our proposed CSD method, that solves the BIVP, considerably outperforms existing LLL and HKZ lattice reduction methods [27]. The gains in outage probability and FER will be shown in Fig. 8 in the next section.

C. Discussion of LNC

In essence, LNC involves two functionalities: 1) channel-code decoding w.r.t. $\mathbf{U}_j$, and 2) multi-user data compression. For 1), based on $\mathbf{Y}_j$, LNC decodes the partial information $\mathbf{U}_j$ by exploiting the 2^m-ary channel code, as detailed in Section IV. For 2), $\mathbf{U}_j$ is a "*compressed version*" of all K users' data $\mathbf{B}$, where $\mathbf{A}_j$ serves as the compression code (or network code). This is in line with the spirit of general network coding [16]. In particular, the network code $\mathbf{A}_j$ is chosen to facilitate the decoding of $\mathbf{U}_j$, and maintain as much information of $\mathbf{B}$ as possible to maximize the network information flow.

The "goodness" of the network code $\mathbf{A}_j$ depends on:

a) The algorithm for identification of $\mathbf{A}_j$. From the above, solving BIVP with CSD yields a better network code $\mathbf{A}_j$ as compared to the LLL and HKZ lattice reduction methods.

b) The wireless channel realization $\mathbf{H}_j$. A "good" $\mathbf{H}_j$ enables more linearly independent NC coefficient vectors within radius $\sqrt{\frac{1}{2^{2R_0}}}$ as in Th.2, resulting in a greater L_j.

In cf-MIMO, some BSs may have "bad" (or "good") channel realization, and outputs a little (or large) number of NC streams. As long as the N BSs can collectively provide sufficient amount of NC streams, the task of CU decoding can be accomplished.

In executing the LNC scheme, each BS can reliably compute L integer linear combinations (equations). If the total rate of these L equations is no greater than its BH rate constraint, all the L equations can be forwarded to the CU. On the other hand, if the total rate of these L equations is greater than its BH rate constraint, the BS just selects the best $L' < L$ equations, where L' is chosen to meet the BH requirement. It is noteworthy that the BH consumption is affected by the BIVP solution of $\mathbf{A}$ matrix. To be specific, the BIVP is set to find as much independent a vector as possible within a bounded region whose radius is determined by the target rate R_0. For different MIMO channel realizations at different BSs, the number of satisfactory vectors $L_1, ..., L_N$ are different, thus the BH consumption of the BSs are different and asymmetrical in general.

VI. Numerical Results

The environment of the simulations follows that in the system model depicted in Section II. For a cf-MIMO with given N, K, n_R and target rate R_0, outage probability (OP) at the CU is served as the performance metric from a theoretical aspect. The frame error rate (FER) of a 2^m-ary LDPC coded system is served as the performance metric from a practical aspect. The OP provides a lower bound for the FER. Here, i.i.d. Rayleigh fading channels for all users and all BSs are considered. The extension to the setup with randomly located users with different path-losses is straightforward. For comparison purpose, we consider decode-forward (DF) [46] and compress-forward [25] [35] [36] as the baseline schemes, whose achievable rate and BH consumption are presented in Appendix. Note that both LNC and DF have a fully digitized usage of the BH, while compress-forward does not. Local CSI with and without channel estimation errors are considered at each BS. In the numerical results, the vertical axis is either FER or outage probability.

A. Moderate numbers of antennas and users

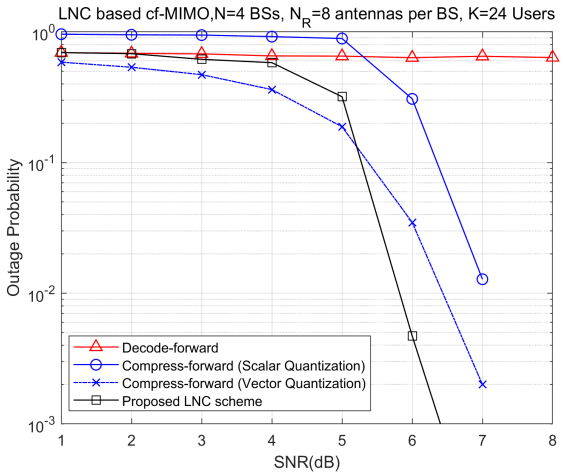

Fig. 3. Comparison of OPs of DF, compress-forward and our proposed LNC based schemes. A cf-MIMO system with $N = 4$ distributed BSs, $n_R = 8$ antenna at each BS, $K = 24$ users, target per-user rate $R_0 = 0.5$ is considered in this figure.

First consider a cf-MIMO with $N = 4$ BSs, $K = 24$ users, $n_R = 8$ antennas at each BS, and a target per-user rate of $R_0 = 0.5$ bit per real-dimension. The spectral efficiency is 12 bits per real-dimension. The channel coefficients are assumed to be i.i.d. and follow a Rayleigh distribution. Fig. 3 shows the OPs of the cf-MIMO system with various schemes including baseline DF, compress-forward and our proposed LNC based scheme. Associate with a certain symmetric rate R_0, the outage probability is defined as the probability that a users achievable rate is less than R0 [37]. For the DF scheme, the OP hardly drops as SNR increases. This is because at each BS, the DF scheme has to rely on only $n_R = 8$ antennas to decode

$K = 24$ users' message. In this case, the BS can only correctly decode and forward a very limited number of users' messages. In contrast, both the compress-forward based scheme and our proposed scheme have OPs that decrease as SNR increases. Here, the BH usage of the compress-forward based scheme and our proposed LNC based scheme are set to be identical for a fair comparison. The achievable rate of the LNC scheme characterized in Theorem. 2 is used to evaluate the OP. At an OP$\leq 10^{-2}$, the proposed LNC based scheme outperforms the compress-forward based scheme with either optimal vector quantization or scalar quantization. Note that to achieve the OP with vector quantization, an optimal vector quantization code that achieves the rate-distortion function with a very long codeword size $n \to \infty$ is required. In practice, such vector quantization code may not be available, or its encoding and decoding may not be affordable. The compress-forward with scalar quantization is thus a more viable solution.

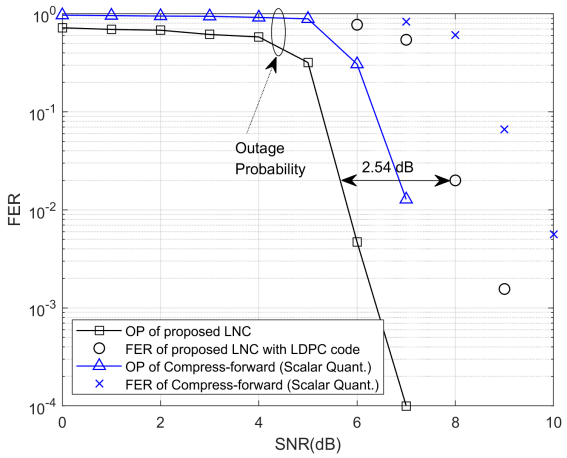

Fig. 4. FER of binary LDPC coded LNC for a cf-MIMO system with $N = 4$ distributed BSs, $n_R = 8, K = 24, R_0 = 0.5$. A rate 1/2 code with BPSK signaling is used. The block size is $k = 480$. The behaviors of FERs agree with that of OPs.

The OP shown in Fig. 3 provides a theoretical upper bound on the FER that any practical coded system can achieve. Fig. 4 presents the FER of the proposed LNC scheme using a practical LDPC code in 5G standard. To meet the target rate of $R_0 = 0.5$ per-user, a rate 1/2 code with BPSK signaling is used. The block size is $k = 480$. The maximum number of iterations of LDPC code decoding is set to 200. In the point-to-point AWGN channel, this code achieves a FER of $= 10^{-2}$ (or 10^{-3}) at 1.68 dB (or 1.99 dB) , which is about 1.69 dB (or 1.99 dB) away from the capacity limit (0 dB for $R_0 = 0.5$). For the proposed LNC scheme in cf-MIMO, our developed soft LNC detection algorithm in Section IV. D is utilized, which yields the input log likelihood ratios (LLRs) to the LDPC decoders to execute BP decoding. At FER=10^{-2}, the performance of the LDPC coded LNC scheme is about 2.54 dB away from the theoretical OP. The FER of the proposed is about 1.4 dB better than that of the compress-forward scheme with scalar quantization. The behaviors of FERs agree with that of OPs. Fig. 5 shows the FER performance of the proposed LNC scheme with various MSE of channel

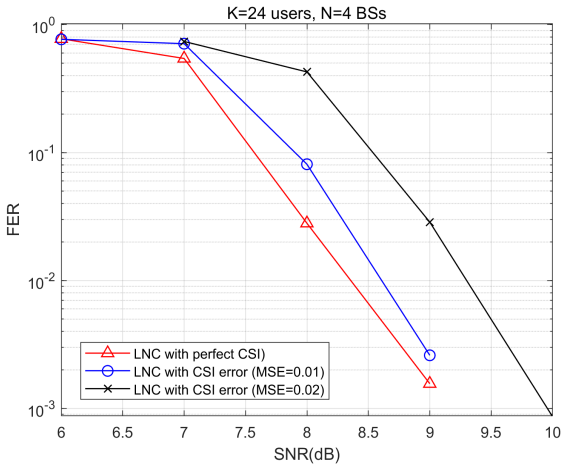

Fig. 5. FER of binary LDPC coded LNC for a cf-MIMO system with channel estimation errors, for $N = 4$ distributed BSs, $n_R = 8, K = 24, R_0 = 0.5$. A rate 1/2 code with BPSK signaling is used. The block size is $k = 480$.

estimation errors. We consider that the estimated channel matrix is given by

$$\widehat{\mathbf{H}} = \mathbf{H} + \mathbf{\Gamma} \tag{52}$$

where $\mathbf{\Gamma}$ denotes the channel estimation error. The mean square error (MSE) w.r.t. is given by

$$E\left(|\widehat{h}_{i,j} - h_{i,j}|^2\right) = E\left(|\tau_{i,j}|^2\right) \tag{53}$$

where $\tau_{i,j}$ is the (i,j)-th entry of $\tau_{i,j}$. The entries of $\mathbf{\Gamma}$ are modelled as following a i.i.d. Gaussian distribution. In obtaining the numerical result in Fig. 5, the MSE $E\left(|\tau_{i,j}|^2\right)$ is set to 0.01 and 0.02. The selection of the NC matrix and the soft LNC detection are then implemented based on the estimated channel matrix $\widehat{\mathbf{H}}$. It is demonstrated that the LNC scheme is subject to losses of about 0.2dB and 0.95 dB respectively, relative to the perfect CSI case.

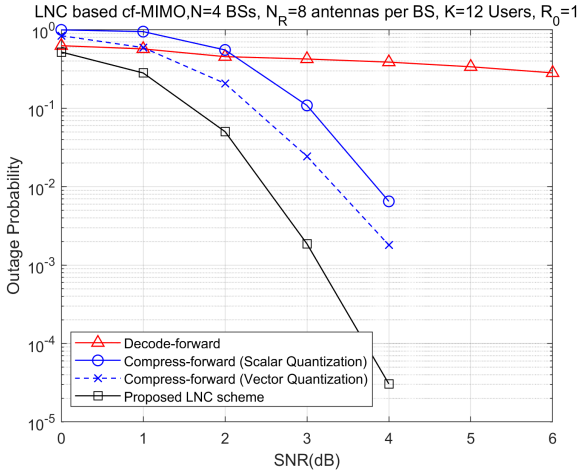

Fig. 6. Outage probability of the cf-MIMO with $N = 4$ distributed BSs, $n_R = 8, K = 12, R_0 = 1$.

Next consider a cf-MIMO with $N = 4$ BSs, $K = 12$ users, $n_R = 8$ antennas at each BS, where the target per-user rate is increased to $R_0 = 1$. The spectral efficiency is 12 bit/per real-dimension, which is the same as in the previously evaluated configuration. In Fig. 6, it is shown that the OP of the proposed LNC scheme considerably outperforms the compress-forward scheme with either optimal vector quantization or scalar quantization, with the same BH consumption. It is also shown that the LNC based scheme outperforms the decode-and-forward scheme, where joint decoding of all users messages by means of iterative APP detection and decoding is utilized.

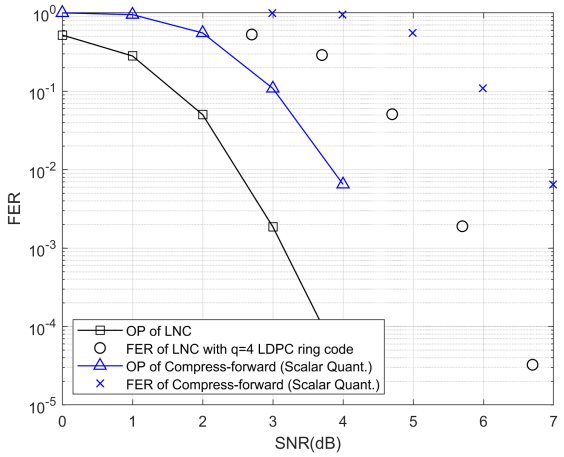

Fig. 7. FER of a $2^m = 2$ LDPC coded cf-MIMO system with $N = 4$ distributed BSs, $n_R = 8, K = 12, R_0 = 1$.

Fig. 7 presents the FER. To meet the target rate of $R_0 = 1$ per-user, a rate 1/2 4-ary LDPC (a doubly irregular repeat-accumulate [43]) ring code with 4-PAM signaling is used, which belongs to the ensemble of lattice code. The block size is $k = 256$. In the point-to-point AWGN channel, this code achieves a FER of $= 10^{-2}$ at 7.61 dB, which is about 2.84 dB away from the capacity limit (4.77 dB for $R_0 = 1$). For the proposed LNC scheme in cf-MIMO, our developed soft LNC detection algorithm in Section IV. D is utilized, which yields the 2^m-level APPs to the LDPC ring code decoders that execute 2^m-ary BP decoding. At FER=10^{-2}, the FER is about 2.96 dB away from the OP upper bound for the proposed scheme. The FER of the proposed LNC scheme is about 1.6 dB better than that of the compress-forward scheme with scalar quantization. Again, the behaviors of FERs agree with that of OPs for the scenario with a higher-level modulation and 2^m-ary code.

Fig. 8 compares the OPs of the LNC based cf-MIMO with various methods in finding $\mathbf{A}_j, j = 1, \cdots, N$. The first baseline is MMSE, which is equivalent to LNC with coefficient matrix $\mathbf{A}$ is set to $\mathbf{I}$ [29]. It has a much inferior performance w.r.t. lattice-based schemes. Our proposed CSD outperforms existing LLL and HKZ methods by 1.2 and 1.5 dBs at OP of 10^{-3}, respectively. This advance is owing to that the proposed CSD is able to provide more NC streams from each distributed BSs, as explained in Fig. 2.

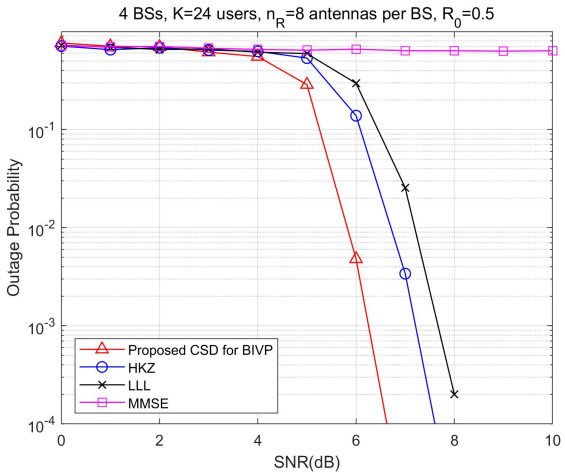

Fig. 8. Comparison to existing LLL and HKZ lattice reduction methods. $N = 4$ distributed BSs,$n_R = 8, K = 24, R_0 = 0.5$.

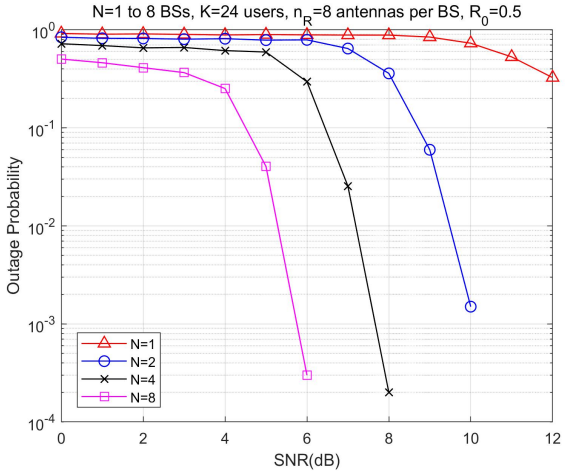

Fig. 9. Comparison of LNC based cf-MIMO with $N = 1, 2, 4, 8$ BSs, $n_R = 8, K = 24, R_0 = 0.5$. The LLL method is used.

Fig. 9 compares the performance of LNC based cf-MIMO with $N = 1$ to 8 distributed BSs.

B. Large numbers of antennas and users

We next consider a cf-MIMO with a larger number of antennas $n_R = 32$ at BS, where $N = 4, R_0 = 1$. Fig. 10 shows the OP (left sub-figure) and the BH consumption (right sub-figure) with various numbers of users $K = 32, 40, 48$. As the number of users increase, the total data-rate increases, and a higher SNR is required to achieve a certain OP requirement. For example, to achieve OP=10^{-2}, the required SNR is -3.72 dB,-1.22 dB, 3.21 dB for $K = 32, 40, 48$, respectively. It is interesting to note to achieve OP=10^{-2}, the BH consumptions does not considerably increase as K increases, for the range of K under consideration. For instance, each BS is required to consume about 22.1, 22.6, 20.5 bits/channel-use of the BH on average for $K = 32, 40, 48$, respectively. We conjecture that this is because for a higher SNR, the NC messages aggregated

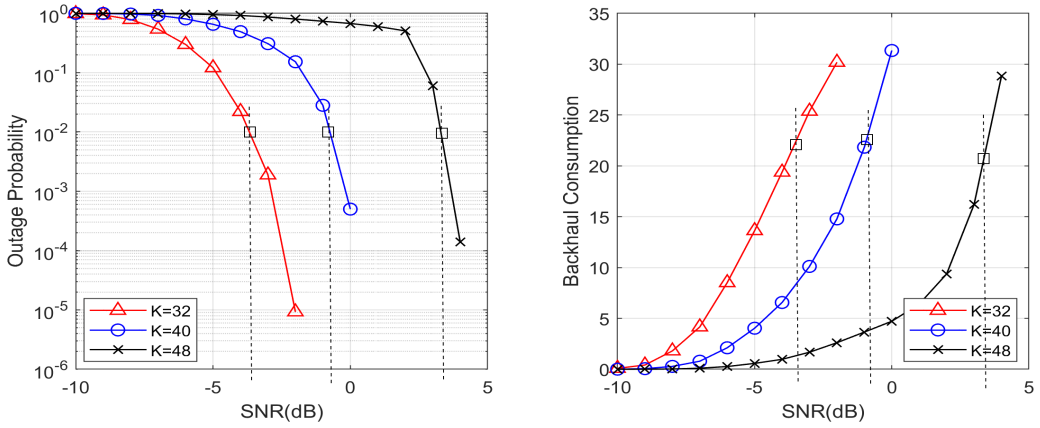

Fig. 10. Outage probability and BH rate consumption with $N = 4$ distributed BSs, $n_R = 32$, $K = 32, 40, 48$, $R_0 = 1$.

from the BSs tend to be more linearly independent, thus a similar number of NC messages from each BS suffices to guarantee the full-rank condition at the CU for an increased number of K.

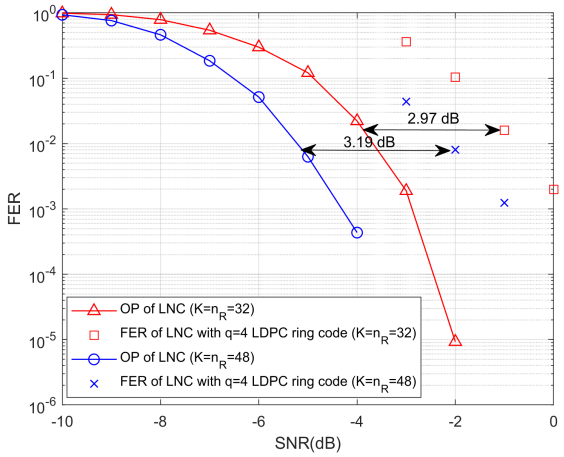

Fig. 11. FER with $N = 4$ distributed BSs, $n_R = 32, 48$, $K = 32, 48$.

To consolidate the result with large n_R and K, Fig. 11 shows the FER of the proposed LNC based scheme with the $2^m = 4$ LDPC code of rate 1/2 and 4-PAM utilized in Fig. 7. At FER=10^{-2}, the gap to the theoretical OP is about 2.97 dB for $n_R = K = 32$ and 3.19 dB for $n_R = K = 48$. The behaviors of FERs agree with that of the OPs. Fig. 12 shows the ϵ-outage rate [37] of LNC-based cf-MIMO system. The ϵ-outage rate is defined as the rate R at which the outage probability of the scheme equals to ϵ. The typical values of ϵ widely used are from 0.1 to 0.001. It is demonstrated that the proposed LNC scheme offers a very much improved outage capacity over the benchmark scheme with decode-forward. As the system load increases, the improvement becomes more significant.

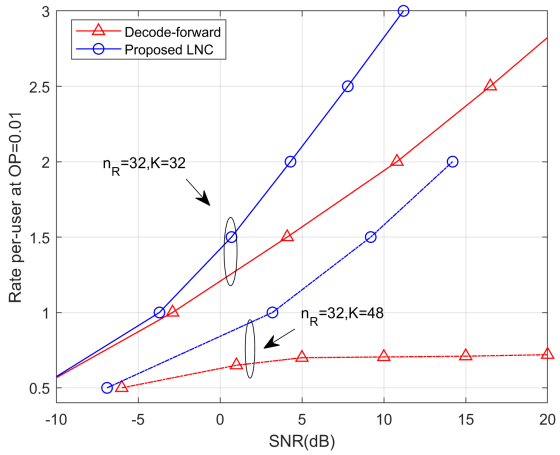

Fig. 12. Rate per-user at OP=0.01. $n_R = 32$, $K = 32, 48$.

TABLE I
THE ORDER OF COMPLEXITY OF LNC BASED CF-MIMO.

	Detection	Decoding	NC Coefficient Identification
LNC	$O\left(L_{CU} n 2^m \cdot E\left(\omega_H\left(\mathbf{a}\right)\right)\right)$	$O(L_{CU} n\left(2^m - 1\right))$	Between $O(NK^3)$ and $O(NK^4)$
DF	$O\left(NKn2^m\right)$	$O(NKn\left(2^m - 1\right))$	Nil

C. Analysis of Implementation Costs

The orders of complexities are shown in Table. I. The computation in LNC for a BS consists of 1) channel-code decoding, 2) LNC soft detection, and 3) identification of $\mathbf{A}_j$. For 1), LNC requires L_j decoding operations at BS j. For the uplink system, the modulation order 2^m is usually not large, where the complexity of ring-code decoding is not considerably greater than that based on conventional binary channel code decoding. For 2), LNC needs to compute L_j streams of APPs w.r.t. the NC messages. The per-symbol detection complexity (of calculating the distance) of stream l is of order $O((2^m - 1)\,\omega_H\left(\mathbf{a}_l\right))$, where $\omega_H\left(\mathbf{a}_l\right) < K$ denotes the weight of the coefficient vector $\mathbf{a}_l$. The average detection

complexity of LNC is thus $O\left(L_j n\left(2^m-1\right) E\left(\omega_H\left(\mathbf{a}\right)\right)\right)$ for BS j. For 3), with LLL, the complexity is between $O(K^3)$ and $O(K^4)$, a polynomial in K. The complexity of HKZ and CSD is moderately higher than LLL. Since $\mathbf{A}_j$ is chosen once per block, for a moderate-to-long block length n (e.g. $n > 480$), this overhead is not significant.

D. Discussion of Backhaul Consumption

The BH consumption of the LNC based cf-MIMO is of the same order of the air-interface capacity, given by $O\left(KR_0\right)$. To see this, note that the entropy of the NC streams is $\frac{1}{n}\sum_{j=1}^{N} H\left(\mathbf{U}_j\right) \leq R_0 \sum_{j=1}^{N} L_j$, which determines the BH consumption. Since $L_j \leq K$, the total BH consumption of N BSs is at most NKR_0, where the typical number of BSs N is not large. Empirically, to achieve FER of 10^{-2} to 10^{-3}, L_j is just a fraction of K. As such, the total BH rate consumption is significantly smaller than NKR_0, i.e., the order is $O\left(KR_0\right)$. Meanwhile, for a properly designed system, the sum-rate KR_0 should be in match with the channel capacity of the air-interface, whose order is also $O\left(KR_0\right)$.

VII. Conclusions

This paper studied lattice network coding based cell-free MIMO system with non-cooperative BSs. We suggested a package of techniques that are essential to its practical implementation, including the 2^m-ary ring-coded modulation, soft detection algorithms, constrained sphere-decoding for solving the BIVP that identifies the optimized coefficient matrix $\mathbf{A}$ at each base stations. Considerable performance enhancement were demonstrated over existing compress-forward and decode-forward schemes.

Appendix

[Rate of Compress-forward]

Each BS compress $\mathbf{y}$ w.r.t the n_R antennas. The description rate is chosen such that the BH capacity constraint is met. The correlation between two entries of $\mathbf{y}$ is

$$E\left(y_\upsilon, y_{\upsilon'}\right) = \sum_{i=1}^{K} h_{\upsilon,i} h_{\upsilon',i} \overset{K\to\infty}{\to} 0, \upsilon, \upsilon' \in \{1,\cdots,n_R\}, \upsilon \neq \upsilon' \tag{54}$$

where Law of large number is used in the last step. In other words, as K becomes large, the correlation among the entries of $\mathbf{y}$ vanishes. Thus, the entropy can be approximated as

$$H\left(\mathbf{y}\right) \overset{K\to\infty}{\approx} \frac{1}{2}\sum_{\upsilon=1}^{n_R} \log_2\left(\sum_{i=1}^{K}\left|h_{\upsilon,i}\right|^2\rho+1\right) + \frac{1}{2} n_R \log_2 2\pi e. \tag{55}$$

It is clear that the entropy of $\mathbf{y}$ is linear in n_R and logarithm in K. For cf-MIMO with large n_R and K under consideration, $H\left(\mathbf{y}\right)$ tends to be much greater than the BH capacity constraint, and thus compression is required. The compression of the signal $\mathbf{y}$ [35] [36] is depicted as follows.

Denote by y_υ the received signal of the υ-th antenna, and

$$\widetilde{y}_\upsilon = y_\upsilon + e_\upsilon, \upsilon \in \{1,\cdots,n_R\}$$

be its quantized version. Here MSE is used as the distortion measurement, i.e. $D_\upsilon = E\left(e_\upsilon^2\right)$. If an optimal *vector quantization code* that achieves the rate-distortion (RD) function is applied, according to the RD theorem [15], the per-antenna compression rate satisfies

$$\begin{aligned} R\left(D_\upsilon\right) &> \min_{p(y_\upsilon|\widetilde{y}_\upsilon)} I\left(y_\upsilon;\widetilde{y}_\upsilon\right) = \frac{1}{2}\log_2\left(\frac{var\left(y_\upsilon\right)}{D_\upsilon}\right) \\ &= \frac{1}{2}\log_2\left(\frac{\sum_{i=1}^{K}\left|h_{\upsilon,i}\right|^2\rho+1}{D_\upsilon}\right) = \frac{C_j^{BH}}{n_R}, \end{aligned} \tag{56}$$

for $\upsilon = 1,\cdots,n_R$. The inequality is owing to the BH constraint, where an equal rate allocation is applied to quantize the n_R antennas' signals. This translates into quantization noise variance

$$D_\upsilon = \left(\sum_{i=1}^{K}\left|h_{\upsilon,i}\right|^2\rho+1\right) 2^{\frac{-2C_j^{BH}}{n_R}}. \tag{57}$$

The quantized version obtained from the N BSs are forwarded to the CU. The CU aggregates

$$\widetilde{\mathbf{y}}_{CU} = \left[\widetilde{\mathbf{y}}_1^T,\cdots,\widetilde{\mathbf{y}}_N^T\right]^T = \mathbf{H}_{CU}\mathbf{x} + \mathbf{z}_{CU} \tag{58}$$

where the entries of $\mathbf{z}_{CU}$ has variances $D_{j,\upsilon}+1$, $\upsilon = 1,\cdots,n_R, j = 1,\cdots,N$. Then, the achievable rate of this compress-forward scheme can be calculated from (58).

The above compress-forward scheme assumes a "perfect" vector quantization RD code for $n \to \infty$. In practice, the optimal vector quantization code would be either unavailable or too expensive. If a simple 1-bit quantization is applied symbol-wisely (without using a RD code) per-antenna, the distortion is calculated to be $0.3633\left(\sum_{i=1}^{K}\left|h_{\upsilon,i}\right|^2\rho+1\right)$. In such case, the total BH consumption is exactly n_R bits.

References

[1] W. Tong and P. Zhu, "6G the next horizon-from connected people and things to connected intelligence," *Cambridge Univ. Press*, 2021.

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *IEEE 16th SPAWC*, pp. 201–205, 2015.

[3] A. Burr, M. Bashar, and D. Maryopi, "Ultra-dense radio access networks for smart cities: Cloud-RAN, fog-RAN and cell-free massive MIMO," *in Proc. PIMRC*, p. 1C5, Sep. 2018.

[4] S. S. Shamai and B. M. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," *in Proc. IEEE VTS 53rd Veh. Technol. Conf.*, p. 1745C1749, May 2001.

[5] S. Zhou, M. Zhao, X. Xu, J. Wang, and Y. Yao, "Distributed wireless communication system: a new architecture for future public wireless access," *IEEE Communications Magazine*, vol. 41, no. 3, pp. 108–113, 2003.

[6] D. Wang, Z. Zhao, Y. Huang, H. Wei, X. Wang, and X. You, "Large-scale multi-user distributed antenna system for 5g wireless communications," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2015.

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

[8] Z. Chen, E. Bjornson, and E. G. Larsson, "Dynamic resource allocation in co-located and cell-free massive MIMO," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 1, pp. 209–220, 2020.

[9] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596–44612, 2019.

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

[11] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, 2017.

[12] T. Yang, X. Yuan, and Q. T. Sun, "A signal-space aligned network coding approach to distributed mimo," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 27–40, 2017.

[13] T. Yang, "Distributed MIMO broadcasting: Reverse compute-and-forward and signal-space alignment," *IEEE Trans. Wireless Comm.*, vol. 16, no. 1, pp. 581–593, 2017.

[14] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.

[15] T. M. Cover and J. A. Thomas, "Elements of information theory," *John Wiley & Sons, Inc.*, 1991.

[16] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[17] S. Lim, Y. H. Kim, A. E. Gamal, and S. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory.*, vol. 57, no. 5, pp. 3132–3152, May. 2011.

[18] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[19] W. Nam, S. Chung, and Y. H. Lee, "Capacity of the Gaussian two-way relay channel to within 1/2 bit," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5488–5494, Nov. 2010.

[20] X. Yuan, T. Yang, and I. Collings, "Multiple-input multiple-output two-way relaying: a space-division approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6421–6440, Oct. 2013.

[21] S. H. Lim, C. Feng, A. Pastore, B. Nazer, and M. Gastpar, "Compute-forward for DMCs: Simultaneous decoding of multiple combinations," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6242–6255, 2020.

[22] Q. T. Sun, J. Yuan, T. Huang, and K. W. Shum, "Lattice network codes based on eisenstein integers," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2713–2725, 2013.

[23] S. Zhang, S. Liew, and P. P. Lam, "Hot topic: physical-layer network coding," *Proceedings of the 12th annual international conference on Mobile computing and networking*, pp. 358–365, 2006.

[24] Q. Chen, F. Yu, T. Yang, and R. Liu, "Gaussian and fading multiple access using linear physical-layer network coding," *IEEE Trans. Wireless Comm.*, May,2023.

[25] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[26] S. Lyu, A. Campello, and C. Ling, "Ring compute-and-forward over block-fading channels," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 6931–6949, 2019.

[27] A. Sakzad, J. Harshan, and E. Viterbo, "Integer-forcing MIMO linear receivers based on lattice reduction," *IEEE Trans. Wireless Comm.*, vol. 12, no. 10, pp. 4905–4915, 2013.

[28] S. Lyu and C. Ling, "Boosted KZ and LLL algorithms," *IEEE Trans. Signal Proc.*, vol. 65, no. 18, pp. 4784–4796, 2017.

[29] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, "Integer-forcing linear receivers," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7661–7685, Dec. 2014.

[30] D. Silva, G. Pivaro, G. Fraidenraich, and B. Aazhang, "On integer-forcing precoding for the Gaussian MIMO broadcast channel," *IEEE Tran. Wireless Comm.*, vol. 16, no. 7, pp. 4476–4488, 2017.

[31] D. Yang and K. Yang, "Multimode integer-forcing receivers for block fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8261–8271, 2020.

[32] O. Ordentlich and U. Erez, "Cyclic-coded integer-forcing equalization," *IEEE Transactions on Information Theory*, vol. 58, no. 9, pp. 5804–5815, 2012.

[33] J. Zhu and M. Gastpar, "Gaussian multiple access via compute-and-forward," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2678–2695, 2016.

[34] Q. Chen, F. Yu, T. Yang, J. Zhu, and R. Liu, "A linear physical-layer network coding based multiple access approach," in *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2803–2808, 2022.

[35] Y. Tan and X. Yuan, "Compute-compress-and-forward: Exploiting asymmetry of wireless relay networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 511–524, 2016.

[36] H. Cheng, X. Yuan, and Y. Tan, "Generalized compute-compress-and-forward," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 462–481, 2019.

[37] D. Tse and P. Viswanath, "Fundamentals of wireless communication," *Cambridge University Press*, 2005.

[38] Q. Huang and A. Burr, "Compute-and-forward in cell-free massive MIMO: Great performance with low backhaul load," *in Proc. IEEE Int. Conf. Commun. Workshops*, pp. 601–606, 2017.

[39] J. Zhang, J. Zhang, D. W. K. Ng, S. Jin, and B. Ai, "Improving sum-rate of cell-free massive MIMO with expanded compute-and-forward," *IEEE Transactions on Signal Processing*, vol. 70, pp. 202–215, 2022.

[40] T. Huang, F. Yu, Q. Chen, and Q. Liu, "On lattice-code based multiple access: Uplink architecture and algorithms," *arXiv: 10.48550/arXiv.2210.00778*, 2022.

[41] T. Yang, L. Yang, Y. J. Guo, and J. Yuan, "A non-orthogonal multiple-access scheme using reliable physical-layer network coding and cascade-computation decoding," *IEEE Trans. Wireless Comm.*, vol. 16, no. 3, pp. 1633–1645, 2017.

[42] N. Sommer, M. Feder, and O. Shalvi, "Low-density lattice codes," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1561–1585, 2008.

[43] F. Yu, T. Yang, and Q. Chen, "Doubly irregular repeat modulation codes over integer rings for multi-user communications," *China Comm. (accepted)*, 2023 (available at: *https://arxiv.org/abs/2210.01330)* .

[44] S. Lin and D. J. Costello, "Error control coding, 2nd edition," *Pearson*, 2004.

[45] M.-C. Chiu, "Bandwidth-efficient modulation codes based on nonbinary irregular repeat–accumulate codes," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 152–167, 2009.

[46] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

Tao Yang (S'07, M'10) received the B.Sc. degree in electronics information engineering from Beihang University (Beijing University of Aeronautics and Astronautics), China, in 2003. He received Ph.D. degrees in electrical engineering from the University of New South Wales (UNSW), Australia, in 2010. He was an OCE Postdoctoral Fellow with Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. He was a faculty member within the Global Big Data Technologies Centre and School of Computing and Communications, University of Technology Sydney (UTS), Australia. He held an Australian Research Council (ARC) Discovery Early Career Research Award (DECRA) Fellowship, and was the recipient of Australian Postgraduate Award (APA), NICTA research project award and Supplementary Engineering Award. He is currently with Beihang University, Beijing, China, where his research works are supported by National Natural Science Foundation of China (NSFC), National Key R&D Program of China and Beijing Natural Science Foundation. He has authored over 90 research articles in IEEE journals and conferences. He has been serving as the reviewer of ARC proposals, NSFC proposals, IEEE journals, and Technical Program Committee (TPC) members of the IEEE ICC and WCNC conferences. His research interests are 5G/6G wireless communication, space-air-ground integrated communication, cell-free networks, coding/signal processing for communications and network information theory. His specialties are multiple access, lattice codes, physical-layer network coding (compute-forward), MIMO and distributed MIMO, and iterative decoding/signal processing techniques.