On Lattice-Code based Multiple Access: Uplink Architecture and Algorithms

Tao Yang, *Member, IEEE*, Fangtao Yu, Qiuzhuo Chen, Rongke Liu, *Senior Member, IEEE*

Abstract

This paper studies a lattice-code based multiple-access (LCMA) framework, and develops a package of processing techniques that are essential to its practical implementation. In the uplink, K users encode their messages with the same ring coded modulation of 2^m-PAM signaling. With it, the integer sum of multiple codewords belongs to the n-dimension lattice of the base code. Such property enables efficient *algebraic binning* for computing linear combinations of K users' messages. For the receiver, we devise two new algorithms, based on linear physical-layer network coding and linear filtering, to calculate the symbol-wise a posteriori probabilities (APPs) w.r.t. the K streams of linear codeword combinations. The resultant APP streams are forwarded to the q-ary belief-propagation decoders, which parallelly compute K streams of linear message combinations. Finally, by multiplying the inverse of the coefficient matrix, all users' messages are recovered. Even with single-stage parallel processing, LCMA is shown to support a remarkably larger number of users and exhibits improved frame error rate (FER) relative to existing NOMA systems such as IDMA and SCMA. Further, we propose a new multi-stage LCMA receiver relying on *generalized matrix inversion*. With it, a near-capacity performance is demonstrated for a wide range of system loads. Numerical results demonstrate that the number of users that LCMA can support is no less than 350% of the length of the spreading sequence or number of receive antennas. Since LCMA relaxes receiver iteration, off-the-shelf channel codes in standards can be directly utilized, avoiding the compatibility and convergence issue of channel code and detector in IDMA and SCMA.

Index Terms

Multiple access, coded modulation, lattice codes, compute-forward, multi-user MIMO, grant-free random access

I. Introduction

Beyond 5G (B5G) and 6G systems are envisaged to support a vast range of massive connectivity, throughput-hungry and latency-sensitive scenarios such as ubiquitous IoT, mobile cloud computing and etc., which calls for advanced multiple access (MA) techniques to achieve higher and flexible system loads, higher

This work is supported by the National Key R&D Program of China (No.2020YFB1807102). The authors are with Beihang University, China

spectral-energy efficiency and lower latency, with affordable complexity and implementation costs. MA is a pivotal component that distinguishes different generations of mobile systems. In 1G to 4G systems, orthogonal MA (OMA) schemes are utilized, where K users are allocated with separated frequency bands, separated time-slots, orthogonal spreading codes, and orthogonal sub-carriers, respectively. OMA endeavors to avoid the existence of multi-user interference, with the hope that users' signals can be separated by employing low-cost receivers. In 5G system, the number of base station antennas that is much larger than the number of users K is implemented. As such, the spatial signatures of the K users are nearly orthogonal, and OMA processing can be largely preserved. Nowadays, it is widely realized that OMA is fundamentally limited by the following aspects. First, the maximum number of supported users is capped at the number of available orthogonal resources, thus it is not possible to meet the connection density requirement envisaged for B5G and 6G. Second, sophisticated dynamic resource allocation protocols and algorithms are required to guarantee orthogonality, whose complexity skyrockets as the connection density levels up. Last but not the least, despite the utilization of orthogonal resources, the wireless channel induces impairments that easily destroy the orthogonality. This may invoke an orthogonality-restoring process with unaffordable implementation cost.

Non-orthogonal MA (NOMA) has been intensively studied in the past two decades. By allowing the existence of multi-user interference, the number of users K in NOMA can go beyond the available time, frequency and spatial resources, and grant-free access becomes possible. The core issue of NOMA is how to deal with multi-user interference. Accompanying with NOMA, interference cancellation and interference suppression based techniques were studied, such as successive interference cancellation (SIC) [1], zero-forcing (ZF) or minimum mean square error (MMSE) filtering. Not long after the discovery of turbo codes in 1993, the "turbo principle" was introduced and devised to solve the multi-user decoding problem, first by Wang&Poor in 1999 [2]. Since 2000, turbo-like iterative detection and decoding has been extensively researched. In "turbo-CDMA" [2], the inner code is a multi-user detector with soft interference cancellation and MMSE suppression, while the outer code is a bank of K convolutional code decoders. Soft probabilities are exchanged among these components iteratively until convergence is achieved. In 2006, Li *et al.* introduced a chip-level interleaved CDMA, named after interleave division multiple-access (IDMA) [3]. The chip interleaver enables uncorrelated chip interference, and thus a simple matched filter optimally combines the chip-level signal to yield the symbol-level soft information. Such an approach exploits the idea of *random-like coding* in dealing with the MA problem. Low-density spreading CDMA/OFMDA and sparse code multiple-access (SCMA) were proposed. It differs from IDMA in that each symbol-level information is spread only to a small number of chips, which forms a sparse matrix in the representation of the multi-user signal that can be depicted using a bi-partite factor graph [4]. SCMA also supports grant-free random access for the massive-connectivity scenario. Spatially coupled codes were also studied for dealing with the MA problem, yielding improved performance for fading MA channels thanks to the better universality [5]. For both IDMA and SCMA, spreading/sparse codes with irregular degree profiles were investigated including the work of ourselves [4], [6], which yielded improved convergence behavior of the multi-user decoding. Other code-domain NOMA techniques such as pattern division MA (PDMA), multi-user shared access (MUSA) and etc. [7] were studied. Rate-splitting MA (RSMA) was studied for closed-loop uplink and downlink systems [8], [9]. The idea is to superimpose a common message on top of the private messages, which is possible to enlarge the rate-region. For grant-free random access, the works on active user identification based on compressive sensing algorithms and coded slotted aloha protocols are also rich in the literature [10]–[12].

Albeit all the potential benefits promised by NOMA, there are several key challenging issues that still prevent NOMA schemes from being deployed in practice. As the uplink MA often utilizes open-loop system, pre-determined information rates apply and the system performance over fading wireless channel is mainly characterized by the frame error rate (FER). In such a setup, NOMA with SIC is subjected to an intrinsic performance loss in terms of outage probability, and the error propagation also severely affects its usefulness. Existing code-domain NOMA schemes require the outer-loop receiver iteration mentioned above, i.e., iterations among the inner multi-user detector and a bank of outer channel code decoders, otherwise the promised performance cannot be achieved. The outer-loop iteration has the following issues: First, it requires a good matching between the multi-user detector and decoders, following the well-known extrinsic information transfer (EXIT) chart principle. As the system load increases, the EXIT curve of the inner detector varies, and the decoder's structure has to be modified accordingly to guarantee that the EXIT curves adapt to each other. Otherwise, the iterative receiver will refuse to converge. This results in the effect that a certain channel code, e.g. off-the-shelf LDPC codes or polar codes in 5G NR standards, may not work in code-domain NOMA as the system load varies. Second, Q receiver iterations involve QK times of decoding operations and Q times of multi-user detection operations, hence the complexity may not be affordable for many use cases. In addition, the receiver iterations are serially processed, causing significant processing delay which is also preferred to be avoided.

From an information theoretic perspective, Zhu and Gastpar showed that any rate-tuple of the entire Gaussian MA capacity region can be achieved using a lattice-code based approach, and the scheme was named compute-forward MA (CFMA) [13]. Almost at the same time, we investigated using practical linear physical-layer network coding (LPNC), that borrows the notion of compute-forward [14], for fading MA [15]–

[17]. In contrast to the random-like coding approaches in existing code-domain NOMA schemes, "*structured codes*" based on lattice were proved to achieve a larger capacity region [13]. The idea of using lattice codes for tackling the MIMO detection and downlink MIMO precoding problems was reported in [18] and [19] under the name of integer-forcing (IF). The latter borrows the notion of reverse compute-forward [20], [21] and exploited the uplink-downlink duality. The design of CFMA and LPNC for the Gaussian MA channel with binary channel codes was studied in [22], [23]. To date, most of the related works have been focusing on achievable rates by proving the existence of "good" nested lattice codes, whereas the practical aspects are not yet sufficiently researched. The impacts of lattice codes on the key performance indicators of practical MA systems, such as the system load, FER, latency , complexity and etc., remain not reported in the literature.

A. Contributions

This paper advocates exploiting lattice coding techniques, which replace the turbo principle, in dealing with the MA problem. We study a lattice code-based multiple-access (LCMA) framework, and develops a package of processing techniques that are essential to its practical implementation. In the uplink, K users encode their messages with the same lattice code. We put forth to utilize a simplified yet powerful lattice coding technique, referred to as *ring coded modulation* (RCM), suitable for widely used 2^m-PAM or 2^{2m}-QAM signaling. Each user's code-modulated sequence may undergo a spreading with its signature sequence of length-N_S. The resultant signal sequences of the K users are transmitted simultaneously. Owing to the underlying ring code, the integer sum of multiple codewords belongs to the extended codebook of the base ring code, i.e., an n-dimension lattice. Such property enables efficient *algebraic binning* for computing linear combinations of users' messages. The operations are over the lattice formed by the extension of the base codebook, which is in contrast to those of existing NOMA schemes.

At the receiver side, a K-by-K coefficient matrix is first selected. To realize algebraic binning, we devise 1) *multi-dimenional linear physical-layer network coding with list sphere decoding* and 2) *linear filtering* based algorithms to calculate the symbol-wise a posteriori probabilities (APPs) w.r.t. the K streams of linear codeword combinations. The resultant APP streams are forwarded to K single-user decoders, which compute the K streams of linear message combinations in parallel. Finally, by multiplying with the inverse of the coefficient matrix, all users' messages are recovered. Even with single-stage parallel processing without receiver iteration, LCMA can support a remarkably larger number of users and exhibits improved FER performance relative to existing code-domain NOMA systems such as IDMA and SCMA.

Further, we propose a new multi-stage receiver. In the first stage, the LCMA receiver is set to compute only $L_{(1)}$, $L_{(1)} < K$, linear message combinations. We introduce *generalized matrix inversion* that can recover a subset of the users' messages from them. Then, the signals of these users are canceled from the received signal. This gives rise to a MA model with less users in the processing of the next stage, leading to enhanced performance of LCMA. The multi-stage processing continues until no further improvement can be achieved. In such a manner, near-capacity performance is demonstrated for a wide range of system loads.

B. Advantages

We demonstrate that the number of users that LCMA can support is no less than 350% of the length of the spreading sequence or the number of receive antennas. This leads to remarkably increased system load, improved FER performance and lower implementation costs compared to state-of-the-art code-domain NOMA schemes. Also, LCMA offers the following advantages that favor the practical implementation:

First, LCMA requires about K single-user decoding operations. In contrast, due to the requirement of outer-loop receiver iteration, IDMA or SCMA requires $Q \cdot K$ decoding operations, where the typical value of iteration number Q is between 4 to 10. In addition, the receiver iterations of IDMA/SCMA are implemented successively, thus LCMA with parallel processing has much lower decoding latency. The operation memory size of LCMA is also drastically reduced as there is no soft information to be fedback in the processing.

Second, for uplink LCMA with BPSK/QPSK signaling, off-the-shelf binary codes in various standards can be directly utilized for any system loads K/N, where N stands for the dimension of the received signal space. In contrast, from the principle EXIT chart or density evolution, IDMA and SCMA have to devise or adopt different codes as the system load K/N changes. Otherwise, the convergence of iterative detection and decoding may not be achieved, leading to overwhelmingly impaired performance or even failed functionality. This is particularly crucial for grant-free random access setup where the system load itself is random.

Furthermore, LCMA offers a unified framework that applies to code-domain NOMA, spatial division MA (SDMA), precoding for the downlink broadcast channel and etc.

II. System Model

Consider a single-cell that consists of $\widetilde{K}$ users and a base station (BS), without interference from other cells[1]. The following assumptions are made for the clarity and conciseness of the model: 1) The users are equipped with single-antenna and the BS is equipped with N_R antennas. The extension to multi-antenna users can be treated by allowing multi-streams for each user as treated in [24]. 2) Orthogonal frequency division multiplexing (OFDM) is presumed for wide-band frequency selective fading channel[2], thus there

[1] Lattice code based methods can also be developed for dealing with inter-cell-interference in a multi-cell setup.

[2] Other advanced waveform techniques such as generalized frequency division multiplexing and OTFS are beyond the scope of this paper.

is *no inter-symbol-interference* in the model. 3) A *block-fading* is assumed, where the channel coefficients remain unchanged for each block while differing over blocks[3]. This is guaranteed by allocating each user a segment of sub-carriers which is no greater than the coherent bandwidth. To justify this, consider the scenario with a large $\widetilde{K}$. All users in the cell are divided into non-overlapping groups. Each group contains K users that are assigned to the same subcarrier segment with flat fading channel coefficients. The users in different groups have non-overlapping sub-carriers, thus their signals are orthogonal to each other. As such, we are safe to consider a specific subcarrier segment that contains K users only and subject to block fading.

In the uplink, K users send signals to a BS. The BS aims to decode all users' messages. Let a row vector $\mathbf{x}_i^T$ denote a length-n coded-modulated symbol-level sequence of user $i, i = 1, \cdots, K$, with normalized average signal power $E\left(\mathbf{x}_i^T\mathbf{x}_i\right)/n = 1$. The symbol-level signal may be multiplied with its designated *spreading-signature* sequence $\mathbf{s}_i$ of length N_S, yielding the chip-level signal $\mathbf{s}_i\mathbf{x}_i^T$. The chip-level signal may or may not undergo an interleaving operation. Then all users' signals are transmitted simultaneously.

Consider coordinated MA and assume receiver-side synchronization[4] as in the convention [3]. In the case with a single-antenna receiver (or single-beam), i.e., $N_R = 1$, the base-band equivalent model is

$$\mathbf{Y} = \sum_{i=1}^{K} \widetilde{h}_i \sqrt{\rho}\mathbf{s}_i\mathbf{x}_i^T + \mathbf{Z} \tag{1}$$

where $\widetilde{h}_i$ denotes the channel coefficient (or the gain of the beam) of user i, $\mathbf{Y}$ and $\mathbf{Z}$ denote the received signal and additive white Gaussian noise (AWGN) with normalized variance $\sigma^2 = 1$, respectively. The per-user SNR is given by ρ. We only consider symmetric SNR. The extension to asymmetric SNR is straightforward.

For a N_R-antenna receiver, let $\widetilde{h}_{i,j}$ denote the channel coefficient from user i to the j-th antenna of the receiver, and let $\widetilde{\mathbf{h}}_i = \left[\widetilde{h}_{i,1}, \cdots, \widetilde{h}_{i,N_R}\right]^T$ be called the *spatial-signature* of user i. Let $\mathbf{h}_i = \left[\widetilde{h}_{i,1}\mathbf{s}_i^T, \cdots, \widetilde{h}_{i,N_R}\mathbf{s}_i^T\right]^T$ be the aggregation of the spreading-signature $\mathbf{s}_i$ and the spatial signature $\widetilde{\mathbf{h}}_i$. The length of $\mathbf{h}_i$ is $N = N_S \times N_R$. The signal model is then given by

$$\mathbf{Y} = \sum_{i=1}^{K} \sqrt{\rho}\mathbf{h}_i\mathbf{x}_i^T + \mathbf{Z} = \sqrt{\rho}\mathbf{H}\mathbf{X} + \mathbf{Z}. \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_K]^T$, $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]$. Note that $\mathbf{h}_i\mathbf{x}_i^T$ consists of $N = N_S \times N_R$ copies of $\mathbf{x}_i^T$.

When the number of antennas N_R is not overwhelmingly small compared to the number of users K, one may opt to set $N_S = 1$, i.e., removing the spreading operation from the MA system. Then the system becomes spatial division MA (SDMA), where only the spatial signatures are exploited to support K users.

[3] This paper focus on presenting LCMA with block fading model, albeit it can be extended to other types of fading models [24].

[4] Receiver-side synchronization can be realized by allowing each user to adjust its kick-off time based on the estimated path delay in synchronization signal block (SSB) broadcasted. The remaining asynchrony are incorporated in the channel coefficients of OFDM sub-carriers.

In this paper we present with a real-valued model. A complex-valued model can be represented by a real-valued model of doubled dimension as treated in [25], [26].

Remark 1 (Problem Statement in Brief): Given the system model above, the problem can be stated as: how to design a transceiver architecture and processing algorithms such that the MA system can support a high number of users whose messages can be reliably decoded in a cost effective manner.

The system model can be slightly modified to depict *grant-free* random-access (GFRA). Denote by $\mathcal{K}_{active} \subset \{1, \cdots, K\}$ a random set of indices of the *active users*. The signal model is given by

$$\mathbf{Y} = \sum_{i \in \mathcal{K}} \sqrt{\rho} \mathbf{h}_i \mathbf{x}_i^T + \mathbf{Z} \tag{3}$$

where $\mathbf{h}_i$ denotes the aggregation of the channel coefficients and the signature for random access, such as that in contention resolution diversity slotted aloha (CRDSA) or coded slotted aloha (SA) [12].

In GFRA, whenever new packets arrive, active users just send their signals out subject to slot-synchronization as in SA. Since BS is not informed of $\mathcal{K}_{active}$, i.e. which users are active, it needs to conduct active-user identification [10]. This is in contrast to the grant-based model where the K users are exactly known. Various coded SA (CSA) techniques are also suggested for resolving the contention due to random access [11], [12].

III. Preliminaries of Lattice Codes and Ring Coded Modulation

This section presents the coding and modulation upon which LCMA is built. For readers whose expertise is not information theory and coding, it would be easier to begin with RCM in Section III. B.

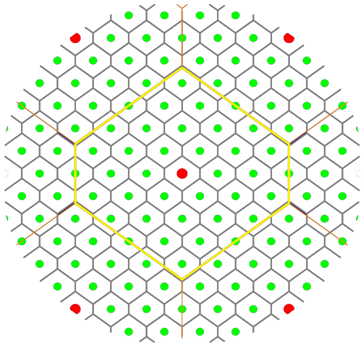


Fig. 1. Illustration of a generic nested lattice code. The green dots denote the points in the fine lattice. The red dots denote the points in the coarse lattice, whose density is smaller than that of the fine lattice points. The fundamental Voronoi regions of the fine and coarse lattices are also drawn. The fine lattice points that are within the Voronoi region of the coarse lattice are used as codewords for delivering messages.

A. Lattices and Lattice Codes

In comparison to off-the-shelf channel codes and widely known digital modulation schemes, lattice codes enjoy a more general mathematical representation for code construction, as well as properties that favour multi-user communication. A n-dimension lattice Λ is defined by a set of real-valued basis vectors $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_k$. All integer combinations of these basis vectors yield all the lattice points, given by

$$\Lambda \triangleq \left\{ \boldsymbol{\lambda} = \sum_{i=1}^{k} \mathbf{g}_i w_i, w_i \in \mathbb{Z}, i = 1, \cdots, k \right\}. \tag{4}$$

Since $w_i \in \mathbb{Z}$, there are infinite number of points in Λ. The basic operations associate with Λ are:

1) The *nearest neighbor vector-quantizer* w.r.t. Λ is defined as

$$Q_\Lambda(\mathbf{x}) \triangleq \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|\mathbf{x} - \boldsymbol{\lambda}\|, \tag{5}$$

which finds in Λ the lattice point that is closed to $\mathbf{x}$.

2) The *Voronoi region* of Λ is defined as

$$\mathcal{V}_\Lambda \triangleq \{\mathbf{x} \in \mathbb{R}^n, Q_\Lambda(\mathbf{x}) = \mathbf{0}\}, \tag{6}$$

which contains all real-valued vectors that are closer to $\mathbf{0}$ than to other lattice points in Λ. See Fig. 1.

3) The *modulo-lattice operation* is defined as

$$\mathbf{x} \bmod \Lambda = \mathbf{x} - Q_\Lambda(\mathbf{x}) \in \mathcal{V}_\Lambda, \tag{7}$$

which computes the difference between $\mathbf{x}$ and its nearest neighbor in the lattice Λ.

Let a subset of the lattice points in Λ be denoted by Λ_S, $\Lambda_S \subset \Lambda$, referred to as a “coarse lattice”. The original lattice Λ is referred to as the “fine lattice”. It is said that Λ_S is nested in Λ. Note that the points in Λ is denser than those in Λ_S, and the Voronoi region of Λ_S is larger than that of Λ, see Fig. 1.

In a nested lattice code for communication, those fine lattice points that are within the Voronoi region of Λ_S is used as codewords, each is associated with a message. The *codebook* is given by $\mathcal{C} = \{\mathbf{c} \in \Lambda \cap \mathcal{V}_{\Lambda_S}\}$, with codebook size of $|\mathcal{C}| = |\Lambda \cap \mathcal{V}_{\Lambda_S}|$ and information rate of $\frac{1}{2}\log_2 |\mathcal{C}|$ bits per symbol. During the online transmission, for a specific message sequence $\mathbf{b}$, the associated codeword $\mathbf{c} \in \Lambda \cap \mathcal{V}_{\Lambda_S}$ is picked with mapping $\mathbf{c} = \phi(\mathbf{b})$. A random dithering vector $\mathbf{d}$, which is uniformly distributed in the Voronoi region Λ_S, is added to the codeword $\mathbf{c}$. A modulo-lattice operation is then performed to ensure that the resultant vector $\mathbf{c} + \mathbf{d}$ is within $\mathcal{V}_{\Lambda_S}$. The transmitted signal sequence is

$$\mathbf{x} = (\mathbf{c} + \mathbf{d}) \bmod \Lambda_S \tag{8}$$

which is uniformly distributed in $\mathcal{V}_{\Lambda_S}$ [27], [28]. With the “*Encrypto Lemma*” [29], the average signal energy is given by $\frac{1}{n}E\left[\|\mathbf{x}\|^2\right] \le P_{\Lambda_S}$ where P_{Λ_S} is the second moment per-dimension of $\mathcal{V}_{\Lambda_S}$.

For AWGN channel, the received signal is $\mathbf{y} = \mathbf{x} + \mathbf{z}$. After the reverse of the dithering operation

$$\mathbf{y}' = (\mathbf{y} - \mathbf{d}) \bmod \Lambda_S, \tag{9}$$

the receiver finds in the fine lattice Λ the point that is closest to $\mathbf{y}'$, that is,

$$\widehat{\mathbf{c}} = \arg\min_{\lambda\in\Lambda} \|\mathbf{y}' - \lambda\| \, . \tag{10}$$

Then the decision on the message $\widehat{\mathbf{b}}$ is obtained from $\phi^{-1}\left(\widehat{\mathbf{c}}\right)$. It is proved that there exists nested lattices $\Lambda_S \subset \Lambda$of rate $R < \max\left[\frac{1}{2}\log_2\left(\rho\right), 0\right]$ such that $\Pr\left(\widehat{\mathbf{b}} \neq \mathbf{b}\right)$ tends to zero as $n \to \infty$. Moreover, with minimum mean sequare error (MMSE) scaling, the rate $R < \max\frac{1}{2}\log_2\left(1+\rho\right)$ is achievable [29].

B. Ring Coded Modulation - A Simplified Lattice Code

The aforementioned lattices and nested lattice codes are conceptual notions, as 1) they require the n-dimension lattice quantization and 2) they are based on the existence of lattices and lattice chains in theory, without given any clue on the construction that can be practically implemented. There are several existing works on practical aspects of nested lattice codes, such as low-density lattice codes in [30]. However, low-density lattice codes are still overly complicated, and its efficient encoding remains largely unsolved. In this paper, we utilize a simplified lattice code, namely q-ary ring coded modulation (RCM) with q-PAM signaling, as the underlying coded-modulation for LCMA.

1) Encoding: Let $\mathbf{b} = \left[b\left[1\right], \cdots, b\left[k\right]\right]^T$ be a column vector denoting a q-ary message sequence[5]. Each entry of $\mathbf{b}$ belongs to an integer ring $\mathbb{Z}_q \triangleq \{0, \cdots, q-1\}$. For a prime q, $\mathbb{Z}_q$ becomes a Galois field denoted by GF(q). For a non-prime q, e.g., $q = 2^m, m = 1, 2, \cdots$, the addition and multiplication rules of $\mathbb{Z}_{2^m}$ are different from those of GF(2^m). In this paper we are primarily interested in $q = 2^m, m = 1, 2, \cdots$, widely used in practice. For presentation purpose, q and 2^m may be used interchangeably.

A q-ary linear code with generator matrix $\mathbf{G}$ is employed to encode $\mathbf{b}$, given by

$$\mathbf{c} = \mathrm{mod}\left(\mathbf{G}\mathbf{b}{,}q\right) = \mathbf{G}\otimes_q\mathbf{b} \tag{11}$$

where "$\otimes_q$" represents the operation of matrix multiplication modulo-q. The generator matrix $\mathbf{G}$ is of size n-by-k and $\mathbf{c} \in \mathbb{Z}_q^n$. Let $\mathcal{C}^n$ denote the codebook which collects all q^k codewords w.r.t. (11).

A random vector $\mathbf{d} \in \mathbb{Z}_q^n$ may be generated and added on $\mathbf{c}$ for the purpose of random permutation. For conciseness, the details are omitted and can be found in [31], [32]. Each entry of $\mathbf{c}$ is *one-to-one* mapped to a symbol that belongs to a constellation of q points. For q-PAM constellation, the mapping is

$$\mathbf{x} = \delta\left(\mathbf{c}\right) = \frac{1}{\gamma}\left(\mathbf{c} - \frac{q-1}{2}\right) \in \frac{1}{\gamma}\left\{\frac{1-q}{2}, \cdots, \frac{q-1}{2}\right\}^n,$$

implemented symbol-wisely. Here γ is a normalization factor to ensure unit average symbol energy. The rate of RCM is $R = \frac{k}{n}\log_2 q$ bits/symbol. For a complex-valued model, two independent streams of q-level

[5] The conversion from a binary message sequence to a q-ary message sequence is straightforward.

RCM, one for the inphase part and the other for the quadrature part, form a RCM with q^2-QAM signaling. When $q = 2$, RCM reduces to conventional binary channel coding with BPSK (or QPSK) signaling.

Remark 2: [*RCM versus conventional coded-modulation*] The RCM differs from conventional *binary coding-oriented* schemes such as bit-interleaved coded-modulation (BICM), trellis coded-modulation (TCM) and multi-level coding with superposition coded-modulation (SCM). In those schemes, binary coded sequence $\mathbf{c}$ is de-multiplexed into $m = \log_2 q$ streams $\mathbf{c}^{(1)}, \cdots, \mathbf{c}^{(m)}$. Then, a *many-to-one* mapping given by $\mathbf{x} = \delta' \left(\mathbf{c}^{(1)}, \cdots, \mathbf{c}^{(m)}\right)$ is employed, e.g. the Grey mapping used for BICM.

2) Integer Additive Property: We next present the key property of RCM to be exploited in LCMA.

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

$$\sum_{i=1}^{K} a_i \mathbf{c}_i, q \in \widetilde{\mathcal{C}}^n \tag{12}$$

for any integer-valued coefficients $[a_1, \cdots, a_K]$, where $\widetilde{\mathcal{C}}^n$=$\mathcal{C}^n + q\mathbb{Z}^n$ denotes the *extended codebook* by replicating the codewords in the base code $\mathcal{C}^n$ over the infinite integer field $\mathbb{Z}^n$. Also,

$$\mathrm{mod}\left(\sum_{i=1}^{K} a_i \mathbf{c}_i, q\right) \in \mathcal{C}^n. \tag{13}$$

That is, the integer-sum of K codewords modulo-q remains as a valid codeword.

Note that this property does not hold in conventional schemes such as BICM, TCM and SCM.

Remark 3: [*Rings versus Galois Fields*] Most existing works on lattice codes and modulation codes focused on prime q where GF(q) and $\mathbb{Z}_q$ are equivalent. The integer additive property holds therein. In practical systems utilizing q-PAM (or q^2-QAM) signaling, non-prime q of values $q = 2^m$ is required. In this case, the integer additive property does not hold for GF(2^m). To see this, recall that GF(2^m) is an extension field of GF(2), which has elements $\left\{0, 1, \beta, \beta^2, \cdots \beta^{2^m-2}\right\}$ [33]. The additive rule w.r.t. these elements is determined based on the primitive element of the polynomials, which is different from the additive rule of integers as in $\mathbb{Z}_{2^m}$. Therefore, to enable the integer additive property for 2^m-PAM signaling, the utilization of ring codes over $\mathbb{Z}_{2^m}$ is indispensable.

C. Relation between RCM and Lattice Codes

The presented RCM is a simplified version of nested lattice codes. Its fine lattice Λ is given by the extended codebook $\widetilde{\mathcal{C}}^n$ with base code given by $\mathbf{c} = \mathbf{G} \otimes_q \mathbf{b}$, i.e.,

$$\Lambda = \left\{ \boldsymbol{\lambda} : \boldsymbol{\lambda} = \mathbf{c} - \frac{q-1}{2} + q\mathbb{Z}^n, \forall \mathbf{c} \in \mathcal{C}^n \right\}. \tag{14}$$

We will use “extended codebook” and “lattice” interchangeably in this paper. The coarse lattice Λ_S is

$$\Lambda_S = q\mathbb{Z}^n. \tag{15}$$

The modulo-Λ_S operation of the shaping lattice is simplified into a one-dimensional symbol-by-symbol modulo-q operation. The above is also referred to as “Construction A” of lattice codes [25].

The presented RCM yields 2^m-PAM signaling, which is in line with a wide range of practical systems. Compared to Gaussian signaling that is of interest in information theory, 2^m-PAM signaling enjoys easier treatment and low peak-to-average power ratio (PAPR), which are of high preference in practical uplink systems. This paper will devote no efforts to shaping (to achieve the shaping gain of at most 1.53 dB).

In current uplink MA systems, power efficiency rather than spectral efficiency may be of a higher priority, thus BPSK (or QPSK) signaling is primarily used. In such a case, RCM boils down to conventional binary coding with BPSK(or QPSK). Yet, even for BPSK, the processing of binary coded LCMA is over the extended codebook (the lattice) and exploited the integer additive property. This is strikingly different from conventional NOMA schemes that operate over the soft information of the binary digits only. For future MA systems, it is envisaged that high spectral efficiency may also be required for the uplink, then RCM with $q > 2$ is required in LCMA in general. For downlink systems featuring adaptive coding and modulation, RCM with $q > 2$ is indispensable to LCMA. This will be reported in a separate paper [34].

IV. LATTICE-CODE BASED MULTIPLE ACCESS (LCMA)

This section presents an uplink LCMA system. Following the convention in studying uplink MA, we consider an open-loop system where there is no feedback link to the transmitter to deliver the CSI or index of adaptive coding and modulation (ACM). Each user transmits at a target (symmetric) information rate R_0.

A. *Transmitter Architecture*

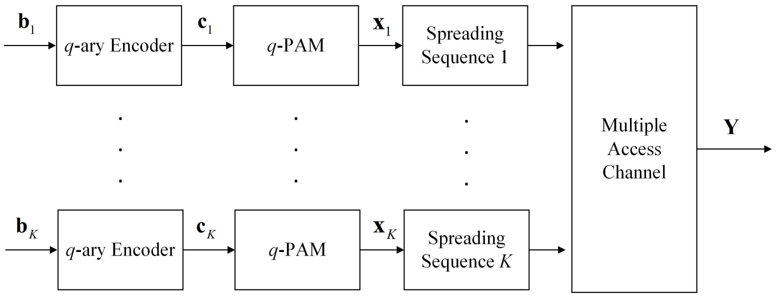


Fig. 2. Block diagram of the transmitters of a K-user LCMA system. All users utilize the same 2^m-ary RCM. No interleaver is implemented.

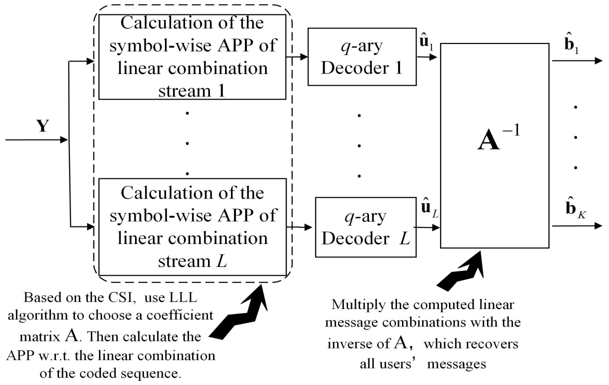
The transmitter architecture is depicted in Fig. 2. K users encode their messages with the same[6] 2^m-ary code as in (11). Each user's encoded digits are one-to-one mapped to 2^m-PAM symbols as in (12), yielding its symbol-level signal sequence $\mathbf{x}_i^T$. The symbol-level signal of user i is multiplied with its designated spreading-signature sequence $\mathbf{s}_i$, yielding the chip-level signal. Then all users' signals are transmitted simultaneously. The signal model was given in (2), repeated below

$$\mathbf{Y} = \sum_{i=1}^{K} \sqrt{\rho}\mathbf{h}_i\mathbf{x}_i^T + \mathbf{Z} =\sqrt{\rho}\mathbf{H}\mathbf{X} + \mathbf{Z}.$$

Comparing to existing NOMA schemes, the distinguishing features of LCMA transmitter involve: 2^m-ary lattice code/ring code, a one-to-one 2^m-PAM mapping, and no symbol-level or chip-level interleaver. For LCMA with $q = 2$ and BPSK (or QPSK) signaling, any conventional binary codes can be utilized.

Remark 4: The spreading module may be removed if the number of receive antenna N_R is sufficiently large to support K users. For scenarios where the spreading module is utilized, this paper confines the discussion to that the entries of $\mathbf{s}_i$ are obtained from $\{0, +1, -1\}$, while $\mathbf{s}_i$ is subject to a power normalization, i.e., $\|\mathbf{s}_i\|^2 = 1$. This is in line with the spreading sequence structure of the SCMA system with the real-valued model, and is in line with that of the IDMA system if the portion of 0-entries in $\mathbf{s}_i$ is set to zero. Any existing spreading signature sequences of SCMA can be used in LCMA. For high system loads that are out of the capability of SCMA, a pragmatic method for generating the spreading sequences is presented in Appendix, Algorithm 2. We note that the generation of $\mathbf{s}_i$ is carried out offline.

B. Receiver Architecture



Based on the CSI, use LLL algorithm to choose a coefficient matrix A. Then calculate the APP w.r.t. the linear combination of the coded sequence.

Multiply the computed linear message combinations with the inverse of A, which recovers all users' messages

Fig. 3. Block diagram of the receiver of a K-user LCMA system with a singel-stage parallel processing.

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

The receiver is set to compute L streams of linear combinations of the K users' messages over the integer ring $\mathbb{Z}_q$, as shown in Fig. 3. These L streams of "*linear message combinations*" are defined as

$$\mathbf{u}_l^T \triangleq \mathrm{mod}\left(\sum_{i=1}^{K} a_{l,i}\mathbf{b}_i^T, q\right) = \mathbf{a}_l^T \otimes_q \mathbf{B}, l = 1, \cdots, L, \tag{16}$$

where $\mathbf{a}_l^T = [a_{l,1}, \cdots, a_{l,K}]$ with entries in $\mathbb{Z}_q$ denotes the *coefficient vector* w.r.t. the l-th stream, $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_K]^T$ stacks up all users' message sequences. Let $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_L]^T$, and let $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_K]^T$ be referred to as the *coefficient matrix*, which is required to have a unique inverse $\mathbf{A}^{-1}$ in $\mathbb{Z}_q$. Then,

$$\mathbf{U} = \mathbf{A} \otimes_q \mathbf{B}. \tag{17}$$

For the MA setting with a common receiver considered in this paper, $L = K$. If all linear message combinations, i.e., $\mathbf{U}$, are reliably computed, all users' messages are successfully recovered by multiplying $\mathbf{U}$ with $\mathbf{A}^{-1}$ over $\mathbb{Z}_q$. For the clarity of physical meanings, we use L (rather than K) to denote the total number of linear message combination streams in the sequel[7].

There are two core tasks for the LCMA receiver: 1) with $\mathbf{Y}$, computes the L linear message combinations $\mathbf{u}_1, \cdots, \mathbf{u}_L$. 2) Identifies $\mathbf{A}_{opt}$ (based on the CSI of $\mathbf{H}$) that minimizes the outage probability or FER of a practical coded system. Now let us concentrate on the former task for a given coefficient matrix $\mathbf{A}$.

1) Optimal Computation Rule: The optimal rule jointly computes $p(\mathbf{U}|\mathbf{Y})$, i.e., the *a posteriori probability (APP) matrix* w.r.t. all L streams of linear message combinations. This is a formidable task even for K of a small dimension, and simplified treatments are required.

2) Parallel Computation Rule: A suboptimal solution to $p(\mathbf{U}|\mathbf{Y})$ is to parallely compute the *APP sequences* of the L streams of linear message combinations, given by

$$p(\mathbf{u}_l|\mathbf{Y}), l \;=\; 1, \cdots, L. \tag{18}$$

Such parallel processing decouples the computation of the L streams of the linear message combinations, and can be implemented in a single-stage. This paper develops two practical methods for implementing the computation in (18), in the next two subsections respectively.

C. The Linear PNC based Approach

Let $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_K]^T$ stacks up all users' coded sequences generated by the q-ary ring code. Define

$$\mathbf{v}_l^T \triangleq \mathrm{mod}\left(\sum_{i=1}^{K} a_{l,i}\mathbf{c}_i^T, q\right) = \mathbf{a}_l^T \otimes_q \mathbf{C} \tag{19}$$

[7] For more general setups such as distributed MIMO, where distributed units (DUs) are connected to a central unit (CU), L of each DU may be smaller than K, as long as all DUs can provide the CU with sufficient numbers of linear message combinations [25].

as the l-th "*linear coded-sequence combination*".

Property 2: With the generator matrix $\mathbf{G}$ in (11) and by applying Property 1, we have

$$\begin{aligned}\mathbf{v}_l &= \bmod\left(\sum_{i=1}^{K} a_{l,i}\mathbf{G}\otimes_q \mathbf{b}_i, q\right) = \mathbf{G}\otimes_q \bmod\left(\sum_{i=1}^{K} a_{l,i}\mathbf{b}_i, q\right)\\ &= \mathbf{G}\otimes_q \mathbf{u}_l. \end{aligned} \tag{20}$$

That is, a linear coded-sequence combination $\mathbf{v}_l$ and a linear message combination $\mathbf{u}_l$ are also related by the multiplication of $\mathbf{G}$ modulo-q.

Property 2 allows for the computation of (18) by implementing: 1) calculating the symbol-wise APPs of $\mathbf{v}_l$ over the extended constellation, to be detailed momentarily; 2) forward the resultant APP sequence to a single-user decoder that exploits the structure of $\mathbf{G}$ to compute $\mathbf{u}_l$. We note that such treatment is impossible for non-lattice code based MA schemes where Properties 1 and 2 do not hold.

1) Symbol-wise APP based on Algebraic Binning: For a given coefficient vector $\mathbf{a}_l$, a specific "*algebraic binning*" [25] structure is formed, explained below. Recall that each user's symbol belongs to a q-PAM constellation. The superposition of the K users' symbols results in an extended constellation with q^K candidates. The candidates are partitioned in to q sets, also called "*bins*". Those candidates whose underlying value of the linear combination are identical (or non-identical), belong to the same (or different) bin. The value of the linear combination, denoted by ω , is referred to as the "bin-index". Such algebraic binning does not apply in non-lattice code based MA schemes. The receiver computes the probabilities of the bin-indices.

Let $v_l\left[t\right]$ and $\mathbf{y}\left[t\right]$ denote the t-th column of $\mathbf{v}_l$ and $\mathbf{Y}$, respectively. The receiver calculates the symbol-wise APPs of the linear coded-sequence combinations $p\left(v_l\left[t\right]\left|\mathbf{y}\left[t\right]\right.\right)$, i.e., the bin-indices. This is implemented in parallel for $l = 1, \cdots, L$. Using the Baye's rule, we obtain

$$p\left(v_l\left[t\right] = \omega|\mathbf{y}\left[t\right]\right) = \sum_{\substack{x_1[t],\cdots,x_K[t]:\\ \mathbf{a}_l^T\otimes_q \mathbf{c}[t]=\omega}} p\left(\mathbf{y}\left[t\right]|\sum_{i=1}^{K}\sqrt{\rho}\mathbf{h}_i x_i\left[t\right]\right), \omega = 0, \cdots, q-1. \tag{21}$$

It equals to the sum of the likelihood functions of the q candidates in the bin with index ω. Here $c_i\left[t\right] = x_i\left[t\right] + \frac{q-1}{2}$ in (12) is utilized. Note that the calculation is over an N-dimension receive signal space, which generalizes the LPNC approach in [23] and is in contrast to integer-forcing (IF) [18].

2) List Sphere Decoding: The order of complexity is up to $O\left(q^K\right)$ if exhaustive search is adopted. To make the complexity manageable, we resort to a list sphere decoding (LSD). Given the received signal vector

$\mathbf{y}[t]$, LSD algorithm is used to form a candidates list $\mathcal{L}$, containing the candidates that are are closed to $\mathbf{y}[t]$ in Euclidean distance, detailed in Appendix, Algorithm 1. Then, (21) is revised into

$$\begin{aligned} p\left(v_l\left[t\right]=\omega|\mathbf{y}\left[t\right]\right) &= \sum_{x_1[t],\cdots,x_K[t]\in\mathcal{L}:v_l[t]=\omega} p\left(\mathbf{y}\left[t\right]|\sum_{i=1}^{K}\sqrt{\rho}\mathbf{h}_i x_i\left[t\right]\right) \\ &= \frac{1}{\eta}\sum_{x_1[t],\cdots,x_K[t]\in\mathcal{L}:\mathbf{a}_l^T\otimes_q\mathbf{c}[t]=\omega} \exp\left(-\frac{\left\|\mathbf{y}[t]-\sum_{i=1}^{K}\sqrt{\rho}\mathbf{h}_i x_i[t]\right\|^2}{2}\right). \end{aligned} \tag{22}$$

The complexity now becomes $O\left(|\mathcal{L}|\right)$. With LSD algorithm, it was suggested the complexity can be made polynomial to the number of users K and modulation size. Note that one is free to adjust the list size $|\mathcal{L}|$ for a suitable performance-complexity tradeoff. In practice, $|\mathcal{L}|=40$ to 50 yields competitive performance.

3) Single-user Decoding: The resultant APP sequence of each stream obtained in (22) is forwarded to a q-ary ring code decoder. Owing to Property 2, the standard single-user decoding operation for the point-to-point channel applies in the decoding of the linear message combinations. The decoding w.r.t. the lth stream yields the hard decision on the linear message combination, denoted by $\widehat{\mathbf{u}}_l$, $l=1,\cdots,L$. Exactly L single-user decoding operations are required, which are performed parallelly in a single-stage.

For $q=2$, off-the-shelf decoding methods apply. For $q=2^m, m=1,2,\cdots$, an iterative q-ary BP algorithm is utilized for q-ary LDPC or irregular repeat-accumulate (IRA) ring codes [35].

4) Achievable Rate: Here we characterize the achievable rate of the LCMA scheme with a full list size of $|\mathcal{L}|=q^K$. Let X_i and V_l denote the random variables (R.V.s) of user i's transmitted signal and the l-th linear combination, respectively. Let $\underline{Y}$ denote the R.V. of the received signal vector of dimension N.

Theorem 1: For a given coefficient matrix $\mathbf{A}$ of a unique inverse in $\mathbb{Z}_q$, an achievable rate region of K-user LCMA with parallel processing is characterized by

$$R_i^{(\mathbf{A})} \leq \log_2 q - \max_l\{\varphi(a_{l,i})H(V_l|\underline{Y})\}$$

for $i=1,\cdots,K$, where

$$\varphi(a)=\begin{cases} 0, & a=0 \\ 1, & a\neq 0 \end{cases},$$

and $H(\bullet)$ denotes the entropy function.

Proof. From Property 2, the probability of $V_l=\omega$ is given by

$$p(V_l=\omega)=\sum_{a_{l1}j_1\oplus\cdots\oplus a_{l,K}j_K=\omega}\prod_{i=1}^{K}p(C_i=j_i)=\frac{q^{K-1}}{q^K}=\frac{1}{q}. \tag{23}$$

Therefore, $H(V_l)=H(X_i)=\log(q)$. For given $\mathbf{A}$, the achievable computation rate of the l-th linear combination is given by [36], [37]

$$R_{l,comp}^{(\mathbf{A})}\leq I(\underline{Y};V_l),\ l=1,\cdots,L,$$

where $I(\bullet)$ denotes the mutual information function. If $a_{l,i} \neq 0$, the l-th linear combination V_l includes the message of user i, which implies that the rate of user i should be no greater than the achievable computation rate of the l-th linear combination. Thus, the rate of user i satisfies

$$\begin{aligned} R_i^{(\mathbf{A})} &\leq \min_{l:a_{l,i}\neq 0}\{R_{l,comp}^{(\mathbf{A})}\} = \min_l\{H\left(V_l\right) - H\left(V_l|\underline{Y}\right)|a_{l,i}\neq 0\} \\ &\overset{(a)}{=} H(V_l) - \max_l\{H\left(V_l|\underline{Y}\right)|a_{l,i}\neq 0\} = \log_2 q - \max_l\{\varphi(a_{l,i})H(V_l|\underline{Y})\} \end{aligned}$$

where step (a) follows from the fact that $V_1, \cdots, V_L$ are independently and uniformly distributed. ■

Corollary 1: A lower bound of the achievable symmetric rate of LCMA is given by

$$R_{sym}^{(\mathbf{A})} < \min_l\{\log_2 q - \varphi(a_{l,i})H(V_l|\underline{Y})\}, \tag{24}$$

which is simply the smallest of the achievable computation rates of all linear message combinations.

In conventional MA, $\mathbf{A}$ is set to be $\mathbf{I}$. With the lattice/ring codes, algebraic binning in LCMA enables the relaxation to computing linear message combinations with any invertible $\mathbf{A}$, resulting $R_{sym}^{(\mathbf{A})} \geq R_{sym}^{(\mathbf{I})}$.

D. The Linear Filtering based Approach

Unlike the LPNC approach, the linear filtering based approach is set to transform the received N-dimension signal into L streams of single-dimension signal streams. Then, one linear message combination is computed from one of these single-dimension streams.

1) Linear Filtering: Let $\mathbf{W}$ of size L-by-N denote a linear filter matrix, with real-valued entries. Let $\mathbf{w}_l^T$ denote the lth row of $\mathbf{W}$,$l = 1, \cdots, L$. We denote the *filtered version of the received signal* by

$$\widetilde{\mathbf{y}}_l^T = \mathbf{w}_l^T\mathbf{Y} = \sqrt{\rho}\mathbf{w}_l^T\mathbf{H}\mathbf{X} + \left(\mathbf{z}_l'\right)^T. \tag{25}$$

Let $\psi_l^T = \mathbf{w}_l^T\mathbf{H}$. For a given coefficient vector $\mathbf{a}_l^T$, we re-arrange the t-th symbol of $\mathbf{r}_l^T, t = 1, \cdots, n$ into

$$\widetilde{y}_l[t] = \sum_{i:a_{l,i}\neq 0}\sqrt{\rho}\psi_{l,i}x_i[t] + \sum_{i:a_{l,i}=0}\sqrt{\rho}\psi_{l,i}x_i[t] + z_l'[t]. \tag{26}$$

Here, we deem the term $\sum_{i:a_{l,i}\neq 0}\sqrt{\rho}\psi_{l,i}x_i[t]$ as the *useful signal part*, which contains the signals of the users whose corresponding coefficients are non-zero. Meanwhile, we deem the term $\sum_{i:a_{l,i}=0}\sqrt{\rho}\psi_{l,i}x_i[t]$ as the *interference*, which contains the signals of the users whose corresponding coefficients are zero. These are irrelevant in computing the linear message combination. We treat the term $\sum_{i:a_{l,i}=0}\sqrt{\rho}\psi_{l,i}x_i[t] + z_l'[t]$ as the *interference-plus-noise*. For moderate-to-large K, it is empirically found that the cardinality of the set

$\{i : a_{l,i} = 0\}$ is usually not very small. According to the central limit theorem, the interference-plus-noise term is regarded as following a Gaussian distribution with zero mean and variance

$$\widetilde{\sigma}_l^2 = \rho \sum_{i:a_{l,i}=0} \psi_{l,i}^2 + 1. \tag{27}$$

With the above arrangement, the symbol-wise APP is calculated by

$$\begin{aligned} p\left(v_l\left[t\right] = \omega | \widetilde{y}_l[t]\right) &= \sum_{x_i[t], i:a_{l,i}=0: v_l[t]=\omega} p\left(\widetilde{y}_l\left[t\right] | \sum_{i:\alpha_{l,i}\neq 0} \sqrt{\rho}\psi_{l,i} x_i\left[t\right]\right) \\ &= \frac{1}{\eta} \sum_{x_i[t], i:\alpha_{l,i}=0: \mathbf{a}_l^T \otimes_q \mathbf{c}[t]=\omega} \exp(-\frac{\left|\widetilde{y}_l[t] - \sum_{i:a_{l,i}\neq 0} \sqrt{\rho}\psi_{l,i} x_i[t]\right|^2}{2\widetilde{\sigma}_l^2}), \end{aligned} \tag{28}$$

evaluated over the one-dimension lattice. The resultant APP sequence of each stream is forwarded to a q-ary ring code decoder, with the same implementation as that in the LPNC based approach.

We note that a well-designed linear filter $\mathbf{W}$ features that the aggregation of the magnitudes of $\psi_{l,i}, i : \alpha_{l,i} = 0$ is much less than that of $\psi_{l,i}, i : \alpha_{l,i} \neq 0$. For example, in the exact IF [18], $\mathbf{W}$ is chosen such that $\psi_{l,i} = 0$ for $i : \alpha_{l,i} \neq 0$. Comparing (22) and (28), it is clear that the LPNC approach directly calculates the symbol-wise APP in the N-dimension signals space, while linear filtering based approach first projects the N-dimension signal, forming a single-dimension signal. Then it calculates the symbol-wise APP. The first approach is subject to a loss due to the reduced list size in LSD, while the second approach is subject to a loss due to the projection operation, relative to the exact optimal solution.

Remark 5: The idea with linear filtering was previously investigated in [18], [19] from an information-theoretic aspect. To the best of our knowledge, the algorithms for a practical coded system that we presented in the above was not reported in the literature. Moreover, our treatment applies to any choice of linear filter, not just confined to the exact IF and regularized IF in [18].

2) Achievable Rate: Let $\widetilde{Y}_l$ denote the R.V. of the l-th stream after the above linear filtering.

Corollary 2: For a given coefficient matrix $\mathbf{A}$, by applying Theorem 1,a lower bound of the achievable symmetric rate is given by

$$R_{sym}^{(\mathbf{A})} < \min_l \{\log_2 q - \varphi(a_{l,i}) H(V_l | \widetilde{Y}_l)\}. \tag{29}$$

Due to *data processing inequality* [38], $H(V_l|\widetilde{Y}_l) \geq H(V_l|\underline{Y})$, hence the symmetric rate of the linear filtering based approach is smaller than or equal to that of the LPNC based approach (with a full list size).

E. On the Optimal Choice of Coefficient Matrix $\mathbf{A}$

For q-PAM signaling in this paper, there is no close-form representation of the entropy. The involvement of the N dimension receive signal space makes the calculation of the multi-dimension entropy even more difficult. As such, the identification of the exact rate maximizing coefficient matrix $\mathbf{A}$ needs to evaluate the entropy for all possible coefficient vectors, which is prohibitive. This paper devotes no effort to find the exact optimal solution. Instead, we will utilize a suboptimal choice given by solving a "shortest lattice point" problem [18]. To make this article self-contained, it is briefly presented below.

Let the eigen-decomposition of the matrix $\rho\mathbf{H}^T\mathbf{H}+\mathbf{I}_K$ be written as

$$\rho\mathbf{H}^T\mathbf{H}+\mathbf{I}_K=\mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Psi}^T. \tag{30}$$

Then, $\widetilde{\mathbf{A}}$ with entries in $\mathbb{Z}$ is given by

$$\arg\min_{\widetilde{\mathbf{A}}}\max_{l}\left\|\mathbf{\Sigma}^{-1/2}\mathbf{\Psi}^T\widetilde{\mathbf{a}}_l\right\|^2 \tag{31}$$

s.t. rank$(\widetilde{\mathbf{A}})=L$. The solution to (31) can be efficiently implemented using an Lenstra-Lenstra-Lovasz (LLL) or Hermite-Korkine-Zolotareff (HKZ) algorithm [39]. The complexity of LLL algorithm, that will be used in the simulation section, is polynomial to the number of streams L. The resultant suboptimal coefficient matrix will be used in the numerical result section.

Remark 6: We note that LCMA is different from lattice-reduction based MIMO processing methods. To be specific, LCMA utilizes n-dimension lattice codes or RCM as the underly coding-modulation, where the lattice is characterized by the code generator matrix $\mathbf{G}$. Lattice-reduction based MIMO processing is dealing with the lattice generated by the N-by-K channel matrix $\mathbf{H}$.

Remark 7 (Hints on LCMA for Grant-free Random Access (GFRA)): LCMA can be applied to GFRA, which is known for its benefits for the massive access scenarios. The spreading sequence can be replaced by random replicas as in the coded slotted aloha (CSA) framework. At the receiver side, active user identification is performed using algorithms such as approximate message passing (AMP) or orthogonal AMP [10]. Then, the linear message combinations can be computed in each slots. Upon sufficient numbers of linear combinations are computed, a subset of the messages can be recovered. The details are not presented due to space limitation.

V. Multi-stage LCMA Receiver

Previously, we presented LCMA system with parallel processing, where the receiver computes $L=K$ linear message combinations in a *single-stage*. If some of these linear message combinations are not correctly computed, decoding errors are incurred. This approach generally suffers from a performance loss relative to the joint processing. It is demonstrated that the loss is not significant for moderate-to-low system loads, but becomes obvious for high system loads. Motivated by this, this section proposes a new *multi-stage* LCMA receiver, whose block diagram is shown in Fig. 4. The idea is briefly illustrated below:

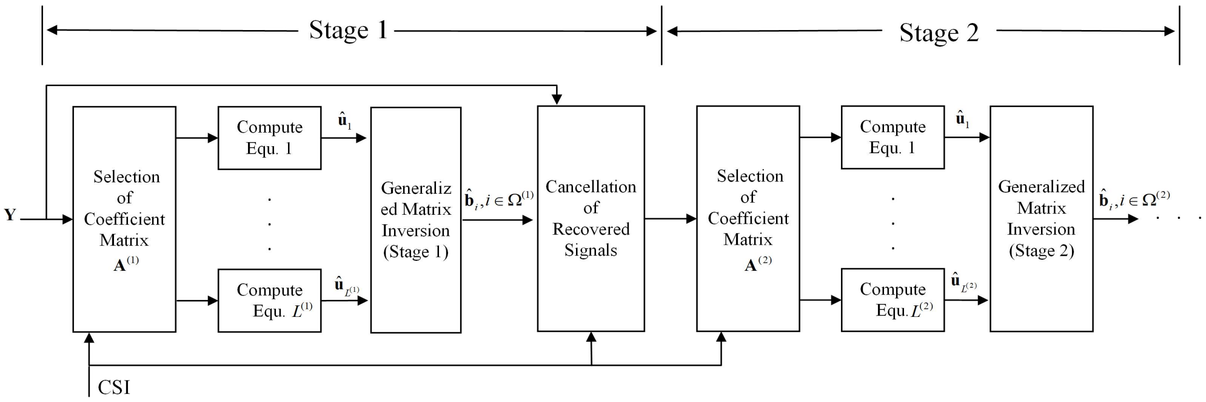

Stage 1
Stage 2
Y
Selection of Coefficient Matrix $\mathbf{A}^{(1)}$
Compute Equ. 1
$\hat{\mathbf{u}}_1$
Compute Equ. $L^{(1)}$
$\hat{\mathbf{u}}_{L^{(1)}}$
Generalized Matrix Inversion (Stage 1)
$\hat{\mathbf{b}}_i, i \in \Omega^{(1)}$
Cancellation of Recovered Signals
Selection of Coefficient Matrix $\mathbf{A}^{(2)}$
Compute Equ. 1
$\hat{\mathbf{u}}_1$
Compute Equ. $L^{(2)}$
$\hat{\mathbf{u}}_{L^{(2)}}$
Generalized Matrix Inversion (Stage 2)
$\hat{\mathbf{b}}_i, i \in \Omega^{(2)}$
CSI

Fig. 4. Block diagram of the multi-stage LCMA receiver.

Consider that $L_{(1)} < K$ linear message combinations are set to be computed in the first stage. Such task is "easier" than computing the entire K linear message combinations, e.g., the decoding SNR threshold becomes lower. We introduce *generalized matrix inversion* (GMI), which recovers a subset of all users' messages from the $L_{(1)}$ linear message combinations. Next, these messages are re-encoded and modulated, forming the signal sequences which are cancelled from the received signal $\mathbf{Y}$. Then, the original K-user MA problem reduces to a $K_{(1)}$-user MA problem with $K_{(1)} < K$, for processing in the next stage. The above process continues for a number of stages until no further improvement can be achieved.

A. Generalized Matrix Inversion

As the GMI operates over the messages sequences, it is safe to omit the position index for the clarity of presentation. That is, the t-th position of $\mathbf{b}_i^T = [b_i[1], \cdots, b_i[k]]$ is denoted by b_i in the sequel. All K users' messages are denoted by $\mathbf{b} = [b_1, \cdots, b_K]^T$. The $L_{(\tau)}$ linear message combinations computed in Stage τ are:

$$\mathbf{u}_{(\tau)}=\mathbf{A}_{(\tau)} \otimes_q \mathbf{b}. \tag{32}$$

Here we show how to extract a subset of messages in $\mathbf{b}$. Denote by $\mathbf{e}_i$ a K-by-1 unit vector whose i-th entry is 1 and the other entries are zero. If there exist $\mathbf{a}_i$ with entries in $\{0, \cdots, q-1\}$ such that

$$\mathbf{a}_i{}^T \otimes_q \mathbf{A}_{(\tau)} = \mathbf{e}_i^T, \tag{33}$$

then from (32) one can write:

$$b_i = \mathbf{e}_i^T \otimes_q \mathbf{b} = \boldsymbol{\alpha}_i{}^T \otimes_q \mathbf{A}_{(\tau)} \otimes_q \mathbf{b} = \mathbf{a}_i{}^T \otimes_q \mathbf{u}_{(\tau)}, \tag{34}$$

i.e., user i's message is recovered by multiplying $\mathbf{a}_i{}^T$ with $\mathbf{u}_{(\tau)}$. The problem at this point is to identify all $\mathbf{a}_i$ satisfying (33). To this end, the step-by-step procedures of GMI are given below:

Algorithm of GMI:

Step 1. Apply Gaussian elimination in GF(q) over $\mathbf{A}_{(\tau)}$. The row transformation yields

$$\mathbf{Q}_{\text{row}} \otimes_q \mathbf{A}_{(\tau)} = \begin{bmatrix} \mathbf{I} & \boldsymbol{\theta} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{35}$$

The column transformation is given by, given by

$$\mathbf{Q}_{\text{row}} \otimes_q \mathbf{A}_{(\tau)} \otimes_q \mathbf{Q}_{\text{col}}= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{36}$$

Step 2. Compute the $\{1\}$-inverse of $\mathbf{A}_{(\tau)}$ given by [40]

$$\mathbf{A}_{(\tau)}^{\{1\}} = \mathbf{Q}_{\text{col}} \otimes \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix} \otimes \mathbf{Q}_{\text{row}}. \tag{37}$$

Here, $\mathbf{I}$ is an identity matrix of the same size as that in (36), $\boldsymbol{\Psi}$ is a randomly chosen matrix in GF(q). The non-unique $\boldsymbol{\Psi}$ will not hinder us from identifying the $\boldsymbol{\alpha}$ vectors.

Step 3. Locate the rows of $\mathbf{A}_{(\tau)}^{\{1\}}\mathbf{A}_{(\tau)}$ that are unit vectors. Let $\Bbbk_{(\tau)}$ collect the indices of these rows.

Step 4. All vectors satisfying (33) are given by the i-th rows of $\mathbf{A}_{(\tau)}^{\{1\}}$, $i \in \Bbbk_{(\tau)}$. Then, by implementing (34), the users' messages with indices in $\Bbbk_{(\tau)}$ are recovered.

B. Signal-level Cancellation

At the end of the first stage, the successfully recovered messages of users, with indices in $\Bbbk_{(1)}$, are re-encoded and modulated, generating $\mathbf{x}_i^T, i \in \Bbbk_{(1)}$. They are cancelled from the original signal. The original K-user MA problem now reduces to a $K_{(1)}$-user MA problem with $K_{(1)} = \left|\Bbbk_{(1)}^c\right| < K$, for processing in the next stage. Note that the coefficient matrix is required to be selected in each stage of the processing. The general expression for the τ-th stage is[8]

$$\mathbf{Y}_{(\tau)} = \mathbf{Y}_{(\tau-1)} - \sum_{i \in \Bbbk_{(\tau)}} \mathbf{h}_i \sqrt{\rho} \mathbf{x}_i^T. \tag{38}$$

The proposed multi-stage LCMA receiver is different from successive IF (SIF) [41]. In SIF, the previously computed linear combination is used as a side-information in computing subsequent linear combinations, via direct cancellation in the n-dimension lattice. To date, there is no practical coding that can realize such lattice-level cancellation. In contrast, by introducing GMI, exact signal sequences is yielded for signal-level cancellation in multi-stage LCMA receiver. The optimized coefficient matrix $\mathbf{A}$ in SIF remains difficult to solve, while the LLL based algorithm applies to the multi-stage LCMA receiver.

[8] Note that there is a rearrangement of the user-indices at the end of each stage, which is not presented for a better readability.

TABLE I

THE ORDER OF COMPLEXITIES OF LCMA, IDMA AND SCMA SYSTEMS

	Detection	Decoding	Coefficient Identification	Interleaver&De-interleaver
LCMA	$O(K \cdot \lvert\mathcal{L}\rvert \cdot n)$	$O(q-1)n$	Between $O(K^3)$ and $O(K^4)$	not required
IDMA	$O(Q \cdot K \cdot log_2^q \cdot N_S \cdot n)$	$O(Q \cdot log_2^q \cdot n)$	not required	$O(2Q \cdot N_S \cdot n)$
SCMA	$O(Q \cdot K^2 \cdot log_2^q \cdot n)$	$O(Q \cdot log_2^q \cdot n)$	not required	$O(2Q \cdot n)$

C. On the Complexities of LCMA Receivers

The computation burden is primarily in 1) the channel-code decoding operations and 2) multi-user detector. For the former, it is clear that LCMA requires only K single-user decoding operations while IDMA/SCMA requires $Q \cdot K$ decoding operations, where the typical value of Q is between 4 to 10. For the latter, LCMA needs to compute K streams of APP-sequences, while IDMA/SCMA requires to compute $Q \cdot K$ streams of APP-sequences. In particular, if the linear filtering based approach is utilized, the per-symbol detection complexity of each stream amounts to q to the power of the number of non-zero positions of the coefficient vector. The additional complexity is not significant for small q and moderate system loads. Recall that LCMA requires to identify a coefficient matrix $\mathbf{A}$ for each block, imposing an extra overhead[9]. Since $\mathbf{A}$ is chosen only once per block, for a moderate-to-long block length (such as $k \geq 256$), the overhead is not significant. We emphasize that off-the-shelf binary codes in various standards can be directly used in LCMA for any load $\frac{K}{N}$. In contrast, as the load $\frac{K}{N}$ varies, IDMA and SCMA have to adopt different codes. Otherwise, the convergence of iterative detection and decoding may not be achieved, leading to overwhelmingly impaired performance or even failed functionality. The orders of complexity are shown in Table. I

It is interesting to note that, albeit a much superior performance of the multi-stage LCMA receiver, its complexity is not necessarily higher than that of the single-stage receiver. The multi-stage receiver involves serial processing, which leads to a higher processing delay, but requires a smaller memory size.

VI. NUMERICAL RESULTS

This section presents numerical results of LCMA. Following the convention of open-loop uplink MA system, the CSI is available to the receiver but not to the transmitters, and there is no feedback link for adaptive coding and modulation or resource allocation. Equal power among the users is applied. No attempts are made to adjust the power/code profile for optimizing the performance. For comparison purpose, we also present the performance of IDMA and SCMA. As open-loop uplink MA system is considered where rate allocation cannot be implemented, rate-splitting MA (RSMA) [9] is not included in the comparison.

[9] With the LLL algorithm, the complexity for this is about cubic to K.

TABLE II

SPREADING SEQUENCE OF LCMA K=10, N_S=4.

1	1	1	0	1	1	1	1	1	1
1	0	-1	-1	1	-1	1	1	1	-1
1	-1	1	-1	1	1	1	-1	0	-1
1	0	-1	0	-1	1	0	1	-1	-1

A. AWGN Multiple-access Channel

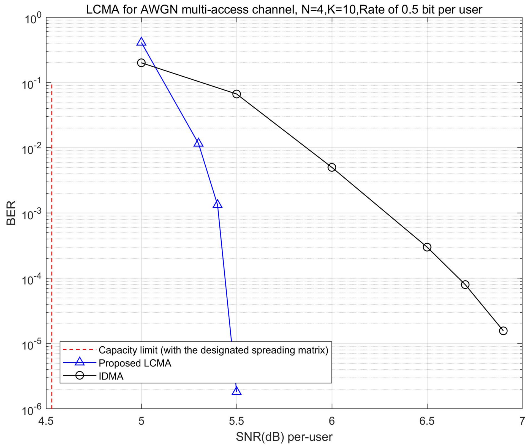


Fig. 5. BER of LCMA with $K = 10$ users and spreading sequence of length $N_S = 4$ in AWGN MA channel.BPSK and rate 1/2 IRA channel codes are used. The sum-rate is 5 bits per channel-use. Multi-stage LCMA receiver with the linear filtering based approach is used.

In the simulations, for the LPNC based approach, the list size of LSD is set to 50. For the linear filtering based approach, the filter matrix $\mathbf{W}$ under consideration is given by

$$\mathbf{W} = \widetilde{\mathbf{A}}\mathbf{H}^T \left(\vartheta \mathbf{H}\mathbf{H}^T + I_N\right)^{-1}$$

where $\widetilde{\mathbf{A}}$ has entries in $\mathbb{Z}$ and $\mathbf{A} = \mathrm{mod}\left(\widetilde{\mathbf{A}}, q\right)$. Here, ϑ is adjusted according to SNR and mutual information of q-PAM signaling, which is slightly different to that in specified [18].

Fig. 5 shows the BER of LCMA for AWGN MA channel where $K = 10$ and the spreading sequence length $N_S = 4$. The system load is $K/N_S = 250\%$. The spreading matrix of LCMA is given in Table II, obtained using Algorithm 2 given in Appendix. The capacity limit with the given spreading matrix is also shown, which provides an performance upper bound. For the purpose of comparison to the capacity limit, a channel code with long block-size is required. Here we use an irregular repeat accumulate (IRA) code of long block-size of 50000 as in [42]. Note that this code is optimized for the single-user AWGN channel, not for MA channel. At the BER of 10^{-5}, LCMA performs within 1 dB the capacity limit of the AWGN MA channel. This suggests that a good single-user channel code also works well in LCMA.

We also present the BER of a baseline IDMA system, with standard chip-level interleaving and iterative ESE detection [3]. We note that at this system load of 250%, the IDMA system with an off-the-shelf capacity approaching channel code does not converge, as the EXIT curves of the inner detector and outer decoder intersect at a low mutual information. Instead, a weak code such as a convolutional code with generator polynomials $[5, 7]_8$ yields the best performance of IDMA, which is shown in Fig. 5. LCMA exhibits a 1.4 dB performance advantage over IDMA with a sufficiently large number or receiver iterations ($Q = 15$). Note that LCMA does not involve receiver iteration, and approximately K single-user decoding operations are required. In contrast, IDMA involves QK single-user decoding operations.

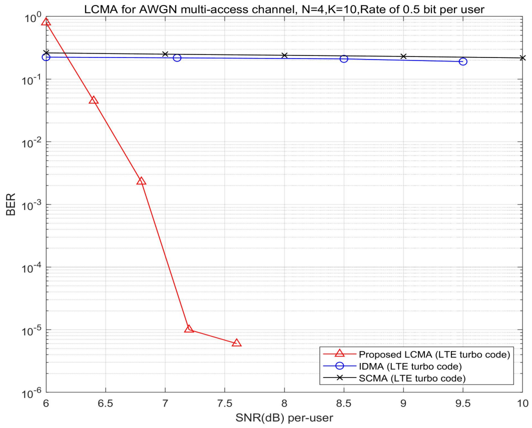


Fig. 6. BER performance of LCMA with $K = 10$ users and spreading sequence of length $N_S = 4$ in AWGN MA channel. BPSK and length-2048 LTE turbo code of rate 1/2 is used for all schemes. Multi-stage LCMA receiver with the linear filtering based approach is used.

Fig. 6 shows the BER performance of LCMA with a length-2048 turbo code in the LTE standard, with the same setup as in Fig. 5. The BER of 10^{-5} is achieved at 7.2 dB by LCMA. In contrast, both IDMA and SCMA lose functionality at this system load of 250%, due to the mismatch between their multi-user detectors and the LTE turbo code decoder. Owing to the iterative receivers in IDMA and SCMA, the outer channel code needs to adapt to the inner ESE or BP detectors, so as to guarantee the convergence. This issue becomes more critical for as the system load becomes relatively high, where strong channel codes tend to fail working. In contrast, LCMA does not involve receiver iteration. Hence it is not subject to such issue and any off-the-shelf channel code can be utilized.

Fig. 7 shows the FER performance of LCMA with various system loads, where $K = 8, 10, 12$ and $N_S = 4$. A rate 1/2 length-960 LDPC code in the 5G NR standard is used. Here we consider LCMA with parallel processing where the multistage LCMA receiver is not implemented. It is observed that LCMA with parallel processing can support all system loads under consideration. In contrast, IDMA fails for system load greater

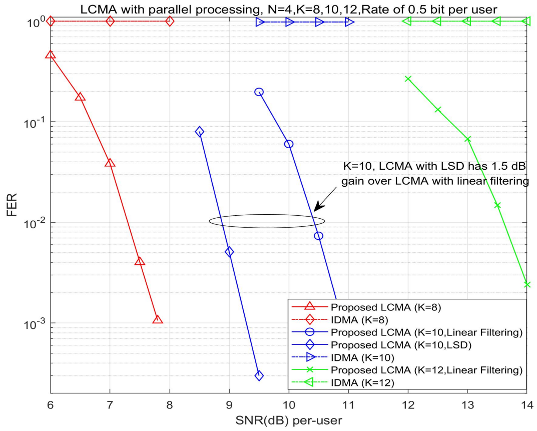


Fig. 7. FER of LCMA with $K = 10$ users and spreading sequence of length $N_S = 4$ in AWGN MA channel, with single-stage parallel processing. BPSK and a length-960 5G NR LDPC code of rate 1/2 is used for LCMA and IDMA. IDMA fails for system load greater than 200%. LCMA with LSD exhibits a 1.5 dB gain over that with the linear filtering for $K = 10$.

than 200%, again due to the poor adaptation of the NR LDPC code with the ESE detector in the iterative receiver. Here we also compared the performance of LCMA based on LSD processing and that based on linear filtering. We observe that LCMA with LSD exhibits a 1.5 dB gain over that with the linear filtering for $K = 10$. Note that with both methods, exactly K single-user decoding operations are required and carried out in parallel. The LSD method is flexible in terms of performance-complexity trade-off, by adjusting the list size in calculating the symbol-wise APPs. The linear filtering method may enjoy easier implementation due to the single-dimension processing.

B. *Fading Multiple-access Channel*

We next consider block fading MA channel where the channel coefficients follows Rayleigh distribution. Fig. 8 shows the FER of LCMA with various system loads, where $K = 10, 14, 16$ and $N_S = 4$. A rate 1/2 length-320 LDPC code in the 5G NR standard is used. The spreading sequences of LCMA are given in Table II. Thanks to the discrepancy of the effective channel gains among users brought by fading, all MA schemes under consideration can support a higher system load relative to the AWGN MA setup. The FER of LCMA, SCMA and IDMA are plotted. $Q = 10$ receiver iterations are implemented in IDMA and SCMA. It is apparent that LCMA outperforms other baseline schemes in terms of the supported system load as well as in FER for the same system load. IDMA and SCMA fail to support $K = 16$ users, while they can hardly achieve FER below 10^{-1} or 10^{-2} for $K = 14$ users and $K = 10$ users, respectively.

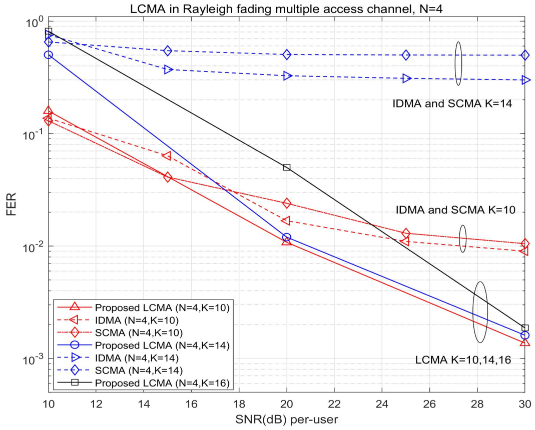


Fig. 8. FER of LCMA with various system loads in Rayleigh fading MA channel, with single-stage parallel processing. LCMA outperforms other baseline schemes in terms of the supported system load as well as in FER performance.

C. Uplink MU-MIMO

We next consider the MU-MIMO setup where the receiver is equipped with N_R antennas. We neglect the spreading process, and the spatial signatures of the receive antenna array play the role of the spreading sequences. In this setup, the iterative ESE or BP algorithms are implemented in the form of an iterative linear MMSE soft cancellation algorithm: the signal of each received antenna can be viewed as a chip-level signal in IDMA/SCMA; the chip-level cancellation with elementary extrinsic information feedback is conducted;the linear MMSE filtering combines all N received antennas signals. Fig. 9 shows the FER of LCMA with various system loads, where $K = 10, 12, 14$ and $N_R = 4$. BPSK and a length-320 5G NR LDPC code of rate 1/2 is utilized. $Q = 10$ receiver iterations are implemented in the baseline scheme with

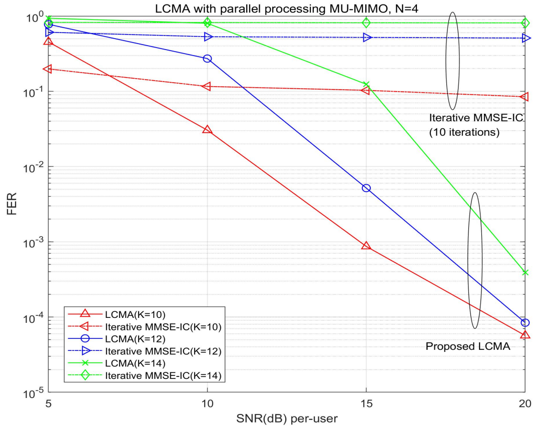


Fig. 9. FER of LCMA with various loads in multi-user MIMO of four receive antennas, using single-stage parallel processing. LCMA can support a system load of no less than 350%, while the baseline scheme with iterative receiver cannot support a load greater than 250%.

MMSE soft cancellation. It is clear that even with parallel processing, LCMA can support a system load of no less than 350%, while the baseline scheme with iterative receiver cannot support a system load greater than 200% where the FER curve flats out. It is also observed that LCMA achieves the full receive diversity of the multi-user MIMO channel for all system loads evaluated.

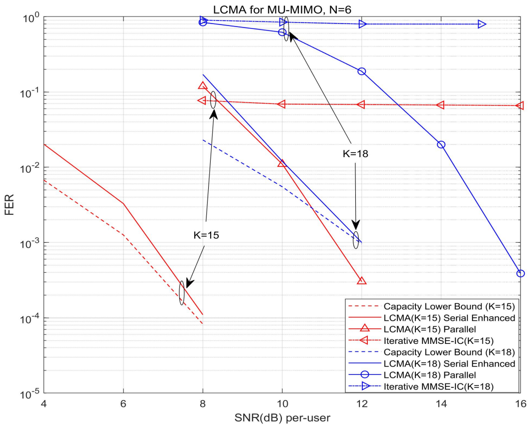


Fig. 10. FER of LCMA with various system loads in MU-MIMO with six receive antennas. LCMA with multistage receiver exhibits a gain more than 4 dB over single-stage parallel processing. The performance gap to the capacity limit is small at FER=10^{-3} for $K = 15$ and $K = 18$.

Fig. 10 shows the FER of LCMA in MU-MIMO setup with various system loads, where $K = 15, 18$ and $N_R = 6$. BPSK and a length-240 5G NR LDPC code of rate 1/2 is utilized. $Q = 10$ receiver iterations are implemented in the baseline scheme with MMSE soft cancellation. Again, even with parallel processing, LCMA can support a system load of no less than 350% (the $K = 21$ curve is not shown due to limited space of the figure), while the baseline scheme with iterative receiver cannot support a system load greater than 200%. We also include the FER with the multistage LCMA receiver, and compare to the outage probability lower bound derived from the capacity region of the MU-MIMO channel. We observe that LCMA with the multistage receiver exhibits a gain more than 4 dB over LCMA with parallel processing. The performance gap to the capacity limit is quite small at FER=10^{-3} for $K = 15$ and $K = 18$. In practice, LCMA with single-stage parallel processing enjoys lower complexity and low processing latency, but requires a greater memory size. LCMA with multistage receiver enjoys significantly improved performance and requires a smaller memory size, at the cost of higher complexity and higher processing latency, where the complexity is much lower than the baseline scheme with iterative receiver.

Previously, we presented numerical results with BPSK/QPSK signaling with $q = 2$. At present, BPSK or QPSK signaling is often used in the uplink MA system owing to its power efficiency, while spectral efficiency is not of a high priority. For future application scenarios that require high spectral efficiency uplink, such

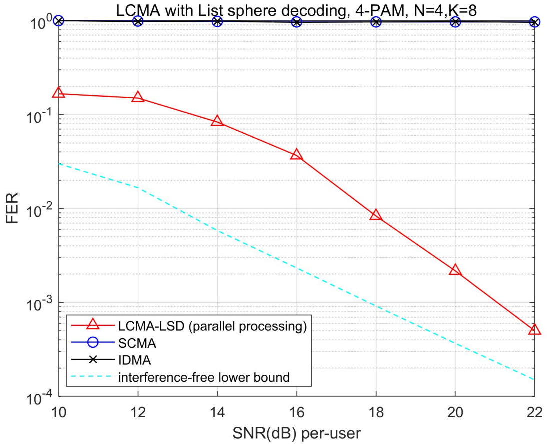


Fig. 11. FER of LCMA with 4-PAM in MU-MIMO, $N_R = 4$, with LSD and parallel processing. The list size of LSD is set to 50. For 4-PAM, LCMA with single-stage parallel processing can support a system load of at least 200%, while SCMA and IDMA lose functionality.

as 4-PAM/16-QAM signaling, q-ary ring codes are required in LCMA. Fig. 11 shows the FER of LCMA in the MU-MIMO setup with $K = 8, N_R = 4$, where each user has 4-PAM signaling. The information rate is 8 bits per channel-use per real dimension. Here we utilize a doubly irregular repeat accumulate (D-IRA) code over integer ring $0, 1, 2, 3$ with coding rate 1/2 in the LCMA system [35]. LSD based algorithm with parallel processing is applied to generate the q-ary APP sequences, which are forwarded to the ring code decoders that compute the linear message combinations. It is demonstrated that, with a higher level modulation, LCMA with parallel processing can support a system load of at least 200%. In contrast, SCMA and IDMA with 4-PAM lose functionality at this system load. We note that to achieve such load and performance advantage in the system with 4-PAM signlling, a q-ary ring code is required. Existing binary code based coded-modulation schemes such BICM, TCM and SCM do not have the structural property of lattice codes and hence the LCMA processing does not apply therein. The details on the design of q-ary ring codes for q-PAM signaling is beyond the scope of this paper, and interested readers may refer to [35].

VII. Conclusions

This paper studied a LCMA framework. We developed a package of processing techniques that help with its practical implementation, including the symbol-wise APP calculation for computing the linear message combination with list sphere detection and linear filtering, the multistage LCMA receiver via generalized matrix inversion and etc.. Significant system load and error probability performance enhancement were demonstrated over existing schemes without using outer loop receiver iteration, with lower complexity and processing delay. Off-the-shelf binary codes such as 5G NR LDPC codes can be directly used in LCMA for any system load, avoiding the issue of adaptation of channel-code and multi-user detector in existing code-domain NOMA schemes. A system load of 350% was witnessed by LCMA, and near-capacity performance was demonstrated for the MU-MIMO scenario. At this stage, there are still many open problems along this research direction, such as the identification of the optimal coefficient matrix, the general design methodology of q-ary ring codes and etc. Furthermore, the notion of lattice-code based MA can be exploited for the downlink scenario, which will be studied in the near future.

References

[1] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, “Application of non-orthogonal multiple access in lte and 5g networks,” *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, 2017.

[2] X. Wang and H. V. Poor, “Iterative (turbo) soft interference cancellation and decoding for coded CDMA,” *IEEE Trans. Comm.*, vol. 47, no. 7, pp. 1046–1061, 1999.

[3] L. Ping, L. Liu, K. Wu, and W. K. Leung, “Interleave division multiple-access,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, 2006.

[4] H. Nikopour and H. Baligh, “Sparse code multiple access,” in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, pp. 332–336, 2013.

[5] S. Kudekar and K. Kasai, “Spatially coupled codes over the multiple access channel,” in *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2816–2820, 2011.

[6] T. Yang, J. Yuan, and Z. Shi, “Rate optimization for IDMA systems with iterative joint multi-user decoding,” *IEEE Trans. Wireless Comm.*, vol. 8, no. 3, pp. 1148–1153, Mar. 2009.

[7] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, “Pattern division multiple access–A novel nonorthogonal multiple access for fifth-generation radio networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, 2016.

[8] B. Rimoldi and R. Urbanke, “A rate-splitting approach to the gaussian multiple-access channel,” *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, 1996.

[9] Y. Mao, B. Clerckx, and V. O. K. Li, “Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis,” *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8754–8770, 2019.

[10] Y. Cheng, L. Liu, and L. Ping, “Orthogonal AMP for massive access in channels with spatial and temporal correlations,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 726–740, 2021.

[11] Z. Sun, Y. Xie, J. Yuan, and T. Yang, “Coded slotted aloha for erasure channels: Design and throughput analysis,” *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4817–4830, 2017.

[12] E. Paolini, G. Liva, and M. Chiani, “Coded slotted aloha: A graph-based method for uncoordinated multiple access,” *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

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

[14] 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.

[15] L. Yang, T. Yang, Y. Xie, J. Yuan, and J. An, “Linear physical-layer network coding and information combining for the K-user fading multiple-access relay network,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5637–5650, 2016.

[16] 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 Commun.*, vol. 16, no. 3, pp. 1633–1645, 2017.

[17] L. Shi, S. C. Liew, and L. Lu, "On the subtleties of q -pam linear physical-layer network coding," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2520–2544, 2016.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] E. Sula, J. Zhu, A. Pastore, S. H. Lim, and M. Gastpar, "Compute–forward multiple access (CFMA): Practical implementations," *IEEE Trans. Comm.*, vol. 67, no. 2, pp. 1133–1147, 2018.

[23] 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.

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

[25] 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.

[26] 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.

[27] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1276, 2002.

[28] 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.

[29] U. Erez and R. Zamir, "Achieving log(1+SNR) on the awgn channel with lattice encoding and decoding," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2293–2314, 2004.

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

[31] 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, 2010.

[32] M. Qiu, L. Yang, Y. Xie, and J. Yuan, "On the design of multi-dimensional irregular repeat-accumulate lattice codes," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 478–492, 2018.

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

[34] T. Yang and et al., "On lattice-code based multiple access: Downlink algorithms and nested ring codes," *In preparation*, 2022.

[35] F. Yu, T. Yang, and Q. Chen, "Doubly irregular repeat modulation codes over integer rings for multi-source networks," *submitted to IEEE Trans. Signal Proc.*, 2022 (available at: *https://shi.buaa.edu.cn/yangtom0403/en/lwcg/32786/content/25308.htm#lwcg*).

[36] S. H. Lim, C. Feng, A. Pastore, B. Nazer, and M. Gastpar, "A joint typicality approach to compute–forward," *IEEE Trans. Inf. Theory*, vol. 64, no. 12, pp. 7657–7685, 2018.

[37] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

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

[39] W. Zhang, S. Qiao, and Y. Wei, "Hkz and minkowski reduction algorithms for lattice-reduction-aided mimo detection," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5963–5976, 2012.

[40] C. R. Roger A.Horn, *Matrix Analysis*. Cambridge University Press, 1990.

[41] O. Ordentlich, U. Erez, and B. Nazer, "Successive integer-forcing and its sum-rate optimality," *Allerton Conference Comm., Control, and Computing*, Oct. 2013.

[42] S. ten Brink and G. Kramer, "Design of repeat-accumulate codes for iterative detection and decoding," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2764–2772, Nov. 2003.

APPENDIX

Algorithm 1 List Sphere Decoding (LSD) for computing the symbol-wise APPs

```
Input: Receive signal vector $\mathbf{y}$, channel matrix $\mathbf{H}$, noise variance $\sigma^2 = 1$, parameters $N$, $K$
Output: List $\mathcal{L}$
  $\mathbf{x} = zeros(K)$, $\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$
  $\mathbf{U}^T\mathbf{U} = QR(\mathbf{H}^T\mathbf{H})$, where $QR(\cdot)$ represents QR decomposition function, and $\mathbf{U}$ is a upper triangular matrix
  $r^2 = 2 - \mathbf{y}^T(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T)\mathbf{y}$
  Search all the points which satisfy (39) below and are in the ball with center $\hat{\mathbf{x}}$ and radius $r$. Then put them in $\mathcal{L}'$.
  $\left\lceil \hat{x}_i - \frac{\sqrt{r^2 - \Phi}}{|u_{ii}|} - \sum_{j=i+1}^{M} \frac{u_{ij}}{u_{ii}}(x_j - \hat{x}_j) \right\rceil \leq x_i \leq \left\lfloor \hat{x}_i + \frac{\sqrt{r^2 - \Phi}}{|u_{ii}|} - \sum_{j=i+1}^{M} \frac{u_{ij}}{u_{ii}}(x_j - \hat{x}_j) \right\rfloor$
  where $\Phi$ is independent of $x_i$ and known from the calculation of $x_{i+1}, \cdots, x_M$.
  $\mathcal{L}$=$SortFront(\mathcal{L}', \Omega)$, where $SortFront()$ is a front-sort function that returns $\Omega$ closest points in $\mathcal{L}'$
  return $\mathcal{L}$
```

Algorithm 2 Spreading Sequence Generation

```
Input: $K$,$N$
Output: $\mathbf{s}_i, i = 1, \cdots K$
1: Generate a Hadamard matrix $\mathbf{S}^*$ of size $K$-by-$K$
2: Randomly truncate $\mathbf{S}^*$ into a $N$-by-$K$ matrix
3: Randomly set the elements into 0 according to a certain degree distribution
4: Normalize the magnitude of each column of the matrix, yielding $\mathbf{s}_i, i = 1, \cdots K$
5: Use Eq.(24) to calculate the symmetric rate
6: while $R_{sym} \leq R_0$ do
7:     return to Step 2
8: end while
9: return $\mathbf{s}_i, i = 1, \cdots K$
```