Capacity-achieving MIMO-NOMA : iterative LMMSE detection

Liu, Lei; Chi, Yuhao; Yuen, Chau; Guan, Yong Liang; Li, Ying

2019

Liu, L., Chi, Y., Yuen, C., Guan, Y. L., & Li, Y. (2019). Capacity-achieving MIMO-NOMA : iterative LMMSE detection. IEEE Transactions on Signal Processing, 67(7), 1758–1773. doi:10.1109/TSP.2019.2896242

https://hdl.handle.net/10356/137251

https://doi.org/10.1109/TSP.2019.2896242

Capacity-Achieving MIMO-NOMA: Iterative LMMSE Detection

Lei Liu, *Member, IEEE*, Yuhao Chi, Chau Yuen, *Senior Member, IEEE*,
Yong Liang Guan, *Senior Member, IEEE*, and Ying Li, *Member, IEEE*

***Abstract*—This paper considers a low-complexity iterative *linear minimum mean square error* (LMMSE) multiuser detector for the *multiple-input and multiple-output* system with *nonorthogonal multiple access* (MIMO-NOMA), where multiple single-antenna users simultaneously communicate with a multiple-antenna base station (BS). While LMMSE being a linear detector has a low complexity, it has suboptimal performance in multiuser detection scenario due to the mismatch between LMMSE detection and multiuser decoding. Therefore, in this paper, we provide the matching conditions between the detector and decoders for MIMO-NOMA, which are then used to derive the achievable rate of the iterative detection. We prove that a matched iterative LMMSE detector can achieve the optimal capacity of symmetric MIMO-NOMA with any number of users, the optimal sum capacity of asymmetric MIMO-NOMA with any number of users, all the maximal extreme points in the capacity region of asymmetric MIMO-NOMA with any number of users, and all points in the capacity region of two-user and three-user asymmetric MIMO-NOMA systems. In addition, a kind of practical low-complexity error-correcting multiuser code, called irregular repeat-accumulate code, is designed to match the LMMSE detector. Numerical results shows that the bit error rate performance of the proposed iterative LMMSE detection outperforms the state-of-art methods and is within 0.8 dB from the associated capacity limit.**

***Index Terms*—MIMO-NOMA, iterative LMMSE, capacity achieving, low-complexity multi-user detection, multi-user code.**

I. Introduction

RECENT investigations have shown that *Multi-user Multiple-Input Multiple-Output* (MU-MIMO), where multiple single-antenna users communicate with a multi-antenna *Base Station* (BS), has become increasing important due to their potential applications in 5G cellular systems and beyond [1]–[6]. In particular, massive MU-MIMO has been shown to be able to bring significant improvement in throughput and energy efficiency [3], [4].

Multiple access schemes, the fundamental techniques of coordinated multi-user communication in the physical layer, play the most important role in each cellular generation. *Frequency Division Multiple Access* (FDMA), *Time Division Multiple Access* (TDMA), *Code Division Multiple Access* (CDMA), and *Orthogonal Frequency-Division Multiple Access* (OFDMA) are the conventional *Orthogonal Multiple Access* (OMA) schemes, which orthogonalize users in time/frequency/code domain to avoid multi-user interference [7], [8]. Due to the orthogonality of OMA, no inter-user interference exists at the receiver side. Hence, simple single-user signal processing in the conventional point-to-point communication can be directly used for OMA. However, there is no free lunch. First, OMA is not able to achieve all points in the capacity region of *multiuser access channel* (MAC). Besides, massive connectivity will be the key scenario in the future wireless communication, and thus the limited radio resources cannot support the massive orthogonal access devices in the OMA any more. Apart from that, user scheduling such as resource allocation is required for orthogonal users in OMA, which leads to heavy additional overhead and results in large latency and high processing complexity in massive connectivity system.

Recently, *Non-Orthogonal Multiple Access* (NOMA), where all the users can be served con-currently in the same time/frequency/code domain, has been identified as one of the key radio access technologies to increase the spectral efficiency and reduce latency in 5G mobile networks [8]–[17]. As opposed to OMA, the key concepts behind NOMA are summarized as follows [16]–[20].

- All the users are allowed to be superimposed at the receiver in the same time/code/frequency domain.
- All points in the capacity region of MAC are achievable.
- Interference cancellation is performed at receiver, either *Successive Interference Cancellation* (SIC) or *Parallel Interference Cancellation* (PIC).

More recently, to enhance spectral efficiency and reduce latency, MIMO-NOMA that employs NOMA techniques over MU-MIMO is considered as a key air interface technology in the fifth-generation (5G) communication system [17]–[23]. Therefore, we focus on MIMO-NOMA in this paper.

Manuscript received April 15, 2018; revised October 22, 2018 and December 27, 2018; accepted January 20, 2019. Date of publication January 30, 2019; date of current version February 15, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pierre Borgnat. This work was supported in part by the National Natural Science Foundation of China under Grant 61671345 and Grant 61750110529 and in part by the joint fund of Ministry of Education of People's Republic of China under Grant 6141A02022338. The work of L. Liu was supported by the China Scholarship Council under Grant 20140690045. This paper was presented in part at the 2016 IEEE International Conference on Communications, Kuala Lumpur, Malaysia, May 22–27, 2016. *(Corresponding author: Ying Li.)*

L. Liu was with the State Key Lab of Integrated Services Networks, Xidian University, Xi'an 710071, China, and also with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: lliu_0@stu.xidian.edu.cn).

Y. Chi and Y. L. Guan are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: chiyuhao1990@163.com; eylguan@ntu.edu.sg).

C. Yuen is with the Singapore University of Technology and Design, Singapore 487372 (e-mail: yuenchau@sutd.edu.sg).

Y. Li is with the State Key Lab of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: yli@mail.xidian.edu.cn).

A. Challenge of Multi-User Detection in MIMO-NOMA

Unlike the MIMO-OMA, signal processing in MIMO-NOMA will cost higher complexity and higher energy consumption at BS [2], [3]. Low-complexity uplink detection for MIMO-NOMA is a challenging problem due to the non-orthogonal interference between the users [3], [11]–[13], especially when the number of users and the number of BS antennas are large. The optimal *multiuser detector* (MUD) for the MIMO-NOMA, such as the *maximum a-posteriori probability* (MAP) detector or *maximum likelihood* (ML) detector, was proven to be an NP-hard and non-deterministic polynomial-time complete (NP-complete) problem [24], [25]. Furthermore, the complexity of optimal MUD grows exponentially with the number of users or the number of BS antennas, and polynomially with the size of signal constellation [25], [26].

B. Background of Low-Complexity Multi-User Detector

Several low-complexity multi-user detectors have been proposed in the literature. They are mainly divided into three categories: uncoded detection, coded SIC detection, and coded PIC detection.

1) Uncoded Low-Complexity Detection: Many low-complexity linear detections such as *Matched Filter* (MF), *Zero-Forcing* (ZF) receiver, *Minimum Mean Square Error* (MMSE), and *Message Passing Detector* (MPD) [7], [27] are proposed for the practical systems. In addition, some iterative methods such as *Jacobi method*, *Richardson method* [28]–[30], *Belief Propagation* (BP) method, and iterative MPD [5], [6], [31], [32] are put forward to further reduce the computational complexity by avoiding the unfavorable matrix inversion in the linear detections. Although being attractive from the complexity view point, these individual detectors are regarded to be sub-optimal MUDs, where decoding results are not fed back to the detector. As a result, the multi-user interference is not cancellated sufficiently.

2) Coded SIC Detection: SIC, where correct decoding results are fed back to the detector for perfect interference cancellation, is one of the key technologies to improve the detection performance. It is well known that for the MAC, the SIC is an optimal strategy and can achieve all points in the capacity region of MIMO-NOMA with time-sharing technology [33], [34]. Besides, the MMSE-SIC detector [37], [38] has been proposed to achieve the optimal performance [7]. Nevertheless, the following disadvantages make SIC infeasible when applying to the practical MIMO-NOMA [3], [7], [35].

- The users are decoded one by one, which greatly increases the time delay.
- The decoding order is required to be known at both the transmitter and receiver, which results in additional overhead cost.
- It assumes that all the previous users' messages are recovered correctly and thus can be completely removed from the received signals. Nevertheless, in practice, the correct recovery is never be possible, which leads to error propagation during the interference cancellation.
- To achieve all points in the capacity region of MIMO-NOMA, time-sharing should be used, which needs cooperation between the users.
- The decoding order of SIC changes with the different channel state and different *Quality of Service* (QoS), which brings a higher overhead cost.

3) Coded PIC Detection: PIC, where users are parallelly recovered and messages exchanged between the detector and decoders are soft, is another promising technique for the practical MIMO-NOMA systems [6], [30], [32], [37]. This technique has been commonly used for the non-orthogonal MAC like the *Code Division Multiple Access* (CDMA) systems [7], [35] and the *Interleave Division Multiple Access* (IDMA) systems [39], [40]. Various iterative detectors,[1] such as iterative *Linear MMSE* (LMMSE) detector, iterative BP detector and iterative MPD [41]–[43]. The advantages of iterative detection are listed as follows.

- The complexity is very low, since the overall receiver is departed into many parallel low-complexity processors.
- Time delay is much lower than SIC, since the users are recovered in parallel.
- Error propagation is greatly mitigated, since user interference are cancellated in soft and thus perfect interference cancellation is not required.
- System overhead is reduced, since the preset decoding order is not required.
- User cooperation is removed, since time-sharing is not required.

The existing PIC detections have a good simulative performance, but are regarded as suboptimal due to a performance gap to the associated capacity limit [35]. This is due to the fact that the detector and the decoders are designed separately and are not matched with each another, which results in performance loss although the decoding feedback is included for the detection.

4) Principles of A Good Iterative Multi-User Detector: From the review above, we conclude the key principles in designing a good iterative multi-user detector.

- Multi-user interference cancellation and discrete signal reconstruction are performed respectively by MUD and user detectors.
- The decoding results should be fed back to the detector for a thorough interference cancellation.
- The detector and multi-user code should be jointly designed and matched with each other to avoid rate loss. In particular, the multi-user channel code should be optimized for the super-channel that encompasses the MIMO-NOMA channel and the multi-user detector.

The achievable rate analysis of such iterative detection for MIMO-NOMA is an intriguing problem.

C. Relationship with Interference Channel and Vector Multiple Access Channel

To clarify the relationship between interference channel (IC), vector multiple-access channel (VMAC) and MIMO-NOMA channel. We first give the definitions of IC and VMAC below.

- IC considers multiple transmitters and multiple receivers, and transmitter cooperation and receiver cooperation are not allowed (i.e. multiple scalar/vector inputs and multiple scalar/vector outputs).

[1] For the uncoded iterative detector in Section I-B1, the iteration is processed inside the detector. However, for the coded PIC detector, the iterative detection is performed between the detector and decoders, i.e., outside the detector.

- VMAC considers multiple transmitters and a single receiver, and both transmitters and receiver are equipped with multiple antennas (i.e. multiple vector inputs and a vector output).

Hence, the MIMO-NOMA channel (multiple scalar inputs and a vector output) discussed in this paper is different from IC because only a single receiver is considered. Moreover, the MIMO-NOMA channel is a special case of VMAC if each transmitter is only equipped with single antenna.

It is well known that the capacity of IC [44] is still an open issue. In addition, the capacity of general VMAC is only solved by a numerical algorithm [45]. In contrast, MIMO-NOMA channel (or VMAC with single-antenna transmitters) has a closed-form capacity region, which has been solved in [52], see also [7] and [34] for more details.

D. Gap Between P2P MIMO and MIMO-NOMA

The *Extrinsic Information Transfer* (EXIT) [46], [47], *MSE-based Transfer Chart* (MSTC) [48], [49], area theorem and matching theorem [46]–[49] are the main methods to analyse the achievable rate or the *Bit Error Rate* (BER) performance of MIMO systems. It is proven that a well-designed single-code with linear precoding and iterative LMMSE detection achieves the capacity of the MIMO systems [43]. However, this results only applies to *point-to-point* (P2P) MIMO systems.

Since there is no user collaboration in MIMO-NOMA, the precoding in P2P MIMO [43] cannot be used. Besides, the singular value decomposition (SVD) and water-filling in [43] are unachievable in multi-user MIMO NOMA too, since there is no channel information at transmitters. Furthermore, only one user rate is analyzed in P2P MIMO [43], but in MIMO-NOMA, the whole achievable rate region that contains all the user rates needs to be established. Apart from that, the non-orthogonal multi-user interference makes the problem be more complicated. For example, the decoding processes of the non-orthogonal users in MIMO-NOMA interfere with each other, which results in a much more complicated MSTC functions and area theorems. In summary, the results in P2P MIMO (e.g. [43]) cannot be cannot be straightforwardly applied to analyze the achievable rates of the iterative detection for MIMO-NOMA.

E. Contributions

In this paper, the achievable rate analysis of the iterative LMMSE detection is provided for MIMO-NOMA, which shows that the low-complexity iterative LMMSE can be rate region optimal if it is properly designed. The contributions of this paper are summarized as follows.[2]

a) Matching conditions and area theorems of the iterative detector are proposed for MIMO-NOMA.

b) Achievable rate analysis of iterative LMMSE detection are provided.

c) Analytical proofs are derived for the designed iterative LMMSE detection to achieve:
    - the capacity of symmetric MIMO-NOMA with any number of users,
    - the sum capacity of the asymmetric MIMO-NOMA with any number of users,
    - all the maximal extreme points in the capacity region of the asymmetric MIMO-NOMA with any number of users, and
    - all points in the capacity region of two-user and three-user asymmetric MIMO-NOMA.

d) We prove that the elementary signal estimator (ESE) of IDMA in Multiple Input and Signal Output (MISO) and the maximal ratio combiner (MRC) in Multiple Output and Signal Input (SIMO) are two special cases of iterative LMMSE receiver. Hence, both ESE of IDMA in MISO and MRC in SIMO are sum capacity achieving.

e) An algorithm is provided to design a practical iterative LMMSE detection.

f) A kind of capacity-approaching multi-user NOMA code for the LMMSE detection, in the form of a special (non-standard) *Irregular Repeat-Accumulate*(IRA) multiuser code, is systematically constructed. This special IRA multi-user code must be designed in conjunction with the LMMSE detection to produce extrinsic transfer functions that satisfy a certain constraint among the different users.

g) Numerical results show that our iterative LMMSE detection with optimized IRA code outperforms the existing methods, and is within 0.8dB from the associated capacity limit.

From the information theoretic point of view, to the best of our knowledge, this is the first work that proves that a proper designed PIC (joint design of the iterative LMMSE detection and the multi-user code) can achieve the capacity of MIMO-NOMA with low complexity. From the practical point of view, the jointly designed iterative LMMSE detection (PIC) has significant improvement in the BER performances over the existing iterative receivers (including both SIC and PIC) in a variety of system loads.

Comments: It is well known that finite-length coding will lead to rate loss. In this paper, when we refer to the proposed iterative LMMSE achieving the capacity (sum capacity or all points in the capacity region) of MIMO-NOMA, infinite-length channel codes are considered by default. Specifically, in this paper, we use an ideal SCM code (with infinite layers and infinite length), which is designed to match the SINR-variance transfer curves of LMMSE detection. The existence of such code is rigorously proved in Appendix D.

This paper is organized as follows. In Section II, the MIMO-NOMA system and iterative LMMSE detection are introduced. The matching conditions and area theorems for the MIMO-NOMA are elaborated in Section III. Section IV provides the achievable rate analysis. Important properties and special cases of the iterative LMMSE detection are given in Section V. Practical multiuser code design is provided in Section VI. Numerical results are shown in Section VII.

II. System Model and Iterative LMMSE Detection

Consider an uplink MU-MIMO system that showed in Fig. 1: N_u autonomous single-antenna terminals simultaneously communicate with an array of N_r antennas of the BS [3], [4]. Here, N_u and N_r can be any finite positive integers. Since all the users interfere with each other at the receiver and are non-orthogonal

[2] In points a, b, c and d, the ideal SCM codes (with infinite layers and infinite length), which are designed to match the SINR-variance transfer curves of LMMSE detection, are used for the multiuser codes.

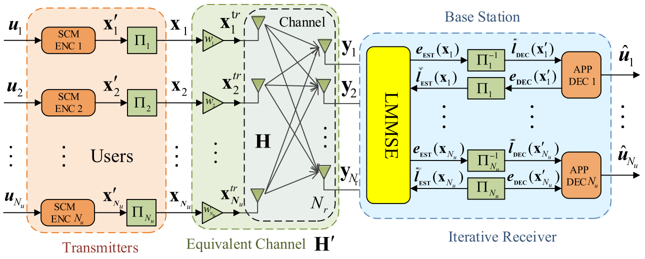

Fig. 1. Block diagram of MIMO-NOMA system. SCM ENC is the superposition coded modulation encoder and APP DEC is the *a-posteriori* probability decoder. Π_i and Π_i^{-1} denotes the interleaver and de-interleaver. LMMSE represents the LMMSE detector. The equivalent channel $\mathbf{H}'$ contains the channel $\mathbf{H}$ and the power parameter of each user w_i, $i \in \mathcal{N}_u$.

in the time, frequency and code domain, it is thus named MIMO-NOMA.[3] The system is represented as

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}^{tr}(t) + \mathbf{n}(t), \quad t \in \mathcal{N}, \;\; \mathcal{N} = \{1, \ldots, N\} \qquad (1)$$

where $\mathbf{H}$ is an $N_r \times N_u$ channel matrix, $\mathbf{n}(t) \sim \mathcal{CN}^{N_r}(0, \sigma_n^2)$ an independent additive white Gaussian noise (AWGN), $\mathbf{x}^{tr}(t) = [x_1^{tr}(t), \ldots, x_{N_u}^{tr}(t)]^T$ the transmission, and $\mathbf{y}_t$ the received vector at time t. In this paper, we consider the block fading channel [7], i.e., $\mathbf{H}$ is fixed during one block transmission and known at the BS. When the channel is block fading, in time-division duplexing (TDD) mode, it is possible for the BS to estimate the downlink channel when receiving message from the uplink. In frequency-division duplexing (FDD) mode, it is possible for the receiver feedback the channel to BS. However, as these are standard assumption for many others in the literature, we will not describe in details.

A. Transmitters

As illustrated in Fig. 1, at user i $(i \in \mathcal{N}_u,\ \mathcal{N}_u = \{1, 2, \cdots, N_u\})$, an information sequence $\mathbf{u}_i$ is encoded by an error-correcting code into an N-length sequence $\mathbf{x}'_i$, which is interleaved by an N-length independent random interleaver[4] Π_i to get $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,N}]^T$. We assume that each $x_{i,t}$ is taken over the points in a discrete signaling constellation $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$. After that, the $\mathbf{x}_i$ is scaled with w_i, and we then get the transmitting $\mathbf{x}_i^{tr}$, $i \in \mathcal{N}_u$. Let $\sigma_{x_i}^2 = 1$ denote the normalized variance of $\mathbf{x}_i$, and $\mathbf{K}_\mathbf{x}$ be power constraint diagonal matrix whose diagonal elements are w_i^2, $i \in \mathcal{N}_u$. Therefore, the system can be rewritten to

$$\mathbf{y}_t = \mathbf{H}\mathbf{K}_\mathbf{x}^{1/2}\mathbf{x}(t) + \mathbf{n}(t) = \mathbf{H}'\mathbf{x}(t) + \mathbf{n}(t), \quad t \in \mathcal{N}, \qquad (2)$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_{N_u}(t)]^T$.

B. Capacity Region of MIMO-NOMA

Let $\mathbf{Y}$ denote the received random vector, and $\mathbf{X}$ represent the transmitting random vector. Assuming $S \subseteq \mathcal{N}_u$, $S^c \subseteq \mathcal{N}_u / S$ and $S \cup S^c = \mathcal{N}_u$, the partial channel matrix is denoted as $\mathbf{H}'_S = [\{\mathbf{h}'_i, i \in S\}]_{N_r \times |S|}$, where $\mathbf{h}'_i$ is the ith column of $\mathbf{H}'$. Similar definition is applied to $\mathbf{X}_S$. Let R_i be the rate of user i and $R_S = \sum_{i \in S} R_i$ represent the sum rate of the users in set S. Then, capacity region[5] $\mathcal{R}_S$ of the MIMO-NOMA system is given by [33], [34]

$$R_S \le I(\mathbf{Y}; \mathbf{X}_S | \mathbf{X}_{S^c}) = \log \left| \mathbf{I}_{|S|} + \frac{1}{\sigma_n^2} \mathbf{H}'^H_S \mathbf{H}'_S \right|, \ \forall S \subseteq \mathcal{N}_u, \qquad (3)$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. The sum rate is

$$R_{sum} = R_{\mathcal{N}_u} = \log \left| \mathbf{I}_{N_u} + \frac{1}{\sigma_n^2} \mathbf{H}'^H \mathbf{H}' \right|. \qquad (4)$$

C. Iterative Receiver

We adopt a joint detection-decoding iterative receiver, which is widely used in the multiple-access systems [31], [37], [43]. The messages $\boldsymbol{e}_{EST}(\mathbf{x}_i), \tilde{\boldsymbol{l}}_{EST}(\mathbf{x}_i), \tilde{\boldsymbol{l}}_{DEC}(\mathbf{x}'_i)$, and $\boldsymbol{e}_{DEC}(\mathbf{x}'_i)$, $i \in \mathcal{N}_u$, are defined as the estimates of the transmissions. As illustrated in Fig. 1, at the BS, the received signals $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ and message $\{\tilde{\boldsymbol{l}}_{EST}(\mathbf{x}_i), i \in \mathcal{N}_u\}$ are passed to a LMMSE detector to estimate message $\boldsymbol{e}_{EST}(\mathbf{x}_i)$ for decoder i, which is then deinterleaved with Π_i^{-1} into $\tilde{\boldsymbol{l}}_{DEC}(\mathbf{x}'_i)$, $i \in \mathcal{N}_u$. The corresponding single-user decoder outputs message $\boldsymbol{e}_{DEC}(\mathbf{x}'_i)$ based on $\tilde{\boldsymbol{l}}_{DEC}(\mathbf{x}'_i)$. Similarly, this message is interleaved by Π_i to obtain $\tilde{\boldsymbol{l}}_{EST}(\mathbf{x}_i)$ for the detector. This process is repeated iteratively until the maximum number of iterations is achieved.

In the rest of this paper, we will not distinguish $\mathbf{x}_i$ and $\mathbf{x}'_i$ as they are same sequences with different permutations, i.e., $\tilde{\boldsymbol{l}}_{DEC}(\mathbf{x}'_i)$ and $\boldsymbol{e}_{DEC}(\mathbf{x}'_i)$ can be denoted with $\boldsymbol{e}_{EST}(\mathbf{x}_i)$ and $\tilde{\boldsymbol{l}}_{EST}(\mathbf{x}_i)$. In fact, the messages $\boldsymbol{e}_{EST}(\mathbf{x}_i)$ and $\tilde{\boldsymbol{l}}_{EST}(\mathbf{x}_i)$ can be replaced by the means and variances respectively.

1) Key Assumptions: For simplicity, we make the following assumptions, which are widely used in iterative decoding and turbo equalization algorithms [41], [43], [47], [51].

Assumption 1: For the LMMSE detector, each $x_i(t)$ is independently chosen from $\mathcal{S}$ for any i and t; the messages $\{\boldsymbol{e}_{EST}(\mathbf{x}_i), i \in \mathcal{N}_u\}$ are independent with each other, and the entries of $\boldsymbol{e}_{EST}(\mathbf{x}_i)$ are i.i.d. given $\mathbf{x}_i$.

Assumption 2: For the decoder, the messages $\boldsymbol{e}_{DEC}(\mathbf{x}'_i)$, $i \in \mathcal{N}_u$ are independent with each other, and the entries of $\boldsymbol{e}_{DEC}(\mathbf{x}'_i)$ are i.i.d. given $\mathbf{x}'_i$.

Assumptions 1 and 2 decompose the overall process into the local processors such as the detector and decoders, which simplifies the analysis of the iterative process. In detail, Assumption 1 simplifies the LMMSE estimation (see Section II-D1, and Assumption 2 simplifies the transfer function of decoders (see Section II-C2.

2) A-posteriori Probability (APP) Decoder: We assume each decoder employees APP decoding[6] at the receiver. The

[3] Here, MIMO-NOMA is different from IC since only a single receiver is considered. Moreover, MIMO-NOMA is also different from VMAC since each transmitter is only equipped with single antenna.

[4] The interleavers improve the system performance by enhancing the randomness of the messages or the channel noise, and avoiding the short cycles in the system factor graph [39], [40], [50].

[5] Different from the interference channel whose capacity is still an open issue and the vector multiple-access channel whose capacity only has a numerical solution, the capacity calculation of MIMO-NOMA is trivial and has been has been well studied in [7], [52].

[6] Although computational complexity of the APP decoding is too high to apply in practical systems, low-complexity message-passing algorithms can be used to achieve near-optimal performance [51]. APP decoding assumption is included to simplify our analysis.

extrinsic variance output of APP decoder is defined as

$$v_{i,t} = \mathrm{MMSE}\big(x_{i,t}|\tilde{\boldsymbol{l}}_{DEC}(\mathbf{x}_{i,\sim t})\big). \tag{5}$$

From Assumption 2, we have $v_{i,t} = v_i, \forall t$. Therefore, we can define the *SINR-Variance* transfer function of the decoders as

$$\mathbf{v}_{\bar{\mathbf{x}}} = \boldsymbol{\psi}(\boldsymbol{\rho}), \tag{6}$$

Where $\boldsymbol{\psi}(\boldsymbol{\rho}) = [\psi_1(\rho_1), \ldots, \psi_{N_u}(\rho_{N_u})]$.

D. LMMSE Detector

In the MIMO-NOMA, the complexity of the optimal MAP detector is too high, and LMMSE detector is an alternative low-complexity detector.

1) A-Posteriori LMMSE Estimation: Message $\tilde{l}_{EST}(x_{i,t})$ is de-mapped to $\bar{x}_{i,t}$ with variance v_i. Assumption 1 indicates that v_i is invariant with respect to t. Hence,

$$\bar{x}_{i,t} = \mathrm{E}\left[x_{i,t}|\tilde{l}_{EST}(x_{i,t})\right], v_i = \mathrm{E}\left[|x_{i,t} - \bar{x}_{i,t}|^2|\tilde{l}_{EST}(x_{i,t})\right], \tag{7}$$

where $\mathrm{E}[a|b]$ denotes the expectation of a given b. Let $\bar{\mathbf{x}}(t) = [\bar{x}_{1,t}, \ldots, \bar{x}_{N_u,t}]$ and $\mathbf{V}_{\bar{\mathbf{x}}(t)} = \mathbf{V}_{\bar{\mathbf{x}}} = \mathrm{diag}(v_1, v_2, \ldots, v_{N_u})$. The *a-posteriori* LMMSE estimation [5], [7], [31], [43] is

$$\hat{\mathbf{x}}(t) = \mathbf{V}_{\hat{\mathbf{x}}}\left[\mathbf{V}_{\bar{\mathbf{x}}}^{-1}\bar{\mathbf{x}}(t) + \sigma_n^{-2}\mathbf{H}'^H\mathbf{y}_t\right], \tag{8}$$

where $\mathbf{V}_{\hat{\mathbf{x}}} = (\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{V}_{\bar{\mathbf{x}}}^{-1})^{-1}$ denotes the *a-posteriori* deviation of the estimation. A derivation of (8) is given in APPENDIX A. For more details of LMMSE, please refer to Section II-C2 and Section IV-F of [5].

2) Extrinsic LMMSE Detector: Let $\hat{x}_{i,t}$ and $v_{\hat{x}_i}$ be the entry and diagonal entry of $\hat{\mathbf{x}}(t)$ and $\mathbf{V}_{\hat{\mathbf{x}}}$, respectively. The LMMSE detector outputs extrinsic[7] mean and variance for $x_{i,t}$ (denoted by $u_{i,t}$ and ϕ_i^{-1}) by excluding the prior message $\tilde{l}_{EST}(x_{i,t})$ with the message combining rule [27]:

$$\phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) = v_{\hat{x}_i}^{-1}(\mathbf{v}_{\bar{\mathbf{x}}}) - v_i^{-1} \quad \text{and} \quad u_{i,t} = \frac{\hat{x}_{i,t}}{\phi_i v_{\hat{x}_i}} - \frac{\bar{x}_{i,t}}{\phi_i v_i}, \tag{9}$$

where $\mathbf{v}_{\bar{\mathbf{x}}} = [v_1, v_2, \ldots, v_{N_u}]$.

3) Extrinsic Transfer Function: The following proposition is proved in Appendix B.

Proposition 1 [53], [54]: Let $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_{N_u}]$, $\boldsymbol{\phi}(\mathbf{v}_{\bar{\mathbf{x}}}) = [\phi_1(\mathbf{v}_{\bar{\mathbf{x}}}), \ldots, \phi_{N_u}(\mathbf{v}_{\bar{\mathbf{x}}})]$. The output of the LMMSE detector is an observation from AWGN channel,[8] i.e., $\mathbf{u}_t = \mathbf{x}(t) + \mathbf{n}_t^*$ with Signal Interference Noise Ratio (SINR) $\boldsymbol{\rho} = \boldsymbol{\phi}(\mathbf{v}_{\bar{\mathbf{x}}})$.

With Proposition 1, we can define the extrinsic LMMSE *SINR-Variance* transfer function of user i as

$$\phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) = v_{\hat{x}_i}^{-1} - v_i^{-1}, \quad \text{for} \quad i \in \mathcal{N}_u. \tag{10}$$

The *a-posteriori* MSE of LMMSE detector for user i is

$$\mathrm{mmse}_{ap,i}^{est}(\mathbf{v}_{\bar{\mathbf{x}}}) = v_{\hat{x}_i}. \tag{11}$$

Furthermore, Proposition 1 will be used to derive the area properties of MIMO-NOMA (see Section III-B.

Remark: The variance v_i varies from 0 to 1, because the signal power is normalized to 1. From (4), the output estimation of user i depends on the input variances of all the users. Thus, the *SINR-Variance* transfer functions of all users interfere with each other. In addition, $\phi_i(\mathbf{v}_{\bar{\mathbf{x}}})$ is monotonically decreasing in $\mathbf{v}_{\bar{\mathbf{x}}}$, which means the lower input variances of the users, the higher the output *SINR* of the detector.

E. Complexity of Iterative LMMSE Detection

From (8), the complexity of LMMSE estimator is $\Xi_{est} = \mathcal{O}\left(\min\{N_r N_u^2 + N_u^3,\ N_u N_r^2 + N_r^3\}\right)$, where $\mathcal{O}(N_u^3)$ (or $\mathcal{O}(N_r^3)$) arises from the matrix inverse calculation, $\mathcal{O}(N_r N_u^2)$ (or $\mathcal{O}(N_u N_r^2)$) from the matrix multiplication, and "min" from *Matrix Inversion Lemma*. Hence, the total complexity of iterative LMMSE detection is $\mathcal{O}\left((\Xi_{est} + N_u \Xi_{dec})N_{ite}\right)$, where N_{ite} is the number of iterations and Ξ_{dec} denotes the single-user decoding complexity per iteration. Note that the complexity of LMMSE detector is much lower than the optimal MUD whose complexity grows exponentially with N_u and N_r, and polynomially with $|\mathcal{S}|$.

III. Matching Conditions and Area Theorems

In [43], [48], [49], the *I-MMSE* theorem and the area theorems for the P2P communication systems are proposed. In this section, these results are generalized to the MIMO-NOMA systems.

A. Matching Conditions of MIMO-NOMA

1) SINR-Variance Transfer Chart: The iterative receiver performs iteration between the detector and the decoders, which are described by $\boldsymbol{\rho} = \boldsymbol{\phi}(\mathbf{v}_{\bar{\mathbf{x}}})$ and $\mathbf{v}_{\bar{\mathbf{x}}} = \boldsymbol{\psi}(\boldsymbol{\rho})$ respectively. Hence, the iteration is tracked by

$$\boldsymbol{\rho}(\tau) = \boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}(\tau - 1)\right), \mathbf{v}_{\bar{\mathbf{x}}}(\tau) = \boldsymbol{\psi}\left(\boldsymbol{\rho}(\tau)\right), \tau = 1, 2, \cdots. \tag{12}$$

Eq. (12) converges to a fixed point $\mathbf{v}_{\bar{\mathbf{x}}}^*$, which satisfies

$$\boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}^*\right) = \boldsymbol{\psi}^{-1}\left(\mathbf{v}_{\bar{\mathbf{x}}}^*\right) \text{ and } \boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right) > \boldsymbol{\psi}^{-1}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right), \\ \text{for } \mathbf{v}_{\bar{\mathbf{x}}}^* < \mathbf{v}_{\bar{\mathbf{x}}} \leq \mathbf{1},$$

where $\boldsymbol{\psi}^{-1}(\cdot)$ denotes the inverse of $\boldsymbol{\psi}(\cdot)$, which exists since $\boldsymbol{\psi}(\cdot)$ is continuous and monotonic [55]. The inequality[9] $\mathbf{v}_{\bar{\mathbf{x}}} \leq \mathbf{1}$ comes from the normalized signal power of $\mathbf{x}(t)$, $t \in \mathcal{N}$.

As shown in Fig. 2, if $\mathbf{v}_{\bar{\mathbf{x}}}^* = \mathbf{0}$, then all the transmissions can be correctly recovered, which means that $\boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right) > \boldsymbol{\psi}^{-1}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right)$ for any available $\mathbf{v}_{\bar{\mathbf{x}}}$, i.e., decoders' transfer function $\boldsymbol{\psi}^{-1}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right)$ lies below that of the detector $\boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right)$.

2) Matching Conditions: The detector and decoders are matched if

$$\boldsymbol{\phi}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right) = \boldsymbol{\psi}^{-1}\left(\mathbf{v}_{\bar{\mathbf{x}}}\right), \quad \text{for} \ \ \mathbf{0} < \mathbf{v}_{\bar{\mathbf{x}}} \leq \mathbf{1}. \tag{13}$$

Therefore, we obtain the following proposition.

[7] The *a-posteriori* estimate in (8) cannot be used directly due to the correlation issue.

[8] The "*" indicates that it is not the channel noise, but an imagined noise including the interference.

[9] In this paper, all the inequalities for the vectors or matrixes correspond to the component-wise inequalities.

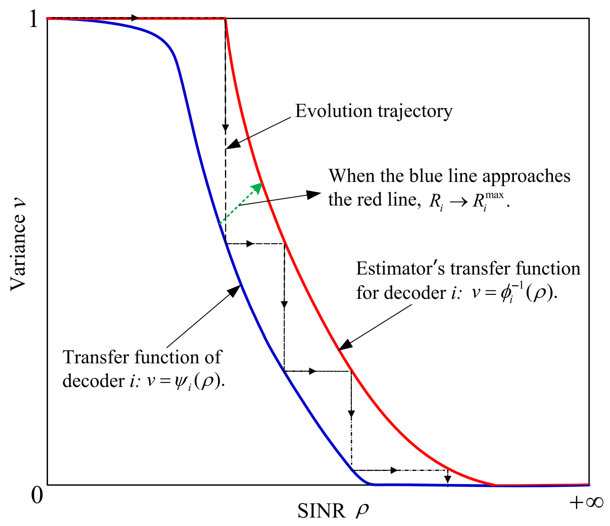

Fig. 2. SINR-variance transfer chart of the iterative receiver.

Proposition 2: For any $i \in \mathcal{N}_u$, the matching conditions of the iterative MIMO-NOMA systems can be rewritten to

$$\psi_i(\rho_i) = \phi_i^{-1}(\phi_i(\mathbf{1})) = 1, \quad \text{for} \quad 0 \le \rho_i < \phi_i(\mathbf{1}); \tag{14}$$

$$\psi_i(\rho_i) = \phi_i^{-1}(\rho_i), \quad \text{for} \quad \phi_i(\mathbf{1}) \le \rho_i < \phi_i(\mathbf{0}); \tag{15}$$

$$\psi_i(\rho_i) = 0, \quad \text{for} \quad \phi_i(\mathbf{0}) \le \rho_i < \infty. \tag{16}$$

Proof: Eq. 13 means that $\phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) = \psi_i^{-1}(v_i)$ for any $i \in \mathcal{N}_u$. First, we have $\phi_i(\mathbf{1}) > 0$, since the detector always uses the information from the channel. Hence, we get $\psi_i(\rho_i) = 1$, for $0 \le \rho_i < \phi_i(\mathbf{1})$. Second, we have $\phi_i(\mathbf{0}) > 1$, since the detector cannot remove the uncertainty introduced by the channel noise. Hence, we get $\psi_i(\rho_i) = 0$, for $\phi_i(\mathbf{0}) \le \rho_i < \infty$. At last, $\psi_i(\rho_i) = \phi_i^{-1}(\rho_i)$ exists due to its monotonicity on $\phi_i(\mathbf{1}) \le \rho_i < \phi_i(\mathbf{0})$. Therefore, we have (14)–(16). ■

Proposition 2 will be used in the area properties and rate analysis of MIMO-NOMA.

B. Area Properties

Let $\mathrm{snr}^{dec}_{pri,i}$ denote the *SNR* of the a-priori message for decoder i, $\mathrm{snr}^{est}_{ext,i}$ be the *SNR* of the extrinsic message for user i at detector, $\mathrm{mmse}^{est}_{ap,i}(\cdot)$ be the *a-posteriori* variance of the message for user i at detector, and $\mathrm{mmse}^{dec}_{ap,i}(\cdot)$ be the *a-posteriori* variance of the message at decoder i. Besides, $\mathbf{snr}^{est}_{ext,i} = [\mathrm{snr}^{est}_{ext,1}, \ldots, \mathrm{snr}^{est}_{ext,N_u}]$. The following proposition gives the area properties of the iterative detection, which will be used to derive the user rate of MIMO-NOMA.

Proposition 3: The achievable rate R_i of user i and an upper bound of R_i are given by

$$R_i = \int_0^{\infty} \mathrm{mmse}^{dec}_{ap,i}(snr^{dec}_{pri,i}) d\mathrm{snr}^{dec}_{pri,i}, \tag{17}$$

$$R_i^{\max} = \int_0^{\infty} \mathrm{mmse}^{est}_{ap,i}(\mathbf{snr}^{est}_{ext}) d\mathrm{snr}^{est}_{ext,i}, \tag{18}$$

where $R_i \le R_i^{\max}$, $i \in \mathcal{N}_u$, where the equality holds if and only if the *SINR-Variance* transfer functions of the detector and user decoders are matched with each other, i.e., the matching conditions (13)∼(16) hold.

From (4), (6) and Proposition 1, we have $\mathrm{snr}^{dec}_{pri,i} = \rho_i$, $\mathrm{snr}^{est}_{ext,i} = \phi_i(\mathbf{v}_{\bar{\mathbf{x}}})$, $\mathrm{mmse}^{dec}_{ap,i}(snr^{dec,i}_{pri}) = \left(\rho_i + \psi_i(\rho_i)^{-1}\right)^{-1}$ and $\mathrm{mmse}^{est}_{ap,i}(\mathbf{snr}^{est}_{ext,i}) = v_{\hat{x}_i}(\mathbf{v}_{\bar{\mathbf{x}}})$. Therefore, we have the following corollary from Proposition 3.

Corollary 1: With the *SINR-Variance* transfer functions $\boldsymbol{\rho} = \boldsymbol{\phi}(\mathbf{v}_{\bar{\mathbf{x}}})$ and $\mathbf{v}_{\bar{\mathbf{x}}} = \boldsymbol{\psi}(\boldsymbol{\rho})$, the achievable rate R_i of user i and an upper bound of R_i are

$$R_i = \int_0^{\infty} \left(\rho_i + \psi_i(\rho_i)^{-1}\right)^{-1} d\rho_i, \tag{19}$$

$$R_i^{\max} = \int_0^{\infty} v_{\hat{x}_i}(\mathbf{v}_{\bar{\mathbf{x}}}) d\phi_i(\mathbf{v}_{\bar{\mathbf{x}}}), \tag{20}$$

respectively, and $R_i \le R_i^{\max}$, $i \in \mathcal{N}_u$, where the equality holds if and only if the matching conditions (13)∼(16) hold.

Now, the achievable rates can be calculated by (20) or (19) together with (13) and the matching conditions (14)∼(16).

IV. Achievable Rate of Iterative LMMSE Detector

User achievable rate is derived for the iterative MIMO-NOMA in this section. The Superposition Coded Modulation (SCM) code is employed for the Forward Error Correction (FEC) code. We show that the achievable rate of iterative LMMSE can achieve the capacity of symmetric MIMO-NOMA and sum capacity of asymmetric MIMO-NOMA.

A. Achieving the Sum Capacity of Asymmetric MIMO-NOMA

For a general asymmetric MIMO-NOMA, achievable rate analysis becomes more complicated due to challenges below.

- All the users' transfer functions interfere with each other at the detector, i.e., the any output of the detector relies on every variance of the input messages from the decoders.
- All the transfer curves of decoders requires to lie below that of the detector.
- The detector and decoders are associated with each other. It is intractable to optimize over an abstract class of decoder transfer functions for each user.

1) Transfer-Constraint Parameter: The area theorem tells us that the achievable rate of every user is maximized if and only if its transfer function matches with that of the detector. Therefore, we can fix the transfer functions of the detector, and then obtain users' achievable rate by matching the decoders' transfer functions with the detector.

To make the analysis feasible, we consider a *transfer constraint* for the input variances of the detector.

$$\gamma_i(v_i^{-1} - 1) = \gamma_j(v_j^{-1} - 1), \quad \text{for any} \quad i, j \in \mathcal{N}_u. \tag{21}$$

Let $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_{N_u}]$ be the *transfer-constraint parameter* of the iterative LMMSE detection. Without loss of generality, we assume $\gamma_1 = 1$ and $\gamma_i > 0$, that is, $v_i^{-1} = 1 + \gamma_i^{-1}(v_1^{-1} - 1)$ for any $i \in \mathcal{N}_u$.

Actually, different values of $\boldsymbol{\gamma}$ give different variance tracks. Furthermore, different variance tracks correspond to different achievable rates of the users, i.e., the user's achievable rate can be adjusted by the *transfer-constraint parameter* $\boldsymbol{\gamma}$.

Fig. 3 and Fig. 4 presents the variance tracks with different values of $\boldsymbol{\gamma}$ for two-users and three-user MIMO-NOMA systems

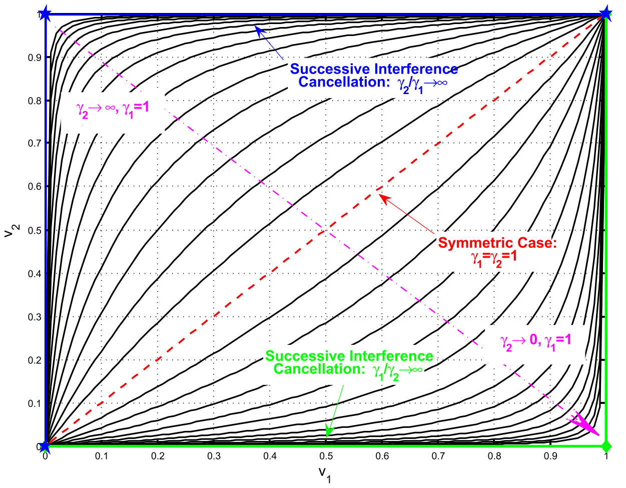

Fig. 3. Variance tracks for different $\boldsymbol{\gamma}$, where $\gamma_1 = 1$ is fixed. v_i denotes the variance of user i, $i = 1, 2$. When γ_2 changes from ∞ to 0, the track changes from the blue curve (SIC case with decoding order: user 1 → user 2) to green curve (SIC case with decoding order: user 2 → user 1). When $\gamma_1 = \gamma_2 = 1$, it degenerates into the symmetric case (red line).

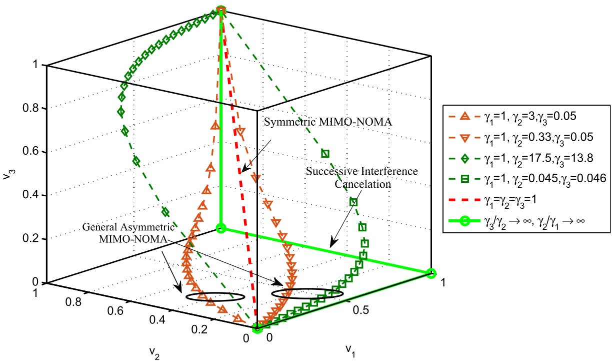

Fig. 4. Variance tracks for different $\boldsymbol{\gamma}$, where $\gamma_1 = 1$ is fixed. v_i denotes the variance of user i, $i = 1, 2, 3$. The variance track changes with γ_2 and γ_3. When $\gamma_3/\gamma_2 \to \infty$ and $\gamma_2/\gamma_1 \to \infty$ (green curve), it degenerates into the SIC case with the decoding order: user 3 → user 2 → user 1. When $\gamma_1 = \gamma_2 = \gamma_3 = 1$, it degenerates into the symmetric case (red line). The other curves are the general asymmetric cases.

respectively. As we can see, (21) includes the symmetric case (i.e. $w_1 = \cdots = w_{N_u}$) and all the SIC points (maximal extreme points of the capacity region). If $\gamma_{k_i}/\gamma_{k_{i-1}} \to \infty$, for any $i \in \mathcal{N}_u/\{1\}$, we obtain the SIC points with the decoding order $[k_1, k_2, \ldots, k_{N_u}]$, which is a permutation of $[1, 2, \ldots, N_u]$. The blue curve and green curves in Fig. 3 and Fig. 4 correspond to the SIC cases.

2) Transfer Function: With the *transfer constraint* in (21), we have

$$\mathbf{V}_{\bar{\mathbf{x}}}^{-1} = \mathbf{I}_{N_u} + \gamma_i(v_i^{-1} - 1)\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}^{-1} = \mathbf{V}_{\bar{\mathbf{x}}}^{-1}(v_i) \tag{22}$$

and

$$\begin{aligned}\mathbf{V}_{\hat{\mathbf{x}}} &= (\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{V}_{\bar{\mathbf{x}}}^{-1})^{-1} \\ &= (\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{V}_{\bar{\mathbf{x}}}^{-1}(v_i))^{-1} \\ &= \mathbf{V}_{\hat{\mathbf{x}}}(v_i)\end{aligned} \tag{23}$$

where $i \in \mathcal{N}_u$, and $\boldsymbol{\Lambda}_{\boldsymbol{\gamma}} = \mathrm{diag}(\boldsymbol{\gamma})$ is a diagonal matrix whose diagonal entries are $\boldsymbol{\gamma}$. Thus, we have

$$\phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) = v_{\hat{x}_i}(v_i)^{-1} - v_i^{-1} = \phi_i(v_i) = \rho_i. \tag{24}$$

For example, if we take $i = 1$, we have

$$\mathbf{V}_{\bar{\mathbf{x}}}^{-1} = \mathbf{V}_{\bar{\mathbf{x}}}^{-1}(v_1), \;\; \mathbf{V}_{\hat{\mathbf{x}}} = \mathbf{V}_{\hat{\mathbf{x}}}(v_1), \;\text{and}\;\; \phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) = \phi_i(v_1). \tag{25}$$

3) Asymmetric Matching Condition: With the *transfer constraint*, the matching conditions are simplified as follows.

Proposition 4: Based on (24), for $i \in \mathcal{N}_u$, the matching conditions (13) can be rewritten to

$$\psi_i(\rho_i) = \phi_i^{-1}(\phi_i(1)) = 1, \quad \text{for} \;\; 0 \le \rho_i < \phi_i(1); \tag{26}$$

$$\psi_i(\rho_i) = \phi_i^{-1}(\rho_i), \quad \text{for} \;\; \phi_i(1) \le \rho_i < \phi_i(0); \tag{27}$$

$$\psi_i(\rho_i) = 0, \quad \text{for} \;\; \phi_i(0) \le \rho_i < \infty. \tag{28}$$

Proof: From (25), we have $\phi_i(\mathbf{1}) = \phi_i(1)$ and $\phi_i(\mathbf{0}) = \phi_i(0)$. Substituting it to (14)–(16), we obtain Proposition 5. ■

4) User Achievable Rate: The users' achievable rates are given by the following lemma.

Lemma 1: For the asymmetric MIMO-NOMA with any N_u and N_r, the achievable rate of user i for iterative LMMSE detection is

$$R_i = \int_{v_1=1}^{v_1=0} \left[v_1 - \gamma_i^{-1}\left[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\right]_{i,i}\right] dv_1^{-1} - \log(\gamma_i), \tag{29}$$

where $\mathbf{V}_{\hat{\mathbf{x}}}(v_1) = (\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{I}_{N_u} + (v_1^{-1} - 1)\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}^{-1})^{-1}$, and $[\cdot]_{i,i}$ denotes the i-th diagonal entry of the matrix.

Proof: See Appendix C. ■

Lemma 1 gives the achievable rate of each user, but it is an complicated integral function and we cannot see the specific relationship between the achievable rates and $\Lambda_{\boldsymbol{\gamma}}$.

Remark: When $\gamma_i = 1$ for $i \in \mathcal{N}_u$, and for a symmetric system with: *(i)* the same rate $R_i = R$ for $i \in \mathcal{N}_u$; *(ii)* the same power $\mathbf{K}_{\mathbf{x}} = w^2\mathbf{I}$, Theorem 1 degenerates to Corollary 2.

5) Achievable Sum Rate: Although it is difficult to give the exact achievable rate region, the iterative LMMSE detection is shown to sum capacity achieving.

Theorem 1: For any N_u and N_r, the iterative LMMSE detection achieves the sum capacity of MIMO-NOMA, i.e., $R_{sum} = \log|\mathbf{I}_{N_u} + \sigma_n^{-2}\mathbf{H}'\mathbf{H}'^H|$.

Proof: See Appendix F. ■

Theorem 1 shows that for a general asymmetric MIMO-NOMA, from the sum rate perspective, the LMMSE detector is an optimal detector without losing any useful information during the estimation.

6) Monotonicity of Achievable Rate: The following lemma shows the monotonicity of achievable rate in (29).

Lemma 2: The achievable rate R_i of user i increases monotonously with γ_i and decreases monotonously with γ_j, where $i, j \in \mathcal{N}_u$ and $j \ne i$.

Proof: It is easy to find that $\mathrm{mmse}_{ap,i}^{est}$ (or $\mathrm{mmse}_{ap,i}^{dec}$) increases monotonously with γ_i and decreases monotonously γ_j for $i, j \in \mathcal{N}_u$ and $j \ne i$. Thus, based on Proposition 3, we have that R_i increases monotonously with γ_i and decreases monotonously γ_j for $j \ne i$. ■

Lemma 2 is important in user rate adjustment, i.e., if we want increase the rate of user i, it only needs to increase γ_i. Besides, the monotonicity is also important for the practical iterative detection design.

B. Achieving the Capacity of Symmetric MIMO-NOMA

Then, we consider a simple symmetric MIMO-NOMA systems, that is the users have the same power and the same rate, i.e., $\mathbf{K_x} = w^2\mathbf{I}$ and $R_i = R_j$, for $i, j \in \mathcal{N}_u$.

1) Transfer Function: Since all the users have the same conditions, we thus obtain that all the users have the same transfer functions, which means $v_i = v$ and $\rho_i = \rho$, for any $i \in \mathcal{N}_u$. Therefore, the transfer functions are derived as:

$$\begin{aligned} v_{\hat{x}_i}(\mathbf{v}_{\bar{\mathbf{x}}}) &\overset{(a)}{=} \frac{1}{N_u}\text{mmse}_{ap}^{est}(\mathbf{v}_{\bar{\mathbf{x}}}) = \frac{1}{N_u}\text{Tr}\left\{\mathbf{V}_{\hat{\mathbf{x}}}\right\} \\ &= \frac{1}{N_u}\text{Tr}\left\{\left(\sigma_n^{-2}w^2\mathbf{H}^H\mathbf{H} + v^{-1}\mathbf{I}_{N_u}\right)^{-1}\right\} \\ &= v_{\hat{x}}(v), \end{aligned} \tag{30}$$

and

$$\begin{aligned} \phi_i(\mathbf{v}_{\bar{\mathbf{x}}}) &\overset{(b)}{=} v_{\hat{x}}(v)^{-1} - v^{-1} \\ &= \frac{1}{N_u}\text{Tr}\left\{\left(\sigma_n^{-2}w^2\mathbf{H}^H\mathbf{H} + v^{-1}\mathbf{I}_{N_u}\right)^{-1}\right\}^{-1} - v^{-1} \\ &= \phi(v) = \rho, \end{aligned} \tag{31}$$

where equations (a) and (b) are obtained from the symmetric assumption. Similarly, we have $\psi_i(\rho_i) = \psi(\rho)$, $i \in \mathcal{N}_u$.

2) Matching Condition: Since all the users are symmetric, Proposition 2 can be simplified as follows.

Proposition 4: The matching conditions of the iterative symmetric MIMO-NOMA system are given by

$$\psi(\rho) = \phi^{-1}(\phi(1)) = 1, \quad \text{for} \quad 0 \le \rho < \phi(1); \tag{32}$$

$$\psi(\rho) = \phi^{-1}(\rho), \quad \text{for} \quad \phi(1) \le \rho < \phi(0); \tag{33}$$

$$\psi(\rho) = 0 \quad \text{for}, \quad \phi(0) \le \rho < \infty. \tag{34}$$

3) Achievable Rate: In this case, the analysis of symmetric MIMO-NOMA degenerates into that of single-user and single-antenna system. From the transfer functions and matching conditions above, we obtain the following theorem.

Corollary 2: For a symmetric MIMO-NOMA with any N_u and N_r that: (i) $R_i = R, \forall i \in \mathcal{N}_u$; (ii) $\mathbf{K_x} = w^2\mathbf{I}$; the iterative LMMSE detection achieves the capacity, i.e., $R_i = \frac{1}{N_u}\log|I_{N_r} + \frac{w^2}{\sigma_n^2}\mathbf{H}\mathbf{H}^H|, \forall i \in \mathcal{N}_u$, and $R_{sum} = \log|I_{N_r} + \frac{w^2}{\sigma_n^2}\mathbf{H}\mathbf{H}^H|$.

Corollary 2 shows that for a symmetric MIMO-NOMA system, the iterative detection structure is optimal, i.e., the LMMSE detector is an optimal detector without losing any useful information during the estimation.

C. Practical Iterative LMMSE Detection Design

It should be noted that the codes design depends also on $\Lambda_{\boldsymbol{\gamma}}$. Since we cannot get a closed-form solution of the user rate with respect to $\Lambda_{\boldsymbol{\gamma}}$, it is hard to obtain the proper $\Lambda_{\boldsymbol{\gamma}}$ for the given user rates. Nevertheless, Algorithm 1 provides a numeric solution of $\Lambda_{\boldsymbol{\gamma}}$ to satisfy user rate requirement.

Algorithm 1: Algorithm for Finding $\Lambda_{\boldsymbol{\gamma}}$.

```
 1: Input: $\mathbf{H}$, $\mathbf{K_x}$, $\sigma_n^2$, $\epsilon > 0$, $\delta > 0$, $N_{max}$,
    $\mathbf{R} = [R_1, \dots, R_{N_u}]$ and calculate $\mathbf{H}'$.
 2: If $\mathbf{R} \in \mathcal{R_S}$ ($\mathcal{R_S}$ is the capacity region given by (3))
 3:     Random choose $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_{N_u}]$, $\gamma_i > 0$,
        $\forall i \in \mathcal{N}_u$,
        Calculate $\mathbf{R}^{(0)}(\boldsymbol{\gamma}) = [R_1^{(0)}, \dots, R_{N_u}^{(0)}]$ by (29)
        and $t = 1$.
 4:     While $\left( ||\mathbf{R}^{(0)} - \mathbf{R}||_1 > \epsilon$ or $t < N_{max} \right)$
 5:         For $i = 1 : N_u$
 6:             fixed $\boldsymbol{\gamma}_{\sim i} = [\gamma_1, \dots, \gamma_{i-1}, \gamma_{i-1}, \dots, \gamma_{N_u}]$,
                find $\gamma_i^*$ for $R_i^{(1)}(\gamma_i = \gamma_i^*) = R_i$, and
 7:             calculate $\mathbf{R}^{(1)}(\boldsymbol{\gamma}_{\sim i}, \gamma_i^*) = [R_1^{(1)}, \dots, R_{N_u}^{(1)}]$.
 8:             While $||\mathbf{R}^{(1)} - \mathbf{R}||_1 > ||\mathbf{R}^{(0)} - \mathbf{R}||_1$
 9:                 $\gamma_i^* = (\gamma_i + \gamma_i^*)/2$ and go to step 7.
10:             End While
11:             $\gamma_i = \gamma_i^*$ and $\mathbf{R}^{(0)} = \mathbf{R}^{(1)}$.
12:         End For
13:         $t = t + 1$.
14:     End While
15:     If $t < N_{max}$
16:         Output: $\boldsymbol{\gamma}$.
17:     Else
18:         $R_i = R_i - \delta, \forall i \in \mathcal{N}_u$.
19:     End If
20: Else $\mathbf{R} \notin \mathcal{R_S}$
21:     Find the projection $\mathbf{R}^*$ of $\mathbf{R}$ on the dominant face of
        $\mathcal{R_S}$.
22:     $\mathbf{R} = \mathbf{R}^*$, and go back to step 2.
23: End If
```

For any N_u and N_r, Algorithm 1 gives a numeric search of $\Lambda_{\boldsymbol{\gamma}}$ given rate $\mathbf{R}$, where N_{max} is the maximum iterative number, ϵ and δ indicate the allowed precision, and $||\cdot||_1$ denotes the 1-norm. It should be noted that γ_i^* in step 6 definitely exists and can be easy searched by dichotomy or quadratic interpolation method as R_i increases monotonously with γ_i (Lemma 2). In addition, steps 8$\sim$10 ensure that the new γ_i^* is always better than the previous one and the search program will not stop until the requirement $\Lambda_{\boldsymbol{\gamma}}$ is got. Experimentally, we find that the points in the system capacity region are always achievable.

V. Important Properties and Special Cases of Iterative LMMSE Detection

Can the iterative LMMSE detection achieve all points in the capacity region of asymmetric MIMO-NOMA? To answer this question, we derive some properties and show that:

- for the two-user MIMO-NOMA, all points in the capacity region can be achieved by iterative LMMSE detection;
- all the maximal extreme points in the capacity region of MIMO-NOMA with any number of users can be achieved by iterative LMMSE detection.

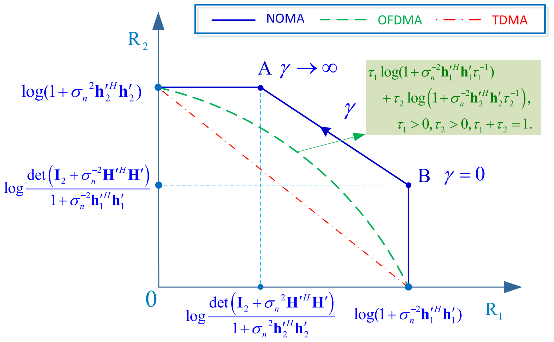
NOMA
OFDMA
TDMA
R_2
$\log(1+\sigma_n^{-2}\mathbf{h}_2'^H\mathbf{h}_2')$
A $\gamma \to \infty$
$\tau_1 \log(1+\sigma_n^{-2}\mathbf{h}_1'^H\mathbf{h}_1'\tau_1^{-1})$
$+\tau_2 \log\left(1+\sigma_n^{-2}\mathbf{h}_2'^H\mathbf{h}_2'\tau_2^{-1}\right),$
$\tau_1 > 0, \tau_2 > 0, \tau_1+\tau_2 = 1.$
γ
$\log\frac{\det\left(\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'\right)}{1+\sigma_n^{-2}\mathbf{h}_1'^H\mathbf{h}_1'}$
B $\gamma = 0$
0
$\log\frac{\det\left(\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'\right)}{1+\sigma_n^{-2}\mathbf{h}_2'^H\mathbf{h}_2'}$
$\log(1+\sigma_n^{-2}\mathbf{h}_1'^H\mathbf{h}_1')$
R_1

Fig. 5. Achievable region of iterative LMMSE detection for two-user MIMO-NOMA system. When the parameter γ changes from 0 to ∞, point (R_1, R_2) moves from maximal extreme point B to maximal extreme point A along segment AB.

Furthermore, MISO and SIMO are discussed as two special cases, which show that the ESE in IDMA and MRC are sum capacity optimal for MISO and SIMO respectively.

A. *Achieving the Maximal Extreme Point*

As it is mentioned in *Capacity Region Domination Lemma* in Appendix G, the system capacity region is dominated by a convex combination of the maximal extreme points, which can be achieved by SIC.

Here, we show that all these maximal extreme points can be achieved by iterative LMMSE detection when the transfer-constraint parameter Λ_γ is properly chosen.

Corollary 3: For any N_u and N_r, all the maximal extreme points in the capacity region of MIMO-NOMA can be achieved by iterative LMMSE detection.

Proof: See Appendix H. ■

This corollary shows that if the parameter Λ_γ is properly chosen, the iterative LMMSE detection degenerates into the SIC methods, i.e., the SIC methods are special cases of the proposed iterative LMMSE detection.

B. *Two-User MIMO-NOMA*

As it is mentioned, it is hard to calculate the specific achievable user rates from (29). However, in two-user case, the achievable rate region can be calculated and it equals to the capacity of MIMO-NOMA.

Theorem 2: Iterative LMMSE detection achieves the whole capacity region of two-user MIMO-NOMA:

$$\begin{cases} R_1 \le \log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_1\mathbf{h}'_1), \\ R_2 \le \log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_2\mathbf{h}'_2), \\ R_1+R_2 \le \log|\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'|. \end{cases} \tag{35}$$

Proof: The pentagon in Fig. 5 indicates the capacity region of two-user MIMO-NOMA system, which is dominated by segment AB, and point A and point B are two maximal extreme points. Without loss of generality, we let $\gamma_1 = 1$ and $\gamma_2 = \gamma \in [0, \infty)$. From Theorem 1, we get

$$R_{sum} = R_1 + R_2 = \log|\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'|, \tag{36}$$

which is the exact sum capacity of the system.

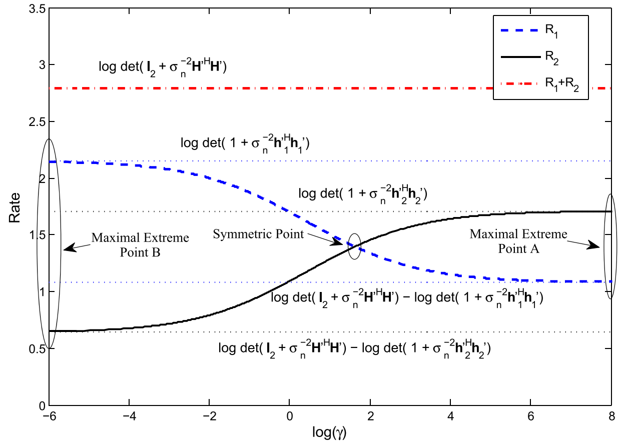

Fig. 6. Relationship between the user rates and parameter γ of the iterative LMMSE detection for two-user MIMO-NOMA system. $N_r = 2$, $N_u = 2$, $\sigma_N^2 = 0.5$ and $\mathbf{H} = [1.32\ -1.31;\ -1.43\ 0.74]$.

In addition, as we discussed in Corollary 3, when γ changes from 0 to ∞, R_1 reduces from $\log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_1\mathbf{h}'_1)$ to $\log|\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'| - \log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_1\mathbf{h}'_1)$, and R_2 increases from $\log|\mathbf{I}_2+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'| - \log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_2\mathbf{h}'_2)$ to $\log(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_2\mathbf{h}'_2)$. As the R_1 and R_2 are both continuous functions of γ, from (36), we can see that when the parameter γ changes from 0 to ∞, the point (R_1, R_2) moves from maximal extreme point B to maximal extreme point A along the segment AB. It means that the iterative LMMSE detection can achieve any point on the segment AB. Therefore, the iterative LMMSE detection achieves all points in the capacity region as it is dominated by the segment AB. ■

Let $\gamma_1 = 1$ and $\gamma_2 = \gamma$, and we can give the specific expressions of R_1 and R_2. The following corollary is derived directly from Lemma 1.

Corollary 4: For two-user MIMO-NOMA with iterative LMMSE detection, the user rates are given by

$$\begin{cases} R_1 = \frac{1}{2}\log(\gamma|A|) + \frac{a_{22}\gamma - a_{11}}{2\eta}\log\frac{a_{22}\gamma + a_{11} - \eta}{a_{22}\gamma + a_{11} + \eta}, \\ R_2 = \frac{1}{2}\log(\gamma^{-1}|A|) - \frac{a_{22}\gamma - a_{11}}{2\eta}\log\frac{a_{22}\gamma + a_{11} + \eta}{a_{22}\gamma + a_{11} + \eta}, \end{cases} \tag{37}$$

where

$$\mathbf{A} = \sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{I}_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and $\eta = \sqrt{a_{22}^2\gamma^2 + 2(2a_{21}a_{12} - a_{22}a_{11})\gamma + a_{11}^2}$. It is easy to find that η is a real number since $\mathbf{A}$ is positive definite and $\gamma \ge 0$.

It should be noted from (37) that R_1 and R_2 are nonlinear functions of γ. It is easy to check that $R_1 + R_2 = \log\det\left(\mathbf{I}_2 + \sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'\right)$, and when $\gamma \to 0$ (or $\gamma \to \infty$), the limit of (R_1, R_2) in (37) converges to the maximal point B (or A) in Fig. 5. When the parameter γ changes from 0 to ∞, the point (R_1, R_2) can achieve any point on the segment AB in Fig. 5. It also shows an alternative proof of Theorem 3. In addition, the achievable rates of TDMA and OFDMA are strictly smaller than that of the iterative LMMSE NOMA.

Fig. 6 presents the rate curves of R_1 and R_2 respect to the parameter γ. It verifies that R_2 increases monotonously with the γ (or γ_2), and $R_1 + R_2$ equals to the sum capacity.

C. *MISO:* $N_r = 1$

Let $N_r = 1$. From (48), (9) can be rewritten to

$$u_{i,t} = x_{i,t} + \frac{v_i^2 {h'_i}^H}{v_i - v_{\hat{x}_i}} \left(\sigma_n^2 + \mathbf{h}'\mathbf{V}_{\bar{\mathbf{x}}}\mathbf{h}'^H\right)^{-1} \times \left[\mathbf{h}'\left(\mathbf{x}_{\backslash i,t} - \bar{\mathbf{x}}_{\backslash i,t}\right) + \mathbf{n}_t\right],$$

$$v_{\hat{x}_i} = v_i - v_i^2 |h_i|^2 \left(\sigma_n^2 + \mathbf{h}'\mathbf{V}_{\bar{\mathbf{x}}}\mathbf{h}'^H\right)^{-1}.$$

Thus,

$$u_{i,t} = x_{i,t} + \frac{{h'_i}^H}{|h'_i|^2} \left[\mathbf{h}'\left(\mathbf{x}_{\backslash i,t} - \bar{\mathbf{x}}_{\backslash i,t}\right) + \mathbf{n}_t\right], \tag{38}$$

$$\rho_i^{-1} = [v_{\hat{x}_i}^{-1} - v_i^{-1}]^{-1} = \frac{1}{|h'_i|^2}\left[\sum_{k \neq i} |h'_k|^2 v_k + \sigma_n^2\right]. \tag{39}$$

Equivalently, it can be rewritten to

$$\mathbf{u}_t = \mathbf{\Lambda}^{-1}_{\mathbf{h}'^H\mathbf{h}'}[\mathbf{h}'^H y_t - \mathbf{\Omega}_{\mathbf{h}'^H\mathbf{h}'}\bar{\mathbf{x}}_t], \tag{40}$$

$$\mathbf{v}^e = \rho.^{-1} = (\sigma_n^2 + \mathbf{h}'\mathbf{V}_{\bar{\mathbf{x}}}\mathbf{h}'^H)|\mathbf{h}'|.^{-2} - \bar{\mathbf{v}}, \tag{41}$$

where $\mathbf{\Lambda}_{\mathbf{A}} = \mathrm{diag}\{\mathbf{A}\}$, $\mathbf{\Omega}_{\mathbf{A}} = \mathbf{A} - \mathbf{\Lambda}_{\mathbf{A}}$, and $|\mathbf{h}'|.^{-2} = [|h'_1|^{-2}, \ldots, |h'_{N_u}|^{-2}]$.

Relation to ESE in IDMA: Note that (40) and (41) are the same as the ESE in IDMA [39], which means that the ESE in IDMA is a kind of LMMSE receiver. This explains that IDMA is a good multiple access scheme, since it can achieve the sum capacity of the MISO system.

D. *SIMO:* $N_u = 1$

Let $N_u = 1$. From (48), (9) can be rewritten to

$$\hat{x}_t = v_{\hat{x}}\left[v^{-1}\bar{x}_t + \sigma_n^{-2}\mathbf{h}'^H\mathbf{y}_t\right], \tag{42}$$

$$v_{\hat{x}} = [\sigma_n^{-2}\|\mathbf{h}'\|^2 + v^{-1}]^{-1}, \tag{43}$$

and

$$u_t = \frac{\mathbf{h}'^H\mathbf{y}_t}{\|\mathbf{h}'\|^2}, \quad v_{\hat{x}} = \frac{\sigma^2}{\|\mathbf{h}'\|^2}. \tag{44}$$

In this case, the iteration between the detector and decoders are trivial.

Relation to MRC: Note that (44) is the exact MRC [56], which means that MRC is a kind of LMMSE receiver. This shows that MRC is optimal and can achieve the capacity of the SIMO system.

VI. Practical Multiuser Code Design for MIMO-NOMA

Recently, Low-Density Parity-Chek (LDPC) codes are optimized to support much higher sum spectral efficiency and user loads for MISO in [57]–[59]. In addition, a LDPC code concatenated with a simple repetition code is constructed to obtain a near MISO capacity performance in [60], . To further support massive users, an IRA code parallelly concatenated with a repetition code is proposed in [61], [62]. However, these design methods do not consider the effect of multiple receive antennas. In this paper, a kind of multi-user IRA code consisting of repetition code and IRA code is optimized for MIMO-NOMA. For more details, please refer to [20]. We will show that the optimized IRA can approaching the MIMO-NOMA capacity (e.g. BER performances are within 0.8dB away from the Shannon limit) for various of system loads. In this section, we give the multi-user IRA code design in detail.

To design suitable multiuser codes for the LMMSE detection, we first derive a transformation between the input-output variance of LMMSE detection and the input-output mutual information of the single-user decoders. Then, based on the EXIT analysis [47], [61]–[63], code parameters can be optimized to match well with LMMSE detection.

To be specific, since the output of LMMSE can be equivalent to the observation from AWGN channel, the *extrinsic* variance associated with the estimated signal from LMMSE is the variance of equivalent noise, such that the *a-priori* mutual information for the decoder is obtained by exploiting the EXIT analysis. For general linear block codes, the EXIT functions can be obtained easily [47], [61]–[63]. For the opposite direction, the *a-priori* variance of LMMSE is determined by the *extrinsic* mutual information from the decoder. The whole iterative process will stop when the decoding is successful or the maximum iteration number is reached. In other words, we can statistically trace the iterative message update between LMMSE detection and a bank of single-user decoders. The detailed process is as follows.

A. *LMMSE → Decoder*

For simplicity, we assume $\mathbf{H}'$ is IID Gaussian, and consider the detection of user k. Let $\bar{x}_k$ and u_k be *a-priori* and *extrinsic* estimations of LMMSE detection associated with x_k. Correspondingly, let v_k and v_k^e be the variances of $\bar{x}_k$ and u_k respectively. We can obtain the *a-posteriori* output variance $v_{\hat{x}_k}$ of LMMSE is [5], [6], [31]

$$v_{\hat{x}_k} = \frac{\sqrt{(snr^{-1} + N_r - N_u)^2 + 4N_u snr^{-1}} - (snr^{-1} + N_r - N_u)}{2N_u(v_k)^{-1}},$$

where $snr = v_k/\sigma_n^2$. *Extrinsic* output variance of LMMSE is

$$v_k^e = [(\hat{v}_k)^{-1} - (v_k)^{-1}]^{-1} = (v_k) \times \frac{\sqrt{(snr^{-1} + N_r - N_u)^2 + 4N_u snr^{-1}} - (snr^{-1} + N_r - N_u)}{(snr^{-1} + N_r + N_u) - \sqrt{(snr^{-1} + N_r - N_u)^2 + 4N_u snr^{-1}}}$$

Based on Proposition 1, we can rewritten $u_k = x_k + \tilde{z}_k$, where $\tilde{z}_k$ is an equivalent Gaussian noise with mean 0 and variance $Var(\tilde{z}_k) = Var(u_k) = v_k^e$. Therefore, *a-priori* mutual information associated with x_k for the DEC can be obtained.

B. *Code Optimization → Detector*

Following the similar methods in [61]–[63], the EXIT function of repetition-aided IRA can be obtained and then *extrinsic* mutual information I_k^e is calculated. According to EXIT analysis [47], [61]–[63], output log-likelihood ratio L_k^e obeys Gaussian distribution $\mathcal{N}((J^{-1}(I_k^e))^2/2, (J^{-1}(I_k^e))^2)$, where

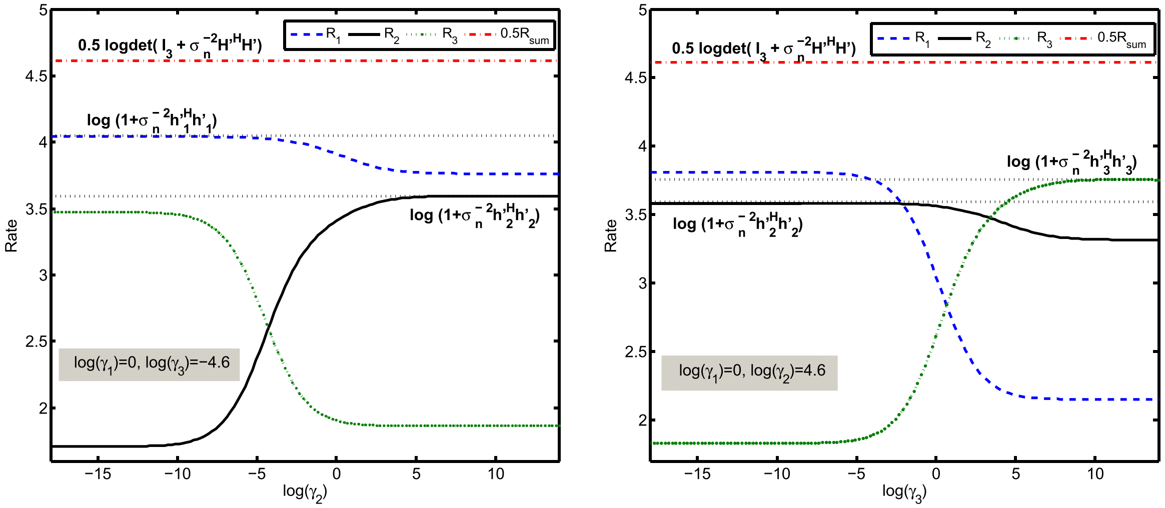

Fig. 7. Relationship between the user rates and parameters $[\gamma_2, \gamma_3]$ of the iterative LMMSE detection for three-user MIMO-NOMA. $N_r = 2, N_u = 3, \sigma_N^2 = 0.5$ and $\mathbf{H} = [0.678\ 0.603\ 0.655;\ 0.557\ 0.392\ 0.171]$.

function $J(\cdot)$ is given in [47]. Since x_k is a BPSK signal, variance $v_k = E_{L_k^e}[1 - (\tanh(L_k^e/2))^2]$ is obtained by Monte Carlo simulations, which is fed back to the LMMSE.

By using this variance-EXIT transfer process between the LMMSE and decoder, we trace statistically the message update and then optimize the parameters of repetition-aided IRA codes to match well with the LMMSE.

VII. Numerical Results

This section presents the numerical results of achievable rate of three-user MIMO-NOMA, and provides the BER simulations for the proposed iterative LMMSE detection with optimized multi-user codes.

A. Three-User MIMO-NOMA

For three-user MIMO-NOMA, it is hard to get a closed-form solution of the user rates. Hence, it is difficult to show the exact achievable rate region of the iterative LMMSE detection. However, the user rates in (29) can be solved numerically.

Fig. 7 shows the relationships between the user rates and $[\gamma_2, \gamma_3]$ with $\gamma_1 = 1$, where $N_r = 2$, $N_u = 3$, $\sigma_N^2 = 0.5$, and $\mathbf{H} = [0.678\ 0.603\ 0.655;\ 0.557\ 0.392\ 0.171]$. Notice that although the user rates change with γ_2 and γ_3, the sum rate R_{sum} is constant and equals to the system sum capacity. Furthermore, the user rate R_2 increases monotonously with γ_2, but R_1 and R_3 decrease monotonously with γ_2. Similarly, the user rate R_3 increases monotonously with γ_3, but R_1 and R_2 decrease monotonously with γ_3.

In Fig. 8, the system capacity region is the polygonal consisted by the red lines, which is dominated by the red hexagonal face. The red points in Fig. 8 are the achievable points of the iterative LMMSE detection. It shows that as we change the values of γ_2 and γ_3, the achievable points can reach any point on the dominated hexagonal face. Therefore, for the three-user MIMO-NOMA, the iterative LMMSE detection can also achieve all points in the capacity region, i.e., the iterative LMMSE detection is an optimal detection. In addition, we can see that the achievable rates of TDMA and OFDMA are strictly smaller than that of the iterative LMMSE NOMA.

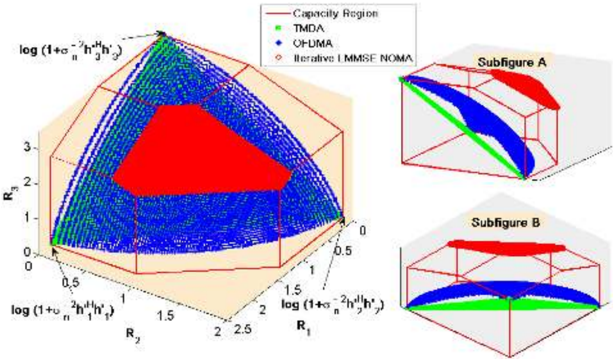

Fig. 8. Achievable rates for all (γ_1, γ_2) of the iterative LMMSE detection for three-user MIMO-NOMA system. $N_r = 3$, $N_u = 3$, $\sigma_N^2 = 0.5$ and $\mathbf{H} = [1.95\ 1.28\ -2.53;\ -0.31\ -0.16\ 2.22;\ 0.55\ 1.08\ -1.98]$. Subfigure A and Subfigure B are the same figure with different rotated viewports.

It should be noted that the results in this paper can also apply to the overloaded MIMO-NOMA systems (like Fig. 7) that the number of users is larger than the number of BS antennas, i.e., $N_u > N_r$.

B. BER Performance With Optimized IRA Codes

Here, we assume that each user employs a repetition-aided IRA code proposed for the Multiple-Access Channel (MAC) [61], [62], which is constructed by parallelly concatenating a repetition code and IRA code. In this paper, we optimize the repetition-aided IRA codes over MIMO-NOMA systems with channel load $\beta = \{0.5, 1, 2, 3\}$, where user number N_u and receive antenna N_r are $(N_u, N_r) = (8, 16), (16, 16), (16, 8)$, and $(24, 8)$, respectively. The corresponding optimized code parameters are given in Table I, which illustrates that these decoding thresholds are very close to the Shannon limits.

To verify the finite-length performance of the repetition-aided IRA codes, we provide the BER performances of the optimized

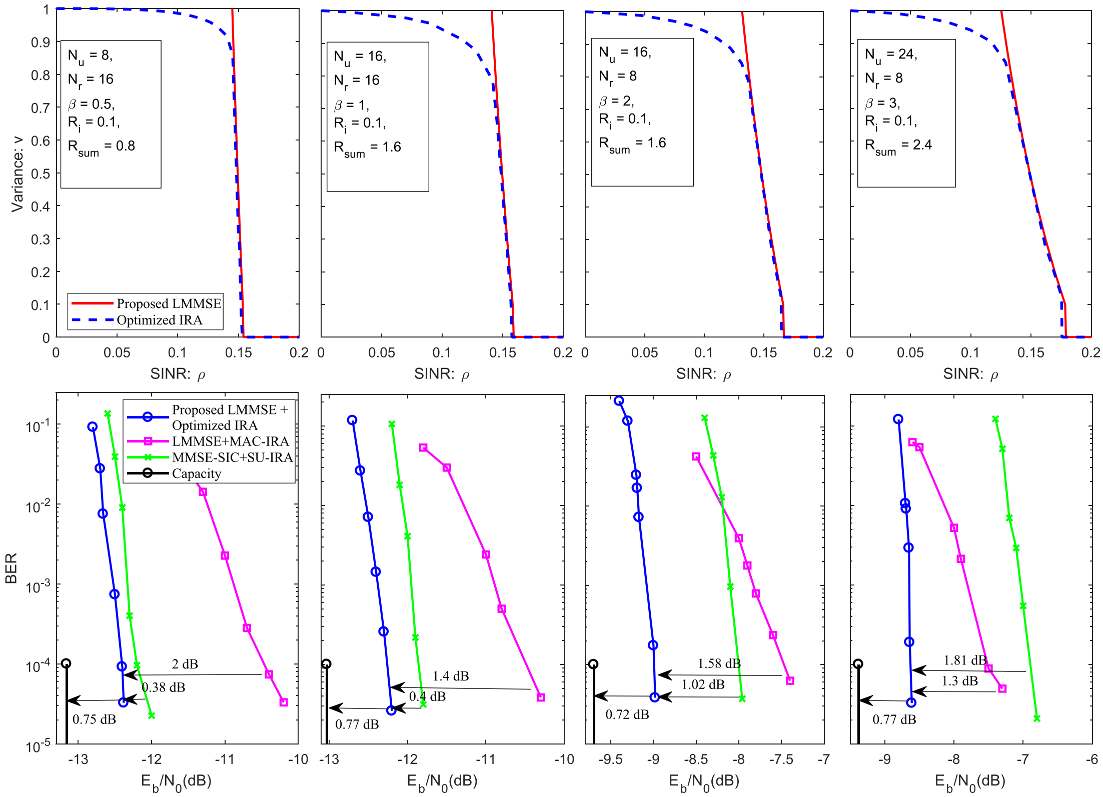

Fig. 9. SINR-variance transfer charts and BER performances of the LMMSE Receiver for MIMO-NOMA with channel load $\beta = \{0.5, 1, 2, 3\}$, where user number N_u and receive antenna N_r are $(N_u, N_r) = (8, 16), (16, 16), (16, 8), (24, 8)$ respectively. Each user is encoded by an optimized IRA code with code rate 0.1 bits/symbol and code length 4.096×10^4. The use rate of MAC-IRA code is 0.08 and decoding threshold is 0.03 dB from the MAC capacity. The rate of SU-IRA is 0.1 and decoding threshold is from 0.05 dB from the single-user capacity.

TABLE I
OPTIMIZED REPETITION-AIDED IRA CODES OVER MIMO-NOMA

β	0.5	1	2	3
N_u	8	16	16	24
N_r	16	16	8	8
R_u	0.1	0.1	0.1	0.1
N	4×10^4	4×10^4	4×10^4	4×10^4
R_{sum}	0.8	1.6	1.6	2.4
q	1	2	2	2
α	2	2	2	2
λ_3 λ_{10} λ_{30} λ_{80} λ_{100}	0.087105 0.138217 0.207022 0.068682 0.498975	0.1016 0.138386 0.262982 0.114347 0.382685	0.107994 0.129009 0.219708 0.141601 0.401687	0.116863 0.127289 0.159387 0.234121 0.36234
$(\frac{E_b}{N_0})^*_{\text{dB}}$	-13.14	-12.95	-9.66	-9.35
S. L.	-13.16	-13.03	-9.7	-9.38

codes. Each user employs a random interleaver and the length of information vector for each user is 4096. The rate of each user is $R_u = 0.1$ bits/symbol, and the sum rate is $R_{sum} = 0.1 * N_u$ bits per channel use. E_b/N_0 is calculated by $E_b/N_0 = \frac{P_u}{2R_u\sigma_n^2}$, where $P_u = 1$ is the power of each user, and σ_n^2 is the variance of the Gaussian noise. The standard sum-product algorithm is used for the single-user decoding, in which the maximum iteration number is 250. Fig. 9 shows that for all β, gaps between the BER curves of the codes at 10^{-5} and the corresponding Shannon limits are about $0.7 \sim 0.8$ dB.

To validate the advantage of the proposed system through matching between LMMSE detector and optimized IRA codes, we provide two state-of-art systems for comparisons, which are LMMSE detector combined with an existing repetition-aided IRA code [61], [62], and MMSE-SIC detector [37], [38] combined with a capacity-approaching Single-User IRA (SU-IRA) code. Note that the parameters of repetition-aided IRA code [61], [62] are $\lambda(x) = 0.063021x + 0.228288x^2 + 0.111951x^9 + 0.226877x^{29} + 0.369864x^{49}$, $q = 5$, and $\alpha = 1$, denoted as MAC-IRA code, whose rate is 0.08 and decoding threshold is 0.03 dB from the MAC capacity. The parameters of SU-IRA are $0.085867x^2 + 0.132226x^9 + 0.198883x^{29} + 0.276011x^{79} + 0.307013x^{99}$, $q = 1$, and $\alpha = 2$, whose rate is 0.1 and decoding threshold is from 0.05 dB from the single-user capacity. As shown as Fig. 9, when the BER curves of three systems are at 10^{-4}, the optimized IRAs have $1.4 \sim 2$ dB performance gains over the un-optimized IRAs, and $0.38 \sim 1.3$ dB performance gains over the systems consisting of MMSE-SIC detector and the SU-IRA code. These comparisons demonstrate that multiuser code optimization provides a promising new treatment for the applications of MIMO-NOMA technologies.

VIII. CONCLUSION

The theoretical limit of the PIC iterative receiver has been an open problem for a long time, especially for the multi-user MIMO channel. This paper analyzes the achievable rate region of the iterative LMMSE multi-user detection for both symmetric and asymmetric MIMO-NOMA. For the symmetric case, it is proved that iterative LMMSE detection achieves the capacity of MIMO-NOMA with any number of users; while for the asymmetric case, it is proved that the iterative LMMSE detection achieves the *sum capacity* of MIMO-NOMA with any number of users. In addition, all the maximal extreme points in the capacity region of MIMO-NOMA with any number of users are achievable, and all points in the capacity regions of two-user and three-user systems are also achievable. Finally, a kind of IRA multiuser code is designed for the iterative LMMSE receiver. Simulation results show that under different channel loads, the BERs of the proposed iterative LMMSE detection are within 0.8dB from the Shannon limits and outperform the existing methods. Furthermore, the improvement is more notable for large system overloads (e.g. $\beta \geq 3$), while for small system overloads (e.g. $\beta \leq 0.5$), the AWGN SU-IRA and the MMSE SIC with SU-IRA is good enough since the user interference is negligible.

How to design a low-complexity iterative receiver to achieve the capacity region of the general vector multiple access channel [45] will be an interesting future work.

APPENDIX A
DERIVATION OF *A-Posteriori* LMMSE

We assume $\mathbf{x}(t) \sim \mathcal{CN}(\bar{\mathbf{x}}(t), \mathbf{V}_{\bar{\mathbf{x}}})$, i.e. $p(\mathbf{x}(t)) \propto e^{-(\mathbf{x}(t)-\bar{\mathbf{x}}(t))^H \mathbf{V}_{\bar{\mathbf{x}}}^{-1}(\mathbf{x}(t)-\bar{\mathbf{x}}(t))}$. Since $\mathbf{n}(t) \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$, we have

$$p(\mathbf{y}_t|\mathbf{x}(t)) \propto e^{-\frac{(\mathbf{y}_t-\mathbf{H}'\mathbf{x}(t))^H(\mathbf{y}_t-\mathbf{H}'\mathbf{x}(t))}{\sigma_n^2}}.$$

Thus, the *a-posteriori* conditional probability of $\mathbf{x}(t)$ given $\mathbf{y}_t$ is

$$\begin{aligned} &p(\mathbf{x}(t)|\mathbf{y}_t) \\ &\quad = p(\mathbf{x}(t))p(\mathbf{y}_t|\mathbf{x}(t)) \\ &\quad \propto e^{-\mathbf{x}(t)^H\left[\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'+\mathbf{V}_{\bar{\mathbf{x}}}^{-1}\right]\mathbf{x}(t)+2\mathbf{x}(t)^H\left[\mathbf{V}_{\bar{\mathbf{x}}}^{-1}\bar{\mathbf{x}}(t)+\sigma_n^{-2}\mathbf{H}'^H\mathbf{y}_t\right]} \\ &\quad \propto e^{-\mathbf{x}(t)^H\mathbf{V}_{\hat{\mathbf{x}}}^{-1}\mathbf{x}(t)+2\mathbf{x}(t)^H\mathbf{V}_{\hat{\mathbf{x}}}^{-1}\hat{\mathbf{x}}(t)} \end{aligned} \tag{45}$$

Therefore, the *a-posteriori* estimation and variance are

$$\hat{\mathbf{x}}(t) = \mathbf{V}_{\hat{\mathbf{x}}}\left[\mathbf{V}_{\bar{\mathbf{x}}}^{-1}\bar{\mathbf{x}}(t) + \sigma_n^{-2}\mathbf{H}'^H\mathbf{y}_t\right], \tag{46}$$

$$\mathbf{V}_{\hat{\mathbf{x}}} = (\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}' + \mathbf{V}_{\bar{\mathbf{x}}}^{-1})^{-1}. \tag{47}$$

Hence, we obtain (8).

APPENDIX B
PROOF OF PROPOSITION 1

The *a-posteriori* LMMSE in Eq. (8) can be rewritten to

$$\hat{\mathbf{x}}(t) = \bar{\mathbf{x}}(t) + V_{\bar{\mathbf{x}}}\mathbf{H}'^H\left(\sigma_n^2\mathbf{I}_{N_r} + \mathbf{H}'V_{\bar{\mathbf{x}}}\mathbf{H}'^H\right)^{-1}(\mathbf{y}_t - \mathbf{H}'\bar{\mathbf{x}}(t)).$$

From (9), we get $u_{i,t} = x_{i,t} + n_{i,t}^*$, and

$$\begin{aligned} n_{i,t}^* = \frac{v_i}{v_{\hat{x}_i}\phi_i}\mathbf{h}_i'^H\left(\sigma_n^2\mathbf{I}_{N_r} + \mathbf{H}'\mathbf{V}_{\bar{\mathbf{x}}}\mathbf{H}'^H\right)^{-1} \cdot \\ \left[\mathbf{H}'\left(\mathbf{x}_{\backslash i}(t) - \bar{\mathbf{x}}_{\backslash i}(t)\right) + \mathbf{n}(t)\right], \end{aligned} \tag{48}$$

where $\mathbf{x}_{\backslash i}(t)$ (or $\bar{\mathbf{x}}_{\backslash i}(t)$) denotes the vector whose ith entry of $\mathbf{x}(t)$ (or $\bar{\mathbf{x}}(t)$) is set to zero. The equivalent noise $n_{i,t}^*$ is independent of $x_{i,t}$. In Eq. (21) of [64] and Theorem 4(b) of [65], a rigorous proof is elaborated to show that $n_{i,t}^*$ is Gaussian distributed, i.e., $n_{i,t} \sim \mathcal{CN}\left(0, 1/\phi_i(\mathbf{v}_{\bar{\mathbf{x}}})\right)$. Hence, we obtain the proposition.

APPENDIX C
PROOF OF LEMMA 1

From (19), the achievable rate of user i is given by

$$\begin{aligned} R_i &= \int_0^\infty \left(\rho_i + \psi_i(\rho_i)^{-1}\right)^{-1} d\rho_i \\ &\overset{(a)}{\leq} \int_{\phi_i(1)}^{\phi_i(0)} \left[\rho_i + \left(\phi_i{}^{-1}(\rho_i)\right)^{-1}\right]^{-1} d\rho_i \\ &\quad + \int_0^{\phi_i(1)} (1+\rho_i)^{-1} d\rho_i \\ &\overset{(b)}{=} \int_{v_i=1}^{v_i=0} \left(v_i^{-1} + \phi_i(v_i)\right)^{-1} d\phi_i(v_i) + \log\left(1 + \phi_i(v_i)\right) \\ &\overset{(c)}{=} \int_{v_i=1}^{v_i=0} v_{\hat{x}_i}(v_i) dv_{\hat{x}_i}(v_i)^{-1} - \int_{v_i=1}^{v_i=0} v_{\hat{x}_i}(v_i) dv_i^{-1} \\ &\quad - \log v_{\hat{x}_i}(v_i = 1) \\ &\overset{(d)}{=} -\int_{v_1=1}^{v_1=0} \gamma_i^{-1}\left[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\right]_{i,i} dv_1^{-1} - \lim_{v_1 \to 0} \log\left[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\right]_{i,i} \\ &\overset{(e)}{=} \int_{v_1=1}^{v_1=0} \left[v_1 - \gamma_i^{-1}\left[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\right]_{i,i}\right] dv_1^{-1} - \log(\gamma_i). \end{aligned} \tag{49}$$

The inequality (a) is derived by (26)∼(28) and the equality holds if and only if there exists such a code whose transfer function satisfies the matching conditions. The equations $(b) \sim (d)$ are given by $\rho_i = \phi_i(v_i)$, (24) and (25), equation (e) comes from (22) and (23). In Appendix D, we show the existence of such codes whose SINR-variance transfer functions match that of the LMMSE detector. In Appendix E, the existence of the infinite integral of (29) is proven.

APPENDIX D
THE CODE EXISTENCE IN LEMMA 1

We first introduce an important property that is established in [43], which builds the relationship between the code rate and its transfer function $\psi_i(\rho_i)$.

Property of SCM Code: Assume $\psi(\rho)$ satisfies

(i) $\psi(0) = 1$ and $\psi(\rho) \geq 0$, for $\rho \in [0, \infty)$;
(ii) monotonically decreasing in $\rho \in [0, \infty)$;
(iii) continuous and differentiable in $[0, \infty)$ except for a countable set of values of ρ;
(iv) $\lim_{\rho\to\infty} \rho\psi(\rho) = 0$.

Let Γ_n be an n-layer SCM code with *SINR-variance* transfer function $\psi^n(\rho)$ and rate R_n. Then, there exists $\{\Gamma_n\}$ such that: (i) $\psi^n(\rho) \leq \psi(\rho), \forall\rho \geq 0, \forall n$; (ii), $R_n \to R\left(\psi(\rho)\right)$ as $n \to \infty$, where $R\left(\psi(\rho)\right)$ denotes code rate of transfer function $\psi(\rho)$.

This property means that there exists such an n-layer SCM code Γ_n whose transfer function can approach $\psi(\rho)$ that satisfies the conditions (i)$\sim$(iv) with arbitrary small error when n is large enough.

From the "*Property of SCM Codes*", we can see that there exist such n-layer SCM codes whose transfer function satisfies (i)$\sim$(iv) when n is large enough. Therefore, it only needs to check the matched transfer function meets the conditions (i)$\sim$(iv) in order to show the existence of such codes. It is easy to see that conditions (i) and (iv) are always satisfied by (26) and (27) respectively. From (24)$\sim$(28), we can see that $\psi_i(\rho_i)$ is continuous and differentiable in $[0,\infty)$ except at $\rho_i=\phi_i(0)$ and $\rho_i=\phi_i(1)$. Thus, Condition (iii) is satisfied. To show the monotonicity of the transfer function, we first rewrite (31) by the random matrix theorem as

$$\begin{aligned}\phi_i(v_i)&=\left[v_i-v_i^2\frac{w^2}{\sigma_n^2}\mathbf{h}_i^H\left(\mathbf{I}_{N_r}+\frac{w^2v_i}{\sigma_n^2}\mathbf{H}\mathbf{H}^H\right)^{-1}\mathbf{h}_i\right]^{-1}-v_i^{-1}\\&=\left[\left(\frac{w^2}{\sigma_n^2}\mathbf{h}_i^H\left(v_i^{-1}\mathbf{I}_{N_r}+\frac{w^2}{\sigma_n^2}\mathbf{H}\mathbf{H}^H\right)^{-1}\mathbf{h}_i\right)^{-1}-1\right]^{-1}\\&=1/\big(f_i^{-1}(v_i)-1\big),\end{aligned}\tag{50}$$

where $f_i(v_i)=\frac{w^2}{\sigma_n^2}\mathbf{h}_i^H\big(v_i^{-1}\mathbf{I}_{N_r}+\frac{w^2}{\sigma_n^2}\mathbf{H}\mathbf{H}^H\big)^{-1}\mathbf{h}_i$. It is easy to check that $f_i(v_i)$ is a decreasing function with respect to v_i, and $\phi_i(v_i)$ is thus a decreasing function of v. With the definition of $\psi(\rho)$ from (26)$\sim$(28), we then see that $\psi_i(\rho_i)$ is a decreasing function in $[0,\infty)$. Therefore, the matched transfer function can be obtained by the SCM code, i.e., there exists such codes that satisfy the matching conditions.

Appendix E
The Existence of Infinite Integral (29)

With (29), we have

$$\begin{aligned}R_i&=-\int_{v_1=1}^{v_1=0}\gamma_i^{-1}[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)]_{i,i}dv_1^{-1}-\lim_{v_1\to0}\ \log(\gamma_iv_1)\\&\overset{(a)}{=}-\int_0^\infty\left[(\mathbf{A}_{\boldsymbol{\gamma}}+s\mathbf{I}_{N_u})^{-1}\right]_{i,i}ds-\lim_{s\to\infty}\ \log(\gamma_is^{-1})\\&\overset{(b)}{=}-\int_0^\infty\mathbf{u}_i{}^H\left(\boldsymbol{\Lambda}_{A_{\boldsymbol{\gamma}}}+s\mathbf{I}_{N_u}\right)^{-1}\mathbf{u}_ids-\lim_{s\to\infty}\ \log(\gamma_is^{-1}),\\&\overset{(c)}{=}-\int_0^\infty\sum_{j=1}^{N_u}\|u_{ij}\|^2\left(\lambda_{\mathbf{A}_{\boldsymbol{\gamma}},j}+s\right)^{-1}ds-\lim_{s\to\infty}\ \log(\gamma_is^{-1})\\&=\sum_{j=1}^{N_u}\|u_{ij}\|^2\log\left(\lambda_{\mathbf{A}_{\boldsymbol{\gamma}},j}\right)-\log(\gamma_i),\end{aligned}\tag{51}$$

where equation (a) comes from $s=v_1^{-1}$ and $\mathbf{A}_{\boldsymbol{\gamma}}=\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}^{1/2}(\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'+\mathbf{I}_{N_u})\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}^{1/2}$; equation (b) is based on $\mathbf{A}_{\boldsymbol{\gamma}}=\mathbf{U}^H\boldsymbol{\Lambda}_{\mathbf{A}_{\boldsymbol{\gamma}}}\mathbf{U}$ and $\mathbf{u}_i$ is the ith column of $\mathbf{U}$; $\lambda_{\mathbf{A}_{\boldsymbol{\gamma}},j}$ is the ith diagonal element of $\boldsymbol{\Lambda}_{A_{\boldsymbol{\gamma}}}$. Thus, we show the existence of the infinite integral (49), i.e., R_i has finite value.

Appendix F
Proof of Theorem 1

With (29), the achievable sum rate is

$$\begin{aligned}R_{sum}&=\sum_{i=1}^{N_u}R_i\\&\overset{(a)}{=}-\int_{v_1=1}^{v_1=0}\sum_{i=1}^{N_u}\left(\gamma_i^{-1}\left[\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\right]_{i,i}\right)dv_1^{-1}\\&\quad-\lim_{v_1\to0}\log\left(v_1^{N_u}\prod_{i=1}^{N_u}\gamma_i\right)\\&=-\int_{v_1=1}^{v_1=0}\mathrm{Tr}\{\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}^{-1}\mathbf{V}_{\hat{\mathbf{x}}}(v_1)\}dv_1^{-1}\\&\quad-\lim_{v_1\to0}\log\left(v_1^{N_u}\prod_{i=1}^{N_u}\gamma_i\right)\\&\overset{(b)}{=}-\lim_{v_1\to0}\log\left(v_1^{N_u}\prod_{i=1}^{N_u}\gamma_i\right)-\big[\log|(v_1^{-1}-1)\mathbf{I}_{N_u}\\&\quad+\left(\mathbf{I}_{N_u}+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'\right)\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}|\big]_{v_1=1}^{v_1=0}\\&=-\lim_{v_1\to0}\log\left(v_1^{N_u}\prod_{i=1}^{N_u}\gamma_i\right)-\lim_{v_1\to0}\log|v_1^{-1}\mathbf{I}_{N_u}|\\&\quad+\log|(\mathbf{I}_{N_u}+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}')\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}|\\&=\log|\mathbf{I}_{N_u}+\sigma_n^{-2}\mathbf{H}'^H\mathbf{H}'|,\end{aligned}$$

which is the exact system sum capacity of MIMO-NOMA system. Equation (a) is derived by (29), and equation (b) is based on (23) and the law $\int\mathrm{Tr}\{(s\mathbf{I}+\mathbf{A})^{-1}\}ds=\log|s\mathbf{I}+\mathbf{A}|$. It means iterative LMMSE detection is sum capacity-achieving.

Appendix G
Capacity Region Domination Lemma

The following lemma is used of the proofs in the rate analyses of iterative LMMSE detection.

Capacity Region Domination Lemma [52]: All the points in the capacity region $\mathcal{R}_\mathcal{S}$ is dominated by a convex combination of the following $(N_u!)$ ***maximal extreme points***.

$$\begin{cases}R_{k_1}=\log\frac{|\mathbf{I}_{N_u}+\frac{1}{\sigma_n^2}\mathbf{H}'^H\mathbf{H}'|}{|\mathbf{I}_{|\mathcal{S}_1^c|}+\frac{1}{\sigma_n^2}\mathbf{H}'^H_{\mathcal{S}_1^c}\mathbf{H}'_{\mathcal{S}_1^c}|},\\ \qquad\vdots\\ R_{k_{N_u-1}}=\log\frac{|\mathbf{I}_{|\mathcal{S}_{N_u-2}^c|}+\frac{1}{\sigma_n^2}\mathbf{H}'^H_{\mathcal{S}_{N_u-2}^c}\mathbf{H}'_{\mathcal{S}_{N_u-2}^c}|}{|1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_{k_{N_u}}\mathbf{h}'_{k_{N_u}}|},\\ R_{k_{N_u}}=\log\left(1+\frac{1}{\sigma_n^2}\mathbf{h}'^H_{k_{N_u}}\mathbf{h}'_{k_{N_u}}\right),\end{cases}\tag{52}$$

where $(k_1,\ldots,k_{N_u})$ is a permutation of $(1,2,\ldots,N_u)$, $\mathcal{S}_i=\{k_1,\ldots,k_i\}$ for $i=1,\ldots,N_u-1$.

Appendix H
Proof of Corollary 3

For any maximal extreme point expressed in (52) with order vector $[k_1,\ldots,k_{N_u}]$, we let $\gamma_{k_i}/\gamma_{k_{i-1}}\to\infty$, for any

$i \in \mathcal{N}_u/\{1\}$. Therefore, similar to the green curves showed in Fig. 3 and Fig. 4, the user k_{N_u} is recovered after all the variances of other users already being zeros as $\gamma_{k_{N_u}}/\gamma_{k_{i-1}} \to \infty$, for any $i \in \mathcal{N}_u/\{1\}$. Thus, from (29), the rate of user k_{N_u} is

$$R_{k_{N_u}} = \log\left(1 + \frac{1}{\sigma_n^2}\mathbf{h'}_{k_{N_u}}^H \mathbf{h'}_{k_{N_u}}\right), \tag{53}$$

which is the same as that in (52). Similarly, when we recovering user k_{N_u-1}, all the users have been recovered except user k_{N_u} and user $k_{N_u} - 1$. Hence, based on Theorem 1, we have

$$R_{k_{N_u-1}} + R_{k_{N_u}} = \log\left|\mathbf{I}_{|\mathcal{S}^c_{N_u-2}|} + \frac{1}{\sigma_n^2}\mathbf{H'}^H_{\mathcal{S}^c_{N_u-2}}\mathbf{H'}_{\mathcal{S}^c_{N_u-2}}\right|. \tag{54}$$

Thus, the rate of user k_{N_u-1} is

$$R_{k_{N_u-1}} = \log\frac{|\mathbf{I}_{|\mathcal{S}^c_{N_u-2}|} + \frac{1}{\sigma_n^2}\mathbf{H'}^H_{\mathcal{S}^c_{N_u-2}}\mathbf{H'}_{\mathcal{S}^c_{N_u-2}}|}{1 + \frac{1}{\sigma_n^2}\mathbf{h'}_{k_{N_u}}^H \mathbf{h'}_{k_{N_u}}}, \tag{55}$$

which is the same as that in (52). Continue this process and we can show all the other users' rates are the same as that of in (52). Therefore, we have Corollary 3.

Acknowledgment

The authors would like to thank Prof. L. Ping for fruitful discussions.

References

[1] D. Argas, D. Gozalvez, D. Gomez-Barquero, and N. Cardona, "MIMO for DVB-NGH, the next generation mobile TV broadcasting," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 130–137, Jul. 2013.

[2] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[4] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[5] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and Y. Su, "Convergence analysis and assurance Gaussian message passing iterative detection for massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6487–6501, Sep. 2016.

[6] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and Y. Su, "A low-complexity Gaussian message passing iterative detection for massive MU-MIMO systems," in *Proc. IEEE 10th Int. Conf. Inf., Commun. Signal Process.*, Singapore, Dec. 2015, pp. 1–5.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, U.K.: Cambridge Univ. Press, 2005.

[8] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[9] METIS, "Proposed solutions for new radio access," *Mobile and Wireless Communications Enablers for the 2020 Information Society*, Rep. TR.-ICT-317669-METIS/D2.4, Feb. 2015.

[10] "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G Whitepaper, Jul. 2014.

[11] B. Kim and W. Chung, "Uplink NOMA with Multi-Antenna," in *Proc. IEEE Veh. Technol. Conf. Spring*, Scotland, U.K., 2015.

[12] S. Chen, K. Peng, and H. Jin, "A suboptimal scheme for uplink NOMA in 5G systems," *IEEE Int. Wireless Commun. Mobile Comput. Conf.*, Aug. 2015, pp. 1429–1434.

[13] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst.*, Barcelona, Aug. 2014, pp. 781–785.

[14] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[15] G. Liu, X. Chen, Z. Ding, Z. Ma, and F. R. Yu, "Hybrid half-duplex/full-duplex cooperative non-orthogonal multiple access with transmit power adaptation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 506–519, Jan. 2018.

[16] B. Di, L. Song, and Y. Li, "Trellis coded modulation for non-orthogonal multiple access systems: Design, challenges, and opportunities," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 68–74, Apr. 2018.

[17] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[18] L. Liu, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving iterative LMMSE detector for MIMO-NOMA systems," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[19] C. Xu, Y. Hu, C. Liang, J. Ma, and L. Ping, "Massive MIMO, non-orthogonal multiple access and interleave division multiple access," *IEEE Access*, vol. 5, pp. 14728–14748, 2017.

[20] Y. Chi, L. Liu, G. Song, C. Yuen, Y. L. Guan, and Y. Li, "Practical MIMO-NOMA: low complexity and capacity-approaching solution," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6251–6264, Sep. 2018.

[21] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[22] H. Wang, R. Zhang, R. Song, and S. Leung, "A novel power minimization precoding scheme for MIMO-NOMA uplink systems," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1106–1109, May 2018.

[23] M. Jiang, Y. Li, Q. Zhang, Q. Li, and J. Qin, "MIMO beamforming design in nonorthogonal multiple access downlink interference channels," *IEEE Trans. Veh. Techn.*, vol. 67, no. 8, pp. 6951–6959, Aug. 2018.

[24] D. Micciancio, "The hardness of the closest vector problem with preprocessing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1212–1215, Mar. 2001.

[25] S. Verdú, "Optimum multi-user signal detection," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Illinois Urbana-Champaign, Urbana, IL, USA, Aug. 1984.

[26] S. Verdú and H. V. Poor, "Abstract dynamic programming models under commutativity conditions," *SIAM J. Control Optim.*, vol. 25, no. 4, pp. 990–1006, Jul. 1987.

[27] H. A. Loeliger, J. Hu, S. Korl, Q. Guo, and L. Ping, "Gaussian message passing on linear models: An update," in *Proc. Int. Symp. Turbo Codes Related Topics*, Apr. 2006, pp. 1–7.

[28] O. Axelsson, *Iterative Solution Methods.* Cambridge, U.K.: Cambridge Univ. Press, 1994.

[29] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Calculation: Numerical Methods.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[30] X. Gao, L. Dai, C. Yuen, and Y. Zhang, "Low-complexity MMSE signal detection based on Richardson method for large-scale MIMO systems," in *Proc. IEEE 80th Veh. Technol. Conf.*, Sep. 2014, pp. 1–5.

[31] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing iterative detection for MIMO-NOMA systems with massive access," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[32] A. Montanari, B. Prabhakar, and D. Tse, "Belief propagation based multiuser detection," in *Proc. Allerton Conf. Comm. Control Comp.*, vol. 43, Sep. 2005.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory-Second Edition*. New York, NY, USA: Wiley, 2006.

[34] A. E. Gamal and Young-Han Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, January 2012.

[35] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[36] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST spacetime communication architecture," *Electron. Lett.*, vol. 35, no. 1, pp. 14–16, Jan. 1999.

[37] X. Wang and H. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, Jul. 1999.

[38] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, Jul. 2011.

[39] L. Ping, L. Liu, K. Y. Wu, and W. K. Leung, "Interleave-division multiple-access (IDMA) communications," in *Proc. Int. Symp. Turbo Codes Related Topics*, Brest, France, Sep. 2003, pp. 173–180.

[40] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006.

[41] Q. Guo and L. Ping, "LMMSE turbo equalization based on factor graphs," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 311–319, Feb. 2008.

[42] A. Sanderovich, M. Peleg, and S. Shamai, "LDPC coded MIMO multiple access with iterative joint decoding," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1437–1450, Apr. 2005.

[43] X. Yuan, L. Ping, C. Xu, and A. Kavcic, "Achievable rates of MIMO systems with linear precoding and iterative LMMSE detector," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7073–7089, Oct. 2014.

[44] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, Jan. 1981.

[45] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[46] A. Ashikhmin, G. Kramer, and S. ten Brink, "Extrinsic information transfer functions: Model and erasure channel properties," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2657–2673, Nov. 2004.

[47] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[48] K. Bhattad and K. R. Narayanan, "An MSE-based transfer chart for analyzing iterative decoding schemes using a Gaussian approximation," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 22–38, Jan. 2007.

[49] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[50] K. S. Andrews, D. Divsalar, S. Dolinar, J. Hamkins, C. R. Jones, and F. Pollara, "The development of Turbo and LDPC codes for deep space applications," *Proc. IEEE*, vol. 95, no. 11, pp. 2142–2156, Nov. 2007.

[51] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[52] T. S. Han, "The capacity region of general multiple-access channel with certain correlated sources," *Inf. Control*, vol. 40, no. 1, pp. 37–60, 1979.

[53] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[54] H. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 835–847, May 1997.

[55] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.

[56] D. G. Brennan, "Linear diversity combining techniques," *Proc. IEEE*, vol. 47, no. 6, pp. 1075–1102, Jun. 1959.

[57] Y. Hu, C. Liang, L. Liu, C. Yan, Y. Yuan, and L. Ping, "Interleave-division multiple access in high rate applications," *IEEE Wireless Commun. Lett.*, doi: 10.1109/LWC.2018.2876538.

[58] J. Song and Y. Zhang, "On construction of rate-compatible raptor-like QC-LDPC code for enhanced IDMA in 5G and beyond," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process.*, Hong Kong, Dec 2018, pp. 1–5.

[59] Y. Zhang, K. Peng, X. Wang, and J. Song, "Performance analysis and code optimization of IDMA with 5G new radio LDPC code," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1552–1555, Aug. 2018.

[60] X. Wang, S. Cammerer, and S. Brink, "Near Gaussian multiple access channel capacity detection and decoding," in *Proc. 10th IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process.*, Hong Kong, Dec. 2018, pp. 1–5.

[61] G. Song and J. Cheng, "Low-complexity coding scheme to approach multiple-access channel capacity," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 2106–2110.

[62] G. Song, X. Wang, and J. Cheng, " A low-complexity multiuser coding scheme with near-capacity performance," *IEEE Trans. Veh. Techn.*, vol. 66, no. 8, pp. 6775–6786, Aug. 2017.

[63] G. Song, J. Cheng, and Y. Watanabe, "Maximum sum rate of repeat-accumulate interleave-division system by fixed-point analysis," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 3011–3022, Oct. 2012.

[64] S. Rangan, P. Schniter, and A. Fletcher, "Vector approximate message passing," arXiv:1610.03082, 2016.

[65] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," arXiv:1701.05284, 2017.

Lei Liu (M'17) received the Ph.D. degree in communication and information system from Xidian University, Xi'an, China, in 2017. He was an exchange Ph.D. student with the Nanyang Technological University, Singapore. From 2016 to 2017, he was a Research Assistant with the Singapore University of Technology and Design (SUTD), Singapore. From March 2017 to July 2017, he was a Postdoctoral Research Fellow with the SUTD, Singapore. He is currently a Postdoctoral fellow with the City University of Hong Kong, Hong Kong. His current research interests include message passing, massive MIMO, NOMA, information theory, compressed sensing, and channel coding. He received the Ph.D. National Scholarship in China in 2015, and the State Scholarship Fund from China Scholarship Council from 2014 to 2016.

Yuhao Chi received the B.S. degree in electronic and information engineering from Shaanxi University of Science and Technology, Xi'an, China, in 2012, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2018. From 2016 to 2017, he received the state scholarship fund from China scholarship council to be an exchange Ph.D. student with Nanyang Technological University, Singapore, and a visiting student with the Singapore University of Technology and Design, Singapore. His research interests include coding theory, multiuser coding and detection, message passing algorithm, and deep learning.

Chau Yuen (SM'12) received the B.Eng. and Ph.D. degree from the Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He is the recipient of the Lee Kuan Yew Gold Medal, Institution of Electrical Engineers Book Prize, Institute of Engineering of Singapore Gold Medal, Merck Sharp and Dohme Gold Medal and twice the recipient of Hewlett Packard Prize. He was a Postdoctoral fellow with the Lucent Technologies Bell Labs, Murray Hill in 2005. He was a Visiting Assistant Professor of Hong Kong Polytechnic University in 2008. From 2006 to 2010, he was with the Institute for Infocomm Research (I2R, Singapore) as a Senior Research Engineer, where he was involved in an industrial project on developing an 802.11n Wireless LAN system, and participated actively in 3Gpp Long Term Evolution (LTE) and LTE− Advanced (LTE− A) standardization. He joined the Singapore University of Technology and Design as an Assistant Professor in June 2010, and received IEEE Asia-Pacific Outstanding Young Researcher Award in 2012.

Dr. Yuen is currently an Editor for the IEEE TRANSACTION ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and awarded as Top Associate Editor from 2009 to 2015.

Yong Liang Guan (SM'17) received the Bachelor of Engineering degree (first class Hons.) from the National University of Singapore, Singapore, and the Ph.D. degree from the Imperial College of London, London, U.K. He is a tenured Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests broadly include modulation, coding and signal processing for communication systems and data storage systems. He is an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. His homepage is at http://www3.ntu.edu.sg/home/eylguan.

Ying Li (M'08) received the B.S. degree in telecommunication engineering and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 1995 and 2005, respectively.

From 2011 to 2012, she was with the University of California, Davis, Davis, CA, USA, as a visiting scholar. She is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. His current research interests include on design and analysis for wireless systems, including channel coding, wireless network communications, interference processing, relay transmission and MIMO techniques.