Constrained Capacity Optimal Generalized Multi-User MIMO: A Theoretical and Practical Framework

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

Abstract—Conventional multi-user multiple-input multiple-output (MU-MIMO) mainly focused on Gaussian signaling, independent and identically distributed (IID) channels, and a limited number of users. It will be laborious to cope with the heterogeneous requirements in next-generation wireless communications, such as various transmission data, complicated communication scenarios, and unprecedented massive user access. Therefore, this paper studies a generalized MU-MIMO (GMU-MIMO) system with more generalized and practical constraints, i.e., practical channel coding, non-Gaussian signaling, right-unitarily-invariant channels (covering Rayleigh fading channel matrices, certain ill-conditioned and correlated channel matrices, etc.), and massive users and antennas. These generalized assumptions bring new challenges in theory and practice. For example, there is no accurate constrained capacity region analysis for GMU-MIMO. In addition, it is unclear how to achieve constrained-capacity-optimal performance with practical complexity. To address these challenges, a unified framework is proposed to derive the constrained capacity region of GMU-MIMO and design a constrained-capacity-optimal transceiver, which jointly considers encoding, modulation, detection, and decoding. Group asymmetry is developed to group users according to their rates, which makes a tradeoff between user rate allocation and implementation complexity. Specifically, the constrained capacity region of group-asymmetric GMU-MIMO is characterized by using the minimum mean-square error (MMSE) optimality of orthogonal/vector approximate message passing (OAMP/VAMP) and the relationship between mutual information and MMSE. Furthermore, a theoretically optimal multi-user OAMP/VAMP receiver and practical multi-user low-density parity-check (MU-LDPC) codes are proposed to achieve the constrained capacity region of group-asymmetric GMU-MIMO. Numerical results demonstrate that the proposed MU-LDPC coded GMU-MIMO systems achieve asymptotic performance within 0.2 dB from the theoretical sum capacity. Moreover, their finite-length performances are about 1~2 dB away from the associated sum capacity of GMU-MIMO.

Index Terms—Generalized multi-user MIMO (GMU-MIMO), right-unitarily-invariant channel matrices, arbitrary signal distributions, constrained channel capacity region, capacity optimal and practical framework, orthogonal/vector approximate message passing (OAMP/VAMP), multi-user LDPC codes.

I. INTRODUCTION

WITH the rapid development of wireless communications, a variety of communication services have emerged, such as heterogeneous vehicular networks [2], in which the number of wireless-enabled devices is predicted to reach 10 million connections per square kilometer in 6G [3]. To support large-scale connectivity, multi-user multiple-input multiple-output (MU-MIMO) [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] is a popular technology that can provide lots of space resources for massive data transmissions. However, due to various data, complex channels, and large numbers of users and antennas, the communication scenarios become more complex, which brings many new challenges to MU-MIMO not only in theory but also in practice.

A. Information-Theoretical Limit of MU-MIMO

A common information-theoretical limit that is frequently used in communication systems is channel capacity. It is well known that capacity is defined by default as the maximum mutual information over all possible choices of the input distribution. We employ a constrained capacity for GMU-MIMO, which is defined as mutual information under a fixed input distribution, due to the arbitrarily fixed input distribution constraint. It should be emphasized that the Gaussian capacity, i.e. Shannon capacity in the strict sense for additive white
Gaussian noise (AWGN) channel, is a special case of the constrained capacity when the signaling is Gaussian.

For available channel state information (CSI) at the transceiver, a kind of water-filling technique was proposed to obtain the Gaussian capacity for Gaussian signaling [6], [8] and constrained capacity of MU-MIMO with arbitrarily distributed input signaling [7]. When CSI is only available at the receiver, the capacity of Gaussian MU-MIMO was approached with Gaussian signaling [14]. For arbitrarily distributed input signaling, only the constrained capacities of point-to-point MIMO (P2P-MIMO) were derived with independent and identically distributed (IID) channel matrices [15], [16] or right-unitarily-invariant channel matrices [17], [18], leveraging random matrix theory [15], [16], [17] and approximate message passing (AMP)-type algorithms [18], respectively. However, the constrained capacity region of MU-MIMO with arbitrarily distributed input signaling and right-unitarily-invariant channel matrices is still unknown.

B. Practical Information-Theoretically Optimal Receivers of MU-MIMO

It was proved that the capacity region of MU-MIMO can be achieved by a successive interference cancellation (SIC) receiver with time-sharing technology [19]. However, the following inherent problems make the SIC receiver impractical in large-scale systems: 1) severe decoding delay, 2) excessive overhead, 3) serious error propagation, and 4) prohibited complexity of optimizations in user decoding order and grouping strategy. To solve the issues of the SIC receiver, a lot of literature focused on parallel interference cancellation (PIC) receivers for MU-MIMO [10], [11], [12], [13].

For CSI only available at the receiver, when properly designed forward error correction (FEC) codes, optimality of the iterative linear minimum mean-square error (Turbo-LMMSE) receiver was proved to achieve the sum capacity of MU-MIMO with Gaussian signaling [10], [11]. To reduce the implementation complexity of LMMSE, Gaussian message passing receivers were proposed in [13] and [12]. Nevertheless, Gaussian signaling is only an ideal concept. In practice, non-Gaussian discrete signaling is generally used, such as quadrature phase-shift keying (QPSK) and quadrature amplitude modulation (QAM). In this case, these Turbo-type receivers [10], [11], [12], [13] are not capacity optimal anymore [18], [20], [21].

To address the above issue, AMP with well-designed FEC codes can achieve the constrained capacity of P2P-MIMO with IID channel matrices and arbitrary input signaling [21]. Moreover, for discrete signaling, the achievable rate of AMP was shown to be higher than that of Turbo-LMMSE. However, AMP is limited to IID channel matrices [20]. For non-IID channels, AMP performs poorly or even diverges [22], [23], [24], such that the results in [21] will become invalid.

To overcome the limitation of AMP on non-IID channels, orthogonal AMP (OAMP) [25] and vector AMP (VAMP) [26] were proposed to offer improved performances in a wide range of right-unitarily-invariant matrices. Compared with AMP, OAMP and VAMP can be applied to more complex and practical communication scenarios, i.e., right-unitarily-invariant matrices, covering certain ill-conditioned and correlated channel matrices [20], [27]. In [18], it provided the rigorously capacity optimality proof for OAMP in P2P coded linear (i.e., MIMO) systems with right-unitarily-invariant matrices and arbitrary input distributions. Furthermore, it was shown that OAMP outperforms AMP in un-coded physical random access channels [28], generalized frequency division multiplexing (GFDM) [29], extra-large-scale massive MIMO [30], and the conventional Turbo receivers in P2P-MIMO channels [18], [20]. Note that the optimal coding design is not investigated in works [28], [29], [30], making it difficult for OAMP to achieve constrained-capacity-optimal performance. Moreover, due to the equivalence of OAMP and VAMP, they are referred to as OAMP/VAMP in this paper.

In a nutshell, current results are restricted to Gaussian signaling [10], [11], [12], [13], IID channels [21], un-coded systems [28], [29], [30], or P2P channels [18], [20].
is presented for MU-OAMP/VAMP to achieve the constrained capacity region of group-asymmetric GMU-MIMO. Moreover, a kind of multi-user low-density parity-check (MU-LDPC) code is designed for MU-OAMP/VAMP.

The major contributions of this paper are summarized as follows.

1) Group-asymmetric GMU-MIMO is developed and its constrained capacity region is characterized accurately.

2) The achievable sum rate and constrained-sum-capacity optimality of MU-OAMP/VAMP are analyzed and proved, based on which the optimal design principle of multi-user codes is presented for GMU-MIMO.

3) A kind of capacity-approaching MU-LDPC code is designed for MU-OAMP/VAMP, whose theoretical detection thresholds are about 0.2 dB away from the constrained sum capacity.

4) Numerical results show that the finite-length performances of the proposed framework with optimized MU-LDPC codes and QPSK modulation are about 1–2 dB from the associated sum capacity. They also outperform the existing state-of-art methods such as Turbo-LMMSE with optimized LDPC codes, and OAMP/VAMP with well-designed irregular P2P-LDPC codes.

In summary, this is the first work to provide a constrained-capacity-optimal framework for GMU-MIMO with practical complexity, which is also the first time to apply the OAMP/VAMP to GMU-MIMO. The main theoretical results of this paper, such as the characterization of the constrained capacity region, the optimal design principle of multi-user codes, and the constrained-capacity optimality proof of MU-OAMP/VAMP, are available for all signal constellations.

Part of the results in this paper has been published in [1]. In this paper, we additionally provide the derivation of the constrained capacity region, detailed proofs, and more numerical results.

D. Connection to Existing Works

1) Relationship With EP: In fact, OAMP/VAMP is algorithmically equivalent to expectation propagation (EP) [35], [36], [37]. For simplicity, we focus on OAMP/VAMP in this paper.

2) Other Related Low-Complexity AMP-Type Algorithms: Recently, to avoid the high complexity LMMSE in OAMP/VAMP, low-complexity Bayes-optimal convolutional AMP (CAMP) [38], memory AMP (MAMP) [39], and generalized MAMP (GMAMP) [40] were proposed for right-unitarily-invariant matrices with arbitrary input distributions. Therefore, they may be good candidates with lower complexity for the proposed framework in this paper.

3) Differences From Turbo-Based MU-MIMO [10], [11], [12], [13]: Although there is a vast amount of works on MU-MIMO, most of their receivers are based on the conventional Turbo receiver [10], [11], [12], [13], which has been proven to be constrained-capacity-optimal for MU-MIMO systems with Gaussian signaling [10]. In this paper, we consider GMU-MIMO with arbitrarily distributed signaling such as Gaussian, QPSK, QAM, etc. We will show that for non-Gaussian signaling, the Turbo-LMMSE is rigorously sub-optimal and worse than the proposed OAMP/VAMP receiver in Section IV-C. Moreover, note that the coding rates of well-designed multi-user codes in [10] and [11] are as low as 0.1. In contrast, based on the OAMP/VAMP receiver, the proposed multi-user codes can support a larger rate range, such as 0.33 ~ 0.67 in Section VI-C. See numerical results for more details.

4) Differences from P2P-MIMO in [21] and [18]: In [21] and [18], AMP and OAMP were proven to be constrained-capacity-optimal for MIMO with arbitrary input signaling, but they are limited to P2P-MIMO that only involve one user with one rate. In contrast, This paper studies a more complicated GMU-MIMO that involves multiple users with different rates. As a result, the constrained-capacity region of GMU-MIMO needs to be established and the design of a constrained-capacity-optimal receiver becomes much more difficult. Therefore, the results in P2P-MIMO [18], [21] cannot be straightforwardly applied to the constrained capacity region analysis and constrained-capacity-optimal receiver design for GMU-MIMO.

5) Bayes Optimality vs Capacity Optimality: MMSE measurement is widely used in signal processing and detection of un-coded systems, which cannot characterize error-free recovery. As a result, MMSE is commonly used to denote the Bayes optimality (MMSE optimality) of an un-coded system. In contrast, the achievable rate is a key measurement for coded systems with error-free recovery. Thus, the maximum achievable rate is commonly used to denote the capacity optimality of a coded system. It is worth noting that Bayesian optimality does not guarantee capacity optimality. This is demonstrated in the numerical results of this paper, where the performances of Bayes-optimal detectors with P2P capacity-approaching LDPC codes are rigorously sub-optimal.

To show the difference between this paper and the existing closely related works, Table I presents the corresponding comparisons. In addition, a list of key abbreviations mainly used throughout the paper is summarized in Table II.

E. Notations

Lowercase letters denote scalars and boldface lowercase letters denote vectors. \([\cdot]^T\) and \([\cdot]^H\) denote transpose and conjugate transpose operations respectively. \(\mathbb{C}\) represents the complex field. \(I\) is the identity matrix. \(I(x, y)\) denotes mutual information between \(x\) and \(y\). \(|S|\) is the cardinality of set \(S\). \(\text{Tr}(A)\) and \(\det(A)\) for the trace and the determinant of \(A\). \(\|a\|\) for the \(\ell_2\)-norm of the vector \(a\). \(E\{\cdot\}\) for the expectation over all random variables included in the brackets. \(E\{a|b\}\) for the conditional expectation of \(a\) for given \(b\). \(\text{mmse}\{a|b\}\) for \(E\{(a - E\{a|b\})^2|b\}\). \(CN(\mu, \Sigma)\) for the circularly-symmetric Gaussian distributions with mean \(\mu\) and covariance \(\Sigma\). The MMSE and constrained capacity of P2P-MIMO with IID Gaussian channel matrices [15], [16] or a sub-class of right-unitarily-invariant channel matrices [17] can be predicted by the replica method [25], [31], [41]. For more types of channel matrices, the rigorous proof of the replica method is still an open issue. In this paper, we assume that the replica method is
reliable and verify it with experimental results. For simplicity, MMSE and constrained capacity are not explicitly stated as the replica method in this paper.

F. Paper Outline

This paper is organized as follows. Section II presents the system model of GMU-MIMO. The constrained capacity region of group-asymmetric GMU-MIMO, achievable sum rate analysis of MU-OAMP/VAMP receiver, and the principle of multi-user code design are derived in Sections III and IV respectively. Section V presents an example of two user-group GMU-MIMO. Numerical simulations are provided in Section VI and the conclusion is presented in Section VII.

II. SYSTEM MODEL AND CHALLENGES

In this section, the model and assumptions of GMU-MIMO systems are provided. Meanwhile, the key challenges of constrained capacity region analysis and multi-user code design are presented.

A. System Model

Fig. 1 illustrates the uplink GMU-MIMO system with \( K \) transmit users and one receiver. Total of \( N \) transmit antennas are employed by the \( K \) users and each user has \( N_u = N/K \) antennas. The receiver has \( M \) receive antennas. Users are equally partitioned into \( G \) groups, where each group has \( K/G \) users. Let \( S_1, \ldots, S_G \) be the sets that include the user indices of the \( G \) groups, with \( S_i \cap S_j = \emptyset \) for \( i \neq j \), \( i, j \in G = \{1, \ldots, G\} \), and \( S_1 \cup S_2 \cdot \cdot \cdot \cup S_G = \mathbb{K} = \{1, \ldots, K\} \). Users in the same group employ the same encoder with the same code rate and that in different groups employ different encoders with different code rates.

Note: The proposed framework can be easily extended to the more general configurations, i.e., each group consists of a different number of users, and each user is equipped with a different amount of antennas. For more details, see Section IV-B.4.

At the transmission side, since the processing of each user’s data is similar, we describe the transmission of user \( i \in S_1 \). Message vector \( d_i \) is encoded by encoder 1 and the output codeword is denoted by \( x_i \). We assume \( x_i \) is a modulated...
signal vector whose entries are from a constellation set $S$. A serial-to-parallel conversion (S/P) is employed to produce the transmit signals over each antenna. Suppose the length of $x_i$ is $N_u L$ for a given integer $L$. Codeword $x_i$ is split into $N_u$ length-$L$ vectors $x_{i,n}, n = 1, \ldots, N_u$, and $x_{i,n} = [x_{i,n,1}, \ldots, x_{i,n,L}]$ is transmitted over antenna $n$. $L$ is the total transmission time for codeword $x_i$. At time $t$, the transmission signals of user $i$ over the $N_u$ antennas are $x^t_i = (x^t_{i,1}, \ldots, x^t_{i,N_u})$. The all transmitted signals of the constrained-capacity-optimal transceiver with practical complexity remains an open issue.

The receiver obtains signal $y^t = [y^t_1, \ldots, y^t_M]^T$, given by

$$\begin{align*}
y^t &= Ax^t + n^t,
&\quad t = 1, \ldots, L,
\end{align*}$$

where $A \in \mathbb{C}^{M \times N}$ is a channel matrix and $n^t \sim \mathcal{CN}(0, \sigma^2 I)$ is an AWGN vector. Without loss of generality, we assume $\frac{1}{N} \text{tr} \{ A^H A \} = 1$ and the signal-to-noise ratio (SNR) is defined as $\text{snr} = \sigma^{-2}$. Based on $y^t$, a multi-user decoder is employed to recover the $K$ users’ messages.

### B. Assumptions and Challenges of GMU-MIMO

#### 1) Assumptions

- A large-scale system is considered including massive users, massive transmitted and received antennas, i.e., $N \to \infty$, $M \to \infty$, and channel load $\beta = \frac{N}{M}$ is fixed.
- The entries of signal $x$ follow an arbitrary distribution $A$ (e.g., BPSK, QPSK, QAM, Gaussian, Bernoulli-Gaussian, etc.).
- Channel matrix $A$ is right-unitarily-invariant, which covers a variety of fading channel models including the commonly used IID random (i.e., Rayleigh fading) matrices and certain ill-conditioned (e.g., correlated) matrices [20], [27]. That is, let the SVD of $A$ be $A = U \Lambda V^H$, where $U \in \mathbb{C}^{M \times M}$ and $V \in \mathbb{C}^{N \times N}$ are unitary matrices, and

#### 2) Challenges

- For arbitrarily distributed $x$ and right-unitarily-invariant $A$, there is no accurate constrained capacity region analysis of the system in (1). The reason is that the existing capacity analyses for conventional MU-MIMO mainly focus on Gaussian signaling [10] or IID channel matrices [21].
- How to achieve the optimal performance of GMU-MIMO with practical complexity is still an open issue. The globally maximum a posteriori (MAP) receiver is the optimal solution [42], but it is unusable for large-scale systems due to its prohibitive complexity. Due to arbitrarily distributed $x$ and right-unitarily-invariant $A$, the existing practical capacity optimal receivers such as Turbo-LMMSE [10] and AMP [21] are sub-optimal.
- The design principle of practical multi-user codes for GMU-MIMO systems is still unclear, particularly for group-asymmetric cases. The conventional capacity-approaching P2P codes [43], [44], [45] are designed specially to overcome channel noises, which ignores the impact of large-scale antennas and users. In [9], [21], and [18], the channel codes are designed for one user in P2P-MIMO, which can not apply to group-asymmetric GMU-MIMO involving different requirements of users. Moreover, the existing multi-user channel codes [10],

---

$A$ is a rectangular diagonal matrix. $U$, $V$, and $\Lambda$ are mutually independent, and $V$ is Haar-distributed [37].

- Channel matrix $A$ is available at the receiver but unknown at the transmitters.$^2$
- The users are group-asymmetric, i.e., all users are partitioned into multiple groups, where user rates of the same group are the same and different groups may have different user rates.

---

$^1$For Gaussian signaling, the information-theoretic limit of the system in (1) is the well-known Gaussian sum capacity $C_{\text{Gau}} = \log |I + \text{snr} A^H A|$ [14], which can be achieved by Turbo-LMMSE [10]. However, for non-Gaussian signaling, the constrained capacity region of the system in (1) is not trivial, and the constrained-capacity-optimal transceiver with practical complexity remains an open issue.

$^2$In large-scale MU-MIMO, CSI at the transmitters is impractical as it brings a huge overhead cost. When CSI is unavailable at transmitters, the conventional precoding and water-filling power allocation are unavailable.

$^3$Due to the limitation of complexity, it is prohibited to design transceivers for a completely asymmetric system that all users may have different rates. Group asymmetry makes a good tradeoff between the system complexity and rate allocation.
III. MULTI-USER OAMP/VAMP RECEIVER AND CONSTRAINED CAPACITY REGION CHARACTERIZATION OF GMU-MIMO

In this section, we first present the multi-user OAMP/VAMP (MU-OAMP/VAMP) receiver for GMU-MIMO. Then, the constrained sum capacity and group capacity region for MU-OAMP/VAMP receiver for GMU-MIMO. Then, the transfer functions are consisted of SNR function \( \Omega_c(\rho) \) and the MMSE (per transmit antenna) functions are defined as

\[
\Omega_c(\rho) = \frac{1}{N} \text{Tr} \left[ |\text{snr}A^H A + \rho I|^{-1} \right], \\
\Omega_{\mathcal{C}}(\rho) = \frac{1}{G} \sum_{g=1}^{G} \Omega_c(\rho), \\
\Omega_{\mathcal{C}}^g(\rho) = \frac{1}{G} \mathbb{E} \left\{ \| x_g - \eta_{\text{mmse}}(\sqrt{\rho} x_g + z) \|^2 \right\},
\]

where \( x_g \) is the signal vector of the users in \( \mathcal{S}_g \) and \( z \sim \mathcal{C}\mathcal{N}(0, I) \) is independent of \( x_g \).

All the asymmetric information is averaged during the iteration by virtue of the right-unitarily-invariant property of \( A \) [46]. As a consequence, the signals input to LD and NLD are effectively averaged, considerably simplifying the design of the MU-OAMP/VAMP receiver. Meanwhile, the orthogonalization in (2) is necessary to make the input-output estimated errors of LD and NLD uncorrelated during the iteration. The orthogonalization is denoted as “Orth” in Fig. 2(a) and proved to ensure the exact state evolution of OAMP [20], [25].

State Evolution: LD and NLD are exactly characterized by the state evolution (SE) as shown in Fig. 2(b), in which the transfer functions are consisted of SNR function \( \psi_c(\rho) \) and MSE function \( \phi_c(\rho) \) as follows.

\[
\text{LD} : \quad \rho = \phi_c(\bar{\rho}) = \left[ \frac{1}{N} \| r - x \|^2 \right]^{-1}, \\
\text{NLD} : \quad \bar{\rho} = \psi_c(\rho) = \frac{1}{G} \sum_{g=1}^{G} \psi_c^g(\rho) = \frac{1}{G} \sum_{g=1}^{G} \| s_g - x_g \|^2.
\]
The following lemma, proved in [26], [37], and [18], shows the approximate IID Gaussianity of MU-OAMP/VAMP, which is critical to simplify the design and analysis of MU-OAMP/VAMP.

**Lemma 1 (Approximate IID Gaussianity):** The input of NLD can be regarded as $r = x + \rho^{-1/2}z$, where $z$ is an AWGN noise, $z \sim CN(0, I)$ independent of $x$, and $x$ is discrete signaling with $x_i \sim P_S(x_i)$. As a result, (6a) and (6c) are rewritten as

$$LD: \rho = \phi_L(v) = \{\Omega_L(v)^{-1}\}^{-1} - v^{-1},$$

$$NLD: \tilde{v} = \tilde{\phi}_L(\rho) = \{(\Omega_L(\rho)^{-1}) - \rho\}^{-1}, \quad (7a)$$

where $v \in [0, 1]$ and $\phi_L(0) = \frac{n}{N} \text{tr}\{A^H A\} = \text{snr}$ due to the normalized singular values of $A$, i.e., $\frac{1}{N} \text{tr}\{A^H A\} = 1$.

Next, we derive the expression of measurement MMSE $\Omega_{Ax}(\text{snr})$ in (11) using the properties of MU-OAMP/VAMP. As shown in Fig. 3, the iterative process between $LD \phi_L(\rho)$ and NLD $\Omega_S(\rho)$ converges to a unique fixed point $(\rho^*, v^*)$, where $v^* = ([\Omega_S(\rho^*)]^{-1} - \rho^*)^{-1}$ based on (7b). The curve $\phi_L(\rho)$ is an upper bound of $\Omega_S(\rho)$ for $0 \leq \rho \leq \rho^*$. According to the iterative process, the following lemma proved in [18, APPENDIX D] is given as follows.

**Lemma 2 (Measurement MMSE):** The constrained sum capacity $C_{\text{GMU-MIMO}}^\text{sum}$ is

$$C_{\text{GMU-MIMO}}^\text{sum} = N^C_{\text{GMU-MIMO}} = N\tilde{C}. \quad (12)$$

**Theorem 1 (Constrained Sum Capacity):** Assume that $\phi_L(\rho)$ and $\Omega_S(\rho)$ has a unique fixed point $(\rho^*, v^*)$. Then, the constrained sum capacity of GMU-MIMO is given by

$$C_{\text{GMU-MIMO}}^\text{sum} = \log |B(v^*)| + N\left(\log \Omega_S(\rho^*) - \int_0^{\rho^*} \Omega_S(\rho) \, d\rho\right), \quad (14)$$

where $B(v) = v^{-1}I + \text{snr}A^H A$. It is worth pointing out that for Gaussian signaling, the constrained sum capacity in (14) is degraded to the Gaussian sum capacity, i.e., $C_{\text{GMU-MIMO}}^\text{sum} = \log |v^{-1}I + \text{snr}A^H A|$ [14], where the MMSE function of Gaussian signaling is $\Omega_{\text{Gaus}}(\rho) = 1/(1 + \rho)$ [7].
C. Constrained-Group-Capacity Region of GMU-MIMO

For convenience, let

\[
\Omega_{S_0}(v) = \frac{1}{|S_Q|} \cdot \text{Tr}\{[\text{snr}A_{S_0}^H A_{S_0} + v^{-1}I_{S_0}]^{-1}\} , \quad (15a)
\]

\[
\phi_{S_0}(v) = [\Omega_{S_0}(v)]^{-1} - v^{-1} , \quad (15b)
\]

\[
\varphi_{S_0}(\rho) = (\rho + [\rho_{S_0,\text{inv}}(\rho)]^{-1})^{-1} , \quad (15c)
\]

where \(\{A_{S_0}, x_{S_0}, n_{S_0}\}\) are the sub-matrix and sub-vectors corresponding to \(S_0\). Using Theorem 1 and multi-user information theory [19, Ch. 15], the constrained capacity region of GMU-MIMO in the form of user group is given as follows.

**Theorem 2 (Constrained-Group-Capacity Region):** Assume that \(\Omega_S(\rho) = \varphi_{S_0}(\rho)\) has a unique positive solution \(\rho^*\), and \(v^* = [\Omega_S(\rho^*)]^{-1} - \rho^*\). The constrained group capacity region \(\{R_{S_0}, \forall Q \subseteq G\}\) of GMU-MIMO in (9) is

\[
R_{S_0} \leq I(Ax + n; x_{S_0}, x_{S_0}^c) = I(A_{S_0}x_{S_0} + n_{S_0}; x_{S_0}) \leq \log [B_{S_0}(v)] + N_u |S_Q| \cdot \log \Omega_S(\rho^*) + \int_0^{v^*} \Omega_S(\rho) d\rho , \quad (16c)
\]

where \(B_{S_0}(v) = v^{-1}I_{S_0} + \text{snr}A_{S_0}^H A_{S_0}\) and \(|S_Q|\) is the number of users in \(S_Q\).

Similarly, for Gaussian signaling, the constrained group capacity region in (16) is degraded to the well-known Gaussian capacity region [14]. As a result, the conventional Gaussian capacity region is a special case of the proposed constrained region when the signaling is Gaussian.

**IV. THEORETICAL SUM CAPACITY OPTIMALITY OF MU-OAMP/VAMP**

In this section, we analyze the achievable sum rate of MU-OAMP/VAMP, which is equal to the constrained sum capacity of GMU-MIMO. Then, the optimal design principle of multi-user codes is given.

A. Achievable Sum Rate and Constrained-Sum-Capacity Optimality of MU-OAMP/VAMP

To verify the optimality of MU-OAMP/VAMP, we investigate the achievable sum rate of MU-OAMP/VAMP with error-free performance.

1) **Code Rate:** The following lemma, proved in [47], shows the connection between the code rate and the MMSE decoding function.

**Lemma 3 (Code-Rate-MMSE):** The code rate \(R_{c_g}\) of the users in \(S_g\) is given by

\[
R_{c_g} = \int_0^\infty \Omega_{C_g}(\rho) d\rho . \quad (17)
\]

Similarly, the average rate \(\bar{R}_{\text{user}}\) of all users is given by

\[
\bar{R}_{\text{user}} = N_u \int_0^\infty \Omega_{C}(\rho) d\rho , \quad (18)
\]

which is shown in Fig. 4. Actually, \(\bar{R}_{\text{user}}\) also denotes the user rate in a symmetric GMU-MIMO.

2) **Upper Bounds of Decoding Function:** The following two lemmas provide two tight upper bounds of the MMSE decoding functions \(\{\Omega_{C_g}(\rho)\}\).

**Lemma 4 (Decoding Gain):** The channel decoding can provide significant coding gains, such that the locally optimal decoding of APP decoders should do better than locally symbol-by-symbol demodulation,

\[
\Omega_{C_g}(\rho) < \Omega_S(\rho) , \quad \forall \rho \geq 0 , \quad \forall g . \quad (19)
\]

Thus,

\[
\Omega_{C}(\rho) < \Omega_S(\rho) , \quad \forall \rho \geq 0 . \quad (20)
\]

According to (7) and (8), Fig. 4 shows that MU-OAMP/VAMP receiver is error-free if and only if the NLD curve \(\Omega_{C}(\rho)\) lies below the LD \(\varphi_{C}(\rho)\) [18]. Therefore, we obtain the error-free condition of MU-OAMP/VAMP in Lemma 5.

**Lemma 5 (Error-Free Decoding):** MU-OAMP/VAMP can achieve error-free decoding if and only if

\[
\Omega_{C}(\rho) < \varphi_{C}(\rho) , \quad \forall \rho \in [0, \text{snr}) \quad (21)
\]

and \(\Omega_{C}(\rho) = 0\) for \(\rho \geq \text{snr}\).

Therefore, as shown in Fig. 4, the MMSE of a feasible coded NLD is defined as

\[
\bar{\Omega}_{C}(\rho) = \min\{\Omega_S(\rho), \varphi_{C}(\rho)\} , \quad \forall \rho \in [0, \text{snr}) , \quad (22)
\]

and \(\bar{\Omega}_{C}(\rho) = 0\) for \(\rho \geq \text{snr}\).

From Lemma 4 and Lemma 5, it is easily obtained as

\[
\bar{\Omega}_{C}(\rho) < \bar{\Omega}_{C}(\rho) , \quad \forall \rho \in [0, \text{snr}) . \quad (23)
\]

3) **Achievable Sum Rate and Constrained-Sum-Capacity Optimality:** Assuming that there exist code-books \(\{C_g\}\) for \(\{S_g\}\), the \(\Omega_{C}(\rho)\) can match \(\varphi_{C}(\rho)\), i.e., \(\Omega_{C}(\rho) = \varphi_{C}(\rho)\), and \(\bar{\Omega}_{C}(\rho) = \bar{\Omega}_{C}(\rho)\). The achievable sum rate of MU-OAMP/VAMP is

\[
R_{\text{MU-OAMP/VAMP}}^\text{sum} = \frac{N}{G} \sum_{g=1}^G R_{c_g} = N \int_0^{\text{snr}} \bar{\Omega}_{C}(\rho) d\rho \quad (24a)
\]

\[
= N \bar{R}_{\text{MU-OAMP/VAMP}}^\text{sum} , \quad (24b)
\]
Assume that the achievable sum rate of MU-OAMP/VAMP is equal to the average constrained capacity, which is the area covered by $\bar{\Omega}_C$. The maximum achievable rate of MU-OAMP/VAMP equals the average rate of MU-OAMP/VAMP, and the optimal MMSE function for Gaussian signaling. For QPSK, 8PSK, and 16QAM modulation, MU-OAMP/VAMP achieves the Gaussian capacity.

**Theorem 3 (Constrained-Sum-Capacity Optimality):**
Assume that $\Omega_S(\rho)=\varphi_C(\rho)$ has a unique positive solution $\rho^*$ and $v^* = (\Omega_S(\rho^*))^{-1} - \rho^*$.

Then, $R_{\text{MU-OAMP/VAMP}} = C_{\text{GMU-MIMO}}.$

**Lemma 6 (Sum Code Rate):** The sum rate of all users is

$$R_{\text{sum}} = \frac{N}{G} \sum_{g=1}^{G} R_{c_g} = N \int_{0}^{\text{snr}} \Omega_c\cdot(\rho) d\rho$$

where $R_{c_g}$ is given in (17).

1) **Symmetric Systems:** Consider $\{\zeta_g(x) = x, g \in \mathbb{G}\}$, which corresponds to the symmetric case, i.e., all users have the same transfer function and rate, i.e.,

$$\Omega_{c_g}(\rho) = \Omega_c\cdot(\rho), \quad 0 \leq \rho \leq \rho^*, \quad \forall g \in \mathbb{G}, \quad (26)$$

and $\Omega_{c_g}(\rho) = 0, \rho > \text{snr}$, where $\zeta_g(\cdot)$ is a variance allocation function.

**B. Multi-User Code Design for MU-OAMP/VAMP**

To maximize the achievable sum rate of MU-OAMP/VAMP, $\{\Omega_{c_g}(\cdot)\}$ is elaborately selected to make $\Omega_c\cdot(\rho)$ match with $\Omega_{c_g}(\rho)$. From (17)-(24), the constraints of $\{\Omega_{c_g}(\cdot)\}$ are given in Property 1.

**Property 1 (Matching Conditions):** The optimal $\{\Omega_{c_g}(\cdot)\}$ for MU-OAMP/VAMP satisfies the following conditions.

- $\{\Omega_{c_g}(\rho), \forall g \in \mathbb{G}\}$ are monotone decreasing in $\rho \geq 0$.
- $\Omega_c(\rho) = \frac{1}{G} \sum_{g=1}^{G} \Omega_{c_g}(\rho) \rightarrow \Omega_c(\rho)$.
- $0 \leq \Omega_{c_g}(\rho) < \Omega_S(\rho), \forall g \in \mathbb{G}$.

Based on Property 1 and (22), we have

$$\Omega_{c_g}(\rho) = \begin{cases} \Omega_S(\rho), & 0 \leq \rho \leq \rho^* \,, \quad \forall g \in \mathbb{G}, \\
\zeta_g(\varphi_C(\rho)), & \rho^* \leq \rho \leq \text{snr} \,. \end{cases}$$

And $\Omega_{c_g}(\rho) = 0, \rho > \text{snr}$, where $\zeta_g(\cdot)$ is a variance allocation function.

**Proof:** See APPENDIX.

As shown in Fig. 5, from Theorem 3, (12), and (24), the average constrained capacity of GMU-MIMO $C$, the average rate of MU-OAMP/VAMP $R_{\text{MU-OAMP/VAMP}}$, and the area covered by the optimal MMSE function of decoder $\bar{\Omega}_C(\rho)$ are equal. This also illustrates the constrained-sum-capacity optimality of MU-OAMP/VAMP.

Fig. 6 shows that the average achievable rate (per transmit antenna) of MU-OAMP/VAMP in GMU-MIMO system with channel load $\beta = N/M = 1.5$, conditional number of channel matrix $\kappa = \{10, 50\}$, and QPSK, 8PSK, 16QAM and Gaussian signaling. MU-OAMP/VAMP achieves the Gaussian capacity for Gaussian signaling. For QPSK, 8PSK, and 16QAM modulations, MU-OAMP/VAMP can achieve the sum constrained capacity under the unique fixed point assumption.

In summary, we have demonstrated the information-theoretical (i.e., constrained-sum-capacity) optimality of MU-OAMP/VAMP over the coded GMU-MIMO with arbitrarily distributed signaling and general right-unitarily-invariant channel matrices. In the next subsection, we will further discuss the rate allocation for each user group based on Theorem 1 and Theorem 3.
\[
\begin{align*}
\{\gamma_i, k \neq i\} & \quad \text{Thus, we obtain the variance allocation function:} \\
\zeta_i(v_g) & = \frac{1}{G} \sum_{g=1}^{G} (b_{ig} v_g^{-1} + c_{ig})^{-1}, \\
\end{align*}
\]

which is a monotone increasing function for \( v_g \). Based on (30), \( v_g = \zeta_i^{-1}(v) \) may not have a close form but it can be obtained numerically. In addition, \( \zeta_i(\cdot) \) is increasing with \( \gamma_i \) and decreasing with \( \{\gamma_i, i \neq g\} \). Following (26) and (17), we have the following lemma.

**Lemma 7 (Monotonicity):** \( \{R_g\} \) is increasing with \( \gamma g \) while decreasing with \( \{\gamma_i, i \neq g\} \).

Lemma 7 provides an important guidance for the rate allocation, i.e., the code rate \( \{R_g\} \) can be changed flexibly by adjusting \( \{\gamma_g\} \), which is useful in the practical multi-user codes design.

3) **Transfer Curves Adjustment:** Following Property 1, \( \{v_g\} \) should satisfy

\[
\begin{align*}
\frac{1}{G} \sum_{g=1}^{G} v_g & = \frac{1}{G} \sum_{i=1}^{G} (b_{ig} v_k^{-1} + c_{ig})^{-1} \\
& \quad \forall \in (\Omega_S(\rho^*), 1],
\end{align*}
\]

where \( g \in \mathbb{G}, i \in S_g, n = 1, \ldots, N_g^i \), \( v_g^{i,n} \) is the variances at the \( n \)-th antenna of user \( i \), \( |S_g| \) denotes the number of users in group \( g \), and \( N_g^i \) is the number of antennas for user \( i \). As a result, the rate of user \( i \) in group \( g \) is

\[
R_g = N_g^i R_{c_g},
\]

where \( R_{c_g} \) is given in (17). On the other hand, the sum rate of group \( g \) is

\[
R_g = |S_g| \sum_{i=1}^{N_g^i} R_{c_g},
\]

which is proportional to the number of users and antennas.

C. **Comparing With Conventional Turbo-LMMSE Receiver**

In the following, we compare the average achievable rate per transmit antenna of MU-OAMP/VAMP and Turbo-LMMSE [9], [10], [11], which shows that Turbo-LMMSE is sub-optimal in the average achievable rate for non-Gaussian signaling. For simplicity, we consider the comparison in symmetric systems.

The main difference between MU-OAMP/VAMP and Turbo-LMMSE is explained as follows. The LMMSE-LD and MMSE-NLD of Turbo-LMMSE are extrinsic that require independent input/output errors. MU-OAMP/VAMP only requires orthogonal input/output errors, which is generally less stringent than the independent requirement.

Assume that the transfer functions of the detector and the decoder in Turbo-LMMSE are matched. The average achievable rate per transmit antenna of Turbo-LMMSE is given in [9]

\[
R_{\text{Turbo-LMMSE}} = \log |S| - \int_{0}^{+\infty} \Omega_S(\rho + \phi_{L}(\Omega_S(\rho)))d\rho.
\]

It rigorously proved that the MSE of MU-OAMP/VAMP is lower than that of Turbo-LMMSE in [20]. Here, we provide an intuitive explanation for this comparison from the perspective of the average achievable rate. Note that the LMMSE-LDs of MU-OAMP/VAMP and Turbo-LMMSE are equivalent for large-scale systems, in which the main difference lies in the NLD output operations. Fig. 8 shows the comparison between the MU-OAMP/VAMP NLD and Turbo-LMMSE NLD, where \( x_{c_{in}} \) is the input of NLD, \( u_{c}^{\text{app}} \) (independent of \( x_{c_{in}} \)) the extrinsic message of the decoder, and \( u_{c}^{\text{ext}} \) the a-priori message of the decoder. Then, the MMSEs of MU-OAMP/VAMP and Turbo-LMMSE are given by

\[
\begin{align*}
\Omega_{c_{\text{Turbo-LMMSE}}} & = \text{var}\{x| x_{c_{in}}^{\text{ext}}, x_{c_{in}}^{\text{in}}\} \\
& = \text{var}\{x| E\{x| u_{c}^{\text{app}}\}, x_{c_{in}}^{\text{in}}\}, \\
\end{align*}
\]

\[
\begin{align*}
\Omega_{c_{\text{MU-OAMP/VAMP}}} & = \text{var}\{x| x_{c_{in}}^{\text{orth}}, x_{c_{in}}^{\text{in}}\} \\
& \overset{(a)}{=} \text{var}\{x| u_{c}^{\text{app}}\} \\
& \overset{(b)}{=} \text{var}\{x| u_{c}^{\text{ext}}, x_{c_{in}}^{\text{in}}\},
\end{align*}
\]

where (a) follows orthogonal operation (2) that \( \{x_{c_{in}}^{\text{orth}}, x_{c_{in}}^{\text{in}}\} \) is a sufficient statistic for \( u_{c}^{\text{app}} \), and (b) follows that \( u_{c}^{\text{app}} \)
receiver. Fig. 2(a) shows that the MU-OAMP/VAMP receiver is consisted of a LMMSE detector and G APP decoders. The complexity of the LMMSE detector is $O(min\{MN^2 + N^3, NM^2 + M^3\})$ [11], where $\tau_{max}$ is the maximum iteration number. The APP decoders include a bank of demodulation and LDPC decoding, which perform the symbol-by-symbol mapping (from estimation to likelihood probability) and sum-product algorithm (SPA) [43] respectively. The complexity of demodulation is $O(K \log_2 |S| \tau_{max})$, where $K$ is the number of users and $|S|$ is the size of modulation constellation $S$. For MU-LDPC code $C_g(\lambda_g(\chi), \mu_g(\chi))$, the number of variable nodes with degree $d$ and check nodes with degree $d$ are denoted by $N_g^V(d)=N_uL\lambda_g^V/d$ and $N_g^C(d)=N_uL\mu_g^C/d$, which require $\sum_{d=2}^{d_{v,max}} 3N_g^V(d)$ additions/subtractions and $\sum_{d=2}^{d_{c,max}} N_g^C(d)(d-2)$ box-plus operations [43] in one iterative detection. In summary, the complexity of the receiver is given in Table III, in which the complexity of LMMSE detection is the decisive factor.

Note that the complexity of the globally optimal maximum a posteriori (MAP) receiver increases exponentially with the number of users and antennas [42]. As a result, compared with the MAP receiver, the MU-OAMP/VAMP receiver is acceptable in practical systems. Compared with Turbo-LMMSE [10], [11], the incremental complexity of MU-OAMP/VAMP is ignorable. The main difference between MU-OAMP/VAMP and Turbo-LMMSE is the output operations of NLD. As shown in Fig. 8, MU-OAMP/VAMP requires additional orthogonal operations and a-posteriori message of the decoder obtained by the sum of the extrinsic and a-priori messages.

Recently, a low-complexity memory AMP (MAMP) has been developed for right-unitarily-invariant matrices [39]. Since MAMP uses a low-complexity memory-matched filter to suppress linear interference, its complexity is comparable to AMP and much lower than that of OAMP/VAMP. Moreover, state evolution can accurately characterize the dynamics of MAMP. Aside from that, the state evolution of the MAMP reaches the MMSE fixed point predicted by the replication method. Therefore, to further achieve the lower complexity, MAMP [39] is a good candidate for the proposed system.

V. TWO USER-GROUP GMU-MIMO

In this section, we take two user-group GMU-MIMO as an example, i.e., all users are partitioned into two groups $S_1$ and $S_2$. The MU-OAMP/VAMP has been first proved to achieve the associated capacity region. Then, practical MU-LDPC codes are designed for two user-group GMU-MIMO.

A. Constrained-Capacity Optimality of MU-OAMP/VAMP in Two User-Group GMU-MIMO

Let $v_1$ and $v_2$ denote the output NLD variances of $S_1$ and $S_2$ respectively. Then

$$v_i = \begin{cases} \Omega_2(\rho), & 0 \leq \rho < \rho^* \\ \zeta_i^{-1}(\varphi_L(\rho)), & \rho^* \leq \rho < \text{snr} \end{cases}, \forall i \in \{1, 2\}. \quad (39)$$
TABLE III

COMPLEXITY OF THE MU-OAMP/VAMP RECEIVER

| Name                  | Complexity                                      |
|-----------------------|------------------------------------------------|
| LMMSE detection       | \(O(\min(MN^2 + N^3, NM^2 + M^3) \tau_{\text{max}})\) |
| MU-LDPC                | Demodulation \(O(K \log_2 |S| \tau_{\text{max}})\) |
| spa decoding           | box-plus \(O(\tau_{\text{max}})\) |
|                       | \(\frac{K}{2} \sum_{G=1}^G \sum_{d=2}^{d_{\text{max}}} 3N_c^G (d) d \tau_{\text{max}}\) |
|                       | \(\frac{K}{2} \sum_{G=1}^G \sum_{d=2}^{d_{\text{max}}} N_c^G (d) (d-2) \tau_{\text{max}}\) |

Fig. 9. SE and achievable rate region of MU-OAMP/VAMP for two user-group GMU-MIMO systems.

Assuming that \(\gamma_1 = 1\), \(b \equiv b_{21} = \gamma_2^{-1}\), and \(c \equiv c_{21} = (1-b)c^*\) according to (30), we can rewrite \(v_2\) as \(v_2 = (bv_1^{-1} + c)^{-1}\). Then, the average output variance of NLD is

\[
v = \frac{v_1 + v_2}{2} = \frac{v_1 + (bv_1^{-1} + c)^{-1}}{2}.
\]

Since \(v = \bar{\Omega}_c^*(\rho)\) is fixed, we can find the desired \(v_1\) by changing \(b\).

As shown in Fig. 9(b), when the value of \(b\) is too large, \(v_1\) will be larger than \(\Omega_S(\rho)\) in certain values of \(\rho\). Thus, the last condition in Property 1 may not hold. In this case, we clip the exceeded part (the black dash line BD) of \(v_1\), and correspondingly allocate this part to \(v_2\) (the purple solid line BC) to satisfy the average constraint (40). By choosing the different values of \(b\), MU-OAMP/VAMP can achieve the whole constrained capacity region of GMU-MIMO, i.e.,

\[
R_{\text{sum}} = R_{\bar{\Theta}_1} + R_{\bar{\Theta}_2} \leq I(Ax + n; x).
\]

As shown in Fig. 9(c), point P, point Q, and segment PQ denote the maximum \(R_{\bar{\Theta}_1}\), the maximum \(R_{\bar{\Theta}_2}\), and the maximum sum rate region, respectively, where point F represents the symmetric system, i.e., \(R_{\bar{\Theta}_1} = R_{\bar{\Theta}_2}\). Meanwhile, different rate allocations can be achieved by adjusting \(b\).

B. Achieving the Maximal Extreme Points of Two User-Group GMU-MIMO

As shown in Fig. 9(c), the achievable rate region of MU-OAMP/VAMP in two user-group GMU-MIMO is dominated by the convex combination of maximal extreme points P and Q. To demonstrate the optimality of MU-OAMP/VAMP, we show all the maximal extreme points that can be achieved by MU-OAMP/VAMP. Due to the symmetry between P and Q, for simplicity, we analyze the transfer curves of each user group for point Q.

At point Q, the achievable sum rates of each user group are

\[
R_{\bar{\Theta}_1} = I(A_{\bar{\Theta}_1}x_{\bar{\Theta}_1} + n_{\bar{\Theta}_1}; x_{\bar{\Theta}_1}),
\]

\[
R_{\bar{\Theta}_2} = I(A_{\bar{\Theta}_2}x_{\bar{\Theta}_2} + n_{\bar{\Theta}_2}; x_{\bar{\Theta}_2}),
\]

\[
R_{\text{sum}} = R_{\bar{\Theta}_1} + R_{\bar{\Theta}_2} \leq I(Ax + n; x).
\]

As shown in Fig. 9(c), the transfer function is

\[
\varphi_{\ell_{\bar{\Theta}_1}}(\rho) = (\rho + [\varphi_{\ell_{\bar{\Theta}_1}}(\rho)]^{-1})^{-1}.
\]

Then, following (22), the MMSE of a feasible coded NLD for user group 1 is

\[
\Omega_{c_1}^* = \min \{\Omega_S(\rho), \varphi_{\ell_{\bar{\Theta}_1}}(\rho)\}, \forall \rho \in [0, \text{snr}],
\]

and \(\Omega_{c_2}^* = 0\) for \(\rho \geq \text{snr}\). Following (22) and (45),

\[
\Omega_{c_2}^* = 2 \times \Omega_{c_1}^* - \Omega_{c_1}^*.
\]
which is MMSE function of a feasible coded NLD for user group 2 with $R_c^g$. As a result, the maximal extreme point Q is achieved. Fig. 10 shows that transfer curves of $\Omega_c^s$ and $\Omega_c^g$ at the maximal extreme point Q are obtained by (46) and (47), which avoids adjusting parameter $b$ in (40). It should be noted that the above strategy is consistent with the rate adjustment strategy discussed in (39) and (40) as $b \to \infty$.

C. Design of MU-LDPC Codes in Two User-Group GMU-MIMO

Similar as [21], the MU-LDPC codes are optimized for two user-group GMU-MIMO, where channel loads $\beta = 1.5$ with $(N, M) = (500, 333)$, and conditional number $\kappa = 10$. We consider that there are $N/G = 250$ antennas in each group.

For a target rate $R_{\text{MU-OAMP/VAMP}} = 1$ in (24), i.e., $R_{\text{MU-OAMP/VAMP}} = 500$, with QPSK modulation, the limit SNR corresponding to the constrained sum capacity is 2.85 dB given in Fig. 6. As a result, the proposed MU-LDPC codes are given in Table IV over the symmetric and group-asymmetric GMU-MIMO systems, where $b$ is the adjustment parameters in (40) for rate allocation, user rate $R_c^g$ of each group is equal to $N_a R_{\text{LDPC}} \log_2 |S_{\text{QPSK}}|$ with the rate of LDPC $R_{\text{LDPC}}$, and QPSK modulation $|S_{\text{QPSK}}| = 4$. Note that the decoding threshold of MU-LDPC codes is $(snr)^{\text{th}}_{\text{B}} = 2.87$ dB and about 0.02 dB away to the constrained sum capacity of GMU-MIMO.

To clearly visualize the above optimization process, Fig. 11 provides the SE curves of desired $\Omega_c^s$ and $\Omega_c^g$, and SE curves of designed $\Omega_c^s$ for MU-LDPC codes over two user-group GMU-MIMO when considering the symmetric and asymmetric scenarios. In Fig. 11, the SE curve of desired $\Omega_c^s$ is given for the symmetric case with $b = 1$ and that of $\Omega_c^g$ is given for the asymmetric cases with $b = 100$ and 0.2. Since $\Omega_c^s$ with $b = 100$ is beyond the constraint of (32), the SE curves of $\Omega_c^g$ and $\Omega_c^s$ with $b = 100$ are adjusted according to (32). Note that the SE curves of designed MU-LDPC codes in Table IV can match with those of desired MU-LDPC codes well, which also illustrates the optimality of the proposed MU-LDPC codes.

VI. NUMERICAL RESULTS

In this section, we provide the practical finite-length bit-error rate (BER) performances of the proposed MU-LDPC codes for MU-OAMP/VAMP in GMU-MIMO. We benchmark the proposed GMU-MIMO framework with P2P-regular LDPC codes, well-designed P2P-irregular LDPC codes, and the state-of-the-art Turbo-LMMSE.

A. Simulation Configuration

1) Signaling Distribution: QPSK modulation is employed in simulations and the corresponding MMSE function $\Omega_{\text{QPSK}}(\rho)$ is given in [21], where QPSK is a modulation scheme widely used in 5G and satellite communications and is also important for cell edge users and low-cost sensor devices, i.e., relatively low SNR and low transmission rate scenarios. However, for higher-order modulations, e.g., 16QAM, code design is very tough, because conventional code design such as EXIT analysis cannot be directly used due to three main reasons: 1) Probability density function (PDF) of the demodulated log-likelihood ratio (LLR) per bit in the constellation points is not Gaussian during the iterations; 2) PDFs of bit LLRs in the same constellation point are asymmetric and correlated; 3) Asymptotic decoding performance cannot be analyzed based on the transmission of all-zero codewords. To address the code design under higher-order modulations, bit-interleaved coded modulation (BICM) and superposition coded modulation (SCM) may be good candidates to approach perfect matching with the optimal code curve, in which BICM can be used to optimize constellation shaping for higher-order modulations during the iterations [48] and SCM has been proven theoretically to asymptotically achieve perfect matching for Gaussian signals of interest [21]. Hence, the design principle of multi-user codes and matching strategy are not limited to QPSK. Detailed discussions on the higher-order modulations are out of the scope of this paper, which is left as future work.

2) User Groups: For simplicity, we consider the symmetric and group-asymmetric GMU-MIMO with two user groups. Each group has 250 transmitted antennas. The total number of transmitted antenna $N = 500$. The coding length of each user in each group is $10^5$, channel is used 400 times for the whole transmission, and the corresponding parameters are given in Table IV. Note that the user rate $R_{c_g}$ of each group is equal to $N_a R_{\text{LDPC}} \log_2 |S_{\text{QPSK}}|$.

3) Ill-Conditioned Channel Matrix: We assume that the channel is quasi-static and channel matrix $A$ is fixed during the transmission, where $A \in C^{M \times N}$ is ill-conditioned, $M = 500$, $M = 333$, and channel load $\beta = N/M = 1.5$. Let the SVD of $A$ be $A = U \Lambda V$. $U$ and $V$ are generated by the orthogonal matrices in the QR decomposition of two IID Gaussian matrices. We set the eigenvalues $\{e_i\}$ in $A$ as $[22]: e_i/e_{i+1} = r^{1/T}, i = 1, \ldots, T - 1$ and $\sum_{i=1}^T e_i^2 = N$, where $T = \min\{M, N\}$ and $\kappa = \{10, 50\}$ denotes the condition number of $A$. Furthermore, note that the advantages of OAMP/VAMP has been demonstrated in spatial correlated channels [20], in which the channel matrices of $L$ channel uses is denoted as $A = \text{diag}\{A_1, A_2, \ldots, A_L\}$, $A_i =$
B. BER Comparison With P2P Regular and Irregular LDPC Codes

Fig. 12 provides the BER simulations of the optimized MU-LDPC codes in Table IV for GMU-MIMO systems. Meanwhile, in order to verify the advantages of the proposed frameworks, we also provide the BER performances of the P2P-regular LDPC codes and the well-designed P2P-irregular LDPC codes. The parameters of the P2P-regular LDPC codes are \((3, 6)\) LDPC codes with coding rate \(R_{\text{LDPC}} = 0.5\) and \((3, 5)\) LDPC codes with coding rate \(R_{\text{LDPC}} = 0.4\) [43]. The degree distributions of one well-designed P2P-irregular LDPC code [44] are \(\lambda(X) = 0.24426x + 0.25907x^2 + 0.01054x^3 + 0.05510x^4 + 0.01455x^5 + 0.01275x^6 + 0.04037x^7\) and \(\mu(X) = 0.25475x^6 + 0.73438x^7 + 0.01087x^8\), whose rate \(R_{\text{LDPC}}\) is 0.5 and the decoding threshold is 0.18 dB away from the P2P-AWGN capacity. The degree distributions of the other P2P-irregular LDPC code [45] are \(\lambda(X) = 0.29472x + 0.25667x^2 + 0.44861x^3\) and \(\mu(X) = x^5\), whose rate \(R_{\text{LDPC}}\) is 0.4 and the decoding threshold is 0.18 dB away from the P2P-AWGN capacity.

For symmetric GMU-MIMO, Fig. 12 shows that the gap between BER curve at \(10^{-4}\) of the optimized MU-LDPC code and the corresponding constrained sum capacity is 1 dB, which verifies capacity-approaching performances of the optimized MU-LDPC codes. In addition, the optimized MU-LDPC codes have about 1.9 dB and 2.75 dB performance gains over the benchmarks P2P-regular LDPC codes and the well-designed P2P-irregular LDPC codes, respectively. Meanwhile, user rate (\(R_{C_0} = 253\)) of the proposed MU-LDPC code is slightly higher than those of the benchmarks P2P-irregular and P2P-regular LDPC codes (\(R_{C_0} = 250\)).

For user group \(S_1\) in Fig. 12, the optimized MU-LDPC codes with \(R_{C_1} = 222.5\) can achieve about 0.45 dB and 0.65 dB performance gains over the P2P-irregular LDPC codes with \(R_{C_1} = 200\), and the well-designed P2P-irregular LDPC codes with \(R_{C_1} = 200\) respectively. For user group \(S_2\) in Fig. 12, the proposed MU-LDPC codes with \(R_{C_2} = 277.5\) can achieve about 0.5 dB and 0.7 dB performance gains over the P2P-irregular and P2P-regular LDPC codes with \(R_{C_2} = 250\), in which our codes also have the higher sum rate. In short, comparing with the benchmarks, the proposed MU-LDPC codes not only has a significant improvement in BER performances, but also has higher transmission rates. The reason is that compared to P2P coding, multi-user coding suppress not only channel noise but also multi-user interference.

To confirm the robustness of the proposed MU-LDPC codes, we consider the BER simulations in GMU-MIMO with imperfect channel estimations, where the standard deviations of estimated channel errors are 0.002 and 0.004. Fig. 12 shows that imperfect channel estimations cause about 0.3 dB performance losses, in which the gaps between BER curves at \(10^{-4}\) of the proposed MU-LDPC codes and the corresponding
closely approaches the rate pair \((R_{Q2}, R_{Q2})\) (achieved by the optimized MU-LDPC codes) on the symmetric point \(F\) and is \((R_{P1}, R_{P2})\) and \((R_{Q1}, R_{Q2})\) for asymmetric points \(P_1\) and \(P_2\), respectively. At point \(Q_1\), the rate pair \((R_{C_1}, R_{C_2}) = (330, 180)\) corresponds to \((R_{LDPC1} = 0.66, R_{LDPC2} = 0.36)\), which closely approaches the rate pair \((R_{C_1} = 335, R_{C_2} = 175)\) at the theoretical extreme point \(Q\) with \((R_{LDPC1} = 0.67, R_{LDPC2} = 0.33)\). Note that the decoding threshold of MU-LDPC codes is \((snr)_{dB} = 5.25\) dB and about 0.02 dB away to the sum capacity of GMU-MIMO, where the symmetric point \(F\) and middle point \(Q_2\) (achieved by the optimized MU-LDPC codes) approach the sum capacity. Due to the symmetry, extreme point \(P_1\), middle point \(P_2\) can be achieved in the same way by the MU-LDPC codes in Table V without re-optimizations. Therefore, the whole region of the maximal sum capacity of GMU-MIMO systems can be achieved by adjusting the parameter \(b\) and optimizing the MU-LDPC codes.

Fig. 14 provides the BER performances of the MU-LDPC codes with MU-OAMP/VAMP in Table V for two user-group GMU-MIMO systems, where \(\beta = 1.5, \kappa = 50,\) and \((snr)_{dB} = 5.25\) dB. For symmetric point \(F\), the MU-LDPC codes with \(R_{C_2} = 250\) have about 3.9 dB and 4.8 dB performance gains over the benchmarks P2P-regular and P2P-irregular LDPC codes respectively.

At extreme point \(Q_1\) and middle point \(Q_2\), for user group \(S_1\), the optimized MU-LDPC codes with \(R_{C_1} = 330\) (\(Q_1\)) and \(R_{C_2} = 280\) (\(Q_2\)) respectively have about 1.5 dB and 1.7 dB BER performance gains over the P2P-regular and P2P-irregular LDPC codes with \(R_{C_1} = 250\). For user group \(S_2\), the optimized MU-LDPC codes with \(R_{C_2} = 180\) (\(Q_1\)) and \(R_{C_2} = 220\) (\(Q_2\)) respectively have about 1.5 dB and 1.7 dB BER performance gains over the P2P-regular and P2P-irregular LDPC codes with \(R_{C_1} = 200\).

D. BER Comparison With Turbo-LMMSE

In this subsection, we compare the proposed framework with the state-of-the-art Turbo-LMMSE [10], [11]. According to [10, Sec. IV], the transfer curves of asymmetric user groups in Turbo-LMMSE can be obtained based on extrinsic LMMSE detector and extrinsic decoders. Based on extrinsic transfer curve matching, the MU-LDPC codes optimized for Turbo-LMMSE are given in Table V, in which the sum rates of symmetric and asymmetric cases are 500, and the gap between the threshold of Turbo-LMMSE and the corresponding limit is 0.3 dB.
As shown in Fig. 14, the proposed MU-LDPC code with MU-OAMP/VAMP has about 3.4 dB performance gain over the Turbo-LMMSE in symmetric cases. For asymmetric cases, the proposed MU-LDPC codes at the extreme point $Q_1$ have about 2.5 dB performance gains. Therefore, the above comparisons verify the advantages of the proposed framework.

VII. CONCLUSION

This paper focuses on GMU-MIMO with the certain general and practical assumptions, i.e., practical channel coding, arbitrary discrete signaling, non-IID channel matrix, massive users and antennas, and available CSI only at the receiver. The information-theoretical limit of GMU-MIMO and the optimal capacity-achieving transceiver with practical complexity are open issues. To solve these issues, this paper proposes a unified framework for GMU-MIMO, jointly considering encoding, modulation, detection, and decoding. Meanwhile, group asymmetry is developed to make a tradeoff between user rate allocation and implementation complexity. Based on the framework, the constrained capacity region of group-asymmetric GMU-MIMO is accurately derived. The MU-OAMP/VAMP receiver with matched multi-user codes is proposed to achieve the error-free recovery performance for GMU-MIMO. The optimal design principle of multi-user code is presented and then a kind of MU-LDPC code is optimized for GMU-MIMO. Numerical results demonstrate that the gaps between theoretical decoding thresholds of the proposed framework with optimized MU-LDPC codes and the constrained sum capacity of GMU-MIMO are about 0.2 dB. Furthermore, their finite-length performances are $1\sim 2$ dB away from the associated sum capacity. This implies that
the proposed framework would be an important candidate for next-generation wireless communications.

**APPENDIX**

**PROOF OF THEOREM 3**

Define $\tilde{\phi}^{-1}(\cdot)$ as the generalized inverse function of $\phi_L(\cdot)$ as following.

$$
\phi^{-1}_L(\rho) = \begin{cases} 
1, & \rho < \phi_L(1), \\
\phi^{-1}_L(\rho), & \phi_L(1) \leq \rho \leq \text{snr}, \\
0, & \rho > \text{snr},
\end{cases}
$$

where $\phi^{-1}_L(\rho)$ is defined as the inverse function of $\phi_L(\cdot)$ for $\rho \in [\phi_L(1), \text{snr}]$.

Then, we have

$$
\varphi_L(\rho) = \left( \frac{1}{\rho + [\phi^{-1}_L(\rho)]^{-1}} \right)^{-1} = \begin{cases} 
(\rho + 1), & \rho < \phi_L(1), \\
(\rho + [\phi^{-1}_L(\rho)]^{-1})^{-1}, & \phi_L(1) \leq \rho \leq \text{snr}, \\
0, & \rho > \text{snr}.
\end{cases}
$$

Define $\tilde{b}(v) = \frac{1}{\text{snr}} \text{Tr} \{ [B(v)]^{-1} \}$. Based on I-MMSE lemma [32],

$$
\tilde{R}_{\text{MU–OAMP/VAMP}} = \int_0^{\rho^*} \Omega_S(\rho) d\rho + \int_{\rho^*}^{\text{snr}} \varphi_L(\rho) d\rho
$$

$$
= \int_0^{\rho^*} \Omega_S(\rho) d\rho + \int_{\rho^*}^{\phi_L(1)} [\rho + (\phi^{-1}_L(\rho))^{-1}]^{-1} d\rho
$$

$$
= \int_0^{\rho^*} \Omega_S(\rho) d\rho + \int_{\phi_L(1)}^{\text{snr}} (\rho + 1) d\rho
$$

$$
= \frac{1}{\text{snr}} \text{Tr} \{ [B(\rho^*)]^{-1} \}.
$$

Inequality (a) is derived by the matching assumption and the equality holds if and only if there exists that code whose transfer function matches the $\Omega_L^*(\rho)$. Equation (b) follows the match assumption, equation (c) is based on the law $\int \text{Tr} \{ sI + A \} d\rho = \log \det \{ sI + A \}$, and equation (d) is derived by $\det \{ I + A B M \} = \det \{ I M + B M X \}$ for any matrices $A_{K \times M}$ and $B_{M \times K}$. Equation (e) follows the fix-point function $\tilde{b}(v) = \Omega_L(v^*) = \Omega_S(v^*)$. Therefore, from Theorem 1 and (12), we can obtain $\tilde{R}_{\text{MU–OAMP/VAMP}} = \tilde{C}$ and $\tilde{R}_{\text{MU–OAMP/VAMP}} = C_{\text{MU–OAMP/VAMP}}$.
[24] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, “Inference for generalized linear models via alternating directions and Bethe free energy minimization,” *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 676–697, Jan. 2017.

[25] J. Ma and L. Ping, “Orthogonal AMP,” *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[26] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.

[27] A. Bereyhi, S. Asaad, B. Güde, R. R. Müller, and H. V. Poor, “Detection of spatially modulated signals via RLS: Theoretical bounds and applications,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2291–2304, Apr. 2022.

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

[29] S. Zhang, C.-K. Wen, K. Takeuchi, and S. Jin, “Orthogonal approximate message passing for GFDM detection,” in Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Jul. 2017, pp. 1–5.

[30] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawan, “Expectation propagation detector for extra-large scale massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2036–2051, Mar. 2020.

[31] K. Takeda, S. Uda, and Y. Kabashima, “Analysis of CDMA systems that are characterized by eigenvalue spectrum,” *Europhys. Lett.*, vol. 76, no. 6, p. 1193, 2006.

[32] D. Guo, S. S. Shitz, 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.

[33] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Oct. 2009.

[34] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[35] T. P. Minka, “Expectation propagation for approximate Bayesian inference,” 2013, arXiv:1301.2294.

[36] M. Opper, O. Winther, and M. J. Jordan, “Expectation consistent approximation,” *J. Mach. Learn. Res.*, vol. 6, no. 12, pp. 2177–2204, 2005.

[37] K. Takeuchi, “Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements,” *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 368–386, Jan. 2020.

[38] K. Takeuchi. “Bayes-optimal convolutional AMP,” *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4405–4428, Jul. 2021.

[39] L. Liu, S. Huang, and B. M. Kurkoski, “Memory AMP,” *IEEE Trans. Inf. Theory*, early access, Jun. 26, 2022, doi: 10.1109/TIT.2022.3186166.

[40] F. Tian, L. Liu, and X. Chen, “Generalized memory approximate message passing,” 2021, arXiv:2110.06069.

[41] A. M. Tulino, G. Caire, S. Verdú, and S. S. Shitz, “Support recovery with sparsely sampled free random matrices,” *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4243–4271, Jul. 2013.

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

[43] R. Gallager, *Information Theory* (2nd ed.). Cambridge, MA: MIT Press, 1968.

[44] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, “Design of capacity-approaching irregular low-density parity-check codes,” *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 619–637, Feb. 2001.

[45] J. Kim, T. Ramamooorthy, and S. W. McLaughlin, “The design of efficiently-encodable rate-compatible LDPC codes,” *IEEE Trans. Commun.*, vol. 57, no. 2, pp. 365–375, Feb. 2009.

[46] L. Liu, Y. Cheng, S. Liang, J. H. Manton, and L. Ping, “On orthogonal approximate message passing,” 2022, arXiv:2203.00224.

[47] 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.

[48] R. Y. S. Tee, R. G. Mauder, and L. Hanzo, “EXIT-chart aided near-capacity irregular bit-interleaved coded modulation design,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 32–37, Jan. 2009.
Ying Li (Member, IEEE) 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 at the University of California at Davis, Davis, CA, USA, as a Visiting Scholar. She is currently a Professor with Xidian University. Her current research interests include design and analysis for wireless systems, including channel coding, wireless network communications, interference processing, and MIMO techniques.

Yong Liang Guan (Senior Member, IEEE) received the B.Eng. degree (Hons.) from the National University of Singapore and the Ph.D. degree from the Imperial College London, U.K. He is currently a Professor of communication engineering with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, where he leads the Continental-NTU Corporate Research Laboratory and led the successful deployment of the campus-wide NTU-NXP V2X Test Bed. He has published an invited monograph, two books and more than 450 journals and conference papers. He has secured over S$70 million of external research funding. He has 15 filed patents and three granted patents, one of which was licensed to NXP Semiconductors. His research interests include coding and signal processing for communication systems and data storage systems. Currently, he is the Associate Vice President of NTU and a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is the Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Chau Yuen (Fellow, IEEE) received the B.Eng. and Ph.D. degrees from Nanyang Technological University (NTU), Singapore, in 2000 and 2004, respectively. He was a Post-Doctoral Fellow at Lucent Technologies Bell Laboratories, Murray Hill, in 2005. From 2006 to 2010, he was at the Institute for Infocomm Research (I2R), Singapore. Since 2010, he has been with the Singapore University of Technology and Design. He is a Distinguished Lecturer of IEEE Vehicular Technology Society. He was a recipient of the Lee Kuan Yew Gold Medal, the Institution of Electrical Engineers Book Prize, the Institute of Engineering of Singapore Gold Medal, the Merck Sharp and Dohme Gold Medal, and twice a recipient of the Hewlett Packard Prize. He has received the IEEE Asia–Pacific Outstanding Young Researcher Award in 2012 and the IEEE Vehicular Technology Society (VTS) Singapore Chapter Outstanding Service Award on 2019. He was the Guest Editor of several special issues, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE Wireless Communications Magazine, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is currently the Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE SYSTEMS JOURNAL, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING.