Vector Similarity Search-Based MCS Selection for Iterative Signal Detection in Massive Multi-User MIMO-OFDM Systems

Fuga Kobayashi *, Takumi Takahashi *, Shinsuke Ibi **, Hideki Ochiai *,
Kazushi Muraoka ***, Takanobu Doi ***, and Naoto Ishii ***
* Graduate School of Engineering, Osaka University, 2-1 Yamada-oka, Suita, 565-0871, Japan
** Faculty of Science and Engineering, Doshisha University, 1-3 Tataramiyakodani, Kyotanabe, 610-0394, Japan
*** NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211–8666, Japan
Email: * {kobayashi-f@wcs., takahashi@, ochiai@} comm.eng.osaka-u.ac.jp,
** sibi@mail.doshisha.ac.jp, *** {k-muraoka, doi-takanobu, naoto_ishii}@nec.com

***Abstract*—This paper proposes a novel vector similarity search (VSS)-based modulation and coding scheme (MCS) selection for massive multi-user multiple-input multiple-output orthogonal frequency division multiplexing (MU-MIMO-OFDM) systems that employ uplink multi-user detection (MUD) based on iterative signal estimation. To maximize the uplink throughput of MU-MIMO-OFDM systems, it is necessary to assign a carefully selected MCS to each user according to an accurate prediction of the mutual information (MI) that can be achieved with MUD based on the knowledge of the estimated channel state information (CSI). However, since the detection accuracy of iterative MUDs, such as expectation propagation (EP), varies depending on the convergence characteristics, it is challenging to analytically predict the achievable MI in the presence of MUD. To address this difficulty, we propose a novel method for predicting the achievable MI by creating a vector database (VDB) offline that stores feature vectors (*keys*) computed from CSI and the actual MI (*values*) achieved with iterative MUD, and then searching this VDB online using approximate nearest neighbors (ANN) search, which enables VSS at ultra-high speed. Simulation results show that the MCS selection based on the proposed MI prediction achieves higher uplink throughput than the conventional schemes in MU-MIMO-OFDM systems using the EP-based MUD.**

***Index Terms*—MU-MIMO-OFDM systems, MCS selection, mutual information, vector database, vector similarity search.**

I. Introduction

Massive multi-user multiple-input multiple-output orthogonal frequency division multiplexing (MU-MIMO-OFDM) is one of the essential technologies to cope with the explosive growth in the number of uplink user equipment (UE) devices in the future wireless networks [1]. In order to maximize the uplink throughput of MU-MIMO-OFDM systems, it is vital to adaptively assign an appropriate modulation and coding scheme (MCS) index to each UE according to its wireless channel quality. As a metric for measuring this channel quality, the estimated (predicted) value of the subcarrier-wise signal-to-interference-plus-noise power ratio (SINR) achievable by each UE with multi-user detection (MUD) has often been adopted [2], [3]. However, in many standardization specifications, channel coding is performed for each transport block (TB); hence, a signal codeword is distributed over multiple subcarriers with different channel quality. For this reason, it is necessary to evaluate the wireless channel quality over all the subcarriers (*i.e.*, allocated resources) that contribute to the transmission of the codeword, and the mutual information (MI), which can be calculated for each TB, has been considered as a more appropriate metric [4], [5].

When using spatial filtering, such as linear minimum mean square error (LMMSE) detection, it is possible to analytically estimate the post-MUD SINR from the channel state information (CSI). Therefore, after converting the estimated SINR of each subcarrier to MI, the predicted MI of the demodulator output can be obtained by averaging over all the subcarriers within the allocated resource (*i.e.*, 1 TB). In contrast, when using (nonlinear) iterative signal detection such as expectation propagation (EP) [6] as MUD, it is difficult to analytically estimate the post-MUD SINR because the detection accuracy varies depending on the iterative convergence characteristics [3]. It should be noted here that if the MI of the demodulator output cannot be predicted with high accuracy, the improvement in MUD accuracy achieved by iterative signal detection cannot be reflected in the MCS selection. Therefore, even if the MUD accuracy at the linklevel is improved, as reported in the literature, it will not lead to an enhancement in the overall system throughput.

Based on the above observations, in this paper, we propose a novel vector similarity search (VSS)-based offline learning approach that employs a vector database (VDB) and approximate nearest neighbors (ANN) search [7]–[9] to predict the achievable MI of the demodulator output in the presence of iterative MUD. The proposed method consists of a VDB construction phase based on offline learning and an online MI prediction phase using the constructed VDB. In the former phase, feature vectors (*keys*), which is a metric that can be calculated from CSI and the average received signal-to-noise power ratio (SNR), are generated and stored in the VDB together with the MIs (*values*) achieved when actual transmissions are performed over those wireless channels. In the subsequent online prediction phase, the key (*i.e.*, feature vector) is generated from the estimated CSI and average received SNR, and the VSS is executed by ANN search on the constructed VDB to obtain the value (*i.e.*, the MI of the demodulator output) that is linked to the key with high similarity. Since this value is the demodulator output MI that has actually been achieved for a similar wireless channel, it is possible to obtain a highly accurate prediction value that also takes into account the convergence characteristics of the iterative MUD. As a learning-based MCS selection scheme, the Q-learning approach [10], [11] has been widely investigated. However, in massive MU-MIMO-OFDM

systems, the size of the table/network becomes enormous, due to channel fluctuations. This may cause a computational challenge as the online learning may not converge within a realistic processing time. In contrast, the proposed method uses knowledge of the CSI and average received SNR as keys to achieve a significant reduction in the search space through offline learning, enabling high-speed, high-precision MI prediction and MCS selection even in massive MU-MIMO-OFDM systems.

ANN search has been used to achieve accurate VSS when similarities or correlations are found in the data being handled, and has been developed in the fields of translation between different languages and classification of high-dimensional images [9]. With the helpful recent development of the ANN library [12], there have been reports of its application to wireless communications, such as the transmission power allocation problem [13] and the meta-learning to a deep unfolding (DU)-aided signal detector [14], and this paper can be positioned as one of the advancements of such studies.

In the rest of this paper, we first confirm that MI prediction in the presence of LMMSE-based MUD can be achieved analytically with high accuracy, and clarify the difference from the achievable MI with EP-based MUD; next show that it is possible to predict MI after EP-based MUD with high accuracy by designing a feature vector (*key*) and creating a VDB. Finally, we demonstrate that system throughput can be significantly improved by MCS selection based on the proposed method.

Notation: Sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. Vectors and matrices are denoted in lower- and upper-case bold-face letters. The transpose and conjugate transpose operators are denoted by $\cdot^{\mathsf{T}}$ and $\cdot^{\mathsf{H}}$, respectively. The $a \times a$ identity matrix is denoted by $\boldsymbol{I}_a$. The diagonal matrix constructed by placing the elements of a vector $\boldsymbol{a}$ on its main diagonal is denoted by $\mathrm{diag}[\boldsymbol{a}]$. The $i-$th row and $j-$th column element of matrix $\boldsymbol{A}$ is denoted by $[\boldsymbol{A}]_{i,j}$. The complex Gaussian distribution with a mean a and a variance b is denoted by $\mathcal{CN}(a, b)$.

II. Preliminaries

A. Signal Model

Consider an uplink massive MU-MIMO-OFDM system in alignment with the fifth generation (5G) new radio (NR) specification as shown in Fig. 1. Each UE has one transmit (TX) antenna, and M UEs perform spatial multiplexing transmission to a base station (BS) receiver equipped with N receive (RX) antennas with uniform rectangular array (URA). In the BS, resource allocation and MCS selection for uplink transmission by each UE are performed based on the acquired CSI, and the results are notified to each UE. Each UE then encodes and modulates the information bit stream using the selected MCS and transmits the resulting signal using OFDM in the allocated time-frequency resource, where the resources allocated to each UE in a given time slot are called TB, and the information bit stream is encoded and modulated for each TB unit. The smallest TB in the 5G NR standard [15] is a resource block (RB) consisting of 14 OFDM symbols (time) $\times$ 12 subcarriers (frequency), and resource allocation is performed adaptively to every UE through scheduling processing. The insertion and removal of cyclic prefix in OFDM transmission are assumed to be carried out ideally, and the receiver performs MUD on the frequency-domain MIMO signal detection, demodulation, and then channel decoding.

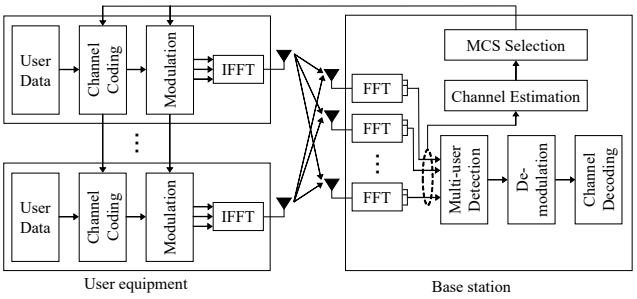


Fig. 1. Configuration of BS and UEs.

Let $x_m[r, \ell]$ be the frequency-domain symbol transmitted by the m-th UE at discrete time r using ℓ-th subcarrier, where each symbol is selected from the set of quadrature amplitude modulation (QAM) symbol candidate points $\mathcal{X} \triangleq \{\chi_1, \cdots, \chi_q, \cdots, \chi_Q\}$ with the average energy E_{s}. Denoting the TX vector spatially multiplexed on the $[r, \ell]$-th time-frequency resource by $\boldsymbol{x}[r, \ell] \triangleq [x_1[r, \ell], \ldots, x_m[r, \ell], \ldots, x_M[r, \ell]]^{\mathsf{T}} \in \mathbb{C}^{M\times 1}$, the RX vector $\boldsymbol{y}[r, \ell] \in \mathbb{C}^{N\times 1}$ can be expressed as

$$\boldsymbol{y}[r, \ell] = \boldsymbol{H}[\ell]\boldsymbol{x}[r, \ell] + \boldsymbol{z}[r, \ell], \tag{1}$$

where $\boldsymbol{H}[\ell] \triangleq [\boldsymbol{h}_1[\ell], \ldots, \boldsymbol{h}_m[\ell], \ldots, \boldsymbol{h}_M[\ell]] \in \mathbb{C}^{N\times M}$ is the frequency-domain MIMO channel matrix of the ℓ-th subcarrier; $\boldsymbol{h}_m[\ell] \in \mathbb{C}^{N\times 1}$ denotes the m-th column vector of $\boldsymbol{H}[\ell]$; and $\boldsymbol{H}[\ell]$ is assumed to be constant during one-time slot transmission ($1 \le r \le R$). In addition, $\boldsymbol{z}[r, \ell] \triangleq [z_1[r, \ell], \ldots, z_n[r, \ell], \ldots, z_N[r, \ell]]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}$ is the additive white Gaussian noise (AWGN) vector, each entry of which follows $\mathcal{CN}(0, N_0)$ with N_0 denoting the noise power density.

Since the MCS used by every UE must be determined prior to actual transmission, the BS must estimate (predict) the MI of the demodulator output that can be achieved in the presence of MUD, based on the estimated values of $\boldsymbol{H}[\ell], \forall \ell$ and the average received SNR $\rho \triangleq E_{\mathrm{s}}/N_0$ without the RX signal $\boldsymbol{y}[r, \ell], \forall r, \forall \ell$.

For simplicity, we assume that the CSI $\boldsymbol{H}[\ell], \forall \ell$ and average received SNR ρ are estimated without error at the BS and the assigned MCS index is notified to each UE without error.

B. EP-based MUD Algorithm

The pseudo-code of the EP-based MUD algorithm designed based on [6] is given in Algorithm 1, where we omit the qualifier $[r, \ell]$ since signal detection is performed independently of indices r, ℓ. The EP detector consists of module A, which performs signal separation using the LMMSE filter, and module B, which calculates the conditional expectation, which is the general MMSE solution, based on the output of module A. The detection accuracy can be gradually improved by exchanging extrinsic information between the two modules. For further details, we refer the reader to, *e.g.*, [6]. The number of iterations is denoted by T, and the qualifier $(\cdot)^{(t)}$ denotes the iteration index for every variable. Finally, α in lines 14 and 15 is a damping factor.

C. Calculation of Bit-Wise LLRs

Continuing from the previous subsection, we will omit the qualifier $[r, \ell]$. Based on the message $q^{(T)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q)$ from module A to B obtained in the final iteration of Algorithm

Algorithm 1 EP-based MUD algorithm

Input: $\boldsymbol{y} \in \mathbb{C}^{N\times 1}, \boldsymbol{H} \in \mathbb{C}^{N\times M}, T$

Output: $q^{(T)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q), \forall m, \forall \chi_q \in \mathcal{X}$

/* Initialization */

$\boldsymbol{x}^{(1)}_{\mathrm{B}\to\mathrm{A}} = \boldsymbol{0} \in \mathbb{C}^{M\times 1}, \boldsymbol{V}^{(1)}_{\mathrm{B}\to\mathrm{A}} = E_{\mathrm{s}}\boldsymbol{I}_M$

/* Iteration */

for $t = 1$ to T **do**

$\boldsymbol{\Psi}^{(t)} = \left(N_0\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{V}^{(t)}_{\mathrm{B}\to\mathrm{A}}\boldsymbol{H}^{\mathsf{H}}\right)^{-1}$

$\boldsymbol{\Xi}^{(t)} = \left(\mathrm{diag}\left[\boldsymbol{h}_1^{\mathsf{H}}\boldsymbol{\Psi}^{(t)}\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M^{\mathsf{H}}\boldsymbol{\Psi}^{(t)}\boldsymbol{h}_M\right]\right)^{-1}$

$\boldsymbol{x}^{(t)}_{\mathrm{A}\to\mathrm{B}} = \boldsymbol{x}^{(t)}_{\mathrm{B}\to\mathrm{A}} + \boldsymbol{\Xi}^{(t)}\boldsymbol{H}^{\mathsf{H}}\boldsymbol{\Psi}^{(t)}\left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^{(t)}_{\mathrm{B}\to\mathrm{A}}\right)$

$\boldsymbol{V}^{(t)}_{\mathrm{A}\to\mathrm{B}} = \boldsymbol{\Xi}^{(t)} - \boldsymbol{V}^{(t)}_{\mathrm{B}\to\mathrm{A}}$

$\forall m, \forall \chi_q \in \mathcal{X} : q^{(t)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q) = \exp\left(-\frac{\left|\chi_q - x^{(t)}_{m,\mathrm{A}\to\mathrm{B}}\right|^2}{v^{(t)}_{m,\mathrm{A}\to\mathrm{B}}}\right)$

$\forall m : x^{(t)}_{m,\mathrm{B}} = \frac{\sum_{\chi_q\in\mathcal{X}} \chi_q \cdot q^{(t)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q)}{\sum_{\chi'_q\in\mathcal{X}} q^{(t)}_{m,\mathrm{A}\to\mathrm{B}}(\chi'_q)}$

$\forall m : v^{(t)}_{m,\mathrm{B}} = \frac{\sum_{\chi_q\in\mathcal{X}} |\chi_q|^2 \cdot q^{(t)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q)}{\sum_{\chi'_q\in\mathcal{X}} q^{(t)}_{m,\mathrm{A}\to\mathrm{B}}(\chi'_q)} - \left|x^{(t)}_{m,\mathrm{B}}\right|^2$

$\forall m : \frac{1}{\overline{v}^{(t)}_{m,\mathrm{B}\to\mathrm{A}}} = \frac{1}{v^{(t)}_{m,\mathrm{B}}} - \frac{1}{v^{(t)}_{m,\mathrm{A}\to\mathrm{B}}}$

$\forall m : \overline{x}^{(t)}_{m,\mathrm{B}\to\mathrm{A}} = \overline{v}^{(t)}_{m,\mathrm{B}\to\mathrm{A}} \cdot \left(\frac{x^{(t)}_{m,\mathrm{B}}}{v^{(t)}_{m,\mathrm{B}}} - \frac{x^{(t)}_{m,\mathrm{A}\to\mathrm{B}}}{v^{(t)}_{m,\mathrm{A}\to\mathrm{B}}}\right)$

$\overline{\boldsymbol{x}}^{(t)}_{\mathrm{B}\to\mathrm{A}} = \left[\overline{x}^{(t)}_{1,\mathrm{B}\to\mathrm{A}}, \ldots, \overline{x}^{(t)}_{m,\mathrm{B}\to\mathrm{A}}, \ldots, \overline{x}^{(t)}_{M,\mathrm{B}\to\mathrm{A}}\right]^{\mathsf{T}}$

$\overline{\boldsymbol{V}}^{(t)}_{\mathrm{B}\to\mathrm{A}} = \mathrm{diag}\left[\overline{v}^{(t)}_{1,\mathrm{B}\to\mathrm{A}}, \ldots, \overline{v}^{(t)}_{m,\mathrm{B}\to\mathrm{A}}, \ldots, \overline{v}^{(t)}_{M,\mathrm{B}\to\mathrm{A}}\right]$

$\boldsymbol{x}^{(t+1)}_{\mathrm{B}\to\mathrm{A}} = \alpha\overline{\boldsymbol{x}}^{(t)}_{\mathrm{B}\to\mathrm{A}} + (1-\alpha)\boldsymbol{x}^{(t)}_{\mathrm{B}\to\mathrm{A}}$

$\boldsymbol{V}^{(t+1)}_{\mathrm{B}\to\mathrm{A}} = \alpha\overline{\boldsymbol{V}}^{(t)}_{\mathrm{B}\to\mathrm{A}} + (1-\alpha)\boldsymbol{V}^{(t)}_{\mathrm{B}\to\mathrm{A}}$

end for

1, the log-likelihood ratios (LLRs) of the coded bits that compose the TX symbol $\boldsymbol{x}_m$ is calculated. When x_m consists of $S = \log_2 Q$ coded bits $c_{m,1}, \ldots, c_{m,s}, \ldots, c_{m,S}$, the bit-wise LLR corresponding to $c_{m,s}$ is calculated as

$$\lambda(c_{m,s}) = \ln\left[\frac{\sum_{\chi_q\in\mathcal{X}|c_s=1} q^{(T)}_{m,\mathrm{A}\to\mathrm{B}}(\chi_q)}{\sum_{\chi'_q\in\mathcal{X}|c_s=0} q^{(T)}_{m,\mathrm{A}\to\mathrm{B}}(\chi'_q)}\right], \tag{2}$$

where $\mathcal{X}|c_s = c$ ($c \in \{0,1\}$) denotes a set of candidate symbol points such that the s-th bit constituting χ_q is c.

D. Calculation of MI for Each TB

With the qualifier $[r,\ell]$, we denote the LLRs of the detector output corresponding to the coded bits that compose $x_m[r,\ell]$ by $\lambda(c_{m,s})[r,\ell], \forall s$, using (2). The number of coded bits $c_{m,s}[r,\ell]$ contained in a single TB is $N_{\mathrm{TB}} = R \times L \times S$, and by using all the corresponding LLRs, the demodulator output MI can be approximately computed as [16]

$$\mathcal{I}_{m,\mathrm{TB}} \approx 1 - \frac{1}{N_{\mathrm{TB}}}\sum_{r=1}^{R}\sum_{\ell=1}^{L}\sum_{s=1}^{S}\eta_{m,s}[r,\ell], \tag{3a}$$

$$\eta_{m,s}[r,\ell] \triangleq \log_2\left(1 + \exp\left[-(2c_{m,s}[r,\ell]-1)\lambda(c_{m,s})[r,\ell]\right]\right). \tag{3b}$$

If the MI of the demodulator output calculated by (3) can be predicted prior to signal transmission, it can be converted to the MI of the decoder output according to the decoding characteristics and used for appropriate MCS selection.

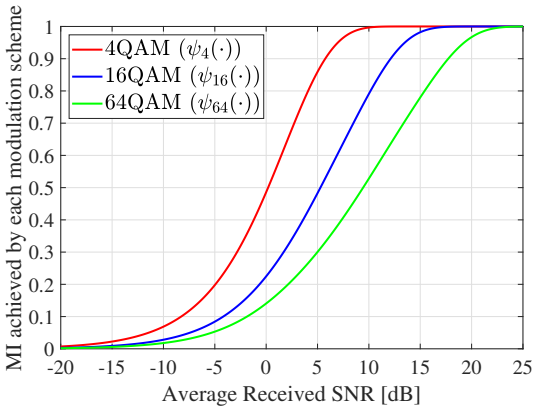


Fig. 2. SNR versus MI in AWGN channel.

III. MI Prediction After LMMSE-Based MUD

A. Analytical MI Prediction for LMMSE-Based MUD

When using the LMMSE detector, the SINR after MUD can be calculated analytically. First, in the detection of $x_m[r,\ell], \forall r$, from the RX signal $\boldsymbol{y}[r,\ell], \forall r$, in (1), the SINR of the LMMSE filter output can be expressed in closed-form as

$$\gamma_m[\ell] = \frac{1}{\left[\left(\rho\boldsymbol{H}^{\mathsf{H}}[\ell]\boldsymbol{H}[\ell] + \boldsymbol{I}_M\right)^{-1}\right]_{m,m}} - 1. \tag{4}$$

Next, by modeling the LMMSE filter output as an AWGN channel output with the same SNR as in (4), the post-MUD SINR for each subcarrier can be converted to the predicted MI of the demodulator output $\tilde{\mathcal{I}}^{Q}_{m}[\ell] = \psi_Q(\gamma_m[\ell])$, where $\psi_Q(\cdot), Q \in \{4,16,64\}$, is the conversion function shown in Fig. 2. By averaging this over all L subcarriers, we can predict the MI of the demodulator output $\mathcal{I}_{m,\mathrm{TB}}$ given by (3), which can be achieved when Q-QAM is used, as

$$\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}} = \frac{1}{L}\sum_{\ell=1}^{L}\tilde{\mathcal{I}}^{Q}_{m}[\ell]. \tag{5}$$

Finally, MCS selection is performed by calculating (5) for all modulation schemes (*e.g.*, $Q \in \{4,16,64\}$), converting each one to decoder output MI according to the decoding characteristics, and then selecting the maximum MCS index that achieves the target (reference) block error rate (BLER).

Based on the above descriptions, in MI-based MCS selection, accurate prediction of the MI of the demodulator output plays a critical role.

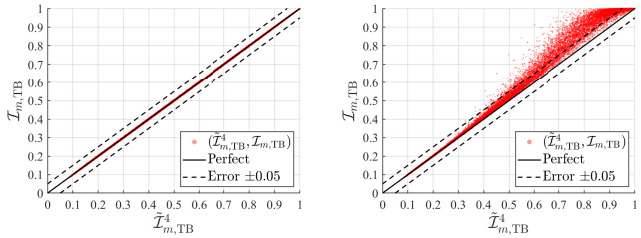
B. Evaluation of Prediction Accuracy

We have simulated a massive MU-MIMO-OFDM system and evaluated the accuracy of the demodulator output MI predicted using (5) with LMMSE-based MUD. The simulation conditions are summarized in Tab. I.

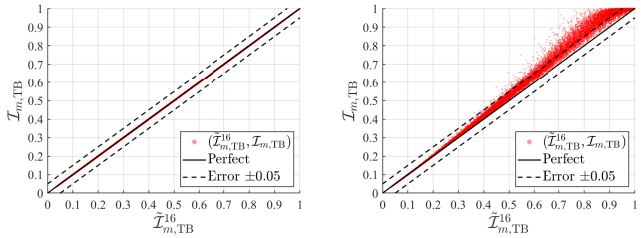
Suppose that we have R OFDM symbols per TB with each TB having N_{RB} RBs, and each RB consisting of N_c subcarriers. The number of subcarriers per TB is then given by $L = N_c \times N_{\mathrm{RB}}$. According to the RB model of the 5G NR standard [15], we have $R = 14$ and $N_c = 12$ (see Section II-A). Assuming that $N_{\mathrm{RB}} = 16$, we have $L = 192$ subcarriers

TABLE I
SIMULATION CONDITIONS IN THIS NUMERICAL SIMULATION

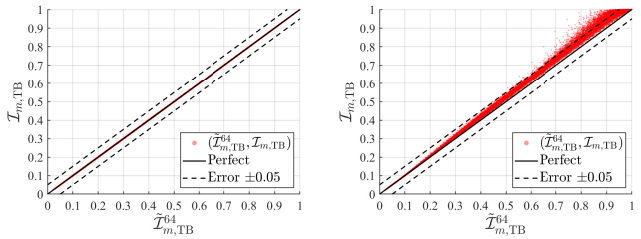
Item	Value
Center frequency	4.7 GHz
Bandwidth	100 MHz
Subcarrier spacing	30 kHz
Number of subcarriers	3072 (L = 192)
Delay spread	100 ns
Delay profile	3GPP CDL-B [17]
Modulation scheme	4QAM. 16QAM, 64QAM [18]
Channel coding scheme	LDPC [18]
MIMO configuration	$(N, M) = (64, 24)$



(a) LMMSE detection, 4QAM (b) EP detection, 4QAM



(c) LMMSE detection, 16QAM (d) EP detection, 16QAM

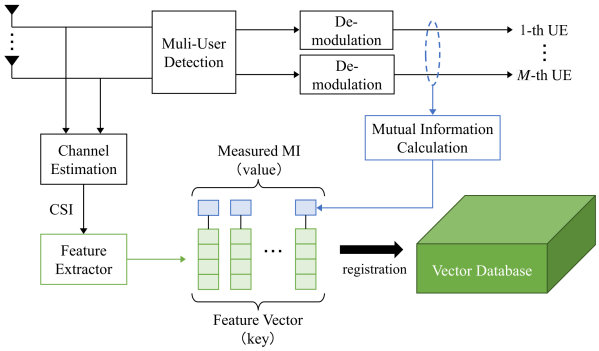


(e) LMMSE detection, 64QAM (f) EP detection, 64QAM

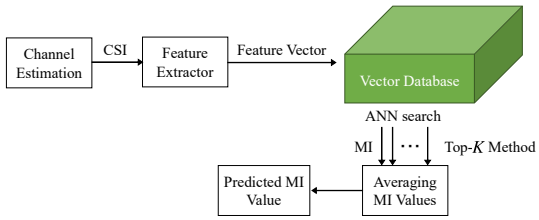
Fig. 3. Scatter plots of $\mathcal{I}_{m,\mathrm{TB}}$ versus $\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$.

and thus $N_{\mathrm{TB}} = R \times L \times S = 2688S$ (see Section II-D). For simplicity, the evaluation was performed assuming that all UEs use the same modulation scheme ($Q \in \{4, 16, 64\}$, and the average received SNR was set to the range from -10 dB to 24 dB.

Fig. 3 shows the scatter plots of the measured MI $\mathcal{I}_{m,\mathrm{TB}}$ calculated using (3) on the vertical axis and the predicted MI $\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$ calculated using (5) on the horizontal axis. Thus, the solid black line, which shows $\mathcal{I}_{m,\mathrm{TB}} = \tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$, represents the ideal case indicating perfect prediction, and the dashed black line, which shows $\mathcal{I}_{m,\mathrm{TB}} = \tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}} \pm 0.05$, represents the reference performance where the prediction error is 0.05. Figs. 3(a), (c), and (e) show the results when LMMSE detection is used as MUD (T = 1 in Algorithm 1) for every modulation scheme. As expected, the samples (red dots) are distributed on the ideal line, indicating that we can achieve highly accurate analytical MI predictions. On the other hand, Figs. 3(b), (d), and (f) show the results when EP detection (T = 4 in



(a) Offline VDB construction phase



(b) Online MI prediction phase

Fig. 4. MI prediction with VDB and ANN.

Algorithm 1) is used as MUD. It can be seen that almost all samples are above the solid line ($\mathcal{I}_{m,\mathrm{TB}} = \tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$) because the measured MI improves by the amount of the iterative gain compared to the predicted MI based on LMMSE detection. The above results suggest the following two vital insights: 1) when MCS selection is performed using (5) and the EP detector is used for MUD, highly reliable communications can be achieved while significantly suppressing the retransmission probability, and 2) if MI prediction with EP-based MUD can be performed with high accuracy, further improvement in system throughput can be expected by selecting an appropriate MCS.

In the following section, we propose a method for accurately predicting the MI with EP-based (*i.e.*, iterative) MUD.

IV. Proposed VSS-Based MI Prediction

Fig. 4 shows a schematic diagram of the proposed VSS-based MI prediction using ANN search on VDB. The proposed method consists of (a) an offline VDB construction phase and (b) an online MI prediction phase. In the construction phase (a), the feature vector (*key*) calculated from the CSI $\boldsymbol{H}$ and average received SNR ρ is linked to the value of MI (*value*), which is paired with the measured MI (*value*) when MU-MIMO-OFDM transmission is performed through that wireless channel and MUD is performed with the EP detector, and they are registered in VDB in vector format. This phase is usually carried out offline, whether in simulations or in real environments, but it is also possible to create VDB online using real data or to update VDB with additional data.

In the prediction phase (b), the key is generated in the same way from the estimated CSI and average received SNR, and the ANN search finds an arbitrary number of neighboring keys, and outputs the associated values. The values obtained in this way are the MIs achieved when actual transmissions are performed over similar wireless channels; therefore, if the feature vector is designed appropriately, it is possible to make highly accurate predictions. In this paper, we used *Faiss* [9],

Algorithm 2 VSS-based MI prediction using Top-K method

Input: Estimated CSI $\acute{\boldsymbol{H}}$, average SNR $\acute{\rho}$, VDB for Q-QAM
Output: Predicted MI $\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$

1: Calculate the feature vector $\acute{\boldsymbol{\upsilon}} = f(\acute{\boldsymbol{H}}, \acute{\rho})$
2: Execute Top-K ANN search with query $\acute{\boldsymbol{\upsilon}}$
3: As a result of the search, we obtain K pairs in order of similarity:
$$(\boldsymbol{\upsilon}_1, \varepsilon_1), (\boldsymbol{\upsilon}_2, \varepsilon_2), \ldots, (\boldsymbol{\upsilon}_K, \varepsilon_K)$$
4: Calculate average of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_K$: $\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}} = \frac{1}{K}\sum_{k=1}^{K} \varepsilon_k$

[12], which is one of the ANN libraries that operate quickly on GPUs.

A. Offline VDB construction Phase

The specific procedure is described as follows. The size of VDB is set to D_{VDB}, which means that the VDB can store D_{VDB} pairs of a feature vector (*key*) $\boldsymbol{\upsilon} \in \mathbb{R}^{D_{\mathrm{key}} \times 1}$ and acquired knowledge (*value*) $\varepsilon \in \mathbb{R}$, in the format of $(\boldsymbol{\upsilon}, \varepsilon)$.

In this paper, keys are generated from $\boldsymbol{H}$ and ρ; hence, the feature extraction process can be expressed as [7]–[9]

$$\boldsymbol{\upsilon} = f(\boldsymbol{H}, \rho), \tag{6}$$

where $f(\cdot)$ is the feature extraction function, and this design greatly affects the prediction performance. In addition, values are the measured MI $\mathcal{I}_{m,\mathrm{TB}}$ in (3).

B. Online MI Prediction Phase

Denoting the obtained CSI and average received SNR by $\acute{\boldsymbol{H}}$ and $\acute{\rho}$, respectively, the key is generated by (6) as

$$\acute{\boldsymbol{\upsilon}} = f\left(\acute{\boldsymbol{H}}, \acute{\rho}\right). \tag{7}$$

It is possible to search for a key that exists near $\acute{\boldsymbol{\upsilon}}$ in (7) and directly employ the associated values as the predicted MI. However, since ANN search is an approximate search, it does not necessarily find the optimal key. Therefore, in this paper, we obtain the predicted MI based on the Top-K method [13] in order to further improve the prediction accuracy.

The Top-K method is a technique that outputs the values associated with the top K keys with high approximate vector similarity, which are calculated internally during ANN search, and in this paper, we improve the prediction accuracy by using the average of the K predicted MI values as the final prediction value. The pseudo-code of the MI prediction using the Top-K method is given in Algorithm 2.

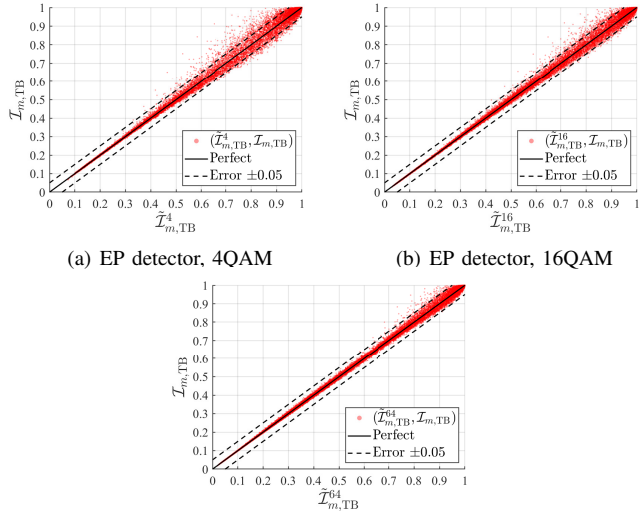
C. Design of Feature Vector

Although wireless channels vary according to stochastic events, there is a strong correlation among these variations in the time, frequency, and spatial domains. If we can design a feature vector (*key*) that can extract these characteristics appropriately, it should be possible to achieve accurate predictions with a much smaller database size than that required for other applications such as natural language processing and high-dimensional picture classification.

In this paper, we consider constructing a feature vector using the predicted MI with LMMSE-based MUD for each subcarrier, *i.e.*, $\tilde{\mathcal{I}}^{Q}_{m}[\ell]$ in (5). In order to reduce the dimensions of the feature vector D_{key}, instead of using all $L = 192$ subcarriers that compose 1 TB, only one subcarrier is used for each RB; thus, the feature vector of size $D_{\mathrm{key}} = 192/12 = 16$ is constructed as

$$\boldsymbol{\upsilon} = \left[\tilde{\mathcal{I}}^{Q}_{m}[12], \tilde{\mathcal{I}}^{Q}_{m}[24], \ldots, \tilde{\mathcal{I}}^{Q}_{m}[192]\right]^{\mathsf{T}} \in \mathbb{R}^{16\times 1}, \tag{8}$$

where ANN search on VDB can be interpreted as a transformation process from the MI with LMMSE-based MUD for every subcarrier to the MI with EP-based MUD for 1 TB.



(a) EP detector, 4QAM (b) EP detector, 16QAM

(c) EP detector, 64QAM

Fig. 5. Scatter plots of $\mathcal{I}_{m,\mathrm{TB}}$ versus $\tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$.

TABLE II
PREDICTION ACCURACY

Modulation scheme	4QAM	16QAM	64QAM
PE ±0.05	96.78 [%]	98.46 [%]	99.57 [%]
MSE	0.000320	0.000197	0.000101

D. Evaluation of Prediction Accuracy

Computer simulations were conducted to evaluate the prediction accuracy with the feature vector in (8). Simulation conditions are the same as those in Section III-B, *i.e.*, the parameters shown in Tab. I. The size of VDB was set to $D_{\mathrm{VDB}} = 134400$ for all modulation schemes. The number of data points (the number of plots) used for accuracy evaluation is 33600, and the wireless channel realizations used for evaluation are generated independently of those used for VDB construction. The parameter for the Top-K method was set to $K = 20$, which turned out to achieve the best accuracy in the pre-simulations.

Fig. 5 shows the scatter plots of pairs of the predicted MI and measured MI, as shown in Fig. 3. The MI of the demodulator output with EP-based MUD is predicted using the proposed method. Tab. II shows the percentage of the sample points that fall within the range of the prediction error (PE) ±0.05 and the mean square error (MSE) between the predicted and measured MIs. From Fig. 5, we can see that the samples are distributed around the ideal line of $\mathcal{I}_{m,\mathrm{TB}} = \tilde{\mathcal{I}}^{Q}_{m,\mathrm{TB}}$, and that the variance decreases at higher order modulation schemes. This can be confirmed numerically in Tab. II, where the percentage of cases where PE is within the range of 0.05 is over 95% for all modulation methods. The above results show that the proposed method using (8) as the key can accurately predict the MI with EP-based MUD.

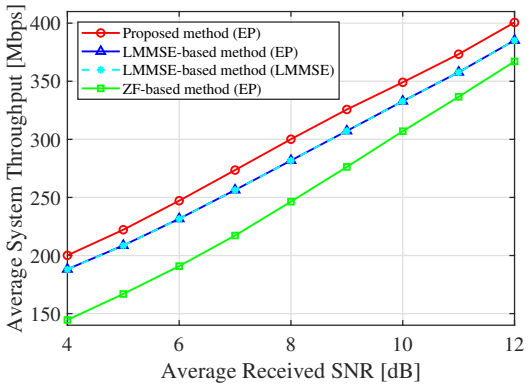


Fig. 6. Average received SNR versus average system throughput.

V. MCS SELECTION AND THROUGHPUT EVALUATION

Finally, we verified the improvement in system throughput when MCS selection is performed based on the MI predicted using the proposed method via computer simulations.

A. MI-Based MCS Selection

As described in Section III-A, MCS selection is achieved by predicting the MI of the demodulator output for all modulation methods listed in the MCS table, converting each of these to the MI of the decoder output according to the specific decoding characteristics, and selecting the MCS index that achieves the reference BLER. Due to space limitations, we omit the details, but the conversion from the demodulator output MI to the decoder output MI is performed using the extrinsic information transfer (EXIT) curve [16].

B. Throughput Evaluation

Simulation conditions are the same as those in Section III-B, as listed in Tab. I, and the VDB used for MI prediction is the same as that in Section IV. The MCS table used is compliant with 5G NR standard [19]. Fig. 6 shows the average system throughput as a function of the average receive SNR, where the following methods are compared:

- ZF-based method (*): MCS is selected based on the SINR predicted based on zero-forcing (ZF)-based MUD [3],
- LMMSE-based method (*): MCS is selected based on the MI predicted using (5) [4], [5],
- Proposed method (*): MCS is selected based on the MI predicted using the proposed VSS-based method,

with (*), specifying the detector used for MUD.

First, with the ZF-based method [3], the poor SINR prediction accuracy indicates that a significant drop in throughput is unavoidable without a correction method such as outer loop link adaptation (OLLA). Next, it is worth noting that the system throughput does not improve even when the MUD scheme is switched from the LMMSE detector to EP detector in the LMMSE-based method. This suggests that it is necessary to select the MCS index according to the MUD scheme in order to reflect the improvement in detection accuracy in the improvement of system throughput. Finally, it can be clearly confirmed that the system throughput improves when the MCS selection is performed using the proposed method in a MU-MIMO-OFDM system that uses the EP-based MUD.

VI. CONCLUSION

In this paper, we proposed a novel VSS-based MI prediction method for MCS selection in massive MU-MIMO-OFDM systems that employ iterative MUD schemes. The proposed method predicts the MI with MUD by performing ANN search on VDB using the feature vector computed from the CSI and average received SNR as the key. Computer simulations show that the proposed method can predict MI in the presence of EP-based MUD with high accuracy by constructing a feature vector from the analytically predicted MI with LMMSE-based MUD. In addition, it was shown that by selecting MCS based on the MI predicted using the proposed method, improvements in detection accuracy using EP-based MUD can be transformed to the improvements in system throughput.

ACKNOWLEDGEMENT

This work was supported in part by JST, CRONOS, Japan Grant Number JPMJCS24N1, and in part by MIC/FORWARD under Grant JPMI240710001.

REFERENCES

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.
[2] P. Bertrand, J. Jiang, and A. Ekpenyong, "Link adaptation control in LTE uplink," in *Proc. IEEE VTC Fall*, 2012, pp. 1–5.
[3] T. Doi, J. Shikida, D. Shirase, K. Muraoka, N. Ishii, T. Takahashi, and S. Ibi, "Outer loop link adaptation based on user multiplexing for generalized approximate message passing in massive MIMO," in *Proc. IEEE WCNC*, 2024, pp. 1–6.
[4] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Trans. Vehi. Technol.*, vol. 59, no. 8, pp. 3766–3778, 2010.
[5] X. Hu and X. Dai, "PER prediction in MIMO-OFDM systems with approximate demodulation: A GMI-based method," *IEEE Trans. Vehi. Technol.*, vol. 72, no. 9, pp. 11 548–11 557, 2023.
[6] 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, 2020.
[7] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.
[8] Y. Cao, H. Qi, W. Zhou, J. Kato, K. Li, X. Liu, and J. Gui, "Binary hashing for approximate nearest neighbor search on big data: A survey," *IEEE Access*, vol. 6, pp. 2039–2054, 2018.
[9] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
[10] L. Zhang, J. Tan, Y.-C. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning-based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3281–3294, 2019.
[11] Y. Liao, Z. Yang, Z. Yin, and X. Shen, "DQN-based adaptive MCS and SDM for 5G massive MIMO-OFDM downlink," *IEEE Commun. Letters*, vol. 27, no. 1, pp. 185–189, 2023.
[12] M. Douze *et al.*, "The Faiss library," 2024. [Online]. Available: ttps://arxiv.org/abs/2401.08281
[13] Z. Yang, A. Nakai-Kasai, and T. Wadayama, "Approximate nearest search with vector databeses for power allocation in interference channels," in *Proc. the 46th SITA*, 2023, pp. 238–243.
[14] L. Wei, T. Wadayama, and K. Hayashi, "Vector similarity search based offline learning for deep-unfolded MIMO signal detector," in *Proc. 2024 IEEE ICMLCN*, 2024, pp. 7–12.
[15] 3GPP, "5G;NR;physical channels and modulation (3GPP TS 38.211 version 16.2.0 release 16)," July 2020.
[16] S. Ibi, T. Matsumoto, R. Thomä, S. Sampei, and N. Morinaga, "EXIT chart-aided adaptive coding for multilevel BICM with turbo equalization in frequency-selective MIMO channels," *IEEE Trans. Vehi. Technol.*, vol. 56, no. 6, pp. 3757–3769, Nov. 2007.
[17] 3GPP, "5G;study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 16.1.0 release 16)," Nov. 2020.
[18] ——, "5G;NR;multiplexing and channel coding (3GPP TS 38.212 version 16.2.0 release 16)," July 2020.
[19] ——, "5G;NR;physical layer procedures for data (3GPP TS 38.214 version 16.5.0 release 16)," April 2021.