Information-Optimum Approximate Message Passing for Quantized Massive MIMO Detection

LIWEN WANG¹, (Member, IEEE), TAKUMI TAKAHASHI¹, (Member, IEEE), SHINSUKE IBI², (Member, IEEE), AND SEIICHI SAMPEI¹, (Fellow, IEEE)
¹Graduate School of Engineering, Osaka University, Suita 565-0871, Japan
²Faculty of Science and Engineering, Doshisha University, Kyotanabe 610-0394, Japan
Corresponding author: Takumi Takahashi (takahashi@comm.eng.osaka-u.ac.jp)

This work was supported in part by the JSPS KAKENHI, Japan, under Grant JP18H03765 and Grant JP19K23516.

ABSTRACT We propose an information-optimum approximate message passing (AMP) for quantized massive multi-input multi-output (MIMO) signal detection. A well-known strategy for realizing low-complexity and high-accuracy massive multi-user detection (MUD) is AMP-based belief propagation (BP). However, when internal operations are conducted with double-precision arithmetic, large memory occupancy and severe processing delay are inevitable in the actual massive MIMO implementation. To address this issue, we replace all operations with a simple look-up table (LUT) search where all messages exchanged between each iteration process are unsigned integers. That is, the proposed signal detection is performed using only simple integer arithmetic. The LUT is designed offline using an information-bottleneck (IB) method, and the probability distribution of messages at each iteration step is required for determining the quantization threshold tracked by discrete density evolution (DDE). Computer simulations demonstrate the validity of the IB LUT-based AMP in terms of bit error rate (BER) performance and memory occupancy. The proposed method allows quantizing the AMP detector with fewer bits while maintaining similar performances, such as that of a typical AMP with double-precision.

INDEX TERMS Massive multi-input multi-output (MIMO), approximate message passing (AMP), iterative detection, quantization, information-bottleneck (IB), look-up table (LUT) search.

I. INTRODUCTION

Compared to typical small-scale multi-antenna systems, massive multi-input multi-output (MIMO), where the base station (BS) is equipped with a massive number of antenna elements, promises significant improvements in spectral efficiency, detection reliability, and energy efficiency [1]–[3]. Increasing the number of BS antenna elements leads to high spatial resolution. This makes it possible to serve a massive amount of wireless links in the same time-frequency resource simultaneously, thereby allowing massive connectivity in uplink multi-user MIMO (MU-MIMO) scenarios. However, high-dimensional signal separation, i.e., large-scale multi-user detection (MUD), raises the computational cost and circuit scale at the receiver considerably [3], [4].

Maximum likelihood detection (MLD) is infeasible in large-scale MUD because of the prohibitive computational burden involved. As low-complexity MUD solutions, linear spatial filters based on least square (LS) and linear minimum mean square error (MMSE) criteria are often utilized while sacrificing optimal detection capability [3].

To improve detection reliability without requiring high computational cost, a message passing (MP) algorithm based on belief propagation (BP) has been investigated [3], [5]–[11]. The BP-based signal detection can take advantage of the law of large numbers to reduce marginalization burden and simplify computations of the conditional symbol expectation. The most well-known BP-based detector is approximate message passing (AMP) [5], which is systematically derived from a strict approximation of Gaussian belief propagation (GaBP) [10], [11] in the large-system limit that occurs when input and output dimensions, N and M respectively, increase to infinity for a given compression rate ρ = N/M. In AMP, a rigorous proof of convergence to the Bayes optimal was presented under independent and identically distributed (i.i.d.) Gaussian measurement matrices [6]. These algorithms can achieve the computational complexity order
of $O(MN)$ for each iteration process because they comprise only scalar-wise tensor product operations.

Thus far, it is difficult to further reduce the computational complexity by simplifying signal processing. Instead, the present study focuses on an approach to avoid internal detection operations with double-precision arithmetic for practical hardware implementation. All signal detection operations are replaced by simple look-up table (LUT) search such that iterative detection performs with only simple integer arithmetic. This makes it possible to reduce memory usage and computational cost significantly, resulting in low-latency signal processing for low-cost receiver design. In this case, the most vital issue is determining how to design the LUT with a limited number of quantization bits without severely sacrificing detection reliability. The most notable method for this approach is information-optimum discrete signal processing [12]–[16], where LUTs are designed using an information-bottleneck (IB) method [17], which is a more generic framework with roots in machine learning and information theory [18]. The idea is to preserve as much information on the variable of interest as possible. Thus, the maximum possible relevant mutual information (MI) needs to be preserved at each iteration step. References [12]–[14] exploit the IB method and discrete density evolution (DDE) scheme [19]–[21] to log-likelihood ratio (LLR) beliefs in the low-density parity-check (LDPC) decoder. Inspired by these methods, we design a novel LUT-based iterative MUD via AMP.

To apply the IB method used in the quantization of LDPC decoding to the LUT design of AMP, there are some difficulties to overcome. First, unlike LDPC decoding, the beliefs exchanged in the AMP-based MUD are fluctuated because of the stochastic behavior of random MIMO channels. Intuitively, the LUT should be constructed adaptively according to the instantaneous channel state information (CSI). However, the CSI-dependent LUT has to be updated at any time the CSI changes, which can lead to a catastrophic processing delay. To remove the dependency of LUTs on the CSI, we focus on the statistical stability of beliefs in AMP under large-system conditions. Because of strong channel hardening effects [3] and well-averaged AMP rule, the CSI-independent LUTs can be realized.

Second, it is difficult to track the probability distribution of beliefs using DDE. In the IB method, the relationship between the discrete numbers to the variable of interest is linked via a joint probability distribution, and its evolution via iterative processing is tracked by DDE. In LDPC decoding, both the check-nodes (CNs) and variable-nodes (VNs) refer to binary values as original referencing values. In this case, the joint probability is simple to track in quantized discrete decoding because a number of possible combinations is well-limited. However, in the AMP-based MUD, the original values referred in the function-nodes (FNs) change continuously and dynamically because of channel coefficients, which implies the number of possible combinations is infinite. Under this condition, the joint probability distribution is not calculable, and the DDE loop is not closing. To tackle this issue, we extend the typical binary class DDE to multi-class, where the probability mass function (PMF) of quantized values is computed based on the probability density function (PDF) of corresponding continuous values, which is unlike the original binary DDE. This is possible because the PDF of beliefs in AMP can be approximated to the Gaussian distribution with high-accuracy in large-system conditions.

Finally, severe performance degradation is inevitable because of the quantization errors even if IB-based quantization is performed faithfully based on the typical AMP algorithm. However, it is difficult to capture the behavior of such non-linear errors in iterative processing. Therefore, we introduce an adjusting parameter to suppress the negative effect of quantization errors. It is possible to modify the average variance of beliefs in AMP by designing the parameter as an equivalent variance of quantization noise.

Given the above, the contributions of this paper are summarized as follows.

- CSI-independent LUT-based AMP for quantized massive MIMO detection, where the LUTs are designed offline using the IB and DDE methods based on long-term channel statistics, is proposed.
- To overcome the limitation of the DDE because of the continuity of the channel coefficient, the quantization methods in [12]–[14] are modified to an application to MUD with the aid of large-system approximations used in the AMP-based signal recovery.
- Adjusting parameters are introduced to suppress the negative effect of inevitable quantization errors and are designed by computer simulations.

We emphasize that, to the best of the authors’ knowledge, a LUT-based MUD mechanism based on the IB method and DDE scheme for massive MIMO systems, which is the key contribution of this paper, has not been proposed yet.

A. RELATED WORKS

As described above, the proposed detector is given by the LUT-based finite-precision MP algorithm. One way of dealing with such a coarse quantization of iterative processing is the finite-alphabet iterative decoding (FAID) [22], [23], where the LUTs are hand-optimized to approximate the actual arithmetic operations as accurately as possible. The FAID approach is restricted to regular LDPC codes, but the resulting quantized decoding was shown to achieve competitive performances on a binary symmetric channel (BSC) with the typical decoding, despite coarse quantization.

The IB method is a fundamentally different way to learn the LUTs in an unsupervised manner [24], where its discrete input-output relationship is designed for maximizing the MI between the quantized message and the corresponding coded bit, and the origin of this approach dates back to [20], [25]. In this line of work, the IB-based decoders for regular LDPC codes were proposed in [12], [16] and found to yield improved decoding capability in comparison
with the aforementioned counterparts with coarse quantization. In [26], the implementation results show that a similar LUT-based decoder delivers throughput up to 0.588-Tb/s with high energy and area efficiency. The IB-based decoders for irregular codes were presented in [15], and the extension to protograph-based raptor-like (PBRL) LDPC codes [27] enabling inherent rate-compatible decoding was proposed in [28]. Besides the above, in [29], the information-optimum decoding for polar codes was presented.

While the finite-precision MP algorithm in the context of channel decoding has been actively investigated, to the best of our knowledge, such a quantized MP algorithm for MUD has not been presented yet in the literature. Therefore, the framework of LUT-based AMP approach is expected to pose a new angle to related research topics.

The rest of this paper is organized as follows: Section II presents the Bayes-optimal AMP algorithm as an equivalent real-valued representation of a complex-valued MU-MIMO system. Section III presents the LUT structure to realize a significant reduction of the required memory occupancy. Section IV proposes CSI-independent LUT construction using IB LUT. In Section V, computer simulations are conducted to characterize the validity of the proposed methods in terms of bit error rate (BER) and memory occupancy. Finally, Section VI concludes the paper with a brief summary.

II. SYSTEM MODEL

A. MATHEMATICAL NOTATIONS

Throughout this paper, vectors and matrices are denoted by lower- and upper-case bold-face letters, respectively. Further, \( T \) and \( H \) are transpose and conjugate transpose operators, respectively. \( P(a|b) \) and \( p(a|b) \) denote the conditional PMF and PDF of the random variable \( a \) given the occurrence of a realization \( b \). \( P(a, b) \) and \( p(a, b) \) denote the joint PMF and PDF, respectively. \( N(a, b) \) and \( CN(a, b) \) indicates i.i.d. real-valued and complex-valued Gaussian distribution with mean \( a \) and variance \( b \), respectively. \( \mathbb{E}\{\cdot\} \) represents an expected value, \( \mathbb{R}^{a\times b} \) and \( \mathbb{C}^{a\times b} \) denote real and complex fields of size \( a \times b \), respectively. \( \mathcal{N}\{\cdot\} \) and \( \mathcal{C}\{\cdot\} \) indicate the real and imaginary parts of complex values, respectively. \( I_D \) is a \( D \times D \) identity matrix. \( [\cdot] \) indicates the floor function.

B. SIGNAL MODEL

Consider an uplink massive MU-MIMO system, where the BS has \( N \) receive (RX) antennas and \( M' \) (\( \leq N \)) user equipment (UE) have a single transmit (TX) antenna. The \( m \)-th UE conveys a modulated symbol \( X_m^c \), which represents one of \( Q \) constellation points \( \mathcal{X} = \{X_1, \ldots, X_q, \ldots, X_Q\} \). The abovementioned complex-valued model can be rewritten using a double-size equivalent real-valued model as

\[
    y = Hx + z,  \tag{2}
\]

where we define \( M = 2M' \) and \( N = 2N \).

For the ease of mathematical notations, we use Gray-coded quadrature phase shift keying (QPSK) signaling (\( \mathcal{X}_c = \{\pm\sqrt{E_c/2}, \pm j\sqrt{E_c/2}\} \)), where the average power density of constellations in the set \( \mathcal{X}_c \) is defined as \( E_c \). Let \( X_c = [X_1^c, \ldots, X_q^c, \ldots, X_Q^c]^T \in \mathbb{C}^{M'\times 1} \) and

\[ y_c = [y_1^c, \ldots, y_q^c, \ldots, y_N^c]^T \in \mathbb{C}^{N\times 1} \]

denote TX and RX symbol vectors, respectively. Assuming flat Rayleigh fading channels, the RX vector \( y_c \) is given by

\[
    y_c = H_c^\dagger x_c + z_c,  \tag{1}
\]

where \( H_c \in \mathbb{C}^{N\times M} \) is an \( N' \times M' \) MIMO channel matrix. The element in the \( n \)-th row and \( m \)-th column, \( H_{n,m}^c \), of \( H_c \) obeys \( \mathcal{CN}(0, \phi_H) \), where \( \phi_H = 1 \). The vector \( z_c = [z_1^c, \ldots, z_q^c, \ldots, z_N^c]^T \in \mathbb{C}^{N\times 1} \) is a complex additive white Gaussian noise (AWGN) vector whose entries obey \( \mathcal{CN}(0, N_0) \), where \( N_0 \) is the noise spectral density; the covariance matrix is given by \( \mathbb{E}\{z_c(z_c)^H\} = N_0 I_N \).

The above-mentioned complex-valued model can be simplified using (2). The LUT designs, which will be explained in the next iteration process is generated by the Bayes-optimal denoiser; further, its variance is computed.
Algorithm 1 Bayes-Optimal AMP [5]

Input: \( y, H, \phi_H, N_0, K \)

Output: \( \hat{x} = [\hat{x}_1, \ldots, \hat{x}_m, \ldots, \hat{x}_M]^T \)

// Initialization
1: \( \mathcal{M} = \{1, \ldots, M\}, \mathcal{N} = \{1, \ldots, N\} \)
2: \( y_n^{(0)} = 0, \forall n \in \mathcal{N}, y_m^{(0)} = 0, \forall m \in \mathcal{M} \)
3: \( \psi(0) = 0, \phi(1) = E_s/2 \)

// Iteration
4: for \( k = 1 \) to \( K \) do
5: \( \hat{j}_n^{(k)} = \sum_{i=1}^{M} h_{n,i} \hat{x}_i^{(k)}, \forall n \in \mathcal{N} \)
6: \( \hat{y}_n^{(k)} = y_n - \hat{j}_n^{(k)} + \frac{\phi_H}{\psi(1)} \phi^{(k)}(k-1), \forall n \in \mathcal{N} \)
7: \( \psi^{(k)}(k) = \frac{N_0}{2} + M \phi_H \phi^{(k)}(k) \)
8: \( \hat{x}_m^{(k)} = \sqrt{\frac{E_s}{2}} \tanh \left[ \frac{N_0 \sqrt{4M}}{4 \phi^{(k)}(k)} \right], \forall m \in \mathcal{M} \)
9: \( \hat{j}_m^{(k+1)} = \sqrt{\frac{E_s}{2}} - \frac{1}{M} \sum_{i=1}^{M} (\hat{x}_i^{(k+1)})^2 \)
10: \( \phi^{(k)}(k+1) = E_s - \frac{1}{M} \sum_{i=1}^{M} (\hat{x}_i^{(k+1)})^2 \)
11: end for

// Decision
12: \( \hat{x}_m = \sqrt{\frac{E_s}{2}} \text{sgn} \left[ \hat{x}_m^{(K)} \right], \forall m \in \mathcal{M} \)

FIGURE 1. Belief propagation in AMP.

The pseudo-code is provided in Algorithm 1. For the ease of notations, the equation in the \( i \)-th line of Alg. 1 is denoted by (A1-\( i \)). In the AMP, the variance of beliefs computed in FNs and VNs, \( \psi^{(k)}(k) \) and \( \phi^{(k)}(k) \), are averaged in the large-system limit, respectively. Furthermore, the harmful effect of self-feedback because of the loopy propagation of beliefs is averaged by \( \frac{M}{2} \phi_H \phi^{(k)}(k-1) \) in (A1-6); this term is referred to as the "\( \alpha \)"-onsager term [5]. Owing to this well-approximated message passing rule, the statistical behavior of beliefs is not largely dependent on instantaneous channel fluctuations; this makes it possible to construct CSI-independent LUT.

D. QUANTIZATION

To clarify how quantization is conducted with predetermined thresholds, we show the quantization process of an arbitrarily variable \( a \in \mathbb{R} \) using Fig. 2. Let \( u \in \mathbb{Q} = \{1, \ldots, 2^B\} \) denote the quantization index of \( a \), where \( B \) is the number of available bits. The threshold is represented by \( \hat{a} [u] \in \mathbb{R} = \mathbb{R} \cup (-\infty, \infty) \) for \( u \in \mathbb{Q} \), where \( \hat{a}[2^B] = \infty \). In this case, the quantization function is defined as

\[
\hat{a} = Q(a), \quad \text{if } \hat{a}[u-1] < a \leq \hat{a}[u],
\]

where we define \( \hat{a}[0] = -\infty \) as the left-sided threshold for the quantization level \( u = 1 \). Here, we introduce the representative value \( \hat{a}[u] \in \mathbb{R} \) to indicate a real number value corresponding to each index \( u \in \mathbb{Q} \), where \( \hat{a}[u-1] < \hat{a}[u] \leq \hat{a}[u+1] \) holds. Hereafter, we present two criteria for determining the quantization threshold: the maximal entropy criterion and the IB-based maximal MI criterion.

1) MAXIMAL ENTROPY CRITERION

In the maximal entropy criterion, \( a \) is quantized such that each quantization level holds equal probability mass, which leads to well-known equiprobable quantization (EPQ) [30]. Because of large-system approximation, the probability distribution of beliefs in Alg. 1 can be regarded as a Gaussian distribution. Therefore, the quantization threshold can be easily determined based on the CDF of Gaussian distribution, namely the error function \( \text{erf}(\cdot) \). This criterion is similar to the FAID approach [22], [23] in that it mimics the results of each arithmetic operation as accurately as possible.

For considering the negative effect of quantization noise, the pseudo-Gaussian noise \( n_q \sim \mathcal{N}(0, N_0) \) is introduced into the reference PDF at the quantization step. That is, the effective noise density is given by

\[
N_0^{\text{REF}} = N_0 + N_q.
\]

2) INFORMATION-BOTTLENECK (IB)-BASED MAXIMAL MUTUAL INFORMATION (MI) CRITERION

Given a joint probability distribution \( p(a, b) \), the IB method proposed in [17], [31] is utilized to determine a compressed representation of \( b, \hat{b} \), which results in the least information loss of the original information \( a \). The target mapper \( p(\hat{b}|b) \) can be found by minimizing the IB function, which is defined as

\[
\mathcal{F} \{ p(\hat{b}|b) \} = I(b; \hat{b}) - \eta I(a; \hat{b}),
\]
where $\eta$ is used to adjust the preference between the compression rate and information loss.

During the construction of the CSI-independent LUT, the quantization setting is not optimal because of the use of an average threshold. Thus, a part of the information is lost each time quantization is conducted. Therefore, to suppress the information loss, it is necessary to utilize not only the distribution of the observed value $b$, but also the original information of interest $a$. In analogy with the rate distortion theory [32], in the IB scenario, by setting the distortion function as the lost mutual information of $b$ and $a$, the IB function (10) is obtained by

$$ R(D) = \min_{p(\hat{b}|b): I(\hat{b}; b) \geq D} I(\hat{b}; b). \quad (11) $$

(10) can also be viewed such that the first term indicates the compression rate and the second term indicates the amount of information $\hat{b}$ contains about $a$. The IB method optimizes the quantization rule that causes minimum MI loss as we set $\eta = \infty$. There are several algorithms proposed in [31] to solve the optimization problem of (10). In this paper, we utilize the sequential IB (sIB) algorithm modified in [33].

### III. LUT WITH SMALL MEMORY USAGE

The simplest approach to building an LUT-based AMP is converting each process of Alg. 1 into a single large-LUT with multiple inputs and outputs. However, in this case, large memory usage is inevitable. For example, when the RX symbol $y_n$ is quantized into $B$ bits, $(N+1)B$ bits are required to represent all possible combinations in (A1-6), which is prohibitively large for practical implementation.

To avoid the impairement, we further divide each process in Alg. 1 into multiple basic arithmetic operations. Thus, the large-LUT with multiple inputs and outputs are divided into the successive computations of two-dimensional LUTs with two inputs and one output. This 2-to-1 LUT is treated as the basic unit of the quantized discrete AMP.

#### A. BASIC UNIT CONSTRUCTION

We define $u_1 \in \mathbb{Q}$, $u_2 \in \mathbb{Q}$, and $w \in \mathbb{Q}$ as quantization indices for the two inputs and one output of the 2-to-1 LUT. The LUT functions $w = \Sigma(u_1, u_2)$ and $w = \Pi(u_1, u_2)$ provide the quantized sum $c = a + b$ and the quantized product $c = a \cdot b$ operations, respectively. In the quantized sum, the 2-to-1 LUT construction is performed as follows:

- All possible combinations of the representative value are listed according to

$$ \hat{c}[(u_1, u_2)] = \hat{a}[u_1] + \hat{b}[u_2] \in \mathbb{R}, \quad (12) $$

where $(u_1, u_2)$ is a two-dimensional index.

- After sorting $\hat{c}[(u_1, u_2)]$ in the ascending order, it is re-indexed as $\hat{c}[\hat{w}]$ with $\hat{w} \in \mathbb{Q} = \{1, \ldots, 2^B\}$.

- According to the quantization threshold, $\hat{c}[\hat{w}]$ is quantized to $B$ bits as $w = Q(\hat{c}[\hat{w}])$.

Fig. 3 shows the 2-to-1 LUT construction in the case of $B = 2$. The entry in the $u_1$-th row and $u_2$-th column of table (d) provides the quantization index of the output value $w$.

#### B. MULTI-LAYER LUT STRUCTURE

Fig. 4 shows the multi-layer 2-to-1 LUT structure for representing the summation of $T$ terms, $\sum_{t=1}^{T} a_t$, where the tree-depth (= Number of layers) is $L$. At the first layer, $T$ terms are divided into $\lfloor T/2 \rfloor$ nodes. Each node contains two terms as inputs and builds the 2-to-1 LUT. At the subsequent layer, the outputs from the previous layer are considered as new inputs, and the same operations are conducted. When the number of inputs is not divisible by two, an additional 2-to-1 LUT operation for one node output and the remaining input is conducted. The expected sum is obtained after the $L$-th layer processing is completed.

Owing to this multi-layer structure, the LUT size required for the operation with $T$ inputs is reduced from $2^{BT}$ to $2^B \cdot (T-1)$, which makes a considerable difference, especially in massive MIMO configurations. Furthermore, when LUTs at the same tree depth are approximated to be identical, memory usage is compressed further. In this case, the required LUT size can be reduced to $2^B \cdot \lceil \log_2 T \rceil$.

### IV. IB LUT FOR MAXIMAL MI USING IB AND DDE

With the 2-to-1 LUTs, we quantize each process in Alg. 1 based on the IB-based maximal MI criterion. The quantization rule is designed to depend on only long-term channel statistics, i.e., the stochastic channel (correlation) model, and therefore, the LUT construction can be pre-designed. Thus, LUT reconstruction requires to be updated when only the propagation environments change widely.
A. PRELIMINARIES

In the IB method, it is necessary to compute the joint probability of values and its referencing original values. In [20], the DDE is applied to track the transition of the joint PMF of beliefs in each iteration step of LDPC decoding. However, there are some issues when applying the IB and DDE to the LUT construction of AMP-based MUD.

In the LDPC decoder, all operations conducted in the FNs and VNs are LLR belief combining, and they can be directly linked to each other in the LLR domain. The AMP-based MUD requires various node processing in addition to the belief combining of (A1-8) and (A1-5). The FNs conduct soft IC using (A1-6) in the symbol domain, and the VNs generate soft symbol replicas using (A1-9), which plays the role of bridging belief-domain and symbol-domain processing. Therefore, some methods are required to link the entire processing as in LDPC decoding.

The RX symbols \( y_n, \forall n \) and the channel coefficients \( h_{n,m}, \forall n, m \) are respectively quantized into \( B \) bits based on the maximal entropy criterion in advance, using \( y_n \sim \mathcal{N}(0, \psi^{(1)}) \) and \( h_{n,m} \sim \mathcal{N}(0, \frac{\phi}{2}) \). Using the predefined quantization index \( u \in \mathbb{Q} \), we denote the corresponding representative values by \( \hat{y}_n[u] \in \mathbb{R} \) and \( h_{n,m}[u] \in \mathbb{R} \).

B. BELIEF COMBINING IN VNS (A1-8)

Let us consider the belief combining in (A1-8). This processing is similar to the LLR combining in VNs of LDPC decoding because the referencing original value is binary. Thus, the joint probability can be obtained easily because the number of possible combinations is sufficiently limited. Here, we focus on the processing in one 2-to-1 LUT, and all LUTs consisting of the multi-layer structure work with the same rule.

Consider a 2-to-1 LUT with two inputs \( f_1[u_1] \) and \( f_2[u_2] \), and one output \( F[w] \) for belief combining in VNs. \( F[w] \) denotes all possible combinations of two inputs with 2B bits. The joint PMF required for the IB method is given by

\[
P[F[w], x_m] = \frac{P[f_1[u_1], x_m] \cdot P[f_2[u_2], x_m]}{P[x_m]}. \tag{13}
\]

Hereafter, we explain how the joint PMFs \( P[f_1[u_1], x_m] \) and \( P[f_2[u_2], x_m] \) are computed. At the first layer processing (\( l = 1 \)), we focus on a part of the first summation-term in (A1-8), given by \( \hat{\alpha}^{(1)}_{n,m}[w] = h_{n,m}[u_1] \tilde{\alpha}^{(1)}_{n,m}[u_2] \), which is the operation for computing all possible MF outputs with the quantized channel and RX symbol inputs. To determine the B-bit quantization threshold for \( \hat{\alpha}^{(1)}_{n,m}[u_2] \) using the IB method, it is necessary to compute the joint PMF \( P[\hat{\alpha}^{(1)}_{n,m}[w], x_m] \) based on the PDF of the corresponding continuous value, i.e., \( \hat{\alpha}^{(1)}_{n,m} = h_{n,m} \gamma_n \).

At the first iteration step \( (k = 1) \) in AMP, \( \alpha^{(1)}_{n,m} \) is given by a simple MF output as

\[
\alpha^{(1)}_{n,m} = h_{n,m} \gamma_n = h_{n,m}^2 x_m + h_{n,m} \left( \sum_{i \neq m} h_{n,i} x_i + z_n \right). \tag{14}
\]

Under sufficient large-system conditions, the PDF of \( \hat{\alpha}^{(1)}_{n,m} \) in (14) approaches an i.i.d. real-valued Gaussian distribution as a result of the CLT, which leads to \( p(\hat{\alpha}^{(1)}_{n,m}|h_{n,m}, x_m) \sim \mathcal{N}(h_{n,m}^2 x_m, h_{n,m}^2 \psi^{(1)}) \); this stochastic behavior is referred to as scalar Gaussian approximation (SGA) [3].

Similarly, at the second iteration step \((k = 2)\), \( \alpha^{(2)}_{n,m} \) can be expressed using \( \tilde{\gamma}^{(1)}_n = y_n \) as

\[
\alpha^{(2)}_{n,m} = h_{n,m} \left( y_n - \sum_{i=1}^M h_{n,i} \tilde{\gamma}^{(2)}_i + \beta^{(2)} \tilde{\gamma}^{(1)}_n \right) = h_{n,m} \left( y_n - \sum_{i=1}^M h_{n,i} \tilde{\gamma}^{(2)}_i + h_{n,m} \beta^{(2)} \left( h_{n,m} x_m + \sum_{i \neq m} h_{n,i} x_i + z_n \right) \right) + h_{n,m} \beta^{(2)} \left( \sum_{i \neq m} h_{n,i} x_i + z_n \right), \tag{15}
\]

where the Onsager-term is denoted by \( \beta^{(k)} \psi^{(k-1)} \) to simplify mathematical notations. \( \beta^{(k)} = \frac{M \gamma}{2} \phi \frac{\psi^{(k-1)}}{\psi^{(k)}} \) is the Onsager-coefficient. Under the SGA conditions, \( p(\hat{\alpha}^{(2)}_{n,m}|h_{n,m}, x_m) \) obeys \( \mathcal{N}(\mu_{\alpha}^{(2)}, \psi_{\alpha}^{(2)}) \) where the equivalent average and variance are respectively given by

\[
\mu_{\alpha}^{(2)} = h_{n,m}^2 \mu_{\alpha}^{(2)}, \quad \psi_{\alpha}^{(2)} = \frac{\phi \mu_{\alpha}^{(2)}}{2} \left( \beta^{(2)} \psi_{\alpha}^{(1)} + \psi^{(2)} \right). \tag{17}
\]

Similarly, at the \( k \)-th iteration step \((k \geq 2)\), we formulate the equivalent average and variance of \( p(\hat{\alpha}^{(k)}_{n,m}|h_{n,m}, x_m) \) by the following recurrence formulas as

\[
\mu_{\alpha}^{(k)} = h_{n,m}^2 x_m \prod_{i=2}^k \beta^{(i)}, \quad \psi_{\alpha}^{(k)} = \frac{\phi \mu_{\alpha}^{(k)}}{2} \left( \beta^{(k)} \psi_{\alpha}^{(k-1)} + \psi^{(k)} \right), \tag{19}
\]

where \( \psi_{\alpha}^{(1)} = \psi^{(1)} \).

Based on the above PDF, the conditional PMF given \( h_{n,m}[u_1] \) and \( x_m \) can be computed by

\[
P[\hat{\alpha}^{(k)}_{n,m}[w]|h_{n,m}[u_1], x_m] = \frac{P[\hat{\alpha}^{(k)}_{n,m}[w] < \alpha^{(k)}_{n,m} \leq \hat{\alpha}^{(k)}_{n,m}[w], h_{n,m}[u_1], x_m]}{P[\hat{\alpha}^{(k)}_{n,m}[w] < \alpha^{(k)}_{n,m} \leq \hat{\alpha}^{(k)}_{n,m}[w], h_{n,m}[u_1]]} = \frac{1}{2} \text{erf} \left( \frac{\hat{\alpha}^{(k)}_{n,m}[w] - \mu_{\alpha}^{(k)}[u_1]}{\sqrt{2} \psi_{\alpha}^{(k)}} \right) - \frac{1}{2} \text{erf} \left( \frac{\hat{\alpha}^{(k)}_{n,m}[w] - 1 - \mu_{\alpha}^{(k)}[u_1]}{\sqrt{2} \psi_{\alpha}^{(k)}} \right), \tag{20}
\]
where we have
\[ \hat{\mu}^{(k)}_{a}(x_1) = x_m \cdot \hat{h}^{2}_{n,m}(u_1) \prod_{i=1}^{k} \alpha^{(i)} \]  \tag{21}

Here, we define \( \gamma^{(k)} \) as the Onsager-coefficient at the \( k \)-th iteration in the discrete LUT-based AMP. The design of \( \gamma^{(k)} \) is presented in Section IV-E.

The dependency on the instantaneous channel coefficient can be removed by marginalization processing as
\[ P\left[ \hat{\omega}^{(k)}_{n,m}[\hat{w}]; x_m \right] = \sum_{a_0=1}^{2^{n}} P\left[ \hat{\omega}^{(k)}_{n,m}[\hat{w}]; \hat{h}_{n,m}[u_1], x_m \right] P\left[ \hat{h}_{n,m}[u_1] \right] \]  \tag{22}

Consequently, we can derive the joint PMF required for the IB method to determine the threshold of \( \hat{\omega}^{(k)}_{n,m}[\hat{w}] \) as
\[ P\left[ \hat{\omega}^{(k)}_{n,m}[\hat{w}]; x_m \right] = P\left[ \hat{\omega}^{(k)}_{n,m}[\hat{w}] | x_m \right] \cdot P\left[ x_m \right] \]  \tag{23}

After determining the threshold, the resultant \( B \)-bit quantized values and its PMF (equivalent to \( P[f_{1}(u_1), x_m] \)) become the input for the 2-to-1 summation LUT at the next layer.

After successively computing the quantized sum based on the multi-layer structure until the first summation-term in (A1-8) is quantized, we achieve the \( B \)-bit quantized belief \( \hat{\gamma}^{(k)}_{m}[w] \) and its joint PMF \( P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_m \right] \), which leads to the multi-layer LUT for (A1-8).

### C. SOFT SYMBOL REPLICA GENERATION (A1-9,10)
Using \( P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_m \right] \) passed from the LUT for (A1-8), the soft symbol replica can be directly computed by the conditional expectation as
\[ \hat{\gamma}^{(k+1)}_{m}[w] = \sum \chi_{\hat{\gamma}} P\left[ \chi_{\hat{\gamma}} | \hat{\gamma}^{(k)}_{m}[w] \right] = \sum \chi_{\hat{\gamma}} \frac{P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_{\hat{\gamma}} \right]}{P\left[ \hat{\gamma}^{(k)}_{m}[w] \right]} \]  \tag{24}

where \( P\left[ \hat{\gamma}^{(k)}_{m}[w] \right] = \sum \chi_{\hat{\gamma}} \frac{P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_{\hat{\gamma}} \right]}{P\left[ \hat{\gamma}^{(k)}_{m}[w] \right]} P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_{\hat{\gamma}} \right] \). Assuming that the corresponding PDF \( P\left[ \hat{\gamma}^{(k)}_{m}[w] \right] \) is Gaussian distributed, the effective variance is given by
\[ \phi^{(k+1)}_{\gamma_{m}[w]} = \zeta + \frac{E_{\gamma}}{2} \left( \sum \chi_{\hat{\gamma}} - \hat{\gamma}^{(k)}_{m}[w] \right)^{2} \]  \tag{25}

where the predetermined parameter \( \zeta \) is introduced as an offset to compensate for quantization errors. Using \( \phi^{(k+1)}_{\gamma_{m}[w]} \), we can compute \( P\left[ \hat{\gamma}^{(k)}_{m}[w]; x_{\hat{\gamma}} \right] \) via Gaussian CDF, and then, the average variance of VNs can be computed as
\[ \phi^{(k+1)}_{\gamma_{m}[w]} = \sum_{w=1}^{2^{n}} P\left[ \hat{\gamma}^{(k)}_{m}[w]; \left( \frac{E_{\gamma}}{2} - \hat{\gamma}^{(k)}_{m}[w] \right)^{2} \right] \]  \tag{26}

### D. BELIEF COMBINING IN FNs (A1-5)
For the reconstruction of the RX symbol replica in (A1-5), the referencing original value for \( \hat{h}_{n,m}^{2} \) is \( h_{n,m}v_{m} \), i.e., a continuous value. Since channel coefficients are quantized with \( B \) bits, the number of possible combinations can be limited to \( 2^{2B} \); however, it remains a large number compared to that in the binary case. In addition, the referencing original value gradually varies and increases in successive summation operations of the \( M \) symbol replicas, and therefore, the typical DDE is not applicable.

To address this issue, we extend the DDE for the binary class [20] to that for multi-class using the quantized values based on its estimated PDF as referencing values. The discrete joint PMF required for determining the quantization threshold is computed based on the PDF of the corresponding continuous variable under SGA conditions.

First, we define \( \omega_{n,m} = h_{n,m}v_{m} \) and \( \hat{\omega}^{(k)}_{n,m} = h_{n,m}^{2} \). Using the quantization index \( v \in \{1, 2^{n}\} \), the representation value of \( \hat{\omega}^{(k)}_{n,m} \) is set to \( \hat{\omega}^{(k)}_{n,m}[w] = h_{n,m}[w]^{2} \) \( \hat{\omega}^{(k)}_{n,m}[v] \). Since \( \hat{h}_{n,m}[u] \) is distributed symmetric around zero and \( \hat{\omega}_{n,m}[v] \), the number of possible candidates of \( \hat{h}_{n,m}[u]^{2} \) is limited to \( 2^{B} \). Therefore, further quantization is not necessary; the 2-to-1 LUT for \( \omega_{n,m} \) does not contain a quantization function.

For subsequent summation, the multi-layer LUT structure is used. At the first layer, the variance of each input is obtained by \( \psi^{(k)}_{\omega_{n,m}[w]} = \phi_{\omega_{n,m}[w]} \).

Here, we focus on two 2-to-1 LUTs for the quantized sum at the \( l \)-th layer, where the outputs are defined by \( J^{(k,l)} \) and \( J^{(k,l)} \), and the corresponding input pairs are denoted by \( \left(J_{1}^{(k,l-1)}, J_{2}^{(k,l-1)} \right) \) and \( \left(J_{1}^{(k,l-1)}, J_{2}^{(k,l-1)} \right) \), respectively. The variance of the node at the \( l \)-th layer is denoted by \( \psi^{(k,l)} \). As the effective noise of two inputs is uncorrelated, the output variance is twice that of the input variance, \( \psi^{(k,l)} = 2\psi^{(k,l-1)} \). Further, the referencing value \( \omega_{n,m} \) is added up through the multi-layer structure; i.e., it has to be quantized with \( B \) bits as \( \hat{J}^{(k,l)}\left[ \hat{w} \right] \), where the maximal entropy criterion is used. Assuming that the corresponding continuous variable is Gaussian distributed, the joint PMF \( P\left[ \hat{J}^{(k,l)}\left[ \hat{w} \right], J^{(k)}\left[ u \right] \right] \) is computed as
\[ P\left[ \hat{J}^{(k,l)}\left[ \hat{w} \right], J^{(k)}\left[ u \right] \right] = \sum_{w=1}^{2^{n}} P\left[ \hat{J}^{(k,l)}\left[ \hat{w} \right]; J^{(k)}\left[ u \right] \right] P\left[ J^{(k)}\left[ u \right] \right] \]  \tag{27}

As iterative processing progresses, the symbol replicas approach hard-decision symbols and the probability mass of the first and last quantization level becomes dominant. If we use all levels, the algorithm does not work well because of the insufficient resolution of the discrete bin.
where the conditional probability follows

\[
P \left[ \hat{j}^{(k,l)}[\tilde{w}] | \hat{j}^{(k,l)}[u] \right] \\
= P \left[ \hat{j}^{(k,l)}[\tilde{w}] - 1 < \hat{j}_{n,m}^{(k,l)} \leq \hat{j}^{(k,l)}[\tilde{w}] | \hat{j}^{(k,l)}[u] \right] \\
= \frac{1}{2} \operatorname{erf} \left( \frac{\hat{j}^{(k,l)}[\tilde{w}] - \hat{j}^{(k,l)}[u]}{\sqrt{2} \psi^{(k,l)}} \right) \\
- \frac{1}{2} \operatorname{erf} \left( \frac{\hat{j}^{(k,l)}[\tilde{w}] - 1 - \hat{j}^{(k,l)}[u]}{\sqrt{2} \psi^{(k,l)}} \right). 
\]

(28)

Based on (28), \( \hat{j}^{(k,l)}[\tilde{w}] \) is quantized using the IB method. Finally, the joint PMF is computed after recursively performing the abovementioned processing based on the multi-layer structure.

**E. SOFT INTERFERENCE CANCELLATION (A1-6,7)**

The Onsager-coefficient \( \beta^{(k)} = \frac{M}{2} \phi_H \frac{\hat{\psi}^{(k)}}{\psi^{(k-1)}} \) in (A1-6) is independent both on the RX antenna and the UE indices, \( n \) and \( m \). This term is theoretically derived based on the Bayes-optimal denoiser (A1-9) in the continuous system. However, it is difficult to compute it in terms of the discrete activation function of (24) because of the severe quantization errors. Owing to the statistical stability of beliefs and the index-invariant characteristics, the Onsager coefficient is often given by appropriate parameters based on learning algorithms [34]. Thus, we may find a proper Onsager correction coefficient to imitate the average effect of self-feedback because of the loopy-propagation even in the quantized system.

For representing the accumulated quantization errors via iteration processing, we provide the Onsager coefficient as

\[
\gamma^{(k)} = \gamma^{(k-1)} + \frac{M}{2} \phi_H \frac{\hat{\psi}^{(k)}}{\psi^{(k-1)}},
\]

(29)

where \( \gamma^{(1)} = 1 \).

Finally, for closing the loop, the joint PMF required for VN processing as discussed in Section IV-B is computed. After the construction of the 2-to-1 LUT for the quantized sum of \( \gamma^{(k)} = \sum_{j=1}^{M} \omega^{(k)}_j \) and \( -\beta^{(k)} \gamma^{(k-1)} \), we define \( \Delta y^{(k)} = \gamma^{(k)} - \beta^{(k)} \gamma^{(k-1)} \). Assuming \( P \left( \Delta y^{(k)} | y_n \right) \sim \mathcal{N} \left( y_n, \psi^{(k)} \right) \) under large-system conditions, the corresponding conditional PMF is given by

\[
P \left[ \Delta y^{(k)}[\tilde{w}] | y[u] \right] \\
= \frac{1}{2} \left( \operatorname{erf} \left( \frac{\Delta y^{(k)}[\tilde{w}] - y[u]}{\sqrt{2} \psi^{(k)}_{\Delta y^{(k)}}} \right) \\
- \operatorname{erf} \left( \frac{\Delta y^{(k)}[\tilde{w}] - 1 - y[u]}{\sqrt{2} \psi^{(k)}_{\Delta y^{(k)}}} \right) \right). 
\]

(30)

After the quantization of \( \hat{j}^{(k)}[\tilde{w}] \) with \( B \) bits using the IB method, \( \tilde{y}[u] \) used in Section IV-C is obtained and the loop is closed. With its PMF, the DDE loop is successfully linked, and we complete the entire processing in IB LUT-based AMP detection. The flow chart of the DDE for the joint PMF is summarized in Fig. 5.

**V. COMPUTER SIMULATIONS**

Computer simulations were conducted to validate the performance of the proposed IB LUT-based AMP detector for massive MUD, where the performance metrics were averaged over 10000 independent channel realizations. The average RX power from each -UE is assumed to be identical based on slow TX power control. Simulation conditions are summarized in Table 1. We assume massive MIMO channels are uncorrelated flat Rayleigh fading, and channel estimation is conducted perfectly at the receiver side. The modulation...
TABLE 2. Predetermined parameter settings ($N_q$, $\zeta$).

| $N' = 80$ | $N' = 96$ |
|-----------|-----------|
| $B = 4$   | (6, 0.07) | (8, 0.09) |
| $B = 5$   | (4, 0.07) | (6, 0.09) |
| $B = 6$   | (4, 0.03) | (6, 0.05) |

scheme is gray-coded QPSK, and the channel coding is not utilized. The number of AMP iterations is $K = 16$. Time and frequency synchronization between TX and RX are assumed to be perfect.

We compare the performances of IB LUT based on the maximal MI criterion, EPQ LUT based on the maximal entropy criterion, and a typical AMP with double-precision. The PDF required for EPQ LUT is computed by marginalizing the histogram of each variable in Alg. 1 over all channel realizations, where the distribution is approximated by a scalar Gaussian distribution under large-system conditions.

Table 2 shows the predetermined parameters $N_q$ for EPQ LUTs and $\zeta$ for IB LUTs. Both parameters are set to achieve the BER performance for minimizing the error floor level in the present massive MIMO configurations. In computer simulations, $N_q$ is determined from 0 to 10 in steps of 0.5 in decimal, and $\zeta$ is determined from 0 to 1 in steps of 0.05.

A. BER PERFORMANCE

Fig. 6 shows the BER performances of different AMP detection schemes in massive MIMO configurations: $(N', M') = (80, 64), (96, 64)$. In [12]–[14], the IB-based LDPC decoders were shown to operate only 0.1 dB away from the double-precision decoding even though all messages were represented with 4 bits. However, when the number of quantization bits is 4 bits, the high-level error floor is inevitable both in the LUT-based detection schemes, as shown in Fig. 6. This is because, in MUD scenarios, the increase in the dynamic range of beliefs due to the random channel coefficients causes severe non-linear signal distortion due to insufficient resolution. The dynamic range becomes smaller because of channel hardening effects [3] as the compression rate $\rho = N/M$ increases, and therefore the error floor level in Fig. 6(a) is less than that in Fig. 6(b).

In the case of 5-bit quantization, we find clear improvements in detection performances due to the improved resolution. However, in both configurations, the EPQ LUT cannot sufficiently suppress the error floor level. This is because, in coarse quantization, the distribution of quantized beliefs deviates significantly from the referencing continuous distribution. In the EPQ LUT that mimics the original arithmetic operations with reference to the continuous distribution, performance degradation due to sever quantization errors is inevitable. In contrast, the IB LUT that maximizes the MI between the quantized beliefs and the corresponding continuous values is less susceptible to the quantization errors caused via each arithmetic operation. Consequently, the IB LUT can sufficiently suppress the error floor level and achieve $BER = 10^{-5}$ because of minimal information loss. The performance degradation at a BER $= 10^{-4}$ compared to the double-precision AMP is about 4.0 dB and about 3.0 dB in Figs. 6(b) and 6(a), respectively.

When the number of quantization bits reaches 6 bits, the resolution becomes high enough to directly mimic the arithmetic operations. Thus, the envelopes of the discretized distribution can be captured with the original continuous Gaussian distribution with high-accuracy. Both LUT-based discrete detection schemes suppress the error floor sufficiently. The performance difference between the EPQ LUT and IB LUT becomes smaller because the negative effect of quantization noise is mitigated. Remarkably, the IB LUT-based AMP can approach the double-precision AMP, and the performance gap is less than 2.0 dB at a BER $= 10^{-4}$.
Let us now consider the BER performances of the EPQ LUT- and IB LUT-based AMP when the number of iterations $K$ is changed. Fig. 7A shows the BER performances of EPQ LUT- and IB LUT-based AMP with different numbers of iterations, where $\xi$ follows Table 2. The transmit $E_s/N_0$ is fixed at 0 dB under the conditions of $(N', M') = (80, 64)$ and $(96, 64)$. In the case of 5-bit quantization, the IB LUT-based AMP can converge with a fewer number of iterations compared to the EPQ LUT-based AMP because of minimal information loss. In the EPQ LUT, the quantized AMP is subject to ill-convergence behavior of iterative detection due to the accumulated quantized errors, resulting in poor detection capability with a small number of iterations. In $(N', M') = (96, 64)$, the number of iterations required for convergence is less than eight iterations; the convergence point is considerably better than EPQ LUT. A similar tendency is observed in the case of 6-bit quantization, and the IB LUT-base AMP can converge to within less than six iterations in both configurations. Further improvements are expected by optimizing $\xi$ for $K$ and $E_s/N_0$.

A significant performance gap is attributed to the difference in the concept of the LUT design. The EPQ is performed to maximize the entropy of quantized values; however, high-accuracy signal detection requires original information; therefore, MI maximization is more appropriate. In addition, the distribution of the quantization error is assumed to be constant throughout all iterative processing in the EPQ LUT. In contrast, in the IB LUT, the PMF is tracked using the DDE. Consequently, the distortion of the probability distribution can be largely mitigated.

### B. MEMORY USAGE

Since CSI-independent LUTs are designed offline based on long-term channel statistics, we do not place our emphasis on the order of the computational complexity to construct the LUTs, instead, we focus on the memory size required for holding LUTs. Table 3 lists all LUTs for constructing the EPQ and IB LUT, where $g$ and $G$ denotes the memory usage for the 2-to-1 LUT and the multi-layer 2-to-1 LUT structure with $T$ terms, in each iteration, respectively; these are are formulated as

$$ g = B \cdot 2^{2B} \cdot K, \quad (32) $$

$$ G_T = B \cdot 2^{2B} \cdot K \cdot [\log_2 2^T]. \quad (33) $$

The averaged variance and Onsager coefficient can be computed directly using the PMF tracked by DDE in the IB LUT. Therefore, LUTs for this computation are no longer necessary. In this case, the mathematical formulation of memory usage required for the EPQ LUT and IB LUT are respectively given by

$$ V_{RF} = 6g + 2GM + GN $$

$$ V_{IB} = 5g + GM + GN $$

In Table 4, memory usage required for EPQ and IB LUT are presented. Indeed, it is extremely small memory usage. With limited memory usage in practical scenarios, the operation with LUTs can be substituted for arithmetic computations in AMP detection.

### TABLE 3. Memory usage of LUT for corresponding arithmetic operations.

| Operation | EPQ | IB |
|-----------|-----|----|
| $h_{m,n}y_n$ | $g$ | $g$ |
| $\sum_{n=1}^{M}h_{m,n}y_n$ | $G_M$ | $G_M$ |
| $y_n - \sum_{m=1}^{M}h_{m,n}y_n - z_n$ | $g$ | $g$ |
| $\beta^{(k)} = \frac{M}{T} \sum_{m=1}^{M} \beta^{(k-1)}$ | $G_M$ | $0$ |

### TABLE 4. LUT memory usage for $(N', M') = (80, 64)$.

| $B$ | EPQ | IB |
|-----|-----|----|
| 4   | 54 KB | 38 KB |
| 5   | 270 KB | 190 KB |
| 6   | 1296 KB | 912 KB |
VI. CONCLUSION

In this paper, a novel CSI-independent LUT-based AMP for massive MIMO detection was proposed to achieve signal detection in integer arithmetic with an LUT designed offline. Maximal entropy criterion and maximal MI criterion are utilized as strategies for quantization threshold design in LUTs; this leads to EPQ and IB quantization. The IB LUT is constructed using the IB method and DDE for minimizing information loss via quantization. It is possible to construct CSI-independent LUTs using the modified IB and DDE-based method by exploiting the statistical stability of beliefs in AMP. The predetermined parameter for suppressing the harmful effect of quantization errors can be designed in addition to the Onsager-term.

Via computer simulations, we confirmed that the proposed IB LUT achieves a BER performance similar to that of the AMP with double-precision. The penalty caused by the quantization error can be well-suppressed. Compared to the EPQ LUT, the LUT built according to IB and DDE manages to track the transition of discrete distribution more accurately and thus design an optimal quantization threshold.

REFERENCES

[1] L. Hanzo, O. Alamri, M. El-Hajjar, and N. Wu, Near-Capacity Multi-Functional MIMO Systems Sphere-Packing, Iterative Detection and Cooperation. Hoboken, NJ, USA: Wiley, May 2009.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Commun. Mag., vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] J. K. S. Lee and J. Thorpe, “Memory-efficient decoding of LDPC codes,” IEEE Trans. Commun., vol. 57, no. 2, pp. 1941–1988, 4th Quart., 2015.

[4] 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, Nov. 2009.

[5] 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.

[6] T. P. Minka, “Expectation propagation for approximate Bayesian inference,” Comput. Res. Repository (CoRR), vol. abs/1301.2294, 1-8, Jan. 2013. [Online]. Available: http://arxiv.org/abs/1301.2294

[7] J. Csiszár, M. Olmos, M. Schneitz-Fernandez, and F. Perez-Cruz, “Expectation propagation detection for high-order high-dimensional MIMO systems,” IEEE Trans. Commun., vol. 62, no. 8, pp. 2840–2849, Aug. 2014.

[8] K. Takeuchi, “Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements,” Jan. 2017, arXiv:1701.05284. [Online]. Available: http://arxiv.org/abs/1701.05284

[9] Y. Kabashima, “A CDMA multiuser detection algorithm on the basis of belief propagation,” J. Phys. A. Math. Gen., vol. 36, no. 43, pp. 11111–11121, Oct. 2003.

[10] T. Takahashi, S. Ibi, and S. Sampei, “Design of adaptively scaled belief in multi-dimensional signal detection for higher-order modulation,” IEEE Trans. Commun., vol. 67, no. 3, pp. 1986–2001, Mar. 2019.

[11] J. Lewandowsky and G. Bauch, “Information-optimum LDPC decoders based on the information bottleneck method,” IEEE Access, vol. 6, pp. 4054–4071, 2018.

[12] F. J. C. Romero and B. M. Kurkoski, “LDPC decoding mappings that maximize mutual information,” IEEE J. Sel. Areas Commun., vol. 34, no. 9, pp. 2391–2401, Sep. 2016.

[13] M. Meidlinger, G. Matz, and A. Burg, “Design and decoding of irregular LDPC codes based on discrete message passing,” IEEE Trans. Commun., vol. 68, no. 3, pp. 1329–1343, Mar. 2020.
TAKUMI TAKAHASHI (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 2016, 2017, and 2019, respectively. From 2018 to 2019, he was a Visiting Researcher with the Center for Wireless Communications, University of Oulu, Finland. In 2019, he joined the Graduate School of Engineering, Osaka University, as an Assistant Professor. His research interests include belief propagation, compressed sensing, signal processing, and wireless communications.

SHINSUKE IBI (Member, IEEE) received the B.E. degree in advanced engineering from the Suzuka College of Technology, Japan, in 2002, and the M.E. and Ph.D. degrees in communication engineering from Osaka University, Japan, in 2004 and 2006, respectively. From 2005 to 2006, he was a Visiting Researcher with the Center for Wireless Communications, University of Oulu, Finland. In 2006, he joined the Graduate School of Engineering, Osaka University. From 2010 to 2011, he was a Visiting Researcher with the University of Southampton, U.K. In 2019, he moved to the Faculty of Science and Engineering, Doshisha University, as an Associate Professor. His research interests include EXIT-based coding theory, iterative detection, digital signal processing, cognitive radio, and communication theory. He received the 64th and 71st Best Paper Awards from IEICE and the 24th Telecom System Technology Award from the Telecommunication Advancement Foundation.

SEICHI SAMPEI (Fellow, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from the Tokyo Institute of Technology, Japan, in 1980, 1982, and 1991, respectively. From 1982 to 1993, he was with the Communications Research Laboratory, Ministry of Posts and Telecommunications. From 1991 to 1992, he was a Visiting Researcher with the University of California, Davis. In 1993, he joined the Faculty of Engineering, Osaka University. He is currently a Professor with the Department of Information and Communications Technology, Osaka University. He has been involved in the development of adaptive modulation, intelligent radio transmission/access, cognitive wireless networking, and wireless distributed network techniques. He is also a member of the Institute of Image Information and Television Engineers (ITE) and a Fellow of the Institute of Electronics, Information and Communication Engineers (IEICE). He received the Shinohara Young Engineering Award, Achievements Award, Communications Society Best Paper Award, Best Paper Award from IEICE, Telecom System Technology Award from the Telecommunication Advancement Foundation, DoCoMo Mobile Science Award from the Mobile Communication Fund, Ericsson Telecommunications Award, and Shida Rinzaburo Award from the President of Council for Info-Communications Promotion Month.