# SCMA Detection in MIMO Systems with Low Complexity EP using QRD

Adam Mekhiche, Antonio Maria Cipriano, Charly Poulliat

# SCMA Detection in MIMO Systems with Low Complexity EP using QRD

Adam Mekhiche
*Thales - IRIT / INP Toulouse*
Gennevilliers, France
adam.mekhiche@thalesgroup.com

Antonio Maria Cipriano
*Thales*
Gennevilliers, France
antonio.cipriano@thalesgroup.com

Charly Poulliat
*IRIT / INP Toulouse*
Toulouse, France
charly.poulliat@toulouse-inp.fr

***Abstract*—This article presents new low complexity message passing based detection algorithms for Sparse Code Multiple Access (SCMA) in Multiple Input Multiple Output (MIMO) systems thanks to the combination of sparser Factor Graph (FG) representation via QR Decomposition (QRD) and enhanced message-passing scheduling. The proposed algorithm achieves better performance to complexity trade-off than state-of-the-art message-passing-based detectors. Our results are presented based on complexity and error rate performance analysis.**

***Index Terms*—Expectation Propagation, Sparse Code Multiple Access, Multiple Input Multiple Output, QR, Scheduling**

## I. Introduction

Non-Orthogonal Multiple Access (NOMA) has been studied for cellular networks like 5G New Radio (NR) [1] to improve the system spectral efficiency, in combination with already existing techniques like MIMO. One technique to achieve NOMA called SCMA [2] uses a sparse multidimensional codebook to spread users across several orthogonal Resource Elements (RE) to overload the system, i.e. having an average of more than one user per RE. Furthermore, to improve performance, SCMA can be combined with a multiple antenna receiver at the cost of a less sparse system. The detection of signals in such systems can be exponentially complex, e.g. detectors with an exact Maximum A Posteriori (MAP) criterion, but the computational complexity can be drastically reduced when the codebook sparsity is taken into account by the detection algorithm. Message Passing Algorithms (MPA) can use this sparsity to achieve, at the same time, near-optimal performance and low complexity detection. MPAs are iterative algorithms designed to exchange messages along the edges of an FG, that represents a factorized version of the target A Posteriori Probability (APP), in order to compute an approximate MAP criterion. One MPA used for SCMA is the Belief Propagation (BP) [2] which achieves near-optimal performance but whose complexity increases rapidly with the connectivity of the graph because it has to enumerate all the possible MIMO symbols. Approximate Message Passing (AMP) [3] is another effective MPA that can achieve great performance, but typically with a low convergence rate. Expectation Propagation (EP) [4] is a very good candidate for SCMA since it can achieve great performance while maintaining a low complexity [5], especially when the associated FG is thoughtfully re-factorized through matrix decomposition, as shown in this article. Previous work [6] studied the pre-processing of the received signal in order to accomplish greater sparsity, when using a multiple antenna receiver, through matrix decomposition. Their method uses QR Decomposition (QRD) to reduce the number of edges in the FG and reorder it to apply a more efficient BP, but BP remains quite computationally expensive. Other works [3], [5] proposed to use low complexity MPA to improve the performance-to-complexity trade off. Low complexity EP applied on modified FG has also been studied in MIMO context [7], [8]. The proposed algorithms in this article achieve a more effective scheduling of the EP message exchange applied to less connected scalar factor graphs thanks to a QRD pre-processing suited for SCMA. These algorithms have improved convergence speed and performance while lowering their overall complexity, compared to other MPA based SCMA detector [3], [5], [6]. The article is organized as follows : Section II introduces the system model used in the paper, then Section III presents the proposed combination of pre-processing and scheduling for EP messages. Section IV and Section V compare the proposed algorithms with the aforementioned MPA, from a computational complexity and a performance point-of-view, respectively.

This work has been partly funded by the French National Research Agency project EVASION, grant reference ANR-20-CE25-0008-01.

## II. System Model

Consider a MU-MIMO uplink transmission of $N_u$ single antenna users toward a base station using $N_r$ receive antennas ($N_u \times N_r$). Each user $i \in [\![1, N_u]\!]$ encodes a stream of bits $\mathbf{b}_i \in \mathbb{F}_2^{K_b}$, using an Error Correcting Code (ECC) of length $N$ and rate $R$, into encoded bits $\mathbf{c}_i \in \mathbb{F}_2^N$. We denote $\mathbb{F}_2$ the finite field of size 2. The encoded bits are randomly interleaved and then mapped to constellation symbols according to its user-specific SCMA codebook. The codebook $\mathbf{B}$ is made of $|\mathbf{B}| = M$ multidimensional symbols, spread across $R_e < N_u$ orthogonal Resource Elements (REs), e.g. different spectral resources of an Orthogonal Frequency Division Multiplexing (OFDM) modulation. A mapping matrix $\mathbf{F} \in [0,1]^{R_e \times N_u}$ is used to represent the codebook's resource allocation of each user :

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{1}$$

We denote $d_v$ the number of users per RE, i.e. the number of ones per line, and $d_f$ the number of RE used per user, i.e. the number of ones per column. At time $n$ and on the RE $k \in [\![1, R_e]\!]$, the vector $\mathbf{s}_{n,k} = [s_1^{n,k}, \dots, s_{N_u}^{n,k}]^T \in \mathbf{B}$ of symbols is sent through a uncorrelated Rayleigh channel $\mathbf{H}_{n,k} \in \mathbb{C}^{N_r \times N_u}$ and the signal $\mathbf{y}_{n,k} \in \mathbb{C}^{N_r}$ is received by the base station using $N_r$ antennas. The corresponding linear model is:

$$\mathbf{y}_{n,k} = \mathbf{H}_{n,k}\mathbf{s}_{n,k} + \mathbf{w}_{n,k} \tag{2}$$

with $\mathbf{w}_{n,k} \in \mathbb{C}^{N_r}$ a vector of Additive White Gaussian Noise (AWGN) samples with the properties $\mathbb{E}(\mathbf{w}_{n,k}) = 0$ and $\mathbb{E}(\mathbf{w}_{n,k}\mathbf{w}_{n,k}^H) = N_0\mathbf{I}_{N_r}$. For the sake of readability, the time index will be omitted except when needed for comprehension.

There are numerous ways to construct an SCMA codebook, with different advantages and drawbacks, but the general rule to build a codebook is given in [9]. In the article, we use constellation rotation codebooks developed in [10]. The advantages of such codebooks are the ability to use the same base constellation for all users, e.g. QPSK, and to account for the rotation of each symbol, for each user and for each resource, directly in an equivalent channel matrix. Let's consider a base constellation $\mathcal{X}$ of size $|\mathcal{X}| = M$, e.g. QPSK, used to map $\mathbf{c} = [\mathbf{c}_1, \dots, \mathbf{c}_{N_u}]^T \in \mathbb{F}_2^{N_u}$ to $\mathbf{x} = [x_1, \dots, x_{N_u}]^T \in \mathcal{X}^{N_u}$. Using a codebook with the mapping matrix (1), we can write

$$\bar{\mathbf{y}} = (\bar{\mathbf{H}} \circ \bar{\mathbf{G}})\mathbf{x} + \bar{\mathbf{w}}. \tag{3}$$

with $\circ$ the Hadamar product. Taking $R_e = 4$ as an example, we denote

$$\bar{\mathbf{G}} = \mathbf{G} \otimes \mathbb{1}_{N_r} \text{with } \mathbf{G} = \begin{bmatrix} \varphi_0 & \varphi_1 & \varphi_2 & 0 & 0 & 0 \\ \varphi_1 & 0 & 0 & \varphi_2 & \varphi_0 & 0 \\ 0 & \varphi_2 & 0 & \varphi_0 & 0 & \varphi_1 \\ 0 & 0 & \varphi_0 & 0 & \varphi_1 & \varphi_2 \end{bmatrix} \in \mathbb{C}^{R_e \times N_t} \tag{4}$$

with $\otimes$ the Kronecker product and $\mathbb{1}_{N_r}$ the vector of ones of size $N_r$ such as $\bar{\mathbf{G}} \in \mathbb{C}^{N_r R_e \times N_t}$. $\bar{\mathbf{H}} \in \mathbb{C}^{N_r R_e \times N_u}$, the rearranged channel matrix, groups receiving antennas and concatenates REs on its lines, with its $i^{\text{th}}$ column:

$\bar{\mathbf{h}}_i = [(h_1)_{1,i} \ \dots \ (h_1)_{N_r,i} \ \cdots \ (h_{R_e})_{1,i} \ \dots \ (h_{R_e})_{N_r,i}]^T$, $\forall a \in [\![0, 2]\!], \varphi_a = e^{ia\Delta}$, with $\Delta$ the difference of phase between to adjacent phase.

The received signal is, similarly to $\bar{\mathbf{H}}$, rearranged as :

$$\bar{\mathbf{y}} = [(y_1)_1 \ \dots \ (y_1)_{N_r} \ \cdots \ (y_{R_e})_1 \ \dots \ (y_{R_e})_{N_r}]^T$$

and the same for the noise. Equation (2) becomes :

$$\bar{\mathbf{y}} = \bar{\mathbf{H}}^{\text{eq}}\mathbf{x} + \bar{\mathbf{w}} \tag{5}$$

with $\bar{\mathbf{H}}^{\text{eq}} = \bar{\mathbf{H}} \circ \bar{\mathbf{G}}$. This new equivalent channel matrix encompasses the contribution of the real channel, the imposed sparsity of the codebook and the constellation rotation of each user. The turbo-iterated receiver computes the a posteriori probability of $\mathbf{x}$ :

$$\mathbb{P}(\mathbf{x}|\bar{\mathbf{y}}, \bar{\mathbf{H}}^{\text{eq}}) \propto \prod_{j=1}^{N_r R_e} \underbrace{\mathbb{P}(\bar{y}_j|\mathbf{x}, \bar{\mathbf{H}}^{\text{eq}})}_{f_j^{\text{EQU}}} \prod_{i=1}^{N_u} \underbrace{\mathbb{P}(x_i|\mathbf{c}_i)}_{f_i^{\text{DEM}}} \underbrace{\mathbb{P}(\mathbf{c}_i)}_{f_i^{\text{DEC}}} \tag{6}$$

where $f_j^{\text{EQU}}$ defines the likelihood function, $f_i^{\text{DEM}}$ the demapping function and $f_i^{\text{DEC}}$ the decoding function which provides an a priori probability of the coded binary vector $\mathbf{c}_i$ thanks to an Error Correcting Code (ECC) decoder. From this a posteriori probability, a Factor Graph (FG) can be drawn to represent the link between each variable (variable nodes) connected through functions (function nodes). The FG is the support on which MPAs are applied to compute an approximation of (6) and this article focuses on EP. MPAs exchange messages along the edges of an FG, and the more edges there are, the more complex the detection becomes, e.g. for BP the computational complexity increase is exponential with $d_v$ the number of users connected to a single RE and $R_e$ the number of RE, while with EP has only a polynomial complexity (square or cube). The next section proposes to reduce the number of exchanged messages through matrix decomposition, as a pre-processing of the received signal on the RE level, and thoughtfully reorder the scheduling of EP messages to enhance the detection performance.

## III. Proposed low complexity SCMA receiver

There is a natural synergy between SCMA and MPA based detectors which rely on the FG sparsity to give a more accurate estimation of $P(\mathbf{x}|\mathbf{y}, \mathbf{H}, N_0)$, e.g. BP returns an exact MAP estimation when applied on a tree FG. Less connected graphs decrease the detection complexity as fewer messages need to be computed. There are several exact or approximate processing to reduce the number of edges in a graph. One way to compute an approximate and less connected FG is to neglect the low-weight edges but at the cost of performance loss [11]. Another exact computation of the equivalent FG can be done through matrix decomposition, such as QR decomposition. This method [6] decomposes each RE channel matrix $\mathbf{H}_k, \forall k \in [\![1, R_e]\!]$ into two matrices, one unitary matrix $\mathbf{Q}_k \in \mathbb{C}^{N_r \times N_r}$ and one upper triangle matrix $\mathbf{R}_k \in \mathbb{C}^{N_r \times N_u}$. Each RE signal is received signal is pre-processed :

$$\mathbf{Q}_k^H\mathbf{y}_k = \mathbf{R}_k\mathbf{s}_k + \mathbf{Q}_k^H\mathbf{w}_k \tag{7}$$

and the proposed algorithms apply the same process but on the equivalent channel matrix that contains the channel perturbation and the codebook constellation rotation. Applying QRD on the RE level ensures that the resulting graph is sparser than before, which would not be the case if the QRD was applied to the global equivalent channel matrix of (6). Since each RE local factor graph is fully connected, there are $N_r$ equalization function nodes ($f^{\text{EQU}}$) and $d_v$ user variable nodes ($x$) so $N_r d_v$ edges per local graph, the QRD processing removes almost half the edges, $d_v(d_v + 1)/2$ edges remain. The rearranged received signal processed by the QRD on the RE level gives the FG of Fig. 1. The proposed algorithms exchange Expectation Propagation [4] messages on this new FG according to a specific scheduling to enhance performance.

EP is an approximate Bayesian inference algorithm that selects a distribution $q \in \mathcal{Q}$ closest to the targeted distribution $p$. The set $\mathcal{Q}$ of distributions in which $q$ lies is the set of complex Gaussian distributions of the exponential family. In order to select the best $q$, EP uses the inclusive Kullback-Leibler divergence which is a divergence measure of the

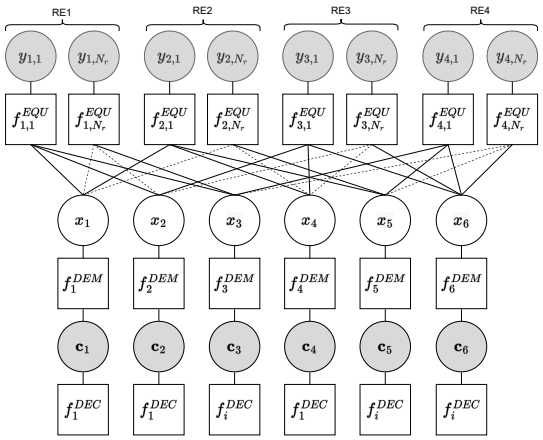


Fig. 1. MIMO SCMA ($N_u = 6$, $R_e = 4$, $d_v = 3$, $d_f = 2$) factor graph representation. Dotted edges are removed through QRD of each RE channel matrix $\mathbf{H}_k$. Gray variables are assumed known or partially known from the detection algorithm perspective, and only the white variables are estimated.

statistical difference between two distributions:

$$D_{KL}(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{8}$$

and the selection of the closest distribution is done as:

$$q = \text{proj}_{\mathcal{Q}}[p] = \arg\min_{\tilde{q}\in\mathcal{Q}} D_{KL}(p||\tilde{q}). \tag{9}$$

EP messages are extrinsic distributions exchanged between a function node and a variable node of an FG. The general expression of an EP message between a function node $f_j$ and a variable node $x_i$ is:

$$m_{f_j\to x_i}(x_i) = \frac{1}{m_{x_i\to f_j}(x_i)} \times \text{proj}_{\mathcal{Q}}\left[ m_{x_i\to f_j}(x_i) \int_{\mathbf{x}} f_j(\mathbf{x}) \prod_{\substack{i'\in\mathcal{N}(f_j)\\ i'\neq i}} m_{x_{i'}\to f_j}(x_{i'}) \right]. \tag{10}$$

The first EP-based receiver that can be derived from this new FG representation is a local QRD Scalar EP (SEP QRD), which exchanges EP messages along the edges of the sparser graph of Fig. 1, similarly to [6], which uses BP messages. This algorithm is less complex than classic SEP [5], [12] (see Sec. IV) but it has better performance and converges more rapidly (see Sec. V). The first proposed receiver combines the use of QRD to remove edges in the FG and the application of the EP algorithm with a modified scheduling of messages to better suit the specific SCMA FG and is called Scheduled Scalar EP with QRD (SSEP QRD). Indeed, the local FG of each RE can be browsed more efficiently by exchanging the EP messages from the last, and most connected user and moving to upper users, which are less connected, one by one instead of the traditional flooding scheduling used for detection. The second algorithm adds a "per-user" schedule of the decoding process to improve soft estimation of symbols and to achieve Successive Interference Cancellation (SIC) and is called SSEP QRD SIC. A study on enhanced EP messages scheduling has been proposed for massive Single User (SU)-MIMO and Multiple User (MU)-MIMO detection [8].

Both algorithms share the same detection scheduling : Each user is treated separately, from the last one ($x_{Nu}$) to the first ($x_1$). The first messages exchanged are the messages from every demapping node to every variable node. During the update of the user $i$, the variable node $x_i$ sends its messages to all the connected equalization function nodes, so there are on average $d_f N_r/2$ messages sent per variable node, some nodes are more connected (e.g. $x_6$ on Fig. 1) and some are less connected (e.g. $x_1$ on Fig. 1). We denote $\mathcal{N}_{f^{\text{EQU}}}(x_i)$ the set of indexes of the equalization function nodes connected to the variable node $x_i$, and conversely with $\mathcal{N}_x(f_j^{\text{EQU}})$. These messages are referred to as $m_{x_i\to f_j^{\text{EQU}}}(x_i)$, $\forall i \in [\![1, N_u]\!], j \in \mathcal{N}_{f^{\text{EQU}}}(x_i)$. Then, all the equalization nodes send their messages $m_{f_j^{\text{EQU}}\to x_i}(x_i)$ back to the variable node $x_i$ and this node propagates it up to the demapping node $f_i^{\text{DEM}}$. Every user is detected according to this scheduling, sequentially from the most connected to the least, and we denote as an "auto iteration" the process of exchanging messages up to the demapping nodes for all users. The first instance of detection is 0 auto iteration. At the last auto iteration, the demapping node computes the Log Likelihood Ratio (LLR) $\lambda_i^e$ which can be sent to a decoder that computes $\hat{\mathbf{b}}_i$ and $P(\mathbf{c}_i)$ as $\lambda_i^a$, used as an a priori probability in the next detection instance. The constellation mapping of the symbol is represented through the function $\varphi : \mathbb{F}_2^{\log_2(M)} \to \mathbb{C}$ and the inverse function $\forall k \in \log_2(M), \varphi_k^{-1} : \mathbb{C} \to \mathbb{F}_2^{\log_2(M)}$ that returns the $k^{\text{th}}$ bits of a base constellation symbol $x$. This step is done once every user has been detected in SSEP QRD. The computation of every auto iteration and a channel decoding step for each user is called a turbo iteration. Compared to SEP [5], [12], this algorithm can be self-iterated (i.e. auto iteration) between turbo iterations instead of just a single instance of detection, and it also saves its EP messages as initialization messages for the next turbo iteration. SSEP QRD scheduling and message computation are detailed in Alg. 1.

Similar scheduling, which provides better SIC, is proposed by decoding each user at the end of its iteration and by propagating the updated messages to the equalization function nodes. Such scheduling has also been proposed in massive MIMO (SSEP QRD SIC) [8] and has shown to outperform classic VEP [13] both in terms of performance and complexity.

The advantages of the proposed algorithms are that they use a QRD of the equivalent channel matrix of each RE to ensure that the resulting FG is sparser than the classic SCMA FG with multiple receive antennas. In addition, QRD needs to be applied once per RE and only when the channel changes, e.g. in a block fading scenario only once per block; not at every auto or turbo iteration which makes this pre-processing computationally thrifty as shown in Sec. IV. Then, unlike VEP [3], [13], the proposed EP can benefit from this novel sparsity to achieve an even lower complexity detection. Finally, the enhanced scheduling increases the convergence speed and improves the performance, see Sec. V.

## IV. Complexity Analysis

The complexity of EP algorithms depends on several factors. On one hand, the number of messages exchanged depends

**Algorithm 1** Scheduled Scalar EP on MIMO-QRD-SCMA FG

**Input:** $\mathbf{y}, \mathbf{H}, N_0$, $T$ = Turbo iteration, $L$ = Auto iteration
**Output:** $\hat{\mathbf{b}}$ - the estimated bits.

1: **for** $k = 1 : R_e$ **do**
2: $\quad \mathbf{H}_k^{\text{eq}} = \mathbf{H}_k \circ \mathbf{G} = \mathbf{Q}_k \mathbf{R}_k, \quad \mathbf{y}_k \leftarrow \mathbf{Q}_k^H \mathbf{y}_k, \quad \mathbf{H}_k^{\text{eq}} \leftarrow \mathbf{R}_k$
3: **end for**
4: $\forall i, j, m \in [\![1, N_t]\!], [\![1, N_r]\!], [\![1, \log_2(M)]\!], \lambda_{i,k}^a = 0$
5: $m_{f_j^{\text{EQU}} \to x_i}(x_i) = m_{x_i \to f_i^{\text{DEM}}}(x_i) = \mathbb{CN}(0, +\infty)$
6: **for** $t = 0 : \text{T}$ **do**
7: $\quad$ **for** $i = 1 : N_t$ **do**
8: $\quad\quad$ **Compute** $m_{f_i^{\text{DEM}} \to x_i}(x_i)$**:**
9: $\quad\quad \tilde{q}(x_i) \propto \exp\left( -\frac{|x_i - \overrightarrow{\mu_i^d}|^2}{\overrightarrow{\nu_i^d}} - \sum_{n=1}^{N} \varphi_n^{-1}(x_i) \lambda_{i,n}^a \right)$
10: $\quad\quad q(x_i) \sim \mathbb{CN}\left( \mu_i^d = \mathbb{E}\left[\tilde{q}(x_i)\right], \nu_i^d = \text{Var}\left[\tilde{q}(x_i)\right] \right)$
11: $\quad\quad \overleftarrow{\mu_i^d} = \overleftarrow{\nu_i^d} \left( \frac{\mu_i^d}{\nu_i^d} - \frac{\overrightarrow{\mu_i^d}}{\overrightarrow{\nu_i^d}} \right)$ and $\overleftarrow{\nu_i^d} = \left( \frac{1}{\nu_i^d} - \frac{1}{\overrightarrow{\nu_i^d}} \right)^{-1}$
12: $\quad$ **end for**
13: $\quad$ **for** $i = 1 : N_t$ **and then for** $j = 1 : N_r R_e$ **do**
14: $\quad\quad$ **Compute** $m_{x_i \to f_j^{\text{EQU}}}(x_i)$**:**
15: $\quad\quad \overleftarrow{\nu_{i,j}^e} = \left( (\nu_i^d)^{-1} - (\overrightarrow{\nu_{i,j}^e})^{-1} \right)^{-1}$
16: $\quad\quad \overleftarrow{\mu_{i,j}^e} = \overleftarrow{\nu_{i,j}^e} \left( \mu_i^d / \nu_i^d - \overrightarrow{\mu_{i,j}^e} / \overrightarrow{\nu_{i,j}^e} \right)$
17: $\quad$ **end for**
18: $\quad$ **for** $l = 0 : L$ **and then for** $i = N_t : -1 : 1$ **do**
19: $\quad\quad$ **for** $j = 1 : N_r R_e$ **do**
20: $\quad\quad\quad$ **Compute** $m_{f_j^{\text{EQU}} \to x_i}(x_i)$**:**
21: $\quad\quad\quad \overrightarrow{\nu_{i,j}^e} = \left( N_0 + \sum_{i' \neq i} |h_{j,i'}|^2 \overleftarrow{\nu_{i',j}^e} \right) / |h_{j,i}|^2$
22: $\quad\quad\quad \overrightarrow{\mu_{i,j}^e} = \left( y_j - \sum_{i' \neq i} h_{j,i'} \overleftarrow{\mu_{i',j}^e} \right) / h_{j,i}$
23: $\quad\quad$ **end for**
24: $\quad\quad$ **Compute** $m_{x_i \to f_i^{\text{DEM}}}(x_i)$**:**
25: $\quad\quad \overrightarrow{\nu_i^d} = \left( \sum_{j' \in \mathcal{N}(x_i)} \frac{1}{\overrightarrow{\nu_{i,j'}^e}} \right)^{-1}, \overrightarrow{\mu_i^d} = \overrightarrow{\nu_i^d} \sum_{j' \in \mathcal{N}(x_i)} \frac{\overrightarrow{\mu_{i,j'}^e}}{\overrightarrow{\nu_{i,j'}^e}}$
26: $\quad\quad$ **if** $l \neq L$ **then**
27: $\quad\quad\quad$ **Compute** $m_{f_i^{\text{DEM}} \to x_i}(x_i)$**:** Alg. 1 line. 8
28: $\quad\quad$ **end if**
29: $\quad\quad$ **for** $j = 1 : N_r R_e$ **do**
30: $\quad\quad\quad$ **Compute** $m_{x_i \to f_j^{\text{EQU}}}(x_i)$**:** Alg. 1 line. 14
31: $\quad\quad$ **end for**
32: $\quad$ **end for**
33: $\quad$ **for** $i = 1 : N_t$ **and then for** $k = 1 : \log_2(M)$ **do**
34: $\quad\quad \lambda_{i,k}^e = \log\left( \frac{\sum_{x_i \in \mathcal{X} : \varphi_k^{-1}(x_i) = 1} \tilde{q}(x_i)}{\sum_{x_i \in \mathcal{X} : \varphi_k^{-1}(x_i) = 0} \tilde{q}(x_i)} \right) - \lambda_{i,k}^a$
35: $\quad$ **end for**
36: $\quad$ Send $\lambda^e$ to decoder and receive $\lambda^a$ and $\hat{\mathbf{b}}$
37: **end for**

on the factor graph representation of the model, i.e. a scalar or vector FG. The use of matrix decomposition to reduce the number of edges, e.g. QRD of the channel matrix, also has an effect on the complexity as such operation can be computationally demanding. On the other hand, EP can be iterated through auto and/or turbo iterations which increases the detection complexity. This section is a study of the detection complexity only, with its pre-processing, and not taking into account the decoding complexity, that can be significant especially when bit-interleaved coded-modulation with iterative decoding (BICM-ID) is used, as it is the same for all the compared algorithms.

If the FG representation used is a vector one [3], [13], i.e. the probability is not fully factorized from the $\mathbf{y}$ standpoint, the detection can be very complex. Either every RE and every antenna are represented using only a single function node $f^{\text{EQU}}$, which leads to the most computationally complex EP algorithm that requires the inversion of a square matrix of size $\min(N_u, N_r R_e)$ per auto iteration which dominates the overall complexity of the detector. The inversion cost is about $\mathcal{O}((L+1) \times \min(N_u, N_r R_e)^3)$. It is possible to factorize only the receiving antennas and having only $R_e$ $f^{\text{EQU}}$ equalization nodes which results in $R_e$ inversion of a $\min(N_u, N_r)$ matrix.

TABLE I
BIG $\mathcal{O}$ COMPLEXITY OF EP ALGORITHMS PER TURBO-ITERATION

| | |
|---|---|
| VEP | $\mathcal{O}((L+1) \times \min(N_u, N_r R_e)^3)$ |
| VEP Local | $\mathcal{O}(R_e \times (L+1) \times \min(N_u, N_r)^3)$ |
| SEP | $\mathcal{O}((L+1) \times N_u \times N_r \times d_f)$ |
| QRD SEP | $\mathcal{O}((L+1) \times d_v(d_v+1)/2 \times R_e) + (d_v^3/3)$ |
| QRD SSEP | $\mathcal{O}((L+1) \times d_v(d_v+1)/2 \times R_e) + (d_v^3/3)$ |

Consider a scalar FG representation [3], [5], [12], without any matrix decomposition. The number of Scalar EP (SEP) [5] messages is greater than with a vector FG, where there are $N_u \times N_r \times R_e$ VEP messages along the edges between the equalization nodes and the variable nodes instead of $N_u$. The SEP complexity is smaller than VEP, especially when the system grows bigger with antennas and users. SEP does not use any auto iteration but instead keeps the memory of the EP messages between turbo iterations. In order to achieve a less expensive EP detection, the removal of edges through QRD is a great candidate when applied carefully. Decomposing the local FGs of each RE, which are fully connected graphs (unlike the overall FG), removes at most half of the edges between the equalization nodes and the variable nodes. Indeed, there are only $d_v(d_v+1)/2$ edges left from $d_v N_r$ initially present before QRD on each local graph. The complexity is lowered and the QRD costs about $\mathcal{O}(d_v^3/3)$ which still preserve the overall complexity gain. Another MPA like BP could be applied on such QRD processed FG [6] but each message is more complex since it depends exponentially on the connectivity of each RE, i.e $\mathcal{O}(R_e \times N_r \times M^{d_v})$.

## V. PERFORMANCE RESULTS

The system simulated has $N_u = 6$ users using the constellation rotation codebook from [10] with $R_e = 4$, $d_v = 3$ and $d_f = 2$ with a QPSK; the codebook uses $\Delta = \pi/6$ and the receiver uses $N_r = 3$ antennas. The mono antenna users use an LDPC encoder and the receiver uses BP iterative decoding. The LDPC decoder makes five inner iterations per turbo iteration, and the inner messages of the decoder are kept and reused as initialization messages for the next turbo iteration. A maximum of 9 turbo iterations (i.e. ten decoder calls) brings the overall LDPC iterations to 50. All the EP algorithms applied on a scalar FG use only one detection step ($L = 0$) and they keep their EP messages throughout the turbo iterations except VEP which uses one more auto iteration ($L = 1$). SEP is extracted from [5] and adapted for MIMO, SEP QRD is the same algorithm but applied on a QRD pre-processed FG and VEP [13] is adapted to the SCMA FG.

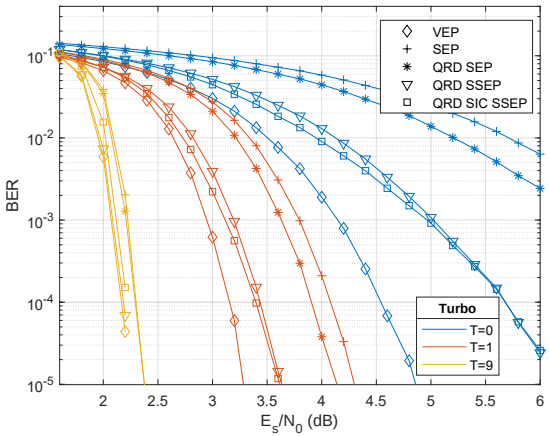


Fig. 2. Performance with an LDPC $N = 4096$ and $R = 1/2$.

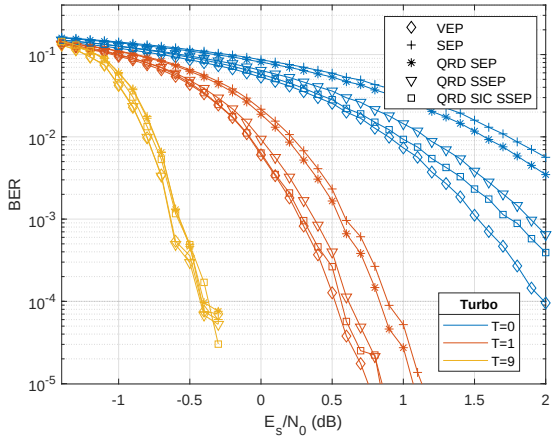


Fig. 3. Performance with an LDPC $N = 3120$ and $R = 1/3$.

Fig. 2 shows the first, second and ninth turbo iterations of five MPA-based detectors using an LDPC $N = 4096$ and $R = 1/2$. The first turbo shows that the most complex algorithm, VEP, achieves the best performance for the first turbo iterations but the two proposed algorithms have the same performance as VEP at $T = 9$ while being far less complex. SEP and SEP QRD are less efficient than the proposed algorithms in all the iterations, and the situation is even worse for SEP as it is also more complex. SSEP QRD SIC converges more rapidly than SSEP QRD, at each turbo the SIC version is slightly better in this scenario using an LDPC of $R = 1/2$.

Fig. 3 shows the same turbo iterations using an LDPC $N = 3120$ and $R = 1/3$. In this scenario using a stronger ECC, VEP remains the best detector but SSEP QRD SIC achieves the same performance from $T = 1$ while SSEP QRD is slightly worse. SEP and SEP QRD still perform worse than the proposed algorithms but achieve the same performance at the last turbo iteration.

Finally, Fig. 4 shows the same turbo iterations using an LDPC $N = 1944$ and $R = 2/3$. This weaker ECC shows that SSEP QRD SIC performs essentially the same as SSEP QRD if the correcting power of the ECC is not sufficient. The proposed algorithms still outperform SEP and SEP QRD but converge slower than VEP as SSEP QRD achieve the same performance than VEP at $T = 9$. SEP and SEP QRD are still the most affected detectors as their performance has worsened the most, even at $T = 9$.

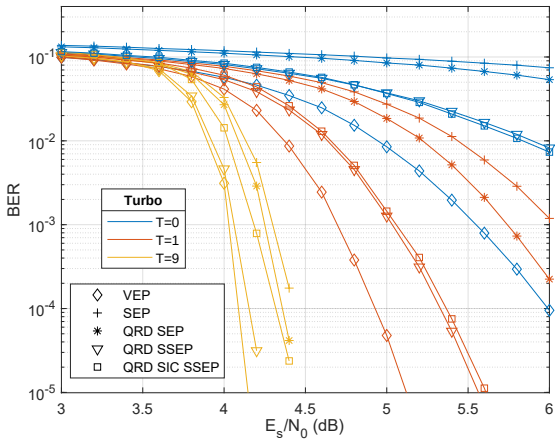


Fig. 4. Performance with an LDPC $N = 1944$ and $R = 2/3$.

## VI. Conclusion

This article is the study of two new algorithms, SSEP QRD and SSEP QRD SIC, that can achieve a better performance-to-complexity trade-off than other EP-based detectors, as they achieve almost VEP performance while being less computationally complex. They rely on scalar FG and QRD to achieve low complexity detection and use enhanced scheduling to improve performance. They are great candidates for efficient MIMO SCMA detection, and they might also be used in other kinds of NOMA schemes like Interleave Division Multiple Access (IDMA) [14] which is closely related to SCMA.

## References


[1] 3GPP, "Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16)," TR 38.812, Dec. 2018.

[2] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE PIMRC*, Sep. 2013, pp. 332–336.

[3] X. Meng *et al.*, "Advanced NOMA Receivers from a Unified Variational Inference Perspective," *IEEE J. Select. Areas Commun.*, pp. 1–1, 2021.

[4] T. P. Minka, "Divergence Measures and Message Passing," Microsoft, TR MSR-TR-2005-173, Jan. 2005.

[5] X. Fu *et al.*, "On Gaussian Approximation Algorithms for SCMA," in *16th IEEE ISWCS*, Oulu, Finland, 2019, pp. 155–160.

[6] G.-M. Kang *et al.*, "Message passing algorithm based on QR decomposition for an SCMA system with multiple antennas," in *IEEE ICTC*, Oct. 2017, pp. 941–944.

[7] Y. Dong *et al.*, "Efficient EP Detectors Based on Channel Sparsification for Massive MIMO Systems," *IEEE Commun. Lett.*, vol. 24, pp. 539–542, 2020.

[8] A. Mekhiche *et al.*, "Low-Complexity Scheduled Expectation Propagation based on QRD and SIC," *IEEE Commun. Lett.*, p. 5, 2023.

[9] M. Taherzadeh *et al.*, "SCMA Codebook Design," in *2014 IEEE 80th VTC-Fall*, Sep. 2014, pp. 1–5.

[10] Y. Zhou *et al.*, "SCMA codebook design based on constellation rotation," in *IEEE ICC*, Paris, France, May 2017, pp. 1–6.

[11] T. Lv and F. Long, "Graph-based low complexity detection algorithms in multiple-input–multiple-out systems: an edge selection approach," *IET Commun.*, vol. 7, pp. 1202–1210, 2013.

[12] S. Wu *et al.*, "Low-Complexity Iterative Detection for Large-Scale Multiuser MIMO-OFDM Systems Using Approximate Message Passing," *IEEE J. Sel. Top. Signal Process.*, vol. 8, pp. 902–915, 2014.

[13] M. Senst and G. Ascheid, "How the Framework of Expectation Propagation Yields an Iterative IC-LMMSE MIMO Receiver," in *IEEE Glob. Commun. Conf.*, Houston, TX, USA, 2011, pp. 1–6.

[14] L. Ping *et al.*, "Interleave division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 938–947, Apr. 2006.