A Tensor-Based Modulation Scheme with Hybrid Optimization for Unsourced Random Access

Ala Baccar[1,2], Alexis Decurninge[1], Sofiane Kharbech[1], Rima Khouja[1], Eric Pierre Simon[3], and Joumana Farah[2]
[1] *Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France*
[2] *University of Rennes, INSA Rennes, CNRS, IETR UMR 6164, Rennes, France*
[3] *University of Lille, CNRS, IEMN UMR 8520, Lille, France*
firstname.lastname@{huawei.com, univ-lille.fr, insa-rennes.fr}

***Abstract*—This paper extends the idea of tensor-based modulation (TBM) for unsourced random access by introducing a new small sub-constellation, enabling a hybrid continuous-discrete optimization approach on the tensor factors. This approach enhances the detection process and extends the tensor decomposition ability, allowing the separation of more active users. Furthermore, this hybrid TBM scheme can be combined with a preamble-based architecture that partitions bits and resources into unsourced and coherent components. This architecture improves decoding efficiency and detection accuracy, providing a robust solution for unsourced random access scenarios. The energy efficiency of the proposed scheme is evaluated under various tensor configurations and compared with state-of-the-art approaches under a quasi-static Rayleigh fading channel model.**

***Index Terms*—Unsourced random access, tensor-based modulation, non-orthogonal multiple access, tensor decomposition, massive connectivity.**

I. Introduction

Building an effective wireless communication system has become essential with the increasing level of hyperconnectivity. Within a dense network, massive random access (MRA) is a recent paradigm of wireless communications designed to serve a massive number of devices efficiently. The leading scenario considers uplink transmission of typically small-sized payloads from numerous randomly activated user equipment (UEs) to a base station (BS). As part of the MRA paradigm, unsourced random access (URA) [1] communication schemes have been demonstrated to substantially reduce the minimum energy per bit needed to ensure robust communication [2]. In the URA concept, each UE uses the same codebook, that is, the user identity cannot be revealed before decoding the signal. Ordinarily, URA does not require a resource request (scheduling), making it a low-latency and scalable design suitable for massive access.

In the literature, several URA-related works have been proposed [3]. Tensor-Based Modulation (TBM) [4] is a promising scheme in URA, demonstrating competitive performance. In TBM, the source signals, sharing the same resources, are structured as rank-one tensors; thus, user separation is performed using tensor-rank decomposition, specifically canonical polyadic decomposition (CPD). This allows signal decoding without relying on additional pilot sequences. However, since the TBM receiver is mainly a two-step decoder, signal separation and demapping, the single-user demapping stage, which relies primarily on log-likelihood ratio (LLR) approximations designed for the Grassmannian constellation, significantly underperforms as the number of bits per tensor mode increases. Coded compressive sensing has also been used in prominent URA architectures. In this context, the message is divided into blocks and each block is encoded through a compressed sensing dictionary. These blocks are concatenated using an outer product [5]. An approximate message passing algorithm is used at the decoder to solve a compressed sensing problem. Fading spread unsourced random access (FASURA) [6] is a remarkable scheme in the URA literature, showing prominent energy efficiency (EE) performance. FASURA is a preamble-based method that relies on splitting the resources into two fragments: (i) the preamble, a small part of the payload coded as unsourced and used for activity detection and channel estimation by employing an energy detection technique, and (ii) the rest of the payload coded using a non-orthogonal multiple access (NOMA) scheme, and decoded through coherent detection. In the same vein as preamble-based techniques, a combination of TBM and coherent modulation, called TBMC, was proposed in [7], where the TBM modulation shapes the unsourced part. TBMC has shown good performance, particularly under low signal-to-noise ratio (SNR) regimes, essentially because of the accurate channel estimation in the preamble.

This paper aims to design a new URA communication scheme that will enable better decoding performance. The contributions of this work are as follows: (i) embedding a small sub-constellation to the TBM structure, which allows us to apply a discrete optimization on the corresponding factor, improving its decoding performance, i.e., increasing the tensor decomposition capacity to separate users; (ii) redesigning the tensor decomposition algorithm at the receiver, referred to as hybrid alternating least squares (Hybrid-ALS), by solving an hybrid optimization problem with discrete constraints on the small sub-constellation and continuous constraints on the other tensor modes.

The rest of the paper is structured as follows. In Section II, we introduce the system model. Section III presents the encoding and decoding parts of the proposed scheme. Section IV analyzes the performance of the introduced approach. Section V summarizes the key conclusions.

Notations — x or X, $\mathbf{x}$, $\mathbf{X}$, $\mathcal{X}$ for scalar, vector (column vector), matrix, n-dimensional array ($n > 2$) . $\mathbf{I}_n$ identity ma-

trix of size n . $\otimes$ Kronecker product . $\odot$ Khatri–Rao product . $[\![.]\!]$ tensor yielded from the factor matrices in argument . $|\{.\}|$ cardinality of the set $\{.\}$. $\mathbb{E}[.]$ expectation operator . $(.)^{\mathsf{H}}$ Hermitian transpose . $(.)^*$ complex conjugate . $(.)^{\mathsf{T}}$ transpose . $(.)^{-1}$ matrix inverse . $\mathsf{vec}(.)$ vectorization operator . $\|.\|$ ℓ^2-norm.

II. System Model

We consider an uplink transmission system between a large number K of single-antenna UEs and one BS equipped with M antennas. A subset of K_{a} UEs ($K_{\mathrm{a}} \ll K$) are randomly activated and connect, in an uncoordinated manner, to the BS. Each UE k ($k = 1, \cdots, K_{\mathrm{a}}$) sends a signal $\mathbf{x}_k \in \mathbb{C}^T$ (normalized to $\|\mathbf{x}_k\|^2 = T$) using the same T available resources over a quasi-static Rayleigh fading channel $\mathbf{h}_k \in \mathbb{C}^M$ (with $\mathbb{E}\left[\|\mathbf{h}_k\|^2\right] = M$). $\mathbf{x}_k$ carries an encoded binary message $\mathbf{m}_k$ of size B. The received signal $\mathbf{Y} \in \mathbb{C}^{T\times M}$ at the M-antennas array is:

$$\mathbf{Y} = \sum_{k=1}^{K_{\mathrm{a}}} \mathbf{x}_k \mathbf{h}_k^{\mathsf{T}} + \mathbf{Z} = \mathbf{X}\mathbf{H}^{\mathsf{T}} + \mathbf{Z}, \tag{1}$$

where $\mathbf{X} \triangleq [\mathbf{x}_1, \cdots, \mathbf{x}_k, \cdots, \mathbf{x}_{K_{\mathrm{a}}}] \in \mathbb{C}^{T\times K_{\mathrm{a}}}$, $\mathbf{H} \triangleq [\mathbf{h}_1, \cdots, \mathbf{h}_k, \cdots, \mathbf{h}_{K_{\mathrm{a}}}] \in \mathbb{C}^{M\times K_{\mathrm{a}}}$, and $\mathbf{Z} \in \mathbb{C}^{T\times M}$ is the white complex-valued Gaussian noise with covariance $\sigma^2\mathbf{I}_M$.

The proposed scheme is preamble-based, as in [6] and [7]; therefore, $\mathbf{x}_k$ is composed of an unsourced part $\mathbf{x}_k^{\mathrm{u}} \in \mathbb{C}^{T^{\mathrm{u}}}$ and a coherent part $\mathbf{x}_k^{\mathrm{c}} \in \mathbb{C}^{T^{\mathrm{c}}}$ (with $T^{\mathrm{u}} + T^{\mathrm{c}} = T$). Consequently, the transmitted signals are formulated as

$$\mathbf{x}_k = \left[\mathbf{x}_k^{\mathrm{u}\mathsf{T}}, \mathbf{x}_k^{\mathrm{c}\,\mathsf{T}}\right]^{\mathsf{T}}. \tag{2}$$

Behind this decomposition, $\mathbf{m}_k \in \{0,1\}^B$ is also split into two disjoint subsets, i.e., before applying any encoding, it is represented by the vectors $\mathbf{m}_k^{\mathrm{u}} \in \{0,1\}^{B^{\mathrm{u}}}$ and $\mathbf{m}_k^{\mathrm{c}} \in \{0,1\}^{B^{\mathrm{c}}}$ for $\mathbf{x}_k^{\mathrm{u}}$ and $\mathbf{x}_k^{\mathrm{c}}$, respectively.

III. The Proposed Scheme

This section describes the steps of the proposed encoding scheme and the corresponding receiver blocks. Before delving into these details, we introduce the TBM scheme, a key technique employed in the proposed method.

A. Preliminary: TBM in brief

In a TBM encoder, the binary message of each UE is split into d disjoint segments that are mapped over d modes, constructing the tensor. Since transmitted signals are rank-one tensor structured, they are expressed as $\mathbf{x}_k = \bigotimes_{i=1}^{d} \mathbf{a}_{k,i}$, where $\mathbf{a}_{k,i} \in \mathscr{C}_i \subseteq \mathbb{C}^{T_i}$ ($i = 1, \cdots, d$) is the ith tensor factor. Thus, $\mathbf{x}_k$ is the vector-shaped d-modes tensor of shape $T_1 \times T_2 \times \cdots \times T_d$, with $\prod_{i=1}^{d} T_i = T$. In this unsourced setting, active UEs share a common sub-constellation codebook, and $\mathscr{C}_i$ is the signal sub-constellation used in mode i. The received signal in TBM is written as

$$\mathbf{y} = \sum_{k=1}^{K_{\mathrm{a}}} \mathbf{a}_{k,1} \otimes \cdots \otimes \mathbf{a}_{k,i} \otimes \cdots \otimes \mathbf{a}_{k,d} \otimes \mathbf{h}_k + \mathbf{z}, \tag{3}$$

where $\mathbf{y} \triangleq \mathsf{vec}(\mathbf{Y}) \in \mathbb{C}^{TM}$ and $\mathbf{z} \triangleq \mathsf{vec}(\mathbf{Z}) \in \mathbb{C}^{TM}$. The TBM decoder detailed in [4] consists of two main stages: multi-user separation via CPD and single-user demapping involving likelihood tests based on the codebook.

B. Encoder design

In the following, we describe separately the encoding steps for unsourced and coherent fragments.

The proposed unsourced encoding: The encoding of $\mathbf{x}_k^{\mathrm{u}}$ is based on the TBM scheme. The binary message $\mathbf{m}_k^{\mathrm{u}}$ of each user k is split into d disjoint subsets that are then mapped to the tensor factors $\mathbf{a}_{k,1}, \cdots, \mathbf{a}_{k,d-1}, \mathbf{u}_k$, of size $T_1^{\mathrm{u}}, \cdots, T_d^{\mathrm{u}}$, respectively, with $\prod_{i=1}^{d} T_i^{\mathrm{u}} = T^{\mathrm{u}}$. The last factor is distinctively denoted because, as the significant part of our contribution, it is the mode that will be concerned with discrete optimization, while the rest of the modes will endure the traditional CPD, that is, the continuous optimization. Hence, we call our approach as hybrid optimization. Thus, $\mathbf{x}_k^{\mathrm{u}}$ is formulated as

$$\mathbf{x}_k^{\mathrm{u}} = \mathbf{a}_{k,1} \otimes \cdots \otimes \mathbf{a}_{k,i} \otimes \cdots \otimes \mathbf{a}_{k,d-1} \otimes \mathbf{u}_k. \tag{4}$$

Let $\mathbf{A}_i \triangleq [\mathbf{a}_{1,i}, \cdots, \mathbf{a}_{k,i}, \cdots, \mathbf{a}_{K_{\mathrm{a}},i}] \in \mathbb{C}^{T_i^{\mathrm{u}}\times K_{\mathrm{a}}}$ ($i = 1, \cdots, d-1$) and $\mathbf{U} \triangleq [\mathbf{u}_1, \cdots, \mathbf{u}_k, \cdots, \mathbf{u}_{K_{\mathrm{a}}}] \in \mathbb{C}^{T_d^{\mathrm{u}}\times K_{\mathrm{a}}}$. Without loss of generality, $\mathbf{U}$ is assumed to be the mode of index d. In the proposed scheme, the U-factor is mapped with a few bits (typically 2 to 5). This choice allows a discrete optimization of this mode within the CPD while maintaining reasonable complexity at the receiver. To manage decoding complexity, the sub-constellation of this factor (denoted by $\mathscr{C}_d$) should remain limited in size, i.e., a low number of bits should be attributed to this mode, which is thus not FEC encoded. In fact, one of the main drawbacks of the state-of-the-art TBM is its reliance on a continuous-domain CPD algorithm, which is suboptimal as the signals belong to a defined sub-constellation, i.e., a discrete domain. The proposed scheme could be extended to include multiple discrete modes. However, this would result in a significantly higher decoding complexity at the receiver. Figure 1 illustrates the mapping process for the unsourced part $\mathbf{m}_k^{\mathrm{u}}$. Note that the set of bits mapped to the U-factor is used to parameterize the FEC encoder to maintain the dependence between modes and increase diversity.

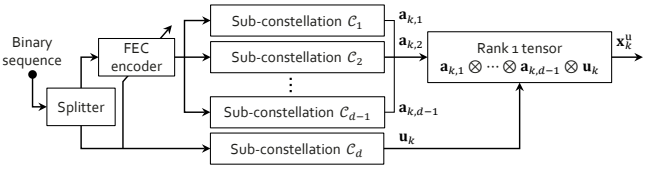

Fig. 1. Encoding diagram for the unsourced fragment.

The coherent encoding: The second fraction of the payload is encoded using NOMA modulation, as was done in the coherent part of FASURA [6]. The concept behind this non-orthogonal transmission is to spread each user's data across the available resources using a complex spreading sequence. More specifically, for each user k, a spreading

matrix $\mathbf{W}_j \in \mathbb{C}^{L \times T^{\mathrm{c}}/L}$ with spreading factor L is used, where $j = 1, \cdots, J$, and J is the number of spreading sequences. Each user picks a spreading sequence based on its unsourced payload. Let $\mathbf{q}_k^{\mathrm{c}} \in \mathbb{C}^{T^{\mathrm{c}}/L}$ be the QPSK-modulated signal mapping $\mathbf{m}_k^{\mathrm{c}}$. The encoded signal is expressed as

$$\mathbf{x}_k^{\mathrm{c}} = \mathsf{vec}\Big(\mathbf{W}_j \odot \mathbf{q}_k^{\mathrm{c}\,\mathsf{T}}\Big). \tag{5}$$

C. Decoder design

We start by describing the proposed unsourced decoder. Recall that one of the main drawbacks of the previous TBM-based schemes [4], [7] is that their tensor decomposition uses a continuous optimization on discrete modes, which is a relaxation of the actual problem. That is why we apply a discrete optimization to one of the tensor modes, namely mode d, to increase the CPD capacity regarding the number of supported users. A continuous optimization is performed on the remaining modes ($i = 1, \cdots, d-1$). The corresponding hybrid optimization problem can be formulated as follows:

$$\{\widehat{\mathbf{a}}_{k,i}\}, \{\widehat{\mathbf{u}}_k\}, \{\widehat{\mathbf{h}}_k\} = \underset{\substack{\{\mathbf{a}_{k,i}\},\{\mathbf{u}_k\},\{\mathbf{h}_k\} \\ \mathbf{u}_k \in \{\mathbf{u}_1, \cdots, \mathbf{u}_{2^n}\}}}{\arg\min} \Bigg\| \mathbf{y} - \sum_{k=1}^{K_{\mathrm{a}}} \mathbf{a}_{k,1} \otimes \cdots \otimes \mathbf{a}_{k,d-1} \otimes \mathbf{u}_k \otimes \mathbf{h}_k \Bigg\|^2, \tag{6}$$

where $\mathbf{a}_{k,i} \in \mathbb{C}^{T_i^{\mathrm{u}}}$, $\mathbf{h}_k \in \mathbb{C}^M$, $i = 1, \cdots, d-1$, $k = 1, \cdots, K_a$, and $\mathbf{u}_k \in \{\mathbf{u}_1, \cdots, \mathbf{u}_{2^n}\}$, where n is the size of the small payload encoded by $\mathbf{u}_k$. While a continuous optimization is used to determine $\{\mathbf{a}_{k,i}\}$ and $\{\mathbf{h}_k\}$, the mode d factor is computed by a projection on the sub-constellation to determine the closest vector of the latter w.r.t. each computed factor. For this reason, a moderate size of the dth sub-constellation is chosen, knowing that it encodes only a small portion of the unsourced payload.

The proposed Hybrid-ALS technique is summarized in Algorithm 2, which includes a partial-ALS procedure (Algorithm 1). Algorithm 2 takes as inputs the received signal in its tensor form $\mathcal{Y}$ and the noise power σ^2, and outputs the estimated factor matrices ($\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \widehat{\mathbf{U}}$) and the channel matrix ($\widehat{\mathbf{H}}$). We define a set $\mathscr{L}$ that contains the indices k of the U-factor that are considered well estimated, based on a pre-fixed threshold η. Note that the criterion used in Step 10 aims at minimizing the difference between the received power of user k and its expected power (supposed to be known in the absence of large-scale fading). An iterative continuous optimization (Steps 6-10 of Algorithm 1) is applied to the first $d-1$ modes of the tensor along with the vectors $\mathbf{u}_k$ whose indices do not belong to the set $\mathscr{L}$. The $\mathbf{u}_k$ vectors of indices in $\mathscr{L}$, stored in matrix $\mathbf{U}_{\mathrm{p}}$, are fixed during the ALS iterations. Then, the columns of the U-factor are projected on the sub-constellation (Step 9 of Algorithm 2) and the set $\mathscr{L}$ is updated (Steps 10-13 of Algorithm 2).

Algorithm 3 summarizes the decoding steps of the full receiver, which works as follows. Regarding the unsourced part of the received signal, the unsourced payloads are decoded and then re-encoded to estimate the channel coefficients of the devices. Afterward, the successfully decoded sequences are employed to identify the users' spreading sequences needed to decode the coherent part. Based on this information, a linear minimum mean squared error (LMMSE) equalizer and FEC decoder are applied to the coherent part of the received signal to estimate the coherent payload. A cyclic redundancy check is then performed to exclude the wrongly decoded coherent sequences. Then, the correctly decoded parts are discarded by successive interference cancellation (SIC). After concatenation of both parts (Step 7 of Algorithm 3), a global SIC is iterated.

Algorithm 1: Partial-ALS

input : $\mathcal{Y}, \{\mathbf{A}_i^{\mathrm{init}}\}_{i=1}^{d-1}, \mathbf{U}_{\mathrm{p}}, \mathbf{H}^{\mathrm{init}}, \sigma^2$
output: $\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \widehat{\mathbf{A}}_{d+1}$

1 $\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1} \leftarrow \{\mathbf{A}_i^{\mathrm{init}}\}_{i=1}^{d-1}$ /* signal factors */
2 Initialize $\widehat{\mathbf{U}}_{\mathrm{p}}$
3 $\widehat{\mathbf{A}}_d \leftarrow \mathbf{U} \triangleq \Big[\mathbf{U}_{\mathrm{p}}, \widehat{\mathbf{U}}_{\mathrm{p}}\Big]$ /* the U-factor */
4 $\widehat{\mathbf{A}}_{d+1} \leftarrow \mathbf{H}^{\mathrm{init}}$ /* channel factor */
5 **while** *convergence criterion not fulfilled and the maximum number of iterations not achieved* **do**
6 **for** $i \leftarrow 1$ **to** $d+1$ **do**
7 $\mathbf{K}_i \leftarrow \widehat{\mathbf{A}}_{d+1} \odot \cdots \odot \widehat{\mathbf{A}}_{i+1} \odot \widehat{\mathbf{A}}_{i-1} \odot \cdots \odot \widehat{\mathbf{A}}_1$
8 $\mathbf{Y}_i \triangleq$ The mode-i unfolding of $\mathcal{Y}$
9 **if** $i \neq d$ **then**
10 $\widehat{\mathbf{A}}_i \leftarrow \mathbf{Y}_i \mathbf{K}_i^* \left(\mathbf{K}_i^{\mathsf{T}} \mathbf{K}_i^* + \sigma^2 \mathbf{I}_{K_{\mathrm{a}}}\right)^{-1}$
11 **else**
12 $\mathbf{K} \leftarrow [\mathbf{K}_d]_{:, K'_{\mathrm{a}}+1 \cdots K_{\mathrm{a}}}$ /* K'_{a} is the number of columns of $\mathbf{U}_{\mathrm{p}}$ */
13 $\mathcal{Y}_{\mathrm{p}} \triangleq \Big[\!\Big[\Big[\widehat{\mathbf{A}}_i\Big]_{:, 1 \cdots K'_{\mathrm{a}}} \Big]\!\Big]_{i=1}^{d+1}$
14 $\mathbf{Y}_{\mathrm{p}} \triangleq$ The mode-d unfolding of $\mathcal{Y}_{\mathrm{p}}$
15 $\widehat{\mathbf{U}}_{\mathrm{p}} \leftarrow (\mathbf{Y}_i - \mathbf{Y}_{\mathrm{p}}) \times \mathbf{K}^* \left(\mathbf{K}^{\mathsf{T}} \mathbf{K}^* + \sigma^2 \mathbf{I}_{K_{\mathrm{a}} - K'_{\mathrm{a}}}\right)^{-1}$
16 $\widehat{\mathbf{A}}_d \leftarrow \Big[\mathbf{U}_{\mathrm{p}}, \widehat{\mathbf{U}}_{\mathrm{p}}\Big]$

IV. Simulation results

A massive random access scenario is simulated with a variable number of active users and different tensor configurations. The performance of the proposed scheme is assessed using two well-known metrics in URA: the probability of miss detection, $p_{\mathrm{md}} = \mathbb{E}\left[\left|\left\{\mathcal{L} \backslash \hat{\mathcal{L}}\right\}\right| / K_{\mathrm{a}}\right]$, and the probability of false alarm, $p_{\mathrm{fa}} = \mathbb{E}\left[\left|\left\{\hat{\mathcal{L}} \backslash \mathcal{L}\right\}\right| / \left|\left\{\hat{\mathcal{L}}\right\}\right|\right]$. $\mathcal{L}$ represents the list of all transmitted messages, i.e., the ground-truth, and $\hat{\mathcal{L}}$ is the list of estimated messages at the receiver.

All tested configurations correspond to a fixed number of $T = 1008$ resources. Fig. 2 presents the results of the first set of simulations with $B = 80$, $M = 8$, and two possible configurations for the proposed scheme:

- Config. 1 (orange plot): the unsourced part encodes only 14 bits on 360 resources, leaving 66 bits on 648

Algorithm 2: Hybrid-ALS

input : $\mathcal{Y}, \sigma^2, \mathscr{C}_d$
output: $\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \widehat{\mathbf{U}}, \widehat{\mathbf{H}}$

Initialize $\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \mathbf{U}_\mathrm{p}, \widehat{\mathbf{H}}$
while *maximum number of iterations not achieved* **do**
  $\{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \widehat{\mathbf{H}} \leftarrow$ Partial-ALS $\left(\mathcal{Y}, \{\widehat{\mathbf{A}}_i\}_{i=1}^{d-1}, \mathbf{U}_\mathrm{p}, \widehat{\mathbf{H}}, \sigma^2\right)$
  $\mathbf{K} \leftarrow \widehat{\mathbf{H}} \odot \widehat{\mathbf{A}}_{d-1} \odot \cdots \odot \widehat{\mathbf{A}}_1$
  $\mathbf{Y}_d \triangleq$ The mode-d unfolding of $\mathcal{Y}$
  $\widehat{\mathbf{U}} \leftarrow \mathbf{Y}_d \mathbf{K}^* \left(\mathbf{K}^\mathsf{T}\mathbf{K}^* + \sigma^2 \mathbf{I}_{K_\mathrm{a}}\right)^{-1}$
  $\mathscr{L} \leftarrow \emptyset$ /* the set of user indices with good U-factor estimation, $K'_\mathrm{a} = |\{\mathscr{L}\}|$ */
  for $k \leftarrow 1$ **to** K_a **do**
    $\widehat{\mathbf{u}}_k \leftarrow \arg\min_{\mathbf{u} \in \mathscr{C}_d} \|\widehat{\mathbf{u}}_k - \mathbf{u}\|^2$ /* project $\widehat{\mathbf{u}}_k$ on the sub-constellation $\mathscr{C}_d$ */
    $c_k = \frac{\left|\prod_{i=1}^{d-1} \|\widehat{\mathbf{a}}_{k,i}\|^2 \|\widehat{\mathbf{u}}_k\|^2 \|\widehat{\mathbf{h}}_k\|^2 - TM\right|}{TM}$ /* criterion to select the best estimated $\widehat{\mathbf{u}}_k$ */
    if $c_k < \eta$ /* in our simulations, we set η to 0.5 */ **then**
      $\mathscr{L} \leftarrow \mathscr{L} \cup \{k\}$
  $\mathbf{U}_\mathrm{p} \leftarrow [\widehat{\mathbf{u}}_k]_{k \in \mathscr{L}}$

Algorithm 3: The full receiver

input : The received signal $\mathcal{Y}$
output : The set of decoded messages $\{\widehat{\mathbf{m}}_k\}$

while *Global SIC stopping criterion not achieved* **do**
  /* Unsourced part */
  while *SIC stopping criterion not achieved* **do**
    Estimate the factors using Algorithm 2.
    Update $\{\widehat{\mathbf{m}}_k^\mathrm{u}\}$, the set of the successfully demapped unsourced fragments of the messages.
    Refine channel estimation for the reconstructed transmitted signals related to $\{\widehat{\mathbf{m}}_k^\mathrm{u}\}$ and substract their product from the unsourced fragment of the received signal.
  /* Coherent part */
  Apply the iterative procedure from [6] to obtain $\{\widehat{\mathbf{m}}_k^\mathrm{c}\}$.
  $\{\widehat{\mathbf{m}}_k\} \leftarrow \left\{\left[\widehat{\mathbf{m}}_k^{\mathrm{u}\mathsf{T}}, \widehat{\mathbf{m}}_k^{\mathrm{c}\mathsf{T}}\right]^\mathsf{T}\right\}$
  Refine channel estimation for the fully reconstructed transmitted signals related to $\{\widehat{\mathbf{m}}_k\}$ and subtract their product from the received signal.

resources for the coherent part. This allows the use of joint maximum likelihood (ML) demapping/decoding at the receiver, which calculates the likelihood of each constellation vector to be mapped to a binary sequence of size 14. The number of resources per mode is $(3, 2, 2, 2, 15)$. The last tensor mode encodes only 5 bits, corresponding to a sub-constellation of size 32, designed by merging canonical and Fourier bases, yielding 30 vectors, which are then completed with two Gaussian vectors. Cubesplit Grassmannian sub-constellations [8] are used for the remaining modes. Polar codes [9] are used for FEC.

- Config. 2 (blue plot): the unsourced part encodes 34 bits on 512 resources, leaving 46 bits on 496 resources. The tensor configuration is $(2, 2, 2, 2, 2, 2, 8)$. FEC is not used in this case and a hard demapper is applied at the receiver. This choice is explained by the fact that with such a big unsourced payload, the complexity of the joint ML receiver becomes prohibitive. Besides, since the number of bits per resource is high in the tensor modes, resorting to LLR approximations would lead to inaccurate values under CubeSplit sub-constellations, and the efficiency of FEC decoding would deteriorate.

Note that, to allow a full comparative analysis of the performance with FASURA, the payload split in Config. 1 is the same as in FASURA. For the coherent part, we set the length spreading factor to $L = 3$ and the number of spreading sequences to $J = 2^{14}$. Therefore, Config. 1 and FASURA differ only in their unsourced part, which will determine their performance comparison.

Fig. 2 presents the EE in terms of the number of active users. The EE is expressed by the required bit energy over noise density ratio to achieve a p_md lower than 0.1. The two configurations of the proposed scheme are compared with the schemes FASURA, TBM, TBMC. For TBMC, $(B^\mathrm{u}, B^\mathrm{c}) = (20, 60)$ and $(T^\mathrm{u}, T^\mathrm{c}) = (5 \times 5 \times 5 \times 4 = 500, 508)$, while for TBM $T = 21 \times 12 \times 4 = 1008$. The results show that Config. 1 achieves nearly the same performance as FASURA. This is because their unsourced detectors perform very closely at low SNRs, while the unsourced receiver of Config. 1 outperforms FASURA only at high SNR (results not presented here for brevity). However, FASURA does not allow a number of active users higher than 160. Using the Config. 2 of our proposed scheme will enable us to serve more users with a better EE starting from $K_\mathrm{a} = 160$.

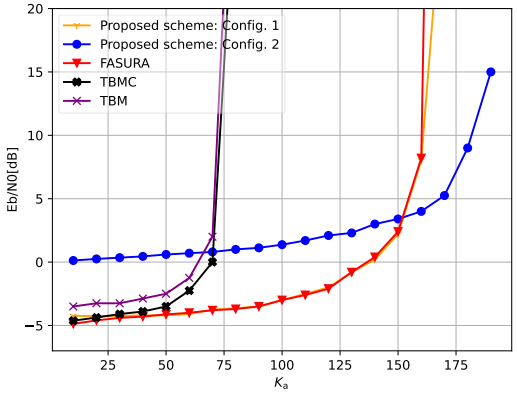

Fig. 2. Energy efficiency ($p_\mathrm{md} < 0.1$) for $B = 80$, and $M = 8$.

In fact, starting from 170 users, the energy detector of FASURA scores a value of p_fa higher than 90%, which overwhelms the coherent part even at a high energy-per-bit to noise ratio. The same problem is observed with our Config. 1 which faces a high decoding error rate in the unsourced part. However, this problem can be alleviated in our scheme

by increasing the number of resources for the unsourced part to reduce the p_{fa}. At the same time, a larger unsourced payload needs to be set, while decreasing the coherent payload. This solution is not possible for FASURA because of complexity reasons since the energy detector performs an exhaustive search over the constellation to estimate the correlation between the received signal and each constellation point. This becomes prohibitive for a large constellation codebook. Nevertheless, this solution is possible in our unsourced detector because it is (i.e., its complexity) not limited by the number of bits. By allocating more resources to the unsourced part, the second configuration (Config. 2) allowed for a reduction in decoding errors at a much higher number of users, thereby improving the decoding performance of the coherent part. However, a loss in performance w.r.t. Config. 1 and FASURA was observed at low energy-per-bit to noise ratio values due to the absence of FEC. Nonetheless, the results demonstrate that the proposed scheme provides greater flexibility than FASURA regarding resource and bit allocation by dynamically adapting the configuration following the number of active users. Note that TBM and TBMC are limited to a maximum of 70 active users.

In Figure 3, a larger payload is considered, with $B = 200$ and a higher number of antennas $M = 16$. In this configuration of the proposed scheme $(B^{\text{u}}, B^{\text{c}}) = (33, 167)$ and $(T^{\text{u}}, T^{\text{c}}) = (384, 624)$ with the same receiver as Config. 2, while for FASURA, $(B^{\text{u}}, B^{\text{c}}) = (14, 186)$ and $(T^{\text{u}}, T^{\text{c}}) = (360, 648)$. The results show that Config. 3 performs better than FASURA overall, demonstrating that the proposed scheme offers better solutions for handling a larger payload.

Decoder complexity analysis: The decoder complexity of the proposed scheme is compared to that of FASURA. The analysis begins with the complexity of the coherent components. Both schemes employ the same NOMA scheme, and the computational load of this part is primarily driven by the LMMSE equalization, which equals $\mathcal{O}(K_{\text{a}}^3 T^{\text{c}}/L)$. FASURA's unsourced part uses an energy detector that estimates a statistical score to evaluate each constellation vector. This yields a complexity of $\mathcal{O}(T^{\text{u}} M 2^{B^{\text{u}}})$. The tensor decomposition and SIC steps mainly drive the complexity of the proposed unsourced detector. Therefore, its complexity is dominated by that of the least squares step in the ALS iterations, multiplied by ρ and α, which are respectively, the number of SIC iterations and the number of iterations in Algorithm 1 and Algorithm 2 combined, yielding $\mathcal{O}(\rho\alpha(K_{\text{a}}^3 + MT^{\text{u}}))$. Contrary to FASURA, the complexity order of our proposed scheme grows cubically in terms of the number of active users, which is disadvantageous when the number of active users is large. However, the size of the unsourced constellation has a negligible impact on the complexity, whereas the exponential growth of the FASURA scheme's complexity in terms of this parameter prevents the consideration of large values of B^{u}.

V. Conclusion

In the context of preamble-based URA systems, this paper introduces a novel unsourced detector capable of detecting a larger number of active users at low SNRs compared to existing schemes. When combined with a preamble-based design, it demonstrates superior performance in terms of energy efficiency. Three tensor configurations are tested: the first yields results comparable to the FASURA baseline, while the other two achieve better decoding performance than FASURA at higher numbers of active users. From a broader perspective, the future goal is to enhance LLR approximations under dense Grassmannian constellation to enable the integration of soft decoding for the case of large unsourced payloads.

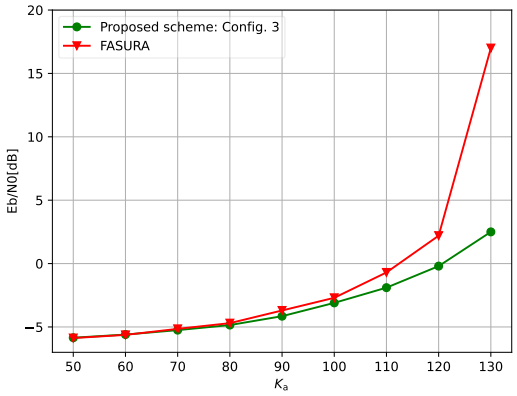

Fig. 3. Energy efficiency ($p_{\text{md}} < 0.1$) for $B = 200$, and $M = 16$.

References

[1] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5g and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 615–637, 2021.

[3] G. Liva and Y. Polyanskiy, "Unsourced multiple access: A coding paradigm for massive random access," *Proceedings of the IEEE*, vol. 112, no. 9, pp. 1214–1229, 2024.

[4] A. Decurninge, I. Land, and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 552–556, 2021.

[5] A. Fengler, P. Jung, and G. Caire, "Sparcs for unsourced random access," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6894–6915, 2021.

[6] M. Gkagkos, K. R. Narayanan, J.-F. Chamberland, and C. N. Georghiades, "Fasura: A scheme for quasi-static massive mimo unsourced random access channels," in *2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*, 2022, pp. 1–5.

[7] A. Rech, A. Decurninge, and L. G. Ordóñez, "Unsourced random access with tensor-based and coherent modulations," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–6.

[8] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured grassmannian constellation for non-coherent simo communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1948–1964, 2020.

[9] V. Bioglio, C. Condo, and I. Land, "Design of polar codes in 5g new radio," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 29–40, 2021.