MASSIVE UNSOURCED RANDOM ACCESS BASED ON BILINEAR VECTOR APPROXIMATE MESSAGE PASSING

Ramzi Ayachi, Mohamed Akrout,Volodymyr Shyianov, Faouzi Bellili, and Amine Mezghani

Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada.

ABSTRACT

This paper introduces a new algorithmic solution to the massive unsourced random access (mURA) problem. The proposed uncoupled compressed sensing (UCS)-based scheme relies on slotted transmissions and takes advantage of the inherent coupling provided by the users' spatial signatures in the form of channel correlations across slots to completely eliminate the need for concatenated coding. As opposed to all existing methods, the proposed solution combines the steps of activity detection, channel estimation, and data decoding into a unified mURA framework. It capitalizes on the bilinear vector approximate message passing (Bi-VAMP) algorithm, tailored to fit the inherent constraints of mURA. Exhaustive computer simulations demonstrate that the proposed scheme outperforms recent coupled and uncoupled mURA schemes in massive connectivity/MIMO setup.

Index Terms— Massive MIMO, Unsourced random access, Bilinear Vector Approximate Message Passing.

1. INTRODUCTION

Next generation wireless systems must satisfy the demand for massive connectivity by providing fast and reliable access to the network for a large number of low-power devices. For the variety of IoT applications, such as massive machine type communication (mMTC), it has become critical to investigate recent aggressive/non-orthogonal connection strategies to fulfill the low latency and high reliability requirements [1]. Conventionally, one of the popular approaches to massive connectivity within the paradigm of grant-based random access advocates the use of a two-phase transmission scheme. In the first phase, known data is transmitted to detect the active users and estimate their channels while in the second phase the identified active users are scheduled to transmit their messages. Within this framework, it was demonstrated that the joint channel estimation and device activity detection can be modeled as an instance of the compressed sensing (CS) problem. Several existent CS recovery techniques, such as the approximate message passing (AMP) algorithm [2] and the covariance-based CS (CB-CS) scheme [3] were studied in the context of massive grant-free access (see [4], [5], and references therein). Another line of work has also investigated the use of random access strategies based on conventional ALOHA [6] and coded slotted ALOHA [7]. Massive unsourced random access is a form of grant-free access that is well suited to a myriad of emerging applications in which the base station (BS) is solely interested in the transmitted messages without necessarily knowing the IDs of the users. The information-theoretic work in [8] introduced a random coding existence bound for mURA using a random Gaussian codebook with maximum likelihood-decoding at the BS. Furthermore, the work in [9] provided an extension of [8] to the case of a MIMO BS with an accurate characterization of the asymptotic spectral efficiency as a function of the number of antennas/users. For computational tractability, most of the recent algorithmic solutions for mURA rely on the slotted transmission framework and CS-based encoding/decoding while adopting different approaches to couple the respective slots. Computationally tractable mURA schemes based on the coupled CS (CCS) framework have been proposed in [10] and [11]. Some effort went into designing efficient algorithms that handle MIMO processing, in the context of CCS-based [3] and uncoupled CS (UCS)-based mURA. We provide in this paper a novel algorithmic solution to the mURA problem which is able to accommodate much more active users than the number of receive antennas elements at the BS. Our UCS-based scheme eliminates completely the need for concatenated coding, which was used in most of the existing work on CS-based URA [3, 11, 12]. The proposed UCS-based technique capitalizes on the recently introduced Bi-VAMP algorithm [13] which provides computationally efficient approximate implementations of both max-sum and sum-product loopy belief propagation.

2. SYSTEM MODEL

Consider a single BS serving a network consisting of K_a active devices, each equipped with a single antenna element. The number of active devices K_a is assumed to be much smaller than the total number of devices K, i.e., $K_a \ll K$. Every communication round, each active device aims to convey B information bits to the BS. Prior to transmission, each user encodes its information bits using a random Gaussian codebook, $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{2^B}\} \subset \mathbb{C}^n$, is the same for all the users in mURA. To contain the exponential growth in the number of codewords so as to limit the computational complexity, each active user divides its B-bit packet into L equal-sized chunks of $J = \frac{B}{L}$ bits, which are separately encoded into $\frac{n}{L}$ symbols each. By utilizing the slotted transmission framework and the new reduced-size codebook, $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{\frac{n}{L} \times 2^J}$ for all users and over all slots, the decoding problem can now be tackled by computationally efficient methods. The columns of $\boldsymbol{\mathcal{A}} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{2^J}]$ are the set of new codewords used by each active user to encode each of its L chunks of the original packet. We model the transmission of the $\{j^{th}\}_{j=1}^{2^J}$ codeword, $\mathbf{a}_j$, by the $\{k^{th}\}_{k=1}^{K_a}$ active user in the $\{l^{th}\}_{l=1}^{L}$ slot by the variable $\delta_{j,k,l}$:

$$\delta_{j,k,l} = \begin{cases} 1 & \text{if user } k \text{ transmits codeword } \mathbf{a}_j \text{ in the } l^{th} \text{ slot} \\ 0 & \text{otherwise} \end{cases}$$

TThe uplink channel vector from the $\{m^{th}\}_{k=1}^{M_r}$ receive antenna element at the BS to the active users is denoted as $\mathbf{h}_m = [h_{1,m}, h_{2,m}, \cdots, h_{K_a,m}]^\mathsf{T} \in \mathbb{C}^{K_a}$, where $[.]^\mathsf{T}$ stands for the transpose of a matrix. By denoting the large- and small-scale fading parameters as β_k and $g_{k,m}$, respectively, we have $h_{k,m} = \sqrt{\beta_k} g_{k,m}$. We also assume Rayleigh block fading in which the channel vector remains constant over the entire observation window which is smaller than the coherence time. The uplink received signal at the $\{m^{th}\}_{m=1}^{M_r}$ BS antenna in the $\{l^{th}\}_{l=1}^{L}$ slot can be expressed as follows:

$$\mathbf{y}_{m,l} = \sum_{k=1}^{K} \sum_{j=1}^{2^J} h_{k,m} \delta_{j,k,l} \mathbf{a}_j + \mathbf{w}_{m,l}. \tag{1}$$

Here $\mathbf{w}_{m,l}$, is a Gaussian noise vector with independent and identically distributed (i.i.d.) components, i.e., $\mathbf{w}_{m,l} \sim \mathcal{N}\left(0, \sigma_w^2 \mathbf{I}\right)$. Let us denote by $\boldsymbol{\Delta}_l$ the assignment matrix pertaining to the $\{l^{th}\}_{l=1}^{L}$ slot whose columns has a single non-zero element which is equal to 1. The position of the non-zero element in the $\{k^{th}\}_{k=1}^{K_a}$ column indicates which codeword was transmitted by the k^{th} user in the l^{th} slot. Moreover, let $\mathbf{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{M_r}] \in \mathbb{C}^{K_a \times M_r}$ denote the channel matrix between all active users and the BS and $\mathbf{Y}_l = [\mathbf{y}_{1,l}, \mathbf{y}_{2,l}, ..., \mathbf{y}_{M_r,l}]$ be the signal measured by the entire antenna array in the l^{th} slot With these notations in mind, it follows from (1) that the input-output relationship of the system is expressed as follows:

$$\mathbf{Y}_l = \boldsymbol{\mathcal{A}} \boldsymbol{\Delta}_l \mathbf{H} + \mathbf{W}_l \quad \text{for} \quad l = 1, \ldots, L. \tag{2}$$

At this point, we have a total of L per-slot equations each of which involving two unknown matrices (namely, $\mathbf{H}$ and $\boldsymbol{\Delta}^l$) with the unknown channel matrix, $\mathbf{H}$, being common to all slots. By further combining all observation matrices in one matrix, $\mathbf{Y} = \left[\mathbf{Y}_1^\mathsf{T}, \mathbf{Y}_2^\mathsf{T}, \cdots, \mathbf{Y}_L^\mathsf{T}\right]^\mathsf{T}$, we obtain:

$$\mathbf{Y} = \mathbf{A} \boldsymbol{\Delta} \mathbf{H} + \mathbf{W}. \tag{3}$$

with $\mathbf{A} \triangleq (\boldsymbol{\mathcal{A}} \otimes \mathbf{I})$ in which $\otimes$ denotes the Kronecker product. Moreover, the overall assignment block matrix $\boldsymbol{\Delta} = \left[\boldsymbol{\Delta}_1^\mathsf{T}, \boldsymbol{\Delta}_2^\mathsf{T}, \cdots, \boldsymbol{\Delta}_L^\mathsf{T}\right]^\mathsf{T}$ obeys a specific structure to ensure some intuitive constraints in mURA. In fact, the number of codewords sent in each $\{l^{th}\}_{l=1}^{L}$ slot must be equal to K_a so that each active user transmits a single codeword per slot. Note here that different users can transmit the same codeword in the same slot but ultimately every user transmits exactly L codewords over the whole transmission period (i.e., over all slots).

3. BILINEAR VECTOR APPROXIMATE MESSAGE PASSING ALGORITHM

In this section, we modify the Bi-VAMP algorithm originally introduced in [13] to fit the specific constraints of the mURA problem. The modified Bi-VAMP algorithm uses a probabilistic observation model to jointly recover the two unknown matrices $\mathbf{X} \triangleq \mathbf{A}\boldsymbol{\Delta}$ and $\mathbf{H}$ from their noise-corrupted product in (3), i.e.:

$$\mathbf{Y} = \mathbf{X}\mathbf{H} + \mathbf{W}. \tag{4}$$

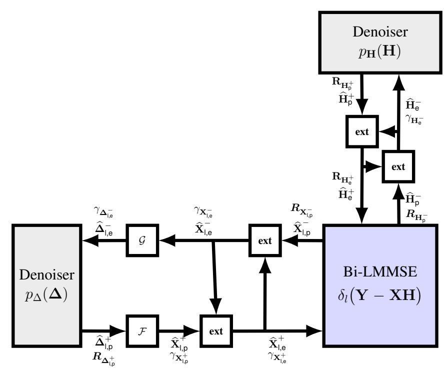

***Fig. 1**: Block diagram of the adapted Bi-VAMP algorithm with its three modules: two denoising modules incorporating the prior information, $p_{\mathbf{H}}(\cdot)$ and $p_{\Delta}(\cdot)$ and the approximate bi-LMMSE module. The three modules exchange extrinsic information/messages calculated through the $\boxed{\textbf{ext}}$ blocks, $\boxed{\mathcal{G}}$ and $\boxed{\mathcal{F}}$ are used to enforce the column-wise structure of $\boldsymbol{\Delta}_l$ in mURA using its row-wise extrinsic information provided by the standard Bi-VAMP algorithm.*

In principle, the standard Bi-VAMP algorithm enables the use of different priors on the columns of $\mathbf{H}$ as well as on the rows of $\mathbf{X}$. As mentioned in Section 2, however, the specific structure of $\boldsymbol{\Delta}$ enforces a prior on the columns of each $\boldsymbol{\Delta}_l$. To appropriately modify Bi-VAMP and make it accommodate the mURA-induced prior on the columns of $\{\boldsymbol{\Delta}_l\}_{l=1}^{L}$ we introduce the functions $\mathcal{F}$ and $\mathcal{G}$ in its block diagram depicted in Fig. 1. We represent $\boldsymbol{\Delta_l}$ by its columns vector $\boldsymbol{\Delta_l} = [\boldsymbol{\delta}_{1,l}, \boldsymbol{\delta}_{2,l}, \cdots, \boldsymbol{\delta}_{K_a,l}]$ and $\mathbf{X}_l = \mathcal{A}\boldsymbol{\Delta_l}$ by its rows $\mathbf{X}_l = [\boldsymbol{x}_{1,l}^\mathsf{T}, \boldsymbol{x}_{2,l}^\mathsf{T}, \cdots, \boldsymbol{x}_{J,l}^\mathsf{T}]^\mathsf{T}$ as illustrated in the factor graph depicted in Fig. 2 associated with the block diagram in Fig. 1. There, we show the variable nodes, $\boldsymbol{\delta}_{k,l}, \boldsymbol{x}_{l,j}, \boldsymbol{h}_m$, the prior factor nodes, $p_{\boldsymbol{\delta}_1}(\boldsymbol{\delta}_1)$ and $p_{\boldsymbol{h}_m}(\boldsymbol{h}_m)$, and the labels, f_{jm}^l and f_g^l, which are used as a shorthand notations, respectively, for the following factor nodes:

$$f^l\left(\boldsymbol{h}_m, \boldsymbol{x}_{j,l}\right) \triangleq \mathcal{N}\left(y^l_{jm}; \boldsymbol{x}_{j,l}\boldsymbol{h}_m, \gamma_w^{-1}\right) \tag{5}$$

$$f_g^l \triangleq \delta\left(\mathcal{A}\boldsymbol{\Delta}_l - \mathbf{X}_l\right) \tag{6}$$

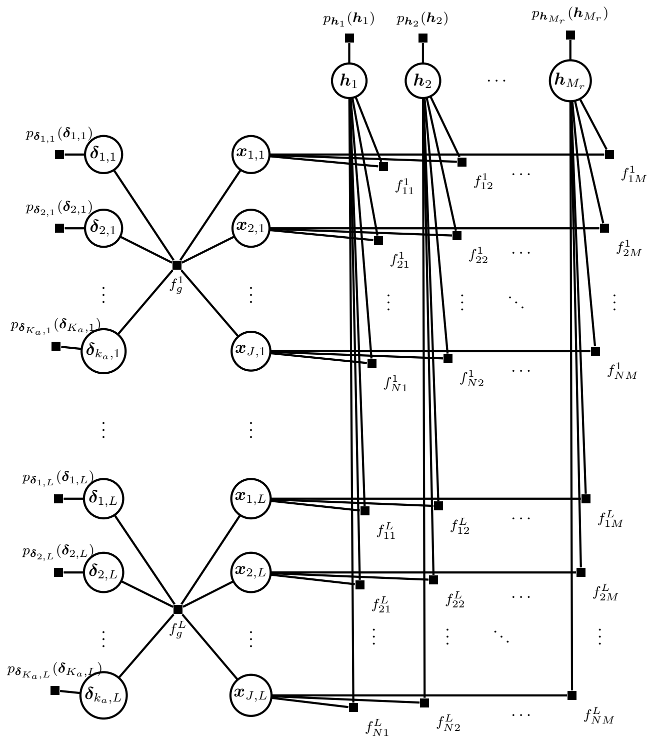

Fig. 2: *Factor graph for our proposed mURA scheme where* $\boldsymbol{\delta}_{k,l}$ *and* $\boldsymbol{x}_{j,l}$ *represent, respectively, the variable nodes pertaining to the columns of* $\boldsymbol{\Delta}$, *and the rows of* $\mathbf{X}$ *while* $\boldsymbol{h}_m$ *represent the variable nodes pertaining to the columns of* $\mathbf{H}$ *columns.*

The function $\mathcal{F}$ computes the posterior mean/precision of the rows of $\mathbf{X}$ from the posterior mean/precision of the columns of $\boldsymbol{\Delta}_l$ as follows:

$$\widehat{\mathbf{X}}_{\mathsf{p}}^{+} = \mathbf{A}\,\widehat{\boldsymbol{\Delta}}_{\mathsf{p}}^{+}, \tag{7a}$$

$$\gamma^{-1}_{\mathbf{X}^{+}_{\mathsf{l,p}}} = \frac{L}{n}\mathrm{Tr}(\boldsymbol{\mathcal{A}}\,\boldsymbol{R}_{\boldsymbol{\Delta}^{+}_{\mathsf{l,p}}}\,\boldsymbol{\mathcal{A}}^{\mathsf{T}}). \tag{7b}$$

In doing so, we then approximate the posterior covariance matrices of the rows of $\mathbf{X}_l$ by a common scaled identity matrix wherein the scaling factor is the mean of their diagonal entries. Such approximation of the messages by their respective mean and scalar variance is a common practice in the message passing paradigm.

Similarly, the function $\mathcal{G}$ finds the extrinsic mean/precision of the columns of $\boldsymbol{\Delta}_l$ from those of the rows of $\mathbf{X}$ which are obtained from the output of the Bi-LMMSE block as depicted in Fig. 1. Indeed, we update the required extrinsic precision as follows:

$$\widehat{\boldsymbol{\Delta}}_{\mathsf{e}}^{-} = \beta\,\mathbf{A}^{\mathsf{T}}\widehat{\mathbf{X}}_{\mathsf{e}}^{-}, \tag{8a}$$

$$\gamma^{-1}_{\boldsymbol{\Delta}^{-}_{\mathsf{l,e}}} = \alpha\,\mathrm{Tr}(\mathcal{A}^{\mathsf{T}}\,\gamma^{-1}_{\mathbf{X}^{-}_{\mathsf{l,e}}}\mathcal{A}) \tag{8b}$$

with α and β being two normalization coefficients which depends on the type of the codebook $\mathcal{A}$. Moreover, the Bi-LMMSE block computes the posterior mean of $\mathbf{H}$ and the associated common precision under i.i.d. Gaussian priors on its columns. Those posterior quantities are used by the extrinsic blocs $\boxed{\textbf{ext}}$ to compute the extrinsic mean and precision ($\widehat{\boldsymbol{H}}_{\mathsf{e}}^{-}$, $\gamma_{\boldsymbol{H}_{\mathsf{e}}^{-}}$) as input to the the denoiser block of $\mathbf{H}$. Owing to the inherent symmetry between $\boldsymbol{X}$ and $\boldsymbol{H}$, the Bi-LMMSE block computes the posterior mean/precision of $\mathbf{X}$ under i.i.d. Gaussian priors on its rows in a similar way.

Now, those posterior quantities are used by the blocks $\boxed{\textbf{ext}}$ and $\boxed{\mathcal{G}}$ to compute the extrinsic means and precisions, $\widehat{\boldsymbol{\Delta}}_{\mathsf{l,e}}^{-}$ and $\gamma_{\boldsymbol{\Delta}^{-}_{\mathsf{l,e}}}$, respectively. Given these extrinsic inputs, the two denoiser blocks find the posterior means, $\widehat{\boldsymbol{H}}_{\mathsf{p}}^{+}$ and $\widehat{\boldsymbol{\Delta}}_{\mathsf{l,p}}^{+}$, along with their respective covariance matrices $\mathbf{R}_{\boldsymbol{H}_{\mathsf{p}}^{+}}$ and $\boldsymbol{R}_{\boldsymbol{\Delta}^{+}_{\mathsf{l,p}}}$ as follows:

$$\widehat{\boldsymbol{h}}_{\mathsf{m,p}}^{+} = \mathbf{g}_{\mathbf{h}}\left(\widehat{\boldsymbol{h}}_{\mathsf{m,e}}^{-}, \gamma_{\boldsymbol{H}_{\mathsf{e}}^{-}}\right), \mathbf{R}_{\boldsymbol{H}_{\mathsf{p}}^{+}} = \frac{\gamma^{-1}_{\boldsymbol{H}_{\mathsf{e}}^{-}}}{M_r}\sum_{m=1}^{M_r}\mathbf{g}'_{\mathbf{h}}\left(\widehat{\boldsymbol{h}}_{\mathsf{m,e}}^{-}, \gamma_{\boldsymbol{H}_{\mathsf{e}}^{-}}\right) \tag{9a}$$

$$\widehat{\boldsymbol{\delta}}_{\mathsf{k,p}}^{l,+} = \mathbf{g}_{\boldsymbol{\delta}}\left(\widehat{\boldsymbol{\delta}}_{\mathsf{k,e}}^{l,-}, \gamma_{\boldsymbol{\Delta}^{-}_{\mathsf{l,e}}}\right), \mathbf{R}_{\boldsymbol{\Delta}^{+}_{\mathsf{l,p}}} = \frac{\gamma^{-1}_{\boldsymbol{\Delta}^{-}_{\mathsf{l,e}}}}{K_a}\sum_{k=1}^{K_a}\mathbf{g}'_{\boldsymbol{\delta}}\left(\widehat{\boldsymbol{\delta}}_{\mathsf{k,l,e}}^{-}, \gamma_{\boldsymbol{\Delta}^{-}_{\mathsf{l,e}}}\right). \tag{9b}$$

Finally, (9a) and (9b) are used to update the extinsic information which is used by the Bi-LMMSE block in the next iteration.

$$\mathbf{g}_{\mathbf{h}}\left(\widehat{\boldsymbol{h}}, \gamma_{\boldsymbol{H}}^{-1}\right) = \frac{\int \boldsymbol{h}p_{\mathbf{h}}(\boldsymbol{h})\mathcal{N}\left(\boldsymbol{h}; \widehat{\boldsymbol{h}}, \gamma_{\mathbf{H}}^{-1}\boldsymbol{I}\right)d\boldsymbol{h}}{\int p_{\mathbf{h}}(\boldsymbol{h})\mathcal{N}\left(\boldsymbol{h}; \widehat{\boldsymbol{h}}, \gamma_{\boldsymbol{H}}^{-1}\boldsymbol{I}\right)d\boldsymbol{h}} \tag{10}$$

$$\mathbf{g}_{\boldsymbol{\delta}}\left(\widehat{\boldsymbol{\delta}}, \gamma_{\boldsymbol{\Delta}}^{-1}\right) = \frac{\int \boldsymbol{\delta}p_{\boldsymbol{\delta}}(\boldsymbol{\delta})\mathcal{N}\left(\boldsymbol{\delta}; \widehat{\boldsymbol{\delta}}, \gamma_{\boldsymbol{\Delta}}^{-1}\boldsymbol{I}\right)d\boldsymbol{\delta}}{\int p_{\boldsymbol{\delta}}(\boldsymbol{\delta})\mathcal{N}\left(\boldsymbol{\delta}; \widehat{\boldsymbol{\delta}}, \gamma_{\boldsymbol{\Delta}}^{-1}\boldsymbol{I}\right)d\boldsymbol{\delta}} \tag{11}$$

$$\mathbf{g}'_{\mathbf{h}}\left(\widehat{\boldsymbol{h}}, \gamma_{\boldsymbol{H}}^{-1}\right) = \frac{\partial\mathbf{g}_{\mathbf{h}}\left(\widehat{\boldsymbol{h}}, \gamma_{\boldsymbol{H}}^{-1}\right)}{\partial\widehat{\boldsymbol{h}}} \tag{12}$$

$$\mathbf{g}'_{\delta}\left(\widehat{\boldsymbol{\delta}}, \gamma_{\boldsymbol{\Delta}}^{-1}\right) = \frac{\partial\mathbf{g}_{\delta}\left(\widehat{\boldsymbol{\delta}}, \gamma_{\boldsymbol{\Delta}}^{-1}\right)}{\partial\widehat{\boldsymbol{\delta}}} \tag{13}$$

3.1. Choice of the priors on $\boldsymbol{\Delta}$ and $\mathbf{H}$

Let us denote the prior on $\boldsymbol{\Delta}_l$ by $p_{\boldsymbol{\Delta}_l}(.)$ which is assumed to be column-wise separable, i.e., $p_{\boldsymbol{\Delta}_l}(\boldsymbol{\Delta}_l) = \prod_{k=1}^{K_a} p_{\boldsymbol{\delta}_{k,l}}(\boldsymbol{\delta}_{k,l})$ since the messages sent by different users are independent. Consider the set of standard unit vectors $\{\mathbf{u}_j\}_{j=1}^{2^J}$ where all element of $\mathsf{u_j}$ are equal to 0 except the j^{th} element which is equal to 1. The prior on the columns of $\mathbf{\Delta}_l$ is then given by:

$$p_{\boldsymbol{\delta}_{k,l}}(\boldsymbol{\delta}_{k,l}) = \frac{1}{2^J} \sum_{j=1}^{2^J} \delta(\boldsymbol{\delta}_{k,l} - \mathbf{u}_j), \tag{14}$$

wherein $\delta(.)$ is the Dirac delta function. Similarly, we assume a column-wise separable prior on $\mathbf{H}$, i.e., $p_{\mathsf{H}}(\mathbf{H}) = \prod_{i=1}^{M_r} p_{\mathsf{h}_i}(\mathbf{h}_i)$ with $p_{\mathsf{h}_i}(\mathbf{h}_i) = \mathcal{N}(\mathbf{h}_i; 0, \sigma^2\mathbf{I})$.

4. SIMULATION RESULTS

In this section, we use exhaustive Monte-Carlo simulations to assess the performance of the proposed URA scheme and benchmark it against the mURA schemes introduced in [3], [14] in terms of spectral efficiency and the error probability. We refer to these two massive MIMO-based mURA schemes in [3], [14] and our scheme by "Covariance-based", "Clustering-based" and "Bi-VAMP-based", respectively. For the "Covariance-based" scheme, we communicate $B = 104$ information bits per packet over $L = 17$ slots for each user. We set the parity allocation for the outer tree code to $p = [0, 8, 8, \ldots, 14]$ and we fix the total rate of the outer code to $R_{out} = 0.437$ by using $J = 14$ coded bits per slot. For the "Clustering-based" scheme, we communicate over $L = 6$ slots with $B = 102$ information bits per packet for each user. For the proposed "Bi-VAMP-based" scheme, we enable each user to communicate $B = 100$ information bits which are conveyed over $L = 10$ slots with 10 bits per slot. We also simulate 5 data points by fixing the ratio $K_a/M_r = 0.6$ with $K_a \in \{50, 75, 100, 150, 200\}$ active users and $M_r \in \{30, 45, 60, 90, 120\}$ antennas at the BS. For the baseline schemes, however, we simulate 3 data points only with $K_a \in \{50, 75, 100\}$ active users and $M_r \in \{30, 45, 60\}$. This is due to their prohibitive computational complexity for 150 and 200 active users. In Fig. 3 we determine the largest per-user spectral efficiency $\mu = B/n$ which leads to a target error probability of 10^{-2}.
Fig. 3 shows that the proposed "Bi-VAMP-based" scheme outperforms both baseline schemes for large numbers of active users as is desirable in massive connectivity setups. Indeed, at the same target error probability, the proposed scheme is able to serve a much higher number of users using the same number of receive antennas by better exploiting the small-scale fading signatures It does so by overcoming the fundamental limitations of EM-based clustering involved in the clustering-based scheme [14] while removing completely the need for concatenated coding which is key to the covariance-based scheme [3]. For instance, in presence of 200 active users it is possible to keep the error probability as low as 10^{-2} at a spectral efficiency around 0.21 bits per channel use with the proposed scheme. In Fig. 4 we fix the number of active users to 200 and the SNR to 30 dB while varying the spectral efficiency from 0.065 to 0.2. From the spectral efficiency, it is possible to calculate the blocklength as $n = B/\mu$. The figure depicts the variation of the error probability as a function of the spectral efficiency for two different number of receive antennas namely $M_r = 64$ and $M_r = 128$. There, it is seen that with $M_r = 128$ the error probability drops below 10^{-2} even for spectral efficiency up to 0.2.

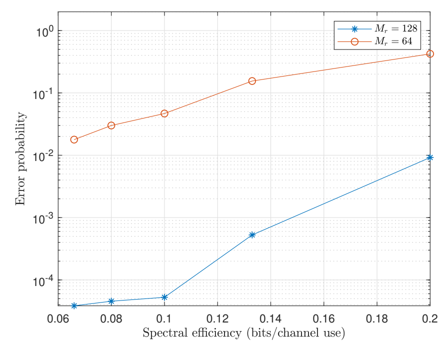

Fig. 4: *Performance of the proposed scheme as a function of the spectral efficiency and number of receive antenna elements, M_r, with a fixed number of active users $K_a = 200$ and SNR = 30 dB.*

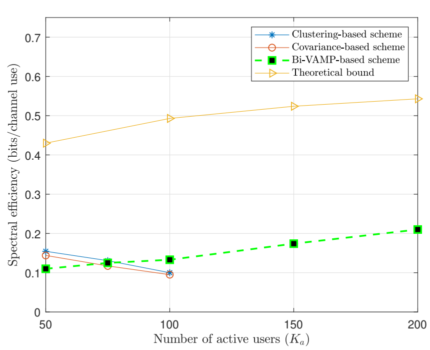

Fig. 3: *Spectral efficiency a function of the number of active users at a target error probability of 10^{-2}.*

5. CONCLUSION

In this work, we proposed an integrated algorithmic solution to the mURA problem. Our method is based on slotted transmissions and the Bi-VAMP algorithm which was tailored to fit the particular constraints of mURA. The proposed approach takes advantage of the inherent coupling provided by the spatial signatures in the form of channel correlations across slots and eliminates the need for concatenated coding. It combines the steps of activity detection, channel estimation and data decoding into a unified URA framework. Computer simulations reveal that the proposed scheme exhibits considerable performance enhancement in massive connectivity as well as massive MIMO set-ups compared to the state-of-the-art methods.

6. REFERENCES

[1] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Processing Magazine*, vol. 35, pp. 88 – 99, 2018.

[2] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 18 914–18 919, 2009.

[3] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, "Massive mimo unsourced random access," *arXiv preprint arXiv:1901.00828*, 2019.

[4] Y. Yuan, Z. Yuan, G. Yu, C.-h. Hwang, P.-k. Liao, A. Li, and K. Takeda, "Non-orthogonal transmission technology in lte evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, 2016.

[5] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.

[6] L. G. Roberts, "Aloha packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review*, vol. 5, p. 28–42, 1975.

[7] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," *IEEE Communications Magazine*, vol. 53, pp. 144 – 150, 2015.

[8] Y. Polyanskiy, "A perspective on massive random-access," *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2157–8117, 2017.

[9] B. Tadele, V. Shyianov, F. Bellili, and A. Mezghani, "Age-limited capacity of massive mimo," *arXiv preprint arXiv:2007.05071*, 2020.

[10] A. Fengler, P. Jung, and G. Caire, "Sparcs for unsourced random access," *arXiv preprint arXiv:1901.06234*, 2019.

[11] R. Calderbank and A. Thompson, "Chirrup: a practical algorithm for unsourced multiple access," *arXiv preprint arXiv:1811.00879*, 2019.

[12] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for uncoordinated multiple access," *arXiv preprint arXiv:1809.04745*, 2019.

[13] M. Akrout, A. Housseini, F. Bellili, and A. Mezghani, "Bilinear generalized vector approximate message passing," *arXiv preprint arXiv:2009.06854*, 2020.

[14] V. Shyianov, F. Bellili, A. Mezghani, and E. Hossain, "Massive unsourced random access based on uncoupled compressive sensing: Another blessing of massive mimo," *IEEE Journal on Selected Areas in Communications*, pp. 820 – 834, 2020.