Improved Bayesian Learning Detectors for Uplink Grant-Free MIMO-NOMA

Boran Yang, Xiaoxu Zhang, *Member, IEEE*,
Li Hao, *Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—Grant-free non-orthogonal multiple access (GF-NOMA) and multiple-input multiple-output (MIMO) techniques are key enablers for massive machine-type communications (mMTC) in 5G cellular Internet of Things. On the other hand, compressed sensing (CS) is widely accepted for multiuser detection, due to the sporadic traffic of mMTC. In this letter, we propose two Bayesian-based CS algorithms for uplink grant-free MIMO-NOMA. Exploiting the iterative expectation maximization (EM) and maximum ratio combining techniques, the spatially enhanced sparse Bayesian learning (SE-SBL) algorithm is developed to alternately update hyperparameter values and combined observation signals. Furthermore, we embed the generalized approximate message passing technique into the SE-SBL and propose a low-complexity computational approach. In particular, the aforementioned Bayesian algorithms do not require any prior knowledge of user activity level and noise power. Simulation results show that the proposed Bayesian methods exhibit a superior performance gain over the state-of-the-art.**

***Index Terms*—Massive machine-type communications, grant-free, non-orthogonal multiple access, multiple-input multiple-output, Bayesian compressed sensing**

I. Introduction

Massive machine-type communications (mMTC) focuses on uplink access of large-scale users with bursty short-time transmissions, where conventional wireless communication networks cannot adequately meet the performance requirements [1]. Non-orthogonal multiple access (NOMA) [2], which provides massive connectivity through non-orthogonal resource allocation among users with limited resources, has been extensively studied. In addition, a grant-free NOMA (GF-NOMA) scheme [3], in which users can arbitrarily transmit data to the base station (BS) without a complicated interactive handshaking process, is widely accepted to mitigate transmission delay and signaling overhead. mMTC has an inherent sporadic traffic characteristic, i.e., only a small percentage of potential users transmit data to the BS [4]. However, the BS cannot identify which users are active at any given time, which motivates additional challenges for multiuser detection (MUD) [5]. Fortunately, sparse activity detection can be formulated as a compressed sensing (CS) problem and efficiently solved by using sophisticated sparse reconstruction algorithms [1]–[5].

Ridge detector (RD) and Lasso detector (LD) proposed in [6] for MUD, by exploiting the sparsity of the transmitted signal. In terms of detection accuracy, RD and LD significantly surpass linear least squares (LS) and minimum mean square error (MMSE) methods [3]. However, both works have not carried a further step toward mining the underlying spatial structure of user activity. Inspired by the above observations, the authors in [7] explored the spatial correlation of user activity in the multiple input multiple output (MIMO) case and proposed an effective MUD approach for multiantenna reception, based on approximate message passing (AMP). The authors in [8] and [9] proposed the variant forms of bilinear generalized AMP algorithm for joint channel estimation and MUD in MIMO-enabled GF-NOMA systems. Block greedy-type algorithms, such as block orthogonal matching pursuit (B-OMP) [10], block compressive sampling matching pursuit (B-CoSaMP) [11], and block subspace pursuit (B-SP) [12], have been proposed to perform MUD by exploiting the block sparsity of a set of sparse vectors. However, the proposed algorithms require accurate knowledge of user activity or noise power, which is generally not practical in mMTC scenarios.

In this letter, we formulate the MUD problem of uplink grant-free MIMO-NOMA system as a block compressed sensing one, by integrating the spatial structure of user activity. The main contributions of this letter are summarized as follows: 1) Bayesian CS algorithms, i.e., spatially enhanced sparse Bayesian learning (SE-SBL) and spatially enhanced generalized approximate message passing sparse Bayesian learning (SE-GAMP-SBL) algorithms, are proposed to enable effective MUD for multiantenna reception; 2) the SE-SBL algorithm alternately obtains estimated values for hyperparameters and inferred value for combined observation signal via iterative expectation maximization (EM) [13] and combining [14] techniques; 3) the SE-GAMP-SBL algorithm significantly reduces the computational complexity of SE-SBL by approximating the posterior distribution without performance penalty; 4) the proposed two algorithms do not require user sparsity and noise power as auxiliary prior information.

II. System Model

We consider a typical uplink grant-free MIMO-NOMA system, where K single-antenna users transmit data to the

Manuscript received July 25, 2023; revised August 21, 2023; accepted September 14, 2023. Date of publication September 14, 2023; date of current version September 14, 2023. This work was supported in part by National Key R&D Program of China under Grant 2018YFB1801104, in part by NSFC Project under Grant 62001399, and in part by Sichuan Science and Technology Program under Grant 2022NSFSC0910. The associate editor coordinating the review of this letter and approving it for publication was Lukas T. N. Landau. *(Corresponding author: Xiaoxu Zhang.)*

B. Yang, X. Zhang, and L. Hao are with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: bryang@my.swjtu.edu.cn; xiaoxuzhang@swjtu.edu.cn; l-hao@swjtu.edu.cn).

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 541-24 Thessaloniki, Greece, and also with the Cyber Security Systems and Applied AI Research Center, Lebanese American University (LAU), Beirut 1102-2801, Lebanon (e-mail: geokarag@auth.gr).

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2023.3317374

BS equipped with M antennas. For a certain time slot, K_a users are activated to transmit data while others remain silent, $K_a \ll K$. The transmitted symbol of the k-th user is denoted as z_k. If the k-th user is active, $z_k \in \mathcal{A}$, otherwise, $z_k = 0$, where $\mathcal{A}$ denotes a constellation set. The spreading sequence for the k-th user is $\boldsymbol{\Psi}_k = (\Psi_{1k}, \Psi_{2k}, \cdots, \Psi_{Nk})^{\mathrm{T}}$ with N subcarriers. To gain insight into the fundamental limit of MUD performance, we assume synchronous reception and perfect channel estimation [3], [4], [6]. In this case, the received signal from the m-th antenna of the BS can be expressed as

$$\boldsymbol{y}^m = \sum_{k=1}^{K} \mathrm{diag}(\boldsymbol{h}_k^m)\boldsymbol{\Psi}_k z_k + \boldsymbol{w}^m = \boldsymbol{H}^m \boldsymbol{z} + \boldsymbol{w}^m, \quad (1)$$

where $\boldsymbol{h}_k^m = (h_{1k}^m, h_{2k}^m, \cdots, h_{Nk}^m)^{\mathrm{T}}$ is the flat Rayleigh fading channel coefficient between user k and BS antenna m, $\boldsymbol{z} = (z_1, z_2, \cdots, z_K)^{\mathrm{T}}$ is the transmitted signal vector for all K users, $\boldsymbol{H}^m \in \mathbb{C}^{N\times K}$ is the equivalent channel matrix, $\boldsymbol{w}^m \in \mathbb{C}^{N\times 1}$ is the noise vector with complex Gaussian distribution $\mathcal{CN}(0\,, \sigma^2\mathbf{I})$, $\mathrm{diag}(\cdot)$ denotes regular diagonalization.

Although the channel coefficients between each user and BS exhibit non-correlation, the user activities are associated for different receiving antennas, as illustrated in Fig. 1. To exploit this special structure, we extend (1) to the block CS model for multiantenna reception,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \quad (2)$$

where $\mathbf{H} = \mathrm{blkdiag}(\boldsymbol{H}^1, \boldsymbol{H}^2, \cdots, \boldsymbol{H}^M) \in \mathbb{C}^{MN\times MK}$ is the block diagonal matrix, $\mathbf{x} = \mathrm{vec}(\boldsymbol{z}, \boldsymbol{z}, \cdots, \boldsymbol{z}) \in \mathbb{C}^{MK\times 1}$ is the block sparse vector, $\mathbf{y} = \mathrm{vec}(\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^M) \in \mathbb{C}^{MN\times 1}$ is the linear observation vector, $\mathbf{w} = \mathrm{vec}(\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^M) \in \mathbb{C}^{MN\times 1}$ is the linear noise vector, $\mathrm{vec}(\cdot)$ denotes column-wise vectorization, $\mathrm{blkdiag}(\cdot)$ denotes block diagonalization.

A key observation of this paper is to detect the active users and the symbols transmitted, which is equivalent to estimating $\boldsymbol{z}$ from the block sparse signal $\mathbf{x}$ recovered from the block CS model (2). Note that $\mathbf{x}$ can be decoupled into M blocks with a common sparse pattern. Fortunately, this unique property can be extracted to enhance MUD performance.

III. The SE-SBL Algorithm

The classical SBL algorithm [3] does not leverage the high correlation of user activity in the spatial domain induced by multiantenna reception. To this end, we introduce the diversity combining technique into the SBL framework to develop a spatially enhanced SBL algorithm named SE-SBL.

Similar to [3], ignoring the finite-alphabet constraints, SE-SBL employs a standard two-layer hierarchical prior model dominated by a set of empirical hyperparameters. Since most of the elements in $\mathbf{x}$ are zero for small user activity factor p_a, in the first layer, a zero-mean Gaussian prior distribution is assigned to each element of $\mathbf{x}$,

$$\begin{aligned} p(\mathbf{x}\,;\boldsymbol{\alpha}) &= \prod_{m=1}^{M}\prod_{k=1}^{K} \mathcal{CN}(x_{mk} \mid 0\,, \alpha_{mk}^{-1}) \\ &= (2\pi)^{-MK}|\boldsymbol{\Lambda}|\exp(-\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x}), \end{aligned} \quad (3)$$

where x_{mk} is the k-th element of the m-th block in $\mathbf{x}$,

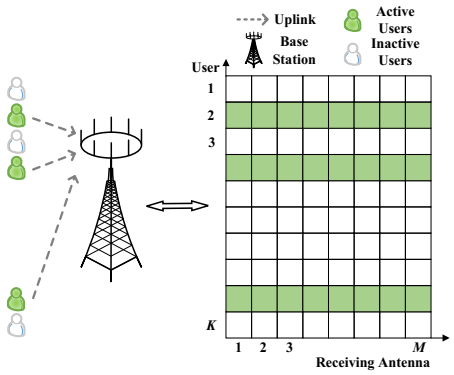

Fig. 1. A typical uplink grant-free MIMO-NOMA system and the spatial structure of user activity.

$\boldsymbol{\alpha} = (\alpha_{11}, \cdots, \alpha_{mk}, \cdots, \alpha_{MK})^{\mathrm{T}}$ is a non-negative hyperparameter designed to regulate the block sparse signal $\mathbf{x}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\alpha})$. Specifically, the element x_{mk} tends to zero when α_{mk} approaches positive infinity. In the second layer, the hyperparameter $\boldsymbol{\alpha}$ follows Gamma prior distribution dominated by the parameters $\{a\,, b\}$,

$$p(\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{k=1}^{K} \Gamma(a)^{-1} b^a \alpha_{mk}^{a-1} e^{-b\alpha_{mk}}. \quad (4)$$

For simplicity, we consider β as the inverse of the noise variance, i.e., $\beta \triangleq 1/\sigma^2$. The Gaussian noise generally satisfies $p(\mathbf{w}) = \mathcal{CN}(\mathbf{w} \mid 0\,, \beta^{-1}\mathbf{I})$. Similarly, the Gamma prior distribution dominated by the parameters $\{c\,, d\}$ is assigned to β, i.e., $p(\beta) = \Gamma(c)^{-1} d^c \beta^{c-1} e^{-d\beta}$. The likelihood function of the received signal $\mathbf{y}$ can be expressed as

$$p(\mathbf{y} \mid \mathbf{x}\,; \beta) = \mathcal{CN}(\mathbf{y} \mid \mathbf{H}\mathbf{x}\,, \beta^{-1}\mathbf{I}). \quad (5)$$

Based on Bayesian principle, the posterior distribution of the block sparse signal $\mathbf{x}$ is calculated as follows

$$p(\mathbf{x} \mid \mathbf{y}\,; \boldsymbol{\alpha}\,, \beta) = \frac{p(\mathbf{y} \mid \mathbf{x}\,; \beta)p(\mathbf{x}\,; \boldsymbol{\alpha})}{\int p(\mathbf{y} \mid \mathbf{x}\,; \beta)p(\mathbf{x}\,; \boldsymbol{\alpha})d\mathbf{x}}. \quad (6)$$

Inserting (3) and (5) into (6), and performing several approximation operations [3], the posterior distribution of $\mathbf{x}$ can be simplified as

$$p(\mathbf{x} \mid \mathbf{y}\,; \boldsymbol{\alpha}\,, \beta) = \mathcal{CN}(\mathbf{x} \mid \boldsymbol{\mu}\,, \boldsymbol{\Sigma}), \quad (7)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the posterior mean and variance as follows

$$\mathbf{x}^{\mathrm{SBL}} = \boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\mathbf{H}^{\mathrm{T}}\mathbf{y}, \quad (8)$$

$$\boldsymbol{\Sigma} = (\beta\mathbf{H}^{\mathrm{T}}\mathbf{H} + \boldsymbol{\Lambda})^{-1}. \quad (9)$$

The posterior mean is treated as the estimated block sparse signal. After obtaining $\mathbf{x}^{\mathrm{SBL}}$, we can acquire the observation signals recovered from each receiving antenna according to

$$\hat{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M) = \mathrm{vec}^{-1}(\mathbf{x}^{\mathrm{SBL}}\,, M). \quad (10)$$

Exploiting the spatial diversity of interference pattern suffered by the transmitted signal and the spatial correlation of user activity induced by multiantenna reception, the combining technique [14] is applied to $\hat{\boldsymbol{X}}$, for creating a combined

observation, on which updating the hyperparameters is attempted. Benefiting from the spatial structure of user activity, the combining process of multiple observed versions can be interpreted as an improvement of the posterior distribution of the transmitted signal $\boldsymbol{z}$. Consider M observations of the transmitted signal, $\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \cdots,\ \boldsymbol{X}_M$, the combined observation signal can be expressed as

$$\hat{\boldsymbol{z}} = \sum_{m=1}^{M} \eta_m \boldsymbol{X}_m, \tag{11}$$

where η_m is the combination weight for the m-th branch.

Any of the combining techniques can be applied to $\hat{\boldsymbol{X}}$, we focus on the maximum ratio combining (MRC) scheme, since it is optimal when the interference on each observed signal is independent [14]. Assuming that the local recovery noises have the same power, the weighting $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_M)^{\mathrm{T}}$ is proportional to the amplitude ratio. In the MRC scheme, each branch is weighted proportional to its root mean square of the observed signal level. Furthermore, the combined signal is expanded into a block-wise structure serving as the improved posterior mean of $\mathbf{x}$,

$$\mathbf{x}^{\mathrm{SE-SBL}} = \hat{\boldsymbol{\mu}} = \mathrm{vec}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{z}}, \cdots, \hat{\boldsymbol{z}}). \tag{12}$$

The iterative EM algorithm is implemented to update the hyperparameters [13]. In the E-step, we first fixed the current values of hyperparameters $\{\boldsymbol{\alpha}^{(t-1)}, \beta^{(t-1)}\}$, followed by calculating the expectation value of the maximum complete likelihood function incorporating the extended value of the combined signal $\mathbf{x}^{\mathrm{SE-SBL}}$, also known as the Q-function. In the M-step, we derive the partial derivatives of Q-function with respect to $\boldsymbol{\alpha}$ and β, and take its values zero to obtain the updated hyperparameters as follows

$$\alpha_{mk}^{(t)} = \frac{2a-1}{2b + \hat{\mu}_{mk}^2 + \Sigma_{mk,mk}}, \tag{13}$$

$$\beta^{(t)} = \frac{NM + 2c - 2}{2d + \|\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\mu}}\|_2^2 + \mathrm{tr}[\boldsymbol{\Sigma}\mathbf{H}^{\mathrm{T}}\mathbf{H}]}. \tag{14}$$

The proposed SE-SBL approach dedicated to MUD with multiantenna reception is summarized in Algorithm 1.

Algorithm 1 SE-SBL

Initializing: $\forall m, k$: $\alpha_{mk}^{(0)} = 1$, $\beta^{(0)} = 1$, $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{0}$, and $t = 1$.
Step 1. Calculating the posterior distribution of block sparse signal $\mathbf{x}$ according to (8) and (9);
Step 2. Creating the combined observation signal $\hat{\boldsymbol{z}}$ according to (11), and extending it into a block-wise pattern according to (12);
Step 3. Updating the hyperparameters according to (13) and (14).
Repeating the above steps until $\left\|\hat{\boldsymbol{\mu}}^{(t)} - \hat{\boldsymbol{\mu}}^{(t-1)}\right\|_2 \leq \varepsilon$, where ε is a prescribed threshold value of 10^{-6}.

IV. The SE-GAMP-SBL Algorithm

SE-SBL is an effective MUD algorithm for multiantenna reception and possesses excellent recovery performance without prior knowledge of user activity factor. However, the inverse operation of the $KM \times KM$ covariance matrix is computed in the process of iteratively updating the block sparse signal $\mathbf{x}$. The SE-SBL algorithm has a computational complexity of $\mathcal{O}(K^3M^3)$, which raises new challenges for MUD in the case of massive connectivity.

Motivated by GAMP-SBL [15], we embed the GAMP technique into the SE-SBL framework to replace matrix inversion and develop a computationally efficient MUD method entitled SE-GAMP-SBL. GAMP is an approximate implementation of loop belief propagation (BP) via central limit theorem and Taylor series expansion.

SE-GAMP-SBL also features a two-layer hierarchical prior model. Let $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \beta\}$ denote the known hyperparameters and it will be updated in M-step. GAMP provides the approximated posterior distributions of $\mathbf{x}$ and $\mathbf{u} = \mathbf{H}\mathbf{x}$. First, GAMP approximates the posterior distribution of x_{mk} as

$$\begin{aligned} p(x_{mk} \mid \mathbf{y}, \hat{r}_{mk}, \tau_{mk}^r, \boldsymbol{\Theta}) &= \mathcal{CN}(x_{mk} \mid \mu_{mk}^x, \phi_{mk}^x) \\ &= \frac{p(x_{mk} \mid \boldsymbol{\Theta})\mathcal{CN}(x_{mk} \mid \hat{r}_{mk}, \tau_{mk}^r)}{\int p(x_{mk} \mid \boldsymbol{\Theta})\mathcal{CN}(x_{mk} \mid \hat{r}_{mk}, \tau_{mk}^r)\mathrm{d}x_{mk}}, \end{aligned} \tag{15}$$

where

$$\mu_{mk}^x \triangleq \frac{\hat{r}_{mk}}{\alpha_{mk}\tau_{mk}^r + 1}, \tag{16}$$

$$\phi_{mk}^x \triangleq \frac{\tau_{mk}^r}{\alpha_{mk}\tau_{mk}^r + 1}. \tag{17}$$

in which $\hat{r}_{mk}$ denotes the approximated Gaussian noise corrupted version of x_{mk} and τ_{mk}^r denotes the variance of noise. Besides, GAMP approximates the posterior distribution of u_{mn} as

$$\begin{aligned} p(u_{mn} \mid \mathbf{y}, \hat{p}_{mn}, \tau_{mn}^p, \boldsymbol{\Theta}) &= \mathcal{CN}(u_{mn} \mid \mu_{mn}^u, \phi_{mn}^u) \\ &= \frac{p(y_{mn} \mid u_{mn}, \boldsymbol{\Theta})\mathcal{CN}(u_{mn} \mid \hat{p}_{mn}, \tau_{mn}^p)}{\int p(y_{mn} \mid u_{mn}, \boldsymbol{\Theta})\mathcal{CN}(u_{mn} \mid \hat{p}_{mn}, \tau_{mn}^p)\mathrm{d}u_{mn}}, \end{aligned} \tag{18}$$

where

$$\mu_{mn}^u \triangleq \frac{\tau_{mn}^p \beta y_{mn} + \hat{p}_{mn}}{\beta\tau_{mn}^p + 1}, \tag{19}$$

$$\phi_{mn}^u \triangleq \frac{\tau_{mn}^p}{\beta\tau_{mn}^p + 1}, \tag{20}$$

in which $\hat{p}_{mn}$ denotes the approximated Gaussian noise corrupted version of u_{mn} and τ_{mn}^p denotes the variance of noise. Consequently, we harvest the approximated posterior distributions of $\mathbf{x}$ and $\mathbf{u}$ replacing the matrix inversion with the GAMP algorithm. The computational complexity is reduced to the order of $\mathcal{O}(NKM^2)$.

SE-GAMP-SBL algorithm performs weighted sum of multiple observed signals by using diversity combining technique. Since multiple observation signals inherently share the same sparse pattern, the combined signal is much closer to the original signal $\boldsymbol{z}$. Following the SE-SBL algorithm, SE-GAMP-SBL adopts the MRC scheme to achieve the diversity gain, i.e., $\hat{\boldsymbol{z}} = (\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_K)^{\mathrm{T}}$ with $\hat{z}_k = \sum_{m=1}^{M} \eta_m \mu_{mk}^x$. Then, the combined observation symbols are expanded into block-wise patterns, $\hat{x}_{mk} = \hat{\mu}_{mk}^x = \hat{z}_k$, $\forall m$, $\forall k$.

The SE-GAMP-SBL algorithm is also concerned with updating the hyperparameters $\boldsymbol{\Theta}$. The Q-function indicating the log-likelihood expectation is calculated and split into two parts based on the required estimated hyperparameters,

$$\begin{aligned} Q(\boldsymbol{\alpha}, \beta \mid \boldsymbol{\alpha}^{(t-1)}, \beta^{(t-1)}) &= \mathrm{E}[\ln p(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}, \beta)] \\ &= \mathrm{E}[\ln p(\mathbf{x}; \boldsymbol{\alpha})p(\boldsymbol{\alpha})] + \mathrm{E}[\ln p(\mathbf{y} \mid \mathbf{x}; \beta)p(\beta)]. \end{aligned} \tag{21}$$

For the Q-function containing only the $\boldsymbol{\alpha}$ term, we have

$$\begin{aligned}\mathrm{E}[\ln p(\mathbf{x}\,;\boldsymbol{\alpha})p(\boldsymbol{\alpha})] = \sum_{m=1}^{M}\sum_{k=1}^{K}\{(a-1)\ln\alpha_{mk} - b\alpha_{mk} \\ + \frac{1}{2}\ln\alpha_{mk} - \frac{1}{2}\alpha_{mk}\mathrm{E}[(\hat{x}_{mk})^2]\} + \text{const.}\end{aligned} \tag{22}$$

Since $p(\hat{x}_{mk} \mid \mathbf{y}\,, \hat{r}_{mk}, \tau^r_{mk}, \boldsymbol{\Theta})$ approximates the Gaussian distribution with posterior mean $\hat{\mu}^x_{mk}$ and variance ϕ^x_{mk}, $\mathrm{E}[(\hat{x}_{mk})^2]$ can be expressed as

$$\begin{aligned}\mathrm{E}[(\hat{x}_{mk})^2] &= (\hat{\mu}^x_{mk})^2 + \phi^x_{mk} \\ &= (\sum_{m=1}^{M}\eta_m\mu^x_{mk})^2 + \frac{\tau^r_{mk}}{\alpha_{mk}\tau^r_{mk}+1}.\end{aligned} \tag{23}$$

Combining (22) and (23), taking the first-order derivative of $\mathrm{E}[\ln p(\mathbf{x}\,;\boldsymbol{\alpha})p(\boldsymbol{\alpha})]$ over $\boldsymbol{\alpha}$ and setting it to zero, we obtain the updated $\boldsymbol{\alpha}$,

$$\alpha^{(t)}_{mk} = \frac{2a-1}{2b+\mathrm{E}[(\hat{x}_{mk})^2]}. \tag{24}$$

For the Q-function containing only the β term, we have

$$\begin{aligned}\mathrm{E}[\ln p(\mathbf{y} \mid \mathbf{x}\,;\beta)p(\beta)] = \frac{NM}{2}\ln\beta + (c-1)\ln\beta \\ - \frac{\beta}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{E}[(y_{mn}-u_{mn})^2] - d\beta + \text{const.}\end{aligned} \tag{25}$$

Since $p(u_{mn} \mid \mathbf{y}, \hat{p}_{mn}, \tau^p_{mn}, \boldsymbol{\Theta})$ approximates the Gaussian distribution with posterior mean μ^u_{mn} and variance ϕ^u_{mn}, $\mathrm{E}[(y_{mn}-u_{mn})^2]$ can be expressed as

$$\begin{aligned}\mathrm{E}[(y_{mn}-u_{mn})^2] &= (y_{mn}-\mu^u_{mn})^2 + \phi^u_{mn} \\ &= (y_{mn} - \frac{\tau^p_{mn}\beta y_{mn} + \hat{p}_{mn}}{\beta\tau^p_{mn}+1})^2 + \frac{\tau^p_{mn}}{\beta\tau^p_{mn}+1}.\end{aligned} \tag{26}$$

Combining (25) and (26), taking the first-order derivative of $\mathrm{E}[\ln p(\mathbf{y} \mid \mathbf{x}\,;\beta)p(\beta)]$ over β and setting it to zero, we obtain the updated β,

$$\beta^{(t)} = \frac{NM+2c-2}{2d+\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{E}[(y_{mn}-u_{mn})^2]}. \tag{27}$$

The proposed SE-GAMP-SBL approach dedicated to MUD with multiantenna reception is summarized in Algorithm 2.

Remark 1: The proposed algorithms use diversity combining technique to improve the posterior distribution without modifying the prior distribution, which amounts to integrating the spatial structure of user activity from a novel perspective.

Remark 2: The proposed algorithms do not require knowledge of user activity factor and noise power.

V. Simulation Results

In this section, the performance of the proposed SE-SBL and SE-GAMP-SBL is analyzed and compared with the classical LS, MMSE, RD [6], LD [6], SBL [3], GAMP-SBL [15], B-OMP [10], B-CoSaMP [11], B-SP [12], and B-SBL [10] algorithms. Oracle-LS is introduced as a benchmark [4], where the actual active users are known at the BS. Similar to the existing literatures [3], [4], [6], the adaptive gain control scheme is adopted by users to compensate the path loss, i.e., $\boldsymbol{h}^m_k \sim \mathcal{CN}(0\,, \mathbf{I})$. The number of potential users is $K = 200$, the user activity factor is $p_a = 0.1$, and the length of the spreading sequence is $N = 120$. Binary phase shift keying is considered in this simulation. The parameter values for the Gamma distribution are $a = 1.5$ and $b = c = d = 10^{-8}$.

Fig. 2 illustrates the effect of receiving antenna number on the SER performance of the proposed algorithms. For comparison, the case of single antenna $M = 1$ is also included. We can observe that increasing the number of antennas M brings significant performance improvement. Moreover, SE-SBL and SE-GAMP-SBL perform approximately the same, indicating that although these two algorithms employ different strategies to compute the posterior distribution, they both approach the true posterior distribution.

Fig. 3 shows the SER performance comparison of the considered algorithms. The proposed SE-SBL and SE-GAMP-SBL significantly outperform the block greedy-type CS algorithms. Moreover, in high SNR regime, the proposed algorithms exhibit a lower SER than the B-SBL and Oracle-

Algorithm 2 SE-GAMP-SBL

Initialization: $\forall m,k$: $\alpha^{(0)}_{mk} = 1$, $\beta^{(0)} = 1$, $\hat{\mu}^{x\,(0)}_{mk} = 1$, $\phi^{x\,(0)}_{mk} = 1$, $\hat{s}^{(0)}_{mn} = 0$, and $t = 1$.

Repeat:

Step 1. $\forall m \in \{1,2,\cdots,M\}$ and $\forall n \in \{1,2,\cdots,N\}$:

$$u^{(t)}_{mn} = \sum_{m=1}^{M}\sum_{k=1}^{K}\mathrm{H}_{mn\,,mk}\hat{\mu}^{x\,(t-1)}_{mk}$$

$$\tau^{p\,(t)}_{mn} = \sum_{m=1}^{M}\sum_{k=1}^{K}\mathrm{H}^2_{mn,mk}\phi^{x\,(t-1)}_{mk}$$

$$\hat{p}^{(t)}_{mn} = u^{(t)}_{mn} - \tau^{p\,(t)}_{mn}\hat{s}^{(t-1)}_{mn}$$

Step 2. $\forall m \in \{1,2,\cdots,M\}$ and $\forall n \in \{1,2,\cdots,N\}$:

$$\hat{s}^{(t)}_{mn} = \frac{1}{\tau^{p\,(t)}_{mn}}\left(\frac{\tau^{p\,(t)}_{mn}\beta^{(t-1)}y_{mn}+\hat{p}^{(t)}_{mn}}{\beta^{(t-1)}\tau^{p\,(t)}_{mn}+1} - \hat{p}^{(t)}_{mn}\right)$$

$$\tau^{s\,(t)}_{mn} = \frac{\beta^{(t-1)}}{\beta^{(t-1)}\tau^{p\,(t)}_{mn}+1}$$

Step 3. $\forall m \in \{1,2,\cdots,M\}$ and $\forall k \in \{1,2,\cdots,K\}$:

$$\tau^{r\,(t)}_{mk} = \left(\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{H}^2_{mn\,,mk}\tau^{s\,(t)}_{mn}\right)^{-1}$$

$$\hat{r}^{(t)}_{mk} = \hat{\mu}^{x\,(t-1)}_{mk} + \tau^{r\,(t)}_{mk}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{H}_{mn\,,mk}\hat{s}^{(t)}_{mn}$$

Step 4. $\forall m \in \{1,2,\cdots,M\}$ and $\forall k \in \{1,2,\cdots,K\}$:

$$\mu^{x\,(t)}_{mk} = \frac{\hat{r}^{(t)}_{mk}}{\alpha^{(t-1)}_{mk}\tau^{r\,(t)}_{mk}+1}$$

$$\phi^{x\,(t)}_{mk} = \frac{\tau^{r\,(t)}_{mk}}{\alpha^{(t-1)}_{mk}\tau^{r\,(t)}_{mk}+1}$$

Step 5. $\forall m \in \{1,2,\cdots,M\}$ and $\forall k \in \{1,2,\cdots,K\}$:

$$\hat{z}^{(t)}_{k} = \sum_{m=1}^{M}\eta^{(t)}_{m}\mu^{x\,(t)}_{mk}$$

$$\hat{x}^{(t)}_{mk} = \hat{\mu}^{x\,(t)}_{mk} = \hat{z}^{(t)}_{k}$$

Step 6. Updating of $\boldsymbol{\alpha}$:

$$\alpha^{(t)}_{mk} = \frac{2a-1}{2b+\mathrm{E}[(\hat{x}^{(t)}_{mk})^2]}$$

Step 7. Updating of β:

$$\beta^{(t)} = \frac{NM+2c-2}{2d+\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{E}[(y_{mn}-u^{(t)}_{mn})^2]}$$

Repeating the above steps until $\sqrt{\sum_{m=1}^{M}\sum_{k=1}^{K}\left|\hat{\mu}^{x\,(t)}_{mk} - \hat{\mu}^{x\,(t-1)}_{mk}\right|^2} \le \varepsilon$, where ε is a prescribed threshold value of 10^{-6}.

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2023.3317374

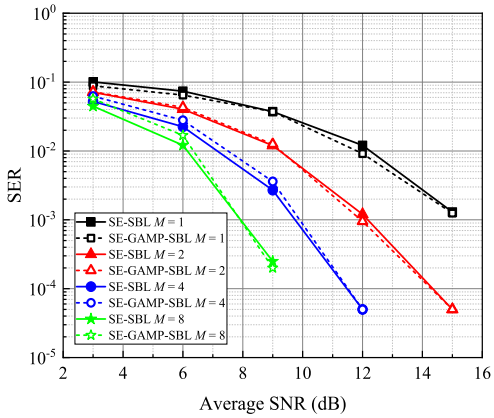

Fig. 2. SER performance of the proposed algorithms using different number of receiving antennas.

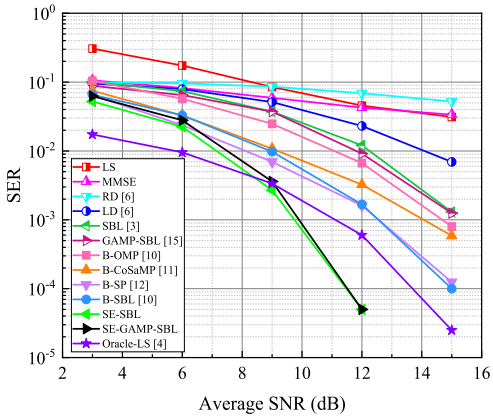

Fig. 3. SER performance comparison of the considered algorithms under receiving antenna number of $M = 4$.

TABLE I
COMPLEXITY COMPARISON

Algorithm	Number of multiplications in each iteration	Runtime (ms)
LS	$NKM^2+2NK^2M^3+K^3M^3$	226.539
MMSE	$KM+NKM^2+2NK^2M^3+K^3M^3$	237.013
RD	$4KM+NKM^2+2NK^2M^3+K^3M^3$	242.793
LD	polynomial time	3238.326
SBL	$NM+2KM+2NKM^2+4NK^2M^3+K^3M^3$	288.708
GAMP-SBL	$9NM+8KM+6NKM^2$	10.818
B-OMP	$K+2KM+N(2K+t)M^2+2Nt^2M^3+t^3M^3$	3.476
B-CoSaMP	$3K+4KM+N(2K+3K_a)M^2+10NK_a^2M^3+9K_a^3M^3$	18.406
B-SP	$3K+4KM+2N(K+K_a)M^2+4NK_a^2M^3+2K_a^3M^3$	14.061
B-SBL	$NM+KM+(2N+1)KM^2+4NK^2M^3+K^3M^3$	295.056
SE-SBL	$(N+1)M+4KM+2NKM^2+4NK^2M^3+K^3M^3$	292.131
SE-GAMP-SBL	$(9N+1)M+10KM+6NKM^2$	12.025
Oracle-LS	$KM+NK_aM^2+2NK_a^2M^3+K_a^3M^3$	5.036

LS counterparts. This implies that improving the posterior distribution brings more benefits than modifying the prior distribution and knowing the actual active users. Particularly, SBL and GAMP-SBL can be regarded as the special case of the proposed algorithms by setting $M = 1$ as shown in Fig. 2 and Fig. 3.

Finally, we compare the computational complexity of the considered algorithms. Table I presents the computational cost and running time required to execute single iteration for $M = 4$. LD is a polynomial time algorithm that can be readily accomplished using the available CVX toolbox. The computational cost of the proposed SE-SBL is higher than the block greedy-type CS algorithms due to $K_a \ll K$. However, the proposed SE-GAMP-SBL has the ability to obtain superiority in computational complexity mitigation. Moreover, the computational costs of SE-SBL and SE-GAMP-SBL are in the same order as those of SBL and GAMP-SBL respectively, because the complexity of Bayesian algorithms mainly stems from computing the posterior distribution.

VI. CONCLUSION

In this letter, we proposed two spatially enhanced Bayesian CS algorithms for efficient MUD in uplink grant-free MIMO-NOMA system. The proposed SE-SBL and SE-GAMP-SBL algorithms exploit the full potential of shared sparsity induced by the spatial structure of user activity, and do not require prior information about user sparsity level and noise power. Simulation results show that the proposed algorithms outperform off-the-shelf methods both in detection performance and computational complexity.

REFERENCES

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. A. Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615-636, Mar. 2021.

[2] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, 2020.

[3] X. Zhang, P. Fan, J. Liu, and L. Hao, "Bayesian learning-based multiuser detection for grant-free NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6317-6328, Aug. 2022.

[4] L. Wu, Z. Wang, P. Sun, and Y. Yang, "Temporal correlation enhanced sparse activity detection in MIMO enabled grant-free NOMA," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2887-2899, Mar. 2022.

[5] T. Hara and K. Ishibashi, "Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design," *IEEE Access.*, vol. 7, pp. 175717-175726, 2019.

[6] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454-465, Feb. 2011.

[7] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933-2946, Jun. 2018.

[8] S. Zhang, Y. Cui, and W. Chen, "Joint device activity detection, channel estimation and signal detection for massive grant-free access via BiGAMP," *IEEE Trans. Signal Process.*, vol. 71, pp. 1200-1215, 2023.

[9] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569-3582, Jul, 2019.

[10] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631-9640, Oct. 2018.

[11] X. Zhang, W. Xu, Y. Cui, L. Lu, and J. Lin, "On recovery of block sparse signals via block compressive sampling matching pursuit," *IEEE Access.*, vol. 7, pp. 175554-175563, 2019.

[12] Y. Du *et al.*, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 682-685, Aug. 2018.

[13] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131-146, Nov. 2008.

[14] F. Clazzer, C. Kissling, and M. Marchese, "Enhancing contention resolution ALOHA using combining techniques," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2576-2587, Jun. 2018.

[15] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "A GAMP-based low complexity sparse Bayesian learning algorithm," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 294-308, Jan. 2018.