Unified Iterative Receiver Design in Uplink Grant-free Massive MIMO SCMA Systems

Yining Li*†, Wenjin Wang*†, Xiaohang Song‡, Xiqi Gao*†, Lei Wang§, Gerhard P. Fettweis‡
*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
†Purple Mountain Laboratories, Nanjing, China, Email: {ynli, wangwj, xqgao}@seu.edu.cn.
‡ Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany,
Email: {xiaohang.song, gerhard.fettweis}@tu-dresden.de
§Huawei Technologies, Co. Ltd., Shanghai, China, Email: wanglei888@huawei.com

***Abstract*—In machine-type communication scenarios, sparse code multiple access (SCMA) is a promising non-orthogonal multiple access (NOMA) scheme owing to shaping gain by combining constellation modulation and spreading patterns together. In this paper, to fully exploit the channel knowledge contained in received data sequences, we propose the joint active user detection (AUD), channel estimation (CE), multi-user detection (MUD), and decoding receiver without knowing users' activity parameters in uplink grant-free massive multiple-input multiple-output (MIMO) SCMA systems. To avoid the permutation and scaling ambiguities of estimation results, the proposed receiver estimates channel based on both received short pilot and data sequences. We introduce auxiliary active state indicators in AUD to describe the sporadic transmission feature. The joint CE and MUD module is constructed as a bilinear inference problem with joint column-wise sparsity. Furthermore, we exploit the SCMA codewords sparsity feature and put Gaussian approximations on modulated symbols in joint CE and MUD module to reduce the receiver complexity. Simulation results show that the proposed unified receiver has substantial performance improvement and lower computational complexity than the conventional two-stage receiver and joint receiver in the literature.**

***Index Terms*—massive MIMO, SCMA, active user detection, channel estimation, bilinear generalized linear problem.**

I. Introduction

Due to the requirements of machine-type communications (MTC), the future wireless network is expected to support the massive connectivity of sensors and machines. However, small data packets and sporadic transmissions characterized in mmTC scenario may cause significant signaling overhead. A potential solution is the grant-free access, where users can send data packets directly without conducting the base station (BS) for grant messages [1]. To accommodate grant-free transmission, various non-orthogonal multiple access (NOMA) schemes have been proposed, in which multiple users access the same time-frequency resources with user-specific codewords. Compared to low-density spreading aided orthogonal frequency division multiplexing (LDS-OFDM) where modulated symbol sequences spread via given signatures, sparse code multiple access (SCMA) obtains constellation shaping gain by mapping coded bit sequence to optimized multi-dimensional codewords directly [2].

The conventional grant-free receiver can be divided into two stages [3]: channel and active states are estimated firstly, then data sequences of active users are detected. Approximate message passing (AMP) was applied in joint active user detection (AUD) and channel estimation (CE) with soft-thresholding denoiser [4] or minimum mean square error (MMSE) denoiser [5]. The multi-user detection (MUD) based on the message passing algorithm (MPA) was proposed by exploiting the sparse structure of SCMA codebooks [6]. To further decrease the complexity, expectation propagation (EP) based multi-user detection for multiple-input multiple-output (MIMO) SCMA system was given in [7]. Except for the two-stage scheme, the joint AUD, CE, and MUD receiver was proposed recently. In [8], the joint estimation and detection problem was transformed into block sparsity single measurement vector (SMV) problem, and solved by the block sparsity adaptive subspace pursuit (BSASP) algorithm. A joint receiver based on belief propagation Gaussian approximation expectation propagation (BP-GA-EP) was introduced in the uplink grant-free SCMA system [9].

Combining massive MIMO with a grant-free SCMA system potentially improve the spectrum efficiency [10]. Multiple-antenna channel estimation can be regarded as multiple measurement vector (MMV) problem and solved by parallel AMP-MMV [3], [11]. Exploiting users' sporadic traffic and channel structural sparsity in the virtual angular domain, Turbo-GMMV-AMP has a performance improvement by performing AUD at the spatial domain and CE at the angular domain [12]. Assuming the data packet carries the user's identity, and zeros are randomly and independently placed in transmitted packets, the blind MUD in time-slotted MIMO NOMA system was modeled as a dictionary learning (DL) problem. It was solved by bilinear generalized approximate message passing (BiG-AMP) [13].

In this paper, we develop the joint AUD, CE, MUD, and decoding framework in massive MIMO grant-free SCMA systems. The joint receiver detects active users' identities by pilot and data aided CE. We construct the joint detection problem as bilinear inference with the joint column-wise sparsity structure, where BiG-AMP cannot be used directly. The auxiliary active indicators are employed to represent structural sparsity. Furthermore, we take advantage of the SCMA codewords sparsity and apply Gaussian approximation

(GA) to SCMA symbols to reduce complexity.

Notations: We use upper (lower) bold-face letters to denote matrices (column vectors). Superscripts ^{T} and * denote transpose and conjugate, respectively. We reserve $\mathrm{E}\left(x|p(x)\right)$ for expectation of $p(x)$, $\|\cdot\|_F$ for normalized Frobenius norm, $|\mathcal{A}|$ for the number of elements in set $\mathcal{A}$. $x \sim \mathcal{CN}\left(x;\mu,\tau\right)$ represents x follows the complex norm distribution with mean μ and variance τ. $[\mathbf{A}]_{:,n}$ denotes the nth column of matrix $\mathbf{A}$.

II. Grant-free SCMA Transmission Model

Consider the uplink grant-free SCMA transmission system consisting of N potential users, and the BS and users are equipped with M antennas and a single antenna, respectively. Assume the users share the resource block consisting of K subcarriers and T time slots in the coherence time. We consider the block fading and rich-scattering channel $\mathbf{g}_{nk} = \left[g_{1nk},\cdots,g_{Mnk}\right]^T$ with the coefficient g_{mnk} being modeled as a Rayleigh fading component $p_g(g_{mnk}) \propto \mathcal{CN}(g_{mnk};0,\sigma_h^2)$. Here, σ_h^2 is decided by large scale fading and assumed to be known at the BS. We assume σ_h^2 is equal to all users for simplicity. In the massive connectivity scenario, the number of potential active users is far larger than the antennas number, i.e., $N \gg M$. But only a fixed small subset of users, denoted as $\mathcal{N}_a$, is active within the coherence time. We introduce active state indicators $\boldsymbol{\xi} = [\xi_1,\cdots,\xi_N]$, where ξ_n is the binary random variable, i.e., $\xi_n = 1$ when $n \in \mathcal{N}_a$ and $\xi_n = 0$ when $n \notin \mathcal{N}_a$. Let $\mathbf{h}_{nk} = \xi_n \mathbf{g}_{nk}$ denote the effective channel, and $\mathbf{H}_k = [\mathbf{h}_{1k},\cdots,\mathbf{h}_{Nk}]$.

The nth user's channel encoded information sequence is denoted by $\mathbf{b}_n$. Define the modulation order as Q_b. We divide $\mathbf{b}_n$ into KJ subsequences of length Q_b and represent each subsequence as $\mathbf{b}_{nkj} = [b_{nkj1},\cdots,b_{nkjQ_b}]^T$. We get the jth SCMA codeword on the kth subcarrier of the nth user $\mathbf{d}_{nkj} = [d_{nkj1},\cdots,d_{nkjQ_d}]^T$ by mapping $\mathbf{b}_{nkj}$ via a user-specific predesigned codebook φ_n, i.e., $\mathbf{d}_{nkj} = \varphi_n(\mathbf{b}_{nkj})$. Note that SCMA has the sparsity feature that only $d_v \ll Q_d$ entries are nonzero in each codeword. The collection of nonzero entries' indices in nth user's codewords is represented as $\mathcal{V}_n$, which satisfies $|\mathcal{V}_n| = d_v$, $\forall n$. Due to the SCMA codebook sparsity, only a small fraction of users collapse in the q_dth symbol, and the set is denoted as $\mathcal{U}_{q_d}$. The transmitted data sequence of the nth user on the kth subcarrier is a T_d-dimension vector $\mathbf{d}_{nk} = [(\mathbf{d}_{nk1})^T,\cdots,(\mathbf{d}_{nkJ})^T]^T$. We use subscript $_{nkt_d}$ or $_{nkjq_d}$, where $t_d = jQ_d + q_d$ considering contexts in the following expressions for simplicity.

In a massive connectivity scenario where the number of users is much larger than pilot length, pilot sequences allocated to different users are non-orthogonal. We generate pilots using i.i.d. Gaussian distribution with unit variance. The nth user's pilot takes on K subcarriers and T_p OFDM symbols. Define the pilot matrix at the kth subcarrier as $\mathbf{P}_k \in \mathcal{C}^{T_p \times N}$. Each frame is composed of pilot and data sequences. The received signal can be written as

$$\mathbf{Y}_k = \mathbf{H}_k\left[\mathbf{P}_k^T, \mathbf{D}_k^T\right] + \mathbf{Z}_k, \tag{1}$$

where $\mathbf{D}_k = [\mathbf{d}_{1k},\cdots,\mathbf{d}_{Nk}]$, elements in $\mathbf{Z}_k \in \mathcal{C}^{M\times T}$ follow $\mathcal{CN}\left(z_{kmt};0,\sigma_z^2\right)$, $T = T_p + T_d$. Define $\mathbf{X}_p \triangleq \mathbf{H}_k\mathbf{P}_k^T$, $\mathbf{X}_d \triangleq \mathbf{H}_k\mathbf{D}_k^T$, $\mathbf{X}_k \triangleq [\mathbf{X}_{p,k},\mathbf{X}_{d,k}]$, and $\mathbf{Y}_k = [\mathbf{Y}_{p,k},\mathbf{Y}_{d,k}]$. Note that for $\forall n \notin \mathcal{N}_a$, both $[\mathbf{H}_k]_{:,n}$ and $[\mathbf{D}_k]_{:,n}$ are zero vectors, which we name joint column-wise sparsity.

Despite that the channel information can be inferred from the received data sequences, the pilot aided CE is still required in the receiver. Ambiguities exist in blind CE results. Assume $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{D}}_k$ are the MMSE estimations $\hat{\mathbf{H}}_k,\hat{\mathbf{D}}_k = \arg\min_{\mathbf{H},\mathbf{D}} \mathrm{E}(||\mathbf{Y}_{d,k} - \mathbf{H}\mathbf{D}^T||_F^2)$. We can find the pair $\{\hat{\mathbf{H}}_k\boldsymbol{\Pi}_k\boldsymbol{\Lambda}_k, (\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\Pi}_k^{-1}\hat{\mathbf{D}}_k^T)^T\}$ also minimizes the MSE, where $\boldsymbol{\Pi}$ and $\boldsymbol{\Lambda} = \mathrm{diag}\left[\alpha_1,\cdots,\alpha_N\right]$ are random permutation and scaling matrix, respectively. The channel estimation based on pilot sequences can help remove the scaling and permutation ambiguities of the inference results in (1). However, the pilot can be relatively short compared to that in the conventional two-stage receiver.

III. Iterative Receiver Design

A. Probability Model and Bayesian Estimation

We drop the index k in the following expressions for simplicity. The joint probability distribution function can be factorized as

$$\begin{aligned} p(\mathbf{X},\mathbf{H},\mathbf{G},\mathbf{D},\mathbf{B},\boldsymbol{\xi}|\mathbf{Y}) \propto P(\boldsymbol{\xi};\alpha)p(\mathbf{G})p(\mathbf{H}|\boldsymbol{\xi},\mathbf{G}) \\ p(\mathbf{X}_p|\mathbf{H})P(\mathbf{B})P(\mathbf{D}|\mathbf{B})p(\mathbf{X}_d|\mathbf{H},\mathbf{D})p(\mathbf{Y}|\mathbf{X}), \end{aligned} \tag{2}$$

where the active states of different users are independent, i.e., $P(\boldsymbol{\xi};\alpha) = \prod_{n=1}^N P(\xi_n;\alpha)$, and ξ_n follows Bernoulli distribution

$$P(\xi_n;\alpha) = \begin{cases} \alpha, & \xi_n = 1, \\ 1-\alpha, & \xi_n = 0. \end{cases}$$

$p(\mathbf{H}|\boldsymbol{\xi},\mathbf{G}) = \prod_{m=1}^M\prod_{n=1}^N \delta(h_{mn} - \xi_n g_{mn})$, $P(\mathbf{B}) = \prod_{n=1}^N\prod_{j=1}^J\prod_{q_b=1}^{Q_b} P(b_{njq_b})$, $\forall b_{njq_b} \in \mathcal{B}$, $P(\mathbf{D}|\mathbf{B}) = \prod_{n=1}^N\prod_{j=1}^J \delta(\mathbf{d}_{nj} - \varphi_n(\mathbf{b}_{nj}))$, $p(\mathbf{X}_p|\mathbf{H}) = \delta(\mathbf{X}_p - \mathbf{H}\mathbf{P}^T)$, $p(\mathbf{X}_d|\mathbf{H},\mathbf{D}) = \delta(\mathbf{X}_d - \mathbf{H}\mathbf{D}^T)$, $p(\mathbf{Y}|\mathbf{X}) = \prod_{m=1}^M\prod_{t=1}^T \mathcal{CN}(x_{mt};y_{mt},\sigma_z^2)$.

We aim to calculate posterior probability $P(\xi_n|\mathbf{Y})$, $P(b_{njq_b}|\mathbf{Y})$, and $p(h_{mn}|\mathbf{Y})$, $\forall m,n,j,q_b$ according to received pilot $\mathbf{Y}_p$ and data sequences $\mathbf{Y}_d$ without knowing hyper-parameter α. However, the marginal probability calculations involve multiple integrations over latent variables and are of high computational complexity, especially in large dimensions. Message passing is a low-complexity method to obtain maximum a posteriori (MAP) or MMSE estimations, in which messages are calculated based on specific principals along the factor graph [14]. The corresponding factor graph of (2) is shown in Fig. 1. Each variable node is connected to its dependent factor nodes. Denote factor nodes corresponding to $P(\mathbf{d}_{nj}|\mathbf{b}_{nj})$, $p(\mathbf{h}_n|\xi_n)$ and $P(\xi_n;\alpha)$ as $f_{b,nj}$, $f_{\xi,n}$ and ψ_n, respectively. For $\forall m,n,t_d,t_p$, denote the beliefs of ξ_n, h_{mn}, d_{nt_d}, b_{x_d,mt_d} and b_{x_p,mt_p} as $g_{\xi,n}$, $b_{h,mn}$, b_{d,nt_d}, b_{x_d,mt_d} and b_{x_p,mt_p}, respectively.

In loopy belief propagation (LBP), the message sent from the variable node to the factor node is the product of messages sent from other adjacent factor nodes, and the message sent

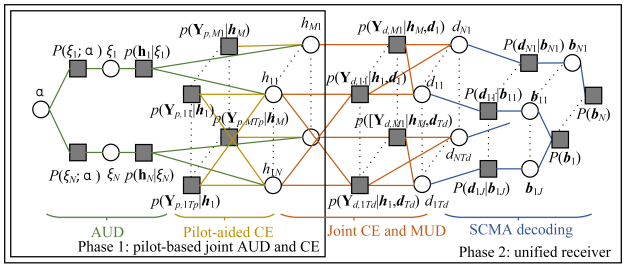

Fig. 1. The factor graph for uplink grant-free massive MIMO SCMA transmission. Squares and circles represent factor and variables nodes, respectively.

from the factor node to the variable node is the integration of the corresponding factor and messages sent from nearby variable nodes [14]. EP projects the posterior probability into Gaussian distribution. Thus, only the first- and second-moments are transferred through the factor graph[7], [9]. AMP simplifies BP in inference from linear transforms by exploiting central-limit-theorem (CLT) and Taylor expansion. Generalized AMP (GAMP) proposed in [15] incorporates arbitrary priors and output distributions. BiG-AMP is a direct extension of GAMP and is also the state-of-the-art algorithm to solve DL, where the dictionary $\mathbf{A}$ and sparse coefficient matrix $\mathbf{X}$ are estimated from observation $\mathbf{Y}$ with $p(y_{mt}|[\mathbf{A}]_{:,m}^{T}[\mathbf{X}]_{:,t})$ [16]. Although the generalized bilinear inference problem shown in (1) has a similar form with DL [17], it also has a specific structural sparsity. In the DL problem, only one of two inferred matrices is sparse, and the non-zero probability of elements are independent. Therefore, the estimation of $\hat{\mathbf{H}}$ and $\hat{\mathbf{D}}$ cannot be solved via BiG-AMP directly.

As shown in Fig. 1, we divide the receiver into four modules. The log-likelihood ratio (LLR) of active state $L_{\xi,n} = \ln(g_{\xi,n}(\xi_n = 1)/g_{\xi,n}(\xi_n = 0))$ is calculated based on channel estimation in AUD. The pilot-based CE can be regarded as an estimation problem with linear mixing. The joint CE and MUD module estimates both $\hat{\mathbf{H}}$ and $\hat{\mathbf{D}}$ based on $\mathbf{Y}_d$. The decoding module get the extrinsic LLR of coded bits based on $\hat{\mathbf{D}}$. According to the characteristics of modules, e.g., discrete or continuous, we adopt different message updating rules.

B. Message Passing Receiver

To avoid the algorithm trapping in solution with ambiguities, we firstly estimate channel coefficients $\hat{\mathbf{H}}$ and active states $\mathbf{L}_\xi$ from the pilot-based joint AUD and CE. Then $\hat{\mathbf{H}}$ is used as initialization in the unified receiver. Factor graphs of the preprocessing and unified receiver are shown in Fig. 1.

1) Pilot based joint AUD and CE: We apply AMP in pilot-based CE, which is a linear mixing estimation problem. Detailed procedures are shown in Alg. 1 line 4-10. The posterior mean r_{mn} and variance τ_{mn}^{r} can be considered as the message sent from hyper-factor node f_p to variable node h_{mn}, i.e., $I_{f_p\to h_{mn}}(h_{mn}) = \mathcal{CN}(h_{mn}; r_{mn}, v_{mn}^{r})$. Note that

$$b_{x_p,mt} \propto \mathcal{CN}\left(x_{p,mt}; o_{p,mt}, v_{p,mt}^{o}\right)\mathcal{CN}\left(x_{p,mt}; y_{p,mt}, \sigma_z^2\right). \tag{3}$$

Since active states are discrete variables taking values of 1 or 0, we use BP to update messages in AUD module. Let $I_{h_{mn}\to f_{\xi,n}}(h_{mn})$ denote the message sent from variable node h_{mn} to $f_{\xi,n}$. We can obtain $I_{h_{mn}\to f_{\xi,n}}(h_{mn}) = I_{f_p\to h_{mn}}(h_{mn})$ based on the BP rules. The message sent from factor node $f_{\xi,n}$ to variable node ξ_n can be calculated by

$$I_{f_{\xi,n}\to\xi_n} = \begin{cases} \prod_{m=1}^{M}\mathcal{CN}(\tilde{h}_{mn}; 0, v_{mn}^{\tilde{h}} + \sigma_h^2), & \xi_n = 1, \\ \prod_{m=1}^{M}\mathcal{CN}(\tilde{h}_{mn}; 0, v_{mn}^{\tilde{h}}), & \xi_n = 0, \end{cases} \tag{4}$$

where $\tilde{h}_{mn} = r_{mn}$ and $v_{mn}^{\tilde{h}} = v_{mn}^{r}$. $g_{\xi,n}$ is the product of $I_{f_{\xi,n}\to\xi_n}(\xi_n)$ and $I_{\psi_n\to\xi_n}(\xi_n)$, and $L_{\xi,n}$ can be written as

$$L_{\xi,n} = \ln\frac{\alpha}{1-\alpha} + \sum_{m=1}^{M}\underbrace{\left(\ln\frac{v_{mn}^{\tilde{h}}}{v_{mn}^{\tilde{h}}+\sigma_h^2} + \frac{\left|\tilde{h}_{mn}\right|^2\sigma_h^2}{\left(v_{mn}^{\tilde{h}}+\sigma_h^2\right)v_{mn}^{\tilde{h}}}\right)}_{\triangleq L_{\xi',mn}}. \tag{5}$$

Denote the message sent from variable node ξ_n to factor node $f_{\xi,n}$ as $I_{\xi_n\to f_{\xi,n}}(\xi_n)$. Based on BP rule, we have $I_{\xi_n\to f_{\xi,n}}(\xi_n) = I_{\psi_n\to\xi_n}(\xi_n)$. Through multiplying messages sent from all adjacent nodes to $f_{\xi,n}$ except for $I_{h_{mn}\to f_{\xi,n}}(h_{mn})$ together and integrating the product over variables except for h_{mn}, $I_{f_{\xi,n}\to h_{mn}}(h_{mn})$ is given by

$$I_{f_{\xi,n}\to h_{mn}} = (1-\rho_{mn})\,\delta(h_{mn}) + \rho_{mn}\mathcal{CN}\left(0, \sigma_h^2\right), \tag{6}$$

where $\rho_{mn} = \exp(L_{\rho,mn})/1+\exp(L_{\rho,mn})$, $L_{\rho,mn} = L_{\xi,n} - L_{\xi',mn}$. $b_{h,mn}$ is the product of $I_{f_p\to h_{mn}}$ and $I_{f_{\xi,n}\to h_{mn}}$. Dropping subscript $_{mn}$ for simplicity, we have

$$b_h \propto (1-\lambda)\,\delta(h) + \lambda\mathcal{CN}(h; \mu_h, v_h), \tag{7}$$

where $\lambda = \exp(L)/(1+\exp(L))$, $\mu_h = \tilde{h}\sigma_h^2/(v^{\tilde{h}}+\sigma_h^2)$, $v_h = v^{\tilde{h}}\sigma_h^2/(v^{\tilde{h}}+\sigma_h^2)$, $L = \ln\frac{\rho v^{\tilde{h}}}{(1-\rho)\left(v^{\tilde{h}}+\sigma_h^2\right)} + \frac{|\tilde{h}|^2\sigma_h^2}{v^{\tilde{h}}\left(v^{\tilde{h}}+\sigma_h^2\right)}$.

Moreover, hyper-parameter α in (5) can be learnt by EM as

$$\alpha = \arg\max_{\alpha}\sum_{n=1}^{N}\mathrm{E}\left(\ln p(\xi;\alpha)\,|g_{\xi,n}(\xi_n)\right). \tag{8}$$

Setting the derivative of $\sum_{n=1}^{N}\mathrm{E}\left(\ln p(\xi;\alpha)\,|g_{\xi,n}(\xi_n)\right)$ with respect to α as zero, we have

$$\alpha = \frac{1}{N}\sum_{n=1}^{N} g_{\xi,n}(\xi_n = 1). \tag{9}$$

2) Unified receiver: The joint CE and MUD can be regarded as a bilinear inference problem with joint column-wise sparsity. We apply BiG-AMP [16] to calculate messages sent from observation $\mathbf{Y}_d$ to $\mathbf{H}$ and $\mathbf{D}$. Let $\hat{h}_{mn}$ and $\tau_{h,mn}$ denote mean and variance with respect to belief $b_{h,mn}(h_{mn})$, respectively. Let $\hat{d}_{nt_d}$ and τ_{d,nt_d} denote mean and variance of $b_{d,nt_d}(d_{nt_d})$, respectively. Given $\{\hat{h}_{mn}, \tau_{h,mn}, \forall m,n\}$ and $\{\hat{d}_{nt_d}, \tau_{d,nt_d}, \forall n,t_d\}$ from the last iteration, the mean and variance estimation $\{o_{d,mt_d}, v_{d,mt_d}^{o}, \forall m,t_d\}$ of $\mathbf{X}_d$ are calculated as Alg. 1, line 20-21. Similar to (3), we have

$$b_{x_d} \propto \mathcal{CN}\left(x_d; o_d, v_d^{o}\right)\mathcal{CN}\left(x_d; y_d, \sigma_z^2\right), \tag{10}$$

Algorithm 1 EM-BP-GA-Bi-AMP

Input: Pilot $\mathbf{P}$, observation $\mathbf{Y}^p$ and $\mathbf{Y}^d$, σ_h^2, σ_z^2.
Output: Extrinsic LLR of coded bits $\mathbf{L}_e$, $\mathbf{L}_\xi$.

1: ***Initialization***: $\hat{\mathbf{H}}=\mathbf{0}$, $\boldsymbol{\tau}^h=\mathbf{1}$, $\hat{\mathbf{D}}=\mathbf{0}$, $\boldsymbol{\tau}^d=\mathbf{1}$, $\mathbf{S}=\mathbf{0}$.
2: ***Pilot-based AUD and CE***:
3: **repeat**
4: $\forall m,t_p$, $v^o_{p,mt_p}=\sum_{n=1}^{N}|p_{nt_p}|^2\tau^h_{mn}$.
5: $\forall m,t_p$, $o_{p,mt_p}=\sum_{n=1}^{N}p_{nt_p}\hat{h}_{mn}-s_{p,mt_p}v^o_{p,mt_p}$.
6: $\forall m,t_p$, update $\hat{x}_{p,mt_p}$ and τ^x_{p,mt_p} of b_{x_p,mt_p} via (3).
7: $\forall m,t_p$, $\tau^s_{p,mt_p}=(1-\tau^x_{p,mt_p}/v^o_{p,mt_p})/v^o_{p,mt_p}$.
8: $\forall m,t_p$, $s_{p,mt_p}=(\hat{x}_{p,mt_p}-o_{p,mt_p})/v^o_{p,mt_p}$.
9: $\forall m,n$, $v^r_{mn}=(\sum_{t_p=1}^{T_p}|p_{nt_p}|^2\tau^s_{p,mt_p})^{-1}$.
10: $\forall m,n$, $r_{mn}=\hat{h}_{mn}+v^r_{mn}\sum_{t_p=1}^{T_p}p^*_{nt_p}s_{p,mt_p}$.
11: $\forall m,n$, $\tilde{h}_{mn}=r_{mn}$, $v^{\tilde{h}}_{mn}=v^r_{mn}$.
12: $\forall n$, update $L_{\xi,n}$ and α via (5) and (9).
13: $\forall m,n$, update $\hat{h}_{mn}$ and τ^h_{mn} of $b_{h,mn}$ via (7).
14: **until** the iteration number meets T_1
15: ***Unified Receiver:***
16: **repeat**
17: **repeat**
18: $\forall m,t$, $\bar{\tau}_{x_d,mt_d}=\sum_{n\in\mathcal{U}_{t_d}}(|\hat{h}_{mn}|^2\tau^d_{nt_d}+|\hat{d}_{nt_d}|^2\tau^h_{mn})$.
19: $\forall m,t$, $\bar{\mu}_{x_d,mt_d}=\sum_{n\in\mathcal{U}_{t_d}}\hat{h}_{mn}\hat{d}_{nt_d}$.
20: $\forall m,t$, $v^o_{d,mt_d}=\bar{\tau}_{x_d,mt_d}+\sum_{n\in\mathcal{U}_{t_d}}\tau^h_{mn}\tau^d_{nt_d}$
21: $\forall m,t$, $o_{d,mt_d}=\bar{\mu}_{x_d,mt_d}-s_{d,mt_d}\bar{\tau}_{x_d,mt_d}$.
22: $\forall m,t$, update $\hat{x}_{d,mt_d}$ and τ^x_{d,mt_d} via (10).
23: $\forall m,t$, $\tau^s_{d,mt_d}=(1-\tau^x_{d,mt_d}/v^o_{d,mt_d})/v^o_{d,mt_d}$.
24: $\forall m,t$, $s_{d,mt_d}=(\hat{x}_{d,mt_d}-o_{d,mt_d})/v^o_{d,mt_d}$.
25: $\forall m,n$, $v^q_{mn}=(\sum_{t\in\mathcal{V}_n}|\hat{d}_{nt_d}|^2\tau^s_{d,mt_d})^{-1}$.
26: $\forall m,n$, $q_{mn}=\hat{h}_{mn}\tilde{q}_{mn}+v^q_{mn}\sum_{t\in\mathcal{V}_n}\hat{d}^*_{nt_d}s_{d,mt_d}$
27: $\forall n,t\in\mathcal{V}_n$, $v^u_{nt_d}=(\sum_{m=1}^{M}|\hat{h}_{mn}|^2\tau^s_{d,mt_d})^{-1}$
28: $\forall n,t\in\mathcal{V}_n$, $u_{nt_d}=\hat{d}_{nt_d}\tilde{u}_{nt_d}+v^u_{nt_d}\sum_{m}\hat{h}^*_{mn}s_{d,mt_d}$
29: $\forall m,n$, update r_{mn} and v^r_{mn} via line 4-10.
30: $\forall m,n$, update $\tilde{h}_{mn}$ and $v^{\tilde{h}}_{mn}$ via (11).
31: $\forall n$, update $L_{\xi,n}$ and α via (5) and (9).
32: $\forall m,n$, update $\hat{h}_{mn}$ and τ^h_{mn} of $b_{h,mn}$ via (7).
33: $\forall n,l_c$, update L_{e,nl_c} and L_{a,nl_c}.
34: $\forall n,t$, update $\hat{d}_{nt_d}$ and $\tau^d_{nt_d}$ via (15) and (16).
35: **until** the inner iteration number meets T_1
36: Decide $\hat{\boldsymbol{\xi}}$ via (12) and reset $\boldsymbol{\tau}^h$ and $\boldsymbol{\tau}^d$.
37: **until** the outer iteration number meets T_2

where subscript $_{mt}$ is dropped for simplification. The residual s_{d,mt_d} and the inverse of residual variance τ^s_{d,mt_d} are calculated in line 23-24. Based on the received data sequences, the mean and variance pair $\{q_{mn},v^q_{mn}\}$ of h_{mn} is calculated in line 25-26, where $\tilde{q}_{mn}=1-v^q_{mn}\sum_{t\in\mathcal{V}_n}\tau^d_{nt}\tau^s_{d,mt}$ in line 26. Denote message sent from hyper-factor node f_d to h_{mn} and d_{nt_d} as $I_{f_d\to h_{mn}}$ and $I_{f_d\to d_{nt_d}}$. $\{q_{mn},v^q_{mn}\}$ can be viewed as the mean and variance of $I_{f_d\to h_{mn}}$, i.e., $I_{f_d\to h_{mn}}=\mathcal{CN}(q_{mn},v^q_{mn})$. Similarly, the mean and variance of d_{nt_d} are given in line 27-28, where $\tilde{u}_{nt}=1-v^u_{nt}\sum_m\tau^h_{mn}\tau^s_{d,mt}$. We also have $I_{f_d\to d_{nt_d}}=\mathcal{CN}\left(d_{nt_d};u_{nt_d},v^u_{nt_d}\right)$.

Accumulating channel information from joint CE and AUD module and that from pilot-based CE, the mean and variance of $\mathbf{H}$ estimation sent to AUD are given by

$$\tilde{h}_{mn}=\frac{r_{mn}v^q_{mn}+q_{mn}v^r_{mn}}{v^q_{mn}+v^r_{mn}},v^{\tilde{h}}_{mn}=\frac{v^q_{mn}v^r_{mn}}{v^q_{mn}+v^r_{mn}}. \quad (11)$$

Active states LLR $L_{\xi,n}$ and belief $b_{h,mn}$ have the same form as (5) and (7), $\forall m,n$, respectively.

The conventional active user detector decides the channel coefficient as non-zero if it is more significant than a preset threshold. Then the nth user is declared to be active if the count of non-zero elements is larger than a certain number [12]. However, the proposed active user detector decides the nth user's activity utilizing both mean and variance of channel coefficients. The activity dector is written as

$$\hat{\xi}_n=\begin{cases}1, & L_{\xi,n}>\epsilon,\\ 0, & L_{\xi,n}\le\epsilon,\end{cases} \quad (12)$$

where ϵ is the threshold. In high SNR regime, for active users, channel estimation with extremely low variance $\mathbf{v}^{\tilde{h}}_n$ can be obtained with joint efforts from pilot-based and data-aided CE. Thus, $L_{\xi,n}$ of the user detected as active is high. For inactive users, data detections $\hat{\mathbf{d}}_n$ are near to zero. According to line 9 and 25 in Alg. 1, $\mathbf{v}^q_n$, the variances of channel coefficients sent from f_d, are far larger than $\mathbf{v}^r_n$. Thus, according to (11), $\mathbf{v}^{\tilde{h}}_n$ is close to $\mathbf{v}^r_n$, and $\tilde{\mathbf{h}}_n$ is close to $\mathbf{r}_n$. The magnitude of $L_{\xi,n}$ is much lower than that of active users. We set a high threshold ϵ_h to make the tradeoff between false alarm rate and miss detection rate. In low SNR regime, L_ξ of active users are low, and we should set a smaller threshold ϵ_l. Thus, considering both high and low SNR scenarios, we have $\epsilon=\min(\epsilon_h,\epsilon_l)$.

Using knowledge of SCMA codebooks' sparsity feature, the message from function node $f_{b,nj}$ to $\mathbf{b}_{nj}$ can be written as

$$I_{f_{b,nj}\to\mathbf{b}_{nj}}\propto\prod_{q_d\in\mathcal{V}_n}\mathcal{CN}\left([\varphi_n(\mathbf{b}_{nj})]_{q_d};u_{njq_d},v^u_{njq_d}\right). \quad (13)$$

Denote $L_{e,njq_b}=\ln\frac{I_{f_b\to b_{njq_b}}(b_{njq_b}=1)}{I_{f_b\to b_{njq_b}}(b_{njq_b}=0)}$ as extrinsic LLR of coded bit b_{njq_b} that sent from detector to decoder, where $I_{f_d\to b_{njq_b}}(b_{njq_b}=1)\propto\sum_{b_{njq_b}=1}I_{f_b\to\mathbf{b}_{nj}}$. Posterior LLR of coded bit L_{p,njq_b} is obtained through decoding iterations. According to BP rules, $L_{a,njq_b}=L_{p,njq_b}-L_{e,njq_b}$ is the prior LLR of coded bit for detector.

Similarly, $\forall q_d\in\mathcal{V}_n$, the message sent from factor node $f_{b,nj}$ to variable node d_{njq_d} is given by

$$I_{f_{b,nj}\to d_{njq_d}}=\sum_{q'_d\neq q_d}\sum_{\mathbf{b}_{nj}}P(\mathbf{d}_{nj}|\mathbf{b}_{nj})\prod_{q_b=1}^{Q_b}I_{b_{njq_b}\to f_{b,nj}}, \quad (14)$$

where $I_{b_{njq_b}\to f_{b,nj}}(b_{njq_b}=1)=1-1/(1+\exp(L_{a,njq_b}))$. Note that $I_{f_{b,nj}\to d_{njq_d}}$ is a probability mass function that takes on SCMA constellations set $\mathcal{C}$. The calculation of belief $g_{d,njq_d}(d_{njq_d})\propto I_{f_{b,nj}\to d_{njq_d}}I_{f_d\to d_{njq_d}}$ requires sampling $I_{f_d\to d_{njq_d}}$ over all possible codewords and calculating the

posterior probability of each codeword. To simplify the procedure, we apply GA on $I_{f_{b,nj}\to d_{njq_d}}$ as $\tilde{I}_{f_{b,nj}\to d_{njq_d}} = \mathcal{CN}(d_{njq_d}; w_{njq_d}, v^w_{njq_d})$, where

$$\begin{aligned} w_{njq_d} &= \mathrm{E}(d_{njq_d}|I_{f_{b,nj}\to d_{njq_d}}), \\ v^w_{njq_d} &= \mathrm{E}(|d_{njq_d}|^2|I_{f_{b,nj}\to d_{njq_d}}) - w^2_{njq_d} \end{aligned} \tag{15}$$

Dropping $_{njq_d}$ for expression simplicity, the posterior mean and variance of d_{njq_d} are shown as

$$\hat{d} = \frac{wv^u + uv^w}{v^u + v^w}, \tau^d = \frac{v^u v^w}{v^u + v^w}. \tag{16}$$

Damped scheme of Alg. 1 is applied to avoid non-convergent situations. We adopt damping in $\bar{\tau}_{x_d,mt_d}$, v^o_{d,mt_d}, v^o_{p,mt_p}, τ^s_{d,mt_d}, s_{d,mt_d}, τ^s_{p,mt_p}, s_{p,mt_p}, $\forall m, n, t_d, t_p$. Moreover, $\hat{h}_{mn}$ and $\hat{d}_{nt}$ in line 25-28 are replaced by damped version $\underline{\hat{h}}_{mn} = \eta\hat{h}_{mn} + (1-\eta)\underline{\hat{h}}_{mn}$ and $\underline{\hat{d}}_{nt} = \eta\hat{d}_{nt} + (1-\eta)\underline{\hat{d}}_{nt}$.

During the message passing iterations, the estimated channel coefficients and data sequences of miss detected users approach zero erroneously. With high variance $\mathbf{v}^{\tilde{h}}_n$ sent from $\mathbf{Y}_d$, the channel estimation $\hat{\mathbf{h}}_n$ cannot help decide users' active state. To avoid trapping in solutions with miss detections, we consider reinitializing the posterior variance of $\hat{\mathbf{H}}$ and $\hat{\mathbf{D}}$ after converging. Keeping $\hat{\mathbf{H}}$ and $\hat{\mathbf{D}}$ unchanged, the variance of detected inactive users' channel coefficients and data sequences are set as maximum value. From the perspective of simulated annealing (SA), resetting variance is analogous to choosing a relative high temperature in SA, which avoids algorithm permatually trapped in local optimum [18]. The reintialization step is shown in Alg. 1 line 36.

C. Complexity Analysis

In BP-GA-EP based receiver, SCMA symbols follow the probability mass function that takes on constellations in the joint CE and MUD module. The complexity of calculating posterior mean and variance for channel coefficients is of order $\mathcal{O}(MN(Jd_v 2^{Q_b} + T_p))$. In the proposed algorithm, we approximate the probability distribution of SCMA symbols as Gaussian distribution. Therefore, the computation complexity of joint CE and MUD is irrelevant to the modulation order. The multiplications number per iteration is on the order of $\mathcal{O}(MN(Jd_v+T_p)+NJd_v 2^{Q_b})$. Thus, the proposed algorithm has lower complexity compared with BP-GA-EP [9].

IV. Simulation Results

For the presented simulation results, we assume $M = 32$, $N = 156$, $K = 1$, $T_d = 216$, $\alpha = 0.1$. The variance of channel coefficients is $\sigma_h^2 = 1$. Information bits are encoded by a code rate $R = 1/2$ low-density parity-check (LDPC) code. The SCMA codebook maps $Q_b = 4$ bits into a codeword of length $Q_d = 8$, where only $d_v = 2$ elements are non-zero in each codeword. The SCMA codebooks can accommodate $C^{d_v}_{Q_d} = 28$ users, and up to $\lceil N/28 \rceil$ users access the same SCMA codebook. However, due to independent channels and pilots, active users can still be accurately detected. Signal to noise ratio (SNR) is defined as $\mathrm{SNR} = 10\log_{10}\frac{||\mathbf{X}||_F^2}{\alpha MT\sigma_z^2}$.

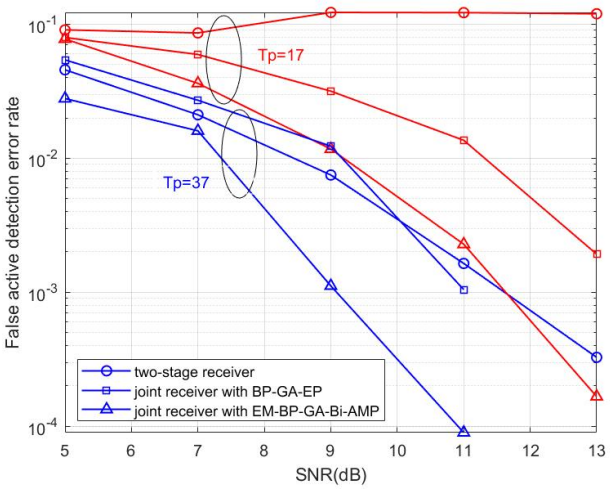

Fig. 2. Active detection error rate comparison between the joint receiver with EM-BP-GA-Bi-AMP, joint receiver with BP-GA-EP and two-stage receiver.

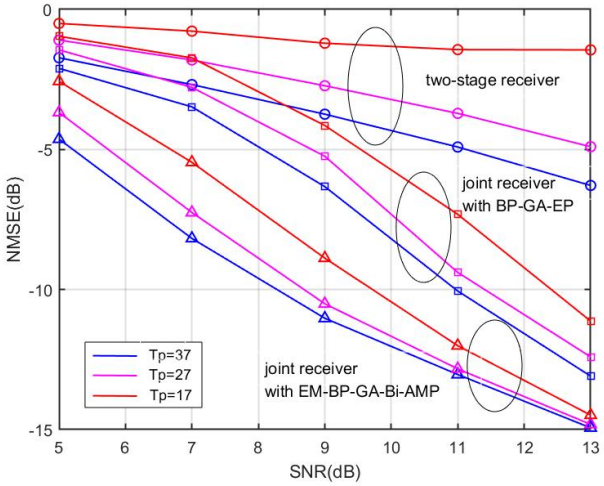

Fig. 3. Channel NMSE comparison between the joint receiver with EM-BP-GA-Bi-AMP, joint receiver with BP-GA-EP and two-stage receiver.

In two-stage receiver, the BS estimates the users' activity states $\hat{\boldsymbol{\xi}}$ and channel $\hat{\mathbf{H}}$ based on received pilots firstly via EM-BP-GAMP [19]. Then, viewing channel estimation as perfect, the BS detects SCMA codewords and decodes coded sequences via GAMP. The maximum iteration number for pilot-aided CE and MUD in two-stage receiver is 100, respectively. Joint receiver with BP-GA-EP [9] assumes channel coefficients follows i.i.d. Gaussian distribution and the maximum iteration number is 200. In proposed receiver, we set $T_1 = 50$, $T_2 = 3$, $\epsilon_h = 50$, $\epsilon_l = \frac{1}{|\mathcal{N}'_a|}\sum_{n\in\mathcal{N}'_a} L_{\xi,n}$, where $\mathcal{N}'_a = \{n|L_{\xi,n} > 0\}$. Damping factor 0.3 is adopted for all algorithms.

The active detection error rate, normalized mean square error (NMSE), and bit error rate (BER) of different receivers are shown in Fig. 2-4. We only consider the NMSE and BER of active users. NMSE is defined as $\mathrm{NMSE}_{[\mathrm{dB}]} =$

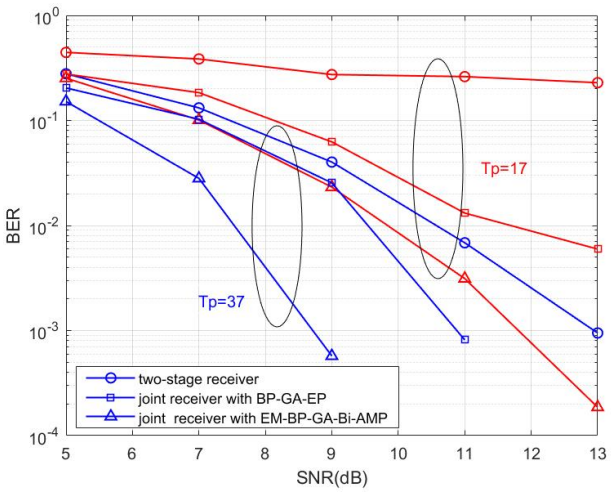

Fig. 4. BER performance comparison between the joint receiver with EM-BP-GA-Bi-AMP, joint receiver with BP-GA-EP and two-stage receiver.

$10\log_{10}(||\hat{\mathbf{H}}_a - \mathbf{H}_a||_F^2/||\mathbf{H}_a||_F^2)$, where $\mathbf{H}_a$ and $\hat{\mathbf{H}}_a$ are exact and estimated channel matrix of active users, respectively.

As shown in Fig. 2 and 3, the conventional receiver cannot detect active users and estimate channel accurately in high SNR when $T_p = 17$. In this circumstance, the number of observations T_p is close to the number of non-zero elements in channel vectors. Thus, GAMP cannot work well in the inference problem with insufficient observations. However, joint CE and MUD module works even with short pilots. This is because the received data sequences also contain channel information, and the correctly detected data sequences can be viewed as pseudo pilots. Compared with joint AUD and pilot-aided CE, the NMSE performance has a substantial improvement in high SNR, where the joint receiver has almost accurate symbol detections.

Fig. 3 also shows that the NMSE performance mainly depends on the length of data sequences in the high SNR regime, and a more extended pilot sequence cannot help estimate the channel more accurately. As can be found in Fig. 4, with better channel estimation, the BER performance of the proposed joint receiver has significant improvement compared with the performance of the two-stage receiver. Moreover, taking advantage of the received data model' structural sparsity, the proposed algorithm also performs better in comparison with the joint receiver with BP-GA-EP.

V. Conclusion

In this paper, we investigated the joint AUD, CE, MUD, and decoding receiver in uplink grant-free massive MIMO SCMA systems. Auxiliary activity indicators were employed to describe structural sparsity in AUD. By exploiting the sparsity feature of users' activity, the joint CE and MUD module was modeled as the joint column-wise sparse bilinear inference problem. According to the modules' characteristics, we adopt different message passing rules and proposed the unified receiver based on EM-BP-GA-Bi-AMP. The proposed scheme has low computational complexity by approximating the probability mass function of SCMA coded symbols to the Gaussian distribution in joint CE and MUD module. Simulation results showed that the proposed unified receiver has a significant performance improvement than counterpart schemes in the literature.

References

[1] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, 2015.

[2] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Commun.(PIMRC)*, Sept. 2013, pp. 332–336.

[3] 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, 2018.

[4] G. Hannak, M. Mayer, A. Jung, G. Matz, and N. Goertz, "Joint channel estimation and activity detection for multiuser communication systems," in *IEEE Int. Conf. Commun. (ICC) Workshop*, June. 2015, pp. 2086–2091.

[5] Z. Chen and W. Yu, "Massive device activity detection by approximate message passing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, March. 2017, pp. 3514–3518.

[6] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June. 2015, pp. 2918–2923.

[7] X. Meng, Y. Wu, Y. Chen, and M. Cheng, "Low complexity receiver for uplink SCMA system via expectation propagation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, March. 2017, pp. 1–5.

[8] Y. Du, B. Dong, W. Zhu, P. Gao, Z. Chen, X. Wang, and J. Fang, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 682–685, 2018.

[9] F. Wei, W. Chen, Y. Wu, J. Ma, and T. A. Tsiftsis, "Message-passing receiver design for joint channel estimation and data decoding in uplink grant-free SCMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 167–181, 2018.

[10] J. Dai, K. Niu, and J. Lin, "Iterative gaussian-approximated message passing receiver for MIMO-SCMA system," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 753–765, 2019.

[11] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, 2018.

[12] 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.

[13] 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, 2019.

[14] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[15] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory ISIT*, Aug. 2011, pp. 2168–2172.

[16] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, 2014.

[17] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854–5867, 2014.

[18] X. Kuai, X. Yuan, W. Yan, H. Liu, and Y. J. Zhang, "Double-sparsity learning based channel-and-signal estimation in massive MIMO with generalized spatial modulation," *IEEE Trans. Commun.*, 2020.

[19] Q. Zou, H. Zhang, D. Cai, and H. Yang, "Message passing based joint channel and user activity estimation for uplink grant-free massive mimo systems with low-precision ADCs," *IEEE Signal Process. Lett.*, vol. 27, pp. 506–510, 2020.