This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

Massive Grant-free OFDMA with Timing and Frequency Offsets

Gangle Sun, Yining Li, Xinping Yi, *Member, IEEE*, Wenjin Wang, *Member, IEEE*, Xiqi Gao, *Fellow, IEEE*, Lei Wang, Fan Wei, Yan Chen, *Member, IEEE*

Abstract—**In the massive grant-free orthogonal frequency division multiple access (OFDMA), the timing and frequency offsets between users impose new challenges on joint active user detection (AUD) and channel estimation (CE) for the subsequent data recovery. In the asynchronous OFDMA, the timing and frequency offset effects can be modeled as the phase-shifting on the pilot matrix. As such, by constructing the measurement matrix with timing and frequency offsets, the joint estimation problem can be formulated as a multiple measurement vector (MMV) recovery problem with structured sparsity. However, such structured sparsity cannot be tackled by the existing compressed sensing (CS) techniques. To address this issue, we develop an efficient structured generalized approximate message passing (S-GAMP) algorithm, which includes the parallel AMP-MMV algorithm as a particular case. To deal with the high dimensionality of the measurement matrix, we propose the dynamic S-GAMP algorithm with a dynamic measurement matrix to reduce the computational complexity. Simulation results confirm the superiority of the proposed algorithms in grant-free OFDMA with both timing and frequency offsets.**

Index Terms—**grant-free, mMTC, channel estimation, active user detection, message passing.**

I. Introduction

MASSIVE machine-type communication (mMTC) is one of the three typical application scenarios in the fifth-generation (5G) mobile communication system, aiming to provide services for massive low-cost and low-energy devices in the Internet of Things (IoT) [2], [3]. In the uplink transmission of the mMTC scenario, the base station (BS) should serve millions of users, of which only sporadic users are active and send short packets [4]–[6].

In the grant-based communication system, the handshaking procedure results in excessive signaling overhead, network congestion, and high transmission latency [7]. The grant-free transmission system is proposed to address this issue, where users send messages to the BS in the pre-allocated collision domain without the scheduling process [7], [8]. To detect transmitted data, the BS must carry out channel estimation (CE) and active user detection (AUD) based on the pilot signals [9].

In current wireless communication systems, the orthogonal frequency division multiple access (OFDMA) technique has been widely adopted for high spectrum efficiency and flexible resource element allocation. Recently, massive grant-free transmission in the OFDMA framework has attracted plenty of research interest where massive users access for transmission on the same resource elements without the scheduling process [10]. In practical OFDMA uplink transmission, e.g., LTE or 5G NR, timing offsets between users are allowed within one cyclic prefix (CP) duration, thus guarantee the orthogonality of subcarriers. However, these allowable users' timing offsets cause the inevitable phase shift of user signals in OFDMA-based grant-free transmission. On the other hand, although frequency offset compensation may be applied in user equipment (UE), there are still slight residual frequency offsets between users. The phase shift of user signals caused by the timing and frequency offsets imposes significant challenges on the joint CE and AUD in the BS. This paper investigates the impact of timing and frequency offsets between users on grant-free transmission and designs efficient algorithms for joint CE and AUD.

A. Prior Work

By exploiting the sporadic activity feature in the mMTC scenario, the joint AUD and CE problem can be formulated as a single measurement vector (SMV) or multiple measurement vector (MMV) problem, depending on the number of receiving antennas [4], [11]. The compressed sensing (CS) techniques have been developed to solve such problems, and can be divided into three categories: convex optimization-based approaches [12], greedy algorithms [13]–[15], and Bayesian methods [16]–[20].

A mixed $l_{2,1}$-regularization penalty function based on the least absolute shrinkage and selection operator (LASSO) was proposed in [21], and the alternating direction method of

Manuscript received May 13, 2021; revised August 28, 2021; accepted October 8, 2021. Date of publication xx, xxxx; date of current version October 14, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61761136016, 61631018 and 61801114, in part by the National Key R&D Program of China under Grant 2019YFB1803102, and in part by the Jiangsu Province Basic Research Project under Grant BK20192002. The work of Xinping Yi was supported in part by the Royal Society International Exchanges Program under Award IEC\NSFC\201080. Part of this work has been presented at the 13th International Conference on Wireless Communications and Signal Processing, Changsha, China, 2021 [1]. The associate editor coordinating the review of this article and approving it for publication was Derrick Wing Kwan Ng. *(Corresponding author: Wenjin Wang.)*

Gangle Sun, Yining Li, Wenjin Wang, and Xiqi Gao are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with Purple Mountain Laboratories, Nanjing 211100, China (e-mail: sungangle@seu.edu.cn; ynli@seu.edu.cn; wangwj@seu.edu.cn; xqgao@seu.edu.cn).

Xinping Yi is with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3BX, U.K. (e-mail: xinping.yi@liverpool.ac.uk).

Lei Wang, Fan Wei, and Yan Chen are with Huawei Technologies Company Ltd., Shanghai 201206, China (e-mail: wanglei888@huawei.com; weifan5@huawei.com; bigbird.chenyan@huawei.com).

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

multipliers algorithm was adopted to handle the large-scale convex joint estimation problem. According to the stability of the user's activity within the frame, [22] proposed the block CS-based sparse signal recovery problem and solved it by two enhanced greedy algorithms, which respectively adopted additive white Gaussian noise floor based threshold, and statistics and machine learning-based cross-validation to determine the termination conditions. By utilizing the sparsity and similarity between neighboring access points, [23] proposed a covariance-based method to perform excellent cooperative activity detection for grant-free massive random access.

More recently, attention has been placed on Bayesian algorithms such as sparse Bayesian learning (SBL) and message passing algorithms to achieve joint estimation. In [16], the SBL algorithm was utilized for CE and data detection in orthogonal frequency division multiplexing (OFDM) systems. In the uplink grant-free scenario, CE and AUD are achieved by the SBL algorithm [24] and the approximate message passing (AMP) algorithm with a soft threshold denoiser [25]. By utilizing the feature of channel sparsity, Yu et al. proposed an AMP algorithm based on the minimum mean squared error (MMSE) denoiser [4]. When the BS is equipped with multiple antennas, the joint estimation problem can be formulated as the MMV problem [7], where the supports of all sparse vectors are identical. This problem can be solved by the vector denoiser-based AMP algorithm [4], [26] and the parallel AMP-MMV algorithm [4], [27]. Furthermore, the generalized approximate message passing (GAMP) algorithm was proposed by [20] to solve the SMV problem in generalized linear systems, and later on, it was developed to solve the MMV recovery problem in [28]. The message passing-based block SBL algorithm was proposed in [29] for joint estimation, which can reduce the computational complexity while achieving similar performance to the block orthogonal matching pursuit algorithm. Based on the expectation maximization and hybrid message passing algorithms, [30] achieved the joint user activity tracking and data detection in the faster-than-Nyquist non-orthogonal multiple access uplink random access.

Considering the different timing offsets caused by the users' different geographical locations and transmission environments, several joint estimation schemes against the timing offset have been proposed in [9], [31], [32]. The authors in [9] transformed the estimation problem from the frequency domain to the time domain to utilize the access delay features. The SBL algorithm and support vector machine (SVM) classifier are proposed for CE and AUD, respectively. [31] introduced the auxiliary preamble structure to detect user activity and proposed the modified interleave-division multiple access receiver to mitigate the interference caused by asynchronous transmission. The joint estimation problem was formulated as the signal recovery problem with the hierarchical sparse structure in [32], and the learned approximate message passing algorithm was proposed to improve the performance without prior information.

B. Motivation and Main Contribution

From the above state-of-the-art overview, it can be found that most current works focus on the research of massive grant-free uplink transmission either in synchronous scenarios or asynchronous scenarios with only timing offset. As a matter of fact, in a practical OFDMA system, both the timing and frequency offsets are inevitable due to the signal transmission distance and variation of oscillators. To the best of our knowledge, little work has been carried out on massive grant-free OFDMA in the presence of timing and frequency offset discrepancy between users. Moreover, when the timing and frequency offsets come to play together, the system model is quite different due to the coupled phase shift of users' signals caused by them. In other words, the previous system models and the corresponding algorithms are not compatible with the practical OFDMA-based massive grant-free system in the presence of both timing and frequency offsets, which results in the unsatisfactory joint estimation performance for grant-free massive access. To deal with the timing and frequency offset in OFDMA-based mMTC scenarios, we are motivated to complete this work. Our contributions can be summarized as follows:

- By decoupling the complex combination of phase shift caused by the timing and frequency offsets, we build the signal model of the grant-free massive access transmission in the asynchronous OFDMA system, covering both the existing synchronous and asynchronous scenarios with only the timing offsets as particular cases. However, due to the random nature of the timing and frequency offsets, the equivalent pilot measurement matrix with phase shift is random and unknown to the BS, making the existing CS algorithms unable to be applied to solve the joint estimation problem. To solve the uncertainty of the measurement matrix, we expand the original pilot measurement matrix with all possible timing offsets and a finite number of discrete frequency offsets and formulate the joint CE and AUD problem as a generalized MMV problem with structured sparsity.
- To solve the above structured sparse MMV problem, we develop an efficient structured generalized approximate message passing (S-GAMP) algorithm, which sets up indicator vectors especially for representing this unique sparsity structure. The proposed S-GAMP algorithm divides the original MMV problem into several independent SMV subproblems and exchanges soft information of indicator vectors from these SMV parts to utilize the prior information of structured sparsity and joint sparsity in the original MMV problem. Moreover, the conventional GAMP algorithm can be seen as a particular case of the S-GAMP in the synchronous transmission scenario.
- To reduce the computational complexity due to the high dimensionality of the measurement matrix, we propose the dynamic S-GAMP algorithm with a low-dimensional dynamic measurement matrix. In each iteration, a few column vectors are extracted from the original measurement matrix to form a low-dimensional measurement matrix to reduce the computational complexity. The optimal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

column vectors estimated in each iteration will be used to extract the column vectors more accurately in the next iteration to reduce the error caused by extraction. As a byproduct, this algorithm can also improve the joint estimation performance because it makes the measurement matrix closer to the independent and identically distributed (i.i.d.) Gaussian matrix so as to meet the requirement of the GAMP algorithm.

The remainder of this paper is organized as follows. We build the system model in Section II. The S-GAMP algorithm is developed in Section III. The method to reduce the algorithm complexity by dynamically updating the measurement matrix is proposed in Section IV. Simulation results of joint estimation performance are included in Section V. Conclusions are drawn in Section VI.

Notation: Throughout this article, uppercase and lowercase bold-face letters denote matrices and column vectors, respectively. In addition, $\mathbf{a}(l)$ represents the l-th element of the vector $\mathbf{a}$ and $\mathbf{A}(m,n)$ represents the element in the m-th row and n-th column of the matrix $\mathbf{A}$. Moreover, $\mathbf{I}_m$ denotes the m-dimensional identity matrix and $(\mathbf{I}_m)_d$ is the matrix generated by cyclically shifting all the row vectors in $\mathbf{I}_m$ to the left by d units simultaneously. The vector $\mathbf{e}_i$ represents the unit vector with the i-th element being one, and its dimension depends on the needs in the calculation process. The symbols $\mathbb{C}$, $\mathbb{R}$, and $\mathbb{N}$ represent the fields of complex numbers, real numbers, and integers, respectively. The expression $\mathcal{CN}\left(x;\mu,\sigma^2\right)$ represents the complex Gaussian distribution function of variable x, with expectation μ and variance σ^2. The superscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ denote the conjugate, transpose and conjugate transpose operations, respectively. $diag(\mathbf{x})$ is a diagonal matrix with elements of $\mathbf{x}$ on its diagonal. $\|\cdot\|_0$, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the l_0, l_2 and Frobenius norm, respectively. Furthermore, the operator $\lfloor\cdot\rfloor$ means to round down the real number, and the operator $\otimes$ represents the Kronecker product operation. $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote mathematical expectation and variance operations, respectively. $\delta(\cdot)$ is the Dirac delta function.

II. Problem Formulation

A. Signal Model

Consider the multiuser uplink transmission with OFDMA, where the M-antenna BS serves N single-antenna potential users, and only N_a of them are active. As shown in Fig. 1, an OFDM symbol contains N_c subcarriers, among which only N_{sub} adjacent subcarriers are allowed to be shared by users. In addition, the transmission of each user in the time domain occupies the same N_{sym} OFDM symbols. The transmitted signals consist of pilots and data, with the pilots occupying S of the allocated N_{sub} subcarriers and G of the N_{sym} OFDM symbols. The indices of the subcarriers and the OFDM symbols occupied by pilots are represented as $\mathbf{s}=[k_1,k_2,\cdots,k_S]^T\in\mathbb{N}^{S\times 1}$ and $\mathbf{g}=[t_1,t_2,\cdots,t_G]^T\in\mathbb{N}^{G\times 1}$, respectively, where $k_1<k_2<\cdots<k_S$ and $t_1<t_2<\cdots<t_G$. In other words, each active user occupies the same resource elements and transmits $L=SG$ pilots simultaneously. The above system parameters are summarized in Table I for reference.

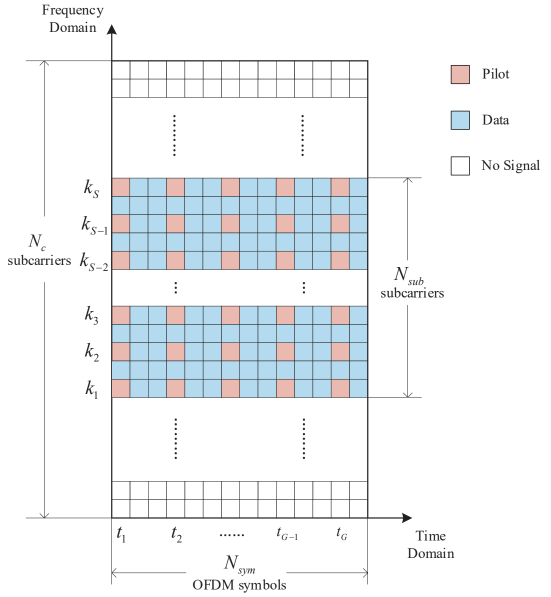


Fig. 1. Time-frequency resource allocation in the mMTC scenario. Each square represents a resource element (RE) occupying an OFDM symbol in the time domain and a subcarrier in the frequency domain. The red and the blue parts represent the pilot and data signal transmitted by the active user, respectively, whereas the white part indicates that the user does not transmit any signal over the RE.

TABLE I
Variable Description

Variables	Meanings
M	The number of the BS antennas
N	The number of single-antenna potential users
N_a	The number of active users
N_c	The number of subcarriers in one OFDM symbol
N_{sub}	The number of adjacent subcarriers assigned to users
$\mathbf{s}$	The index vector of subcarriers occupied by pilots
N_{sym}	The number of OFDM symbols assigned to users
$\mathbf{g}$	The index vector of OFDM symbols occupied by pilots
L	The length of pilot sequences

Suppose only a small fraction of subcarriers are occupied by each user, such that the considered system is a narrowband OFDMA. Therefore, the channel coefficient between each user and each antenna can be modeled as an independent Gaussian distributed random variable. That is, the uplink channel vector $\mathbf{h}_n\in\mathbb{C}^{M\times 1}$ between the n-th user and the BS is given by

$$\mathbf{h}_n\sim\mathcal{CN}(\mathbf{0},\sigma_h^2\mathbf{I}_M), \tag{1}$$

where σ_h^2 is the variance of channel coefficient that is identical across users and antennas. We consider the block fading channel model where the channel coefficients remain static within each block.

Each user is assigned with unique but not necessarily orthogonal pilot sequences for CE and AUD. We denote the frequency-domain pilots of the n-th user on the t-th OFDM symbol as $\mathbf{x}_n^t\in\mathbb{C}^{S\times 1}$ and it is modulated to the time-domain pilot signal $\tilde{\mathbf{x}}_n^t\in\mathbb{C}^{N_c\times 1}$ as follows:

$$\tilde{\mathbf{x}}_n^t=\mathbf{W}_{\mathbf{s}}^H\mathbf{x}_n^t, \tag{2}$$

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. 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/TWC.2021.3121066

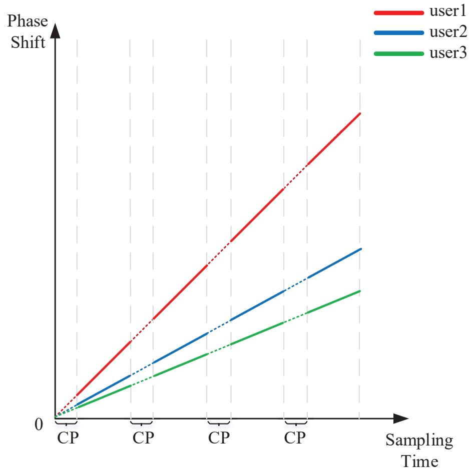


Fig. 2. The effect of frequency offset on the phase of the time-domain sampled signal is linearly related to the sampling time, and the rate of phase shift will be different for users with different values of frequency offset. In addition, the removal of CP sequences results in a discontinuous phase shift between adjacent OFDM symbols.

where $\mathbf{W}_\mathbf{s} \in \mathbb{C}^{S\times N_c}$ represents the matrix composed of S row vectors ($S \ll N_c$) extracted from the N_c-point discrete Fourier transform (DFT) matrix $\mathbf{W}$ indexed by the vector $\mathbf{s}$. Since the pilot and data signals occupy different subcarriers that are sufficiently apart, the inter-carrier interference between them is negligible. Therefore, we place our focus on the pilot signals in the rest of the paper.

The time-domain signal can be generated by adding the CP sequences of length N_{CP} in front of the signal $\tilde{\mathbf{x}}_n^t$. Since the CP sequences are the same as the last N_{CP} samples of $\tilde{\mathbf{x}}_n^t$, the influence of the timing offset on the signal $\tilde{\mathbf{x}}_n^t$ is equivalent to cyclically shifting $\tilde{\mathbf{x}}_n^t$ to the right by τ_n sampling intervals. On the other hand, the phase shift of the time-domain signal caused by the frequency offset ε_n has a linear relationship with sampling time [33], [34], as shown in Fig. 2. It is worth noting that the phase shift of adjacent OFDM symbols is discontinuous due to the removal of the CP sequences after receiving the signal.

After removing the CP sequences, the time domain received signal $\tilde{\mathbf{Y}}^t \in \mathbb{C}^{N_c\times M}$ on the t-th OFDM symbol can be written as

$$\tilde{\mathbf{Y}}^t = \sum_{n=1}^{N} \xi_n \mathbf{\Lambda}^t_{\varepsilon_n} (\mathbf{I}_{N_c})_{\tau_n} \tilde{\mathbf{x}}_n^t \mathbf{h}_n^T + \tilde{\mathbf{Z}}^t, \tag{3}$$

where $\xi_n \in \{0,1\}$ is the activity indicator of the n-th user and it is equal to 1 when the n-th user is active and 0 otherwise. We assume that the activity probabilities of all users are identical and represented by α. The matrix $\mathbf{\Lambda}^t_{\varepsilon_n} \triangleq \phi^t \text{diag}(1,\omega,\cdots,\omega^{N_c-1})$ represents the phase shift matrix caused by the frequency offset ε_n, where $\omega = e^{\frac{j2\pi\varepsilon_n}{N_c}}$ and $\phi^t = \omega^{N_{CP}+(t-1)(N_{CP}+N_c)}$ is the cumulative phase shift in the previous sequences. Moreover, the cyclically shifted matrix $(\mathbf{I}_{N_c})_{\tau_n}$ shows the effect of timing offset τ_n. $\tilde{\mathbf{Z}}^t \in \mathbb{C}^{N_c\times M}$ is the circularly symmetric complex white Gaussian noise matrix.

Plugging equations (2) into (3), we obtain the demodulated signal $\mathbf{Y}^t \in \mathbb{C}^{S\times M}$, which is given by

$$\begin{aligned}\mathbf{Y}^t =& \mathbf{W}_\mathbf{s}\tilde{\mathbf{Y}}^t \\ =& \sum_{n=1}^{N} \xi_n \underbrace{\mathbf{W}_\mathbf{s}\mathbf{\Lambda}^t_{\varepsilon_n} (\mathbf{I}_{N_c})_{\tau_n} \mathbf{W}_\mathbf{s}^H}_{\triangleq \mathbf{P}_n^t} \mathbf{x}_n^t \mathbf{h}_n^T + \mathbf{Z}^t \\ =& \mathbf{X}^t\mathbf{H} + \mathbf{Z}^t,\end{aligned} \tag{4}$$

where $\mathbf{P}_n^t \in \mathbb{C}^{S\times S}$ represents the phase shift matrix of the n-th user caused by the timing and frequency offsets on the t-th OFDM symbol, $\mathbf{X}^t \triangleq [\mathbf{P}_1^t\mathbf{x}_1^t, \mathbf{P}_2^t\mathbf{x}_2^t, \cdots, \mathbf{P}_N^t\mathbf{x}_N^t] \in \mathbb{C}^{S\times N}$ denotes the equivalent pilot signal matrix, and $\mathbf{H} \triangleq [\xi_1\mathbf{h}_1, \xi_2\mathbf{h}_2, \cdots, \xi_N\mathbf{h}_N]^T \in \mathbb{C}^{N\times M}$ is the equivalent channel matrix.

Further, the received signal over G OFDM symbols can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{H} + \mathbf{Z}, \tag{5}$$

where $\mathbf{Y} \triangleq [(\mathbf{Y}^{t_1})^T, (\mathbf{Y}^{t_2})^T, \cdots, (\mathbf{Y}^{t_G})^T]^T \in \mathbb{C}^{L\times M}$ is the demodulated received signal on G OFDM symbols with $L = SG$ as defined earlier, $\mathbf{Z} \in \mathbb{C}^{L\times M}$ is the system noise matrix whose elements follow the Gaussian distribution with mean value of 0 and variance of σ_z^2, and $\mathbf{X} \triangleq [(\mathbf{X}^{t_1})^T, (\mathbf{X}^{t_2})^T, \cdots, (\mathbf{X}^{t_G})^T]^T \in \mathbb{C}^{L\times N}$ is the equivalent pilot matrix of N users on the G OFDM symbols. To simplify the subsequent analysis of the phase shift caused by the timing and frequency offsets, we rewrite the matrix $\mathbf{X}$ as follows

$$\mathbf{X} = [\mathbf{P}_1\mathbf{x}_1, \mathbf{P}_2\mathbf{x}_2, \cdots, \mathbf{P}_N\mathbf{x}_N], \tag{6}$$

where $\mathbf{x}_n \triangleq [(\mathbf{x}_n^{t_1})^T, (\mathbf{x}_n^{t_2})^T, \cdots, (\mathbf{x}_n^{t_G})^T]^T \in \mathbb{C}^{L\times 1}$ denotes the unique frequency-domain pilot sequences sent by the n-th user and satisfies the i.i.d. zero-mean complex Gaussian distribution with the variance of $\frac{1}{L}$, i.e, $\mathbf{x}_n \sim \mathcal{CN}(\mathbf{0}, \frac{1}{L}\mathbf{I}_L), \forall n = 1,2,\cdots,N$. The matrix $\mathbf{P}_n \in \mathbb{C}^{L\times L}$ represents the phase shift matrix of the n-th user resulted from the timing and frequency offsets on the G OFDM symbols. The phase shift matrix $\mathbf{P}_n$ is defined as

$$\mathbf{P}_n \triangleq \begin{bmatrix} \mathbf{P}_n^{t_1} & & & \\ & \mathbf{P}_n^{t_2} & & \\ & & \ddots & \\ & & & \mathbf{P}_n^{t_G} \end{bmatrix}. \tag{7}$$

In the next section, we will analyze and approximate the phase shift matrix $\mathbf{P}_n$ to simplify the subsequent analysis.

B. Approximation of Phase Shift Matrix

Before proceeding further, we take a closer look at the elements in $\mathbf{P}_n^t$. For the k_u-th and k_v-th subcarriers, the (u,v)-th element of $\mathbf{P}_n^t(u,v)$ can be expressed as

$$\begin{aligned}\mathbf{P}_n^t(u,v) =& \frac{1}{N_c}\omega^{N_{CP}+(t-1)(N_{CP}+N_c)}\psi^{\varepsilon_n-k_u+1} \\ &\cdot \sum_{i=0}^{N_c-1} e^{\frac{j2\pi(k_v-k_u+\varepsilon_n)i}{N_c}+j2\pi(k_u-1-\varepsilon_n)\left\lfloor\frac{i+\tau_n}{N_c}\right\rfloor},\end{aligned} \tag{8}$$

where $\psi = e^{\frac{j2\pi\tau_n}{N_c}}$, and k_u, k_v are the u-th and v-th elements of the vector $\mathbf{s}$.

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

When $u \neq v$, $\mathbf{P}_n^t(u,v)$ indicates the coefficient of the inter-carrier interference (ICI). In practical communications, e.g., LTE or 5G NR, the users must estimate the frequency offsets by detecting the downlink synchronization signals in cell search procedure [35] and utilize the frequency offset compensation methods [36], [37] to compensate the large frequency offsets, so that the frequency offsets can be controlled within a slight range. It is worth noting that the 3rd Generation Partnership Project (3GPP) 38.101-1 specifies the minimum radio frequency requirements for new radio users, where the frequency error is within ± 0.1 ppm, and the subcarrier spacing must be 15 kHz, 30 kHz or 60 kHz [38]. Consequently, the maximum frequency offset $\varepsilon_{\max}$ is between 0.27% and 4.75%. Given the slight residual frequency offsets in the practical communication systems, we can figure out the amplitude of $\mathbf{P}_n^t(u,v)$ as

$$\begin{aligned} \left|\mathbf{P}_n^t(u,v)\right| &= \frac{1}{N_c}\left|\frac{1-e^{j2\pi\varepsilon_n}}{1-e^{\frac{j2\pi(k_v-k_u+\varepsilon_n)}{N_c}}}\right| \\ &\approx \left|\frac{\varepsilon_n}{k_v-k_u+\varepsilon_n}\right|, \quad u \neq v, \end{aligned} \tag{9}$$

where $|\mathbf{P}_n^t(u,v)|$ is not related to the timing offset τ_n, and the maximum amplitude of ICI coefficient is around 0.01. As such, the ICI caused by the slight residual frequency offsets is negligible and the matrix $\mathbf{P}_n^t$ can be approximated as a diagonal matrix for the sake of tractability in modeling.

Further, given the infinitesimal ε_n, $\mathbf{P}_n^t(u,u)$ can be approximated by utilizing the Taylor expansion, with its summation term in (8) being expressed as

$$\begin{aligned} &\sum_{i=0}^{N_c-1}\omega^i e^{j2\pi(k_u-1-\varepsilon_n)\left\lfloor\frac{i+\tau_n}{N_c}\right\rfloor} \\ &= \frac{N_c+j\pi\left(N_c-2\tau_n\right)\varepsilon_n+o\left(\varepsilon_n\right)}{1+\frac{j\pi}{N_c}\varepsilon_n+o\left(\varepsilon_n\right)} \\ &\approx N_c\omega^{\frac{N_c-1}{2}-\tau_n}, \end{aligned} \tag{10}$$

where $o\left(\varepsilon_n\right)$ is the high-order infinitesimal of ε_n. Combining equation (8) and (10), the main diagonal element $\mathbf{P}_n^t(u,u)$ can be approximated as

$$\mathbf{P}_n^t(u,u) \approx \omega^{(N_{CP}+N_c)t-\frac{N_c+1}{2}}\psi^{1-k_u}, \tag{11}$$

where the two parts of the above expression represent the phase shift caused by the frequency offset ε_n and timing offset τ_n, respectively.

By decoupling the phase shift caused by the timing and frequency offsets, the phase shift matrix $\mathbf{P}_n$ can be expressed as follows:

$$\mathbf{P}_n = \mathbf{\Gamma}_{\varepsilon_n} \otimes \mathbf{\Gamma}_{\tau_n}, \tag{12}$$

where the diagonal matrices $\mathbf{\Gamma}_{\varepsilon_n}$ and $\mathbf{\Gamma}_{\tau_n}$ are the phase shift matrices caused by frequency offset ε_n and timing offset τ_n, respectively, whose expressions are shown in equation (13) and (14).

$$\mathbf{\Gamma}_{\tau_n} \triangleq \operatorname{diag}\left[\psi^{1-k_1},\psi^{1-k_2},\cdots,\psi^{1-k_S}\right]. \tag{14}$$

TABLE II
THE DEVIATION CORRESPONDING TO DIFFERENT FREQUENCY OFFSETS

Frequency Offset ε_n	**Deviation** E_a
$\pm 0.27\%$	1.20×10^{-5}
$\pm 1.10\%$	1.99×10^{-4}
$\pm 2.00\%$	6.58×10^{-4}
$\pm 3.38\%$	1.88×10^{-3}
$\pm 4.75\%$	3.71×10^{-3}

To verify the accuracy of the above approximation, we calculate the deviation E_a of the above approximation for the main diagonal elements can be defined as

$$\begin{aligned} E_a &\triangleq \left|\mathbf{P}_n^t(u,u)-\omega^{(N_{CP}+N_c)t-\frac{N_c+1}{2}}\psi^{1-k_u}\right| \\ &= \left|1-\frac{\sin\left(\pi\varepsilon_n\right)}{N_c\sin\left(\frac{\pi\varepsilon_n}{N_c}\right)}\right|. \end{aligned} \tag{15}$$

It can be found that the deviation E_a is independent of the timing offset τ_n and increases as $|\varepsilon_n|$ does within the range of $[0,\varepsilon_{\max}]$. Table II shows the value of the deviation E_a corresponding to different value of the frequency offset ε_n, and it can be found that the deviation of the above approximation is negligible.

C. Structured Sparse Model

Differently from the conventional CS models, the measurement matrix $\mathbf{X}$ in (5) is unknown to the receivers due to the random nature of timing and frequency offsets. Without knowing such a matrix, it is challenging for CS algorithms to achieve excellent joint estimation performance. To address this issue, we expand the system model (5), where the known complete measurement matrix is expanded from $\mathbf{X}$ with all possible timing offsets and a finite number of discrete frequency offsets within $[-\varepsilon_{\max},\varepsilon_{\max}]$. Correspondingly, the joint activity detection and channel estimation problem is formulated as a generalized MMV problem with structured sparsity, which means each block in the evaluated matrix has at most one non-zero row, and the positions of non-zero rows indicate the values of timing and frequency offsets.

For tractability, we assume that the possible timing offsets are within $\mathbf{d} = [1,2,\cdots,D]^T \in \mathbb{N}^{D\times 1}$, and the frequency offsets are within $\mathbf{q} = \left[\varepsilon^{(1)},\varepsilon^{(2)},\cdots,\varepsilon^{(Q)}\right]^T \in \mathbb{R}^{Q\times 1}$, which is uniformly sampled from the range $[-\varepsilon_{\max},\varepsilon_{\max}]$. A proper choice of Q compromises between computational complexity and tractability. Given the small value of $\varepsilon_{\max}$, the uniform sampling with a small Q does not cause much deviation. Therefore, the system model (5) can be rewritten as follows:

$$\mathbf{Y} = \mathbf{X}_e\mathbf{H}_e + \mathbf{Z}, \tag{16}$$

where $\mathbf{X}_e \triangleq [\mathbf{X}_{e,1},\mathbf{X}_{e,2},\cdots,\mathbf{X}_{e,N}] \in \mathbb{C}^{L\times(NDQ)}$ is the expanded known measurement matrix, $\mathbf{X}_{e,n} \in \mathbb{C}^{L\times(DQ)}$ is given by

$$\mathbf{X}_{e,n} = \left[\mathbf{P}_{n,1,1}\mathbf{x}_n,\mathbf{P}_{n,1,2}\mathbf{x}_n,\cdots,\mathbf{P}_{n,D,Q}\mathbf{x}_n\right], \tag{17}$$

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

$$\mathbf{\Gamma}_{\varepsilon_n} \triangleq \omega^{-\frac{N_c+1}{2}} \mathrm{diag}\left[\omega^{(N_{CP}+N_c)t_1}, \omega^{(N_{CP}+N_c)t_2}, \cdots, \omega^{(N_{CP}+N_c)t_G}\right], \tag{13}$$

where $\mathbf{P}_{n,d,q}$ represents the phase shift matrix $\mathbf{P}_n$ with the timing offset $\tau_n = \mathbf{d}(d)$ and the frequency offset $\varepsilon_n = \mathbf{q}(q)$, i.e.,

$$\mathbf{P}_{n,d,q} \triangleq [\mathbf{\Gamma}_{\varepsilon_n}]_d \otimes [\mathbf{\Gamma}_{\tau_n}]_q,$$

where $[\mathbf{\Gamma}_{\varepsilon_n}]_d$ and $[\mathbf{\Gamma}_{\tau_n}]_q$ represent the matrix $\mathbf{\Gamma}_{\varepsilon_n}$ when $\tau_n = \mathbf{d}(d)$ and the matrix $\mathbf{\Gamma}_{\tau_n}$ when $\varepsilon_n = \mathbf{q}(q)$, respectively. In addition, $\mathbf{H}_e = \left[\mathbf{H}_{e,1}^T, \mathbf{H}_{e,2}^T, \cdots, \mathbf{H}_{e,N}^T\right]^T \in \mathbb{C}^{(NDQ)\times M}$ is the structured sparse channel matrix, where $\mathbf{H}_{e,n}$ contains at most one non-zero row vector, as shown in Fig. 3. The submatrix $\mathbf{H}_{e,n} \in \mathbb{C}^{(DQ)\times M}$ corresponding to the n-th user can be expressed as

$$\mathbf{H}_{e,n} = \boldsymbol{\eta}_n \mathbf{h}_n^T, \tag{18}$$

where $\boldsymbol{\eta}_n \in \{0,1\}^{DQ\times 1}$ is the indicator vector and there is at most one non-zero element in $\boldsymbol{\eta}_n$. To be specific, $\boldsymbol{\eta}_n = \mathbf{0}$ means that the n-th user is inactive with the probability of $1-\alpha$, i.e. $p(\boldsymbol{\eta}_n = \mathbf{0}) = 1-\alpha$; In addtion, $\boldsymbol{\eta}_n = \mathbf{e}_i$ means that the n-th user is active, and $i \in \{1,2,\cdots,DQ\}$ indicates the position of the non-zero row in the $\mathbf{H}_{e,n}$, which can subsequently represent the timing and frequency offset. By assuming that the user's timing and frequency offsets are chosen from the vectors $\mathbf{d}$ and $\mathbf{q}$ with equal probability, respectively, we can get $p(\boldsymbol{\eta}_n = \mathbf{e}_i) = \frac{\alpha}{DQ}$. Therefore, the probability density function $p(\boldsymbol{\eta}_n)$ of the indicator vector $\boldsymbol{\eta}_n$ can be written as

$$\begin{aligned} p(\boldsymbol{\eta}_n) &= \sum_{\bar{\boldsymbol{\eta}} \in \mathcal{S}_{\bar{\boldsymbol{\eta}}}} p(\boldsymbol{\eta}_n = \bar{\boldsymbol{\eta}})\, \delta(\boldsymbol{\eta}_n - \bar{\boldsymbol{\eta}}) \\ &= \frac{\alpha}{DQ} \sum_{i=1}^{DQ} \delta(\boldsymbol{\eta}_n - \mathbf{e}_i) + (1-\alpha)\,\delta(\boldsymbol{\eta}_n), \end{aligned} \tag{19}$$

where $\mathcal{S}_{\bar{\boldsymbol{\eta}}} = \{\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{DQ}\}$.

By expanding the measurement matrix $\mathbf{X}_e$, we transform the system model (5) into a new one (16) with a structured sparse $\mathbf{H}_e$. It can be recognized as a CS model and solved by employing message passing algorithms, which will be detailed in Section III.

III. Joint Active User Detection and Channel Estimation with Timing and Frequency Offset

This section focuses on the joint CE and AUD with both timing and frequency offsets, which can be formulated as a generalized MMV problem with structured sparsity. While such a formulation can be solved by using classical approaches such as the parallel AMP-MMV algorithm in the synchronous transmission system with the sparsity structure ignored, the joint estimation performance is severely degraded because their assumption on the random position of non-zero rows does not match the unique sparse structure here. To address this issue, we propose a novel approach (S-GAMP) by combining the GAMP with the belief propagation (BP) algorithm, setting up a new factor especially for representing this unique sparse structure. The soft information of structured sparsity is exchanged on the new factor, which makes the prior information of structured sparsity be utilized effectively, thus improving the joint estimation performance.

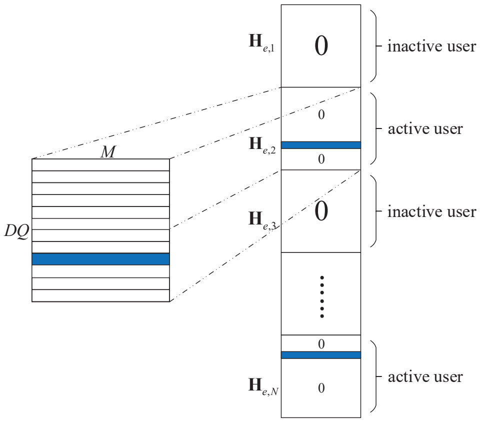


Fig. 3. Structured sparsity of matrix $\mathbf{H}_e$. The white parts represent zero matrices, and the blue parts represent non-zero vectors. There is at most one non-zero row vector in the matrix $\mathbf{H}_{e,n}$. If there is a non-zero row vector in the $\mathbf{H}_{e,n}$, the n-th user is active, and the position of that indicates the values of timing offset τ_n and frequency offset ε_n; Otherwise, the n-th user is inactive if $\mathbf{H}_{e,n}$ is a zero matrix.

A. Derivation of the S-GAMP Algorithm

In what follows, we detail the derivation behind the algorithm. The derivation starts with the joint posterior probability $p(\mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta}|\mathbf{Y})$, which can be expressed as

$$p(\mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta}|\mathbf{Y}) = \frac{1}{p(\mathbf{Y})} p(\mathbf{Y}, \mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta}), \tag{20}$$

with $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T, \cdots, \boldsymbol{\eta}_N^T]^T \in \{0,1\}^{NDQ\times 1}$, where $p(\boldsymbol{\eta}) = \prod_{n=1}^N p(\boldsymbol{\eta}_n)$ because the active states and values of timing and frequency offsets of different users are independent with each other. The matrix $\mathbf{R} = \mathbf{X}_e\mathbf{H}_e$ represents the received signal matrix without noise. Given the knowledge of $\mathbf{Y}$, the maximization of posterior probability function $p(\mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta}|\mathbf{Y})$ is equivalent to the maximization of joint probability density function $p(\mathbf{Y}, \mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta})$. The expression of $p(\mathbf{Y}, \mathbf{R}, \mathbf{H}_e, \boldsymbol{\eta})$ can be factorized as equation (21), where $\mathbf{h}_m^e = \left[(\mathbf{h}_{m1}^e)^T, (\mathbf{h}_{m2}^e)^T, \cdots, (\mathbf{h}_{mN}^e)^T\right]^T \in \mathbb{C}^{(NDQ)\times 1}$ and $\mathbf{h}_{mn}^e \in \mathbb{C}^{(DQ)\times 1}$ are the m-th column vectors of matrix $\mathbf{H}_e$ and $\mathbf{H}_{e,n}$, respectively, with the following probability relationship $p(\mathbf{h}_m^e) = \prod_{n=1}^N p(\mathbf{h}_{mn}^e)$. Furthermore, the posteriori probability function $p(\mathbf{H}_e|\boldsymbol{\eta}) = \prod_{m=1}^M \prod_{n=1}^N p(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n)$ because the channel coefficients between different users and different antennas are independent. The function $p(\mathbf{R}|\mathbf{H}_e) = \prod_{l=1}^L \prod_{m=1}^M p(\mathbf{R}(l,m)|\mathbf{h}_m^e)$ due to the independence between different $\mathbf{h}_{mn}^e$ and the independence between pilot sequences from the same user. In addition, the expression of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

$$\begin{aligned} p\left(\mathbf{Y},\mathbf{R},\mathbf{H}_e,\boldsymbol{\eta}\right) &= p\left(\mathbf{Y}|\mathbf{R}\right)p\left(\mathbf{R}|\mathbf{H}_e\right)p\left(\mathbf{H}_e|\boldsymbol{\eta}\right)p\left(\boldsymbol{\eta}\right) \\ &= \prod_{l=1}^{L}\prod_{m=1}^{M} p\left(\mathbf{Y}\left(l,m\right)|\mathbf{R}\left(l,m\right)\right)\prod_{l=1}^{L}\prod_{m=1}^{M} p\left(\mathbf{R}\left(l,m\right)|\mathbf{h}_m^e\right)\prod_{m=1}^{M}\prod_{n=1}^{N} p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right)\prod_{n=1}^{N} p\left(\boldsymbol{\eta}_n\right), \end{aligned} \tag{21}$$

the conditional probability function $p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right)$ in equation (21) is given by

$$\begin{aligned} &p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right) \\ &= \int \delta\left(\mathbf{h}_{mn}^e - \boldsymbol{\eta}_n\mathbf{h}_n\left(m\right)\right)\mathcal{CN}\left(\mathbf{h}_n\left(m\right);0,\sigma_h^2\right)d\mathbf{h}_n\left(m\right), \end{aligned} \tag{22}$$

where the variable $\mathbf{h}_n\left(m\right) \sim \mathcal{CN}(0,\sigma_h^2)$ and the function $\delta\left(\mathbf{h}_{mn}^e - \boldsymbol{\eta}_n\mathbf{h}_n\left(m\right)\right)$ is to constrain the vector $\mathbf{h}_{mn}^e$ to satisfy the structural sparsity described in equation (18). Given the equation (21), the corresponding factor graph can be drawn in Fig. 4, based on which the message passing algorithm will be presented accordingly [39].

Inspired by [4] and [27], our idea of solving the generalized MMV problem is to break it into M independent SMV subproblems, each of which will be solved in parallel. Unlike conventional algorithms using user activity indicator variables, our proposed S-GAMP algorithm sets up indicator vectors $\boldsymbol{\eta}_n$, which contains both the user activity and structured sparsity information. The prior information of structured sparsity and joint sparsity in the original MMV problem can be utilized in the S-GAMP algorithm through exchanged soft information of indicator vector $\boldsymbol{\eta}_n$ among different SMV subproblems. Among many SMV solutions, the BP algorithm is one of the most efficient approaches to perform statistical inference and achieve excellent convergence when the factor graph is tree-like [27]. However, lots of circles between the $\mathbf{h}_{mn}^e$ and $f_{\mathbf{R}(l,m)}$ nodes make it difficult to guarantee the convergence of the BP algorithm. To solve this problem, considering the same cycle structure as in the factor graph of the GAMP algorithm [4], [20], we leverage the GAMP algorithm as a part of the S-GAMP algorithm to guarantee the convergence by approximating the messages passed between the nodes $\mathbf{h}_{mn}^e$ and the nodes $f_{\mathbf{R}(l,m)}$. Therefore, the convergence of the S-GAMP algorithm is the same as the GAMP algorithm [4], [20], [40], which is verified in Fig. 6 in Section V. The proof of the convergence of the GAMP algorithm can be referred to [20].

In particular, the messages passed between the $\boldsymbol{\eta}_n$ and $\mathbf{h}_{mn}^e$ nodes can be derived as follows. Note that $I_{A\to B}\left(x\right)$ denotes the message passed from node A to node B, which is a function of x, and $b_x\left(x\right)$ represents the belief of variable x.

1) Message Computations between Nodes: Since the message $I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right)$ is the output of the GAMP algorithm, and each element of the vector $\mathbf{h}_{mn}^e$ is assumed to be independent of each other, the message $I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right)$ can be expressed as

$$I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right) = \prod_{i=1}^{DQ}\mathcal{CN}\left(\mathbf{h}_{mn}^e\left(i\right);\mu_{mni}^h,\sigma_{mni}^h\right), \tag{23}$$

where μ_{mni}^h and σ_{mni}^h are the mean and variance of the variable $\mathbf{h}_{mn}^e\left(i\right)$, respectively.

Messages from all SMV subproblem parts are sent to node $\boldsymbol{\eta}_n$ to exchange soft information about the unique sparse structure. With the BP rules, the message $I_{f_{mn}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$ related to $\boldsymbol{\eta}_n$ can be expressed by integrating $p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right)I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right)$ with respect to $\mathbf{h}_{mn}^e$, and its expression is given by

$$\begin{aligned} I_{f_{mn}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right) &= \frac{\int p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right)I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right)d\mathbf{h}_{mn}^e}{Z_{f_{mn}\to\boldsymbol{\eta}_n}} \\ &= \frac{\sum\limits_{i=1}^{DQ}\lambda_{mni}\delta\left(\boldsymbol{\eta}_n - \mathbf{e}_i\right) + \delta\left(\boldsymbol{\eta}_n\right)}{\sum\limits_{i=1}^{DQ}\lambda_{mni} + 1}, \end{aligned} \tag{24}$$

where the normalization constant $Z_{f_{mn}\to\boldsymbol{\eta}_n}$ is expressed as equation (25) and the constant λ_{mni} is defined as

$$\lambda_{mni} = \frac{\mathcal{CN}\left(0;\mu_{mni}^h,\sigma_{mni}^h + \sigma_h^2\right)}{\mathcal{CN}\left(0;\mu_{mni}^h,\sigma_{mni}^h\right)}. \tag{26}$$

The exchanged information is fed back to these SMV subproblem parts as input to achieve better joint estimation. The feedback message $I_{\boldsymbol{\eta}_n\to f_{mn}}\left(\boldsymbol{\eta}_n\right)$ can be represented by $I_{f_{m'n}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$ and $p\left(\boldsymbol{\eta}_n\right)$ as

$$\begin{aligned} &I_{\boldsymbol{\eta}_n\to f_{mn}}\left(\boldsymbol{\eta}_n\right) \\ &= \frac{p\left(\boldsymbol{\eta}_n\right)\prod\limits_{m'\neq m} I_{f_{m'n}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)}{Z_{\boldsymbol{\eta}_n\to f_{mn}}} \\ &= \frac{\alpha\sum\limits_{i=1}^{DQ}\delta\left(\boldsymbol{\eta}_n - \mathbf{e}_i\right)\prod\limits_{m'\neq m}\lambda_{m'ni} + \left(1-\alpha\right)DQ\delta\left(\boldsymbol{\eta}_n\right)}{\alpha\sum\limits_{i=1}^{DQ}\prod\limits_{m'\neq m}\lambda_{m'ni} + \left(1-\alpha\right)DQ}, \end{aligned} \tag{27}$$

where $Z_{\boldsymbol{\eta}_n\to f_{mn}}$ is the normalization constant and its expression can be written as

$$\begin{aligned} Z_{\boldsymbol{\eta}_n\to f_{mn}} &= \int p\left(\boldsymbol{\eta}_n\right)\prod_{m'\neq m} I_{f_{m'n}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)d\boldsymbol{\eta}_n \\ &= \frac{\frac{\alpha}{DQ}\sum\limits_{i=1}^{DQ}\prod\limits_{m'\neq m}\lambda_{m'ni} + 1 - \alpha}{\prod\limits_{m'\neq m}\left(\sum\limits_{i=1}^{DQ}\lambda_{m'ni} + 1\right)}. \end{aligned} \tag{28}$$

The message $I_{f_{mn}\to\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right)$ can be written as equation (29) and it is considered as the input of the GAMP algorithm for the next iteration as shown in Fig. 4, where the normalization constant $Z_{f_{mn}\to\mathbf{h}_{mn}^e}$ can be written as

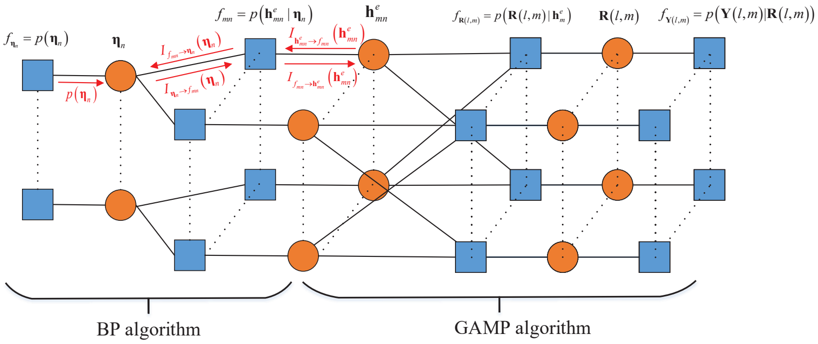
$$Z_{f_{mn}\to\mathbf{h}_{mn}^e} = \int\int I_{\boldsymbol{\eta}_n\to f_{mn}}\left(\boldsymbol{\eta}_n\right)p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right)d\boldsymbol{\eta}_n d\mathbf{h}_{mn}^e = 1. \tag{30}$$

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. 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/TWC.2021.3121066

$$\begin{aligned} Z_{f_{mn}\to\boldsymbol{\eta}_n} &= \int\int p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right) I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right) d\mathbf{h}_{mn}^e d\boldsymbol{\eta}_n \\ &= \sum_{i=1}^{DQ} \mathcal{CN}\left(0;\mu_{mni}^h,\sigma_{mni}^h+\sigma_h^2\right) \prod_{i'\neq i} \mathcal{CN}\left(0;\mu_{mni'}^h,\sigma_{mni'}^h\right) + \prod_{i=1}^{DQ} \mathcal{CN}\left(0;\mu_{mni}^h,\sigma_{mni}^h\right), \end{aligned} \tag{25}$$

$$\begin{aligned} I_{f_{mn}\to\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right) &= \frac{\int I_{\boldsymbol{\eta}_n\to f_{mn}}(\boldsymbol{\eta}_n) p\left(\mathbf{h}_{mn}^e|\boldsymbol{\eta}_n\right) d\boldsymbol{\eta}_n}{Z_{f_{mn}\to\mathbf{h}_{mn}^e}} \\ &= \frac{\alpha \sum\limits_{i=1}^{DQ} \mathcal{CN}\left(\mathbf{h}_{mn}^e(i);0,\sigma_h^2\right) \prod\limits_{i'\neq i} \delta\left(\mathbf{h}_{mn}^e(i')\right) \prod\limits_{m'\neq m} \lambda_{m'ni} + (1-\alpha) DQ\delta\left(\mathbf{h}_{mn}^e\right)}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m'\neq m} \lambda_{m'ni} + (1-\alpha) DQ}, \end{aligned} \tag{29}$$



$f_{\boldsymbol{\eta}_n} = p(\boldsymbol{\eta}_n)$
$\boldsymbol{\eta}_n$
$p(\boldsymbol{\eta}_n)$
$I_{f_{mn}\to\boldsymbol{\eta}_n}(\boldsymbol{\eta}_n)$
$I_{\boldsymbol{\eta}_n\to f_{mn}}(\boldsymbol{\eta}_n)$
$f_{mn} = p(\mathbf{h}_{mn}^e | \boldsymbol{\eta}_n)$
$I_{\mathbf{h}_{mn}^e\to f_{mn}}(\mathbf{h}_{mn}^e)$
$I_{f_{mn}\to\mathbf{h}_{mn}^e}(\mathbf{h}_{mn}^e)$
$\mathbf{h}_{mn}^e$
$f_{\mathbf{R}(l,m)} = p(\mathbf{R}(l,m) | \mathbf{h}_m^e)$
$\mathbf{R}(l,m)$
$f_{\mathbf{Y}(l,m)} = p(\mathbf{Y}(l,m)|\mathbf{R}(l,m))$
BP algorithm
GAMP algorithm

Fig. 4. Factor graph of the system model. The orange circles and blue squares denote variable nodes and function nodes, respectively. We use the BP algorithm to transmit messages in the left tree-like structure and the GAMP algorithm to transmit messages in the circle structure on the right.

2) Probability Density Function of Variable Nodes: According to the BP rules and the messages in (23) and (29), we derive the belief $b_{\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right)$ of $\mathbf{h}_{mn}^e$ as equation (31), where the normalization constant $Z_{\mathbf{h}_{mn}^e}$ is given by

$$\begin{aligned} & Z_{\mathbf{h}_{mn}^e} \\ &= \int I_{f_{mn}\to\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right) I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right) d\mathbf{h}_{mn}^e \\ &= \frac{\left(\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ\right) \prod\limits_{i=1}^{DQ} \mathcal{CN}\left(0;\mu_{mni}^h,\sigma_{mni}^h\right)}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m'\neq m} \lambda_{m'ni} + (1-\alpha) DQ}. \end{aligned} \tag{32}$$

For simplicity, we abbreviate $b_{\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right)$ as $b_{\mathbf{h}_{mn}^e}$ in the following equations. Furthermore, the posterior mean $\mathrm{E}\left[\mathbf{h}_{mn}^e(i)|b_{\mathbf{h}_{mn}^e}\right]$ and posterior variance $\mathrm{Var}\left[\mathbf{h}_{mn}^e(i)|b_{\mathbf{h}_{mn}^e}\right]$ of $\mathbf{h}_{mn}^e(i)$ are given by

$$\begin{aligned} \mathrm{E}\left[\mathbf{h}_{mn}^e(i)|b_{\mathbf{h}_{mn}^e}\right] &= \int \mathbf{h}_{mn}^e(i) b_{\mathbf{h}_{mn}^e} d\mathbf{h}_{mn}^e \\ &= \frac{\alpha \frac{\sigma_h^2 \mu_{mni}^h}{\sigma_{mni}^h+\sigma_h^2} \prod\limits_{m=1}^{M} \lambda_{mni}}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ}, \end{aligned} \tag{33}$$

and

$$\begin{aligned} & \mathrm{Var}\left[\mathbf{h}_{mn}^e(i)|b_{\mathbf{h}_{mn}^e}\right] \\ &= \mathrm{E}[|\mathbf{h}_{mn}^e(i)|^2|b_{\mathbf{h}_{mn}^e}] - \left|\mathrm{E}\left[\mathbf{h}_{mn}^e(i)|b_{\mathbf{h}_{mn}^e}\right]\right|^2, \end{aligned} \tag{34}$$

where $\mathrm{E}[|\mathbf{h}_{mn}^e(i)|^2|b_{\mathbf{h}_{mn}^e}]$ can be expressed as

$$\begin{aligned} \mathrm{E}[|\mathbf{h}_{mn}^e(i)|^2|b_{\mathbf{h}_{mn}^e}] &= \int \left|\mathbf{h}_{mn}^e(i)\right|^2 b_{\mathbf{h}_{mn}^e} d\mathbf{h}_{mn}^e \\ &= \frac{\alpha\left(\left|\frac{\sigma_h^2\mu_{mni}^h}{\sigma_{mni}^h+\sigma_h^2}\right|^2 + \frac{\sigma_h^2\sigma_{mni}^h}{\sigma_{mni}^h+\sigma_h^2}\right) \prod\limits_{m=1}^{M} \lambda_{mni}}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ}. \end{aligned} \tag{35}$$

The probability density function $b_{\boldsymbol{\eta}_n}(\boldsymbol{\eta}_n)$ of the vector $\boldsymbol{\eta}_n$ is utilized to detect active users and estimate the timing and frequency offset, which can be written as

$$\begin{aligned} b_{\boldsymbol{\eta}_n}(\boldsymbol{\eta}_n) &= \frac{p(\boldsymbol{\eta}_n) \prod\limits_{m=1}^{M} I_{f_{mn}\to\boldsymbol{\eta}_n}(\boldsymbol{\eta}_n)}{Z_{\boldsymbol{\eta}_n}} \\ &= \frac{\alpha \sum\limits_{i=1}^{DQ} \delta\left(\boldsymbol{\eta}_n - \mathbf{e}_i\right) \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ\delta(\boldsymbol{\eta}_n)}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ}, \end{aligned} \tag{36}$$

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

$$\begin{aligned} b_{\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right) &= \frac{I_{f_{mn}\to\mathbf{h}_{mn}^e}\left(\mathbf{h}_{mn}^e\right) I_{\mathbf{h}_{mn}^e\to f_{mn}}\left(\mathbf{h}_{mn}^e\right)}{Z_{\mathbf{h}_{mn}^e}} \\ &= \frac{\alpha \sum\limits_{i=1}^{DQ} \mathcal{CN}\left(\mathbf{h}_{mn}^e(i); \frac{\sigma_h^2 \mu_{mni}^h}{\sigma_{mni}^h+\sigma_h^2}, \frac{\sigma_h^2 \sigma_{mni}^h}{\sigma_{mni}^h+\sigma_h^2}\right) \prod\limits_{i'\neq i} \delta\left(\mathbf{h}_{mn}^e(i')\right) \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ \delta\left(\mathbf{h}_{mn}^e\right)}{\alpha \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + (1-\alpha) DQ}. \end{aligned} \tag{31}$$

where $Z_{\boldsymbol{\eta}_n}$ is the normalization constant and can be expressed as

$$\begin{aligned} Z_{\boldsymbol{\eta}_n} &= \int p\left(\boldsymbol{\eta}_n\right) \prod_{m=1}^{M} I_{f_{mn}\to\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right) d\boldsymbol{\eta}_n \\ &= \frac{\frac{\alpha}{DQ} \sum\limits_{i=1}^{DQ} \prod\limits_{m=1}^{M} \lambda_{mni} + 1 - \alpha}{\prod\limits_{m=1}^{M} \left(\sum\limits_{i=1}^{DQ} \lambda_{mni} + 1\right)}. \end{aligned} \tag{37}$$

The estimated indicator variable $\hat{\xi}_n$, timing offset $\hat{\tau}_n$ and frequency offset $\hat{\varepsilon}_n$ of n-th user can be estimated through the distribution probability $b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$, which can be expressed as follows:

$$\hat{\xi}_n = \begin{cases} 0 & \text{, if } \int b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right) \delta\left(\boldsymbol{\eta}_n\right) d\boldsymbol{\eta}_n > T_{th} \\ 1 & \text{, otherwise.} \end{cases} \tag{38}$$

$$\hat{\tau}_n = \mathbf{d}\left(\lfloor \frac{\hat{i}_n - 1}{Q} \rfloor + 1\right), \tag{39}$$

$$\hat{\varepsilon}_n = \mathbf{q}\left(\hat{i}_n - \left(\hat{\tau}_n - 1\right) Q\right), \tag{40}$$

where T_{th} is the threshold used to determine user activity with its value compromising the miss detection rate and the false alarm rate, and the variable $\hat{i}_n$ is defined as

$$\hat{i}_n = \underset{i\in\{1,\cdots,DQ\}}{\arg\max} \int b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right) \delta\left(\boldsymbol{\eta}_n - \mathbf{e}_i\right) d\boldsymbol{\eta}_n. \tag{41}$$

Building upon the factor graph and the message expressions derived above, we come up with the S-GAMP algorithm, as shown in Algorithm 1, where $\mathbf{X}_e\left(l, (n-1)DQ + i\right)$ is expressed as x_{inl} for simplicity, μ_{ml}^r, σ_{ml}^r, μ_{ml}^s and σ_{ml}^s are the mean and variance of the variables $\mathbf{R}\left(l, m\right)$ and s_{ml}, respectively. During the iterations, the damping factor can be leveraged to prevent our algorithm from diverging [28], [41]. In Algorithm 1, lines 7-16 represent the GAMP algorithm, and lines 17-21 are expressions derived using BP algorithm rules.

Remark 1: The parallel AMP-MMV algorithm mentioned in [4], [27] is a particular case of our S-GAMP algorithm in the synchronous scenario. It means that the system model is no longer structured sparse, and the joint estimation problem degenerates into a canonical MMV problem.

Remark 2: The AMP decoder for sparse superposition coding proposed in [42], [43] is a special case of the S-GAMP algorithm when $\alpha = 1$, $M = 1$, $\mathbf{h}_n\left(m\right) = c$, and all complex Gaussian variables degenerate into real Gaussian variables, where c is a fixed constant used to constrain transmit power.

Algorithm 1: S-GAMP Algorithm

```
1  Input: $\mathbf{Y}$, $\mathbf{X}_e$, $\mathbf{d}$, $\mathbf{q}$, $\alpha$, $\sigma_h^2$, $\sigma_z^2$ and Number of
     iterations $T_{\max}$.
2  Output: Estimated channel matrix $\hat{\mathbf{H}}$, estimated user
     activity vector $\hat{\mathbf{u}}$, estimated user timing and frequency
     offset vector $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\varepsilon}}$, and $b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$.
3  Initialize:$\mathrm{E}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right] = 0$,
     $\mathrm{Var}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right] = \sigma_h^2$, $\forall i, m, n$.
4  for $t = 1, \cdots, T_{\max}$ do
5      for $m = 1, \cdots, M,\ l = 1, \cdots L$ do
6          $\sigma_{ml}^r = \sum_{n=1}^{N} \sum_{i=1}^{DQ} |x_{inl}|^2 \mathrm{Var}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right]$
7          $\mu_{ml}^r =$
             $-\mu_{ml}^s \sigma_{ml}^r + \sum_{n=1}^{N} \sum_{i=1}^{DQ} x_{inl} \mathrm{E}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right]$
8          $\mathrm{E}\left[\mathbf{R}(l,m) | b_{\mathbf{R}(l,m)}\right] = \frac{\mathbf{Y}(l,m)\sigma_{ml}^r + \mu_{ml}^r \sigma_z^2}{\sigma_{ml}^r + \sigma_z^2}$
9          $\mathrm{Var}\left[\mathbf{R}(l,m) | b_{\mathbf{R}(l,m)}\right] = \frac{\sigma_{ml}^r \sigma_z^2}{\sigma_{ml}^r + \sigma_z^2}$
10         $\mu_{ml}^s = \frac{\mathrm{E}\left[\mathbf{R}(l,m) | b_{\mathbf{R}(l,m)}(\mathbf{R}(l,m))\right] - \mu_{ml}^r}{\sigma_{ml}^r}$
11         $\sigma_{ml}^s = \frac{\sigma_{ml}^r - \mathrm{Var}\left[\mathbf{R}(l,m) | b_{\mathbf{R}(l,m)}(\mathbf{R}(l,m))\right]}{\left(\sigma_{ml}^r\right)^2}$
12     end
13     for $m = 1, \cdots, M,\ n = 1, \cdots, N,\ i = 1, \cdots, DQ$
         do
14         $\hat{h}_{mni} = \mathrm{E}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right]$
15         $\sigma_{mni}^h = \left(\sum_{l=1}^{L} \left(|x_{inl}|^2 \sigma_{ml}^s\right)\right)^{-1}$
16         $\mu_{mni}^h = \sigma_{mni}^h \sum_{l=1}^{L} x_{inl}^* \mu_{ml}^s + \hat{h}_{mni}$
17         Update $\mathrm{E}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right]$ and
             $\mathrm{Var}\left[\mathbf{h}_{mn}^e(i) | b_{\mathbf{h}_{mn}^e}\right]$ via (33) and (34)
18     end
19 end
20 for $n = 1, \cdots, N$ do
21     Calculate $b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$, $\hat{\xi}_n$, $\hat{\tau}_n$, $\hat{\varepsilon}_n$ and $\hat{i}_n$ via (36),
         (38), (39), (40) and (41).
22 end
23 for $m = 1, \cdots, M,\ n = 1, \cdots, N$ do
24     $\hat{\mathbf{H}}(n,m) = \hat{\xi}_n \mathrm{E}\left[\mathbf{h}_{mn}^e(\hat{i}_n) | b_{\mathbf{h}_{mn}^e}\right]$
25 end
26 $\hat{\mathbf{u}} = [\hat{\xi}_1, \cdots, \hat{\xi}_N]^T$, $\hat{\boldsymbol{\tau}} = [\hat{\tau}_1, \cdots, \hat{\tau}_N]^T$,
     $\hat{\boldsymbol{\varepsilon}} = [\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_N]^T$.
```

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

B. State Evolution Analysis

One significant feature of the AMP framework is that the state evolution can measure the per-iteration performance [4], [44]. Under large system limits, the expression of the variance of the variable to be estimated is exactly the state evolution function [4], and the smaller the value, the more accurate the estimated variable value is. For generality, we unify the variance σ_{mni}^{h} of the element $\mathbf{h}_{mn}^{e}\left(i\right), \forall m, n, i$, in the matrix $\mathbf{H}_e$ to be estimated into τ_h, and let $\tau_h^{(t)}$ represent the variance of the channel coefficient in the t-th iteration. In the asymptotic regime where $L, NDQ \rightarrow \infty$ with fixed ratio $\frac{NDQ}{L}$, the general state evolution can be expressed as

$$\tau_h^{(t+1)} = \sigma_z^2 + \frac{NDQ}{L}\,\mathrm{E}\{|f_\Theta(X+\sqrt{\tau_h^{(t)}}V,\Theta)-X|^2\}, \quad (42)$$

where the random variable X captures the distribution of the entries of the matrix $\mathbf{H}_e$, $V \sim \mathcal{CN}\left(0,1\right)$, and $\Theta = \left\{\{\boldsymbol{\eta}_n\}_{n=1}^{N}\right\}$ denotes the set of structured sparsity indicator vectors. The function $f_\Theta\left(\cdot,\Theta\right)$ is the denoiser with the information of Θ, and the expectation is taken over X, V and Θ. $\mathrm{E}\{|f_\Theta(X+\sqrt{\tau_h^{(t)}}V,\Theta)-X|^2\}$ characterizes the MSE of the denoiser at the t-th iteration. Note that for MMSE denoiser used in the GAMP algorithm, the MSE of the denoiser can be rewritten as

$$\mathrm{E}\{|f_\Theta(X+\sqrt{\tau_h^{(t)}}V,\Theta)-X|^2\} = \mathrm{E}\left\{\mathrm{Var}\left(X^{(t)} \mid \Phi^{(t)},\Theta\right)\right\}, \quad (43)$$

where $\Phi^{(t)} \triangleq X + \sqrt{\tau_h^{(t)}}V$. With the decomposition of variance [4], [40], we can get

$$\begin{aligned} &\mathrm{E}\left[\mathrm{Var}\left(X \mid \Phi^{(t)}\right)\right] \\ &= \mathrm{E}\left[\mathrm{Var}\left(X \mid \Phi^{(t)},\Theta\right)\right] + \mathrm{E}\left[\mathrm{Var}\left(\mathrm{E}\left[X \mid \Phi^{(t)},\Theta\right] \mid \Phi^{(t)}\right)\right] \\ &\geq \mathrm{E}\left[\mathrm{Var}\left(X \mid \Phi^{(t)},\Theta\right)\right], \end{aligned} \quad (44)$$

which shows that the MSE of the denoiser can be reduced by $\mathrm{E}\left[\mathrm{Var}\left(\mathrm{E}\left[X \mid \Phi^{(t)},\Theta\right] \mid \Phi^{(t)}\right)\right]$ with the knowledge of the structured sparsity indicator set Θ. Therefore, the introduction of the structured sparsity informtion Θ helps to improve the estimation performance in each iteration.

IV. The S-GAMP Algorithm with Dynamic Measurement Matrix

As mentioned in Section II-C, the high dimensionality of the expanded measurement matrix $\mathbf{X}_e$ incurs potentially high computational complexity. To overcome this problem, we propose to dynamically update the measurement matrix $\mathbf{X}_e$ to reduce the complexity.

The joint estimation process is divided into several cascaded steps, as shown in Fig.5. The receiver runs the S-GAMP algorithm in these steps, but the dynamic measurement matrix constructed in different steps is different. In each step, the BS selects a few timing and frequency offsets to construct the dynamic low-dimensional measurement matrix, which substantially reduces the computational complexity. The estimated timing and frequency offsets in each step are used to determine the construction of the dynamic measurement matrix in the next step to ensure accuracy.

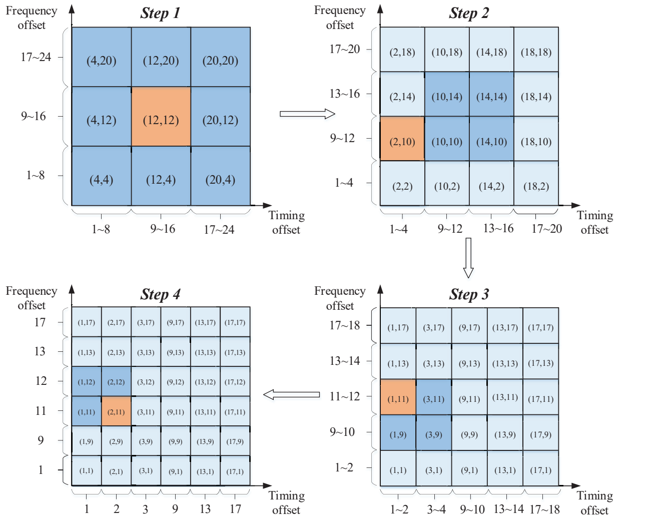


Fig. 5. The diagram of dynamically updating measurement matrix. The two-dimensional coordinates are used to represent the timing and frequency offset combinations. The dark blue area and the light blue area represent the regions where the timing offset and frequency offset are most likely and less likely, respectively. The orange area is obtained from equations (47) and (48), indicating the region where the timing and frequency offset are most likely among all the blue regions.

Take the n-th user as an example. All possible combinations of timing and frequency offsets are represented by two-dimensional coordinates (d, q), where d indicates the timing offset $\tau_n = \mathbf{d}\left(d\right)$ and q the frequency offset $\varepsilon_n = \mathbf{q}\left(q\right)$. Initially, the timing offset vector $\mathbf{d}$ and the frequency offset vector $\mathbf{q}$ are evenly divided into D_1 and Q_1 parts, respectively as shown in Fig. 5. Median values of these parts can constitute $\mathbf{d}_n^1 = \{\tau_{n,1}^1, \tau_{n,2}^1, \cdots, \tau_{n,D_1}^1\}$and $\mathbf{q}_n^1 = \{\varepsilon_{n,1}^1, \varepsilon_{n,2}^1, \cdots, \varepsilon_{n,Q_1}^1\}$, respectively. The D_1Q_1 coordinates composed of $\mathbf{d}_n^1$ and $\mathbf{q}_n^1$ can be used to represent these D_1Q_1 uniform regions.

Then, we use Algorithm 1 to estimate the region where the actual timing and frequency offsets are most likely to be. Assume that the dynamic timing and frequency offset vectors in the j-th step are $\mathbf{d}_n^j = \{\tau_{n,1}^j, \tau_{n,2}^j, \cdots, \tau_{n,D_j}^j\}$ and $\mathbf{q}_n^j = \{\varepsilon_{n,1}^j, \varepsilon_{n,2}^j, \cdots, \varepsilon_{n,Q_j}^j\}$, respectively. Thus, the dynamic measurement matrix $\mathbf{X}_e^j \in \mathbb{C}^{L \times (ND_jQ_j)}$ in the j-th step can be generated by the following equation:

$$\mathbf{X}_e^j = \left[\mathbf{X}_1^j, \mathbf{X}_2^j, \cdots, \mathbf{X}_N^j\right], \quad (45)$$

where the matrix $\mathbf{X}_n^j$ is given by

$$\mathbf{X}_n^j = \left[\mathbf{P}_{n,1,1}^j\mathbf{x}_n, \mathbf{P}_{n,1,2}^j\mathbf{x}_n, \cdots, \mathbf{P}_{n,D_j,Q_j}^j\mathbf{x}_n\right], \quad (46)$$

where $\mathbf{P}_{n,a,b}^j$ represents the phase shift matrix $\mathbf{P}_n$ with the timing and the frequency offsets being $\mathbf{d}_n^j\left(a\right)$ and $\mathbf{q}_n^j\left(b\right)$, respectively. By substituting the dynamic measurement matrix $\mathbf{X}_e^j$ into Algorithm 1 and replacing $\mathbf{d}$ and $\mathbf{q}$ with $\mathbf{d}_n^j$ and $\mathbf{q}_n^j$, we can estimate that the actual timing and frequency offsets are most likely to be in the region corresponding to the $k_{n,t}^j$-th timing offset part and the $k_{n,f}^{j}$-th frequency offset part, where $k_{n,t}^{j}$ and $k_{n,f}^{j}$ can be expressed as

$$k_{n,t}^{j} = \underset{k\in\{1,\cdots,D_j\}}{\arg\max} \int b_{\boldsymbol{\eta}_n}^{j}\left(\boldsymbol{\eta}_n\right) \sum_{i=(k-1)Q_j+1}^{kQ_j} \delta\left(\boldsymbol{\eta}_n - \mathbf{e}_i\right) d\boldsymbol{\eta}_n, \tag{47}$$

$$k_{n,f}^{j} = \underset{k\in\{1,\cdots,Q_j\}}{\arg\max} \int b_{\boldsymbol{\eta}_n}^{j}\left(\boldsymbol{\eta}_n\right) \sum_{i=1}^{D_j} \delta\left(\boldsymbol{\eta}_n - \mathbf{e}_{k+(i-1)Q_j}\right) d\boldsymbol{\eta}_n. \tag{48}$$

where $b_{\boldsymbol{\eta}_n}^{j}\left(\boldsymbol{\eta}_n\right)$ is the $b_{\boldsymbol{\eta}_n}\left(\boldsymbol{\eta}_n\right)$ output by algorithm 1 in the j-th step. As shown in Fig.5, the most likely regions in each step are shown in orange.

In addition to the initial timing offset vector $\mathbf{d}_n^1$ and frequency offset vector $\mathbf{q}_n^1$, the choice of timing offset and frequency offset vectors in other steps is also critical. Each timing offset part in the j-th step is further evenly divided into $(D_{j+1} - D_j + 1)$ smaller parts, and each frequency offset part in the j-th step is further evenly divided into $(Q_{j+1} - Q_j + 1)$ smaller parts. Take the timing offset as an example. The timing offset is most likely to be in the $k_{n,t}^{j}$-th part among the D_j parts in j-th step. Therefore its corresponding $(D_{j+1} - D_j + 1)$ smaller parts are all reserved for the $(j+1)$-th step. On the other hand, only one smaller part in the middle of each remaining unlikely $(D_j - 1)$ parts is kept to the $(j+1)$-th step. Therefore, we can get D_{j+1} smaller timing offset parts in the $(j+1)$-th step, with $(D_{j+1} - D_j + 1)$ smaller parts from the most likely part and $(D_j - 1)$ smaller parts from the $(D_j - 1)$ unlikely parts. Median values of these smaller parts are used to form dynamic timing offset vector $\mathbf{d}_n^{j+1}$. Similarly, the frequency offset vector $\mathbf{q}_n^{j+1}$ is generated in the same way.

Algorithm 2: Dynamic S-GAMP Algorithm

```
Input: the timing offset vector $\mathbf{d}$, frequency offset
   vector $\mathbf{q}$, and the number of steps $T_{\max}^{dyn}$.
Output: The output of Algorithm 1 when $j = T_{\max}^{dyn}$.
for $j = 1, \cdots, T_{\max}^{dyn}$ do
    if $j = 1$ then
        Generate $\mathbf{d}_n^1$ and $\mathbf{q}_n^1$ using the vectors $\mathbf{d}$ and $\mathbf{q}$,
          $n = 1, 2, \cdots, N$.
    else
        Update $\mathbf{d}_n^j$ and $\mathbf{q}_n^j$ according to $\mathbf{d}_n^{j-1}$, $\mathbf{q}_n^{j-1}$,
          $k_{n,t}^{j-1}$ and $k_{n,f}^{j-1}$, $n = 1, 2, \cdots, N$.
    end
    Generate the measurement matrix $\mathbf{X}_e^j$ via (45).
    Substitute matrix $\mathbf{X}_e^j$ into Algorithm 1, replace $\mathbf{d}$
      and $\mathbf{q}$ with $\mathbf{d}_n^j$ and $\mathbf{q}_n^j$, and calculate $k_{n,t}^{j}$ and
      $k_{n,f}^{j}$ via (47) and (48), $n = 1, 2, \cdots, N$.
end
```

Algorithm 2 summarizes the above dynamic programming process. Table III shows the computational complexity comparison of the proposed algorithms with other baseline algorithms. It is shown that the complexity of the S-GAMP algorithm is similar to that of the parallel AMP-MMV algorithm. Furthermore, compared with the S-GAMP and the parallel AMP-MMV algorithms, the dynamic S-GAMP algorithm has significant advantages in computational complexity. Notice that as the poor AUD performance of both the cross-correlation and the linear MMSE methods cannot meet the requirement in practical systems, as shown in Section V, it is of little significance to compare them with the proposed algorithms in terms of complexity.

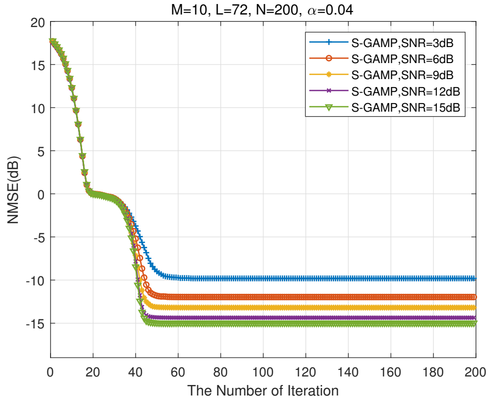


Fig. 6. The convergence of the S-GAMP algorithm under different SNR scenarios while $M = 10$ and $\alpha = 0.04$.

V. Simulation Results

In this section, we evaluate the performances of our proposed algorithms compared with parallel AMP-MMV algorithm, linear MMSE algorithm and cross-correlation algorithm, which are classical algorithms used for synchronization in OFDM system. The performance metrics are defined as follows:

- Channel Estimation Mean Square Error (MSE):

$$\mathrm{MSE} = 10\log_{10}\frac{\|\hat{\mathbf{H}} - \mathbf{H}\|_2^2}{\|\mathbf{H}\|_2^2}. \tag{49}$$

- User False Alarm Rate (UFAR) and User Detection Miss Rate (UDMR):

$$\mathrm{UFAR} = \frac{\|U\left(\hat{\mathbf{u}} - \mathbf{u}\right)\|_0}{N - N_a}, \quad \mathrm{UDMR} = \frac{\|U\left(\mathbf{u} - \hat{\mathbf{u}}\right)\|_0}{N_a}, \tag{50}$$

where $\mathbf{u} = \left[\xi_1, \xi_2, \cdots, \xi_N\right]^T$ is the user activity vector. The function $U(\cdot)$ is the step function, and the operation is componentwise. The UFAR and UDMR are contradictory and sensitive to the threshold, which means that the threshold can be adjusted to reduce the UFAR at the cost of increasing the UDMR, and vice versa. In practical systems, the choice of the threshold depends on the desired value of UDMR or UFAR. In our simulation, to fairly compare the active user detection performance among different algorithms, we adjusted the threshold values under different conditions to make their UFAR equal to 10^{-3} [45] and then compared their UDMR.

- Average timing offset estimation error (ATEE) and Average frequency offset estimation error (AFEE):

$$\mathrm{ATEE} = \frac{\mathbf{u}^T|\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}|}{N_a}, \quad \mathrm{AFEE} = \frac{\mathbf{u}^T|\boldsymbol{\varepsilon} - \hat{\boldsymbol{\varepsilon}}|}{N_a}, \tag{51}$$

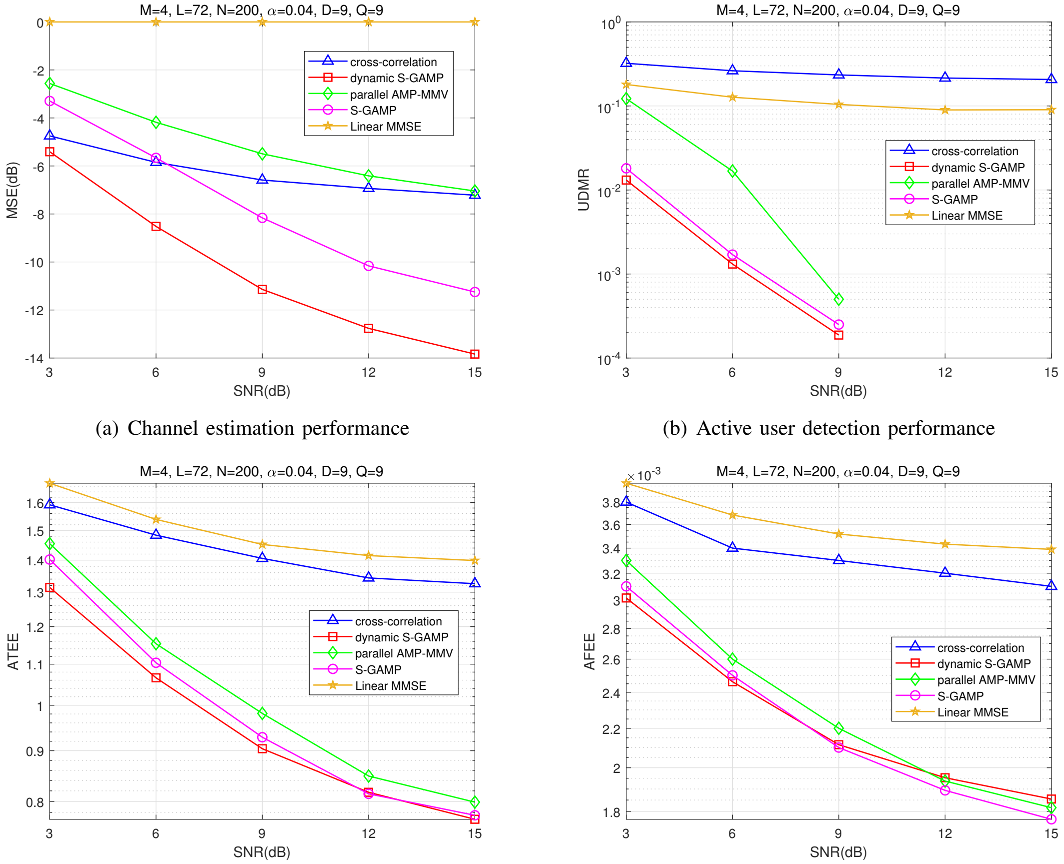
This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066

TABLE III
COMPUTATIONAL COMPLEXITY COMPARISON

	Multiplications	**Additions**
S-GAMP	$\mathcal{O}\left(T_{\max}M^2ND^2Q^2\right)$	$\mathcal{O}\left(T_{\max}M^2ND^2Q^2\right)$
dynamic S-GAMP	$\mathcal{O}(T_{\max}M^2N\sum_{j=1}^{T_{\max}^{dyn}}D_j^2Q_j^2)$	$\mathcal{O}(T_{\max}M^2N\sum_{j=1}^{T_{\max}^{dyn}}D_j^2Q_j^2)$
parallel AMP-MMV	$\mathcal{O}\left(T_{\max}MNDQL\right)$	$\mathcal{O}\left(T_{\max}MNDQL\right)$
cross-correlation	$\mathcal{O}\left(MNDQ(L+1)\right)$	$\mathcal{O}\left((LM-1)NDQ\right)$
Linear MMSE	$\mathcal{O}\left(NDQL^2\right)$	$\mathcal{O}\left(NDQL^2\right)$

TABLE IV
SYSTEM PARAMETERS

System Parameters	**Values**	**System Parameters**	**Values**
Number of subcarriers	$N_c = 2048$	Number of discrete frequency offset	$Q = 9$
Number of pilots	$L = 72$	Subcarrier index vector of pilots	$\mathbf{s} = [1, 3, 5, \cdots, 71]^T$
Number of users	$N = 200$	OFDM symbol index vector of pilots	$\mathbf{g} = [3, 12]^T$
Length of CP	$N_{CP} = 144$	Maximum frequency offset	$\varepsilon_{\max} = 0.0133$
Channel coefficient variance	$\sigma_h^2 = 1$	Number of the dynamic extraction steps	$T_{\max}^{dyn} = 2$
Maximum timing offset	$D = 9$	Dimensions of extraction in each step	$D_1 = Q_1 = 3, D_2 = Q_2 = 4$



(a) Channel estimation performance

(b) Active user detection performance

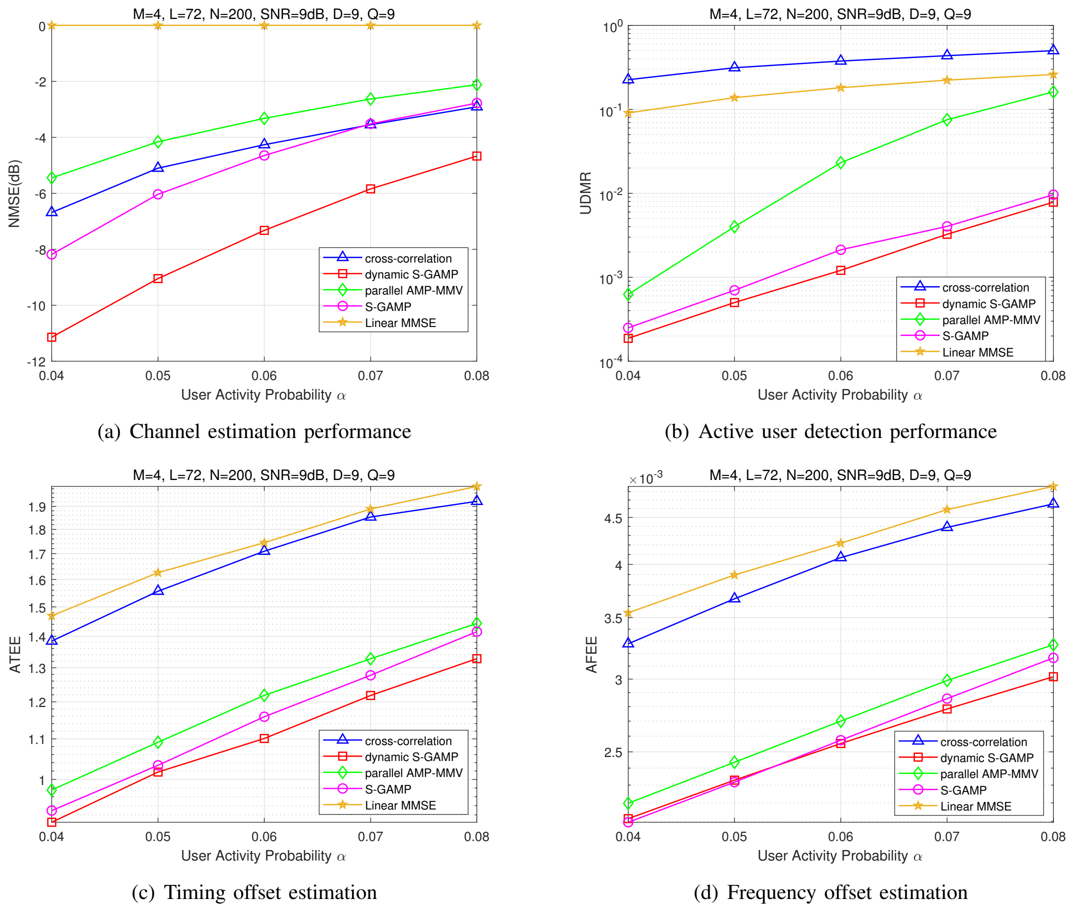
(c) Timing offset estimation

(d) Frequency offset estimation

Fig. 7. The performance comparison under different SNR scenarios while $\alpha = 0.04$ and $M = 4$.

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. 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/TWC.2021.3121066



(a) Channel estimation performance

(b) Active user detection performance

(c) Timing offset estimation

(d) Frequency offset estimation

Fig. 8. The performance comparison under different user activity probability while $\mathrm{SNR} = 9\,\mathrm{dB}, M = 4$.

where $\boldsymbol{\tau} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^{N\times 1}$ represent users' actual timing and frequency offset vectors, respectively.

The parameter settings of the following simulation scenarios are shown in Table IV. Besides, the signal-to-noise ratio (SNR) in the simulations is defined as follows:

$$\mathrm{SNR} = 10\log_{10}\frac{\|\mathbf{XH}\|_F^2}{LM\sigma_z^2}. \tag{52}$$

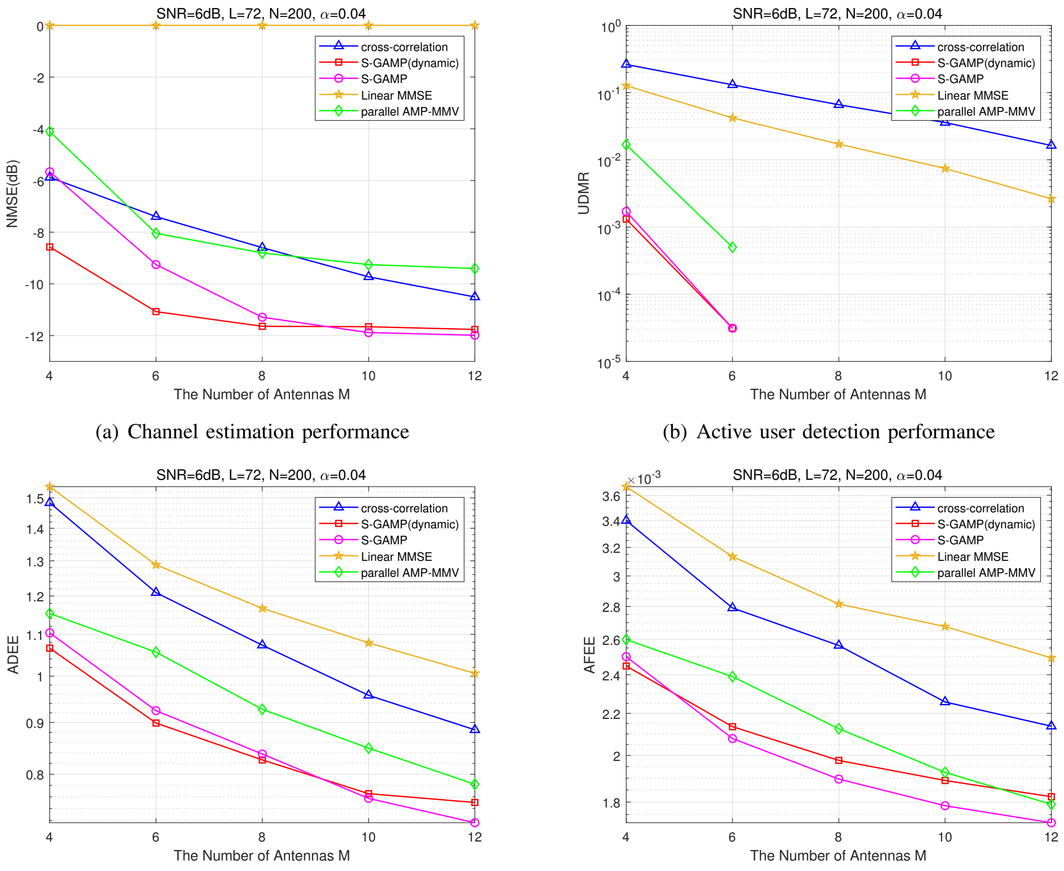
In Fig. 6, we simulated the CE performance of the S-GAMP algorithm in each iteration under different SNR to verify the convergence of the S-GAMP algorithm. It is shown that the performance of the S-GAMP algorithm is improved as the number of iterations increases until convergence.

In Fig. 7, we compare the CE, AUD, and offset estimation performance with SNR among these algorithms under $\alpha = 0.04, M = 4$. As shown in Fig. 7, the proposed dynamic S-GAMP and S-GAMP algorithms perform better joint estimation than other baseline algorithms because they make full use of the prior information of structured sparsity. Notice that the dynamic S-GAMP algorithm can perform better than the S-GAMP algorithm because the extracted measurement matrix is closer to the i.i.d. Gaussian matrix, which is required to guarantee the performance of the GAMP algorithm [4], [12], [20]. However, since the extracted measurement matrix is incomplete after all, the excellent performance of the dynamic S-GAMP algorithm cannot be guaranteed in all scenarios, such as the high SNR region, as shown in Fig. 7(c) and 7(d). Furthermore, the CE performance of the S-GAMP algorithm is slightly worse than the cross-correlation algorithm under low SNR in Fig. 7(a), because the performance of the S-GAMP algorithm degrades when the elements of the expanded measurement matrix are no longer i.i.d. Gaussian distributed, or the noise interference is high. It is worth mentioning that the poor AUD performance of the cross-correlation algorithm makes it unable to be applied in practical systems, resulting in its CE performance of little significance.

Fig. 8 illustrates the joint estimation performance comparison among these algorithms with different values of user activity probability under $\mathrm{SNR} = 9\,\mathrm{dB}, M = 4$. As shown in Fig. 8(a), the dynamic S-GAMP algorithm can achieve better CE performance than other algorithms in the same SNR regime. Furthermore, we can observe from Fig. 8(b) that the UDMR of the dynamic S-GAMP algorithm is slightly better than the S-GAMP algorithm, and their performance is far better than the other three algorithms. Fig. 8(c) and 8(d) show that the dynamic S-GAMP algorithm performs superior offset estimation in the same SNR regime. Although the joint estimation performance of these algorithms decreases with the increase of the user activity probability, the dynamic S-GAMP is still superior to others.

Fig. 9 provides the joint estimation performance comparison of the considered algorithms versus the number of antennas

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.



(a) Channel estimation performance

(b) Active user detection performance

(c) Timing offset estimation

(d) Frequency offset estimation

Fig. 9. The performance comparison under different numbers of antennas while $\alpha = 0.04, \mathrm{SNR} = 6\,\mathrm{dB}$.

under $\alpha = 0.04, \mathrm{SNR} = 6\,\mathrm{dB}$. It is shown that the proposed algorithms outperform other baseline algorithms, and their joint estimation performance becomes better as the increase of the number of antennas thanks to the more information of joint structured sparsity from the increasing SMV parts.

VI. Conclusion

This paper considered the massive grant-free transmission in the asynchronous OFDMA system and analyzed the impact of timing and frequency offsets on joint active user detection and channel estimation. To deal with the structured sparsity introduced by timing and frequency offsets, we proposed an efficient message passing algorithm (S-GAMP), leveraging the properties of the structured sparsity. In addition, we proposed the dynamic S-GAMP algorithm by updating the measurement matrix dynamically to reduce the computational complexity, which also improves the robustness. It is expected that the proposed S-GAMP algorithms could pave the way for the deployment of massive grant-free OFDMA in the mMTC scenarios.

References

[1] G. Sun, Y. Li, X. Yi, W. Wang, X. Gao, and L. Wang, "OFDMA based massive grant-free transmission in the presence of timing offset," in *Proc. 13th Int. Conf. Wireless Commun. Signal Process.*, Changsha, China, Oct. 2021.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, 2021.

[4] 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.

[5] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, 2017.

[6] 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.

[7] 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 Process. Mag.*, vol. 35, no. 5, pp. 88–99, 2018.

[8] Z. Zhang, X. Wang, Y. Zhang, and Y. Chen, "Grant-free rateless multiple access: A novel massive access scheme for Internet of Things," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2019–2022, 2016.

[9] J. Fu, G. Wu, Y. Zhang, L. Deng, and S. Fang, "Active user identification based on asynchronous sparse Bayesian learning with SVM," *IEEE Access*, vol. 7, pp. 108 116–108 124, 2019.

[10] 3GPP TS 38.211 V16.6.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; Physical channels and modulation (Release 16) ," Tech. Rep., June 2021.

[11] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," *IEEE Access*, vol. 3, pp. 195–208, 2015.

[12] J. Zhu, L. Han, and X. Meng, "An AMP-based low complexity gen-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3121066, IEEE Transactions on Wireless Communications

eralized sparse Bayesian learning algorithm," *IEEE Access*, vol. 7, pp. 7965–7976, 2019.

[13] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *J. Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[14] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[15] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.

[16] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[17] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun, pp. 211–244, 2001.

[18] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. Inf. Theory. Workshop*, 2010, pp. 1–5.

[19] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Gao, "Unifying message passing algorithms under the framework of constrained bethe free energy minimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4144–4158, 2021.

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

[21] Q. He, T. Q. S. Quek, Z. Chen, Q. Zhang, and S. Li, "Compressive channel estimation and multi-user detection in C-RAN with low-complexity methods," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3931–3944, 2018.

[22] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, 2018.

[23] X. Shao, X. Chen, D. W. K. Ng, C. Zhong, and Z. Zhang, "Cooperative activity detection: Sourced and unsourced massive random access paradigms," *IEEE Trans. Signal Process.*, vol. 68, pp. 6578–6593, 2020.

[24] 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, 2018.

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

[26] J. Kim, W. Chang, B. Jung, D. Baron, and J. C. Ye, "Belief propagation for joint sparse recovery," 2011. [Online]. Available: http://arxiv.org/abs/1102.3289

[27] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, 2013.

[28] 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.

[29] Y. Zhang, Z. Yuan, Q. Guo, Z. Wang, J. Xi, and Y. Li, "Bayesian receiver design for grant-free NOMA with message passing based structured signal estimation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8643–8656, 2020.

[30] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan, and L. Hanzo, "Iterative joint channel estimation, user activity tracking, and data detection for FTN-NOMA systems supporting random access," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2963–2977, 2020.

[31] S. Kim, H. Kim, H. Noh, Y. Kim, and D. Hong, "Novel transceiver architecture for an asynchronous grant-free IDMA system," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4491–4504, 2019.

[32] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, 2021.

[33] D. D. Lin, R. Pacheco, T. J. Lim, and D. Hatzinakos, "Joint estimation of channel response, frequency offset, and phase noise in OFDM," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3542–3554, 2006.

[34] J. Armstrong, "Analysis of new and existing methods of reducing intercarrier interference due to carrier frequency offset in OFDM," *IEEE Trans. Commun.*, vol. 47, no. 3, pp. 365–369, 1999.

[35] 3GPP TS 38.213 V16.6.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; Physical layer procedures for control (Release 16)," Tech. Rep., June 2021.

[36] H. C. Nguyen, E. de Carvalho, and R. Prasad, "Multi-user interference cancellation schemes for carrier frequency offset compensation in uplink OFDMA," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1164–1171, 2014.

[37] T. Yucek and H. Arslan, "Carrier frequency offset compensation with successive cancellation in uplink OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3546–3551, 2007.

[38] 3GPP TS 38.101-1 V16.5.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Tech. Rep., Sept 2020.

[39] 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.

[40] J.-C. Jiang and H.-M. Wang, "Massive random access with sporadic short packets: Joint active user detection and channel estimation via sequential message passing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4541–4555, 2021.

[41] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar, "On the convergence of approximate message passing with arbitrary matrices," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5339–5351, 2019.

[42] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity achieving sparse superposition codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4894–4927, 2017.

[43] J. Barbier, M. Dia, and N. Macris, "Universal sparse superposition codes with spatial coupling and GAMP decoding," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5618–5642, 2019.

[44] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[45] 3GPP TS 38.104 V16.8.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; Base Station (BS) radio transmission and reception (Release 16) ," Tech. Rep., June 2021.

Gangle Sun received the B.E. degree in communication engineering from Tianjin University, Tianjin, China, in 2019. He is currently pursuing the Ph.D. degree with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His research interests include wireless communication, signal processing and massive connectivity.

Yining Li received the B.E. and M.E. degree in information engineering from Southeast University, Nanjing, China, in 2018 and 2021, respectively. She is currently pursuing the Ph.D. degree at The Ohio State University. Her research interests focus on wireless communications and machine learning.

© 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Xinping Yi (Member, IEEE) received the Ph.D. degree in electronics and communications from Télécom ParisTech, Paris, France, in 2015. He is currently a Lecturer (Assistant Professor) with the Department of Electrical Engineering and Electronics, University of Liverpool, U.K. Prior to Liverpool, he was a Research Associate with Technische Universität Berlin, Berlin, Germany, from 2014 to 2017; a Research Assistant with EURECOM, Sophia Antipolis, France, from 2011 to 2014; and a Research Engineer with Huawei Technologies, Shenzhen, China, from 2009 to 2011. His main research interests include information theory, graph theory, machine learning, and their applications in wireless communications and artificial intelligence.

Wenjin Wang (Member, IEEE) received the Ph.D. degree in communication and information systems from Southeast University, Nanjing, China, in 2011. From 2010 to 2014, he was with the School of System Engineering, University of Reading, Reading, U.K. He is currently a Professor with the National Mobile Communications Research Laboratory, Southeast University. His research interests include advanced signal processing for future wireless communications and satellite communications. He was awarded a Best Paper Award at IEEE WCSP'09. He was also awarded the first grade Technological Invention Award of the State Education Ministry of China in 2009.

Xiqi Gao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 1997.

He joined the Department of Radio Engineering, Southeast University, in April 1992. Since May 2001, he has been a professor of information systems and communications. From September 1999 to August 2000, he was a visiting scholar at Massachusetts Institute of Technology, Cambridge, and Boston University, Boston, MA. From August 2007 to July 2008, he visited the Darmstadt University of Technology, Darmstadt, Germany, as a Humboldt scholar. His current research interests include broadband multicarrier communications, massive MIMO wireless communications, satellite communications, optical wireless communications, information theory and signal processing for wireless communications. From 2007 to 2012, he served as an Editor for the IEEE Transactions on Wireless Communications. From 2009 to 2013, he served as an Associate Editor for the IEEE Transactions on Signal Processing. From 2015 to 2017, he served as an Editor for the IEEE Transactions on Communications.

Dr. Gao received the Science and Technology Awards of the State Education Ministry of China in 1998, 2006 and 2009, the National Technological Invention Award of China in 2011, the Science and Technology Award of Jiangsu Province of China in 2014, and the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communications theory.

Lei Wang received the B.Sc. degree in communication engineering and the Ph.D. degree in communication and information systems from the Nanjing University of Aeronautics and Astronautics in 2005 and 2012, respectively. He is currently a Senior Researcher with Huawei Technologies Company Ltd., Shanghai, China. His research interests include wireless communication and signal processing.

Fan Wei received the B.S. degree from the Xi”an University of Posts and Telecommunications, Xi'an, China, in 2011, and the M.S. degree from Xidian University, Xi'an, in 2014, and the Ph.D. degree from the Network Coding and Transmission Laboratory, Shanghai Jiao Tong University, Shanghai, China, in 2019. He joined Huawei Technologies Co., Ltd., Shanghai, in 2019, where he has been working on various research topics on 5G and 6G air interface design. His research interests include wireless communication with focus on nonorthogonal multiple access schemes and grant-free transmission procedures.

Yan Chen (Member, IEEE) received the B.Sc. degree from the Chu Kochen Honored College, Zhejiang University, in 2004, and the Ph.D. degree from the Institute of Information and Communication Engineering, Zhejiang University, in 2009. She was a Visiting Researcher with The Hong Kong University of Science and Technology from 2008 to 2009. She joined Huawei Technologies Company Ltd., Shanghai, in 2009. From 2010 to 2013, she was the Project Leader of Huawei Internal Green Radio Project studying energy efficient solutions for wireless networks and the Project Leader of the umbrella GTT project in GreenTouch Consortium. Since 2013, she has been the Project Leader of Huawei research on multiple access, including NOMA, mission critical and massive IoT, and the related standardization in 3GPP. Her research interests include massive connectivity in IoT networks, NOMA transceiver design, grant-free transmissions, compressive sensing, and artificial intelligence enabled IoT communications. She won the IEEE Communication Society Award for Advances in Communication in 2017.