Generalized channel estimation and user detection for massive connectivity with mixed-ADC massive MIMO


Published in:
IEEE Transactions on Wireless Communications

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2019 IEEE. This work is made available online in accordance with the publisher’s policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:15. Mar. 2020
Generalized Channel Estimation and User Detection for Massive Connectivity With Mixed-ADC

Massive MIMO

Ting Liu, Shi Jin, Chao-Kai Wen, Michail Matthaiou, and Xiaohu You

Abstract—This paper aims to provide a partial-DFT pilot sequence assisted joint channel estimation and user activity detection scheme for massive connectivity, in which a large number of devices with sporadic transmission communicate with a base station (BS) in the uplink. The joint channel estimation and device detection problem can be formulated as a compressed sensing single measurement vector (SMV) or multiple measurement vector (MMV) problem depending on whether the BS is equipped with single or large number of antennas. Due to high hardware cost and power consumption in massive multiple-input multiple-output (MIMO) systems, a mixed analog-to-digital converter (ADC) architecture is considered. In order to accommodate a large number of simultaneously transmitting devices, the joint channel estimation and active user detection are formulated as a MMV problem for the massive connectivity scenario; and the proposed GTurbo-MMV algorithm can precisely estimate the channel state information (CSI) and detect active devices with relatively low overhead. Furthermore, we study the state evolution (SE) for the MMV problem to obtain achievable bounds on channel estimation and detection performance, in which both the missing and false detection probabilities can be made tend to zero in the massive MIMO regime. Simulation results confirm the theoretical accuracy of our analysis.

Index Terms—Channel estimation, GTurbo-MMV, mixed-ADC architecture, partial-DFT pilots, sporadic transmission, state evolution (SE), user activity detection.

I. INTRODUCTION

Massive machine-type communication (mMTC) is a key technology for 5G mobile communication networks [1], which is expected to support different applications, as shown in Fig. 1. In such a massive device connectivity scenario for 5G, the use of massive MIMO is required,1 in which a BS equipped with a large number of antennas can serve a massive number of devices. The massive connectivity of mMTC aims at supporting only a small fraction of active potential devices out of an extended number of devices to communicate with the base station (BS); a technology known as massive sporadic transmission [2–4]. For such a scheme, small packets may cause great burden by increasing the ratio of signaling overhead to useful payload under current media access control (MAC) protocols [5] where a user’s transmission commences upon the reception of uplink grant in the physical downlink control channel. Therefore, contention based uplink transmission has been introduced by using identical signature sequences for each active user [6], which remarkably reduces the significant overhead caused by the large number of sporadically active users. However, this technique is limited by the resolution of contention and the collision caused by the large number of online devices, and it is impractical to assign orthogonal pilot sequences to all potential terminal devices. Fortunately, grant-free transmission scheme with non-orthogonal pilot sequences has been introduced to avoid the overhead in signaling of activity, and reliable communication can be achieved after joint channel estimation and activity detection [7–9].

A. Related Work

There are many reported attempts on addressing massive sporadic transmission problems by using tools of compressed sensing (CS). In [8], Xu et al. considered the case when the BS has perfect knowledge of the large scale fading parameters of different online users and modified the Bayesian CS algorithm to detect active users and estimate their channel responses. By further exploiting the sparsity of user activity, [10] proposed a sparsity-based detector by maximizing the posterior prob-

Fig. 1. Massive device communication, where only a fraction of devices are active at each time slot.

1Massive MIMO networks are naturally suited to support mMTC, due to their: (a) unprecedented processing gains; (b) improved contention in resource access [4]; (c) effective improvement in the successful probability of active device detection [13].
ability with finite alphabet restrictions. Moreover, [11] performed joint information decoding without prior knowledge of channel state information (CSI) based on a greedy algorithm. Recently, in [12, 13], an approximate message passing (AMP) [14] based algorithm has been introduced to massive sporadic transmission where an independent identically distributed (i.i.d.) Gaussian pilot sequence was assigned to each user, when the BS is equipped with single antenna and multiple antennas, which transforms the problem to single measurement vector (SMV) based AMP and vector AMP for multiple measurement vector (MMV) (VAMP-MMV) [15], respectively. The statistical characteristics of channel estimation and user activity detection were studied via the so called state evolution (SE). From an information theory perspective, an overall achievable rate was characterized in closed-form with maximal-ratio combing (MRC) and minimum mean square error (MMSE) beamforming at the BS for massive connectivity with massive MIMO in [16].

A common limitation of the massive MTC (mMTC) with massive MIMO is that a large number of RF chains used for antennas at the BS would significantly complicate the hardware design for massive connectivity. Such systems require an ADC unit for each receiver antenna. However, the implementation cost and power consumption are exponentially increased by using high-resolution ADCs. Nevertheless, this problem can be resolved by using low-resolution ADCs (e.g., 1-3 bits) to replace high-resolution ADCs, resulting quantized massive MIMO systems [17–19]. Unfortunately, low-resolution ADC systems face formidable challenges, such as data rate loss, error floor, and high pilot overhead caused by the nonlinear characteristics of the low-resolution quantization [20–22]. Motivated by the above concerns, a mixed-ADC architecture was proposed for massive MIMO in [23, 24], where a few number of antennas are equipped with high-resolution ADCs while others are connected with low-resolution ADCs. Particularly, the reserved high-resolution ADCs can be utilized to perform accurate channel estimation in a round-robin manner with reduced-complexity algorithm and affordable pilot overhead. In [25], the mixed-ADC architecture was proposed that can help maintain the promised detection performance while significantly reducing the computational complexity. Moreover, this architecture has inherent practical relevance for massive MIMO, while still offering a large proportion of the channel capacity of the ideal ADC architecture. The spectral and energy efficiency of the mixed-ADC framework have been analyzed in [26], showing a better balance between energy consumption and throughput compared to the pure low-resolution and ideal high-resolution ADC architectures. In this process, the mixed-ADC architecture is practically useful because it provides comparable performance with the ideal infinite-resolution ADC topology, while the signal processing complexity and power consumption can be reduced at the same time. On these grounds, the Bayesian optimal data detector in hybrid broadband mmWave systems with mixed-ADC architecture was investigated in [27]. We also note that the computational complexity of massive connectivity regime is still an open problem. This paper shows that adopting the discrete Fourier transformation (DFT) pilot matrix as a pilot sequence pool can significantly reduce the computational complexity of massive MIMO supporting massive connectivity.

B. Main Contributions

This paper considers the uplink of a single-cell massive connectivity scenario, where a plethora of online users or devices are served by multiple antennas at the BS. We emphasize that massive MIMO has been widely studied in the literature [28–30]. However, the biggest body of related research is focused on the regime of small number of online users as compared to the number of BS antennas. In contrast, this paper articulates that the channel estimation and user detection can be jointly achieved under the mixed-ADC architecture, when a massive number (larger than the number of BS antennas) of devices is deployed. Due to the sporadic characteristic of massive connectivity, principles and techniques from the CS literature [31] can be leveraged for activity detection and channel estimation. In our previous study, the generalized turbo signal recovery algorithm (GTurbo-SR) was applied in solving CS problems [32] under non-linear measurements with a partial DFT pilot matrix. The partial DFT pilot matrix not only reduces the hardware cost and implementation complexity, but also provides stronger guarantee of convergence compared to the AMP algorithm [33–35]. Our specific contributions are the following:

1. For the case where the BS is equipped with a single antenna, we first propose the GTurbo-SMV algorithm to investigate the signal detection for the pure low-resolution measurement and analyze the scalar-wise SE [32]. To utilize the extra degrees of freedom (DoF) provided by multiple antennas, we extend the GTurbo-SMV to GTurbo-MMV. By fully exploiting the joint sparsity property of the channel response, the GTurbo-MMV algorithm can be proved to be capable of identifying active users with enhanced accuracy. We also analyze the SE for GTurbo-MMV in the large system regime, which provides valuable insights into the channel estimation of our algorithm, and further asymptotic activity detection accuracy conclusion is derived.

2. A mixed ADC architecture can assist in channel estimation and active device detection due to the available high-resolution ADCs at the existing BS. What is more, a mixed ADC architecture is economically beneficial because it can be implemented by deploying some low-resolution ADCs at the BS. In this paper, GTurbo-MMV can be shown to be capable of achieving strong performance and strong robustness in solving user detection and channel estimation problems with a mixed-ADC architecture. On the other hand, we also derive theoretical bounds on the channel estimation mean square error (MSE) by using our previous SE analysis.

3. Instead of using the traditional i.i.d. Gaussian sequences (non-orthogonal sequences) or some other orthogonal sequences as pilots, a row-orthogonal matrix, e.g., a partial DFT pilot matrix, is allocated to different online users to reduce the cost of pilot overhead. Most importantly, the DFT pilot matrix is not required to be stored, which reduces the buffering requirements to a large extent. Moreover, with the implementation of the Cooley-Tukey FFT algorithm, channel
estimation and user activity detection can be implemented with low complexity.

The rest of this paper is organized as follows. The problem under investigation is formulated in Section II, along with the pilot allocation scheme. In Section III, we first introduce the GTurbo-MMV algorithm for user activity detection and channel estimation, and then use our SE analysis to derive several propositions and theorems. Section IV presents the numerical results of our proposed algorithm. The paper concludes in Section V.

Notations: A scalar variable is denoted by lowercase normal letter \(a\). Vectors and matrices are denoted by bold lowercase \(\mathbf{a}\) and uppercase letters \(\mathbf{A}\), respectively. A \(K \times K\) identity matrix is denoted as \(\mathbf{I}_K\). For a given vector \(\mathbf{a}\), its conjugate transpose is denoted as \(\mathbf{a}^\dagger\), and \(\mathbf{a}^T\) denotes its transpose; \(\|\mathbf{a}\|_2\) denotes the Euclidean norm of vector \(\mathbf{a}\); \(\|\mathbf{A}\|_F\) is the Frobenius norm of matrix \(\mathbf{A}\). We let \(\mathbf{a}_i\) denote the \(i\)-th column of \(\mathbf{A}\), \(\mathbf{a}[i]\) denotes the \(i\)-th row of \(\mathbf{A}\) while \(\delta_{i,j}\) represents the \((i,j)\) entity of \(\mathbf{A}\). For a given square matrix \(\mathbf{A}\), \(\text{diag}(\mathbf{A})\) would get all elements in its diagonal and form a new vector in original order. For a vector \(\mathbf{a}\), \(\text{diag}(\mathbf{a})\) would construct a diagonal matrix with \(\mathbf{a}\) as the diagonal elements. The italic \(I\) denotes the set and \(|I|\) is its cardinality. A random vector \(\mathbf{x}\) drawn from the complex Gaussian distribution of mean \(\mathbf{a}\) and covariance \(\mathbf{A}\) is denoted as \(CN(\mathbf{x}; \mathbf{a}, \mathbf{A})\) with probability density function \(e^{-\frac{1}{2}(\mathbf{x} - \mathbf{a})^\dagger \mathbf{A}^{-1}(\mathbf{x} - \mathbf{a})}/\det(\pi\mathbf{A})\). The discrete set with \(B\)-bit quantizer is \(\mathcal{R}_B\), and the set of natural numbers without zero is \(\mathbb{N}^+\).

II. SYSTEM MODEL

We consider the uplink massive sporadic transmission of a single-cell wireless cellular system, as shown in Fig. 2, which consists of \(N\) online devices with single-antenna and the one BS has \(M_R\) antennas. Under the contention-based transmission scheme, user terminals are independent to each other and across each frame, and assumed to have the same probability \(p_a\) to transmit data. An indicator function \(\beta_n\) is used to indicate the activity state of online user \(n\), i.e., \(P(\beta_n = 1) = p_a\), and \(P(\beta_n = 0) = 1 - p_a\). Therefore, in massive sporadic transmission where \(p_a \ll 1, N \gg 1\), the number of active users within a frame is a Poisson random variable \(N_a\) (\(N_a = \sum_{n=1}^{N} \beta_n\)). To efficiently recover the information transmitted by the active users, pilot-based channel estimation and active user detection have to be executed within each frame.

Define \(s_n \in \mathbb{C}^L\) to be the pilot sequence of user \(n\), where \(L\) is the length of pilot sequence. Each pilot sequence \(s_n\) is extracted from the sequence pool \(\mathcal{S}\),

\[
\mathcal{S} = \mathcal{Y} \mathcal{F},
\]

where \(\mathcal{S} = [s_1, s_2, \ldots, s_N] \in \mathbb{C}^{L \times N}\), \(\mathcal{Y} \in \mathbb{C}^{L \times M_R}\) is a row-selection matrix with measurement ratio \(\alpha = L/N\), and \(\mathcal{F} \in \mathbb{C}^{N \times N}\) is a unitary DFT matrix.

Then, the received signal at the BS can be given by

\[
\mathbf{Y}^T = \mathbf{S} \mathbf{H}^T + \mathbf{Z}^T,
\]

where \(\mathbf{Y}^T = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{M_R}] \in \mathbb{C}^{L \times M_R}\), \(\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]\) with \(\mathbf{h}_n \in \mathbb{C}^{M_R}\) representing the channel response of the \(n\)-th user. In this paper, we assume that a block-fading is adopted, and all channels follow the independent flat fading within a coherence block, where \(\mathbf{h}_n\)‘s remain constant, but vary independently from block to block; \(\mathbf{Z} \in \mathbb{C}^{M_R \times L}\) denotes the complex white Gaussian noise with element-wise variance \(\sigma^2\). Denote \(I_{N_a}\) as the active users’ index, which is the subset of online user set \(I_N = \{1, 2, \ldots, N\}\) with \(|I_{N_a}| = N_a\), i.e., \(\mathbf{h}_n = \mathbf{0}\) for \(n \notin I_{N_a}\). We assume that \(\mathbf{h}_n\) is i.i.d., and can be seen as drawn from the joint Bernoulli-Gaussian (BG) distribution

\[
P(\mathbf{h}_n) = (1 - p_a)\delta(\mathbf{h}_n) + p_a \mathcal{CN}(\mathbf{h}_n; \mathbf{0}, \mu_n \mathbf{I}),
\]

where \(\delta(\cdot)\) represents the Dirac delta for a vector, and \(\mu_n\) is the large-scale fading factor of user \(n\). The large-scale components are assumed to be known at the BS.

By taking the distribution of \(\mathbf{h}_n\) in (3), \(\mathbf{H}^T\) can be characterized with joint sparsity, which makes (2) a MMV problem within the field of CS [31]. To further simplify our system model, we denote

\[
\mathcal{E} = \mathbf{F}^H, \quad (4)
\]

where \(\mathcal{E} = [\xi_1, \xi_2, \ldots, \xi_{M_R}]\). Then, Eq. (2) can be equivalently expressed as

\[
\mathbf{Y}^T = \mathbf{Y} \mathcal{E} + \mathbf{Z}^T. \quad (5)
\]

III. CHANNEL ESTIMATION AND USER DETECTION

In this section, we first discuss the GTurbo algorithm for SMV. Based on the joint sparsity of \(\mathbf{H}^T\) in (2), we then propose a GTurbo algorithm for MMV (GTurbo-MMV), which iterates the extrinsic information for each antenna and combines them within the joint estimator of \(\mathbf{H}^T\), to estimate the channel responses and detect user activity.

A. Bayesian Detector Formulation

A main point of this paper is to exploit the massive connectivity by adopting the Bayesian inference under a mixed-ADC massive MIMO architecture. We assume that some of the BS antennas adopt low-resolution ADCs (e.g., 1-3 bits) while the remaining BS antennas adopt high-resolution ADCs (i.e., infinite-resolution ADCs). The complex-valued
quantizer \( Q(.) \) consists of two real-valued quantizers \( Q \), and the resulting quantized signal is expressed as

\[
\tilde{Y} = Q_c(Y_T^T) \triangleq Q(Y_R^T) + jQ(Y_I^T),
\]

where \( Y_T^T \) and \( Y_I^T \) are the real and imaginary part of \( Y_T \) respectively.

The quantized output of each element \( \tilde{y}_{m,l} \) is assigned the value \( \tilde{y}_{m,l} = \tilde{Y}_{R,(m,l)} + j\tilde{Y}_{I,(m,l)} \) by judging whether each part, \( \tilde{Y}_{R,(m,l)} \) or \( \tilde{Y}_{I,(m,l)} \), falls into the interval \( [\tilde{y}_{low}^{R(m,l)}, \tilde{y}_{up}^{R(m,l)}] \) or \( [\tilde{y}_{low}^{I(m,l)}, \tilde{y}_{up}^{I(m,l)}] \). For a typical uniform B-bit quantizer with quantization step size \( \Delta \), the quantized output is given by

\[
\tilde{y}_{R,(m,l)}(\tilde{y}_{m,l}, \tilde{Y}_T) \in \left\{ \left( -\frac{1}{2} + b \right) \Delta; b = -\frac{2^B}{2} + 1, \ldots, \frac{2^B}{2} \right\} \triangleq \mathcal{R}_B,
\]

where \( m = 1, 2, \ldots, M_R, l = 1, 2, \ldots, L \), and \( B \) is the number of quantization bits. The associated lower and upper thresholds of \( \tilde{Y}_{R,(m,l)} \) are respectively expressed as

\[
\begin{align*}
\tilde{y}_{low}^{R(m,l)} &= \tilde{Y}_{R,(m,l)} - \frac{\Delta}{2}, & \text{for } \tilde{Y}_{R,(m,l)} \geq -\left( \frac{2^B}{2} - 1 \right) \Delta, \\
\tilde{y}_{up}^{R(m,l)} &= \tilde{Y}_{R,(m,l)} + \frac{\Delta}{2}, & \text{for } \tilde{Y}_{R,(m,l)} \leq \left( \frac{2^B}{2} - 1 \right) \Delta, \\
\tilde{y}_{low}^{I(m,l)} &= \tilde{Y}_{I,(m,l)} - \frac{\Delta}{2}, & \text{for } \tilde{Y}_{I,(m,l)} \geq -\left( \frac{2^B}{2} - 1 \right) \Delta, \\
\tilde{y}_{up}^{I(m,l)} &= \tilde{Y}_{I,(m,l)} + \frac{\Delta}{2}, & \text{for } \tilde{Y}_{I,(m,l)} \leq \left( \frac{2^B}{2} - 1 \right) \Delta.
\end{align*}
\]

For linear measurements in (5) and nonlinear (quantized) measurements in (6), we both consider the Bayesian optimal inference to estimate \( H_T \) and \( \Xi \) from \( Y_T \) or \( \tilde{Y} \) in terms of MSE. Toward this end, the posterior joint distribution under nonlinear measurements can be divided into

\[
P(\tilde{Y} | \Xi) = \prod_{m=1}^{M_R} \prod_{l=1}^{L} P(\tilde{y}_{m,l} | \xi_{\Theta(l)}(m)),
\]

where \( \Theta(l) \) is a mapping function to map the selected row index to the original row number representing as the consequence of \( Y \). The right hand side of (9) results from the product of real and imaginary counterparts is given by

\[
P(\tilde{y}_{m,l} | \xi_{\Theta(l)}(m)) = \frac{1}{\sqrt{\pi \sigma^2}} \int_{\tilde{y}_{low}^{up}(m,l)}^{\tilde{y}_{up}^{R}(m,l)} e^{-\frac{(y - \xi_R(\Theta(l),m))^2}{\sigma^2}} dy \times \frac{1}{\sqrt{\pi \sigma^2}} \int_{\tilde{y}_{low}^{up}(m,l)}^{\tilde{y}_{up}^{I}(m,l)} e^{-\frac{(y - \xi_I(\Theta(l),m))^2}{\sigma^2}} dx,
\]

where \( \tilde{y}_{low}^{R(m,l)} \in (\tilde{y}_{low}^{R(m,l)}, \tilde{y}_{up}^{R(m,l)}) \) and \( \tilde{y}_{up}^{I(m,l)} \in (\tilde{y}_{low}^{I(m,l)}, \tilde{y}_{up}^{I(m,l)}) \). According to the cumulative Gaussian distribution function, we obtain

\[
P(\tilde{y}_{m,l} | \xi_{\Theta(l)}(m)) = \Psi \left( \frac{\tilde{y}_{R,(m,l)} - \xi_R(\Theta(l),m)}{\sigma} \right) \times \Psi \left( \frac{\tilde{y}_{I,(m,l)} - \xi_I(\Theta(l),m)}{\sigma} \right),
\]

with

\[
\Psi \left( \frac{\tilde{y}_{up} - \xi}{\sigma} \right) \triangleq \Phi \left( \frac{\tilde{y}_{up} - \xi}{\sigma} \right) - \Phi \left( \frac{\tilde{y}_{low} - \xi}{\sigma} \right),
\]

where \( \Phi(\cdot) \) denotes the cumulative Gaussian distribution function.

The Bayesian estimation of \( H_T \) is a CS problem due to the joint sparsity of channel response. Further, as the sparsity pattern is sensed at BS with multiple antennas, this becomes a MMV compressed sensing setup. In our conference paper [32], an innovative GTurbo signal recovery (GTurbo-SR) algorithm was proposed as an iterative scheme to recover CSI from quantized measurements. We will briefly outline the GTurbo-SR algorithm and extend to the MMV scenario with mixed-ADC architectures in the following subsections.

B. GTurbo-SMV

GTurbo-SR algorithm was proposed by [32] as an iterative algorithm for CS. We first introduce the GTurbo algorithm for the SMV problem. The basic idea in using GTurbo algorithm is to perform signal recovery based on the turbo principle in iterative decoding within each iteration, which is shown in Fig. 3, where \( \gamma, \tilde{y}, \xi \) and \( h \) are the vector representations of \( Y, \tilde{Y}, \Xi \), and \( H \) when the BS has only one antenna. During each iteration of Algorithm 1, module A estimates \( \xi \) based on the observation of \( \gamma \) or \( \tilde{y} \) through posterior estimation, while module B estimates the channel response \( h_{\text{post}} \) in the same manner. According to the turbo principle, the extrinsic output of one module is fed back to the other module as the priori input and the two modules are executed iteratively until they meet the convergence criteria. Finally, in the last iteration, the output of module B is considered as the channel estimation of \( h \). The expectations and variances of lines 1 and 2, as well as lines 8 and 9 are taken with respect to (w.r.t.) the following posterior probabilities,

\[
\begin{align*}
\mathbb{E}[\xi] &= \frac{P(\xi_n | y_n) P(\xi_n)}{P(y_n | \xi_n) P(\xi_n)}, & \text{linear}, \\
\mathbb{E}[\xi] &= \frac{P(\xi_n | y_n) P(\xi_n)}{P(\xi_n | y_n)}, & \text{nonlinear}, \\
\mathbb{E}[h_{\text{post}}] &= \frac{\mathbb{E}[\xi_A] P(\xi_n | y_n) P(h_{\text{post}} | \xi_A) P(h_{\text{post}})}{\mathbb{E}[\xi_A]}, & \text{MMV}.
\end{align*}
\]

where \( \mathbb{E}[\xi] = \mathbb{E}[\xi | y_n], P(\xi_n | y_n) = \mathbb{E}[\xi_A] P(\xi_n | y_n), P(h_{\text{post}} | \xi_A) P(h_{\text{post}})dh_{\text{post}} \), please refer to [32] for the explicit expressions of lines 1 and 2, lines 8 and 9.

C. GTurbo-MMV for massive connectivity with mixed-ADC architectures

We now extend the GTurbo-SMV algorithm to the GTurbo-MMV algorithm for the massive connectivity problem in the multiple-antenna case with mixed-ADC topologies. Compared with SMV, the major difference of the MMV problem in (2) mainly comes from the joint sparse property of \( H_T \). To apply the GTurbo scheme in the MMV scenario, we first derive the

\[
\begin{align*}
&\Psi \left( \frac{\tilde{y}_{up} - \xi}{\sigma} \right) \triangleq \Phi \left( \frac{\tilde{y}_{up} - \xi}{\sigma} \right) - \Phi \left( \frac{\tilde{y}_{low} - \xi}{\sigma} \right),
\end{align*}
\]

where \( \Phi(\cdot) \) denotes the cumulative Gaussian distribution function.
Algorithm 1: GTurbo-SMV

**Input:** Observations $y$ or $\hat{y}$, pilot matrix $F$, likelihood $P(y|\xi)$ or $P(\hat{y}|\xi)$, and prior distributions $P(h)$

**Output:** Recovered channel $h$

**Initialize:** Current iteration $t \leftarrow 1$, $\xi^{\text{pri}}_A \leftarrow 0$, $\nu^{\text{pri}}_A \leftarrow E[|h|^2]$

while $t \leq T$

1) **Output steps:**
Compute the posterior mean and variance of $\xi$


\[
\xi^{\text{post}}_A \leftarrow E[\xi | \xi^{\text{post}}, \nu^{\text{post}}_A];
\]

\[
\nu^{\text{post}}_A \leftarrow \text{Var}[\xi | \xi^{\text{post}}, \nu^{\text{post}}_A];
\]

2) **Estimate $h$ via $F\xi$ and its corresponding variance**


\[
h^{\text{post}}_A \leftarrow F^\dagger \xi^{\text{post}};
\]

\[
\nu^{\text{post}}_A \leftarrow \frac{1}{N} \sum_{n=1}^N \nu^{\text{post}}_{a,n};
\]

Compute the extrinsic mean and variance of $h$


\[
h^{\text{pri}}_A \leftarrow E[h | h^{\text{post}}_A];
\]

\[
v^{\text{pri}}_A \leftarrow \frac{1}{N} \sum_{n=1}^N v^{\text{post}}_{h,n};
\]

3) **Input steps:**

Compute the posterior mean and variance of $h$


\[
h^{\text{post}}_A \leftarrow E[h | h^{\text{post}}_B, v^{\text{post}}_B];
\]

\[
\nu^{\text{post}}_A \leftarrow \text{Var}[h | h^{\text{post}}_B, v^{\text{post}}_B];
\]

4) **Estimate $\xi$ via $Fh$ and its corresponding variance**


\[
\xi^{\text{post}}_B \leftarrow Fh^{\text{post}};
\]

\[
\nu^{\text{post}}_B \leftarrow \frac{1}{N} \sum_{n=1}^N \nu^{\text{post}}_{h,n};
\]

Compute the extrinsic mean and variance of $\xi$


\[
\xi^{\text{pri}}_B \leftarrow Fh^{\text{pri}};
\]

\[
v^{\text{pri}}_B \leftarrow \frac{1}{N} \sum_{n=1}^N v^{\text{ext}}_{h,n};
\]

\[
\xi^{\text{ext}}_B \leftarrow \xi^{\text{pri}}_B - \xi^{\text{ext}}_B (\xi^{\text{ext}}_B^\dagger v^{\text{ext}}_B - \xi^{\text{ext}}_B v^{\text{ext}}_B);\]

Update for the next iteration:

$t \leftarrow t + 1$

Proposition 1: The explicit expressions of the posterior mean and variance as well as the corresponding $\text{mse}$ of model (14) are respectively given by

\[
E[h | r] = \frac{P_{\text{a}}}{P(r)}CN(r; 0, \mu I + V)\left(I + \frac{1}{\mu}V\right)^{-1}r,
\]

\[
\text{Var}[h | r] = \frac{P_{\text{a}}}{P(r)}CN(r; 0, \mu I + V)[(AV^{-1}r)(AV^{-1}r)^H + \Lambda]
\]

\[
- E[h | r] E[h | r]^H
\]

\[
\text{mse} = p_a \mu I - p_a^2 (AV^{-1})
\]

\[
\times \left[\int p_a + (1 - p_a)CN(r; 0, V)/CN(r; 0, \mu I + V) dr\right]
\]

\[
\times (AV^{-1})^H,
\]

where $\Lambda = (\mu^{-1}1 + (V^{-1})^{-1}$, $P(r) = (1 - p_a)CN(r; 0, V)$, and $\text{mse} \triangleq E[\text{Var}[h | r]]$.

Proof: Please refer to Appendix A.

From Proposition 1, we can then construct the GTurbo algorithm for MMV as in Fig. 4. The overall workflow of GTurbo-MMV is to make a row-vector-wise MMSE estimation of $H$, and maximum posterior probability (MAP) estimation of active user set $I_{\text{active}}$ in module B within each iteration. Given the $M_R$ pairs of extrinsic information $H^{\text{ext}}_{B \rightarrow A}$, $v^{\text{ext}}_{h,B}$ that are calculated in parallel on each antenna in module A, module B then estimates the MMSE of $H$, and then $H^{\text{ext}}_{B \rightarrow A}$, $v^{\text{ext}}_{h,B}$ are fed back to module A as the priori inputs, which completes the entire loop of one iteration. The reason why GTurbo-MMV will outperform $M_R$ parallel GTurbo-SMV is that GTurbo-MMV takes full knowledge of the joint sparsity of $H^T$, and then provides additional information through the posterior estimation. The proposed GTurbo-MMV is summarized in Algorithm 2, and we omit the superscript $t$ at each iteration for brevity in this subsection.

In Algorithm 2, step 1) contains $M_R$ parallel running stages. Each stage computes one pair of extrinsic information, $h^{\text{ext}}_{B \rightarrow A}$, $v^{\text{ext}}_{h,B}$, in the manner of calculation of lines 1 to 7 in Algorithm 1. Subsequently, the channel estimation part in module B computes the posterior mean and variance. The explicit expressions of MMSE estimation $h^{\text{post}}_{B,n}$ and its corresponding variance $v^{\text{post}}_{h,B,n}$ in lines 2-3 of Algorithm 2 are directly taken from (16) in Proposition 1 and can be
evaluates the extrinsic information of $H$. After the processing of module B, step 3) of Algorithm 2

Consequently, when the unknown matrix is jointly sparse. In this section, SE analysis can be extended to MMV case

given as

$$
E \left[ h_B^\text{pri}, v_B^\text{pri} \right] = \frac{p_n}{P(h_B^\text{pri})} \mathcal{CN} \left( h_B^\text{pri}, 0, \mu_n I + V^\text{pri} \right) 
\times \left( I + \frac{1}{\mu_n} V^\text{pri} \right)^{-1} h_B^\text{pri},
$$

where $V^\text{pri} = \text{diag}(v_B^\text{pri})$, $P(h_B^\text{pri}) = (1 - p_n)CN(0; h_B^\text{pri}, V^\text{pri}) + p_n CN(h_B^\text{pri}, 0, \mu_n I + V^\text{pri})$.

In line 4 of Algorithm 2, we employ the MAP method to detect the user activity. The activity of user $n$, i.e., $\beta_n$, is inferred from the ratio between $P(\beta_n = 1|h_B^\text{pri})$ and $P(\beta_n = 0|h_B^\text{pri})$, that is,

$$
P(\beta_n = 1|h_B^\text{pri}) = \frac{P(h_B^\text{pri}|\beta_n = 1)P(\beta_n = 1)}{P(h_B^\text{pri}|\beta_n = 0)P(\beta_n = 0)},
$$

where $P(h_B^\text{pri}|\beta_n = 1) = \mathcal{CN}(h_B^\text{pri}, 0, \mu_n I + V^\text{pri})$ and $P(h_B^\text{pri}|\beta_n = 0) = \mathcal{CN}(h_B^\text{pri}, 0, V^\text{pri})$ for active devices and sleeping devices. Therefore, $\hat{\beta}_n$ can be judged according to the following criterion:

$$
\hat{\beta}_n = \begin{cases} 
1, & \text{if } P(\beta_n = 1|h_B^\text{pri}) > P(\beta_n = 0|h_B^\text{pri}), \\
0, & \text{otherwise}.
\end{cases}
$$

Consequently, $\hat{I}_{N_a}$ in line 4 of GTurbo-MMV can be obtained. After the processing of module B, step 3) of Algorithm 2 evaluates the extrinsic information of $\Xi$ based on the linear transform $\Xi = FH^T$, and then feeds them back to module A as the priori information of next iteration.

IV. PERFORMANCE ANALYSIS OF GTURBO-MMV

We now analyze the performance when a large number of online users are served by the BS, that is the so called SE [35, 36]. The SE analysis can be extended to MMV case when the unknown matrix is jointly sparse. In this section, we first use SE to derive achievable bounds on the channel estimation accuracy, then apply these bounds to determine the asymptotic performance of user detection when massive antennas are deployed at the BS.

A. SE Analysis of Channel Estimation

The performance of channel estimation can be analyzed by the SE equations, and we first focus on a general SE trajectory of GTurbo-MMV in Proposition 2.

Proposition 2: The SE of the GTurbo-MMV algorithm is
characterized by
\[
\vartheta^t = \left[ f_1(v_{1,1}^t), f_2(v_{2,2}^t), \ldots, f_{M_R}(v_{M_R,M_R}^t) \right],
\]
\[
\Omega^{t+1} = \text{diag} \left( \frac{1}{(\alpha \vartheta_1^t)^{-1} - v_{1,1}^t}, \ldots, \frac{1}{(\alpha \vartheta_{M_R}^t)^{-1} - v_{M_R,M_R}^t} \right)
\]
\[
V^{t+1} = \left( \frac{1}{N} \sum_{n=1}^{N} \text{mse}_n(\Omega^{t+1}, \mu_n) \right)^{-1} - \Omega^{t+1}^{-1},
\]
where the initial values are \( v_{1,1}^0 = v_{2,2}^0 = \ldots = v_{M_R,M_R}^0 = E[|h_{i,i}|^2] \), and
\[
f_m(v^t) = \left\{ \begin{array}{ll}
\frac{1}{\sigma^2} & \text{if } v^t = \frac{1}{\sigma^2} \sum_{i=1}^{n} \xi_i^2, \text{ when } \Delta_m > 0, \\
\Delta_m, & \text{if } v^t = \frac{1}{\sigma^2} \sum_{i=1}^{n} \xi_i^2, \text{ when } \Delta_m = 0,
\end{array} \right.
\]
(21)

with \( \xi \in \mathbb{R} \). When \( \Delta_m > 0 \), the definition of \( \Psi(\cdot) \) is given by (12), and \( \Psi'(\cdot) \) of \( f_m(v^t) \) is expressed as
\[
\Psi'(\tilde{y}; z, c^2) = \frac{\partial \Psi(\tilde{y}; z, c^2)}{\partial z} = \frac{e^{-\frac{(\tilde{y} - z)^2}{2c^2}} - e^{-\frac{(\tilde{y} - z + 1)^2}{2c^2}}}{\sqrt{2\pi}c^2}.
\]
(22)
The iteration index and measurement ratio are defined as \( t \) and \( \alpha \) respectively, and \( \text{mse}_n(\cdot, \cdot) \) is the mse of device \( n \).

**Proof:** The SE of the GTurbo-SMV algorithm is characterized by [32],
\[
\vartheta^t = f_m(v^t),
\]
\[
\eta^{t+1} = \frac{1}{(\alpha \vartheta)^{-1} - v^t},
\]
\[
v^{t+1} = \left( \frac{1}{\text{mse}(\eta^{t+1})} - \eta^{t+1} \right)^{-1},
\]
(23)
with the initialization \( v^0 = E[|h_{i,i}|^2] \).

By inspection of step 1) and step 4) in Algorithm 2, the first two equations in (24) can be extended to the multiple measurement case by calculating them in parallel based on the antennas’ different quantization resolutions. That is,
\[
\vartheta^t = \left[ \vartheta_1^t, \ldots, \vartheta_{M_R}^t \right]
\]
\[
= \left[ f_1(v_{1,1}^t), f_2(v_{2,2}^t), \ldots, f_{M_R}(v_{M_R,M_R}^t) \right],
\]
\[
\Omega^{t+1} = \text{diag} \left( \frac{1}{(\alpha \vartheta_1^t)^{-1} - v_{1,1}^t}, \ldots, \frac{1}{(\alpha \vartheta_{M_R}^t)^{-1} - v_{M_R,M_R}^t} \right)
\]
(25)
and \( \text{mse}_n(\Omega^{t+1}, \mu_n) \) in the last equation of (24) for each user \( n \) is shown in (26) at the top of the next page, with \( (\Omega^{t+1})^{-1} \) being a real diagonal matrix. Therefore, the SE of GTurbo-MMV are extensions of (24) by replacing scalar values with vectors and matrices in (25) and (26).

However, the integral part of (26) is computational challenging since the diagonal entries of \( \Omega^{t+1} \) are not identical. From [24–26], we know that the MMSE of mixed-ADC architecture can be modeled as linear combinations of MMSEs from different estimators, where the antennas with distinct quantization accuracy contribute relevant portions to the overall performance. This observation motivates us to use the same quantization model within (22) to derive bounds for our state evolution analysis, which results in Proposition 3 as follows.

**Proposition 3:** Define \( f_{\text{lower}}(v^t) = f_{\Delta_{\text{max}}}(v^t) \), \( f_{\text{upper}}(v^t) = f_{\Delta_{\text{max}}}(v^t) \). The SE of the GTurbo-MMV algorithm in channel estimation is bounded by
\[
\vartheta^t = f_{\text{bound}}(v^t),
\]
\[
\eta^{t+1} = \frac{1}{(\alpha \vartheta)^{-1} - v^t},
\]
\[
v^{t+1} = \left( \frac{1}{\text{mse}(\eta^{t+1})} - \eta^{t+1} \right)^{-1},
\]
(27)
where \( \text{mse}_n(\eta^{t+1}, \mu_n) \) in (27) is given by
\[
\text{mse}_n(\eta^{t+1}, \mu_n) = p_a \mu_n
\]
\[
= \left( \frac{R_R g(r, \mu_n + \frac{1}{\eta^{t+1}})}{p_a + (1 - p_a)} - \frac{R_R g(r, \mu_n + \frac{1}{\eta^{t+1}})}{p_a + (1 - p_a)} \right) dr,
\]
(28)
with \( g(a, b) = CN(a; 0, B) \). The lower bound of channel estimation is calculated by setting \( f_{\text{bound}}(v^t) = f_{\text{lower}}(v^t) \), with upper bound \( f_{\text{bound}}(v^t) = f_{\text{upper}}(v^t) \).

**Proof:** Please refer to Appendix B.

Due to the bit-resolution \( B \) in \( \vartheta^t \), \( \text{mse}_n(\eta^{t+1}, \mu_n) \) can be considered as \( \text{mse}_n(B, \mu_n) \). This relationship exemplifies the fundamental tradeoff between hardware cost/implementation complexity and system performance. In other words, the big picture is that wireless system designers should not simply strike to maximize the system throughput, but optimize hardware cost scaling to achieve specific throughput targets. Observing that the asymptotic behavior of channel estimation in Algorithm 2 can be characterized by MMV equations of SE, thus a large class of performance metrics (i.e., MSE or BER) can be computed easily. Moreover, the analytical framework can be applied to arbitrary postulated measures by changing the prior distribution of channel response.

**B. Asymptotic Performance of Active Device Detection**

In this subsection, we first bound the probability of the GTurbo-MMV-based active device detection in terms of the missing and false detection in Theorem 1. Then, further let \( M_R \rightarrow \infty \) for an insightful study in the massive MIMO regime.

**Theorem 1:** In the large system regime, within each iteration \( t \) in GTurbo-MMV, the probabilities of missing \( \text{Pr}_{\text{Miss}_n} \) and false \( \text{Pr}_{\text{False}_n} \) detection of user \( n \) are bounded separately,
\[
\text{Pr}_{\text{Miss}_n}(M_R, r_{1,n}^t(k_{\text{lower}}^t)) \leq \text{Pr}_{\text{Miss}_n,1},
\]
\[
\leq \text{Pr}_{\text{Miss}_n,1}(M_R, r_{1,n}^t(k_{\text{upper}}^t)),
\]
\[
\leq \text{Pr}_{\text{False}_n,2}(M_R, r_{2,n}^t(k_{\text{lower}}^t)),
\]
\[
\leq \text{Pr}_{\text{False}_n,2}(M_R, r_{2,n}^t(k_{\text{upper}}^t)),
\]
(29)

\[
\text{mse}_n(\Omega^{t+1}, \mu_n) = p_a \mu_n I - p_a^2 \left( \frac{1}{\mu_n} I + \Omega^{t+1} \right)^{-1} \Omega^{t+1} \left[ \int_{p_a + (1 - p_a)CN(\mathbf{r}; 0, (\Omega^{t+1})^{-1})} \right] \left( \Omega^{t+1} \left( \frac{1}{\mu_n} I + \Omega^{t+1} \right)^{-1} \right).
\]  

where \( P_{\text{Missing},n}^t \triangleq \mathbb{P}(\beta^t_n \neq 1 | \beta_n = 1) \), \( P_{\text{False},n}^t \triangleq \mathbb{P}(\beta^t_n = 1 | \beta_n \neq 1) \); \( \kappa_{\text{upper}}^t = (\alpha f_{\text{upper}}(\nu^{t-1}))^{-1} - \nu^{t-1} \); \( \kappa_{\text{lower}}^t = (\alpha f_{\text{lower}}(\nu^{t-1}))^{-1} - \nu^{t-1} \) defined in Proposition 3:

\[
R_{1,n}(M_R, r_{1,n}^t(\kappa^t_n)) = 1 - \frac{\Gamma(M_R, r_{1,n}^t(\kappa^t_n)M_R)}{\Gamma(M_R)},
\]

\[
R_{2,n}(M_R, r_{2,n}^t(\kappa^t_n)) = \frac{\Gamma(M_R, r_{2,n}^t(\kappa^t_n)M_R)}{\Gamma(M_R)},
\]

with the upper incomplete gamma function \( \Gamma(\cdot, \cdot) \), gamma function \( \Gamma(\cdot) \); and

\[
r_{1,n}^t(\kappa^t_n) = \frac{\kappa^t_n}{\mu_n} \left( 1 - \frac{1}{M_R} \ln \frac{1 - p_a}{p_a} + \ln \frac{\mu_n + \kappa^t_n}{\kappa^t_n} \right),
\]

\[
r_{2,n}^t(\kappa^t_n) = (1 + \kappa^t_n/\mu_n) \left( 1 - \frac{1}{M_R} \ln \frac{1 - p_a}{p_a} + \ln \frac{\mu_n + \kappa^t_n}{\kappa^t_n} \right).
\]

Proof: Please refer to Appendix C.

Based on Theorem 1, we can further deduce an asymptotic user detection performance theorem when \( M_R \) grows large.

Corollary 1: When the number of BS antennas \( M_R \to \infty \), we have

\[
\lim_{M_R \to \infty} P_{\text{Missing},n}^t = \lim_{M_R \to \infty} P_{\text{False},n}^t = 0.
\]

Proof: Please refer to Appendix D.

It is noted that Theorem 1 demonstrates that the device detection error probability approaches zero as \( M_R \to \infty \) for arbitrary \( t \). Interestingly, we found that Theorem 1 would hold true even if the SNR is relative small. That is to say, by leveraging the properties of massive MIMO, our algorithm can exactly detect all active users in massive connectivity with probability 1. This advantage mainly comes from the diversity and processing gains provided by massive antennas at the BS. Another remarkable property of Theorem 1 is that the accurate device detection is guaranteed as long as \( M_R \) is large, regardless of the values of \( N, N_a, L \).

C. Computational Complexity Analysis

The vector-wise manner at each iteration of GTurbo-MMV algorithm significantly reduces the computational complexity. Besides, the DFT operation outputs do not need to be stored, which is hardware-cost friendly to MMV in massive connectivity. The computational complexity of the GTurbo-MMV algorithm mainly comes from the calculation of the DFT and IDFT of size \( N \), which is \( \mathcal{O}(M_R N \log N) \) at each iteration. Besides, vector-wise operations of module A are \( \mathcal{O}(M_R L) \) and vector multiplications of module B are \( \mathcal{O}(M_R^2 N) \) in Algorithm 2 at each iteration. As a result, the overall computational complexity of the GTurbo-MMV algorithm is \( \mathcal{O}(TM_R (a_4 N \log(N) + b_4 L + c_4 M_R N)) \) with constants \( a_4, b_4, c_4 \) and \( T \) iterations (see Table I). It can be shown in the following analysis that the convergent MSE performance is normally achieved within few iterations (7-10). Since the first term contributes more to the overall computational complexity, we can conclude that the computational complexity of GTurbo-MMV is lower than that of AMP-MMV due to \( \log(N) < L \).

V. Numerical Results

In this section, we will evaluate the performance of our proposed GTurbo-MMV in channel estimation and user activity detection. We use missing, false and successful probability of user detection (MPUD, FPUD and SPUD) to measure the accuracy in user activity detection. MPUD or FPUD is defined to be the mean ratio of missing or false detected active users to the number of active or sleeping users, and SPUD is defined to be the probability of exact recovery of active users’ indices. Meanwhile, the performance of channel estimation is evaluated by the estimation MSE, defined as,

\[
\text{MSE} \triangleq \| \hat{H} - H \|^2_p / N M_R.
\]

The SNR of this system is defined as \( \text{SNR} = 1/\sigma^2 \). The possibility \( p_a \) that each \( N = 2000 \) online users transmit data within a frame is set to be 0.05, and \( L = 200 \) normalized partial DFT matrix is used as pilot sequence for each online user. Let \( D = [d_1, ..., d_N] \), and \( d_n \) denotes the distance between device \( n \) and the BS. It is assumed that \( D \) are randomly distributed in the regime \( [A_1, A_2] \)km, and the path loss model of the wireless channel for device \( n \) can be expressed as \( \mu_n = -128.1 - 36.7 \log_{10}(d_n) \) in dB with \( A_1 = 0.05, A_2 = 0.1 \). We assume that an ideal power control is performed to obtain the same large-scale fading factors from Fig. 5 to Fig. 11, while Fig. 12 demonstrates the performance with different large-scale components.

We first consider the iterative performance of our proposed GTurbo-MMV by replacing some low-resolution ADCs with high-resolution ADCs in Fig. 5. Here, \( M_R^\infty \) represents the number of high-resolution ADCs, and \( M_R - M_R^\infty = M_R^\infty \) is the number of B-bit ADCs. We set \( M_R = 16 \), and SNR = 20dB. As we can see in Fig. 5, the channel estimation and active device detection can converge in no more than 7 iterations for linear measurements and different mixed-ADC architectures. This means that our proposed algorithm converges fast under linear and nonlinear scenarios. As for the nonlinear measurements, both mixed 2-bit architecture and mixed 3-bit architecture are considered in Fig. 5. It should be noted that the performance loss in typical massive MIMO
TABLE I. Computational Complexity Analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>SMV</th>
<th>MMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP</td>
<td>$O[T(a_1LN + b_1L + c_1N)]$</td>
<td>$O[TM_R(a_2LN + b_2L + c_2M_RN)]$</td>
</tr>
<tr>
<td>GTurbo</td>
<td>$O[T(a_3N\log(N) + b_3L + c_3N)]$</td>
<td>$O[TM_R(a_4N\log(N) + b_4L + c_4M_RN)]$</td>
</tr>
</tbody>
</table>

Fig. 5. Iterative performance of the proposed algorithm under different measurement architecture with $p_0 = 0.05$, online users $N = 2000$, SNR = 20dB. The length of pilot sequence is $L = 200$.

Fig. 6. MSEs over different measurement architectures and BS antennas against SNRs when $p_0$ of each $N = 2000$ online users is 0.05 and the pilot length $L = 200$.

systems is mainly caused by low-resolution ADCs, which can be compensated by employing high-resolution ADCs at the BS to form a mixed-ADC architecture. It can be seen that the performance of channel estimation and active user detection is relatively poor in Fig. 5 (a) and Fig. 5 (b) when 6 high-resolution ADCs are installed under the mixed 2-bit architecture, and if we further increase $M_R^\infty$ to 10, then the MSE and the MPUD under the mixed 2-bit architecture will be improved and will approach that of linear measurement. Moreover, when it comes to the mixed 3-bit architecture, the performance of both MSEs and MPUDs are much better than that of the mixed 2-bit architecture in the same scenario. We also observe that MPUDs in Fig. 5 (b) improve significantly if 8 or more high-resolution ADCs are deployed at the BS. For example, the SPUD is about 0.90, and the MPUD is approximately 0.0006 under the mixed 3-bit architecture when $M_R^\infty = 10$, let alone the scenario of $M_R^\infty > 10$ or linear measurements. There is a point that should be mentioned here that the performances of channel estimation and active user detection are almost identical to that of linear measurements if the number of high-resolution ADCs increases to 12 or even larger with a mixed 3-bit architecture.

Fig. 6 demonstrates the performance of the GTurbo-MMV algorithm in terms of channel estimation with different number of BS antennas under mixed 2-bit and mixed 3-bit architectures. As expected, GTurbo-MMV has its performance gains in terms of channel estimation with the growth of SNR, and MSEs are generally improved by increasing $M_R^\infty$ or the resolution of ADCs. Moreover, when SNR=11dB, the MSE of mixed 3-bit architecture with $M_R^\infty = 40$ only exhibits 0.05dB loss when compared with linear measurement in Fig. 6 (a). Particularly, the gaps of MSE with mixed 3-bit architecture are not negligible when $M_R^\infty$ varies from 24 to 40 in Fig. 6.
(a), from 12 to 20 in Fig. 6 (b), and from 6 to 10 in Fig. 6 (c), respectively. The point is that GTurbo-MMV can achieve nearly linear channel estimation performance with low-resolution ADCs at the BS under the mixed 3-bit architecture, which is more pronounced when $M_R = 64$.

It can be observed in Fig. 7 that for all values of SNR, when the BS is equipped with $M_R = 64$ antennas, the SPUDs are significantly higher than those in the cases when there are $M_R = 16$ and $M_R = 32$ antennas at the BS. Further, the values of SPUD can approach 1 when the SNR is around 14 dB with 64 antennas at the BS, or 18 dB with 32 BS antennas and 22 dB with 16 BS antennas for linear and nonlinear measurements, respectively. On the other hand, all SPUDs with mixed 3-bit architecture are better than that with mixed 2-bit architecture irrespective of the value of $M_R$. Consequently, it is indicated that the GTurbo-MMV grant-free access technique can estimate CSI and detect device activity with extremely high accuracy by increasing $M_R^2$ or the quantization bits of low-resolution ADCs in massive MTC connectivity systems.

Fig. 7. SPUDs versus SNRs under different measurement architectures with different number of high resolution ADCs when $p_a$ of each $N = 2000$ online users is 0.05 and the pilot length $L = 200$.

Fig. 8. Missing detection performance comparisons of GTurbo-MMV and AMP-MMV when $p_a$ of each $N = 2000$ online users is 0.05 and the pilot length $L = 200$.

Fig. 9. Successful and error (includes missing and false) probability of active user detection performances for different online users under the nonlinear and linear measurements with active access probability $p_a = 0.01$ among $N = 2000$ online users. The length of pilot sequence is $L = 120$ and SNR = 20 dB.

Fig. 8 investigates the missing detection performance when there are different numbers of antennas at the BS for linear measurement and mixed 3-bit architecture, respectively. The AMP-MMV algorithm for linear measurement with partial DFT and i.i.d. Gaussian pilot sequences are also illustrated in Fig. 8 as comparison. Here, the variance of each entry in the i.i.d. Gaussian pilots is normalized to $1/N$ for a fair comparison. We observe that the MPUDs of linear AMP-MMV with different pilot sequences are poorer than that of linear GTurbo-MMV when there are $M_R = 16, 32, 64$ antennas at the BS. The performance difference between linear AMP-MMV with i.i.d. Gaussian pilots and partial DFT pilots indicates that the SE of linear AMP-MMV cannot be applied to partial DFT matrices directly. This is mainly because AMP or AMP-MMV is developed for i.i.d. Gaussian matrices [33]. Furthermore, when $M_R = 64$, all MPUDs are much lower than those when $M_R = 16$ or $M_R = 32$ in the same regime of SNR, which indicates that the performance of channel estimation and active device detection can be improved significantly when a large number of BS antennas is deployed at the BS.

We also consider the performance of GTurbo-MMV algorithm in terms of successful detection and error detection with different online users in Fig. 9. The number of BS antennas is $M_R = 16$. The possibility $p_a$ is set to be 0.01 and $L = 120$, but it has to be greater than the number of active devices to ensure no performance impact, or else a larger number of BS
Fig. 10. Performance bound of GTurbo-MMV with different number of high-resolution ADCs, SNR = 17dB, the number of active users is 100 for each online users \( N = 2000 \), and the pilot length is \( L = 200 \).

antennas are needed. As we can see in Fig. 9 (a) and Fig. 9 (b), both relatively high SPUDs and relatively low MPUDs can be achieved in a wide range of online devices \( N \). Moreover, Fig. 9 (a) demonstrates that the SPUD of a mixed 2-bit architecture can be improved when \( M_R^\infty \) is increased from 8 to 10. It is also observed that the SPUD with \( M_R^\infty = 8 \) under the mixed 3-bit architecture is higher than that of the mixed 2-bit framework with 10 high-resolution ADCs. Fig. 9 (b) depicts that the FPUD can approach zero when the number of online users is no more than 900 for both linear and nonlinear scenarios. It is indicated in Fig. 9 (b) that the FPUDs and MPUDs with mixed 3-bit architecture are lower than that of the mixed 2-bit architecture even though the number of high-resolution ADCs of mixed 3-bit architecture is smaller than that of the mixed 2-bit framework, which agrees with the analysis of SPUDs in Fig. 9 (a).

Next, we discuss the simulation results and analytical solutions in Fig. 10. As for the mixed 3-bit architecture with different number of high-resolution ADCs, the performance curves of channel estimation and active device detection are between the corresponding lower bounds (i.e., linear measurements when \( M_R^\infty \)) and upper bounds (i.e., 3-bit nonlinear measurements when \( M_R^\infty = M_R^3 \)). It can be inferred that the performance gap will decrease with the increase of \( M_R^\infty \) and will eventually be eliminated within the limit of lower bound. Therefore, the mixed architecture can help maintain the promised performance while significantly reducing the computational complexity.

Fig. 11 illustrates the analytical performance of channel estimation when the number of BS antennas varies from 16 to 64. It can be shown in Fig. 10 and Fig. 11 that all MSEs and error probabilities can converge in no more than 6 iterations. Simulations are conducted to verify the accuracy of analytical results. In particular, we compare the SE of MMV in (27) and (28) with those obtained by simulations. Especially, it can also be observed that the numerical results can match the SE equations characterized in Proposition 2 very well at a relatively low SNR when there are \( M_R = 64 \) antennas at the BS, which means that the simulation results of channel estimation obtained from GTurbo-MMV algorithm can perfectly match those predicted by the equations of SE when there are a large number of antennas at the BS.

Finally, Fig. 12 studies the scenario where power control is not performed in order to reduce the power consumption. We observe that a low missing probability of active device detection can still be achieved when \( M_R = 16 \) under the mixed 3-bit architecture. However, the signals from a long-distance device will be dominated to a large extent by the other signals with strong strength if the interval of distance is large, such as \( A_1 = 0.05, A_2 = 1 \). Fortunately, power control can be applied to make the effective path loss factors identical across different devices. Consequently, both scenarios can achieve satisfactory performance.

VI. CONCLUSION

In this paper, we presented a new scheme in allocating non-orthogonal pilot sequences to different online users in massive, small-packet sporadic transmission, which can offer potential solution to the emerging mMTC applications. To exactly identify active users and estimate their channel responses, we then proposed a GTurbo-MMV algorithm to perform joint channel estimation and user activity detection for linear and mixed ADC architectures. Our simulation results demonstrated a nearly linear performance of our proposed algorithm in channel estimation for nonlinear measurements, along with an approximately accurate active user performance even when the number of online users was far greater than the pilot
APPENDIX

A. Proof of Proposition 1

The posterior probability $P(h|r)$ of (14) can be computed following the Bayes’ rule,

$$P(h|r) = \frac{P(r|h)P(h)}{\int P(r|h)P(h)dh},$$

(34)

where $P(r|h) = CN(h; r, V)$ and $P(h)$ is defined in (15).

Then, based on (34), we have the posterior mean $E[h|r]$ and its corresponding variance $Var[h|r]$:

$$E[h|r] = \frac{\int hP(r|h)P(h)dh}{P(r)},$$

$$Var[h|r] = E[h^2|r] - E[h|r]E[h|r]^H,$$

$$= \frac{p_a}{P(r)}CN(r; 0, \mu I + V)(I + \frac{1}{\mu}V)^{-1}r,$$

(35)

where $A = (\mu^{-1}I + (V)^{-1})^{-1}$ and $P(r) = (1 - p_a)CN(0; r, V) + p_aCN(r; 0, \mu I + V)$.

To further derive the mse of estimator (35), we rewrite $Var[h|r]$ as

$$Var[h|r] = \frac{A_2}{P(r)} - \frac{A_1}{P(r)}\left(\frac{A_1}{P(r)}\right)^H,$$

(36)

where $A_1 = \int hP(r|h)P(h)dh$ and $A_2 = \int hh^H P(r|h)P(h)dh$ for the notation convenience. Then, $E[Var[h|r]]$ is given by

$$E[Var[h|r]] = \int Var[h|r]P(r)dr = \int A_2dr - \int \frac{A_1A_1^H}{P(r)}dr,$$

(37)

as for the first part in (37), we have

$$\int A_2dr = p_a \int hh^H CN(h; \Lambda V^{-1}r, \Lambda) CN(r; 0, \mu I + V)dr$$

$$= p_a (\Lambda V^{-1}) (\mu I + V)(\Lambda V^{-1})^H + p_a \Lambda,$$

(38)

while the second part is shown in (39) at the top of the next page. By substituting (38) and (39) into (37), we have $E[Var[h|r]]$ in (40) at the top of the next page. Consequently, (35) and (37) are equations given in Proposition 1.

B. Proof of Proposition 3

We first observe that when each antenna at the BS has the same quantization model, the $f_m(v')$ in (22) are then identical for all $m = 1, 2, \ldots, M_R$, which results in the same diagonal elements of $\Omega^{+1}$.

Consider $\Omega^{+1} = \theta^{+1}I$ with identical diagonal entries, $mse_n(\Omega^{+1}, \mu_n)$ in (26) is then simplified to (41) at the top of the next page, where $mse_n(\Omega^{+1}, \mu_n)$ has the same diagonal entries due to symmetry property. And the $(i, i)$th element of $mse_n(\Omega^{+1}, \mu_n)$ can be further simplified by using the matrix’s trace property,

$$mse_n(\theta^{+1}, \mu_n) = p_a \mu_n - \frac{(p_a \theta^{+1} + \frac{1}{\mu_n} \theta^{+1})^2}{M}$$

$$\times \int r^H r g(r, \mu_n + \frac{1}{\mu_n} \theta^{+1}) dr,$$

(42)

where $g(a, b) = CN(a; 0, 0, b)$. Note that the integral can be obtained through spherical coordinates instead of Cartesian coordinates. This completes the proof for Proposition 3.

C. Proof of Theorem 1

To complete the proof of Theorem 1, we first consider a special case in channel (14) when $V = vI$. Define an indicator function $\nabla(\cdot)$ as

$$\nabla(h) = \begin{cases} 1, & h \neq 0, \\ 0, & h = 0. \end{cases}$$

(43)

Under this circumstance, the MAP estimator of $h$ (i.e., $\hat{h}$) can be simplified in Proposition 4.

Proposition 4: By specifying $V = vI$ in (14), the MAP inference of $\nabla(\hat{h})$ can be simplified to a threshold detector, i.e.,

$$\nabla(\hat{h}) = \begin{cases} 1, & ||\hat{h}||^2 > \varepsilon, \\ 0, & \text{otherwise}, \end{cases}$$

(44)

with $M$ being the dimension of $h$, $\varepsilon = \frac{v(\mu + v)}{\rho} \left(\ln \frac{1 - p_a}{p_a} + M \ln \frac{\mu + v}{\mu} \right)$. Consequently, the error
\[
\int \frac{A_1 A_1^H}{P(r)} \, dr = p_a^2 \int \frac{CN(r; 0, \mu I + V)^2 (\Delta V^{-1} r) (\Delta V^{-1} r)^H}{P(r)} \, dr
\]
\[
= p_a^2 \left( \Delta V^{-1} \right) \left[ \int \frac{r r^H CN(r; 0, \mu I + V)}{p_a + (1 - p_a)CN(r; 0, V)/CN(r; 0, \mu I + V)} \, dr \right] (\Delta V^{-1})^H, \quad (39)
\]
\[
E[\text{Var}[h|r]] = p_a \mu I - p_a^2 \left( \Delta V^{-1} \right) \left[ \int \frac{r r^H CN(r; 0, \mu I + V)}{p_a + (1 - p_a)CN(r; 0, V)/CN(r; 0, \mu I + V)} \, dr \right] (\Delta V^{-1})^H, \quad (40)
\]
\[
\text{mse}_n(\Omega^{t+1}, \mu_n) = p_a \mu_n I - \frac{(p_a \theta^{t+1})^2}{(1/\mu_n + \theta^{t+1})^2} \int \frac{r r^H CN(r; 0, (1/\mu_n + \theta^{t+1}) I)}{p_a + (1 - p_a)CN(r; 0, 1/\mu_n + \theta^{t+1}) I} \, dr. \quad (41)
\]

Proof: Threshold \( \varepsilon \) in (44) can be obtained by calculating \( ||h||_2^2 \) in equation \( P(h|\nabla(h) = 0)P(\nabla(h) = 0) = P(h|\nabla(h) = 1)P(\nabla(h) = 1) \).

Since \( ||h||_2^2 \) can be modeled as a chi-squared distributed random variable with 2M degrees of freedom and \( P(||h||_2^2 \leq \varepsilon) = 1 - \Gamma(M, x/2)/\Gamma(M) \), it consequently follows that,
\[
P(\nabla(h) = 0|\nabla(h) = 1) = \frac{\Gamma(\frac{x}{2}, M)}{\Gamma(M)},
\]
\[
P(\nabla(h) = 1|\nabla(h) = 0) = \frac{\Gamma(\frac{x}{2}, M)}{\Gamma(M)}, \quad (45)
\]

Proof: Threshold \( \varepsilon \) in (44) can be obtained by calculating \( ||h||_2^2 \) in equation \( P(h|\nabla(h) = 0)P(\nabla(h) = 0) = P(h|\nabla(h) = 1)P(\nabla(h) = 1) \).

Since \( ||h||_2^2 \) can be modeled as a chi-squared distributed random variable with 2M degrees of freedom and \( P(||h||_2^2 \leq \varepsilon) = 1 - \Gamma(M, x/2)/\Gamma(M) \), it consequently follows that,
\[
P(\nabla(h) = 0|\nabla(h) = 1) = P\left( \frac{||h||_2^2 \leq 2\varepsilon}{\mu + v} \right) = 1 - \frac{\Gamma(M, r_1 M)}{\Gamma(M)},
\]
\[
P(\nabla(h) = 1|\nabla(h) = 0) = P\left( \frac{||h||_2^2 \geq 2\varepsilon}{\mu + v} \right) = \frac{\Gamma(M, r_2 M)}{\Gamma(M)}, \quad (46)
\]

where
\[
r_1 = \frac{\varepsilon}{M(\mu + v)} = \frac{v}{\mu} \left( \frac{1}{M} \ln \frac{1 - p_a}{p_a} + \ln \frac{\mu + v}{v} \right),
\]
\[
r_2 = \frac{\varepsilon}{M v} = (1 + v/\mu) \left( \frac{1}{M} \ln \frac{1 - p_a}{p_a} + \ln \frac{\mu + v}{v} \right). \quad (47)
\]

Based on Proposition 3, the missing \( P_{\text{Missing}, \eta}^t \) and false \( P_{\text{False}, \eta}^t \) probabilities can be bounded separately by evaluating (45) on either upper or lower bounds in (27), which completes the proof.

D. Proof of Corollary 1

 Lemma 1: For \( a \in \mathbb{R}^+ \),
\[
\lim_{M_R \to \infty} \frac{\Gamma(M_R, a M_R)}{\Gamma(M_R)} = 1 - \lim_{M_R \to \infty} e^{-a M_R} \sum_{k=M_R}^{\infty} \frac{(a M_R)^k}{k!} = \begin{cases} 
1, & a \in (0, 1), \\
0, & a \in (1, \infty).
\end{cases} \quad (48)
\]

Proof: Note that \( e^{-a M_R} \sum_{k=M_R}^{\infty} \frac{(a M_R)^k}{k!} \) is the tail probability, i.e., \( P(X \geq M_R) \) of \( a M_R \)-mean Poisson random variable \( X \). The bounds are derived by using Chernoff-bound technique [38],
\[
P(X \geq M_R) \leq (ae^{1-a})^{M_R} a \in (0, 1),
\]
\[
P(X \geq M_R) > P(X > M_R) \geq 1 - (ae^{1-a})^{M_R} a \in (1, \infty), \quad (49)
\]

where we denote \( f(a) \triangleq ae^{1-a} \), and it can be proved that \( f(a) \in (0, 1] \) for \( a \in \mathbb{R}^+ \) and achieves the maximum only when \( a = 1 \). Then, we can obtain
\[
\lim_{M_R \to \infty} \frac{\Gamma(M_R, a M_R)}{\Gamma(M_R)} = 1 - \lim_{M_R \to \infty} P(X \geq M_R) = \begin{cases} 
1, & a \in (0, 1), \\
0, & a \in (1, \infty).
\end{cases} \quad (50)
\]

To complete the proof for Corollary 1, we can conclude from (28) that,
\[
p_a \mu_n > \text{mse}_n(\eta^t, \mu_n) > p_a \mu_n - \left( \frac{p_a \eta^t}{1/\mu_n + \eta^t} \right)^2 \frac{M_R}{R_n} \times \int r^H r g (r, \mu_n + \eta^t) \, dr = p_a c^t, \quad (51)
\]

where \( c^t = \mu_n - p_a (p_a \theta^t)^2 (1/\mu_n + \eta^t)^2 \). Also, based on the monotonically decreasing property of function \( f_m \left( v^{t-1} \right) \) in (22) when \( \Delta_m \) is fixed, \( \kappa^t = (\alpha f_m \left( v^{t-1} \right))^{-1} - v^{t-1} \) can be proven to be monotonically increasing, which yields
\[
(\alpha f_m \left( p_a \mu_n \right))^{-1} - p_a \mu_n > \kappa^t > (\alpha f_m \left( p_a c^t \right))^{-1} - p_a c^t. \quad (52)
\]

Next, combined with Lemma 1 and \( \kappa^t / \mu_n \ln(1 + \mu_n / \kappa^t) < 1 \), \( (1 + \kappa^t / \mu_n) \ln(1 + \mu_n / \kappa^t) > 1 \) when \( \kappa^t > 0 \), we have
\[
\lim_{M_R \to \infty} R_1(n) \left( M_R, r_1^t(n) \right) = 1 - \lim_{M_R \to \infty} e^{-a M_R} \sum_{k=M_R}^{\infty} \frac{(a M_R)^k}{k!} = 0, \quad (53)
\]

Proof of Corollary 1
Consequently, (32) can be proved by applying the squeeze theorem on equations (29).

REFERENCES


Shi Jin (S’06-M’07-SM’17) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from the Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He serves as an Associate Editor for the IEEE Transactions on Wireless Communications, and IEEE Communications Letters, and IET Communications. Dr. Jin and his co-authors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and a 2010 Young Author Best Paper Award by the IEEE Signal Processing Society.

Chao-Kai Wen (S’00–M’04) received the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Taiwan, in 2004. He was with Industrial Technology Research Institute, Hsinchu, Taiwan and MediaTek Inc., Hsinchu, Taiwan, from 2004 to 2009, where he was engaged in broadband digital transceiver design. Since 2009, he jointed the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, where he is currently Professor. His research interests center around the optimization in wireless multimedia networks.

Michail Matthaiou (S’05–M’08–SM’13) was born in Thessaloniki, Greece in 1981. He obtained the Diploma degree (5 years) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2004. He then received the M.Sc. (with distinction) in Communication Systems and Signal Processing from the University of Bristol, U.K. and Ph.D. degrees from the University of Edinburgh, U.K. in 2005 and 2008, respectively. From September 2008 through May 2010, he was with the Institute for Circuit Theory and Signal Processing, Munich University of Technology (TUM), Germany working as a Postdoctoral Research Associate. He is currently a Reader (equivalent to Associate Professor) in Multiple-Antenna Systems at Queen’s University Belfast, U.K. after holding an Assistant Professor position at Chalmers University of Technology, Sweden. His research interests span signal processing for wireless communications, massive MIMO, hardware-constrained communications, and performance analysis of fading channels.

Dr. Matthaiou and his co-authors received the 2017 IEEE Communications Society Leonard G. Abraham Prize. He is currently awarded the prestigious 2018/2019 Royal Academy of Engineering/The Leverhulme Trust Senior Research Fellowship and will receive the 2019 EURASIP Early Career Award. His team was the Grand Winner of the 2019 Mobile World Congress Challenge. He was the recipient of the 2011 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East and Africa Region and a co-recipient of the 2006 IEEE Communications Chapter Project Prize for the best M.Sc. dissertation in the area of communications. He has co-authored papers that received best paper awards at the 2018 IEEE WCSP and 2014 IEEE ICC and was an Exemplary Reviewer for IEEE COMMUNICATIONS LETTERS for 2010. In 2014, he received the Research Fund for International Young Scientists from the National Natural Science Foundation of China. In the past, he was an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, Associate Editor/Senior Editor for IEEE COMMUNICATIONS LETTERS and was the Lead Guest Editor of the special issue on “Large-scale multiple antenna wireless systems” of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.

Xiaohu You (F’11) was born in 1962. He received the master’s and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, China, in 1985 and 1988, respectively. From 1999 to 2002, he was the Principal Expert of C3G Project, responsible for organizing China’s 3G mobile communications research and development activities. From 2001 to 2006, he was the Principal Expert of the China National 863 5G Project. Since 1990, he has been with the National Mobile Communications Research Laboratory, Southeast University, where he is currently the Director and a Professor. He has contributed over 100 IEEE journal papers and two books in the areas of adaptive signal processing, and neural networks and their applications to communication systems. His research interests include mobile communication systems, and signal processing and its applications. Dr. You was a recipient of the National First Class Invention Prize in 2011. He served as the General Chair of the IEEE WCNC 2013 and the IEEE VTC 2016. He is a Secretary General of the Future Forum and the Vice Chair of the China IMT-2020 Promotion Group and the China National Mega Project on New Generation Mobile Network.