IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 26, 2020, accepted September 21, 2020, date of publication September 29, 2020, date of current version October 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027523

Detection Techniques for Massive Machine-Type Communications: Challenges and Solutions

ROBERTO B. DI RENNA[1], (Graduate Student Member, IEEE),
CARSTEN BOCKELMANN[2], (Member, IEEE),
RODRIGO C. DE LAMARE[1], (Senior Member, IEEE),
AND ARMIN DEKORSY[2], (Senior Member, IEEE)
[1]Centre for Telecommunications Studies (CETUC), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro 22453-900, Brazil
[2]Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany

Corresponding author: Roberto B. Di Renna (robertobrauer@cetuc.puc-rio.br)

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES)-Finance code: 001.

ABSTRACT Massive machine-type communications (mMTC) is one of the key application scenarios of fifth generation (5G) and beyond cellular networks. Bringing the unique technical challenge of supporting a huge number of MTC devices (MTCD) in cellular networks, how to efficiently estimate the channel, detect the active users and data in this scenario is an open research topic. In this regard, this paper aims to present an overview of different techniques to address the problem of channel estimation, activity and data detection specifically for the mMTC scenario. In order to highlight potential solutions and to propose new research directions, we discuss the performance of the state-of-the-art techniques in the literature using a unified evaluation framework.

INDEX TERMS 5G, channel estimation, detection, massive access, mMTC, random access.

I. INTRODUCTION

Massive machine-type communications (mMTC) has become a key communication paradigm for various emerging smart services. Industrial automation, public safety, the Internet of Things (IoT), health-care, utilities, transportation, smart metering, remote manufacturing, and numerous other applications [1] are some examples that may coexist with the mobile devices communication among humans.

Different from the conventional human-type communications (HTC), mMTC for IoT have unique service features, as (i) uplink traffic dominated by very short packets: divided in metadata (preamble) and data (payload), the aggregation node should identify the active devices and estimate the channel with the metadata [2]; (ii) uncoordinated access: using a grant-less or grant-free data transmission, active devices transmit frames without preceding scheduling process to eliminate the need for round-trip signaling. The absence of large overheads avoids degradation of spectral and energy efficiency [3]; (iii) sparse user activity: despite the expected huge number of connected devices, each machine has a small probability of being active at the same time instant [4]; (iv) high energy efficiency: requirement of very long battery life where each device has to operate in an ultra-low power mode [5]; (v) low data rates: contrary to critical IoT or URLLC, (applications as tele-surgery, intelligent transportation, etc) massive IoT use cases may have some degree of tolerance on data reliability and latency constraints [6], [7].

Although detection techniques have been investigated for more than 50 years, at each new emerging application, novel schemes are required. In a scenario where a massive number of devices access a BS without coordination, how to deal with the sparsity pattern of transmissions, in order to address device identification, accurate channel state information (CSI) and to detect the data are challenging issues. The main focus of this work is on detection techniques for mMTC, their scenarios and possible supporting solutions. Thus, we first present the state-of-the-art and then our contributions.

A. RELEVANT PRIOR ART

We have carried out an extensive overview of techniques for mMTC and we introduce them in this section. Initially, activity and data detection techniques are presented, dividing the solutions into regularized, greedy and message-passing

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Huo.

algorithms. Then, techniques for channel estimation and activity detection are introduced, focusing on message-passing and machine learning techniques. Finally, a review of the surveys published so far is carried out and a table gathers the contributions of each work.

Given the expected sporadic device access at the base station, multi-user detection (MUD) algorithms should be reformulated to address the mMTC scenario [9]. Considering a perfect channel state information (CSI) at the BS, Zhu and Giannakis [10] proposed the Sparse Maximum a Posteriori Probability (S-MAP) detection which performs MAP detection of the new sparse problem, considering a zero-augmented finite alphabet. This scheme applies a signal processing approach adding a regularization parameter into the cost function. Originally from the compressed sensing (CS) field, the regularization exploits the sparsity of the scenario. In a way to reduce the complexity of the MAP detector, the authors relaxed the constraint so that optimization tools could be applied directly to obtain the solutions [11], [12].

Several techniques followed the pioneering work in [10], using CS approaches. Employing the QR decomposition, [13] incorporated a sparsity constraint in the successive interference cancellation (SA-SIC) reaching the performance of the S-MAP detector. The well-known K-Best detector was modified in [14], rewriting the signal model as a sparse model suitable for CS approaches. [15] reports a detection order method based on the activity probability of devices and the sorted QR decomposition (SQRD), in order to increase the efficiency of SA-SIC. The work in [16] investigates the reliability of each soft-estimate obtained by a regularized MMSE-SIC detector while [17] proposed an algorithm based on the direction method of multipliers (ADMM). The ones described in [18] and [19] are iterative and belong to the class of Bayesian inference algorithms. There are also solutions based on approximate message passing (AMP) [20]–[22], as [23] and [24], while the works in [25] and [26] developed a solution based on AMP and expectation minimization (EM) [27].

As envisaged mMTC networks support a massive number of devices, a common assumption is that the system is underdetermined. That is, the quantity of devices capable to access the BS at the same time instant is much higher than the number of resources at the BS. Owing the sporadic traffic pattern, greedy algorithms emerged as CS-MUD techniques. The well-known orthogonal matching pursuit (OMP) [28] and orthogonal least squares (OLS) [29] were first applied in the mMTC context in [30]. Seeking better performance, there are in the literature modifications of OMP, such as compressive sample matching pursuit (CoSaMP) algorithm [31], detection-based orthogonal matching pursuit (DOMP) algorithm [32], Weighted Group Orthogonal Matching Pursuit (wGOMP) [33], [34] and detecting-based group orthogonal matching pursuit (DGOMP) algorithm [35]. Rewriting the signal model in a way to increase the sparsity of the transmitted vector, the performance of the OLS is improved in [36]. By exchanging extrinsic information between active user detector and symbol detector, the schemes in [37]–[42] propose adaptive and iterative detectors that also employ channel coding.

Focusing the work on activity detection and channel estimation, [43]–[49] use the AMP, verifying the missed device detection and false alarm performance. Based on EP, [50] propose a solution with the factor graph approach. [50] and [51] uses a Bayesian message passing algorithm while the message passing of Ahn *et al.* [52] uses the EP and computes iteratively the moment matching. Lehmann derived in [53] an inference algorithm based on message-passing resulting in an iterative code-aided receiver. The work in [54] compares different approaches of the Hierarchical Hard Thresholding Pursuit (HiHTP) algorithm. Machine learning approaches are also suggested, as in [55]–[59].

Given the advancement of technology, several survey papers have been published with different focus of MTC and mMTC. Meanwhile, the existing literature lacks a comprehensive survey of detection techniques for MTC that discusses performance, complexity and future directions. Covering general aspects of MTC, [60]–[67] discusses technologies, opportunities and objectives. Since the provision of massive access is one of the main issues for mMTC, surveys that discuss physical and medium access layers like [68]–[72] have been published. Focusing exclusively on high-priority applications as remote surgery, industrial automation and autonomous vehicles [73] highlights diverse challenges and future aspects of mission critical MTC (mcMTC) on 5G-enabling technologies. Salam *et al.* present in [4] recent developments in data aggregation techniques, including application scenarios, design and limitations. [74] and [75] discuss security issues while the traffic is the main topic of Soltanmohammadi *et al.*, in [76]. In order to present a review of machine learning-assisted solutions, Sharma *et al.* categorized the different approaches in [77] while [78] is focused on challenges.

The main goal of this paper is to provide a survey of detection techniques that have been proposed over the last years for mMTC. We discuss the different schemes and compare their performance under the same framework, identifying strengths and weaknesses of each one of them, while drawing future trends to steer the efforts along the same line.

B. CONTRIBUTION AND OUTLINE

Table 1 lists the existing surveys related to MTC networks. The current published surveys either focus on particular aspects of MTC or do not cover the detection in a holistic way. As seen in Table 1, the surveys have provided contributions into the network architectures, data aggregation, application scenarios, enabling technologies, standards, security issues, traffic and limitations. Although the particularities of the MTC scenario inspired a lot of researchers to propose different solutions for channel estimation and activity and data detection problems. Nonetheless, to the best of our knowledge, this is the first survey to categorize and provide a

TABLE 1. Main related surveys on M2M communications in the literature.

Year	Paper	Main issues addressed	Related content	Enhancements and differences in this paper
2019	[4]	Data aggregation methods	Sections II and III-A	We detail one of the aggregation methods to develop the evaluation framework
2018	[60]	IoT requirements, communication standards, 5G new radio	Section II-A	We discuss the points related to MTC including new information
2017	[61]	MTC design requirements, applications, economic considerations		We debate the design requirements and applications
2015	[62]	Network architectures, challenges and solutions as power consumption, massive access, coverage and security	Section II-C	We review the challenges to implement most of mMTC scenarios and discuss the main points
2017	[63]	Communications standards, implementation challenges		We discuss the challenges holistically
2016	[64]	Categorization of MTC applications	Section II-A	We review MTC applications and include new solutions
2016	[65]	Technical requirements, standards, network architecture, business models	Section II	We debate the technical requirements in the challenges and applications sections, adding new information
2015	[66]	Network architecture, challenges, applications, open research issues		
2014	[67]	Platforms and communication standards		N/A
2016	[68]	Discussion of physical and medium access techniques	Section II-B	We focus in evaluate solutions for channel estimation and activity and data detection, considering one physical access technique
2018	[69]	Evaluation of throughput and access latency of physical and medium access techniques in a common simulation framework		
2015	[70]	Medium access control issues and protocols for MTC	Section III-A	We focus in physical access techniques in order to compare and discuss literature solutions
2014	[71]	Random access channel, discussing LTE and LTE-A limitations	Section II-C	We discuss the main points of the paper including new information
2018	[72]	Radio resource management	Section III-A	N/A
2019	[73]	General view of mission-critical MTC	Section II-A	N/A
2014	[74]	Security aspects and requirements	Section II-C	We discuss the points related to MTC including new information
2016	[75]	Security issues and network architecture		
2016	[76]	Traffic issues	Section II-B	We discuss the main points including new information
2019	[77]	Machine learning solutions to mMTC in an Ultra Dense Network scenario	Section IV-B2	Evaluation and discussion of the performance of machine learning solutions of channel estimation and activity and data detection
2020	[78]	Machine learning and deep learning techniques for resource managements of IoT networks		

comprehensive overview of the detection techniques specifically for mMTC. The main contributions of this paper are:

1) Extensive categorization of state-of-the-art for mMTC detection: There are a variety of solutions proposed to address many challenges of MUD in mMTC networks. In this paper, we classify and briefly review the main detection techniques for mMTC scenarios.
2) Performance evaluation of presented approaches in the same evaluation framework: In terms of complexity, frame error rate, false alarms and missed detection, we analyze the performance of main literature solutions.
3) Summarize the overall mMTC challenges and open issues specifically for detection techniques.

This paper is organized as follows. Section II outlines the main application scenarios, traffic aspects and general challenges. Section III describes the signal model adopted, while Section IV provides the categorization of existing solutions, detailing the different approaches. Finally, in Section V we discuss the results and compare different solutions in terms of their requirements and advantages, given future research directions. Section VI concludes the paper.

Notations: For a vector $\mathbf{a}$ (a matrix $\mathbf{A}$), $\mathbf{a}^*$ ($\mathbf{A}^*$) and $\mathbf{a}^T$ ($\mathbf{A}^T$) denote the complex conjugate and the transpose of a ($\mathbf{A}$), respectively. For an invertible matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ denotes the inverse of $\mathbf{A}$. $\mathbf{a}[i]$ denotes the ith element of a vector $\mathbf{a}$, and $[\mathbf{A}]_{i,j}$ is the element in the ith row and the jth column of $\mathbf{A}$. For a length-n vector $\mathbf{a}$, p-norm ($p \leq 1$) is defined as $\|\mathbf{a}\|_p = \left(\sum_{i=1}^{n} |\mathbf{a}[i]|^p\right)^{1/p}$, while the 0-norm $\|\mathbf{a}\|_0$ is the number of nonzero elements in $\mathbf{a}$. diag($\mathbf{a}$) denotes the $n \times n$ diagonal matrix whose ith diagonal element is $\mathbf{a}[i]$. For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$, and $\mathcal{S}_1 \cap \mathcal{S}_2$ is the intersection of $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}_1 \backslash \mathcal{S}_2$ denotes the relative complement of $\mathcal{S}_2$ with respect to $\mathcal{S}_1$. For a scalar a, $|a|$ is the absolute value of a. $\mathcal{E}[\cdot]$ stands for the expectation operator, and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ represents the multivariate complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma}$.

II. M2M COMMUNICATIONS

This section is composed by an introduction to application scenarios of mMTC followed by a brief description of the various traffic features and challenges.

A. APPLICATION SCENARIOS OF M2M COMMUNICATIONS

MTC devices can be applied to various scenarios, enabling real-time monitoring and control of any physical environment. The major application scenarios are [66], [79]:

1) E-health: Applications as tracking or monitoring a patient, identification and authentication of patients, diagnosing patient conditions and providing real-time information on patients' health related data to the remote monitoring center;
2) Smart-environment: This category encompasses all forms of automation, whether in home/office, agriculture, environmental monitoring, lighting;
3) Intelligent transportation: This field is related to services as smart parking, smart car counting, M2M assisted driving and e-ticketing;
4) Security and public safety: Remote surveillance, personal tracking and public infrastructure alarm protection from disasters such as fire, earthquakes, hazardous spills or crimes. Collaboration among relevant organizations, including medical support, police, military and fire department;
5) Smart-grid: Mainly related to power monitoring, this category also includes applications as meter reading, electricity distribution and transmission tower protection;
6) Industrial automation: Productivity enhancement possible by communications among machines and supply chain automation applications. Other example of applications are production on demand, quality control, optimization of packaging and inventory tracking.

There are other futuristic applications, as robotic applications and ''information-ambient society''. The first is an application related to the improvement of the quality of humans lifes, saving costs and resources. Dangerous tasks would be coordinated by robots, as fire-fighting, disable explosive devices, perform surgery and driverless vehicles [66]. ''Information-ambient society'' is related to devices be capable to deal with humans information provided by the Big Data technology. The idea is to enhance our society in terms of its intelligence and innovation level [80].

B. TRAFFIC FEATURES IN M2M COMMUNICATIONS

In general, MTC devices access the network sporadically to transmit frames with a few bits. Despite that, mMTC traffic comprises specific patterns due to diversity in the application scenarios. For instance, an agriculture sensor network sends few bits of data periodically while a smart-grid application consumes high bandwidth and requires connection with a higher frequency. In case of a catastrophe event, the network must be prepared to receive simultaneous transmissions of emergency data. Unlike HTC communications, mMTC traffic is mainly in the uplink and can be generated any time of the day. While the human-type communication traffic follows a certain data volume, session length, and interaction frequency during daytime and evening, mMTC should have an infrastructure that handles different traffic patterns. A few works as [76] and [81] investigate the difference between HTC and mMTC and the competition for resources. The study in [82] considers different traffic patterns and their impact with detection algorithms.

Therefore, mMTC have challenging traffic aspects such as scalability, periodic, low frame size and data rates, no mobility and deals mostly with the uplink, which requires special attention to design the infrastructure and coexist with the established cellular networks [66], [76].

C. CHALLENGES

3G and LTE networks can support a few MTC applications but not all of them. Thus, it is expected that 5G handles the massive number of MTCD and the services already available. In the following, we mostly focus on the challenges on the PHY and MAC layers but shortly discuss other issues.

One of the open problems is the limitation of available pilot sequences. Due to the huge number of devices, the reuse of pilots significantly increases the frame collision probability. Furthermore, it may causes the need for retransmissions leading to network congestion. In addition, as each transmitted frame has a few bits, very high signalling overhead per data frame becomes another critical issue. Thus, an efficient signalling reduction technique is required.

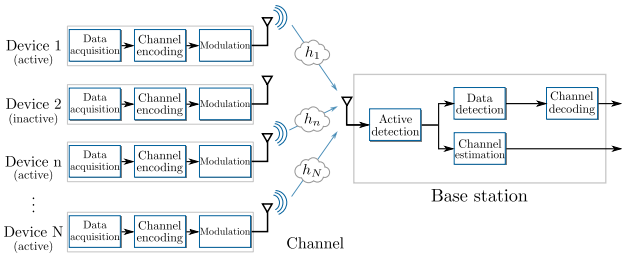


FIGURE 1. System model represented by a block diagram.

Many MTC devices can be coined ''low-cost and will have batteries as the main source of power. Therefore, energy efficiency is a concern for mMTC. Since it is required that MTCD operate autonomously for a few years, the communication design should be power efficient. Since long distance communications to the BS is challenging the use of relays could help. Another issue related to energy efficiency is the geographic positions of MTCDs. As they can be located anywhere in the cell, at edges or shadow areas, it would be hard for the BS to serve all devices at any time.

In higher layers, there are also open challenges. For instance, a new transport protocol is required for mMTC, as the Transmission Control Protocol (TCP) is not efficient for mMTC traffic features. The connection setup of TCP is unnecessary and the congestion control would probably degrade the performance as mMTC uses a wireless medium and the amount of data is very small. As it was not designed for it, real-time applications would not work properly with TCP as it requires data to be stored in a memory buffer. Furthermore, low-cost MTCDs have limited capabilities to implement security algorithms. In this way, authentication and data integrity may be a security concern in mMTC.

Some detection techniques take into consideration most of those challenges and propose solutions to deal with them. The next section details the main schemes in the literature.

III. SYSTEM MODEL

While the existing data aggregation technologies support various applications, there are still open problems to be investigated. How to deal with the massive access of hundreds of billion devices with small-sized transmission payloads and sporadic features is one of the main challenges of this kind of network [4], [69]. Indeed, promising techniques such as compressed sensing (CS), non-orthogonal multiple access (NOMA) and massive multiple-input multiple-output (mMIMO) based random access that can effectively address the lack of spectrum resources and enhance spectrum usage efficiency have been proposed. In this way, we present the system model used to compare the performance of the detection techniques in the literature.

A. GENERAL ASSUMPTIONS

In order to reduce signalling overhead, grant-free random access (GFRA) has been proposed [83]. In the uplink of such systems, each device transmits metadata along with data. This massive uplink connectivity scenario is illustrated in Fig. 1, where N devices with a single antenna each access a single base station (BS), equipped with M antennas.

In general, there are two types of metadata considered in GFRA, namely orthogonal [84], [85] and non-orthogonal metadata [9], [48]. Compared with the non-orthogonal counterpart, orthogonal metadata detection is much simpler and effective and channel estimation is more accurate thanks to the orthogonality of metadata. Nevertheless, frame collision restricts its performance due to the limited number of orthogonal-metadata sequences. On the other hand, non-orthogonal metadata can alleviate metadata collisions since it has a larger number of sequences, but its channel estimation would be affected due to non-orthogonality of metadata. Since it is expected to MTC handle a massive number of connections, the insufficient number of resources in a orthogonal metadata approach implies the usage of non-orthogonal metadata in GFRA. In the literature, there are works that address the GFRA in different ways. In [86], sparse sequences were used instead of binary sequences for data signal spreading in order to increase the number of MTC devices and allow device identification. With the aim to reduce the metadata collision and improve the GFRA throughput, the work in [87] suggests the usage of multiple resource blocks. In [88], another GFRA scheme was proposed where each device's channel impulse response is used as a unique signature to differentiate signals that are simultaneously transmitted and the works in [89], [90] studies where the wireless signal of each device is spread by a unique sequence.

Using GFRA with non-orthogonal sequences, we have established a common system model to compare the performance of the detection algorithms. We define that when a device has data to transmit, it splits the codeword in multiple frames and transmit them in multiple transmission slots. In each time slot, each active device selects randomly a non-orthogonal metadata sequence from a predetermined codebook and sends the rest of the codeword. Since in practice the BS would have a list of devices that are associated with it, and their unique identifiers, we assume that the metadata sequences are known at the BS. Since these unique identifiers are known to the BS, the metadata sequence is also known at the BS. Given the sporadic activity of devices, they will communicate to the BS only when it is needed, so not all of them will be active during the same coherence time.

As illustrated in Fig. 1, we consider that devices are synchronized in time, i.e., devices are turned on or turned off in the same transmission slot. This assumption is valid since the frame size of mMTC is typically very small (between 10 and 100 bytes) [4]. We consider that the whole transmitted frame experience the same channel in such a way that the duration of a transmission slot ($\tau = \tau_\phi + \tau_x$) is smaller than the coherence time and coherence bandwidth of the channel. The time index t indicates each transmitted vector in the same transmission slot. As we consider a grant-free random access model, each frame has metadata and data. Thus, the time index indicates how each frame is divided.

In a given coherence time, at the t-th symbol interval, the received signal $\mathbf{y}[t]$ is organized in a $M \times 1$ vector that contains the transmitted metadata ($\boldsymbol{\phi}[t]$) or the data ($\mathbf{x}[t]$), as

$$\mathbf{y}[t] = \begin{cases} \mathbf{H}\sqrt{\tau_\phi}\,\mathbf{B}\,\boldsymbol{\phi}[t] + \mathbf{v}[t], & \text{if} \quad 1 \le t \le \tau_\phi \\ \mathbf{H}\sqrt{\tau_x}\,\mathbf{B}\,\mathbf{x}[t] + \mathbf{v}[t], & \text{if} \quad \tau_\phi < t \le \tau \end{cases} \tag{1}$$

where $\mathbf{H}$ is the $M \times N$ channel matrix, $\mathbf{B}$ is the $N \times N$ transmission power matrix, $\mathbf{v}$ is the $M \times 1$ noise vector, while τ_ϕ and τ_x are the number of metadata and data symbols, respectively. For each time instant t, the metadata and data are represented by the $N \times 1$ vectors

$$\boldsymbol{\phi}[t] = \quad \boldsymbol{\Delta}\,\boldsymbol{\varphi}[t] = [\delta_1\,\varphi_1[t], \ldots, \delta_N\,\varphi_N[t]]^T \quad \text{and} \tag{2}$$

$$\mathbf{x}[t] = \quad \boldsymbol{\Delta}\,\mathbf{s}[t] = [\delta_1\,s_1[t], \ldots, \delta_N\,s_N[t]]^T, \tag{3}$$

where $\boldsymbol{\varphi}[t]$ and $\mathbf{s}[t]$ are $N \times 1$ vectors of symbols from a regular modulation scheme denoted by $\mathcal{A}$, as quadrature phase-shift keying (QPSK). The $N \times N$ diagonal matrix $\boldsymbol{\Delta}$ controls the activity of each device in the specific transmission slot with

$$\begin{cases} \Pr(\delta_n = 1) = \rho_n, \\ \Pr(\delta_n = 0) = 1 - \rho_n. \end{cases} \tag{4}$$

Thus, each transmitted vector ($\boldsymbol{\phi}[t]$ or $\mathbf{x}[t]$) is composed by the augmented alphabet $\mathcal{A}_0$, where $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$. The $N \times N$ diagonal $\mathbf{B}$ matrix gathers the transmission power component b of each device, as in mMTC systems each MTCD has a different power level [1], [4]. The noise vector $\mathbf{v}$ is modelled as an independent zero-mean complex-Gaussian $M \times 1$ vector with variance σ_v^2.

In this work, we consider the block fading model, where the values change independently from slot to slot, that is, the channel is constant over a transmission slot duration. The $M \times N$ channel matrix $\mathbf{H}$ corresponds to the channel realizations between the BS and devices as modeled by

$$\mathbf{H} = \mathbf{A}\,\boldsymbol{\mathcal{N}}^{1/2}, \tag{5}$$

where $\mathbf{H}$ gathers independent fast fading, geometric attenuation and log-normal shadow fading. $\mathbf{A}$ is the $M \times N$ matrix of fast fading coefficients circularly symmetric complex Gaussian distributed, with zero mean and unit variance. We also the consider the effects of path loss and shadowing experienced by each MTCD, modelling them in the $N \times N$ diagonal matrix $\boldsymbol{\mathcal{N}}$, where each component is given by $10\log_{10}(\chi) + \omega$, where χ is the signal-to-noise ratio (SNR) and ω is a Gaussian random variable with zero mean and variance σ_ω^2 [91]. Thus, each vector $\mathbf{h}_n$ can be written as

$$\mathbf{h}_n = \mathbf{a}_n\sqrt{\beta_n}, \quad \forall n = 1, \ldots, N. \tag{6}$$

The β_n coefficients are assumed to be known at the BS and changes very slowly, reaching a new value just in a new transmission slot. Given the features of mMTC scenarios, the number of devices N is larger than that of antennas M at the base station, in a way that it consists of an underdetermined system. All signal model parameters are described in Table 2.

TABLE 2. Description of signal model parameters.

Parameter	Description
M	Number of base station antennas;
N	Number of devices;
K	Number of active devices;
τ	Number of transmitted symbols per trans. slot, given by $\tau = \tau_x + \tau_\phi$, where τ_x represents the data and τ_ϕ the metadata;
ρ_n	Random variable with a beta distribution that represents the probability of being active of the n-device;
$\mathbf{y}[t]$	$M \times 1$ received symbol vector of the time instant t;
$\boldsymbol{\phi}[t]$	$N \times 1$ metadata vector of the time instant t composed by the augmented alphabet $\mathcal{A}_0$;
$\mathbf{x}[t]$	$N \times 1$ data vector of the time instant t composed by the augmented alphabet $\mathcal{A}_0$;
$\boldsymbol{\Delta}$	$N \times N$ diagonal matrix that controls each device activity in the specific transmission slot;
$\mathbf{B}$	$N \times N$ diagonal matrix that gathers the transmission power of each device;
$\mathbf{v}[t]$	noise component, modelled as a independent zero-mean complex-Gaussian $M \times 1$ vector with variance σ_v^2;
$\mathbf{H}$	$M \times N$ channel matrix, where $\mathbf{H} = \mathbf{A}\,\boldsymbol{\mathcal{N}}^{1/2}$;
$\mathbf{A}$	$M \times N$ matrix of fast fading coefficients;
$\boldsymbol{\mathcal{N}}$	$N \times N$ diagonal matrix that gathers the path loss and shadowing experienced by each device;

B. KEY PERFORMANCE INDICATORS

To evaluate the performance of detection techniques, we consider three key performance indicators (KPI):

- The **Frame Error Rate** (FER) denotes the total number of frames incorrectly detected by the BS;
- The **Missed Detection Rate** (MDR) denotes the total number of symbols that have been transmitted in a specific time instant that the detector judged as zero, divided by the number of active devices;
- The **False Alarm Rate** (FAR) is the number of symbols detected as different from zero, divided by the difference between the total number of devices and the number of active devices at a time instant;

Considering $S_{\mathbf{x}}$ as the true support set, that is, the list of active devices in $\mathbf{x}$, a Venn diagram in Fig. 2 represents the key performance indicators. For the techniques that perform channel estimation, we evaluate the efficiency by the normalized mean-squared error (NMSE).

C. GENERAL SIMULATION PARAMETERS

In order to evaluate the performance of the cited algorithms, we consider $N = 128$ MTCDs connected to a single base-station equipped with $M = 64$ antennas. The evaluated solutions experience an independent and identically-distributed (i.i.d.) random flat-fading channel model and the values $a_{m,n}$ of (6) are taken from complex Gaussian distribution of $\mathcal{CN}(0, 1)$. When the device is active, it radiates QPSK symbols with power values drawn uniformly at random in 0.1 W to 0.3 W and the probability of being

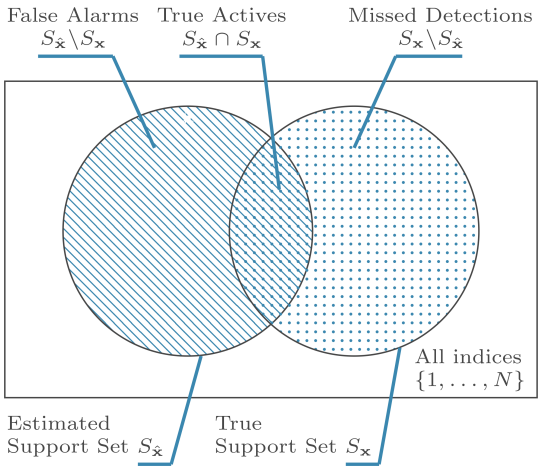


FIGURE 2. Venn Diagram of False Alarm and Missed Detection Errors.

active of each device ρ_n is drawn uniformly at random in [0.1, 0.3]. Each frame has 128 symbols, split into 60 metadata and 68 data. This balance between pilots and data is suggested in [9].

IV. DETECTION TECHNIQUES

In this section we present and compare relevant state-of-the-art algorithms divided in four different classes of detectors: regularized, greedy, message-passing and machine learning based. First, we start with the algorithms that aim to perform activity and data detection, discussing the results based on the KPIs. For the activity detection and channel estimation techniques, beyond the missed detection and false alarm rates, we provide the NMSE performance. In order to increase the readability, in the beginning of each subsection we present a table (Tables 3 and 4) that concisely introduces the main ideas of each algorithm reproduced.

A. ACTIVITY AND DATA DETECTION

1) REGULARIZED DETECTORS

In order to perform MAP detection of the mMTC sparse problem, the Sparse Maximum a Posteriori Probability (S-MAP) detection was proposed in [10]. This approach was the first that applied a regularization parameter into the cost function, inspiring the following works to propose suboptimal algorithms. The authors of [10] considered a simplified version of (1), that ignores the number of symbols, transmission power and time instant, as given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}. \tag{7}$$

Another assumption of [10] is to consider the zero as part of the modulation alphabet $\mathcal{A}$ (as QPSK), this way including the zero in the detection problem. With $\mathcal{A}_0 = \mathcal{A} \cup 0$ being the augmented modulation alphabet and ρ_n the activity probability for the n-device, [10] described the prior distribution of $\mathbf{x}$ as

$$\Pr(\mathbf{x}) = \prod_{n=1}^{N} \Pr(x_n) = \prod_{n=1}^{N} (1-\rho_n)^{1-|x_n|_0} \left(\frac{\rho_n}{|\mathcal{A}|}\right)^{|x_n|_0}, \tag{8}$$

where $|\mathcal{A}|$ is the cardinality of the modulation alphabet and $|x_n|_0$ is the element-wise l_0-norm that is equal to 1 if x_n is a non-zero value, otherwise it is zero. The output of the S-MAP detector maximizing the *a posteriori* probability $\Pr(\mathbf{x}|\mathbf{y})$ is given by Bayes' rule as

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_{\mathbf{x} \in \mathcal{A}_0^N} \Pr(\mathbf{x}|\mathbf{y}) \\ &= \arg\min_{\mathbf{x} \in \mathcal{A}_0^N} -\ln \Pr(\mathbf{y}|\mathbf{x}) - \ln \Pr(\mathbf{x}), \end{aligned} \tag{9}$$

which leads to

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathcal{A}_0^N} \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \sigma_v^2 \sum_{n=1}^{N} \lambda_n |x_n|_0, \tag{10}$$

where $\lambda_n = \ln[(1-\rho_n)/(\rho_n/|\mathcal{A}|)]$ is the regularization parameter for the n-th symbol detection.

As the main objective of the S-MAP detection is to find a vector in $\mathcal{A}_0^N$ that maximizes the cost function in (10), naturally, the complexity of S-MAP is tremendous. In [10] itself the authors propose two relaxing approaches, called Ridge detector (RD) and Lasso detector. RD regularizes the least squares (LS) solution using the l_2-norm, while LD uses the l_1-norm.

SA-SIC

For the sake of achieving an acceptable detection performance with much lower complexity compared to other optimal but complex S-MAP detectors, [13] proposed the sparsity-aware successive interference cancellation (SA-SIC). Considering that the BS has the perfect knowledge of probability of being active of each device and perfect CSI, SA-SIC recovers transmitted symbols in a sequential manner, incorporating the regularization into the problem. As SA-SIC uses the QR decomposition of $\mathbf{H} = \mathbf{Q}\mathbf{R}$, we have

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathcal{A}_0^N} \frac{1}{2} \sum_{n=1}^{N} \left[\left| \tilde{y}_n - \sum_{l=n}^{N} R_{nl} x_l \right|^2 + \sigma_v^2 \lambda_n |x_n|_0 \right], \tag{11}$$

where $\tilde{\mathbf{y}} = \mathbf{Q}^H \mathbf{y}$. Like any SIC technique, SA-SIC is sensitive to the error propagation from the early layers. Therefore, ordering techniques should be applied to mitigate the error propagation.

SA-SIC with A-SQRD

Considering the same assumptions of the SA-SIC, the authors of [15] proposed a permutation of columns of channel matrix $\mathbf{H}$ based on channel gains. The idea of SA-SIC with A-SQRD is to replace the l_0-norm with the l_2-norm in (10), incorporate

TABLE 3. Concise description of simulated activity and data detection techniques.

Type	Category	Algorithm	Main ideas
Activity and data detection	Regularized	SA-SIC [13]	• Lower complexity alternative to the S-MAP detectors; • Sequentially recover transmitted symbols, incorporating the regularization into the problem. Sensitive to error propagation.
		SA-SIC with A-SQRD [15]	• Replaces the l_0-norm with the l_2-norm, incorporated into the channel $\mathbf{H}$. • To reduce the error propagation, it permutates the columns of $\mathbf{H}$ based on channel gains.
		AA-MF-SIC [16]	• Iteratively updates a MMSE filter using a constellation-list scheme; • The MMSE filter is regularized with a l_1-norm.
		AA-RLS-DF [42]	• Scheme with implicit channel estimation, that uses a l_0-norm regularized recursive least squares (RLS) adaptive algorithm; • Uses the metadata to update the weights; • Exhibits an adaptive detection order and cancels the interference of the previously detected symbols; • The scheme also considers a decoding scheme for mMTC.
	Greedy	wGOMP [33]	• Refines the activity detection, exploiting the channel code; • With the knowledge of the number of active devices, wGMOP iteratively updates weights to acquire the likelihood of activity for each node, improving the activity detection.
		bcSIC [34]	• Computes the activity estimation as in wGOMP and performs the LS estimation for the chosen node and its channel decoding. • Each residual is updated with the most likely codeword of the chosen node for interference cancellation.
		mSOMP-EXT [39]	• Computes the average LLR of each device, where an approximation that does not requires the knowledge of the sparsity level and the noise variance is introduced; • Since it is an iterative algorithm, each LLR creates and transfers the extrinsic information through iterations to support the data detection.
		TA-BSASP [99]	• Adaptive algorithm based on the classical subs pace pursuit; • Reconstructs the sparse vector by exploiting the inherent block sparsity, as the authors vectorized all data transmitted in different time slots.
	Message passing	M-AMP [48]	• Considers a non-coherent transmission scenario; • Incorporates the sparsity not only in the transmitted vector, but also in $\mathbf{H}$. • Designs a denoiser capable to suppress the metadata sequences that do not belong to the evaluated device, using a soft-thresholding function.
		NSD-AMP [49]	• Develops a section-wise equivalent model, that decouples the estimation in different sections. This allows the design of the section-wise Bayes-optimal denoiser for the AMP, minimizing the MSE section by section.
		Joint-EM-AMP [25]	• Uses the EM algorithm to estimate the activity of devices. • Updates the activity estimate with the detection of each received symbol.
		CS-MPA [31]	• Uses a CS approach to realize the user activity and a message-passing method to the data detection. • CoSaMP algorithm is used to estimate the number of active users and the maximum iteration number. • Performs a LS approach and computes a residual in order to aid in the next iteration.
		EM-BSBL [18]	• Does not require the priori knowledge of the activity factor; • Modifies the BSBL algorithm that considers the prior of the row sparsity property and the column coherence of each device. • Makes use of reasonable priors and a set of hyperparameters to control estimated signals that can be learned from the training progress via EM.

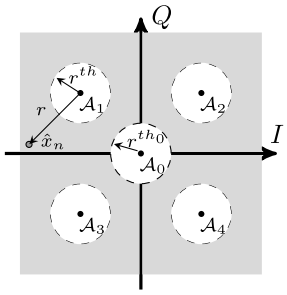


FIGURE 3. Shadow area constraint for QPSK modulation. Estimates within the white circles are deemed reliable and quantized to the alphabet, outside the circle a list estimation scheme is used.

the regularization factor into the channel matrix $\mathbf{H}$, as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathcal{A}_0^N} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_N \end{bmatrix} - \underbrace{\begin{bmatrix} \mathbf{H} \\ \sigma_v \operatorname{diag}\left(\sqrt{\boldsymbol{\lambda}}\right) \end{bmatrix}}_{\mathbf{H}'} \mathbf{x} \right\|_2^2, \quad (12)$$

find the QR decomposition of the augmented $\mathbf{H}'$, employ the modified Gram-Schmidt algorithm [92] and reorder the columns of $\mathbf{H}'$ before each orthogonalization step.

AA-MF-SIC

Focusing on the filter refinement, the authors in [16] incorporated an l_1-norm regularization in the linear MMSE filter ($\mathbf{W}$) and a constellation-list scheme to increase the detection performance. The main idea of the AA-MF-SIC algorithm is to iteratively update the regularized linear MMSE filter at each new symbol detection n, as

$$\mathbf{w}_n = \left(\overline{\mathbf{H}}_n \overline{\mathbf{H}}_n^H + \frac{\sigma_v^2}{\sigma_x^2}\mathbf{I} + \frac{2\lambda_n}{\sigma_x^2}\Lambda\right)^{-1} \overline{\mathbf{H}}_n \boldsymbol{\delta}_n, \quad (13)$$

where $\Lambda = \operatorname{diag}\left\{\frac{1}{|w_{n,1}|+\epsilon}, \frac{1}{|w_{n,2}|+\epsilon}, \ldots, \frac{1}{|w_{n,M}|+\epsilon}\right\}$, $\boldsymbol{\delta}_n$ is a $N \times 1$ zero column vector with 1 at the n-th position and Λ_n is the regularization parameter defined by [10]. As this algorithm also performs the successive interference cancellation, $\overline{\mathbf{H}}_n$ denotes the matrix obtained by taking the columns $n, n+1, \ldots, N$ of the channel. In order to avoid the error propagation, the authors in [16] evaluate the reliability of each soft symbol estimate. The constellation-list scheme consists of a shadow area constraint as shown in Fig. 3, with the augmented alphabet of a modulation scheme, that compares the distance r between the soft estimate and all the possible constellation symbols with

$$\arg\min_{i\in 1,\cdots,|\mathcal{A}|+1} \|\mathcal{A}_{0_i} - \tilde{x}_n\|^2. \quad (14)$$

If the soft estimate falls into the shadow area ($r > r^{th}$ or $r > r^{th_0}$), the estimate is considered unreliable and then $\tilde{x}_n$ proceeds to the list scheme detailed below. Otherwise, it is just quantized ($\mathcal{Q}[\cdot]$) to the nearest symbol of the augmented constellation $\mathcal{A}_0$, as $\tilde{x}_n = \mathcal{Q}\left[\mathbf{w}_n^H \mathbf{y}_n\right]$. The radius of each reliability region are defined by the probability of being active of each device and the radius of the region around the zero (inactive device) is the complement of the radius of the regions around the constellation symbols, $r^{th_0} = 1 - r^{th}$. The list scheme employed in the shadow area is used to select the best constellation symbol candidate, according to

$$\kappa_{\text{opt}} = \arg\min_{i\in 1,\cdots,|\mathcal{A}_0|} \|\mathbf{y}_n - \mathbf{h}_n \mathcal{A}_{0_i}\|^2, \quad (15)$$

where $\mathbf{y}_n$ is the received vector after the SIC operation and the vector $\mathbf{h}_n$ contains the estimate of the channel between the device that performs symbol detection and the BS. The index κ_{opt} indicates which candidate of the list $\mathcal{A}_{0_i}$ will replace the quantized version of the unreliable soft symbol estimate $\tilde{x}_n$. After the detection, the algorithm proceeds with SIC.

AA-RLS-DF

Since prior techniques do not perform channel estimation, the work in [42] builds on previous decision feedback techniques [93], [94] and proposes a scheme with implicit channel estimation. The AA-RLS-DF uses a regularized recursive least squares (RLS) adaptive algorithm that relies on the metadata to update the weights. More sophisticated algorithms [95] can also be considered. The detection order is updated at each new layer, using the least squares estimation (LSE) criterion. The adaptive receive filter can be decomposed into feedforward and feedback filters. The feedforward one is updated at every new received vector by the l_0-norm regularized RLS algorithm. The feedback filter is a component that is concatenated to the feedforward filter in order to cancel the interference of the previously detected symbols. The feedforward and feedback receive filters are written as

$$\mathbf{w}_{\psi_n}[t] = \begin{cases} \mathbf{w}_{\psi_n}^f[t], & n = 1; \\ \left[\mathbf{w}_{\psi_n}^{f\ T}[t], \mathbf{w}_{\psi_n}^{b\ T}[t]\right]^T, & n = 2, \ldots, N. \end{cases} \quad (16)$$

As the length of the filter increases at each new detection, the received vector also increases, concatenating the previous detected symbol, as $\mathbf{y}_{\psi_n}[t] = \left[\mathbf{y}^T[t], \hat{\mathbf{x}}_{\psi_n}^T[t]\right]^T$. As this algorithm takes into account each part of the received block per time, the time index t is necessary, as is the detection order index ψ_n. The detection order is updated with the minimum argument of the regularized cost function given by

$$\mathcal{J}_j[t] = \sum_{l=0}^{t} \mu^{t-l} |\hat{x}_j[l] - \mathbf{w}_j^H[t]\,\mathbf{y}_{\psi_n}[t]|^2 + \gamma \|\mathbf{w}_j[t]\|_0, \quad (17)$$

where μ is the forgetting factor of the RLS algorithm, j is the index of the procedure to decide the next symbol to be detected and γ is a non-zero positive constant to balance the regularization and, consequently, the estimation error. After a few steps, the l_0-norm regularized RLS adaptive expression becomes

$$\begin{aligned} \mathbf{w}_j[t] = \mathbf{w}_j[t-1] + \mathbf{k}[t]\,\epsilon_n^*[t] \\ -\gamma\xi \operatorname{sgn}\left(w_{j,p}[t]\right) f_\xi\left(w_{j,p}[t]\right), \end{aligned} \quad (18)$$

where $\mathbf{k}$ is the gain vector and ξ is a positive parameter that regulates the range of the attraction to zero on small

coefficients of the filter. Moreover, the function $f_\xi\left(w_{j,p}[t]\right)$ is given by

$$f_\xi\left(w_{j,p}[t]\right) = \begin{cases} \xi^2\left(w_{j,p}[t]\right) + \xi, & -1/\xi \le w_{j,p}[t] < 0; \\ \xi^2\left(w_{j,p}[t]\right) - \xi, & 0 \le w_{j,p}[t] \le 1/\xi; \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

After the filters weights are computed with the metadata symbols, the algorithm uses the same procedure to compute the data soft estimates.

Performance evaluation

Initially, Fig. 4, shows the Frame Error Rate (SER) performance averaged over 10^5 runs. Considering the average SNR as $10\log\left(N\,\sigma_x^2/\sigma_v^2\right)$, the linear mean squared error (LMMSE), unsorted SA-SIC [14], SA-SIC with A-SQRD [15], AA-MF-SIC [16], AA-RLS [42] and AA-RLS-DF [42] are compared. As a lower bound, the Oracle LMMSE detector, which has the knowledge of the index of nonzero entries, is considered. Since the schemes of [42] do not require explicit channel state information, in order to perform a fair comparison, we take into account an imperfect channel estimation to the other approaches. We considered $\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E}$, where $\mathbf{H}$ represents the channel estimate and $\mathbf{E}$ is a random matrix corresponding to the error for each link. Each coefficient of the error matrix follows a Gaussian distribution, i.e., $\sum \mathcal{CN}\left(0, \sigma_e^2\right)$ where $\sigma_e^2 = \sigma_v^2/5$.

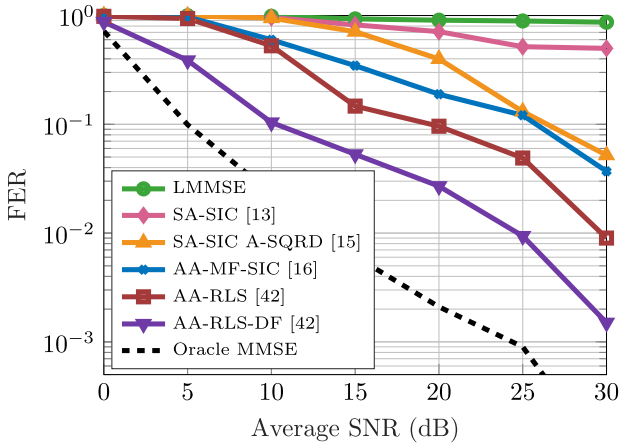
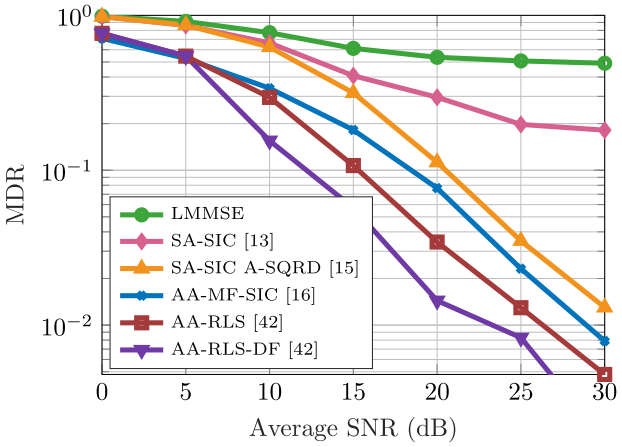
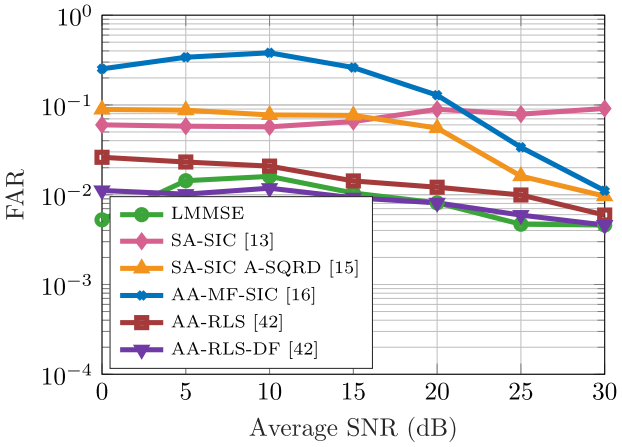


FIGURE 4. Frame error rate vs. Average SNR. Comparison of regularized algorithms for $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols, split into 60 metadata and 68 data. 10^5 Monte Carlo trials.

As the linear MMSE is not designed for the sparse scenario, it presents a poor performance. The unsorted SA-SIC is susceptible to error propagation, thus it does not perform well. A-SQRD and AA-MF-SIC are effective since both consider the activity probabilities, but under imperfect CSI conditions, their performance strongly degrades. On the other hand, as AA-RLS and AA-RLS-DF doe not need an explicit channel estimation, they are more efficient. The interference cancellation performed by the decision-feedback scheme leads to an evident FER gain. The activity error rates are shown in Figs. 5a and 5b, much of the FER gain of the schemes of [42] is due to the high activity detection accuracy of the regularized RLS filters illustrated by the MDR performance.



(a) Missed detection rate vs. Average SNR.



(b) False alarms rate vs. Average SNR.

FIGURE 5. Activity error rates for comparison of regularized algorithms. Simulation parameters: $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols, split into 60 metadata and 68 data. 10^5 Monte Carlo trials.

2) GREEDY DETECTORS

Widely studied in the compressive sensing field, greedy algorithms have been applied as potential solutions to mMTC activity and data detection. As this approach has low complexity and generally only requires termination tuning, that is, the termination of the reconstruction (transmitted vector) has to be adapted for the specific problem instance in order to avoid inaccurate results. The pioneering work [30] applied the well-known OLS and OMP algorithms to an uplink sparse scenario as mMTC. As OMP has a better performance, a lot of improvements of it are available in the literature and some of them were specifically designed for mMTC.

wGOMP

Drawing inspiration from the block-sparse variant of OMP, the GOMP [96], the authors in [33] propose an improvement

where they refine the activity detection, exploiting the channel code. Considering perfect CSI and that the BS has the knowledge of the number of active devices, structuring the receiver in an iterative feedback approach, the main idea of wGOMP is to pass weights $\mathbf{w}$ based on the channel decoding output to the multi-user detection. Repeating iterations until $\mathbf{w}$ values no longer change and the feedback process has converged, at each new step, the weights give the likelihood of activity for each node, improving the activity decision. The weights are introduced in each correlation in the block selection step of GOMP, as

$$\tilde{k} = \arg\max_{k \in \overline{\mathcal{B}}^{(u-1)}} \frac{1}{|\gamma(k)|} \sum_{j \in \gamma(k)} \mathbf{w}_{\{j\}} \frac{|\mathbf{H}_{\{j\}}^{H} \mathbf{r}^{(u-1)}|}{\|\mathbf{H}_{\{j\}}\|_2}, \tag{20}$$

where k is the index of the block, $\mathbf{r}$ is the residual, u the iteration index and $\mathbf{H}$ is the channel matrix. The list of inactive devices is given by $\mathcal{B}$ and γ is the part of the channel matrix that should be considered. These weights allow the choice of devices which are likely active, due to information from channel coding. If no weights are applied, ($\mathbf{w} = 1$), the wGOMP and GOMP are identical.

In order to reduce the complexity of the problem, the work in [33] also considers independent subproblems. The main idea is to apply parallel CS-MUD detectors to each subproblem, detecting each part of the transmitted vector separately. Once all parts have been detected, the estimated symbols can be sorted per node, resulting in the node-specific data vectors, later decoded by the channel decoder. As originally in [33] the authors considered a Viterbi decoder, the weights computation is given by

$$\mathbf{w}_m = 0.5 \frac{\xi_{\mathcal{C}}\left(\tilde{\mathbf{d}}_k\right) - \xi_{\varepsilon}\left(\tilde{\mathbf{d}}_k\right)}{2L}, \forall m \in \gamma(k), \tag{21}$$

where

$$\begin{cases} \xi_{\mathcal{C}}\left(\tilde{\mathbf{d}}_k\right) = \min\limits_{\tilde{\mathbf{d}}_k \in \mathcal{C}} \left|\left|\tilde{\mathbf{d}}_k - \hat{\mathbf{d}}_k\right|\right|_2 & \text{, for all } \mathcal{C} \text{ and} \\ \xi_{\varepsilon}\left(\tilde{\mathbf{d}}_k\right) = \left|\left|\tilde{\mathbf{d}}_k - \mathbf{0}\right|\right|_2 = \left|\left|\tilde{\mathbf{d}}_k\right|\right|_2 & \text{, for zero.} \end{cases} \tag{22}$$

The metric used by the decoder to decide the true hypothesis is the smallest Euclidean distance, returning either the most likely codeword, as determined by the channel decoding, or the all-zero word accordingly.

bcSIC

In order to improve the last approach, the authors in [34] incorporate the most likely codeword in the iterative feedback scheme. Known as block-correlation SIC (bcSIC), the idea is to compute the activity estimation with (20) and performs the LS estimation for the chosen node $\tilde{k}$ followed by the channel decoding for this node. Subsequently, the residual is updated with the most likely codeword of node $\tilde{k}$ for interference cancellation, as

$$\mathbf{r}^{(u)} = \mathbf{r}^{(u-1)} - \mathbf{H}_{\{\gamma(\tilde{k})\}} \hat{\mathbf{d}}_{\tilde{k}} \tag{22}$$

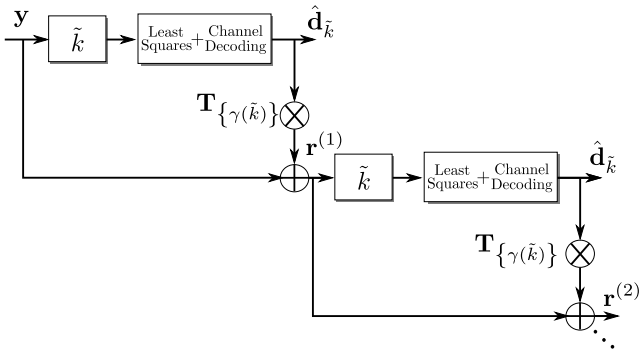


FIGURE 6. Structure of the bcSIC, where each block $\tilde{k}$ contains a GOMP activity detection step.

where, due to this modification, the estimates $\tilde{\mathbf{d}}_k = \mathbf{H}_{\{\gamma(k)\}}^{\dagger} \mathbf{r}^{(u-1)}$ are never re-evaluated, unlike wGOMP that performs an LS estimation for all active nodes $\mathcal{B}^{(u)}$ in each iteration. The structure of bcSIC is shown in Fig. 6.

mSOMP-EXT

Drawing inspiration in other modification of the OMP, the simultaneous orthogonal matching pursuit (SOMP) [97] with extrinsic information transfer (SOMP-EXT) [98], the work in [39] developed an algorithm that performs the joint activity and data detection with no prior knowledge of sparsity and noise levels, just the channel gains. Named mSOMP-EXT [39], this algorithm computes the average LLR of each n-th device, where $\mathcal{L}\left(z_{l,n}^{(t)}\right) = \log\left(\Pr\left[z_{l,n}^{(t)} \,|\, n \in \mathcal{S}\right] / \Pr\left[z_{l,n}^{(t)} \,|\, n \notin \mathcal{S}\right]\right)$. $\mathcal{S}$ indicates the list of active devices, τ is the number of symbols in the frame and t is the time index. The scheme repeats until l is an iteration marker surpasses the number of subcarriers M. As part of the essence of SOMP-EXT, the LLR creates and transfers the extrinsic information through iterations to support detection as, $\forall n \in \{1, \ldots, 2N\} \backslash \hat{\mathcal{S}}_{l-1}$,

$$Z_{l,n} = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathcal{L}\left(z_{l,n}^{(t)}\right), \tag{23}$$

$$E_{l,n} = (\tau - 1) Z_{l,n}, \tag{24}$$

$$A_{l,n} = \begin{cases} 0 & \text{, if } l = 1, \\ \frac{1}{l-1} \sum_{l'=1}^{l-1} E_{l',n} & \text{, if } l \geq 2, \end{cases} \tag{25}$$

$$\Lambda_{l,n} = Z_{l,n} + E_{l,n} + A_{l,n}. \tag{26}$$

As this scheme works with real values, the estimated support list $\hat{\mathcal{S}}_l$ has $2N$ elements, which is updated as follows

$$\hat{n}_l = \arg\max_{n \in \{1, \cdots, N\} \backslash \hat{\mathcal{S}}_{l-1}} \left(\Lambda_{l,n} + \Lambda_{l,n+N}\right), \tag{27}$$

$$\hat{\mathcal{S}} = \hat{\mathcal{S}}_{l-1} \cup \{\hat{n}_l, \hat{n}_l + N\}. \tag{28}$$

Using the estimated list of active devices, the soft estimation and the residual are computed as $\tilde{\mathbf{x}}_{\hat{\mathcal{S}}_l}^{(t)} = (\mathbf{H}_{\hat{\mathcal{S}}_l}^{(t)})^{\dagger} \mathbf{y}^{(t)}$ and $\mathbf{r}_l^{(t)} = \mathbf{y}^{(t)} - \mathbf{H}_{\hat{\mathcal{S}}_l}^{(t)} \tilde{\mathbf{x}}_{\hat{\mathcal{S}}_l}^{(t)}$. In the next iteration, the vector $z_{l,n}^{(t)}$ is computed using the last residual and the procedure repeats.

The authors in [39] used an LLR approximation in order to not require the knowledge of the sparsity level K and the

noise variance σ_v^2. As in the work [39] was considered Υ-ary QAM symbols, the following LLR approximation is taken into account:

$$\mathcal{L}\left(z_{l,n}^{(t)}\right) = \log\left(\frac{\frac{1}{\sqrt{\Upsilon}}\sum_{\upsilon=0}^{\sqrt{\Upsilon}-1}\frac{1}{\sqrt{2\pi}\sigma_{l,1}}\exp\left(-\frac{\left(z_{l,n}^{(t)}-q_\upsilon\right)^2}{2\sigma_{l,1}^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_{l,0}}\exp\left(-\frac{\left(z_{l,n}^{(t)}\right)^2}{2\sigma_{l,0}^2}\right)}\right), \tag{29}$$

where $q_\upsilon = \frac{2\upsilon-\sqrt{\Upsilon}+1}{\sqrt{2}}$, $\upsilon = 0, 1, \ldots, \sqrt{\Upsilon}-1$, is the in-phase component of an Υ-ary QAM symbol that corresponds to a nonzero element of $\tilde{\mathbf{x}}^{(t)}$.

For a sufficiently large number of subcarriers M, $\sigma_{l,0}^2$ and $\sigma_{l,1}^2$ are approximated as $\sigma_{l,0}^2 \approx \sigma_{l,1}^2 \approx \sigma_x^2$, where

$$\sigma_x^2 = \frac{1}{\sqrt{\Upsilon}}\sum_{\upsilon=0}^{\sqrt{\Upsilon}-1} q_h^2 = \frac{\Upsilon-1}{6} \tag{30}$$

is the average power of nonzero elements of $\mathbf{x}^{(t)}$. This LLR approximation and a threshold parameter empirically obtained for a stopping criterion, composes the modification of mSOMP-EXT.

TA-BSASP

An improvement of the classical subspace pursuit (SP) algorithm is presented in [99]. The threshold aided block sparsity adaptive subspace pursuit (TA-BSASP) reconstructs the sparse vector by exploiting the inherent block sparsity, as the authors vectorized all data transmitted in different time slots. TA-BSASP uses a stopping criterion based on the AWGN noise, given by

$$\min\left\{\left|\left|\tilde{\mathbf{c}}^{(l)}[m]\right|\right|_2^2\right\} \leq \tau\, P_{th}, \tag{31}$$

where $\tilde{\mathbf{c}}^{(t)}$ is the estimated solution of the vectorized transmitted vector, as shown in Fig. 7, m is an index of the support set, τ is the number of elements in the same block and P_{th} is the AWGN noise floor, selected experimentally.

Until the stopping criterion is met, TA-BSASP updates the support estimate list, with the time index t, as

$$\Lambda = \Gamma^{(t-1)} \bigcup \Xi\left(\|\mathbf{D}^H[n]\,\mathbf{r}^{(t-1)}\|_2, s\right), \forall\, n = 1, 2, \ldots, N \tag{32}$$

where Ξ is a set, whose elements are the indices of the largest s elements of its argument. s is initialized as one and determines how many devices the algorithm will deal per iteration. $\mathbf{D}$ is a sparse version of the channel matrix $\mathbf{H}$, given by

$$\mathbf{D} = \begin{bmatrix} H(1,1)\,\mathbf{I}_\tau & H(1,2)\,\mathbf{I}_\tau & \ldots & H(1,\tau N)\,\mathbf{I}_\tau \\ H(2,1)\,\mathbf{I}_\tau & H(2,2)\,\mathbf{I}_\tau & \ldots & H(2,\tau N)\,\mathbf{I}_\tau \\ \vdots & \vdots & \ddots & \vdots \\ H(N,1)\,\mathbf{I}_\tau & H(N,2)\,\mathbf{I}_\tau & \ldots & H(N,\tau N)\,\mathbf{I}_\tau \end{bmatrix} = \mathbf{H} \otimes \mathbf{I}_\tau. \tag{33}$$

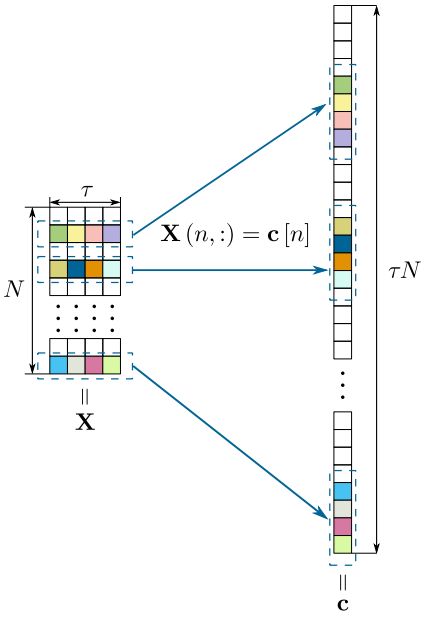


FIGURE 7. Vectorized version of transmitted symbols in the TA-BSASP signal model.

Following this step, TA-BSASP computes the LS estimate $\mathbf{w}$ of the set Λ and performs support pruning, as

$$\hat{\Gamma}^{(t)} = \Xi\left(\|\mathbf{w}[n]\|_2, s\right), \quad \forall\, n = 1, 2, \ldots, N. \tag{34}$$

Then, the algorithm proceeds with the signal estimate $\tilde{\mathbf{c}}^{(t)}\left[\hat{\Gamma^{(t)}}\right]$, update of residual $\mathbf{r}^{(t)}$ and if $\|\mathbf{r}^{(t)}\|_2 < \|\mathbf{r}^{(t-1)}\|_2 <$, the support pruning is updated ($\Gamma^{(t)} = \hat{\Gamma}^{(t)}$) as the iterative index ($t = t+1$). Otherwise, the sparsity level is updated, with $s = s+1$. When the stopping criteria in (31) is met, the algorithm stops and the data are recovered.

Performance evaluation

In order to verify the efficiency of the greedy algorithms, we modify the system model of (7), as the schemes does not consider the metadata, but include in the signal model spreading sequences. In this way, the general received vector for this performance analysis is given by

$$\mathbf{y}^{(t)} = \sum_{n=1}^{N} \mathrm{diag}\left(\mathbf{h}_n^{(t)}\right)\mathbf{s}_n x_n^{(t)} + \mathbf{v}^{(t)} = \mathbf{G}^{(t)}\mathbf{x}^{(t)} + \mathbf{v}^{(t)}, \tag{35}$$

where at a time slot l ($1 \leq l \leq \tau$), a transmitted symbol $x_n^{(t)}$ of active user n is spread onto M subcarriers using a unique spreading sequence $\mathbf{s}_n \in \mathbb{C}^M$. The channel gain $h_{m,n}$ and noise vector are computed as described in Subsection III-A.

Each frame has 128 data symbols and for simplicity, we assume as a stopping criterion for wGOMP and bcSIC the number of active devices, even if this is unrealistic. For mSOMP-EXT we choose $v_{th} = -0.4$, as in [39]. In TA-BSASP, the P_{th} used for the stopping criterion is the same as in [99], 0.68, 0.51, 0.48, 0.38, and 0.28, respectively, at the SNR of 0 dB, 2 dB, 4 dB, 6 dB, 8 dB.

As none of the greedy solutions perform channel estimation, for this comparison we considered perfect CSI. The Frame Error Rate (SER) performance averaged over 10^5 runs is depicted in Fig. 8. Although wGOMP and bcSIC originally use channel coding, we do not use it so that the comparison with other techniques is fair. As each frame has 128 data symbols, we divided wGOMP in $N_{\text{subp}} = 16$ subproblems, thus considering 8 symbols for each device ($\tau_{\text{seg}} = 128/N_{\text{subp}} = 8$).

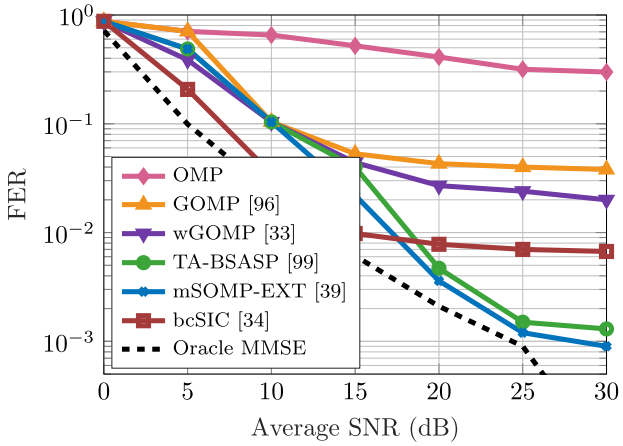
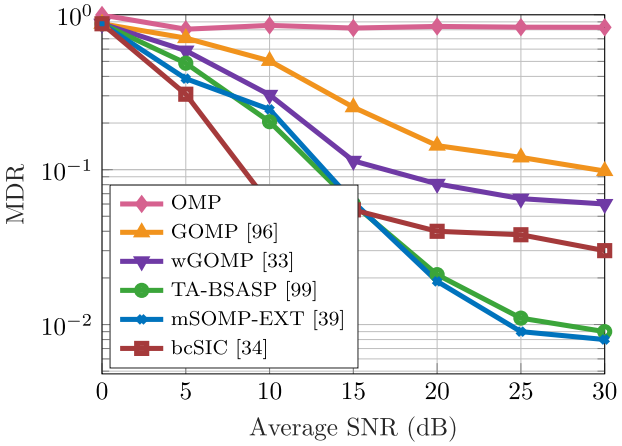


FIGURE 8. Frame error rate vs. Average SNR. Comparison of greedy algorithms for $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols, split into 60 metadata and 68 data. 10^5 Monte Carlo trials.

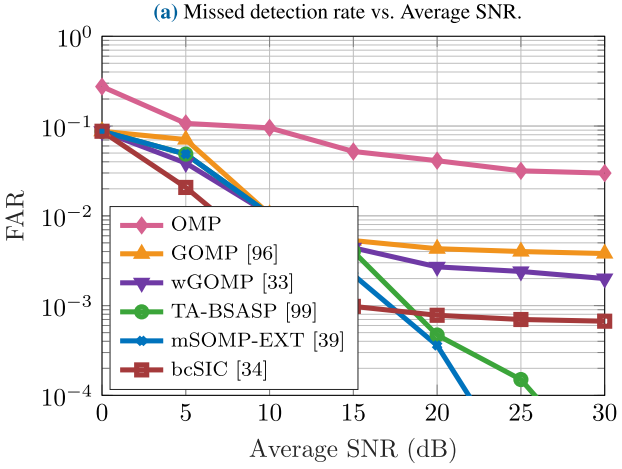
It is possible to observe that the well-known OMP has the worst performance, as it does not include any refined activity detection scheme. As wGOMP exploits block-sparsity across all subproblems via the feedback of activity estimation based on the output of the channel decoding (or quantization of the soft estimation) wGOMP has a lower FER than GOMP for high SNRs. Since each subproblem only considers changed partial block-sparsity, it enables the correction of some activity errors in the feedback process. The interference cancellation incorporated in bcSIC algorithm reduces its FER comparing to wGOMP and reaches the error floor, due to error propagation, in a low SNR value, approaching the curve to the lower bound. Showing better results in high SNRs, mSOMP-EXT and TA-BSASP have worse performance than bcSIC for low SNR due to poor activity detection compared to the bcSIC as seen in Figs. 9a and 9b. We can also conclude that the FER performance of OMP, GOMP and wGOMP are primarily limited by activity errors, while bcSIC is primarily limited by error propagation.

3) MESSAGE PASSING DETECTORS

Initially proposed by Donoho, Maleki and Montanari [20]–[22], the application of factor graphs to CS problems inspired many other works. As this class of iterative thresholding algorithms considers the posteriori distribution of the signal to be reconstructed, the usage of factor graphs to



(a) Missed detection rate vs. Average SNR.



(b) False alarms rate vs. Average SNR.

FIGURE 9. Activity error rates for comparison of greedy algorithms. Simulation parameters: $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols, split into 60 metadata and 68 data. 10^5 Monte Carlo trials.

marginalize the joint probability distribution of the received vector enabled its application in the communications area.

As the sensing matrix in CS problems is a dense matrix, the fundamental factor graph is fully connected. Accordingly, messages in fully connected graphs are functions and their computation is tricky. Still, it is common practice in loopy Belief Propagation to approximate messages by prototype functions that take after Gaussian density functions which can be described only by its mean and variance. Thus, message passing summarizes to the exchange of the parameters of a function instead of the function itself.

Non-coherent scenario

The works in [48] and [49] propose a modification in the AMP in a non-coherent transmission scenario. The main idea of the non-coherent approach is that the transmitted data bits are embedded in the index of the transmitted pilot sequence of each active device. Therefore, if the algorithm correctly detects the active devices, consequently it will detect the data. The disadvantage of this method is that the BS is required to allocate for each device not just one metadata sequence per

frame, but a set of 2^J sequences when J bits are transmitted by each device. Due to the massive number of devices requiring connection at the same time and non-orthogonal metadata sequences, the probability of two devices having identical sequences is seen as the probability of frame collisions, which should be taken account in the performance. Considering a system model similar to (1), both works [48] and [49], incorporate the sparsity not only in the transmitted vector, but in the channel matrix. In this way, the sparse structure of $\mathbf{H}$ has the rows corresponding to inactive users are zero. Thus, the activity detection problem reduces to finding the non-zero rows of the channel matrix $\mathbf{H}$. In this way, the signal model in this case is given by

$$\mathbf{Y} = \sqrt{\tau_\phi}\mathbf{\Phi}\mathbf{H} + \mathbf{V}, \tag{36}$$

where the metadata $\mathbf{\Phi}$ and channel $\mathbf{H}$ matrices are

$$\mathbf{\Phi} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_N] \in \mathbb{C}^{\tau \times N2^J} \text{ and } \mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_N]^H,$$

in which the channel matrix between the n-th device and the BS is $\mathbf{H}_k = \left[\delta_{k,1}\mathbf{h}_k, \ldots, \delta_{k,2^J}\mathbf{h}_k\right] \in \mathbb{C}^{M\times 2^J}$ and δ is the parameter that defines if the device is active or inactive, as in (2). Due to the rewritten signal model, the channel matrix now has 2^J more rows than the previous one but with the same number of active devices, which increases the sparsity level of the system. In order to exploit these properties, the works in [48] and [49] proposes a modified AMP algorithm.

M-AMP

The authors in [48] presents the M-AMP where, for the active device n, the estimate of the row of $\overline{\mathbf{H}} = \left[\overline{\mathbf{h}}_{k,1}, \ldots, \overline{\mathbf{h}}_{k,2^J}\right] \in \mathbb{C}^{M\times 2^J}$ corresponding to the i-th metadata sequence is $\hat{\overline{\mathbf{h}}}_{k,i}$, with $\overline{\mathbf{h}}_{k,i} = \delta_{k,i}\mathbf{h}_k$. Then, with the k-th device transmitting the pilot sequence i', the estimate is

$$\overline{\mathbf{h}}^l_{k,i} = \begin{cases} \left(\mathbf{h}_k + \left(\mathbf{\Sigma}^l\right)^{\frac{1}{2}}\mathbf{w}\right) \sim \mathcal{CN}\left(0, \eta_k\mathbf{I} + \mathbf{\Sigma}^l\right), & i = i', \\ \left(\left(\mathbf{\Sigma}^l\right)^{\frac{1}{2}}\mathbf{w}\right) \sim \mathcal{CN}\left(0, \mathbf{\Sigma}^l\right), & i \neq i'. \end{cases} \tag{37}$$

where the update of the state evolution is given by $\mathbf{\Sigma}^{l+1} = \frac{\sigma_v^2}{\tau_\phi}\mathbf{I} + \frac{N}{\tau_\phi}\mathbb{E}\{\mathbf{e}\mathbf{e}^H\}$, $\mathbf{e} = \eta(\mathbf{h}_\beta - (\mathbf{\Sigma}^l)^{\frac{1}{2}}\mathbf{w}) - \mathbf{h}_\beta$ and $\mathbf{w}$ is a complex Gaussian vector with unit variance and t is the iteration index. $\mathbf{h}_\beta$ has the following distribution:

$$p_{\mathbf{h}_\beta} = (1-\rho)\,\delta + \rho\,\mathcal{CN}(0, \beta\mathbf{I}), \tag{38}$$

$\mathcal{CN}(0, \beta\mathbf{I})$ is the distribution of the channel vector of the active device and δ is the dirac Delta at zero corresponding to the inactive device channel distribution. The expectation in the update of the state evolution expression is taken with respect to β, parameter of (6). Thereby, the resulting modified denoiser is

$$\tilde{\eta}_{l,n}\left(\hat{\overline{\mathbf{h}}}^l_n\right) = f\left(\psi\left(\hat{\overline{\mathbf{h}}}^l_k\right)\right)\eta_{l,n}\left(\hat{\overline{\mathbf{h}}}^l_n\right), \tag{39}$$

where $\eta_{l,n}$ is the denoising function equivalent for the MMSE [48]. The idea of the modification is to design a denoiser capable to suppress the metadata sequences that do not belong to the evaluated device. For this, a soft-thresholding function is used, the sigmoid function, defined by

$$f\left(\psi\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right)\right) = \frac{1}{1 + \exp\left(-c\left(\psi\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right) - \frac{1}{2}\right)\right)}, \tag{40}$$

where c is a parameter that determines the sharpness of the sigmoidal transition, the coefficient $\psi\left(\hat{\overline{\mathbf{x}}}^l_{k,i}\right)$ is seen as a measure of the proportional likelihood of a given sequence allocated to device k and is given by

$$\psi\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right) = \frac{\Lambda\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right)}{\sum_{i'=1}^{2J}\Lambda\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right)}, \tag{41}$$

where the likelihood function is

$$\Lambda\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right) = \frac{|\Sigma^l|}{|\beta_k\mathbf{I} + \Sigma^l|}q\left(\hat{\overline{\mathbf{h}}}^l_{k,i}; \Sigma^l\right)^{-1}, \text{ and} \tag{42}$$

$$q\left(\hat{\overline{\mathbf{h}}}^l_{k,i}; \Sigma^l\right) = \exp\left(-\left(\hat{\overline{\mathbf{h}}}^l_{k,i}\right)^H\left(\left(\Sigma^l\right)^{-1} - \left(\Sigma^l + \beta_n\mathbf{I}\right)^{-1}\right)\hat{\overline{\mathbf{h}}}^l_{k,i}\right). \tag{43}$$

As the main idea is that only a single row corresponding to a device may be non-zero as it is impossible for a device to transmit both metadata sequences concurrently, the authors of [48] proposed M-AMP.

NSD-AMP

Also focusing on the non-coherent transmission, the work in [49] proposes another modification to the AMP. The idea is to develop a section-wise equivalent model, that decouples the estimation in different sections, in this way allowing the design of the section-wise Bayes-optimal denoiser for the AMP, minimizing the MSE section by section. Starting from the classical AMP,

$$\mathbf{H}^{l+1}_n = \eta_{l,n}\left((\mathbf{\Phi}_n)^H\mathbf{R}^l + \mathbf{H}^l_n\right), \quad n = 1, \cdots, N, \tag{44}$$

$$\mathbf{R}^{l+1} = \mathbf{Y} - \mathbf{\Phi}\mathbf{H}^{l+1} + \frac{\mathbf{R}^l}{\tau}\sum_{n=1}^{N}\eta'_{l,n}\left((\mathbf{\Phi}_n)^H\mathbf{R}^l + \mathbf{H}^l_n\right), \tag{45}$$

and using the analysis presented in [100], the authors argue that the output of the denoiser applied to the residual $(\mathbf{\Phi}_n)^H\mathbf{R}^l + \mathbf{H}^l_n$, as in (44), is statistically equivalent to the output of applying the denoiser to

$$\hat{\mathbf{H}}_n = \mathbf{H}_n + \mathbf{V}_n\mathbf{\Sigma}_l^{\frac{1}{2}}, \tag{46}$$

which is called section-wise equivalent model. Based on the equivalent model in (46), the section-wise MMSE denoiser, in other words, the equivalent of (44) and (45) is given by

$$\eta_{l_n}\left(\hat{\mathbf{H}}^l_n\right) = \left[\overline{\omega}_{n,1}\mathbf{\Theta}_n\hat{\mathbf{h}}^l_{n,1}, \cdots, \overline{\omega}_{n,1}\mathbf{\Theta}_n\hat{\mathbf{h}}^l_{n,1}\right]^T \tag{47}$$

in which

$$\boldsymbol{\Theta}_n = \beta_n \left(\beta_n \mathbf{I} + \boldsymbol{\Sigma}_l\right)^{-1}, \tag{48}$$

$$\overline{\omega}_{n,i} = \frac{\exp\left(M\left(\pi_{n,i} - \phi_n\right)\right)}{\sum_{j=1}^{2^J} \exp\left(M\left(\pi_{n,j} - \phi_n\right)\right) + 2^J\left(\frac{1-\rho}{\rho}\right)}, \tag{49}$$

$$\pi_{n,i} = \frac{\left(\hat{\mathbf{h}}_{n,i}^{l}\right)^H \left(\boldsymbol{\Sigma}_l^{-1} - \left(\boldsymbol{\Sigma}^{-1} + \beta_n \mathbf{I}\right)^{-1}\right) \hat{\mathbf{h}}_{n,i}^{l}}{M}, \tag{50}$$

$$\phi_n = \frac{\log\left(\left|\mathbf{I} + \beta_n \boldsymbol{\Sigma}_l^{-1}\right|\right)}{M}. \tag{51}$$

where ρ refers to the probability of being active of each device (equal for all) and β_n are the large scale coefficients given by (6). Besides the section-wise MMSE denoiser, the authors of [49] describe how to decode the embedded data. After t iterations with (44) and (45) the following threshold is computed

$$\mathcal{M}_{n,i} = \left(\frac{1}{\varkappa_t^2}\frac{1}{\beta_n + \varkappa_l^2}\right)\frac{\boldsymbol{\zeta}_{n,i}^H \boldsymbol{\zeta}_{n,i}}{M} - \phi_n, \qquad \forall i, n \tag{52}$$

$$\phi_n = \log\left(1 + \frac{\beta_n}{\varkappa_l^2}\right) \tag{53}$$

where $\boldsymbol{\zeta}_{n,i}$ denotes the i-th row of the matrix $\left(\boldsymbol{\Phi}_n\right)^H \mathbf{R}^l + \mathbf{H}_n^l$, as in (44) and $\varkappa_t^2$ is iteratively obtained using the scalar form of state evolution equations given by

$$\varkappa_0^2 = \frac{1}{\mathrm{SNR}} + \frac{\rho}{\epsilon}\mathbb{E}\left[\beta\right] \tag{54}$$

$$\varkappa_{l+1}^2 = \frac{1}{\mathrm{SNR}} + \frac{1}{\epsilon}\sum_{i=1}^{2^J}\mathbb{E}\left[\overline{\omega}_{\beta,i}\theta_\beta \varkappa_l^2\right] + \frac{1}{\epsilon}\sum_{i=1}^{2^J}\mathbb{E}\left[\Upsilon_{\beta,i}^{l}\right] \tag{55}$$

with all expectations with respect to β, $\mathbf{H}_\beta$ and $\mathbf{V}$ and

$$\theta_\beta = \frac{\beta}{\beta + \varkappa_l^2}, \tag{56}$$

$$\Upsilon_{\beta,i}^{l} = \overline{\omega}_{\beta,i}\left(1 - \overline{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta + \varkappa_l^2}\right)^2 \frac{\left(\hat{\mathbf{h}}_{\beta,i}^{l}\right)^H \hat{\mathbf{h}}_{\beta,i}^{l}}{M}, \tag{57}$$

$$\hat{\mathbf{H}}_\beta^l = \left[\hat{\mathbf{h}}_{\beta,1}^{l}, \cdots, \hat{\mathbf{h}}_{\beta,2^J}^{l}\right]^T + \varkappa_l \mathbf{V}, \tag{58}$$

where $\epsilon = \tau/N$ is a positive value when $(\tau, N) \to \infty$. Thus, $\mathcal{M}_{n,i}$ is evaluated as

$$i_n^* = \arg\max_i \mathcal{M}_{n,i}, \qquad \forall i, n. \tag{59}$$

Then the support vector $\boldsymbol{\delta}_n = \left[\delta_{n,1}, \cdots, \delta_{n,2^J}\right]^T$ of the n-th device is a vector of zeros if $\mathcal{M}_{n,i_n^*} \geq 0$ or a vector of zeros with an 1 in the i_n^* element, otherwise.

Performance evaluation

As in this section we have two different approaches to deal with the problem, we have also two simulation scenarios, in order to have the fairest possible comparison. Firstly, we compare the activity detection performance of the non-coherent algorithms. As the goal of these algorithms is to improve the denoiser, we simulated a scenario with 100 devices, 50 receive antennas and 200 metadata sequences for M-AMP and NSD-AMP, while the well-known AMP has 100 metadata sequences, as suggested in [48] and [49]. If the detector determines that one of the metadata sequences corresponds to one assigned to a device, then that device is detected as active, independently of whether an information bit is transmitted. In all algorithms, the number of iterations is fixed in 30, the number of embedded data bits transmitted by each active device is 4($J = 2$) and each device has a activity probability drawn uniformly at random in [0.1, 0.3]. As shown in Fig. 10, AMP has a better performance than the M-AMP, as expected. Besides deal with more metadata sequences, the modification of the denoiser is a function outside of the standard denoiser, which carries less information than the scheme of NSD-AMP, that incorporates the system statistics in the new denoiser. Another important point is that the NSD-AMP has the information of the probability of being active of each device, which is not available for M-AMP and AMP. Since the setup to include the bits of each algorithm is different and this is not the main goal of the paper, the evaluation of the non-coherent boils down to the activity detection performance.

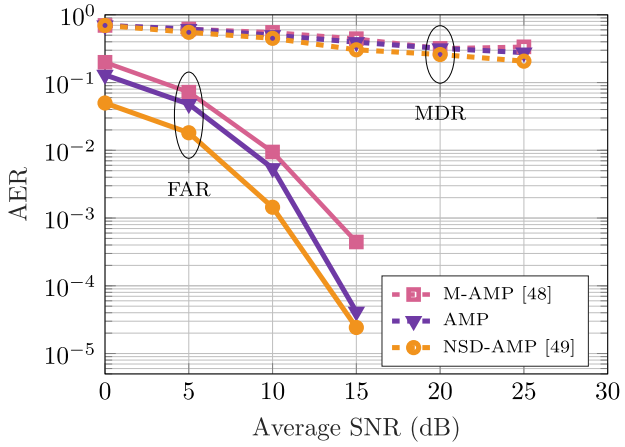


FIGURE 10. Activity error rate vs. Average SNR. Comparison of message-passing algorithms for $N = 100$, $M = 50$ with ρ_n drawn uniformly at random in [0.1, 0.3] which M-AMP and NSD-AMP consider 200 metadata sequences and detect devices along with a single embedded information bit. 10^5 Monte Carlo trials.

Coherent scenario

Returning to the coherent scenario, the works in [18], [25], [31] and uses the message-passing approach to achieve the activity and signal detection.

Joint-EM-AMP

The authors of [25] proposes the Joint-EM-AMP, solution that uses the expectation maximization (EM) algorithm to estimate the activity of devices. As EM is an iterative algorithm that increases the likelihood probability of each

iteration, it guarantees convergence to at least a local maximum of the likelihood function $f\left(\mathbf{y}_j|\lambda_{n,j}\right)$. Thus, this approach updates the activity estimate with the detection of each j symbol of the received frame. Considering a system model similar to (7) with $r_{n,j}$ being the estimated mean of $x_{n,j}$ by decoupling of AMP and $\phi_{n,j}$ the effective noise variance, the posterior probability of $x_{n,j}$ is expressed as

$$f\left(x_{n,j}|r_{n,j},\lambda_{n,j}\right)=\frac{f\left(r_{n,j}|x_{n,j}\right)p\left(x_{n,j}|\lambda_{n,j}\right)}{f\left(r_{n,j}|\lambda_{n,j}\right)} \tag{60}$$

where

$$f\left(r_{n,j}|\lambda_{n,j}\right)=\sum_{x_{n,j},\in\mathcal{A}} f\left(r_{n,j}|x_{n,j}\right)p\left(x_{n,j}|\lambda_{n,j}\right), \tag{61}$$

$$f\left(r_{n,j}|x_{n,j}\right)=\mathcal{CN}\left(r_{n,j}-x_{n,j},\phi_{n,j}\right), \tag{62}$$

$$\begin{aligned}p\left(x_{n,j}|\lambda_{n,j}\right)&=\left(1-\lambda_{n,j}\right)\delta\left(x_{n,j}\right)\\&\quad+\lambda_{n,j}\sum_{i=1}^{|\mathcal{A}|}p_{\mathcal{A}}^{i}\delta\left(x_{n,j}-d_i\right).\end{aligned} \tag{63}$$

Here, $\delta(\cdot)$ is the Dirac delta function and (63) is the prior information of the transmitted discrete symbols conditioned on user activity parameter $\lambda_{n,j}$, obtained iteratively by the expectation maximization algorithm. d_i is the i-th symbol of the modulation constellation $\mathcal{A}$, $\mathcal{CN}(\cdot)$ indicates a complex Gaussian distribution and the estimates of the posterior mean and variance of $x_{n,j}$ are given by

$$\hat{x}_{n,j}=\sum_{x_{n,j}\in\mathcal{A}} x_{n,j}f\left(x_{n,j}|r_{n,j},\lambda_{n,j}\right), \tag{64}$$

$$v_{n,j}=\sum_{x_{n,j}\in\mathcal{A}}|x_{n,j}|^2x_{n,j}f\left(x_{n,j}|r_{n,j},\lambda_{n,j}\right)-|\hat{x_{n,j}}|^2. \tag{65}$$

Thus, the updated $\lambda_{n,j}$ is then obtained as

$$\lambda_n^{t+1}=\frac{1}{\tau}\sum_{j=1}^{\tau}\sum_{x_{n,j}\in\mathcal{A}} f\left(x_{n,j}|\mathbf{y}_{n,j},\lambda_{n,j}^{t}\right) \tag{66}$$

where

$$f\left(x_{n,j}|\mathbf{y}_{n,j},\lambda_{n,j}^{t}\right)=\frac{\sum_{i=1}^{\mathcal{A}}p_{\mathcal{A}}^{i}\delta\left(x_{n,j}-d_i\right)-\delta\left(x_{n,j}\right)}{p\left(x_{n,j}|\lambda_{n,j}\right)}. \tag{67}$$

Thus, the Joint-EM-AMP performs the classical AMP algorithm and iteratively estimates the activity of devices with the expectation propagation approach. After t predefined iterations, if λ_n is greater than 0.5 the k-th device is considered active and inactive, otherwise. The data of active devices are recovered from the estimated values of $x_{n,j}$ in this approach.

EM-BSBL

The work in [18] applies the expectation maximization in a different way. Without requiring the priori knowledge of the activity factor, the work in [18] modifies the block sparse Bayesian learning (BSBL) [101] that considers the prior of the row sparsity property and the column coherence of each device. This Bayesian inference method make use of reasonable priors and a set of hyperparameters to control the estimated signals which can be learned from the training progress via expectation maximization.

This method vectorizes the transmitted frame, from the simplified signal model of (7) as shown in Fig.7. In order to estimate the vectorized transmitted frame $\bar{\mathbf{x}}=\text{vec}(\mathbf{X})$,two hyperparameters are included in the system, $\boldsymbol{\gamma}$ and $\mathbf{C}$. The first one, $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_N)$ is a non-negative hyperparameter that controls the row sparsity property of $\mathbf{X}$ and $\mathbf{C}$ is a positive definite matrix that controls the column coherence of $\mathbf{X}$. The likelihood of the received signal $\mathbf{y}$ is

$$p\left(\mathbf{y}|\mathbf{x};\sigma_v^2\right)=\mathcal{CN}\left(\mathbf{y}|\mathbf{H}\mathbf{x},\sigma_v^2\mathbf{I}\right). \tag{68}$$

The authors assume zero-mean Gaussian *prior* for the transmitted signals $\mathbf{x}$, and the variance is composed of $\boldsymbol{\gamma}$ and $\mathbf{C}$, therefore $p(\mathbf{x};\gamma_n,\mathbf{C}_n)=\mathcal{CN}(\mathbf{x}|0,\boldsymbol{\Sigma}_0),\forall n$, where the variance $\boldsymbol{\Sigma}_0=\text{diag}(\gamma_1\mathbf{C}_1,\cdots,\gamma_N\mathbf{C}_N)$. Then, the posterior distribution of $\mathbf{x}$ is computed as

$$p\left(\mathbf{x}|\mathbf{y};\sigma_v^2,\gamma,\mathbf{C}\right)=\frac{p\left(\mathbf{y}|\mathbf{x};\sigma_v^2\right)p(\mathbf{x};\gamma,\mathbf{C})}{\int p\left(\mathbf{y}|\mathbf{x};\sigma_v^2\right)p(\mathbf{x};\gamma,\mathbf{C})\,d\mathbf{x}} \tag{69}$$

$$=\mathcal{CN}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{70}$$

where the posterior mean and covariance are respectively

$$\hat{\mathbf{x}}=\boldsymbol{\mu}=(\sigma_v^2)^{-1}\boldsymbol{\Sigma}\mathbf{H}^H\mathbf{y} \tag{71}$$

$$\boldsymbol{\Sigma}=\left(\Sigma_0^{-1}+\sigma_v^2\mathbf{H}^H\mathbf{H}\right)^{-1}, \tag{72}$$

and the hyperparameters $\boldsymbol{\gamma}$, $\mathbf{C}$, and σ_v^2 estimation via the iterative EM method are summarized as follows

$$\begin{aligned}\gamma_i&=\frac{1}{\tau}\text{tr}\left(\mathbf{C}^{-1}\boldsymbol{\Sigma}_x^i+\boldsymbol{\mu}\hat{\mathrm{i}}_x(\boldsymbol{\mu}_x^i)^H\right),\\\mathbf{C}&=\frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{\Sigma}_{\boldsymbol{x}}^{\boldsymbol{i}}\,\boldsymbol{\mathcal{C}}\,\boldsymbol{\mu}_{\boldsymbol{x}}^{\boldsymbol{i}}(\boldsymbol{\mu}_{\boldsymbol{x}}^{\boldsymbol{i}})^{\boldsymbol{H}}}{\gamma_i}\\\sigma_v^2&=\frac{\|\mathbf{y}-\mathbf{D}\boldsymbol{\mu}\|_2^2+\sigma_2^v\left(N\tau-\text{tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\right)\right)}{M\tau},\end{aligned} \tag{73}$$

where $\boldsymbol{\mu}_x^i$ is a $\tau\times 1$ vector and denotes the i-th block of $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_x^i$ is a $\tau\times\tau$ matrix and denotes the i-th block of $\boldsymbol{\Sigma}$. As a stopping criterion, if the mean $\|\boldsymbol{\mu}^t-\boldsymbol{\mu}^{t-1}\|_2$ is lower than a prescribed threshold, as 10^{-6}, the algorithm stops.

CS-MPA

The work in [31] proposed a mixture of techniques, a CS approach to realize the user activity and a message-passing method to the data detection. Named CS-MPA detector, the algorithm is divided in two parts. Initially, the signal model considered is similar to (36) in terms of the channel matrix contains the sparsity of the system. The compressive sampling matching pursuit (CoSaMP) [102] algorithm is used to estimate the number of active users and the maximum iteration number. Thereby, the algorithm perform a least squares approach and computes a residual in order to aid in the next iteration. After all iterations, the estimated channel columns different from zero are considered to be related to an active device.

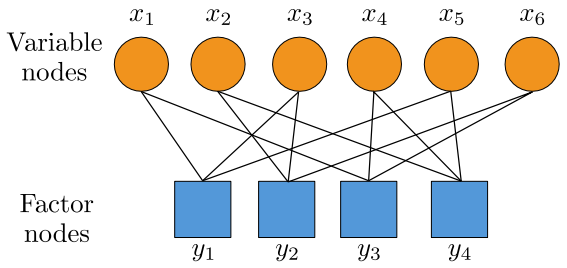


FIGURE 11. Factor graph representation of MPA-based receiver including variable and factor nodes.

For the data detection, CS-MPA considers a simple factor graph, as in Fig. 11, in which transmitted symbols for all devices are variable nodes $\{x_n\}_{n=1}^{N}$ and the observations are factor nodes $\{y_m\}_{m=1}^{M}$. In the factor graph, there exists an edge between a variable node and a factor node if and only if the device is active. In MPA, the marginal distribution of a variable node can be regarded as the product of the messages received by that node as it is represented by

$$\mu_{m\to n}^{t}(x_n) \propto \sum_{x_i|i\in M(m)\backslash n} \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{1}{2\sigma_v^2}\|y_m - h_{m,n}x_n - \sum_{i\in M(m)\backslash n} h_{m,i}x_i\|^2\right\} \prod_{i\in M(m)\backslash n} \mu_{i\to m}^{t-1}(x_i), \quad (74)$$

$$\mu_{n\to m}^{t}(x_n) \propto \prod_{i\in M(n)\backslash m} \mu_{i\to n}^{t-1}(x_n), \quad (75)$$

where $\mu_{m\to n}^{t}(x_n)$ denotes the message passed from factor node y_m to variable node x_n in the i-th iteration, $\mu_{n\to m}^{t}(x_n)$ is the message passed from the other direction and and $M(m)\backslash n$ presents all elements in $M(m)$ except for n. After all iterations, the (approximate) marginal probability distribution of x_n is computed by

$$p(x_n) \propto \prod_{i\in M(n)} \mu_{i\to n}^{T}(x_n). \quad (76)$$

Therefore, each estimated symbol of active users is taken from the modulation alphabet with the maximum marginal probability.

Performance evaluation

For the coherent scenario, AMP, CS-MPA, EM-BSBL and Joint-EM-AMP are evaluated. Considering 128 devices transmitting frames with 128 QPSK symbols to a single base station equipped with 64 antennas, its is possible to notice that EM-BSBL and Joint-EM-AMP shows better performance than other schemes. Due to the expectation-maximization procedure, the activity estimation of each algorithm is more refined. Despite being based on AMP, the consideration of posterior probability of the transmitted symbols to reach the update equation of the activity parameter improves considerably the estimation performance. On the other hand, CS-MPA uses a simple scheme to estimate the activity and the classical message-passing algorithm to detect the transmitted symbols, which results in intermediate performance.

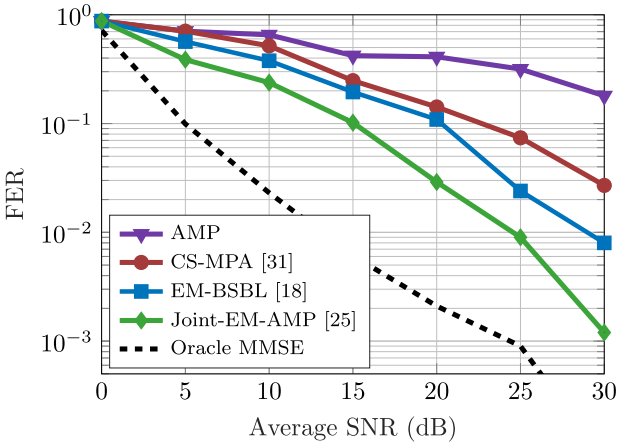


FIGURE 12. Frame error rate vs. Average SNR. Comparison of message-passing algorithms for $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3] and active devices transmitting frames with 128 QPSK symbols. 10^5 Monte Carlo trials.

B. ACTIVITY DETECTION AND CHANNEL ESTIMATION

1) MESSAGE PASSING SOLUTIONS

The message-passing approach has been also applied to obtain accurate channel estimation with a low computational complexity. As most of channel models consider Gaussian approximations, the use of message passing approaches is attractive.

AMP with MMSE denoiser

One of the first works [43] consists of a robust technique, where a minimum mean-squared error (MMSE) denoiser in the vector AMP algorithm is designed for user activity detection and channel estimation based on statistical channel information. Considering a signal model similar to (36), the main difference between the MMSE denoiser in (47) is the component of (48), which is given by

$$\overline{\omega}_n = \frac{1}{1 + \frac{1-\epsilon}{\epsilon}\exp\left(-M\left(\pi_n - \phi_n\right)\right)} \quad (77)$$

being updated in each state. For the device activity detection, a threshold strategy is adopted as

$$\begin{cases} 1, \text{ if } \left(\left(\mathbf{R}^t\right)^H \mathbf{\Phi}_n + \hat{\mathbf{h}}_n^t\right)^H \left(\left(\mathbf{R}^t\right)^H \mathbf{\Phi}_n + \hat{\mathbf{h}}_n^t\right) > \theta_{t,n}, \\ 0, \text{ if } \left(\left(\mathbf{R}^t\right)^H \mathbf{\Phi}_n + \hat{\mathbf{h}}_n^t\right)^H \left(\left(\mathbf{R}^t\right)^H \mathbf{\Phi}_n + \hat{\mathbf{h}}_n^t\right) < \theta_{t,n}, \end{cases} \quad (78)$$

with the threshold as $\theta_{t,n} = M\log(1 + \frac{\beta_n}{\varkappa_t^2})/(\frac{1}{\varkappa_t^2} - \frac{1}{\varkappa_t^2+\beta_n})$ where $\varkappa_t$ is iteratively obtained as in (54) and (55). Thus, if the device is considered active then the estimated channel is given by the corresponding $\hat{\mathbf{h}}$.

MP-BSBL

Based on Bayesian learning, the work in [51] uses a block sparse approach with belief propagation (BP) and mean field

TABLE 4. Concise description of simulated activity detection and channel estimation techniques.

Type	Category	Algorithm	Main ideas
Activity detection and channel estimation	Message passing	AMP with MMSE denoiser [43]	• Robust technique where a minimum mean-squared error (MMSE) denoiser in the vector AMP algorithm is designed for user activity detection and channel estimation based on statistical channel information.
		MP-BSBL [51]	• Based on Bayesian learning, MP-BSBL uses a block sparse approach with belief propagation (BP) and mean field (MF), to achieve low complexity. • Considering that the metadata of each device is composed by Zadoff-Chu sequences, MP-BSBL is build on the idea of separating the problem in blocks as the well-known BOMP. • A two-layer hierarchical structure, factorizes the joint a posteriori pdf of the block sparse channel vector.
		Iterative EP [52]	• This iterative algorithm uses expectation propagation and the method of moments to compute the mean and variance of the distribution of the channel. The activity detection is realized based on the channel estimations.
	Machine learning	BRNN [55]	• Consists in a feedforward neural network with interleaved fully connected layers and non-linear transformation layers. • A batch normalization is added for initialization and residual connection is used to avoid vanishing/exploding gradients. The main goal of this approach is to detect the active devices and estimate the channel using an MMSE estimator.
		DNN-MP-BSBL [58]	• Motivated by the convergence speed of MP-BSBL, DNN-MP-BSBL impose weights on the Gaussian messages represented on the factor graph of the scenario. The idea is to simultaneously use the weights on the MF message update and further train it to improve the activity detection accuracy.

(MF), to achieve low complexity. Considering a vector version of (36) and that the metadata of each device is composed by Zadoff-Chu sequences, in the signal model of [51], both the metadata and channel matrices are sparse. Based on the idea of separating the problem in blocks as the well-known block orthogonal matching pursuit (BOMP) [103], a two-layer hierarchical structure, shown in Fig. 14, factorizes the joint a posteriori pdf of the block sparse channel vector $\overline{\mathbf{h}}$:

$$\begin{aligned} p\left(\overline{\mathbf{h}}, \boldsymbol{\gamma}, \lambda|\mathbf{y}\right) &\propto p\left(\mathbf{y}|\overline{\mathbf{h}}, \boldsymbol{\lambda}\right) p\left(\overline{\mathbf{h}}|\boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) p\left(\lambda\right) \\ &= p\left(\lambda\right) \prod_{m=1}^{\tau_\phi M} p\left(y_m|\overline{\mathbf{h}}, \boldsymbol{\lambda}\right) \\ &\quad \prod_{n=1}^{M} \prod_{i=1}^{(M/N)K} p\left(\overline{h}_{n,i}|\gamma_n\right) p\left(\gamma_n\right), \end{aligned} \tag{79}$$

where $p(\overline{\mathbf{h}}) = \int_\gamma p\left(\overline{\mathbf{h}}|\boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) \mathrm{d}\boldsymbol{\gamma}$ is the prior pdf of $\overline{\mathbf{h}}$ given by the product of a conditional prior pdf $p\left(\overline{\mathbf{h}}|\boldsymbol{\gamma}\right)$ and a hyperprior pdf $p\left(\boldsymbol{\gamma}\right)$. Since the parameter λ here is the noise precision $1/\sigma_w^2$ and the joint a posteriori pdf is composed of complex Gaussian pdfs, a Gamma distribution ($p\left(\gamma_n\right)$), the factor graph in Fig. 14 can be build. The BP rule is used at the function nodes and MF rule is used at other function nodes. After a predetermined number of iterations, the channel is estimated with the mean $m_{\overline{h}_{n,i}}$ computed by the exchange of messages, most of them approximated by complex Gaussian pdfs between the factor and function nodes. The activity detection is performed by the evaluation of the inverse of the estimated hiperprior γ with a predefined threshold.

Iterative EP

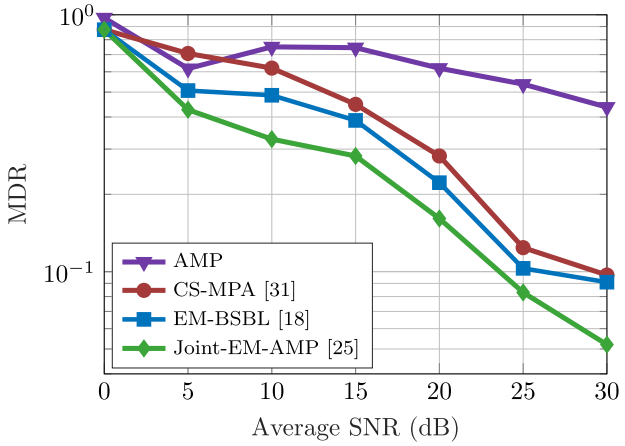
With a different approach, the work in [52] proposes an iterative algorithm that uses expectation propagation (EP) to perform active user detection and channel estimation. Consider the signal model as (36):

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{h} + \mathbf{v}, \tag{80}$$

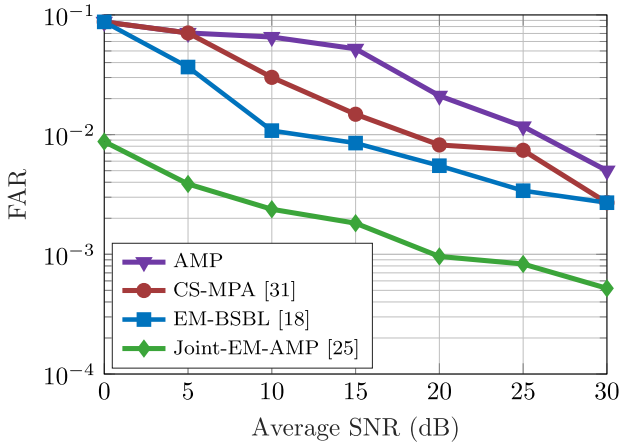
where $\boldsymbol{\Phi}$ is the metadata matrix and the prior distribution of the channel vector $\mathbf{h}$ is given by

$$p(\mathbf{h}) = \prod_{n=1}^{N} \left[(1-\rho_n)\,\delta(h_n) + \rho_n \mathcal{CN}\left(h_n|0, \beta_n\right)\right], \tag{81}$$

where $\delta(\cdot)$ is the dirac function. With this, the idea of the algorithm is to approximate the target distribution $f(\mathbf{h}) = p(\mathbf{y}|\mathbf{h})p(\mathbf{h})$ to the Gaussian distribution $q(\mathbf{h}) = \mathcal{CN}(\mathbf{h}|\tilde{\mathbf{m}}, \tilde{\mathbf{V}})$. After that, the algorithm proceeds in order to match the mean vector true target $\tilde{\mathbf{m}}$ and covariance matrix $\tilde{\mathbf{V}}$ to those of the true target distribution $f(\mathbf{h})$ based on the iterative EP algorithm.



(a) Missed detection rate vs. Average SNR.



(b) False alarms rate vs. Average SNR.

FIGURE 13. Activity error rates for comparison of message-passing algorithms in a coherent scenario. Simulation parameters: $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols;. 10^5 Monte Carlo trials.

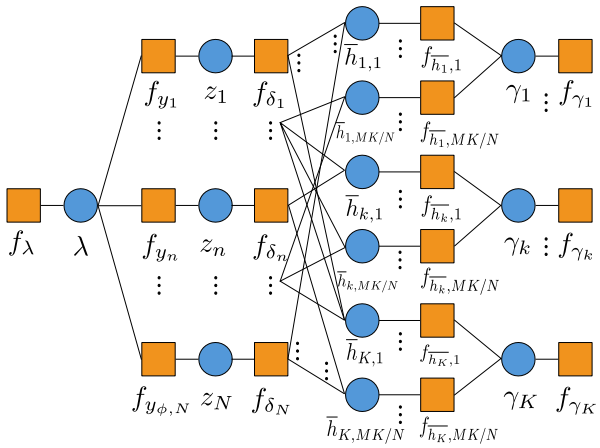


FIGURE 14. Factor graph of MP-BSBL, where the auxiliary variable z_n and extra constrains δ, denoted by f_{δ_n} and f_{y_n} are introduced. This auxiliary variables are function of the channel, metadata vector and λ.

Thus, using the Kullback-Leibler (KL) divergence criterion and approximating the target distribution as

$$f(\mathbf{h}) = f_1(\mathbf{h}) f_2(\mathbf{h}) = p(\mathbf{y}|\mathbf{h}) p(\mathbf{h}),$$

$$f_1(\mathbf{h}) = p(\mathbf{y}|\mathbf{h}) \approx q_1(\mathbf{h}) = \mathcal{CN}(\mathbf{h}|\tilde{\mathbf{m}}_1, \tilde{\mathbf{V}}_1), \quad (82)$$

$$f_2(\mathbf{h}) = p(\mathbf{h}) \approx q_1(\mathbf{h}) = \mathcal{CN}(\mathbf{h}|\tilde{\mathbf{m}}_1, \tilde{\mathbf{V}}_1), \quad (83)$$

where, after reconstructing the unnormalized global Gaussian approximation as $f(\mathbf{h}) \approx q(\mathbf{h}) = q1(\mathbf{h})q2(\mathbf{h}) = \mathcal{CN}(\mathbf{h}|\tilde{\mathbf{m}}, \tilde{\mathbf{V}})$, the mean vector $\tilde{\mathbf{m}}$ and the covariance matrix $\tilde{\mathbf{V}}$ are given by

$$\tilde{\mathbf{V}} = \left(\sigma_w^{-2} \mathbf{\Phi}^H \mathbf{\Phi} + \tilde{\mathbf{V}}_2^{-1}\right)^{-1} \quad (84)$$

$$\tilde{\mathbf{m}} = \tilde{\mathbf{V}} \left(\sigma_w^{-2} \mathbf{\Phi}^H \mathbf{y} + \tilde{\mathbf{V}}_2^{-1} \tilde{\mathbf{m}}_2\right), \quad (85)$$

where $\tilde{\mathbf{m}}_1$ and $\tilde{\mathbf{V}}_1$ where approximated. Then, the task of the iterative EP is to compute the $\tilde{\mathbf{m}}_2$ and $\tilde{\mathbf{V}}_2$ parameters. The update equations are given by

$$\tilde{v}_{2,n}^{t+1} = \left[V_q^t[h_n]^{-1} - (\tilde{v}_{\setminus 2,n}^t)^{-1}\right]^{-1}, \quad (86)$$

$$\tilde{m}_{2,n}^{t+1} = \left[V_q^t[h_n]^{-1} E_q^t[h_n] - (\tilde{v}_{\setminus 2,n}^t)^{-1} \tilde{m}_{\setminus 2,n}^t\right], \quad (87)$$

where $\tilde{v}_{\setminus 2,n}^t$ corresponds to the contribution of the n-th marginal of $q^t(\mathbf{h})$ and $E_q^t[h_n]$ and $V_q^t[h_n]$ are the mean and variance of the distribution to be match and are computed by method of moments. The algorithm proceeds until a predefined number of iterations is reached and the mean vector $\tilde{\mathbf{m}}$ is the estimated channel $\hat{\mathbf{h}}$. Once $\hat{\mathbf{h}}$ is obtained, by performing the likelihood test on $\hat{\mathbf{h}}$, the active devices are detected. Considering H_1 as the hypothesis for the active device and H_0 as inactive, the log-likelihood ratio test is obtained after the thresholding of each element of $\hat{\mathbf{h}}$ as

$$\begin{cases} H_1 \text{ if } |\hat{h}_n|^2 \geq \theta_n, \\ H_0 \text{ if } |\hat{h}_n|^2 < \theta_n, \end{cases} \quad (88)$$

where $\theta_n = \log(1 + \frac{\beta_n}{\tilde{V}_{nn}})/(\frac{1}{\tilde{V}_{nn}} - \frac{1}{\beta_n + \tilde{V}_{nn}})$ and $\tilde{V}_{nn}$ is the n-th diagonal of $\tilde{\mathbf{V}}$.

Performance evaluation

With the same framework of the previous simulations, the message-passing algorithms are evaluated in the scenario of $N = 128$ MTCDs connected to a single base-station equipped with $M = 64$ antennas. Each frame is composed of 128 metadata symbols in order to estimate the channels, considered as the same of (6). Initially, Fig. 15, shows the Frame Error Rate (SER) performance averaged over 10^5 runs. Considering the average SNR as $10 \log\left(N \sigma_x^2 / \sigma_v^2\right)$, the AMP with MMSE denoiser in [43] outperfoms the traditional AMP [20]. On the other hand, the techniques in [51] and [52] consider the prior distribution of the channel vecto and show better performance, as shown in Figs.16a and 16b referred to as the activity detection performance. Using the method of moments, the Iterative EP [52] has a much better NMSE performance compared to the other schemes, while its computational complexity is quadratic and that of MP-BSBL [51] is linear. The well-known BOMP [103] algorithm with knowledge of the the number of active devices is used as a lower bound for MP-BSBL, as done in the original paper.

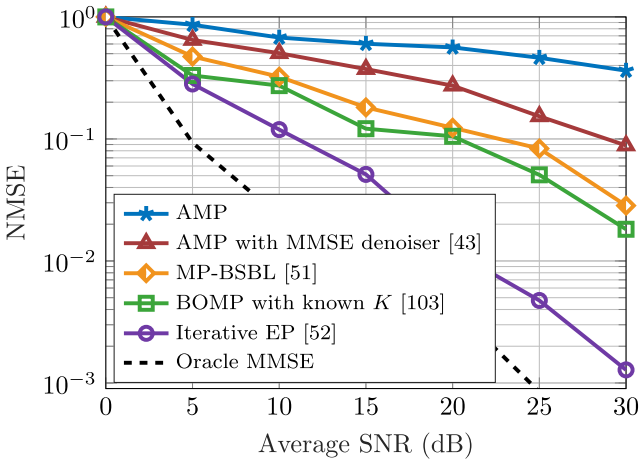
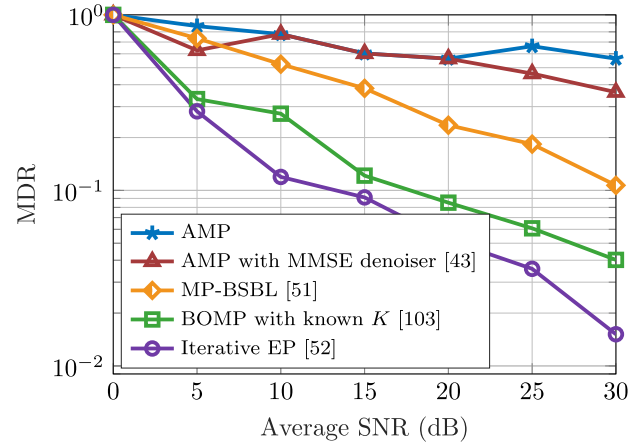
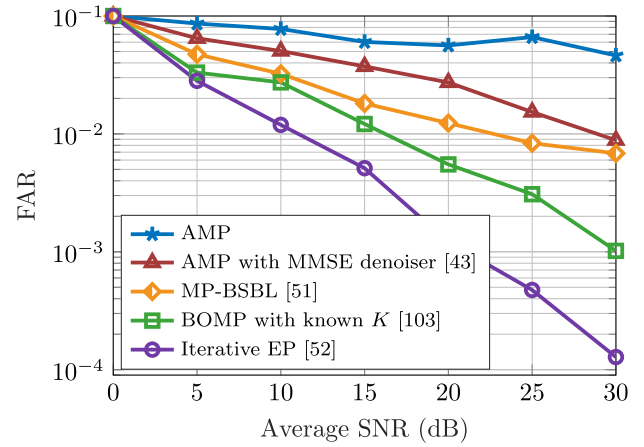


FIGURE 15. Normalized mean squared error vs. Average SNR. Comparison of message-passing algorithms for $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3] and active devices transmitting frames with 128 QPSK symbols. 10^5 Monte Carlo trials.



(a) Missed detection rate vs. Average SNR.



(b) False alarms rate vs. Average SNR.

FIGURE 16. Activity error rates for comparison of message-passing algorithms for channel estimation. Simulation parameters: $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols;. 10^5 Monte Carlo trials.

2) MACHINE LEARNING SOLUTIONS

Recently, some works investigated the use of machine-learning approaches to channel estimation. Deep Learning (DL) [104] is a branch of machine-learning which is also referred to as deep neural networks (DNNs), uses a large amount of training data to learn parameters in a neural network. In order to analyze the complex channel characteristics, the work in [59] uses a typical branch of DNN, the long short-term memory (LSTM). By implicitly reproducing the behavior of the channel with the LSTM algorithm, this approach considers the greedy SISD algorithm [40] to perform activity and data detection. On the other hand, the works in [55] and [58] uses the DL approach to explicitly estimate the channels. Using the metadata vectors as the training set, both works map each received vector $\mathbf{y}$ as an input and considers as the loss function the mean square error (MSE) $\|\hat{\mathbf{h}}_{\text{DNN}} - \mathbf{h}\|_2^2$, where $\hat{\mathbf{h}}_{\text{DNN}}$ is the estimated channel gain and $\mathbf{h}$ is the known channel gain in the training set.

BRNN

The work in [55] consists of a feedforward neural network with interleaved fully connected layers and non-linear transformation layers. A batch normalization is added for initialization and residual connection is used to avoid vanishing/exploding gradients. The t-th layer of the network can be expressed as

$$\mathbf{h}^{t+1} = f\left(\mathbf{W}^t \mathbf{h}^t + \mathbf{b}^t\right), \tag{89}$$

where the parameters to be learned $\mathbf{W}^t$ and $\mathbf{b}^t$ are the weight matrix and the bias vector, respectively, while $f(\cdot)$ denotes the non-linear operator given by

$$f\left(h_1, \ldots, h_M\right) = \text{sign}\left(0, h_1, \ldots, h_M\right) \cdot \left[h_1, \ldots, h_M\right], \tag{90}$$

where sign$(\cdot)$ denotes the sign function. In the back propagation, the gradient of (90) is 1 if $f(h) = h$ and 0 if $f(h) = 0$. The main goal of this approach is to detect the active devices and estimate the channel using an MMSE estimator. Thus, the loss function of the designed network is the cross entropy loss function given by

$$\mathcal{L}(\hat{\mathbf{z}}) = -\sum_{n=1}^{N} z_n \log(\hat{z}_n), \tag{91}$$

where $\hat{z}_n$ is the n-th element of the activation vector $\mathbf{z}$ with $z_n = 1$ if $\|x_n\|2 > 0$ and $z_n = 0$ otherwise. This network uses 6 nodes for training and test phases, while the optimizer adopted for training neural networks is a stochastic gradient algorithm with a momentum 0.9 and a learning rate 0.01.

DNN-MP-BSBL

The work in [58] transfers the iterative message-passing process of MP-BSBL [51], depicted in the last section, to a neural network. Motivated by the convergence speed of MP-BSBL, the authors of [58] impose weights on the Gaussian messages represented on the factor graph depicted in Fig. 14. The idea is to simultaneously use the weights on the MF message update and further train it to improve the activity detection accuracy. Under the argument that the training of the network is conducted offline, just a small computational complexity

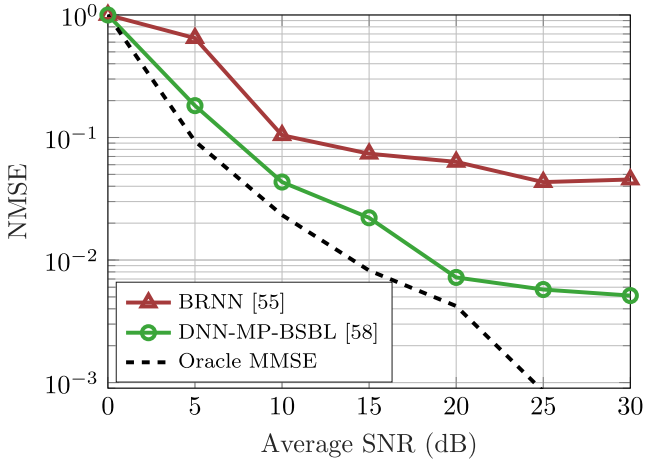


FIGURE 17. Normalized mean squared error vs Average SNR, for comparison of machine-learning algorithms for channel estimation.

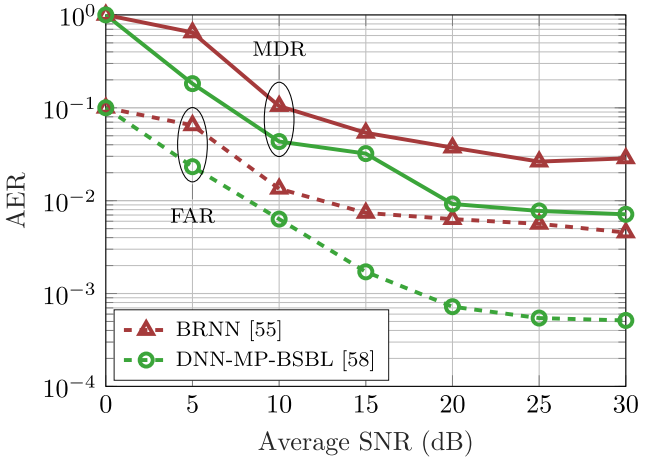


FIGURE 18. Activity error rate vs. Average SNR. Comparison of machine-learning algorithms for channel estimation. Simulation parameters: $N = 128$, $M = 64$ with ρ_n drawn uniformly at random in [0.1, 0.3]. Each frame is composed by 128 QPSK symbols. 10^5 Monte Carlo trials.

is added to the online process, the authors argue that it is possible to use the network even though it is necessary to use 10^5 training sequences. With 9 layers within each iteration block, the received vector is primarily used as the input and then, at each layer, the quantities present in the joint a posteriori pdf in (79) are updated, as its weights. After a predefined number of iterations, the activity detection decision is made by comparing the variable γ with a threshold. If the device is detected as active, the estimated channel gain is attributed to the device.

Performance evaluation

Differently from the other simulation scenarios, the networks need an offline training, before the real transmissions. Although the scenario matches with the previous ones, with $N = 128$, $M = 64$, MTCDs sporadically active with an activity probability drawn uniformly at random in [0.1,0.3] and the channel modelled as in (6), for both simulations, was necessary a training set with the size of 10^5. The rest of the parameters considered specifically for each algorithm, are consistent with those cited in the original works, as for DNN-MP-BSBL [58], the threshold to decide the activity of the device is 0.1, the epoch number is 20, the learning rate is 10^{-3} and 20 iterations. For BRNN, we used 10^6 samples for training, and the rest of the parameters followed the description in the section and the original paper. To evaluate the channel estimates, we used the normalized mean square error (NMSE),

$$\text{NMSE} = \frac{\|\hat{\mathbf{h}} - \mathbf{h}\|_2^2}{\|\mathbf{h}\|_2^2}. \tag{92}$$

We consider only the channels associated with active devices. Fig. 17 shows the simulation results of the NMSE performance of BRNN and DNN-MP-BSBL under different SNR scenarios. With the results averaged under 10^5 runs, it is possible to realize that DNN-MP-BSBL achieves an efficiency closer to the lower bound, the oracle linear MMSE.

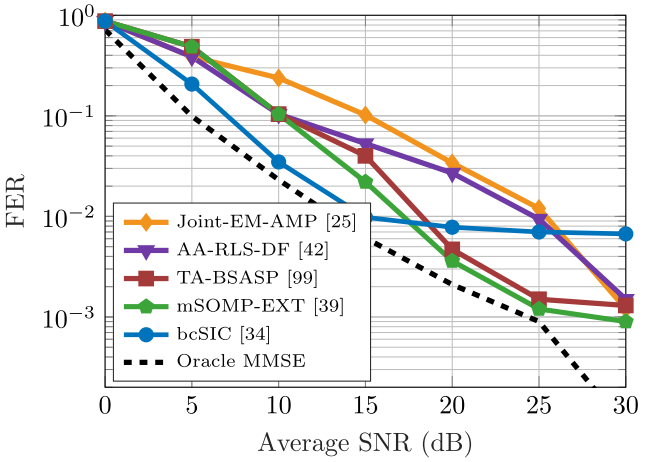


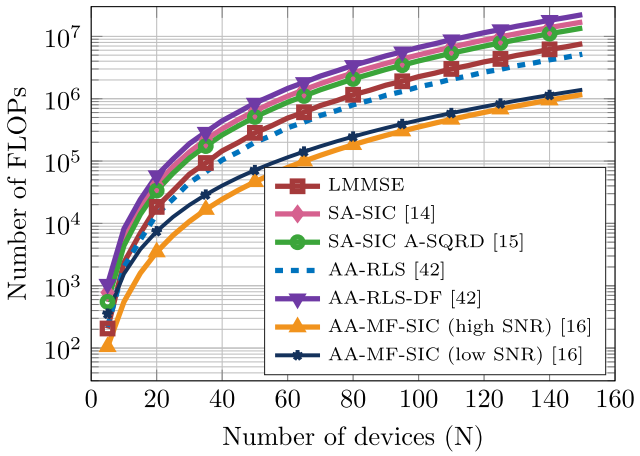
FIGURE 19. Frame error rate vs Average SNR. Comparison between better performance algorithms.

The weight update scheme and the joint a posteriori pdf of the model, considerably outperforms BRNN, since that network only estimates the activity of devices. The superiority of the DNN-MP-BSBL scheme is even more evident in Fig.18, where the activity detection is shown.

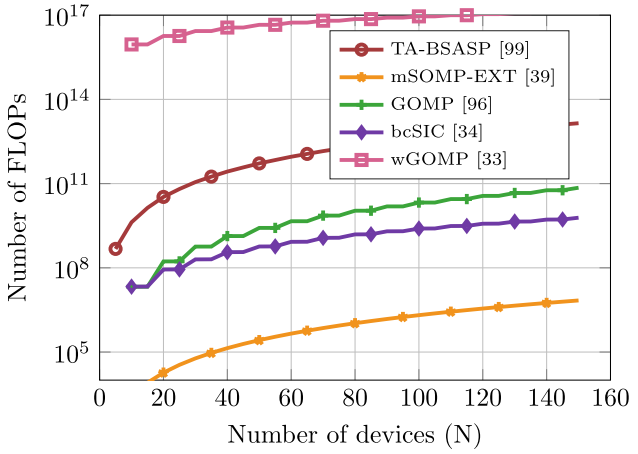
V. DISCUSSIONS

The results demonstrate that each solution provides a direct relation between activity detection rates and data detection. In order to verify the best approach, Fig. 19 compares the schemes with the best performance in each family of algorithms. Considering the unified evaluation framework and perfect channel estimation, we notice that the regularized and greedy detectors show better performance with a few techniques being as efficient as the lower bound.

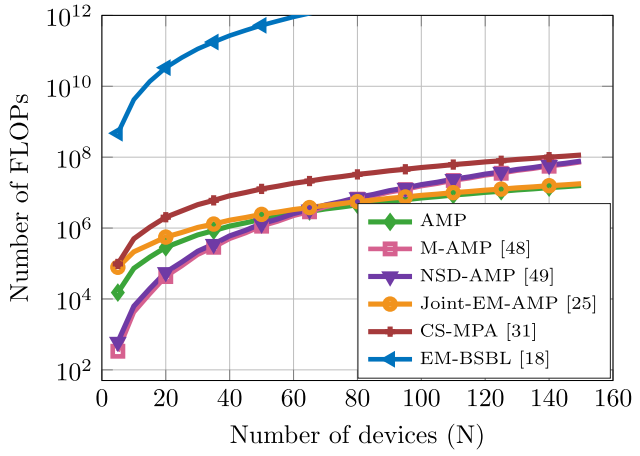
As part of the performance analysis, each technique should have their computational cost evaluated.



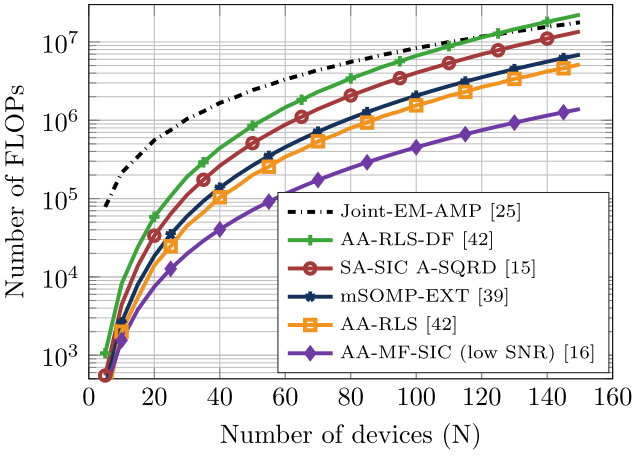
(a) Regularized algorithms for activity and data detection.



(b) Greedy algorithms for activity and data detection.



(c) Message-passing algorithms for activity and data detection.



(d) Comparison of algorithms with lowest FLOPs for activity and data detection.

FIGURE 20. **Floating-point operations vs. Number of devices. Simulation parameters: Number of receiver antennas M is $N/2$ and the number of active devices K is 10% of N. The number of symbols in each frame $\tau = 128$, $\tau_\phi = 64$, constants $c = 5$ and $c' = 10$ of GOMP, bcSIC and wGOMP, the latter, still with $N_{\text{subp}} = 8$, $\tau_{\text{seg}} = 128$ and $N_{\text{fb}} = \tau_{\text{seg}}/N_{\text{subp}}$. Regarding the non-coherent approaches, the number of bits J in the NSD-AMP is 2.**

Fig. 20 depicts the detailed complexity analysis of simulated approaches based on required floating-point operations (FLOPs). Fig. 20a shows that the regularization approaches exhibits a competitive performance as compared to the greedy techniques (in Fig. 20b) exhibit a high computational complexity as those schemes require matrix inversions. Since the expected number of devices is huge, a cubic computational complexity order would be an issue to be dealt with at the BS. As for the message-passing techniques, Fig 20c shows a competitive performance as the number of iterations have greater influence on the computational cost. In particular, the algorithms that exhibit lower computational cost are compared in Fig. 20d. One can see that there is a trade-off between computational cost and data detection performance as the algorithms with lowest computational cost are not those with lowest frame error rates. Table 5 details the FLOPs counting of the analyzed techniques, for each detected vector. Parameters that have not been previously presented, such as T_1, T_2, c, and c', are constants determined in the original papers.

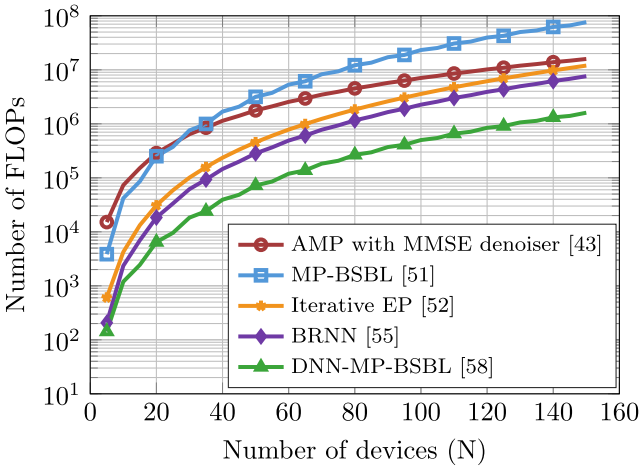


FIGURE 21. **Floating-point operations vs. Number of devices of activity detection and channel estimation algorithms.**

Concerning the channel estimation algorithms, as shown in Fig. 21, Iterative EP shows the best performance, but requires a cubic complexity order. As for machine learning

TABLE 5. FLOPs counting of considered techniques in detail.

Type	Category	Algorithm	Complexity
Activity and data detection	Linear	MMSE	$2M^3 + 4\left(N+1\right)M^2 + 2\left(N^2+N+1\right)M - \left(N^2+N\right)$
	Regularized	SA-SIC [13]	$\left\|\mathcal{A}_0\right\|\left(N^3+N^2+6\right)$
		SA-SIC with A-SQRD [15]	$2N^3 + 4\left(M+1\right)N^2 + \left(M-1\right)N$
		AA-MF-SIC (high SNR) [16]	$\leq (1/6)\left(2N^3+11N^2+21N-2\right)$
		AA-MF-SIC (low SNR) [16]	$\geq (1/6)\left(2N^3+11N^2+21N-2\right)+10N^2$
		AA-RLS [42]	$\left(6M^2+10M\right)N$
		AA-RLS-DF [42]	$\sum_{i=1}^{N}\left[6\left(M+i\right)^2+10\left(M+i\right)\right]$
	Greedy	TA-BSASP [99]	$5N + 3N\tau + 3NM\tau^2 + 2M\tau\left(\tau K\right)^2 + \left(\tau K\right)^3$
		mSOMP-EXT [39]	$2N^3 + 2N^2 + \left(1+2M\right)4N - 2M - 1$
		OMP	$\sum_{i=1}^{K} cMi^2 + c'i^3$
		GOMP [96]	$\sum_{i=1}^{K} 2M\left(i\tau\right)^2 + c'\left(i\tau\right)^3$
		bcSIC [34]	$\sum_{i=1}^{K} K\left(cM\tau^2 + c'\tau^3\right)$
		wGOMP [33]	$\sum_{u=1}^{N_{\text{fb}}K} N_{\text{fb}}N_{\text{subp}}\sum_{i=1}^{K} c\lfloor M/N_{\text{subp}}\rfloor\left(i\tau_{\text{seg}}\right)^2 + c'\left(i\tau_{\text{seg}}\right)^3$
	Message passing	AMP	$\tau M\left(11N+4\right) + 4\left(\left\|\mathcal{A}\right\|+1\right)$
		M-AMP [48]	$N\left(M^3+7M^2+2N+2MN+1\right)$
		NSD-AMP [49]	$N\left(M^3+9M^2+12M+4MN+5\right)+2^{J+1}\left(N+1\right)$
		Joint-EM-AMP [25]	$M\tau\left(4\left\|\mathcal{A}\right\|^2+17\left\|\mathcal{A}\right\|+11N+21\right)+3\tau\left\|\mathcal{A}\right\|\left(\left\|\mathcal{A}\right\|+1\right)$
		CS-MPA [31]	$\left[T_1\left(3N-1\right)2M + T_2\left(MN\right)\right]\tau$
		EM-BSBL [18]	$(N\tau)^3 + 2(M\tau+3/2)(N\tau)^2 + (M\tau+1)N\tau$
Activity detection and channel estimation	Message passing	AMP with MMSE denoiser [43]	$\tau M\left(11N+4\right) + 4\left(\left\|\mathcal{A}\right\|+1\right)$
		MP-BSBL [51]	$\left[\left(10\tau_\phi+10\right)MK + \left(7K+14\right)\tau_\phi M + N + 2\right]$
		Iterative EP [52]	$\left[3N^3+\frac{5}{2}N^2+\left(M+1\right)N^2+\left(7M+\frac{5}{2}\right)N-2M+15\right]$
	Machine learning	BRNN [55]	$2M^3 + 4\left(N+1\right)M^2 + 2\left(N^2+N+1\right)M - \left(N^2+N\right)$
		DNN-MP-BSBL [58]	$\left[M(K(17\tau_\phi+16)+27\tau_\phi)+N+2\right]$

schemes, even though BRNN has high complexity due to the use of the linear MMSE channel estimator, DNN-MP-BSBL has linear complexity at least in the online process. Compared with message-passing schemes, machine learning approaches present better performance but require the offline process, which consumes a large amount of training data to learn parameters in the neural network.

In addition to the techniques analyzed here, it is important to highlight that recent relevant papers were published that jointly perform all the three tasks in the same scheme, channel estimation and activity and data detection. Given the performance and computational cost shown before, it is no surprise that those works are message-passing schemes that employ a factor graph representation for the problem. The work in [105] considers the uplink SCMA scenario and uses the expectation propagation to project the intractable distributions into Gaussian families in order to obtain a linear complexity decoder. With the aim of investigating the time-slotted and non-time-slotted grant-free NOMA, the work in [106] applies the bilinear generalized approximate message passing (BiG-AMP) [107] algorithm to mMTC. In order to address the overhead problem, the work in [108] proposes a

Bayesian receiver design for grant-free low density signature orthogonal frequency division multiplexing (LDS-OFDM). This approach is composed by the belief propagation (BP), expectation propagation (EP) and mean field (MF) techniques, so that the scheme jointly estimates the channels and performs activity and data detection, avoiding the use of metadata signals.

VI. CONCLUDING REMARKS

This paper investigated detection techniques for mMTC. We have provided a brief scenario introduction, where applications, traffic features and challenges are discussed. We described the signal model used in the unified evaluation framework adopted, highlighting the grant-free random access model and the key performance indicators used to evaluate the efficiency of the techniques.

Subsequently, detection techniques have been presented, where relevant works were categorized as regularized, greedy and message-passing detectors, which have the objective of performing activity and data detection, were explained and discussed along with their simulation results. Moreover, activity detection and channel estimation schemes classified as message-passing and machine-learning techniques were presented and had its simulation results compared. In the discussions section, the simulation results were evaluated along with a complexity analysis of each simulated approach.

As for future research, given the massive access request expected for the next generation of wireless systems, we suggest the investigation of asynchronous grant-free random access systems. The assumption that active devices are synchronized at the frame level introduces additional overhead. Under this scenario and as just a few relevant schemes that jointly perform channel estimation, activity and detection were reported, improvements and new approaches with message-passing algorithms are welcome since their trade-off between performance and computational cost is attractive.

REFERENCES

[1] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[2] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 12–18, Jun. 2014.

[3] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 184–192, Jul. 2012.

[4] T. Salam, W. U. Rehman, and X. Tao, "Data aggregation in massive machine type communication: Challenges and solutions," *IEEE Access*, vol. 7, pp. 41921–41946, 2019.

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[6] Ericsson, "Cellular networks for massive IoT," Ericsson, Stockholm, Sweden, Tech. Rep. Uen 284 23-3278, Jan. 2016. [Online]. Available: https://www.ericsson.com/48ff1f/assets/local/reports-papers/white-papers/massive_iot_whitepaper.pdf

[7] *Study on Communication Services for Critical Medical Applications (Release 17)*, document TR 22.826 V1.0.0, 3GPP, May 2019.

[8] Y. Han and J. Lee, "Uplink pilot design for multi-cell massive MIMO networks," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1619–1622, Aug. 2016.

[9] 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, Sep. 2018, doi: 10.1109/MSP.2018.2844952.

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

[11] S. Boyd, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[13] B. Knoop, F. Monsees, C. Bockelmann, D. Wuebben, S. Paul, and A. Dekorsy, "Sparsity-aware successive interference cancellation with practical constraints," in *Proc. 17th Int. ITG Workshop Smart Antennas (WSA)*, Stuttgart, Germany, 2013, pp. 1–8.

[14] B. Knoop, F. Monsees, C. Bockelmann, D. Peters-Drolshagen, S. Paul, and A. Dekorsy, "Compressed sensing K-best detection for sparse multi-user communications," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 1726–1730.

[15] J. Ahn, B. Shim, and K. B. Lee, "Sparsity-aware ordered successive interference cancellation for massive machine-type communications," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 134–137, Feb. 2018.

[16] R. B. Di Renna and R. C. de Lamare, "Activity-aware multiple feedback SIC for massive machine-type communications," in *Proc. 12th Int. ITG Conf. Sys. Commu. Coding (SCC)*, Rostock, Germany, Feb. 2019, pp. 1–6.

[17] A. C. Cirik, N. M. Balasubramanya, and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46–49, Feb. 2018.

[18] X. Zhang, Y.-C. Liang, and J. Fang, "Novel Bayesian inference algorithms for multiuser detection in M2M communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7833–7848, Sep. 2017.

[19] X. Zhang, F. Labeau, Y.-C. Liang, and J. Fang, "Compressive sensing-based multiuser detection via iterative reweighed approach in M2M communications," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 764–767, Oct. 2018.

[20] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[21] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Cairo, Egypt, Jan. 2010, pp. 1–5, doi: 10.1109/ITWKSPS.2010.5503193.

[22] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: II. Analysis and validation," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Cairo, Egypt, Jan. 2010, pp. 1–5, doi: 10.1109/ITWKSPS.2010.5503228.

[23] Z. Chen, F. Sohrabi, and W. Yu, "Multi-cell sparse activity detection for massive random access: Massive MIMO versus cooperative MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4060–4074, Aug. 2019.

[24] K. Senel and E. G. Larsson, "Device activity and embedded information bit detection using AMP in massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Singapore, Dec. 2017, pp. 1–6.

[25] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, Mar. 2017.

[26] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2019, doi: 10.1109/TWC.2018.2878720.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[28] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[29] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.

[30] H. F. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," in *Proc. 8th Int. Symp. Wireless Commun. Syst.*, Nov. 2011, pp. 291–295.

[31] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Boston, MA, USA, Sep. 2015, pp. 1–5.
[32] W. Xiong, J. Cao, and S. Li, "Sparse signal recovery with unknown signal sparsity," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–8, Dec. 2014.
[33] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Improving greedy compressive sensing based multi-user detection with iterative feedback," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Las Vegas, NV, USA, Sep. 2013, pp. 1–5.
[34] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Efficient detectors for joint compressed sensing detection and channel decoding," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2249–2260, Jun. 2015, doi: 10.1109/TCOMM.2015.2424414.
[35] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "Blind detection of uplink grant-free SCMA with unknown user sparsity," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.
[36] H. F. Schepker and A. Dekorsy, "Compressive sensing multi-user detection with block-wise orthogonal least squares," in *Proc. IEEE 75th Veh. Technol. Conf. (VTC Spring)*, Yokohama, Japan, May 2012, pp. 1–5.
[37] B. K. Jeong, B. Shim, and K. B. Lee, "A compressive sensing-based active user and symbol detection technique for massive machine-type communications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 6623–6627.
[38] C. Bockelmann, "Iterative soft interference cancellation for sparse BPSK signals," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 855–858, May 2015.
[39] N. Y. Yu, "Multiuser activity and data detection via sparsity-blind greedy recovery for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 2082–2085, Nov. 2019.
[40] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, Jul. 2016.
[41] R. B. Di Renna and R. C. de Lamare, "Adaptive activity-aware constellation list-based decision feedback detection for massive machine-type communications," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 2111–2115, doi: 10.1109/IEEECONF44664.2019.9049081.
[42] R. B. Di Renna and R. C. de Lamare, "Adaptive activity-aware iterative detection for massive machine-type communications," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1631–1634, Dec. 2019, doi: 10.1109/LWC.2019.2932674.
[43] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.
[44] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018, doi: 10.1109/TSP.2018.2795540.
[45] Z. Sun, Z. Wei, L. Yang, J. Yuan, X. Cheng, and L. Wan, "Exploiting transmission control for joint user identification and channel estimation in massive connectivity," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6311–6326, Sep. 2019.
[46] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Near-optimum sparse channel estimation based on least squares and approximate message passing," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 754–757, Dec. 2017, doi: 10.1109/LWC.2017.2739154.
[47] 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, doi: 10.1109/TSP.2020.2967175.
[48] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.
[49] Z. Tang, J. Wang, J. Wang, and J. Song, "Device activity detection and non-coherent information transmission for massive machine-type communications," *IEEE Access*, vol. 8, pp. 41452–41465, 2020, doi: 10.1109/ACCESS.2020.2976824.
[50] W. Dai, H. Wei, J. Zhou, and W. Zhou, "An efficient message passing algorithm for active user detection and channel estimation in NOMA," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, Sep. 2019, pp. 1–6.
[51] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, Oct. 2018.
[52] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178–5189, Jul. 2019.
[53] F. Lehmann, "Joint user activity detection, channel estimation, and decoding for multiuser/multiantenna OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8263–8275, Sep. 2018.
[54] G. Wunder, I. Roth, R. Fritschek, and J. Eisert, "Performance of hierarchical sparse detectors for massive MTC," 2018, *arXiv:1806.02754*. [Online]. Available: http://arxiv.org/abs/1806.02754
[55] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, Sep. 2019, pp. 1–5.
[56] D. A. Awan, R. L. G. Cavalcante, M. Yukawa, and S. Stanczak, "Detection for 5G-NOMA: An online adaptive machine learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6, doi: 10.1109/ICC.2018.8422449.
[57] K. Senel and E. G. Larsson, "Joint user activity and non-coherent data detection in mMTC-enabled massive MIMO using machine learning algorithms," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, Bochum, Germany, 2018, pp. 1–6.
[58] Z. Zhang, Y. Li, C. Huang, Q. Guo, C. Yuen, and Y. L. Guan, "DNN-aided block sparse Bayesian learning for user activity detection and channel estimation in grant-free non-orthogonal random access," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12000–12012, Dec. 2019.
[59] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.
[60] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.
[61] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.
[62] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.
[63] A. Ali, G. A. Shah, M. O. Farooq, and U. Ghani, "Technologies and challenges in developing Machine-to-Machine applications: A survey," *J. Netw. Comput. Appl.*, vol. 83, pp. 124–139, Apr. 2017.
[64] Y. Cao, T. Jiang, and Z. Han, "A survey of emerging M2M systems: Context, task, and objective," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1246–1258, Dec. 2016.
[65] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.
[66] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE–A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2nd Quart., 2015.
[67] J. Kim, J. Lee, J. Kim, and J. Yun, "M2M service platforms: Survey, issues, and enabling technologies," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 61–76, 1st Quart., 2014.
[68] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[69] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, C. Stefanovic, P. Popovski, Q. Wang, M. Schellmann, E. Kosmatos, P. Demestichas, M. Raceala-Motoc, P. Jung, S. Stanczak, and A. Dekorsy, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28969–28992, 2018.
[70] A. Rajandekar and B. Sikdar, "A survey of MAC layer issues and protocols for Machine-to-Machine communications," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 175–186, Apr. 2015.

[71] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE–A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.
[72] N. Xia, H.-H. Chen, and C.-S. Yang, "Radio resource management in machine-to-machine communications—A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 791–828, 1st Quart., 2018.
[73] N. A. Mohammed, A. M. Mansoor, and R. B. Ahmad, "Mission-critical machine-type communication: An overview and perspectives towards 5G," *IEEE Access*, vol. 7, pp. 127198–127216, 2019.
[74] J. Cao, M. Ma, H. Li, Y. Zhang, and Z. Luo, "A survey on security aspects for LTE and LTE–A networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 283–302, 1st Quart., 2014.
[75] A. Barki, A. Bouabdallah, S. Gharout, and J. Traoré, "M2M security: Challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1241–1254, 2nd Quart., 2016.
[76] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in Machine-to-Machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, Dec. 2016.
[77] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.
[78] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2nd Quart., 2020.
[79] Y. Saleem, N. Crespi, M. H. Rehmani, and R. Copeland, "Internet of Things-aided smart grid: Technologies, architectures, applications, prototypes, and future research directions," *IEEE Access*, vol. 7, pp. 62962–63003, 2019.
[80] O. Shigeru, "M2M and big data to realize the smart city," *NEC Tech. J.*, vol. 7, no. 2, pp. 67–71, Sep. 2012.
[81] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1960–1973, Dec. 2013.
[82] R. B. D. Renna and R. C. D. Lamare, "Iterative list detection and decoding for massive machine-type communications," *IEEE Trans. Commun.*, early access, Jul. 6, 2020, doi: 10.1109/TCOMM.2020.3007525.
[83] C. Schlegel, R. Kempter, and P. Kota, "A novel random wireless packet multiple access method using CDMA," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1362–1370, Jun. 2006.
[84] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive MIMO," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 506–516, Feb. 2019.
[85] J. Ding, D. Qu, and H. Jiang, "Optimal preamble length for spectral efficiency in grant-free RA with massive MIMO," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2019, pp. 1–5.
[86] J. Choi, "Two-stage multiple access for many devices of unique identifications over frequency-selective fading channels," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 162–171, Feb. 2017.
[87] J. Choi, K. Lee, and N. Y. Yu, "Compressive random access using multiple resource blocks for MTC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, USA, Dec. 2016, pp. 1–5, doi: 10.1109/GLOCOMW.2016.7848865.
[88] J. Choi and N. Y. Yu, "Compressive channel division multiple access for MTC under frequency-selective fading," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2715–2725, Jun. 2017.
[89] G. Wunder, Č. Stefanović, P. Popovski, and L. Thiele, "Compressive coded random access for massive MTC traffic in 5G systems," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2015, pp. 13–17.
[90] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. 10th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Ilmenau, Germany, 2013, pp. 1–5.
[91] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[92] Å. Björck, "Numerics of gram-Schmidt orthogonalization," *Linear Algebra Appl.*, vols. 197–198, pp. 297–316, Jan./Feb. 1994.
[93] R. De Lamare and R. Sampaio-Neto, "Minimum mean-squared error iterative successive parallel arbitrated decision feedback detectors for DS-CDMA systems," *IEEE Trans. Commun.*, vol. 56, no. 5, pp. 778–789, May 2008.
[94] P. Li, R. C. de Lamare, and R. Fa, "Multiple feedback successive interference cancellation detection for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2434–2439, Aug. 2011.
[95] R. C. de Lamare and R. Sampaio-Neto, "Adaptive reduced-rank processing based on joint and iterative interpolation, decimation, and filtering," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2503–2514, Jul. 2009.
[96] A. Majumdar and R. Ward, "Fast group sparse classification," *Can. J. Electr. Comput. Eng.*, vol. 34, no. 4, pp. 136–144, 2009.
[97] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.
[98] N. Y. Yu, "A fast and noise-robust algorithm for joint sparse recovery through information transfer," *IEEE Access*, vol. 7, pp. 37735–37748, Mar. 2019.
[99] 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, Dec. 2018, doi: 10.1109/TWC.2018.2872594.
[100] R. Berthier, A. Montanari, and P. M. Nguyen, "State evolution for approximate message passing with non-separable functions," *Inf. Inference*, vol. 9, no. , pp. 33–79, Jan. 2019, doi: 10.1093/imaiai/iay021.
[101] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 912–926, Sep. 2011.
[102] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.
[103] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.
[104] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[105] F. Wei, W. Chen, Y. Wu, J. Ma, and T. A. Tsiftsis, "Message-passing receiver design for joint channel estimation and data decoding in uplink grant-free SCMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 167–181, Jan. 2019.
[106] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, Jul. 2019, doi: 10.1109/TWC.2019.2915955.
[107] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.
[108] 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, Aug. 2020, doi: 10.1109/TVT.2020.2998507.

ROBERTO B. DI RENNA (Graduate Student Member, IEEE) was born in Niterói, Brazil, in 1991. He received the Diploma and M.Sc. degrees in telecommunications engineering from Fluminense Federal University, Niterói, Rio de Janeiro, Brazil, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in communications systems with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil. He has been a Visiting Researcher with the Department of Communications Engineering, Universität Bremen, Bremen, Germany. His research interests include next generation wireless communication systems, compressive sensing theories, multi-user detection, the Internet of Things (IoT), and machine-type communications (MTC).

CARSTEN BOCKELMANN (Member, IEEE) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Bremen, Bremen, Germany, in 2006 and 2012, respectively. Since 2012, he has been a Senior Research Group Leader with the University of Bremen, coordinating research activities regarding the application of compressive sensing/sampling to communication problems. His research interests include communications in massive machine communication, ultra reliable low latency communications (5G) and industry 4.0, compressive sensing, channel coding, and transceiver design.

RODRIGO C. DE LAMARE (Senior Member, IEEE) was born in Rio de Janeiro, Brazil, in 1975. He received the Diploma degree in electronic engineering from the Federal University of Rio de Janeiro, in 1998, and the M.Sc. and Ph.D. degrees in electrical engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2001 and 2004, respectively. Since January 2006, he has been with the Communications Group, Department of Electronic Engineering, University of York, U.K., where he is currently a Professor. Since April 2013, he has also been a Professor with PUC-RIO. He is an elected member of the IEEE Signal Processing for Communications and Networking Committee. He currently serves as an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. His research interests include communications and signal processing, areas in which he has published over 450 articles in international journals and conferences.

ARMIN DEKORSY (Senior Member, IEEE) received the Dipl.-Ing. degree in communications engineering from Fachochschule Kontanz, Kontanz, Germany, in 1992, the Dipl.-Ing. degree in communications engineering from the University of Paderborn, Paderborn, Germany, in 1996, and the Ph.D. degree in communications engineering from the University of Bremen, Bremen, Germany, in 2000. He is currently the Head of the Department of Communications Engineering, University of Bremen. He is distinguished by his more than ten years of industrial experience in leading research positions with Deutsche Telekom, Alcatel-Lucent (Bell Labs), and Qualcomm successfully conducting (inter)national research projects (18 BMBF/BMWI/EU projects). He authored/coauthored more than 160 journal and conference publications and holds more than 19 patents in the area of wireless communications. He investigates new lines of research in wireless communications and signal processing for transmitter baseband design which can readily be transferred to industry. His research interests include cooperative and distributed communications, compressive sensing, and in-network processing. He is a Vice-Chairman of the VDE/ITG expert committee Information and System Theory and represents the ETSI, University of Bremen, and NetWorld2020 ETP. He was an Editor of the IEEE Communications Letter.