Sparse Signal Processing for Grant-Free Massive Connectivity A future paradigm for random access protocols in the Internet of Things

Liang Liu, Erik G Larsson, Wei Yu, Petar Popovski, Cedomir Stefanovic and Elisabeth de Carvalho

Tweet

LIU LINKÖPING UNIVERSITY

Sparse Signal Processing for Grant-Free Massive IoT Connectivity

Liang Liu, Erik G. Larsson, Wei Yu, Petar Popovski,
Čedomir Stefanović, and Elisabeth De Carvalho

Abstract

The next wave of wireless technologies will proliferate in connecting sensors, machines, and robots for myriad new applications, thereby creating the fabric for the Internet of Things (IoT). A generic scenario for IoT connectivity involves a massive number of machine-type connections. But in a typical application, only a small (unknown) subset of devices are active at any given instant, thus one of the key challenges for providing massive IoT connectivity is to detect the active devices first and then to decode their data with low latency. This article outlines several key signal processing techniques that are applicable to the problem of massive IoT access, focusing primarily on advanced compressed sensing technique and its application for efficient detection of the active devices. We show that massive multiple-input multiple-output (MIMO) is especially well-suited for massive IoT connectivity in the sense that the device detection error can be driven to zero asymptotically in the limit as the number of antennas at the base station goes to infinity by using the multiple-measurement vector (MMV) compressed sensing techniques. The paper also provides a perspective on several related important techniques for massive access, such as embedding of short messages onto the device activity detection process and the coded random access.

I. INTRODUCTION

Last decades witnessed the remarkable achievements of the wireless technologies towards offering connectivity to the people. Recently there has been a growing interest to provide ubiquitous connectivity for machines and objects, many of which do not require interactions with humans [1]. This is being driven by the rapid advance of Internet of Things (IoT) that will significantly benefit the way we conduct business, deliver education, health care and government services, and the way we live everyday lives [2]. Targeting at the growing market of IoT, the 5G cellular technologies roadmap has already identified massive machine-type communications (mMTC) as one of the three main use cases, along with enhanced mobile broadband (eMBB) and ultra-reliable and low latency communications (URLLC).

Liang Liu and Wei Yu are with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada (e-mails:lianguot.liu@utoronto.ca; weiyu@comm.utoronto.ca). Erik G. Larsson is with the Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden (e-mail: erik.g.larsson@liu.se), Petar Popovski, Čedomir Stefanović, and Elisabeth De Carvalho are with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark (e-mails: {petarp,cs,edc}@es.aau.dk)
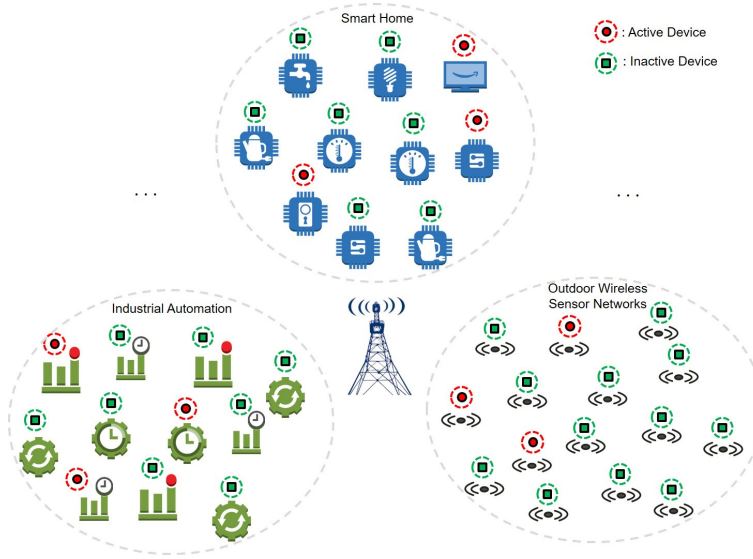
Fig. 1. A typical IoT network with a massive number of devices, e.g., sensors, actuators, etc. Arising from the sporadic IoT data traffic, at each time slot only a subset of devices in the network are active.

The fundamental challenge of mMTC for IoT is to enable data transmission from a massive number of devices in an efficient and timely manner. However, the key characteristic of the IoT traffic is that the device activity patterns are typically *sporadic*, so that at any given time only a small and random fraction of all devices are active, as shown in Fig. 1. The sporadic traffic pattern may be, e.g., due to the fact that often devices are designed to sleep most of the time in order to save energy and are activated only when triggered by external events, as is typically the case in a sensor network. In these scenarios, the active users need to be dynamically identified along with the reception of their data, which is a challenging task.

A. Grant-Based Random Access Scheme

The common user access approach in cellular systems is to have the uncoordinated devices perform random access using the dedicated physical random access channel to contend for the resource blocks for their data transmission [3], as shown in Fig. 2. Specifically, in the first stage each active device picks a random preamble, sometimes referred to as a pilot sequence, from a predefined set of orthogonal preamble sequences, to notify the base station (BS) that the user has become active. In the second stage, the BS sends a response corresponding to each activated preamble as a grant for transmitting in the next step. In the third stage, each device that has received a response to its preamble transmission sends a connection request in order to demand resources for subsequent data transmission. In case a preamble has been selected by a single device, the connection request of the device is granted by the BS, which in turn sends contention resolution message informing device about the resources reserved for the pending
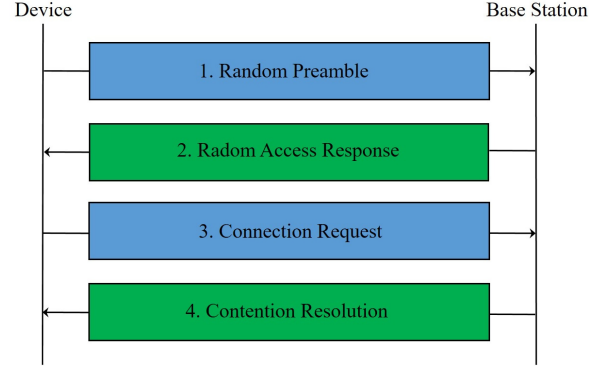
Fig. 2. Grant-based random access procedure: due to the lack of coordination, collisions occur when two or more devices select the same pilot and new access attempts may be needed in this case.

data transmission. However, if two or more devices have selected the same preamble in the first stage, their connection requests collide. When the BS detects a collision, it does not reply with a contention resolution message; the affected devices restart the random access procedure after a timer expires. In the above procedure, the messages sent by the active devices in the first and third phases correspond to metadata, since they belong to control information for establishing the connection without containing any data information.

This access mechanism can be seen as an instance of the classical ALOHA, imposing a limit on the number of active devices that can get the grant to access the network. Recently, extensive efforts have been devoted to different variants of the random access schemes with advanced contention resolution strategies [4], [5]. However, due to the large number of collisions in the massive IoT scenarios, still many users cannot access the network even if some of the colliding connection requests could be resolved, as shown in the following example.

Example 1: Consider a cellular network consisting of one BS and 2000 users. Let L denote the length (and thus the number) of the orthogonal preambles available for the devices to choose. Assume that in each time slot, 100 out of these 2000 devices are active, each picking one of the L orthogonal pilots at random. The coherence bandwidth and the coherence time of the wireless channel are 1MHz and 1ms, respectively, thus in each coherence block 1000 symbols can be transmitted. Moreover, we assume both the scenario in which the contention resolution is not performed and the scenario in which even when there is a collision, the BS can always grant access to one of the colliding devices[1]. Under this setup, the average numbers of devices that are granted the permission to access the network, for both the cases with and without contention resolution, versus different values of L are plotted in Fig. 3. The plot is obtained by Monte Carlo simulations. It is observed that to guarantee 90% success rate, at least $L = 470$ and

---

[1]This could happen, e.g., due to the *capture effect*, in random access networks.
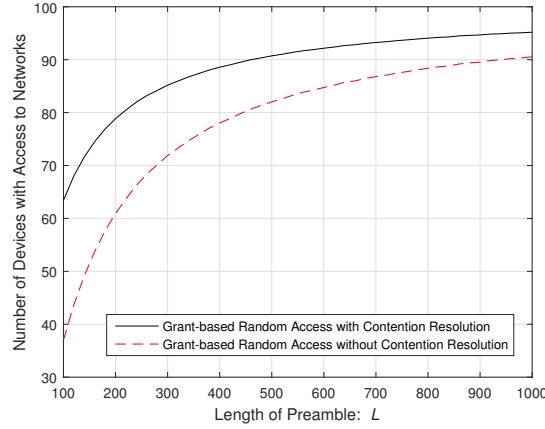
Fig. 3. Grant-based random access with orthogonal pilots and capture: How many users can access the network?
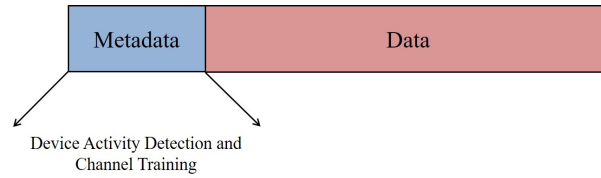

Fig. 4. Grant-free transmission strategy: Metadata contains preamble for device activity detection and channel estimation, and data is directly transmitted after metadata without waiting for the grant from BS.

$L = 930$ out of 1000 symbols are needed, respectively, as pilot for the cases with and without contention resolution.

A question arising from the above example is how can we accommodate more devices with low latency requirement in the future massive IoT connectivity systems. One promising solution is the *grant-free* device access scheme based on the advanced *compressed sensing* techniques.

B. Grant-Free Access Scheme

Under the grant-free access scheme, each active device directly transmits its metadata and data to the BS without waiting for any permission, as shown in Fig. 4. Specifically, in contrast to the random access scheme in which pilot sequences are randomly selected at each time slot, each device under the grant-free access scheme is pre-assigned with a unique pilot sequence used for all the time slots. This pilot sequence thus also serves as the ID for this user and is reminiscent of the role that code-division multiple-access (CDMA) sequence plays in facilitating the extraction of a user data under interference from other users. At each time slot, the BS first detects the active devices by detecting which pilot sequences are used. Next, the BS estimates their channels based on the received metadata, and then decodes the data with the estimated channels [6], [7].

Obviously, the very fact that both metadata and data in the grant-free access are sent in a single step offers the possibility to decrease the access latency compared to the grant-based access. However, device activity detection is now more challenging, because due to the massive number of devices in the network as well as the limited channel coherence time, it is not possible to assign orthogonal pilot sequences to all the devices. The difference with the classical CDMA systems is that the activation dynamics covers a much larger population, placing this problem in the realm of sparse signal processing.

This article aims to pave the way for a theoretical investigation on how the sparse signal processing technologies can enable accurate and efficient active device detection under the grant-free access scheme. We first point out that the device activity detection can be cast into a compressed sensing problem. Next, a random pilot sequence design is introduced, and the use of approximate message passing (AMP) algorithm [8] is proposed for detecting the active devices. Further, we show that massive multiple-input multiple-output (MIMO) [9], [10], which has already exhibited outstanding performance for enhancing the spectrum efficiency in human-type communications, provides an opportunity to leverage the so-called *multiple measurement vector* (MMV) compressed sensing technique [11], [12] to achieve asymptotically perfect device activity detection accuracy in the massive IoT machine-type communications. Another important fact about mMTC is that it dominantly relies on short packet transmissions. We elaborate on a new method to embed a small number of information bits in the short packets that can be decoded in the device activity detection process. This is enabled by letting each active device randomly select one pilot from a predefined set and letting the BS detect which pilot is used by each active device using AMP. Finally, this article discusses the related technique of coded ALOHA [13] for device activity detection.

II. DEVICE ACTIVITY DETECTION AS A COMPRESSED SENSING PROBLEM

As discussed in the above, it is the sporadic IoT traffic and device activity detection that impose the greatest challenge in the design of the grant-free device access protocol. Interestingly, it is also the sporadic IoT traffic itself that provides a promising opportunity for tackling this challenge. As only a small subset of users are active at each time slot, user activity detection amounts to a sparse signal recovery problem.

Suppose that there are N users in the system, which are denoted by the set $\mathcal{N} = \{1, \cdots, N\}$. Further, assume that the BS is equipped with one antenna, and the channel from user n to the BS is denoted by h_n. In each coherent time slot, define the user activity indicator function as

$$\alpha_n = \begin{cases} 1, & \text{if user } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \quad \forall n \in \mathcal{N}. \tag{1}$$

Assume that each device n decides in each coherence block whether or not to access the channel with probability ϵ_n in an independent manner. Then, α_n can be modeled as a Bernoulli random variable so that $\Pr(\alpha_n = 1) = \epsilon_n$, $\Pr(\alpha_n = 0) = 1 - \epsilon_n$, $\forall n$. The sparse activity level ϵ_n depends on the specific

applications. The model is sufficiently general so that it can capture a variety of applications, e.g., a sensor fusion network in which the sampling rates at different sensors may even be different.

Suppose that each device n is assigned with one pilot sequence $\boldsymbol{a}_n \in \mathbb{C}^{L \times 1}$ with $\|\boldsymbol{a}_n\|^2 = 1$, where L denotes the length of device pilot sequence. Furthermore, we assume that the active users are synchronized within the cyclic prefix, and accurately enough in frequency such that the block fading assumption yields a legitimate model for the channel. This is justified by having the BS send a beacon that invites uplink transmissions from the active devices. The received signal at the BS for device activity detection is then

$$\boldsymbol{y} = \sqrt{\xi} \sum_{n \in \mathcal{N}} \alpha_n h_n \boldsymbol{a}_n + \boldsymbol{z} = \sqrt{\xi} \boldsymbol{A} \boldsymbol{x} + \boldsymbol{z}, \tag{2}$$

where $\boldsymbol{y} = [y_1, \cdots, y_L]^T \in \mathbb{C}^{L \times 1}$ is the received signals over L symbols, ξ is the total transmit energy of the pilot for each active device, $\boldsymbol{z} \in \mathbb{C}^{L \times 1} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ is the independent additive white Gaussian noise (AWGN) at the BS, $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N]$ is the collection of pilot sequences of all the devices, and $\boldsymbol{x} = [x_1, \cdots, x_N]^T$ with $x_n = \alpha_n h_n$ denoting the effective channel of device n. The goal for the BS is to detect the active devices and detect their channels by recovering $\boldsymbol{x}$ based on the noisy observation $\boldsymbol{y}$.

Restricted by the limited coherence time in a practical massive IoT connectivity scenario, the length of device pilot sequence is much smaller than the number of devices, i.e., $L \ll N$. Hence, (2) describes an underdetermined linear system with more unknown variables than equations. However, since $\boldsymbol{x}$ is sparse with many zero entries based on (1), such a reconstruction problem is a sparse optimization problem that can be possibly solved via non-linear compressed sensing techniques.

There are two main theoretical questions in compressed sensing. First, how to design the sensing matrix $\boldsymbol{A}$ so as to capture almost all the information about $\boldsymbol{x}$ with a minimal cost L? Second, given a sensing matrix $\boldsymbol{A}$, how to recover $\boldsymbol{x}$ from the noisy observation $\boldsymbol{y}$ even if $L < N$? In fact, these two questions are coupled: a good design of the sensing matrix $\boldsymbol{A}$ leads to an easier algorithm for recovering the sparse signal $\boldsymbol{x}$. For the massive IoT connectivity setting, this indicates that the device pilot sequences should be carefully designed to enable efficient activity detection schemes at the BS side.

Although a number of desirable properties for a good sensing matrix are known, e.g., restricted isometry property (RIP), optimizing the sensing matrix design is quite a challenging problem. This magazine article mainly focuses on simple ways to construct the sensing matrix $\boldsymbol{A}$ that are easy to be implemented for practical pilot design. In Sections III and IV, we consider the case when each entry of $\boldsymbol{A}$ is i.i.d. randomly generated based on Gaussian distribution and review the AMP algorithm [8] to recover $\boldsymbol{x}$ [6], [14]. Later in Section VI, we will briefly review the other choices of the sensing matrix $\boldsymbol{A}$ and the corresponding compressed sensing algorithms, e.g., the sparse-graph based algorithm with a sparse $\boldsymbol{A}$ [15], and their applications in device activity detection, e.g., coded slotted ALOHA [13].

III. AMP-BASED DEVICE ACTIVITY DETECTION

AMP, proposed in the seminal work [8], is an efficient iterative thresholding method designed for large-scale compressed sensing problems, making it appealing in the massive IoT connectivity scenario of interests. An attractive feature of the AMP framework is that it allows an analytic performance characterization via the so-called *state evolution* [16]. In the following, we introduce how the AMP algorithm works for device activity detection in massive IoT connectivity.

A. Device Pilot Sequence Design

We assume in this section that the entries of user pilots are generated from i.i.d. complex Gaussian distribution with zero mean and variance $1/L$, i.e.,

$$a_{n,l} \sim \mathcal{CN}(0, 1/L), \quad \forall n, l. \tag{3}$$

This particular choice of user pilot sequence is convenient for use with the AMP algorithm for two reasons: first, the convergence of the AMP algorithm for device activity detection is guaranteed if $\boldsymbol{A}$ is generated in this way [8]; second, with such a Gaussian sensing matrix, the state evolution of the AMP algorithm is well established [16], based on which detection performance, e.g., *missed detection probability* (probability that an active device is not detected) and *false alarm probability* (probability that an inactive device is declared to be active), can be analytically characterized in the asymptotic limit.

B. Algorithm Design and Performance Analysis

1) General Form of AMP Algorithm: The AMP algorithm aims to provide an estimate $\hat{\boldsymbol{x}}(\boldsymbol{y})$ based on $\boldsymbol{y}$ that minimizes the mean-squared error (MSE)

$$\text{MSE} = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y}} ||\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}||_2^2. \tag{4}$$

Based on an approximation of the message passing algorithm and starting with $\boldsymbol{x}^0 = \boldsymbol{0}$ and $\boldsymbol{r}^0 = \boldsymbol{y}$, the AMP algorithm proceeds at each iteration as [8], [17]:

$$x_n^{t+1} = \eta_{t,n}((\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t), \tag{5}$$

$$\boldsymbol{r}^{t+1} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{t+1} + \frac{N}{L}\boldsymbol{r}^t \sum_{n=1}^{N} \frac{\eta'_{t,n}((\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t)}{N}, \tag{6}$$

where $t = 0, 1, \cdots$ is the index of the iteration, $\boldsymbol{x}^t = [x_1^t, \cdots, x_N^t]^T$ is the estimate of $\boldsymbol{x}$ at iteration t, $\boldsymbol{r}^t = [r_1^t, \cdots, r_L^t]^T \in \mathbb{C}^{L \times 1}$ denotes the corresponding residual, $\eta_{t,n}(\cdot) : \mathbb{C} \to \mathbb{C}$ is the so-called denoiser, and $\eta'_{t,n}(\cdot)$ is the first-order derivative of $\eta_{t,n}(\cdot)$. The basic intuition is that since the solution should minimize $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$, algorithm makes progress in (5) by moving in the direction of the gradient of $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^t\|^2$, i.e., $(\boldsymbol{r}^t)^H \boldsymbol{a}_n$, $n = 1, \cdots, N$, and then promotes sparsity by applying an appropriately designed denoiser $\eta_{t,n}(\cdot)$. The residual is then updated in (6), but corrected with a so-called Onsager term involving $\eta'_{t,n}(\cdot)$.

2) State Evolution: An important analytical result about the AMP algorithm is the so-called state evolution in the asymptotic regime when $L, K, N \to \infty$, while their ratios converge to some fixed positive values $N/L \to \omega$ and $K/N \to \epsilon = \lim_{N \to \infty} \sum_n \epsilon_n / N$ with $\omega, \epsilon \in (0, \infty)$. In systems for massive IoT connectivity, these assumptions indicate that the length of the pilot sequence is in the same order of the number of active users or total users. After the tth iteration of the AMP algorithm, define a set of random variables $\hat{X}_n^t$'s as

$$\hat{X}_n^t = X_n + \tau_t V_n, \quad \forall n, \tag{7}$$

where the random variables X_n's capture the distributions of x_n's, V_n follows the normal distribution, i.e., $V_n \in \mathcal{CN}(0, 1)$, and is independent of X_n as well as $V_j, \forall j \neq n$, and τ_t is the state variable, which changes from iteration to iteration as modeled by a simple scalar iterative function known as the MSE map:

$$\tau_{t+1} = \frac{\sigma^2}{\xi} + \omega \mathbb{E}[|\eta_{t,n}(X_n + \tau_t V_n) - X_n|^2]. \tag{8}$$

Here, the expectation is over the random variables X_n's and V_n's and over all n. Under the aforementioned asymptotic regime, [16] shows that applying the denoiser to $(r^t)^H a_n + x_n^t$ in (5) is statistically equivalent to applying the denoiser to $\hat{X}_n^t$ as shown in (7).

The statistical model of AMP as given in (7) and (8) can be utilized to design the denoiser functions $\eta_{t,n}(\cdot)$'s in (5) and to quantify the performance of the AMP algorithm.

3) Minimax Framework for Denoiser Design: The flexibility in the AMP algorithm design lies in the denoiser $\eta_{t,n}(\cdot)$ in (5). In the AMP literature, the prior distribution of x is in general assumed to be unknown. In this case, the denoiser $\eta_{t,n}(\cdot)$ is designed under the minimax framework to optimize the AMP algorithm performance for the worst-case or least-favorable distribution of x [18]. Such a design leads to a soft thresholding denoiser for promoting sparsity even for x with the worst-case distribution [8]:

$$\eta_{t,n}(\hat{x}_n^t) = \left(\hat{x}_n^t - \frac{\theta_n^t \hat{x}_n^t}{|\hat{x}_n^t|} \right) \mathbb{I}(|\hat{x}_n^t| > \theta_n^t), \tag{9}$$

where the distribution of $\hat{x}_n^t$ is captured by $\hat{X}_n^t$, and $\theta_n^t > 0$ is the threshold for device n for the tth iteration of the AMP algorithm, which can be optimized based on the state evolution (8) to minimize the MSE as given in (4). With this denoiser, after the tth iteration of the AMP algorithm as shown in (5) and (6), device n is declared to be active if $|(r^t)^H a_n + x_n^t| > \theta_n^t$, and declared to be inactive otherwise. Note that AMP with soft thresholding implicitly solves the LASSO problem [19], i.e., the sparse signal recovery problem as an ℓ_1-penalized least squares optimization.

4) Bayesian Framework for Denoiser Design: On the other hand, if the distribution of x is known in (2), we can design the minimum mean-squared error (MMSE) denoiser via the Bayesian approach to
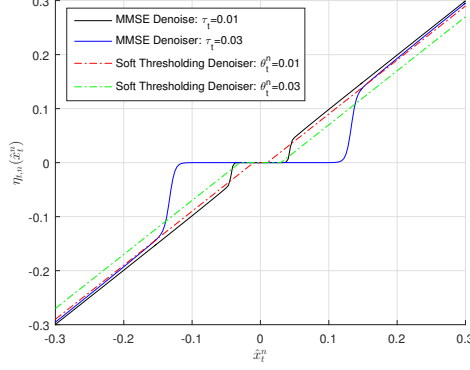
Fig. 5. MMSE denoiser versus soft thresholding denoiser in the AMP algorithm.

minimize the MSE for the estimation of $\boldsymbol{x}$ as given in (4) [18]. Considering the equivalent signal model (7), the MMSE denoiser is given as the following conditional expectation

$$\eta_{t,n}(\hat{x}_n^t) = \mathbb{E}[X_n | \hat{X}_n^t = \hat{x}_n^t], \quad \forall t, n, \tag{10}$$

where the expectation is over X_n and $\hat{X}_n^t$.

For example, if we assume a Rayleigh fading channel such that $h_n \sim \mathcal{CN}(0, \beta_n)$, where β_n denotes the path-loss and shadowing component of user n and is assumed to be known by the BS, then the effective channel $x_n = \alpha_n h_n$ follows a Bernoulli-Gaussian distribution. Under this particular distribution of $\boldsymbol{x}$, an analytical expression of the above MMSE denoiser can be found in [14], which is in general non-linear and has complicated form. Similar to the soft thresholding denoiser case, with the MMSE denoiser (10), we can detect the user activity based on whether the magnitude of $(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t$ is larger than or smaller than a carefully designed threshold θ_n^t.

A comparison between the soft thresholding and MMSE denoisers with a Bernoulli-Gaussian distributed $\boldsymbol{x}$ is given in Fig. 5. It can be observed that the MMSE denoiser is also a thresholding based denoiser, but more "soft" around the regime around the threshold. Moreover, the threshold for the MMSE denoiser is obtained by calculating (10) to minimize the MSE (4), while the design of the threshold for the soft thresholding denoiser follows a minimax framework, which is not optimal given a particular distribution of $\boldsymbol{x}$ in general.

5) Analytical Performance Characterization: The state evolution also allows an analytical performance characterization of the AMP algorithm. For example, with both the soft thresholding and MMSE denoisers, missed detection event happens if one user is active but $|(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t| < \theta_n^t$, while false alarm event happens if one user is inactive but $|(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t| > \theta_n^t$. Since $\hat{x}_n^t$ defined in (7) captures the statistical distribution of $(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t$, the probabilities of missed detection and false alarm for device n after the
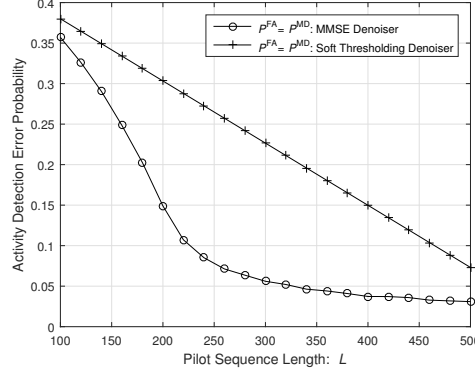
Fig. 6. Probabilities of missed detection and false alarm versus pilot sequence length L.

tth iteration of the AMP algorithm thus can be expressed as

$$P_{t,n}^{\text{MD}} = \text{Pr}(\hat{x}_n^t < \theta_n^t | \alpha_n = 1), \tag{11}$$

$$P_{t,n}^{\text{FA}} = \text{Pr}(\hat{x}_n^t > \theta_n^t | \alpha_n = 0), \tag{12}$$

respectively.

Given the distribution of x and denoiser $\eta_{t,n}(\cdot)$, we can track the values of τ_t's over iterations based on the state evolution (8), then calculate the probabilities of missed detection and false alarm based on (11) and (12).

Example 2: Here we provide a numerical example to show the probabilities of missed detection and false alarm achieved by the AMP algorithm, under the same setup that is used in Example 1. The $N = 2000$ devices are assumed to be randomly located in a cell with a radius 1000 meters, while each device accesses the channel with an identical probability $\epsilon_n = 0.05$, $\forall n$, i.e., $\epsilon = 0.05$ and $K = 100$ of the $N = 2000$ devices are active at any given time. The transmit power of each user for sending its pilot is $\rho^{\text{pilot}} = 23\text{dBm}$. The power spectral density of the AWGN at the BS is assumed to be -169dBm/Hz. Moreover, we define the system-level missed detection and false alarm probabilities as $P^{\text{MD}} = \sum_{n=1}^N P_n^{\text{MD}}$ and $P^{\text{FA}} = \sum_{n=1}^N P_n^{\text{FA}}$, where P_n^{MD} and P_n^{FA} denote the missed detection and false alarm probabilities of device n achieved by AMP after its convergence. Hence, $\epsilon N P^{\text{MD}}$ and $(1-\epsilon)N P^{\text{FA}}$ are the average numbers of missed detection and false alarm events at each time slot in a system with N devices. In addition, under both the soft thresholding and MMSE based AMP algorithms, the thresholds θ_n^t's are carefully selected such that $P^{\text{MD}} = P^{\text{FA}}$.

Fig. 6 shows the device activity detection accuracy achieved by the AMP algorithm with the soft thresholding denoiser and the MMSE denoiser. It is observed that with the MMSE denoiser, 90% active devices can be detected with the AMP algorithm when the length of pilot sequence satisfies $L > 220$. Recall that in Example 1 of the random access scheme, such a performance can be achieved only when
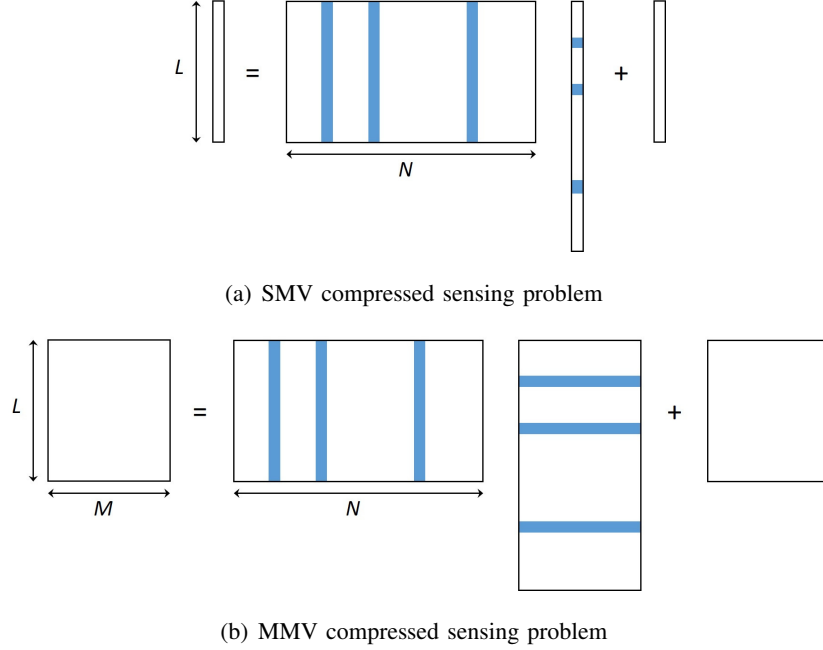
(a) SMV compressed sensing problem

(b) MMV compressed sensing problem

Fig. 7. SMV versus MMV compressed sensing problem: "group sparsity" provides additional information about X since if one entry is zero, the other entries on the same row should be also zero.

the pilot sequence length is longer than 470 even if contention resolution is performed. Moreover, with a careful design of the MMSE denoiser, the MMSE denoiser based AMP algorithm outperforms the soft thresholding denoiser based AMP algorithm in terms of device activity detection.

IV. FROM SMV TO MMV: MASSIVE MIMO FOR MASSIVE IoT CONNECTIVITY

As compared to most other applications of compressed sensing such as imaging, a unique and essential opportunity provided by the wireless massive IoT connectivity system design lies in the potential for utilizing the MMV technique for compressed sensing [11], thanks to the multi-antenna technologies nowadays used ubiquitously in cellular networks. The previous section deals with the application of compressed sensing technique for user activity detection when the BS is equipped with one antenna. In the literature of compressed sensing, the case with one measurement vector is referred to as a single measurement vector (SMV) problem. Recently, massive MIMO has emerged as a revolutionary technology for dealing with the future data deluge for human-type communications. This section shows that massive MIMO is also a natural solution for accommodating a huge number of IoT devices for the future machine-type communications. From the compressed sensing perspective, device activity detection in massive MIMO systems corresponds to the MMV problem, which generalizes the sparse signal recovery problem to the case with a group of measurement vectors for a group of signal vectors that are assumed to be

jointly sparse and share a common support. It is of both theoretical and practical importance to investigate the role of massive MIMO on massive IoT connectivity, which is the aim of this section.

Suppose that the BS is equipped with M antennas. In this case, the channel from user n to the BS is $\boldsymbol{h}_n \in \mathbb{C}^{M \times 1}$. Then, the signal model given in (2) is generalized to

$$\boldsymbol{Y} = \sqrt{\xi}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{Z}, \tag{13}$$

where $\boldsymbol{Y} \in \mathbb{C}^{L \times M}$ is the matrix of received signals across M antennas over L symbols, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^T$ with $\boldsymbol{x}_n = \alpha_n \boldsymbol{h}_n$ denoting the effective channel of user n, and $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_M]$ with $\boldsymbol{z}_m \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\forall m$, is the independent AWGN at the BS.

As compared to the SMV signal model (2), the main difference lies in the fact that $\boldsymbol{X}$ in (13) is a row-sparse matrix, i.e., if one entry of one particular row of $\boldsymbol{A}$ is zero, the other entries of that row must be also zero. This information can be utilized to improve the user detection accuracy. A comparison between the SMV model (2) and MMV model (13) is illustrated in Fig. 7. In the following, we discuss how to generalize the AMP based algorithm in Section III to the massive MIMO scenario and to quantify its significant improvement in device activity detection accuracy over the single-antenna BS case.

A. Algorithm Design

With massive MIMO at the BS, the user pilot sequence assignment still follows (3), which is the same as the case with one antenna at the BS. However, the AMP algorithm is modified as [12]

$$\boldsymbol{x}_n^{t+1} = \eta_{t,n}((\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t), \tag{14}$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}^{t+1} + \frac{N}{L}\boldsymbol{R}^t \sum_{n=1}^{N} \frac{\eta_{t,n}'((\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t)}{N}. \tag{15}$$

As compared to (5) and (6), the dimensions of the signals are now $\boldsymbol{x}_n^t \in \mathbb{C}^{M \times 1}$ and $\boldsymbol{R}^t \in \mathbb{C}^{L \times M}$; moreover, the denoiser is a mapping in higher dimension, i.e., $\eta_{t,n}(\cdot) : \mathbb{C}^{M \times 1} \to \mathbb{C}^{M \times 1}$.

The state evolution of the AMP algorithm still holds for MMV in the asymptotic regime that $L, K, N \to \infty$ with fixed ratios $N/L \to \omega$ and $K/N \to \epsilon$. Specifically, define [12]

$$\hat{\boldsymbol{X}}_n^t = \boldsymbol{X}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{V}_n, \tag{16}$$

where the random vector $\boldsymbol{X}_n$ captures the distribution of $\boldsymbol{x}_n$, $\boldsymbol{V}_n \in \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ is the independent Gaussian noise, and $\boldsymbol{\Sigma}_t$ can be tracked over iterations as follows

$$\boldsymbol{\Sigma}_{t+1} = \frac{\sigma^2}{\xi}\boldsymbol{I} + \omega\mathbb{E}\left[(\eta_{t,n}(\boldsymbol{X}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{V}_n) - \boldsymbol{X}_n)(\eta_{t,n}(\boldsymbol{X}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{V}_n) - \boldsymbol{X}_n)^H\right]. \tag{17}$$

Here, the expectation is over $\boldsymbol{X}_n$'s and $\boldsymbol{V}_n$'s and over all n. Then, in (14), applying denoiser to $(\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t$ is statistically equivalent to applying denoiser to

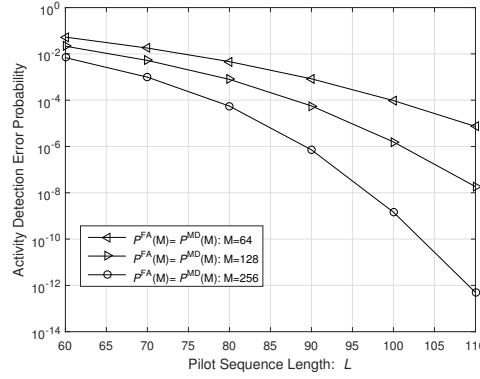$$\hat{\boldsymbol{x}}_n^t = \boldsymbol{x}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{v}_n, \tag{18}$$

Fig. 8. Probabilities of missed detection and false alarm versus pilot sequence length L with massive MIMO.

where the distributions of $\hat{\boldsymbol{x}}_n^t$ and $\boldsymbol{v}_n$ are captured by $\hat{\boldsymbol{X}}_n^t$ and $\boldsymbol{V}_n$.

Based on the above state evolution, denoisers of the MMV-based AMP algorithm can be designed based on different criteria as for the SMV case. For example, the soft thresholding denoiser is

$$\eta_{t,n}(\hat{\boldsymbol{x}}_n^t) = \left(\hat{\boldsymbol{x}}_n^t - \frac{\theta_n^t \hat{\boldsymbol{x}}_n^t}{\|\hat{\boldsymbol{x}}_n^t\|} \right) \mathbb{I}(\|\hat{\boldsymbol{x}}_n^t\|_2 > \theta_n^t), \tag{19}$$

Further, assuming Bernoulli Gaussian distributed $\boldsymbol{x}_n$'s, the MMSE denoiser

$$\eta_{t,n}(\hat{\boldsymbol{x}}_n^t) = \mathbb{E}[\boldsymbol{X}_n | \hat{\boldsymbol{X}}_n^t = \hat{\boldsymbol{x}}_n^t], \tag{20}$$

is characterized in [6]. With both the soft thresholding and MMSE denoisers, after the tth iteration of the AMP algorithm, user n can be declared to be active if $\|(\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t\|_2 > \theta_n^t$, and declared to be inactive otherwise, where θ_n^t is the carefully designed threshold for device detection.

B. Asymptotically Perfect Device Activity Detection

Fix the number of antennas at the BS, M, the missed detection and false alarm probabilities from the MMSE denoiser based AMP algorithm, denoted by $P_{t,n}^{\mathrm{MD}}(M)$ and $P_{t,n}^{\mathrm{FA}}(M)$ (reducing to (11) and (12) when $M = 1$), are characterized in [6]. Interestingly, perfect device activity detection is achieved in the asymptotic regime of $M \to \infty$ if the thresholds for device detection, i.e., θ_n^t's, are properly selected (c.f. [6, Theorem 4]):

$$\lim_{M \to \infty} P_{t,n}^{\mathrm{MD}}(M) = \lim_{M \to \infty} P_{t,n}^{\mathrm{FA}}(M) = 0, \quad \forall t, n. \tag{21}$$

This important result implies that in a massive MIMO system, in which M can be larger than 100, the AMP-based grant-free access scheme is able to detect device activity with extremely high accuracy in the massive IoT connectivity systems.

Example 3: Here we provide a numerical example to show the power of massive MIMO for massive IoT connectivity, under the same setup that is used in Examples 1 and 2. Fig. 8 shows the probabilities

of missed detection and false alarm (which are made equal by adjust the detection threshold) versus pilot sequence length L, with $M = 64, 128, 256$ antennas at the BS. Here, similar to Example 2, we define the system-level missed detection and false alarm probabilities as $P^{\text{MD}}(M) = \sum_{n=1}^{N} P_n^{\text{MD}}(M)$ and $P^{\text{FA}}(M) = \sum_{n=1}^{N} P_n^{\text{FA}}(M)$, where $P_n^{\text{MD}}(M)$ and $P_n^{\text{FA}}(M)$ denote the missed detection and false alarm probabilities of device n achieved by AMP after its convergence. As compared to Fig. 6, it is observed that with $M = 256$ antennas at the BS, both the missed detection and false alarm probabilities can be driven down to 10^{-12} when the pilot sequence is $L = 120$, several orders of magnitude lower than the SMV case.

This article mainly focuses on the device activity detection performance under the grant-free access scheme. However, as shown in Fig. 4, besides device activity detection, channel estimation is performed as well via the metadata; moreover, data also should be decoded. Fortunately, the state evolution of the AMP algorithm enables us to characterize the channel estimation performance analytically, thus making it possible to quantify the user achievable rate with the effect of device activity detection into consideration [7]. Readers interested in information-theoretical studies on the capacity of the massive IoT connectivity systems with randomly active devices (also known as many-access channel) can refer to [20], [21]. These references provide a justification for our proposed strategy to first detect the user activity via preambles then decode the user messages, i.e., the grant-free access scheme shown in Fig. 4.

V. AMP-BASED DEVICE ACTIVITY DETECTION WITH EMBEDDED INFORMATION

The AMP algorithm is introduced above for device activity detection. In this section, we show how a modified version of AMP may be used for non-coherent detection of information bits embedded in the pilot transmission. Although the two-phase grant-free access scheme shown in Fig. 4 works very well for most of the cases when the user messages are of moderate and large size [7], [20], [21], as discussed at the end of last section, the strategy discussed in this section can be an effective alternative in the special case when very short messages (1 or several bits) are transmitted.

A. Motivation

In many applications the amount of data to be transmitted per block may comprise only a small number of information bits, or even a single bit. This situation is particularly common in control signaling, where the message may contain acknowledgment (ACK/NACK) bits in a retransmission protocol, or simply a concise request for a particular kind of response from the BS.

The transmission of extremely short packages is a challenging problem from two perspectives. First, fundamentally, the protection of very short packets against transmission errors is very expensive. For a single bit, repetition coding is the only possible strategy and for short blocks, block codes with low coding gains must be used. Second, as only error probability matters, capacity is an irrelevant metric. In

fact, for extremely short blocks even finite-block-length information theory becomes inapplicable as the corresponding bounds and approximations are too loose to be of practical value.

As an aside, it is noteworthy that most academic work tend to deal with the transmission of long coded blocks, with Shannon capacity as the primary performance metric. Contrarily, much of the effort invested in standardization and system design is concerned with the transmission of short data blocks on the control plane, for which Shannon capacity is, mostly, an illegitimate performance measure. An explanation for this situation might be that digital transmission on the control plane is too hard to model and tackle with rigorous information theory: there is no "Shannon theory" available for its analysis – whereas in contrast, established recipes are available for capacity analysis of long-block transmission. A contributing reason might also be that many academic researchers simply are unaware of the importance and the magnitude of the problem.

There are practical solutions for transmission of a single bit of control information. For example, [22] considers the joint transmission of linearly coded payload data and a single "additional bit". The transmitter uses the additional bit, through a one-to-one mapping, to select one out of two possible codebooks for the encoding of the payload. The receiver uses a fast algorithm to detect which codebook that was used, such that the additional bit can be detected before attempting to decode the payload data.

B. Algorithm Design

In the context of grant-free random access with non-orthogonal pilots, the main focus of our discussion, a small number, say J, of bits $b_1, \ldots, b_J$ may be encoded as follows [23], [24]. Each terminal is assigned, a priori, 2^J distinct, typically non-orthogonal, pilots. Upon transmission, the terminal uses the bits $\{b_i\}$ to select one of these 2^J pilots; specifically, it selects pilot number $1 + b_1 + 2b_2 + 4b_3 + \cdots + 2^{J-1}b_J$ – which depending on the bits $\{b_i\}$ ranges from 1 to 2^J. The BS detects "activity" using the AMP algorithm – but now, importantly, "activity" means the combination of the event that a particular terminal is active, and that a particular string of J bits is being communicated. One may think of the resulting communication scheme as non-coherent transmission.

The analytical model for device activity detection with embedded information in a massive MIMO system is given by:

$$\boldsymbol{Y} = \bar{\boldsymbol{A}}\bar{\boldsymbol{X}} + \boldsymbol{Z}, \tag{22}$$

where $\bar{\boldsymbol{A}} = [\boldsymbol{a}_{1,1}, \cdots, \boldsymbol{a}_{1,2^J}, \cdots, \boldsymbol{a}_{N,1}, \cdots, \boldsymbol{a}_{N,2^J}] \in \mathbb{C}^{L \times 2^J N}$ denotes the collection of all the $2^J N$ pilots that can be used by the devices, and $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_{1,1}, \cdots, \bar{\boldsymbol{x}}_{1,2^J}, \cdots, \bar{\boldsymbol{x}}_{N,1}, \cdots, \bar{\boldsymbol{x}}_{N,2^J}]^T \in \mathbb{C}^{2^J N \times M}$ denotes the collection of all the $2^J N$ effective channels of the devices. Specifically, the effective channel is modeled as $\bar{\boldsymbol{x}}_{n,i} = \alpha_{n,i}\boldsymbol{h}_n$, $n = 1, \cdots, N$ and $i = 1, \cdots, 2^J$, where

$$\alpha_{n,i} = \begin{cases} 1, & \text{if user } n \text{ is active and its } i\text{th pilot is used,} \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

As compared to (13) for sole device activity detection, the dimensions of sensing matrix $\bar{A}$ and effective channels $\bar{X}$ are enlarged by a factor of 2^J to embed J bits information.

The AMP for device activity detection in the form described above in principle could be directly applied to this problem as it stands. However, significantly, it is suboptimal because the BS knows *a priori* that among the 2^J pilots assigned to each terminal, only one can be active at a time, i.e., if $\alpha_{n,i} = 1$, then $\alpha_{n,j} = 0$, $\forall j \neq i$.

Here we discuss the modified AMP algorithm for joint detection of user activity and embedded information bits, as proposed in [23]. For conciseness of the exposition we focus here on the case of a single embedded bit b (for which we omit the index) – i.e., $J = 1$; then each user is assigned one out of two unique, but generally non-orthogonal pilot sequences. The modification of the AMP should introduce the constraint that out of the two possible pilots, at most one may be transmitted at at a time; the possible options are that either none of these pilots are sent (device silent), the first one is sent (device active and communicates "0"), or the second one is sent (device active and communicates "1"). The overarching idea is to modify the AMP denoiser function, $\eta_{t,n}(\cdot)$, to take into this constraint into account.

In more details, similar to (16), let $\hat{x}_{n,1} = \bar{x}_{n,1} + \Sigma^{\frac{1}{2}} v_n$ and $\hat{x}_{n,2} = \bar{x}_{n,2} + \Sigma^{\frac{1}{2}} v_n$ be the two vectors associated with the two possible pilots (for information bit "0" respectively "1") for device n; we omit the iteration index t of the AMP algorithm here for brevity. The statistical characterization of $\hat{x}_{n,1}$ and $\hat{x}_{n,2}$ is

$$\hat{x}_{n,i} \sim \begin{cases} \mathcal{CN}(\mathbf{0}, \beta_n \mathbf{I} + \Sigma), & \text{if } \alpha_{n,i} = 1, \\ \mathcal{CN}(\mathbf{0}, \Sigma), & \text{if } \alpha_{n,i} = 0, \end{cases} \quad i = 1, 2. \tag{24}$$

Based on these characterizations, we construct the following likelihood ratios:

$$\lambda_{n,i} = \frac{p(\hat{x}_{n,i} | \alpha_{n,i} = 1)}{p(\hat{x}_{n,i} | \alpha_{n,i} = 0)} = \frac{|\Sigma|}{|\beta_n \mathbf{I} + \Sigma|} \exp\left(-\hat{x}_{n,i}^H ((\beta_n \mathbf{I} + \Sigma)^{-1} - \Sigma^{-1}) \hat{x}_{n,i}\right). \tag{25}$$

We now re-work the denoiser, such that in each time update, the constraint is taken into account that at most one of the vectors $\bar{x}_{n,1}$ and $\bar{x}_{n,2}$ can be nonzero. Suppose that device n is detected to be active. In principle, a comparison of $\lambda_{n,1}$ and $\lambda_{n,2}$ to a threshold would yield a hypothesis test, that could be used to discriminate between the two possibilities $\alpha_{n,1} = 1$ and $\alpha_{n,2} = 1$ or equivalently $b = 0$ and $b = 1$; one of $\bar{x}_{n,1}$ and $\bar{x}_{n,2}$ could then be set to zero based on the outcome of this test. In this process, taking a soft decision as given in (19) instead is preferable in order to avoid taking premature incorrect decisions on the embedded bits, which may propagate to subsequent iterations. Experimentation in [23] has shown that a good heuristic is to use soft decision obtained by taking the original soft thresholding denoiser function given in (19) and multiplying the denoisers for $\bar{x}_{n,1}$ and $\bar{x}_{n,2}$ by $\gamma(\lambda_{n,1}/(\lambda_{n,1} + \lambda_{n,2}))$ and $\gamma(\lambda_{n,2}/(\lambda_{n,1} + \lambda_{n,2}))$, respectively, where $\gamma(x) = 1/(e^{-c(x-0.5)} + 1)$ is a modified sigmoid function with its inflection point at $x = 0.5$, where c is a parameter to control the sharpness of the sigmoid
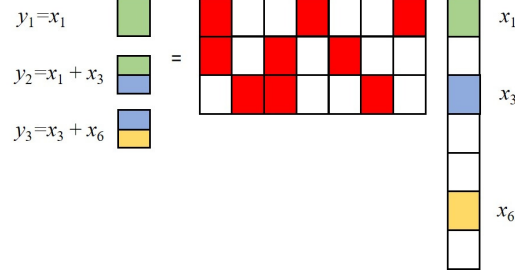
Fig. 9. Example to illustrate the philosophy of the sparse-graph based compressed sensing algorithm [15]. In this example, $\boldsymbol{x}$ is a 3-sparse signal, in which the zero entries are colored in white, while the non-zero entries, i.e., x_1, x_3, and x_6, are colored in green, blue, and yellow, respectively. Moreover, the sensing matrix $\boldsymbol{A}$ is sparse, in which the zero entries are colored in white, while the non-zero entries are colored in red. x_1 is detected from y_1. Then, x_1 is subtracted from y_2 so that x_3 is detected. Similarly, x_6 is detected after x_3 is removed from y_3.

function. The intuition is that the larger the likelihood ratio $\lambda_{n,1}$ is relative to $\lambda_{n,2}$, the more likely it is that $\alpha_{n,1} = 1$ or $b = 0$ and the closer the weight for $\bar{\boldsymbol{x}}_{n,1}$ is to unity, and the closer the weight for $\bar{\boldsymbol{x}}_{n,2}$ is to zero. This way, the effect of the denoiser on $\bar{\boldsymbol{x}}_{n,1}$ is similar to the effect of the soft thresholding (19) as used in the original AMP algorithm solely for device activity detection, whereas on $\bar{\boldsymbol{x}}_{n,2}$ is instead pushed down towards zero. A similar interpretation holds for the opposite case when $\lambda_{n,1} < \lambda_{n,2}$.

Importantly, while the modified denoiser outlined here yields good results in numerical experiments, it is not optimal in any known sense. Research opportunities are available to find improved denoisers that can make a better utilization of the constraint that at most one of $\bar{\boldsymbol{x}}_{n,1}$ and $\bar{\boldsymbol{x}}_{n,2}$ can be nonzero. An extension of the modified AMP denoiser to the case of multiple embedded bits is available [24].

A final remark is that the embedding of one or several information bit(s) of course incurs an expense of storage of more pilot sequences at the device and at the BS. Also for given coherence block length, more resources need be dedicated to pilot transmission in order to maintain the same error probability performance. Yet, in the case to transmit very short messages, the embedding scheme has been shown to be efficient compared to conventional scheme consisting of pilot-based channel estimation (using the sparsity/AMP-based techniques proposed here) followed by coherent detection [24].

VI. OTHER COMPRESSED SENSING TECHNIQUES FOR DEVICE ACTIVITY DETECTION

Besides AMP, researchers with diverse backgrounds have developed many other powerful algorithms to reconstruct sparse signals from low dimensional linear measurements as given in (2) and (13). These compressed sensing algorithms can also be leveraged in our considered massive IoT connectivity setting for device activity detection, e.g., coded slotted ALOHA.

One powerful algorithm of low complexity is the so-called sparse-graph based compressed sensing algorithm [15], where the sensing matrix $\boldsymbol{A}$ is designed by sparsifying each row of the measurement matrix with zero patterns guided by sparse-graph codes. The reason for such a sparse sensing matrix design is to disperse the signal into single-tons that only contain one non-zero element in $\boldsymbol{x}$ and peel

them off from multi-tons that contain two or more non-zero elements in $\boldsymbol{x}$ such that they can become single-tons.

Fig. 9 gives a simple example to briefly illustrate how this algorithm works to recover $\boldsymbol{x}$ in the ideal case without noise in (2), in which the dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$ are $N = 7$ and $L = 3$, respectively, and x_1, x_3, x_6 are non-zero entries in $\boldsymbol{x}$. Moreover, the sensing matrix is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{26}$$

Due to the sparsity in both $\boldsymbol{A}$ and $\boldsymbol{x}$, in the received signal $\boldsymbol{y} = [y_1, y_2, y_3]^T$, y_1, y_2, and y_3 only contain information about x_1, $x_1 + x_3$, and $x_3 + x_5$, respectively. Thus, x_1 can be detected from the single-ton y_1. Next, x_1 is removed from y_2 and y_3, which become single-tons so that x_3 and x_6 can be decoded. Note that the effect of the channels is not taken into account in this example.

Density evolution, a powerful tool in modern coding theory, tracks the average density of remaining edges that are not decoded after a fixed number of peeling iteration. The convergence of the above algorithm is guaranteed by showing the convergence of the density evolution towards zero.

We remark that the above "successive interference cancellation" procedure is the principle of coded slotted ALOHA, a powerful multiuser access scheme in which the active devices transmit replicas of their packets in randomly chosen slots that contain both metadata (i.e., pilot sequences) and data. A successful detection of a packet replica in some slot enables removal of the related replicas from the slots in which they occur. This, in turn, lowers the number of colliding packets in the affected slots and boosts their detection probability, instigating new rounds of successive interference cancellation etc. If a single user packet can be detected in a slot, then the entries in $\boldsymbol{x}$ denote packets of active users, and entries in the sparse sensing matrix $\boldsymbol{A}$ denote the choice of the slots where the packets are repeated. On the other hand, the possibility to decode multiple user packets in a slot is also discussed in [13], [25] to improve the detection performance.

We remark that besides AMP and the sparse-graph based algorithm, many powerful compressed sensing algorithms exist in the literature, including LASSO [26], Orthogonal Matching Pursuit (OMP) [27], and so on. Further, the group-sparsity in the MMV model (13) can also be utilized in LASSO [28], in which the ℓ_1/ℓ_2 penalty, i.e., the sum of ℓ_2 norm penalty, is used to promote the desired sparsity pattern. The potential to apply these advanced compressed sensing techniques for user activity detection has been discussed in [29], [30]. It would be of great significance to investigate which compressed sensing algorithm is best suited for device activity detection in the massive IoT connectivity setting, in terms of the complexity of pilot sequence design, the pilot sequence length required to achieve reasonable device detection accuracy, the corresponding missed detection and false alarm probabilities performance and channel estimation performance, etc.

VII. Conclusions

A key feature of the future IoT network is the massive number of devices, e.g., sensors, actuators, and etc., each with sporadic data traffic. Facilitating the data transmission from so many IoT devices with extremely low latency poses plenty of new research challenges to the signal processing community. To embrace the upcoming era of IoT, this article advocates a grant-free access scheme that mitigates the delay arising from the contention resolution in the current random access scheme, and outlines a compressed sensing based approach for device activity detection to enable the grant-free access scheme to work. Most notably, the massive MIMO technology, originally proposed for improving the spectrum efficiency of human-type communications, is able to boost the device activity detection accuracy remarkably for massive IoT connectivity as well, with the aid of the MMV-based AMP algorithm. We have also discussed about the potential to decode some short messages along with the device activity detection process.

References

[1] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711-1726, Sep. 2016.

[2] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233-2243, Nov. 2014.

[3] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86-93, June 2013.

[4] O. Y. Bursalioglu, C. Wang, H. Papadopoulos, and G. Caire, "RRH based massive MIMO with on the "fly" pilot contamination control," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016.

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

[6] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part I: Device activity detection and channel estimation," to appear in *IEEE Trans. Signal Process.*, 2018. [Online] Available: https://arxiv.org/abs/1706.06438.

[7] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part II: Achievable rate characterization," to appear in *IEEE Trans. Signal Process.*, 2018. [Online] Available: https://arxiv.org/abs/1706.06433.

[8] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914-18918, Nov. 2009.

[9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[10] T. L. Marzetta, E. G. Larsson, H. Yang and H. Q. Ngo, *Fundamentals of Massive MIMO*, Cambridge, UK: Cambridge University Press, Nov. 2016.

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

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

[13] E. Paolini, C. Stefanović, G. Liva, and P. Popovski, "Coded random access: How coding theory helps to build random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, June 2015.

[14] 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.

[15] X. Li, S. Pawar, K. Ramchandran, "Sub-linear time time compressed sensing using sparse-graph codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2015, pp. 16451649.

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

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

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

[19] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6920-6941, Oct. 2011.

[20] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516-3519, Jun. 2017.

[21] W. Yu, "On the fundamental limits of massive connectivity," in *Inf. Theory and Appl. (ITA), Workshop*, Feb. 2017.

[22] E. G. Larsson and R. Moosavi, "Piggybacking an additional lonely bit on linearly coded payload data," *IEEE Wireless Commun. Lett.*, pp. 292-295, Aug. 2012.

[23] K. Senel and E. G. Larsson, "Device activity and embedded information bit detection using AMP in massive MIMO," in *Proc. IEEE Global Commun. Conf. (Globecom) workshops*, 2017.

[24] 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. of International ITG Workshop on Smart Antennas (WSA)*, Mar. 2018.

[25] G. Wunder, C. Stefanović, P. Popovski and L. Thiele, "Compressive coded random access for massive MTC traffic in 5G systems," in *Proc. of 49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2015.

[26] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, 1996.

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

[28] L. Jacob, G. Obozinski, and J. Vert, "Group Lasso with overlap and graph Lasso," in *Proc. Int. Conf. on Machine Learning, Montreal*, QC, Canada, Jun. 2009.

[29] 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.

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