An Opportunistic Thresholding Detector for IoT Random Access in Massive MIMO

Shahrokh Farahmand and Mohammad Naseri Tehrani

Abstract—Introduction of internet of things (IoT) into the massive MIMO random access scheme has generated many complications. Specifically, the number of IoT devices is significantly greater than mobile broadband. On the other hand, their traffic is intermittent. Given the different nature of IoT traffic, on-off random access channel can model it accurately. An extension of this access model has been recently applied to massive MIMO scenarios where it was demonstrated to entail common support recovery in a jointly sparse multiple measurement vector (MMV) problem. Several sparse recovery algorithms have also been proposed. However, we are not particularly bothered by the choice of algorithm but pose a new question which has been previously overlooked. While measurements size is severely constrained and dictated by the environment and the physical nature of communication channel, which is out of our control, the number of antennas is at our disposal to increase at will. The question is can we relieve the burden on number of measurements by increasing number of antennas and still maintain a satisfactory recovery performance. The so-called trivial pursuit (TP) gives a positive answer for independent sensing matrices. Motivated by TP, we introduce a novel opportunistic thresholding detector (OTD) which exploits partial de-correlation of sensing matrices to improve support recovery performance with minimum modifications and thus keeping the low-complexity nature of the algorithm. Extensive simulations have corroborated the superior performance of OTD compared to ordinary thresholding.

Index Terms—Sparsity, Support Recovery, Massive MIMO, On-Off Channel, Random Access, Internet of Things (IoT).

I. INTRODUCTION

According to predictions, by 2022, there will be around 18 billion IoT devices connected to the network while there will be only around 11.6 billion connected smart phones, tablets, pcs, laptops, and fixed phones overall [1] p.33. Subsequently, any wireless technology envisioned for 5G and beyond should have built-in potential to deal with the type of traffic IoT devices generate. Compared to persistent high data-rate mobile broadband which communicates voice and video, most IoT traffic is low-rate and sporadic in nature but a very large number of devices should be served. Before proceeding, let us introduce the technology we want to enable to successfully tackle IoT traffic.

Marzetta was the first to investigate multi-user massive MIMO where a BS with a large number of antennas communicated with a number of single antenna user terminals (UTs) [2]. [3]. He proved a major result which is now known as channel hardening. Basically, it states that with an unlimited number of antennas at BS and using orthogonal training sequences to learn the channels between BS and UTs, using simple precoding/combining methods at downlink/uplink, the channel between the BS and different UTs decouple into single-user deterministic channels free from fast-fading, interference, and noise [3]. This comes as a consequence of law of large numbers [4]. While pilot contamination problem (see [5] and references therein) has been successfully attacked by various researchers, initial access to a massive MIMO BS had not received due attention until two years ago. In massive MIMO, one typically assume that orthogonal pilot sequences are already assigned uniquely to UTs. Therefore, a separate MAC is not needed because channel hardening decouples UTs’ channels. However, the number of orthogonal sequences is limited and cannot be allocated to UTs on a permanent basis. Rather, some dynamic allocation is needed on a per demand basis and the process for this random access to pilots (RAP) has been recently investigated [6]. Some versions of the algorithm is able to recover data even in the presence of collisions [7], [8]. This scheme is reminiscent of slotted Aloha and has been considered for 4G LTE random access as well [9].

The orthogonal pilots shortage is even more pronounced for IoT connections as their numbers is considerably larger. These devices are inactive most of the time but become active and communicate at a low data rate once in a while. In an extreme case, these IoT devices do not communicate any data but only announce their activity to their corresponding application. For example, smoke detector only goes off when there is a fire. Motion detection sensors in a home security application go off when someone passes in front of them and so on. This transmission scheme is desirable as devices with no data do not transmit at all thus preserving their battery power. Consequently, massive MIMO BSs should be able to detect and distinguish various IoT devices when they announce their activity. Accordingly, we consider the so-called on-off random access channel which is a good approximation of IoT devices behavior. This channel has been investigated in the SISO case by several authors [10], [11], [12].

In the on-off random access channel, users are assigned independent pseudo-random non-orthogonal codes. These codes exist in abundance and can be assigned to individual users on a permanent basis. As a consequence of non-orthogonality, active users are sure to interfere. When a particular user becomes active it transmits its code. We prefer the term code over pilot to emphasize non-orthogonality. Allowing for error-free detection in spite of interference, is the fact that active users form a small subset of total users. Hence, exploiting sparsity-enabling techniques, one can recover all the active users form a small subset of total users.
users correctly provided that code length is sufficiently large. This scenario is reminiscent of overloaded CDMA systems but with a low activity factor [13]. An extension of on-off random access channel which involved simultaneous activity detection and data decoding was looked at by [14] and [15] in the massive MIMO context. These references realized that user activity detection in such channel amounts to a sparse common support recovery problem for multiple measurement vectors (MMVs). Subsequently, [14] proposed CoSaMP and MMV thresholding while [15] applied group least absolute shrinkage and selection operator (group-LASSO) to detect active terminals. In this manuscript, we are not particularly bothered by the choice of algorithm. Instead, we pose a novel problem that have not been looked at before. In massive MIMO, the number of measurements is limited by the length of a coherence interval [16] which is defined as the product of coherence frequency and coherence time. This limitation is dictated by the environment such as number and position of scatterers and Tx/Rx systems, their relative speeds, and so on and is completely out of our control. On the other hand, the number of channels in MMV setup, which amounts to the number of antennas here, can be arbitrarily large. The question we pose is that can we relieve the burden on number of measurements by increasing the number of antennas?

In order to find a convenient answer, we surveyed fundamental sufficient and/or necessary measurement inequalities derived from upper bounds on probability of support recovery error for a fixed sensing matrix in the presence of noise. Unfortunately, the results are inconclusive meaning that sufficient conditions does not offer a desirable positive answer while necessary conditions point to the contrary [17]. There is a gap between the two which has not been closed up to now. Our interest in the independent sensing matrices model were raised by the trivial pursuit (TP) algorithm proposed in [18], where it is demonstrated that even with a single measurement, TP, which is a simple thresholding algorithm, can recover the true support error-free provided that number of antennas is large. Amazingly, the proof is a simple application of law of large numbers (LLN) to the decision statistics. In the case of independent sensing matrices, there also exists a non-zero gap between the necessary and sufficient conditions for optimal detection in noisy settings. However, certain sufficient conditions suggest that number of measurements can be considerably lowered by increasing number of antennas [19]. Furthermore, TP is a simple proof of concept for this phenomenon.

Given the aforementioned results, our contributions towards addressing the posed question is summarized as follows:

1. We conduct a survey to summarize the existing results on the performance of optimal recovery algorithms for fixed sensing matrices, which is the MMV model occurring in MIMO on-off random access, to check if the posed question can be addressed. We show that the results are inconclusive.

2. Motivated by TP in [18], we offer intuition on why a model change towards independent sensing matrices is well-warranted.

3. Based on item 2, we propose a novel low-complexity algorithm that can bring about a partial model change in the MMV problem and thus significantly improve performance in the small measurement regime, while requiring minimum modifications.

4. Corroborate improved performance of the proposed algorithm 3 via comprehensive simulations.

5. Analyze the proposed algorithm to derive the so-called measurement inequality which determines the rate at which system parameters should increase to ensure probability of failure in exact support recovery goes to zero. This item is reported in an accompanying paper [20].

The rest of the paper is organized as follows. Section II presents the problem formulation. Section III performs a survey of analytical fundamental limits for MMV problem with fixed sensing matrices. Section IV provides intuition on why a shift towards independent sensing matrices is well-warranted. Based on the observations of Section IV, Section V introduces the novel opportunistic thresholding detector (OTD). Finally, Section VI provides numerical results and Section VII concludes the paper.

**Notation:** Uppercase boldface letters are used for matrices and lowercase boldface letters are used for vectors. Calligraphic letters are used to represent index sets. If $S$ is an index set, $A_S$ denotes the submatrix generated by keeping only those columns of $A$ whose index belongs to $S$. Similarly, $A_S^S$ denotes a submatrix generated by keeping only rows of $A$ corresponding to the index set $S$. $\|A\|_F$ denotes the Frobenius norm of $A$ and $\|x\|_\ell$ represents $\ell$ norm of vector $x$. Finally, $\|A\|_{\ell,\ell,k}$ denotes the application of vector norm $\ell$ to the columns of $A$ separately and reporting the result as a row vector. Then applying vector norm $k$ to this row vector. It is usually referred to as a mixed norm.

**II. Problem Formulation**

Let us consider the uplink of a single-cell with a single base station (BS) where an $M$-antenna massive MIMO BS serves $N$ single antenna users. In an IoT scenario, $N$ is large but each user transmits only sporadically. In the extreme case, which is considered here, each active user transmits a single bit and inactive users do not transmit at all. It is assumed that $K$ users out of $N$ are active at each random access slot where $K \ll N$. Note that we need not know $K$ exactly but an upper bound is sufficient. We consider the extension of on-off random access channel to massive MIMO. Our model is almost the same as that of [11], [14] and [15] with a few minor differences to be specified later on.

Continuing with our setup, one considers a coherence interval [16] which consists of $T_c$ time slots, where each time slot equals one OFDM symbol length. Furthermore, $T_c$ represents the coherence time and $B_c$ denotes the number of OFDM sub-carriers falling into a single coherence bandwidth. This coherence interval will offer $T_cB_c$ channel uses or a resource block of length $T_cB_c$ among which channel gain matrix is both fixed and flat. Note that size of a coherence interval is a function of physical channel properties such as number of scatterers, their relative distances to BS and users, and user/scatterer speeds which are all given and out of our control.
Typical values for coherence interval in various scenarios can be found in [16]. Now, define $L$ to be the pseudo-random code length which is assigned independently and permanently to each individual user. Note that these codes are not orthogonal as $N$ is very large and there are not a sufficient number of orthogonal codes for assignment on a permanent basis. We use $c_n$ which is a vector of size $L \times 1$ to represent the code corresponding to user $n$. Here, we chose Rademacher sequences as codes, where each entry $c_n(i)$ takes values $\{\pm 1/\sqrt{L}\}$ equiprobably. The choice of Rademacher codes is made based on a combination of satisfactory simulation performance and convenience of mathematical analysis. For example, random Gaussian codes lend themselves easily to mathematical analysis. However, their performance was not very appealing in the simulations. The exact opposite was true for Gaussian on a unit sphere or normalized length Gaussian codes which performed satisfactorily in simulations while proving to be a challenge to analysis. Setting $L$ equal to a positive integer multiple of $B_c$ proves to be convenient. Upon defining the positive integer $a \geq 1$, one has $L = aB_c$ and $T_a := \lfloor T_a B_c / L \rfloor = \lfloor T_a / a \rfloor$. Note that the same random access procedure is utilized $T_a$ times per coherence interval, but each time a new independent set of $K$ users are active and transmitting. Therefore, a single coherence time is divided into $T_a$ random access slots. To determine the active/inactive status of each user an $N \times 1$ vector $q$ is defined whose $n$’th component correspond to user $n$. If user $n$ is active in a particular random access slot, $q_n = q(n) = 1$ otherwise $q_n = q(n) = 0$. As a result, $q$ consists of $K$ ones and $N-K$ zeros. In addition, we consider a block fading model on fast-fading channel gains meaning that channel is constant for one coherence time and changes independently afterwards.

Next, define $H$ of size $M \times N$ to be the channel gain matrix between BS and users where its $(m,n)$’th entry $H_{m,n}$ represents channel gain from user $n$ to antenna $m$. It is assumed that $H_{m,n}$ is independent identically distributed (IID) Gaussian with zero-mean and unit variance. The path loss is absorbed into the power term to be defined later on. At a particular random access slot, each active user transmits $\sqrt{P_n c_n}$, and inactive users remain silent. BS’s $m$’th antenna receives $y_m = X_C W + y_m$. In the single measurement vector (SMV) setup, $M = 1$, it is well-known that if $L$ increases faster than a certain rate, recovery error can be made arbitrarily small [21]. Two main criterion for recovery error are Mean Square Error (MSE) also known as $\ell_2$-norm and probability of exact support recovery. Note that the non-zero support of $X$ represents the active users. Hence, we are interested to know if one can recover $X$ from $Y$ and $C$ with a small $L$ and large $M$. In the single measurement vector (SMV) setup, i.e. $M = 1$, it is well-known that if $L$ decreases faster than a certain rate, recovery error can be made arbitrarily small [21]. Two main criterion for recovery error are Mean Square Error (MSE) also known as $\ell_2$-norm and probability of exact support recovery. Note that the non-zero support of $X$ represents the active users. Hence, in our setup one only needs to recover the support accurately. As a result, MSE is not a good criterion because MSE can be small while support is reconstructed incorrectly [21]. Therefore, probability of exact support recovery will be our choice of criterion. In this manuscript, we try to address
the previously posed question. First, a survey is conducted to look for a sure positive or negative answer. Upon offering the required intuition, the novel OTD is proposed.

III. PERFORMANCE ANALYSIS SURVEY

In this section, performance of best and worst recovery algorithms are looked at in detail in order to derive general conclusions on the performance of support recovery via model (5). The best recovery algorithm is an exhaustive search which maintains prohibitive complexity but nevertheless determines the fundamental performance limit for (5). As the worst recovery algorithm one can select a simple low-complexity maximum-correlation or thresholding detector. In addition, performance differs depending on noiseless/noisy setting and worst-case versus average-case analysis. The difference between the two latter cases needs some clarification. For the single BS antenna scenario or multiple user SISO, a complete comparison of performance of various algorithms can be found in [10].

A. Average versus Worst Case Analysis

In the worst case analysis, one treats \( \mathbf{X} \) as deterministic and tries to find a measurement inequality such that all possible instances of \( \mathbf{X} \) with a particular sparsity degree or lower are recovered correctly. The worst-case analysis is valid but rather pessimistic because it should take care of all instances of \( \mathbf{X} \). On the other hand, simulations suggest that the true support recovery performance is considerably better than predicted by worst case analysis. Hence, average-case analysis treats \( \mathbf{X} \) as random with a usually continuous distribution over the non-zero components. Then, it provides a measurement inequality that offers correct support recovery for ALMOST all possible instances of \( \mathbf{X} \). Thus, there exist a set of \( \mathbf{X} \) instances with that particular sparsity degree which can not be correctly recovered. However, they form a set of measure zero. Therefore, average-case analysis is able to recover \( \mathbf{X} \) almost surely.

Here, we consider an illuminating example. Let us consider the noiseless SMV model

\[
y = \mathbf{A} \mathbf{x},
\]

where \( \mathbf{y} \in \mathbb{R}^{L}, \mathbf{x} \in \mathbb{R}^{N} \) and \( \mathbf{A} \in \mathbb{R}^{L \times N} \). It is desirable to recover \( \mathbf{x} \) from \( \mathbf{y} \) and \( \mathbf{A} \). Note that \( L < N \) but \( \mathbf{x} \) is \( K \)-sparse. Worst-case analysis suggests that if \( \mathbf{A} \) satisfies 2\( K \)-RIP (Restricted Isometry Property), then any \( \mathbf{x} \) with \( \| \mathbf{x} \|_0 \leq K \) can be uniquely recovered via exhaustive search. The proof is simple. Firstly, 2\( K \)-RIP means that

\[
(1 - \delta)\| \mathbf{x} \|_2 \leq \| \mathbf{A} \mathbf{x} \|_2 \leq (1 + \delta)\| \mathbf{x} \|_2
\]

for a \( 0 \leq \delta < 1 \) and for all \( \mathbf{x} \) such that \( \| \mathbf{x} \|_0 \leq 2K \). Now, suppose there exists two different vectors \( \mathbf{x}_1 \neq \mathbf{x}_2 \) such that \( \| \mathbf{x}_1 \|_0 \leq K \), \( y = \mathbf{A} \mathbf{x}_1 \) and \( \| \mathbf{x}_2 \|_0 \leq K \), \( y = \mathbf{A} \mathbf{x}_2 \). Then subtracting these two equations one gets \( \mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) = 0 \). However, \( (\mathbf{x}_1 - \mathbf{x}_2) \) is at most 2\( K \) sparse. Therefore, (7) suggest that \( \| \mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) \|_2 \geq (1 - \delta)\| \mathbf{x}_1 - \mathbf{x}_2 \|_2 \) which is a contradiction. Thus, any \( K \)-sparse \( \mathbf{x} \) can be uniquely recovered. Hence, any deterministic or random \( \mathbf{A} \) satisfying 2\( K \)-RIP should ensure recovery of \( K \)-sparse \( \mathbf{x} \). Now, if \( \mathbf{A} \) is a random matrix made up of IID Gaussian entries then any \( L \times L \) submatrix of \( \mathbf{A} \) will be full-rank almost surely over \( \mathbf{A} \). Thus, \( \mathbf{A} \) satisfies \( L \)-RIP almost surely. Therefore, it is sufficient for \( L \) to be greater than or equal to 2\( K \). Measurement inequality yields \( L \geq 2K \).

Now, average case analysis for (6) assumes that non-zero elements of \( \mathbf{x} \) come from a continuous, lets say Gaussian, distribution. Given that only \( K \) elements of \( \mathbf{x} \) are non-zero, \( \mathbf{A} \mathbf{x} \) is made up of a linear combination of \( K \) columns of \( \mathbf{A} \). Provided that any \( L \times K \) sub-space formed from a random selection of columns of \( \mathbf{A} \) are distinct from other such subspaces then \( \mathbf{x} \) will be recovered unless it falls into the intersection of two such formed subspaces. However, the intersection will be a subspace of lower dimension and thus of measure zero, meaning that the probability that a randomly generated \( \mathbf{x} \) falls into the intersection of two distinct subspaces of dimension \( K \) equals zero. Thus \( \mathbf{x} \) can be recovered almost surely over distribution of \( \mathbf{x} \) for a deterministic \( \mathbf{A} \). However, if \( \mathbf{A} \) is also random, say made up of IID Gaussian entries, then any \( L \times (L - 1) \) sub-space formed from columns of \( \mathbf{A} \) is distinct with probability one over \( \mathbf{A} \). Hence, it suffices for perfect recovery to have \( L - 1 \geq K \) or \( L \geq K + 1 \). Given this measurement inequality, perfect recovery using exhaustive search occurs almost surely over \( \mathbf{x} \) and \( \mathbf{A} \).

This subsection is concluded with an example where \( L = 3 \), \( N \) is arbitrary and \( K \) can be \( K = 1 \) or \( K = 2 \). If \( K = 1 \), Fig. [1] demonstrates that in a 3 dimensional space, subspaces of dimension 1 are lines and those distinct lines passing through origin maintain an empty intersection set aside the origin. Hence, worst-case analysis suggests that perfect recovery for \( K = 1 \) is possible. Remember that \( L = 3 \geq 2K = 2 \). Next, consider \( K = 2 \) and Fig. [2] Subspaces of dimension 2 form planes in \( \mathbb{R}^3 \). Distinct planes passing through origin have non-zero intersections which are lines. Therefore, worst case analysis tells us that we can not recover all instances of \( \| \mathbf{x} \|_0 \leq 2 \) correctly. Specifically, if one selects \( \mathbf{x} \) to be a point on one of these intersection lines, unique recovery is impossible as the line belongs to two viable planes. Remember that \( L = 3 \geq 2 \times 2K = 4 \). On the other hand, average-case analysis suggests that if we select a point on the plane with a continuous probability, then the probability of falling on a specific line is zero. Hence, a random \( \mathbf{x} \) will not fall on any of the intersection lines with probability 1. Note that \( L = 3 \geq K + 1 = 3 \). Therefore, recovery is possible for \( K = 2 \) using average case analysis.

B. Performance Limits of an Exhaustive Search Decoder

First, one looks at the upper performance limit via the model (5). This limit can be achieved using an exhaustive search decoder. The decoder is defined in Table I.

Worst-case noiseless case has been analyzed by [22] where it was proved that perfect recovery is guaranteed provided that

\[
\text{spark}(\mathbf{C}) - 1 + \text{rank}(\mathbf{Y}) \geq 2K.
\]

Later [23] showed that this condition is also necessary hence it is a fundamental limit. To simplify (8), note that spark of
a matrix is defined to be the smallest number of columns that are linearly dependent. For random square matrices of dimension \( L \) with IID entries whose second and fourth moments are bounded, such as Gaussian and Rademacher, \cite{24} has proved that the smallest eigenvalue scales as \( \frac{1}{\sqrt{L}} \) with high probability. Therefore, it is bounded away from zero for finite \( L \). Thus, any \( L \times L \) submatrix of \( \mathbf{C} \) in \((5)\) is full-rank almost surely. Thus, it is concluded that \( \text{spark}(\mathbf{C}) = L + 1 \). Note the specific relation between \( \text{spark} \) and RIP property in a matrix. If a matrix has \( \text{spark} \) equal to \( L + 1 \), then it satisfies \( L\text{-RIP} \) property. As for the rank of \( \mathbf{Y} \), we know that it is bounded by the smallest dimension of the matrix. However, in the noiseless case, each column of \( \mathbf{Y} \) is made up of linear combination of \( K \) particular columns of \( \mathbf{C} \). Hence, rank of \( \mathbf{Y} \) is guaranteed to be less than or equal \( K \). Consequently, \((8)\) amounts to

\[
L + \min\{L, M, K\} \geq 2K \tag{9}
\]

Several important observations can be made from \((9)\). Firstly, \( M \) does offer a positive rule as long as \( M < \min\{L, K\} \). Once \( M \) becomes greater than or equal to \( \min\{L, K\} \), it will no longer benefit us and performance will be limited by \( \min\{L, K\} \). Secondly, for \( M \geq \min\{L, K\} \) the measurement inequality becomes \( L \geq K \) where one needs only half the measurements needed in the SMV worst-case noiseless scenario. Thirdly, similar to SMV case, \( N \) does not have an impact in the fundamental limit \((9)\).

Fig. 2. Distinct planes passing through origin intersect on a line

Fig. 1. Distinct lines passing through origin have an empty intersection set aside the origin

Focusing on noisy worst-case setup in \cite{24}, one defines the matrix \( \mathbf{Z} \in \mathbb{R}^{K \times M} \) to be \( \mathbf{Z} := \mathbf{X}^S \) where \( S \) denotes the nonzero support of rows of \( \mathbf{X} \) or the active user set. On par with worst-case analysis, \( \mathbf{Z} \) is treated as deterministic. The main theorem suggests that the following measurement inequality provides a necessary and sufficient condition for perfect recovery \cite{25, Thm 1}:

\[
L \ c(\mathbf{Z}) \geq N, \tag{10}
\]

\[
c(\mathbf{Z}) := \min_{T \subseteq S} \frac{1}{2|T|} \log \det \left( \mathbf{I}_M + \frac{1}{L \sigma_{w}^2} \mathbf{Z}^T \mathbf{Z} \right). \tag{11}
\]

Before interpreting this result, limitations of the analysis in \((10)\) is mentioned. Firstly, information-theoretic tools have been utilized which deal with very long Gaussian codes. Thus \( \mathbf{C} \) is made up of Gaussian entries which is different from our Rademacher model. Furthermore, \( L \) is assumed to be large which is not of much interest here. Unfortunately, as we see later on, almost all mathematical analysis tools available requires \( L \) to be large and small \( L \) proves to be a challenge.

To gain insight from \((10)\), one considers two special cases, originally considered in \cite{25}. In the high SNR regime where \( L \sigma_{w}^2 \ll 1 \), measurement inequality reduces to \cite{25, Eq (15)}

\[
L \frac{\min\{K, M\}}{2K} \log \left( \frac{1}{L \sigma_{w}^2} \right) \geq \log(N) \tag{11}
\]

Similar to \((9)\), it can be observed that \( M \) is helpful up to a point and offers only a minor improvement. Unlike the noiseless setup, \( N \) does have an impact. When all the \( M \) columns of \( \mathbf{Z} \) are the same, i.e., \( \mathbf{Z} := [\mathbf{z} | \mathbf{z} | \cdots | \mathbf{z}] \), measurement inequality reduces to \cite{25, Eq (18)}

\[
L \ min_{T \subseteq S} \frac{1}{|T|} \log \left( 1 + \frac{M}{L \sigma_{w}^2} \|\mathbf{z}_T\|_2^2 \right) \geq 2 \log(N) \tag{12}
\]

The effect of \( M \) is enduring in this case, meaning that as \( M \) increases, \( L \) decreases proportionally. This result can be deceiving because a particular \( \mathbf{Z} \) in the format \( \mathbf{Z} := [\mathbf{z} | \mathbf{z} | \cdots | \mathbf{z}] \) need not necessarily yield the worst possible selection of \( \mathbf{Z} \). To elaborate further, this scenario is equivalent to having \( M \) observations of the same signal vector in independent noise. Thus, a simple averaging of these

| Table I. Exhaustive Search Decoder |
|-----------------------------------|
| Initialization. For a \( K \) jointly sparse matrix \( \mathbf{X} \) we have \( \binom{N}{K} \) possible choices for support |
| Repeat for \( j = 1, \ldots, \binom{N}{K} \) |
| Pick the set \( S_j \), with cardinality equal to \( K \) as the true support |
| Run LS for the linear model \( Y = C_{S_j} \mathbf{X}_{S_j}^T + \mathbf{W} \) resulting respectively in the estimate \( \hat{\mathbf{X}}_j \) and error \( E_j \); |
| \( \mathbf{X}_j := (C_{S_j}^T C_{S_j})^{-1} \mathbf{C}_{S_j}^T Y \), |
| \( E_j := \left\| \mathbf{I} - C_{S_j} \left( C_{S_j}^T C_{S_j} \right)^{-1} C_{S_j}^T \right\|_F \) |
| End Repeat |
| Select \( j^* := \arg \min_j E_j \) to be the true hypothesis. Thus, \( S_{j^*} \) will represent the true support and \( \hat{\mathbf{X}}_{j^*} \) denotes the ML estimate |

| \( L \ c(\mathbf{Z}) \) | \( N \) |
|-----------------------------------|
| \( c(\mathbf{Z}) := \min_{T \subseteq S} \frac{1}{2|T|} \log \det \left( \mathbf{I}_M + \frac{1}{L \sigma_{w}^2} \mathbf{Z}^T \mathbf{Z} \right) \). |

| \( L \min_{T \subseteq S} \frac{1}{|T|} \log \left( 1 + \frac{M}{L \sigma_{w}^2} \|\mathbf{z}_T\|_2^2 \right) \geq 2 \log(N) \) |
vectors leads to a noiseless SMV scenario when \( M \) is large. Accordingly, for large \( M, L \geq 2K \) will be sufficient. This is the type of inequality we are interested in. That is, a large \( M \) that reduces the burden on \( L \). Unfortunately, it is not known if the structure \( Z := \{ z_1 \mid z_2 \mid \cdots \mid z_N \} \) does yield a worst-case choice or not. Provided \( M \) becomes large, \( L \) can be driven even lower to one in (12). Unfortunately, this enticing conclusion is not valid, because a large \( L \) is assumed in the derivations of [28].

A noisy average-case analysis has been carried out by [17]. Compared to the maximum-likelihood decoder in Table I, [17] applies a Bayesian detector to a combinatorial number derivations of [25].

\[
M \geq \kappa M \log \left( \frac{1}{\sigma_w^2} \right) \geq \log \left( K(N - K) \right) \tag{13}
\]

where \( \kappa \) is a distribution parameter and treated as constant. Then, true support can be recovered with an arbitrarily small probability of error. Note that while large \( M \) does help, it does not remove the burden on \( L \) as \( L \) still needs to be larger than \( O(K \log (N/K)) \). Turning to the necessary condition, [17] proposes

\[
M \geq (1 - \epsilon - \delta) \frac{2\sigma_w^2}{\kappa} \log \left( \frac{N}{K} \right) \tag{14}
\]

is needed to ensure correct support recovery. Here, \( \epsilon, \delta \ll 1 \) are arbitrarily small constants. This result is remarkable in the sense that it is independent of \( L \) and remains valid as long as \( M \) is larger than a certain threshold. Unfortunately, there exist a gap between sufficient and necessary conditions in [17] as \( L \) should be large in (13) but can be small in (14). To the best of our knowledge no one has closed this gap up to now.

Summarizing the aforementioned performance limits, one can conclude that results are rather inconclusive for small \( L \) set aside the noiseless worst-case scenario.

\[\text{C. Performance Limits of Thresholding / Maximum Correlation Decoders}\]

This decoder has two versions that are generally referred to as thresholding and maximum correlation. The two algorithms are equivalent in the sense that if one achieves perfect recovery with a particular tuning of its parameter (threshold or number of active users), the parameter for the second algorithm can be tuned to guarantee correct recovery as well [26]. Maximum correlation is reported in Table II while thresholding is presented in Table III.

### Table II. Maximum Correlation Decoder

- Repeat for \( j = 1, \ldots, N \)
  - Evaluate \( E_j = \| c_j \|_2^2 \)
- End Repeat
- Select the \( K \) largest \( E_j \) values and return their indices as the estimated support

### Table III. Thresholding Decoder

- Repeat for \( j = 1, \ldots, N \)
  - Evaluate \( E_j = \| c_j \|_2^2 \)
  - If \( E_j \geq \mu \) where \( \mu \) is a threshold add \( j \) to the estimated support.
  - Otherwise decide \( j \) was not active.
- End Repeat

While [27] has performed noisy worst-case and average-case analysis, noiseless average-case analysis has been carried out by [28]. For the sake of completeness, the performance of SMV thresholding has been investigated by [26].

Let \( S \) be the true support. Then nonzero elements of \( X \) are modelled as follows [28]:

\[
X^S := \Sigma \Phi \tag{15}
\]

where \( \Sigma \) is a diagonal deterministic matrix with positive diagonals \( \sigma_i^2, \ i = 1, \ldots, K \) and \( \Phi \) is a random matrix whose columns are selected as normalized Gaussian vectors also known as Gaussian on a unit sphere. Also define

\[
R = \max_{i,j=1,\ldots,N, i\neq j} | c_i^T c_j | \tag{16}
\]

Here, \( R \) represent the eigenvalue spread or condition number of \( \Sigma \), and \( \mu \) is referred to as coherence. According to [28], given

\[
\mu < \frac{1}{R \sqrt{K}} \tag{17}
\]

and with \( \theta := \mu R \sqrt{K} < 1 \), the probability of support recovery error is upper bounded by

\[
N \exp \left( -M (\theta^{-2} - \log(\theta^{-2}) - 1)/2 \right) \tag{18}
\]

Let us interpret this result. Note that \( \mu \) is a decreasing function of \( L \). If elements of \( C \) are chosen IID zero-mean as is the case in our setup, \( \mu \) converges to zero almost surely as \( L \to \infty \) by the strong law of large numbers (SLLN). Therefore, the bound in (17) demands that \( L \) be large. Provided \( L \) is large enough then probability of error can be driven to zero by increasing \( M \). Given our interest in low \( L \) and large \( M \) regime, this result provides mixed conclusions. It verifies that increasing \( M \) indefinitely can be helpful to thresholding but it also demands that \( L \) is not very small. Unfortunately, this is only a sufficient condition and does not state if recovery is still possible when \( L \) falls below the given threshold.

Focusing on the noisy worst-case setup, [27] suggests that following condition is sufficient for perfect support recovery:

\[
\| C^S \|_2 \leq \| C^S \|_2 \leq \min_{i \in S} \lambda_2^i - \max_{i \in S} \lambda_2^i (2K - 1) \mu, \tag{19}
\]

where

\[
\lambda_2^i := \| X^i \|_2
\]

and \( \mu \) is defined similar to (16). While this result seems difficult to interpret, indeed it offers a simple explanation. To see this, set \( W = 0 \) and check that (19) basically asks for

\[
\mu < \frac{1}{2K - 1} \left( \frac{\min_{i \in S} \lambda_2^i}{\max_{i \in S} \lambda_2^i} \right) \tag{20}
\]
Comparing (20) with (17), one realizes that they are almost the same. One main difference is that $R$ which was the statistic for random $X$ that determined spread of non-zero $X$ entries in the average-case in (17) is replaced by the true spread in the worst-case analysis in (20). However, there exist a major difference as $M$ does appear in the average-case analysis but it is not present in (20).

Finally, the noisy average-case is investigated. It should be noted that (27) treats $\Phi$ as having IID standard Gaussian entries which is different from the Gaussian on a unit sphere assumed by (28). According to (27) the following condition is sufficient for perfect recovery:

$$\|C^T S W\|_{2,\infty} \leq C(M) \left( \min_{i \in S} \sigma_i^2 - \max_{i \in S} \sigma_i^2 (\mu \sqrt{K}) \right),$$  \hspace{1cm} (21)

Here, $C(M)$ is an increasing function of $M$. If (21) holds then probability of support recovery error is upper bounded by $K \exp(-M \gamma^2 / \pi)$ where $\gamma$ is a positive constant. By inspection, one can find a similar pattern between (21) and (17). Note that the right hand side parentheses in (21) should be positive, which translates to (17). However, one can observe that noise is easily tackled by $M$. As noise increases, left hand side in (21) will increase. But one can compensate for noise by increasing $M$ and thus $C(M)$ on the right hand side.

To summarize, the reported results does not confirm if one can correctly recover support with small $L$. However, the worst-case noiseless scenario suggests $L \geq K$ is needed even for the optimal decoder. As a final note, it should be mentioned that thresholding performance can equal that of the combinatorial optimal decoder! For example, (29) has proven that in the average-case analysis and for low SNR, thresholding is the optimum detector for both complete and partial support recovery criteria.

IV. INTUITIVE OBSERVATIONS

Two observations are presented here which form the foundation for the opportunistic thresholding algorithm that will be proposed in the next section. They both point to the same direction. Specifically, they suggest that if the constant sensing matrix $C$ in (5) is replaced by IID sensing matrices $C_m$ that is

$$y_m = C_m x + w_m, \quad m = 1, \ldots, M$$  \hspace{1cm} (22)

then support recovery performance can be greatly improved. First, these two observations are presented. Then, fundamental limits of a constant sensing matrix versus IID sensing matrices are compared.

A. Observation I

Suppose system model is given by (22) where $C_m$’s, which are sensing matrices, are IID with independent Gaussian entries. Furthermore, suppose that $X := [x_1 \ | \ x_2 \ | \ \cdots \ | \ x_M]$ enjoys a common sparse support, i.e., many rows are identically zero, but its nonzero entries assume values that are IID Gaussian and independent from $C_m$’s. Then the trivial pursuit proposed by (18) can recover the support of $X$ even with $L = 1$ provided that $M$ grows large. Note that trivial pursuit is nothing but the low-complexity thresholding algorithm applied to the independent $C_m$ case. This great result is remarkable in the sense that with $L = 1$ it is impossible to estimate the nonzero entries in $X$ even if one knew the true support. Yet, $L = 1$ is sufficient to recover the common support for large $M$. Finally, it should be noted that the proposed decoder is not optimal and indeed very simple. Yet, it achieves such an impressive performance.

B. Observation II

Let us consider the model (5) but assume that channel gain matrix $H$ is known. Furthermore, choose $L = 1$ which means that there is no pseudo-random codes $c_n$ but just a scalar $c_n = \pm 1$. Without loss of generality, assume all $c_n = +1$. Subsequently, the $C$ matrix becomes a row vector $c_T := [1 \ 1 \ \cdots \ 1]^T \in \mathbb{R}^{1 \times n}$ and measurements per antenna become a scalar. If we stack the scalar measurements per antenna in a row vector $y_T := [y_1, y_2, \ldots, y_M]^T$, we obtain:

$$y_T = c^T X + w_T, \quad y_m = c^T x_m + w_m, \quad \forall m = 1, \ldots, M$$

Now, if one moves the known channel gains from the unknown $\{x_m\}$’s to the $c^T$ vector, the following equivalent equation ensues.

$$y = H \tilde{x} + w$$  \hspace{1cm} (23)

where we have defined:

$$\tilde{x} := [q_1 \sqrt{T_1} \mid q_2 \sqrt{T_2} \mid \cdots \mid q_N \sqrt{T_N}]^T \in \mathbb{R}^{N \times 1}$$

Note that with $L = 1$ and known $H$, $y$ becomes of size $M$ and the vector $\tilde{x}$ which is of size $N$ becomes $K$-sparse. Given that entries of $H$ are independent standard Gaussians, one can apply the known results in SMV case to conclude that if $M$ grows larger than a certain threshold then perfect recovery is possible using either the optimal decoder (21) or even a simple thresholding algorithm (10). Indeed, $H$ can be seen as a natural code assigned to the different users by the environment. Therefore, if we knew $H$, we could have used it to distinguish different users and there was no need for codes of length $L$. To signify the relationship between observation II and IID sensing matrices $C_m$’s, note that according to (23), the $m$’th antenna receives $y_m = h^m \tilde{x} + w_m$ where $h^m$ represents the $m$’th row of $H$. For a given $m$, $h^m$ can be thought of as of $C_m$ and then we can see that $C_m$’s are IID and that is the reason we obtain such a good performance with small $L$ and large $M$.

C. PERFORMANCE COMPARISON OF FIXED versus INDEPENDENT SENSING MATRICES

Selecting probability of correct support recovery as their criterion of choice, (19) proved that measurement inequalities can be improved for IID sensing matrices compared to a fixed sensing matrix. Before replicating their final results here, it should be mentioned that if MMSE is selected as criterion, the performance of two cases will be almost the same (30). In particular, (19) suggests that sufficient conditions for vanishing probability of error in support recovery are

$$L > K \left( 1 + \frac{1}{M f (\text{SNR}_{\text{min}})} \right)$$  \hspace{1cm} (24)
for the linear sparsity regime, i.e., \( \lim K/N = \beta > 0 \) and
\[
L > K \left( 1 + \frac{1}{Mf(\text{SNR}_{\text{min}})} \log \left( \frac{N}{K} \right) \right)
\]
(25)
for the sublinear sparsity regime, i.e., \( \lim K/N = 0 \). Note that \( f(.) \) is an increasing function of its argument which is the minimum SNR. \( \text{SNR}_{\text{min}} \) is defined as the ratio of the smallest nonzero entry of \( X \) squared divided by noise variance. Furthermore, a necessary condition to ensure vanishing probability of error in support recovery is given by
\[
L > \frac{2K \log \left( \frac{N}{K} \right) - 2 \log(2)}{M \log(1 + K \text{SNR}_{\text{min}})}
\]
(26)
Note that again there exists a nontrivial gap between necessary and sufficient conditions. Specifically, one can let \( M \) go to infinity in (26) and obtain an \( L \) as small as \( L = 1 \) which is desired. However, the sufficient condition demands that \( L > K \) even for very large \( M \). Given the result of trivial pursuit presented in [13], we conjecture that the necessary condition (26) is also sufficient because the optimal decoder is guaranteed to outperform trivial pursuit and hence should be able to achieve trivial pursuit performance.

As [19] carries out a noisy worst-case analysis, its results are comparable to (25). Comparison of (24), (25) with that of (11), one observes that while for a constant sensing matrix, \( M \) helps as long as \( M \leq K \) and still we need \( L > \mathcal{O}(\log N) \) (c.f. (11)), for independent sensing matrices \( M \) can greatly reduce the burden on \( L \). That is, \( L \) can be lowered to as much as its noiseless bound \( L > K \) by increasing \( M \). Given this proven result for independent sensing matrices, and Observations I and II, it can be deduced that independent sensing matrices ensure support recovery with small \( L \) and large \( M \). Such a result has not been obtained for the case of fixed sensing matrix. Armed with these observations, we move on to introduce our Opportunistic Thresholding Detector (OTD) which offers improved performance compared to the model in (5).

V. OPPORTUNISTIC THRESHOLDING DETECTOR

Previous section advised to use independent sensing matrices across antennas instead of a fixed sensing matrix, as the former has proven performance guarantees for small \( L \) and large \( M \) regime. Given a fixed code matrix \( C \), active user will transmit its corresponding code and all antennas will receive the same code multiplied by unknown channel gains. Therefore, code matrix is fixed across antennas. However, channel gains are independent across antennas and users. Therefore, we can de-correlate the code matrix \( C \) by multiplying it with the channel gain matrix \( H \). Unfortunately, we do not know \( H \). So, we propose the following: We have \( T \) random access slots in a coherence interval. In the first random access slot, we use the model (5) to detect the active user set. Once the active user set is determined, we apply a reduced dimension least-squares (LS) to estimate the corresponding channel gains. Then, we move the channel gains that are estimated into the sensing matrix and treat them as known. This way, we partially de-correlate the sensing matrices. As we proceed with more random access slots, more users become active and hence their channels will be estimated and known. Therefore, our sensing matrices become more and more de-correlated and we move from the fixed sensing matrix setup at the first random access slot to completely uncorrelated sensing matrices at the end of coherence interval. Note that user terminals’ speeds should be lower than a certain threshold to ensure sufficiently long coherence interval which allows for complete de-correlation. However, if the coherence interval is not long enough partial de-correlation is still achieved. Before proceeding any further, we present the mathematical formulation for the proposed OTD.

Suppose we are at the first random access slot, and we do not know any of the channel gains. Then, once the measurement matrix \( Y \) is observed, we perform thresholding on model (5) as follows. For \( n = 1, 2, \ldots, N \) define the decision statistic \( \theta_n \):
\[
\theta_n = \frac{1}{M} \| c_n^T Y \|_2^2 = \frac{1}{M} \sum_{m=1}^{M} (c_n^T y_m)^2
\]
(27)
Then, select the \( K \) largest \( \theta_n \) values and set their indices to be the support estimate. Let us refer to this set as \( \hat{S} \). Then, apply LS to the over determined problem obtained from keeping only the indices corresponding to \( \hat{S} \):
\[
\hat{X}^\hat{S} := (C_S^T C_S)^{-1} C_S^T Y
\]
(29)
where we can de-correlate the code matrix \( C \) by multiplying it with the channel gain matrix \( \hat{H} \). Unfortunately, we do not know \( \hat{H} \). So, we propose the following: We have \( T \) random access slots in a coherence interval. In the first random access slot, we use the model (5) to detect the active user set. Once the active user set is determined, we apply a reduced dimension least-squares (LS) to estimate the corresponding channel gains. Then, we move the channel gains that are estimated into the sensing matrix and treat them as known. This way, we partially de-correlate the sensing matrices. As we proceed with more random access slots, more users become active and hence their channels will be estimated and known. Therefore, our sensing matrices become more and more de-correlated and we move from the fixed sensing matrix setup at the first random access slot to completely uncorrelated sensing matrices at the end of coherence interval. Note that user terminals’ speeds should be lower than a certain threshold to ensure sufficiently long coherence interval which allows for complete de-correlation. However, if the coherence interval is not long enough partial de-correlation is still achieved. Before proceeding any further, we present the mathematical formulation for the proposed OTD.

Suppose we are at the first random access slot, and we do not know any of the channel gains. Then, once the measurement matrix \( Y \) is observed, we perform thresholding on model (5) as follows. For \( n = 1, 2, \ldots, N \) define the decision statistic \( \theta_n \):
\[
\theta_n = \frac{1}{M} \| c_n^T Y \|_2^2 = \frac{1}{M} \sum_{m=1}^{M} (c_n^T y_m)^2
\]
(27)
Then, select the \( K \) largest \( \theta_n \) values and set their indices to be the support estimate. Let us refer to this set as \( \hat{S} \). Then, apply LS to the over determined problem obtained from keeping only the indices corresponding to \( \hat{S} \):
\[
\hat{X}^\hat{S} := (C_S^T C_S)^{-1} C_S^T Y
\]
(29)
Once this estimate is obtained, one notes that from (3), (4), \( X_{i,j} = q_i \sqrt{P_i} h_{j,i} \). Given that \( q_i = 1 \) for active users, and assuming known \( P_i \)'s at the BS, one can estimate the channels between user \( n \) and BS as
\[
\hat{h}_{n} := \frac{(X^n)^T}{\sqrt{P_n}}
\]
(30)
Given such an initial random access slot at the beginning of each coherence interval, we move on to define OTD for any subsequent random access slots. Suppose, we are at an arbitrary random access slot greater than one. Let \( \Lambda \) represent the index set of users that have been active at least once in each coherence interval. Then, define the following parameters:
\[
\hat{H}_{i,j} := \begin{cases} H_{i,j}, & j \in \Lambda \\ 1, & \text{O.W.} \end{cases}
\]
(31)
and
\[
\hat{H}_{i,j} := \begin{cases} 1, & j \in \Lambda \\ H_{i,j}, & \text{O.W.} \end{cases}
\]
(32)
Upon defining \( \forall m = 1, \ldots, M \)
\[
x_m := \begin{bmatrix} q_1 \hat{H}_{m,1} \sqrt{P_1} & q_2 \hat{H}_{m,2} \sqrt{P_2} & \cdots & q_N \hat{H}_{m,N} \sqrt{P_N} \end{bmatrix}
\]
(33)
one arrives at the following set of equations \( \forall m = 1, \ldots, M \) which are equivalent to (5):
\[
y_m = A_m x_m + w_m,
\]
\[
A_m := \begin{bmatrix} \hat{H}_{m,1} c_1 & \hat{H}_{m,2} c_2 & \cdots & \hat{H}_{m,N} c_N \end{bmatrix}
\]
(34)
Note that compared to the fixed sensing matrix in (5), we now have a set of sensing matrices $A_m$ which are partially uncorrelated. Next, we apply thresholding to the model (34). Let us define the decision statistics:

$$\theta_n = \frac{1}{M} \sum_{m=1}^{M} \left( \hat{H}_{m,n} c_n^T y_m \right)^2$$

(35)

Then, we select the $K$ largest $\theta_n$'s and choose their indices to represent the true support. We refer to this set as $\hat{S}$ and use the LS estimate (29) and channel estimate in (30) to obtain channel estimates of the active users. Note that as we progress over random access slots, more users become active and we know more users channels. Hence, the sensing matrices $A_m$'s become more and more de-correlated. Thus, we expect to successively obtain better support recovery performance as we move towards the end of a coherence interval. Before proceeding any further, several remarks are in order.

**Remark 1.** We assume known powers in OTD. This assumption is reasonable and practical because the only factor that changes over time in $P_n$ comes from the path-loss which varies slowly over time. Path-loss variations occur on a much longer scale compared to small scale fading, hence $P_n$'s can be assumed constant over many coherence intervals and thus can be easily estimated beforehand.

**Remark 2.** OTD is opportunistic in the sense that it does not schedule users or force them to send data/training/code for channel estimation. Instead, it relies on the information it obtains on-demand from the active users which are random and transmit at will. Indeed, OTD exploits the measurements and random activity patterns to estimate active users’ channels from the data that is made available to it. Amazingly, OTD does not know which users are active in advance, yet it proceeds to estimate their channels. Then, the partial information OTD obtains in a random access slot is exploited to improve its performance in future random access slots.

**Remark 3.** OTD operates independently in different coherence intervals. The independent operation arises because we assume a block-fading channel model where the channel changes independently across blocks. Consequently, all channel state information (CSI) obtained in a coherence interval becomes obsolete for the future coherence intervals. Therefore, the same process should be repeated for future coherence intervals.

**Remark 4.** One might suggest that instead of moving channels of users with known CSI into the sensing matrix, separate them completely from the random access process and decode them by applying a e.g., zero-forcing (ZF) or maximum ratio combining (MRC) to the BS measurements. There exist three limitations to this approach. Firstly, the channel estimates obtained on a single random access slot may not be very accurate. As we progress across random access slots and the same user becomes active multiple times, we improve our channel estimate. This process can not be carried out if users with known CSI are separated. Secondly, if we incur an error in support estimation, it leads to an erroneous channel estimate. While we can recover from such errors by averaging CSI at different random access slots that a particular user becomes active, as done by OTD, separating users with known CSI leads to error propagation. Thirdly, if $N$ is larger than $M$, ZF is not practical as inverse of $H^T H$ either does not exist or is on the order of $M$ which is too complex to compute. Besides, these users transmit only sporadically and therefore our efforts in implementing ZF can be completely wasted. On the other hand, MRC is what we are approximately doing in our proposed OTD with a modification of an added pseudo-random code.

**Remark 5.** Note that if CSI for all users is available, then vector $x_m$ becomes element-wise positive. Hence, there will be no need to square the correlations in (35). The following decision statistic will suffice:

$$\theta_n = \frac{1}{M} \sum_{m=1}^{M} \hat{H}_{m,n} c_n^T y_m$$

**Remark 6.** There exists a notable difference between the model to which [18] applies trivial pursuit (TP) and our model in (34). Note that TP is applied to recover support for $x_m$ whose support is $K$, same as our model, but whose non-zero entries are IID Gaussian. Therefore, for a fixed $M, L, N, K$, [18] has $LM$ measurements and $NM$ unknowns. Note that number of unknowns always exceeds that of equations and therefore LS is not applicable. Indeed, first support should be recovered via TP and then provided that $L > K$, one can apply LS to the reduced dimension problem to recover the nonzero values. On the other hand, for our problem, in the extreme case, where all users’ CSI is known and moved into the sensing matrix, [33] suggests $\tilde{x}_1 = \tilde{x}_2 = \cdots = \tilde{x}_M$ and hence we get $N$ unknowns and $LM$ equations. For $LM > N$, this model can be solved via LS but note that its complexity will be of order $O(N^3)$ which is significant. OTD, on the other hand, offers a complexity that is linear in both $M, N$.

Finally, we offer an intuition on why OTD works. Note that rigorous performance analysis for OTD is carried out in [20]. Through analysis, we realized that OTD performs as illustrated in Fig. 3. Initially, when no CSI is available, distribution of decision statistics $\theta_n$ is plotted in the top plot. Note that the intersection area in the middle is where errors occur. On the other hand, OTD moves the mean of $\theta_n$ for active users to a higher value as in the bottom subplot of Fig. 3 while slightly increasing the variance. However, the
positive effect of higher mean outdoes the negative effect of small increase in variance. As a result, performance improves. Note that this phenomenon occurs only for those users who have been active at least once in this coherence interval, and decision statistics for inactive users will remain as in top subplot of Fig. 3. An alternative viewpoint on OTD emphasizes its resemblance to decision feedback equalizer (DFE). In fact, OTD first performs support recovery. Then, it uses support information to estimate active users channels. Afterwards, it applies the newly acquired CSI to improve its support recovery performance in the next random access slot. In a sense, OTD alternates between support detection and channel estimation which is a characteristic of DFE-type methods.

VI. NUMERICAL RESULTS

Our evaluation of OTD consists of mathematical analysis and numerical simulations. Space limitations prevented us from reporting the mathematical analysis in here, so full proofs and derivations are presented in the accompanying paper [20]. To give a flavor of [20], note that we can neither use law of large numbers utilized in TP in [13], due to lack of independence, nor the existing results on fixed sensing matrices such as [28] and [27]. Our model falls somewhere in between these two. While our sensing matrices are uncorrelated in the limit, the decision statistics summands \( \left( \hat{H}_{m,n} e_n^T y_m \right)^2 \) are neither independent nor uncorrelated. Therefore, OTD demands a performance analysis of its own. In [20], we decompose \( \theta_j \)'s in (35) into conditionally independent components by applying Martingale theory. Then, we prove that each component is sub-exponential or sub-Gaussian and can be bounded using Bernstein or Hoeffding inequalities. Finally, per term results are fused to obtain measurement inequalities that define the relationship needed between \( M, L, N, K \) that ensures perfect recovery. Unfortunately, mathematical analysis demands \( L \) to be large, while our main claim regarding OTD is in the low \( L \) regime. To deal with this dilemma, we have performed extensive simulations to demonstrate the effectiveness of OTD in the small \( L \) regime. These results are reported here.

A. Perfect CSI

To isolate the effect of imperfect CSI due to noise or errors in support recovery from the effectiveness of the OTD in general, we divide the numerical results into two subsections. In the first part, we assume that active users’ channels have been estimated error-free. That is, \( H_{i,j} = \hat{H}_{i,j} \) for \( j \in \Lambda \). Therefore, the effects of noise and support recovery error in estimating \( h_n \)'s is overlooked. However, both noise and support recovery errors do have a direct impact on active users detection performance even if we set aside errors in CSI estimation. We adopt partial support recovery error defined below as Error Type I:

\[
P(\text{Error Type I}) := \frac{|S - \hat{S}|}{K}
\]

where \( S \) represents the true support and \( \hat{S} \) represents the estimated support. Note that error type I is a sort of soft error criteria in the sense that it partially penalizes errors in support.

Each single miss in support recovery contributes a factor \( 1/K \) to error type I. This error is sometimes referred to as distortion [29]. Error type II is more strict and will equal 0 only if all the support is correctly recovered. That is \( \hat{S} = S \). It is equal to one otherwise meaning that all support errors are penalized by the same weight. The reason behind selection of error type II is that it is easier to analyze mathematically and it corresponds to the case of distortion-free support recovery. Therefore, it has been widely exploited in sparse recovery literature. See e.g., [21], [27], [28], and so on. Extensive simulations suggests that both error types offer the same performance patterns. Due to space limitations, we report error type II numerical results only.

In the first simulation, we fix \( N = 150, K = 15, \) and \( L = 40, \) and we vary \( M \) from one to 1000. Then, we plot the error type II for various percentage of users with known CSI. Note that when known CSI ratio equals 0, we are basically using ordinary thresholding, and when known CSI ratio equals 1, it amounts to knowing all users channel gains and thus we can apply MRC instead of OTD. With complete CSI, OTD and MRC are basically doing the same operation with the exception that OTD has a code matrix \( C \) also while MRC exploits only \( H \). In between, lies the true potential of OTD, and the two cases of no CSI and complete CSI correspond to special cases of OTD in the two extremes. First, we consider the noiseless setting with equal power for all users. That is, \( P_1 = P_2 = \cdots = P_N = 1 \) and \( \sigma_n^2 = 0 \). Fig. 4 illustrates probability of error type II. Upon observing this figure, one realizes no CSI performs better than low ratios of known CSI which are 20 percent and 50 percent. In the limit of large \( M \) they perform similarly. However, this pattern reverses when known CSI ratio is large, e.g., equals 90 percent. Then, we observe a considerable gain in OTD versus the algorithm that does not exploit active users’ CSI. A comparison between no CSI and known CSI ratio of 90 percent reveals that OTD can successfully lower the error probability by increasing \( M \) for a small fixed \( L \). However, OTD needs a large ratio of users with known CSI to be effective. This was expected as OTD improves slot by slot as it obtains more and more CSI.
Second simulation deals with noise and unequal users’ powers. It is assumed that minimum SNR=0 dB and power spread across users is 7 dB. It should be noted that in order to have a fair comparison for various $L$, the code matrix $C$ entries variance should be scaled as $1/\sqrt{L}$. In fact, we spread the same power across a larger number of measurements as $L$ increases. Therefore, noise variance should be scaled similarly if a minimum SNR of 0 dB is desired. Fig. 5 plots probability of error type II, where a different pattern emerges compared to Fig. 4. Specifically, OTD with no CSI performs the worst and OTD improves as we add CSI. Curiously, even low and moderate amounts of CSI such as 20 percent and 50 percent do outperform the no CSI setup. Armed with the knowledge of mathematical derivations in [20], we know that both power spread and minimum SNR do appear in the measurement inequality for OTD. This can explain the difference between noiseless and noisy setups. Again, these simulations are testament to the fact that by increasing $M$, OTD will outperform the ordinary thresholding for a fixed small $L$.

In the last simulation of this subsection, we fix $N = 150$, $K = 15$, and $M = 1000$, and we let $L$ increase from one to 100. Fig. 6 plots probability of type II error versus $L$ in the noiseless equal power scenario. Subsequently, Fig. 7 illustrates the same error versus $L$ but in the noisy unequal power regime. Same parameters as before have been chosen. Note the significant gain that OTD with partial CSI brings about compared to the ordinary thresholding which amounts to OTD with no CSI. Again, the results are more significant in the noisy unequal power setup. For example, Fig. 7 suggests that to achieve a probability of error equal $10^{-2}$, OTD with 90 percent CSI requires 30 percent less measurements ($L$) than ordinary thresholding. In the noiseless case, we observe that OTD improvement is more profound in the low $L$ regime. For large values of $L$, OTD performs similarly to ordinary thresholding. Note that OTD was designed mainly with the small $L$ regime in mind.

**B. Imperfect CSI**

In this subsection, OTD is examined with all possible sources of imperfection included. To be specific, we plot OTD performance in a coherence interval. It is assumed that the coherence interval consists of 100 random access slots. While 100 might be too large for smart vehicles and that sort of applications, it is well-suited to pedestrian-based IoT or even immobile wireless setups such as smart homes. Simulation parameters are selected as $N = 150$, $K = 15$, $L = 40$, $M = 1000$ and the noiseless equal-power setup is considered first. To obtain the presented results, 10000 Monte Carlo runs are performed. To remind you of our setup, at each random access slot $K$ users become active independent of previous and future slots. The aim is to detect the active users. While ordinary thresholding relies on the different users codes to detect active users, OTD exploits a combination of codes and available CSI opportunistically obtained during the previous random access slots. Fig. 8 demonstrates the error type II where two curves are plotted. The first is thresholding without channel gain knowledge which is the ordinary thresholding presented in [14]. The second curve corresponds to OTD. Note
that OTD begins in a similar fashion as ordinary thresholding but it successively applies the channel knowledge it opportunistically obtains to improve on each slot. Note that ordinary thresholding performance is almost constant across various random access slots as expected. However, OTD improves as time goes on. Curiously, at first OTD demonstrates a higher error than ordinary thresholding. But after the initial slots have passed it achieves a performance considerably better than ordinary thresholding. The full CSI error turned out to be zero for 10000 Monte Carlo runs for the given parameters. Fig. 10 is a demonstration of percentage of users who have been active at least once in this coherence interval up to that particular random access slot and hence their CSI is available to OTD. This figure can be thought of as the cumulative distribution function (CDF) of active users over random access slots.

The next set of simulations focus on the major limiting factor in OTD which is the error propagation in erroneous channel estimates. Note that we first detect the support and then we apply LS to estimate channel gains. If active elements are missing from the detected support, then corresponding channel estimates will incur large errors and these errors propagate into future access slots as OTD exploits them for support detection. Let us consider the equal power noisy setting with minimum SNR=0 dB. All other parameters are same as before. First, we assume true support is known in channel estimation phase. Fig. 11 plots error versus random access slot. Note the significant improvement in OTD compared to ordinary thresholding. Specifically, around slot 30, which corresponds to 95 percent known CSI ratio, we get an order of magnitude improvement in performance compared to ordinary thresholding. Dropping the perfect support assumption, we arrive at Fig. 11 which demonstrates that OTD performance suffers somehow as a consequence of support recovery errors. However, OTD is
still found to significantly outperform ordinary thresholding as random access slots increases. Finally, it should be noted that when full CSI is assumed 10000 Monte Carlo simulations yielded zero errors and hence this curve does not appear in any of the figures.

VII. CONCLUSION

An integration of on-off random access channel and massive MIMO was investigated, where it is known that active users detection amounts to support recovery for a multiple measurement vector (MMV) problem with a fixed sensing matrix and common sparse support. Given that we have physical limitations on the channel coherence interval dictated by the environment, we focused on the scenario where number of measurements is limited but number of antennas can be arbitrarily large. Our first contribution was to look at upper and lower performance limits of MMV with common support with arbitrarily large. Our first contribution was to look at upper and lower performance limits of MMV with common support with the aforementioned observation in mind. Then, we presented a model exploiting independent sensing matrices which allow same low-complexity properties. Intuition behind OTD was to correlate sensing matrices to gain an improved active user detection performance. OTD is derived with certain modifications to the ordinary thresholding algorithm and enjoys the same low-complexity properties. Intuition behind OTD was also presented. While mathematical performance was relegated to an accompanying paper, extensive numerical simulations determined the benefits and drawbacks of OTD. It was finally demonstrated that for IoT applications with a limited speed such as pedestrian or immobile scenarios, proposed OTD is practical and can significantly improve upon the ordinary thresholding algorithm available in the literature.

REFERENCES

[1] U. Ewaldsson. (2016, Nov) Ericsson mobility report. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2016/emr-november-2016-live-streaming-joins-social-media.pdf
[2] T. L. Marzetta, “How much training is required for multiuser mimo?” in Proc. of Asilomar Conf. on Signals, Systems, and Computers, Pacific Grove, CA, USA, Oct/Nov 2006, pp. 290–294.
[3] “—, “Noncooperative cellular wireless with unlimited number of base station antennas,” IEEE Trans. on Wireless Communications, vol. 9, no. 11, pp. 3590–3600, Nov 2010.
[4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive mimo for next generation wireless systems,” IEEE Communications Magazine, vol. 52, no. 2, pp. 186–195, Feb 2014.
[5] S. Buzzi, C. L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, “A survey of energy-efficient techniques for 5G networks and challenges ahead,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 697–709, April 2016.
[6] E. d. Carvalho, E. Bjornson, J. H. Sorensen, P. Popovski, and E. G. Larsson, “Random access protocols for massive mimo,” IEEE Communications Magazine, vol. 55, no. 5, pp. 216–222, May 2017.
[7] J. H. Sorensen, E. de Carvalho, and P. Popovski, “Massive mimo for crowd scenarios: A solution based on random access,” in 2014 IEEE Globecom Workshops (GC Wkshps), Austin, TX, USA, Dec 2014, pp. 352–357.
[8] E. de Carvalho, E. Bjornson, J. H. Sorensen, E. G. Larsson, and P. Popovski, “Random pilot and data access in massive mimo for machine-type communications,” IEEE Transactions on Wireless Communications, vol. 16, no. 12, pp. 7703–7717, Dec 2017.
[9] E. Bjornson, E. de Carvalho, E. G. Larsson, and P. Popovski, “Random access protocol for massive mimo: Strongest-user collision resolution (sucr),” in 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[10] A. K. Fletcher, S. Rangan, and V. K. Goyal, “On-off random access channels: A compressed sensing framework,” arXiv, vol. abs/0903.1022, Mar 2009. [Online]. Available: http://arxiv.org/abs/0903.1022
[11] C. Bockelmann, H. F. Schepker, and A. Dekorsy, “Compressive sensing based multi-user detection for machine-to-machine communication,” Transactions on Emerging Telecommunications Technologies, vol. 24, no. 4, pp. 389–400, Jun 2013.
[12] 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.
[13] H. Zhu and G. B. Giannakis, “Exploiting sparse user activity in multiuser detection,” IEEE Transactions on Communications, vol. 59, no. 2, pp. 454–465, February 2011.
[14] Y. D. Beyene, R. Juntti, and K. Ruttik, “Random access scheme for sporadic users in 5G,” IEEE Transactions on Wireless Communications, vol. 16, no. 3, pp. 1823–1833, March 2017.
[15] M. Ghayani and A. Shahzadi, “Multiple random access for massive mimo framework: A unified compressive sensing based approach,” Computers and Electrical Engineering, vol. 64, pp. 524 – 536, Nov 2017.
[16] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, Fundamentals of Massive MIMO. Cambridge University Press, 2016.
[17] G. Tang and A. Nehorai, “Performance analysis for sparse support recovery,” IEEE Transactions on Information Theory, vol. 56, no. 3, March 2010.
[18] D. Baron, M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, “Distributed compressive sensing,” arXiv, vol. abs/0901.3403, Jan 2009. [Online]. Available: http://arxiv.org/abs/0901.3403
[19] S. Park, N. Y. Yu, and H. Lee, “An information-theoretic study for joint sparsity pattern recovery with different sensing matrices,” IEEE Transactions on Information Theory, vol. 63, no. 9, pp. 5559–5571, Sept 2017.
[20] M. N. Taheri, F. Karray, and S. Farahmand, “An asymptotic analysis of opportunistic thresholding detector via martingale theory,” IEEE Transactions on Signal Processing (to be submitted), Dec. 2018.
[21] M. J. Wainwright, “Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting,” IEEE Transactions on Information Theory, vol. 55, no. 12, pp. 5728–5741, Dec 2009.
[22] J. Chen and X. Huo, “Theoretical results on sparse representations of multiple-measurement vectors,” IEEE Transactions on Signal Processing, vol. 54, no. 12, pp. 4634–4643, Dec 2006.
[23] M. E. Davies and Y. C. Eldar, “Rank awareness in joint sparse recovery,” IEEE Transactions on Information Theory, vol. 58, no. 2, pp. 1135–1146, Feb 2012.
[24] M. Rudelson and R. Vershynin, “The least singular value of a random square matrix is $O(n^{1/2})$,” Comptes Rendus Mathematique, vol. 346, no. 15, pp. 893 – 896, Aug 2008.
[25] Y. Jin and B. D. Rao, “Support recovery of sparse signals in the presence of multiple measurements,” constructions of information and coding theory, vol. 59, no. 5, pp. 3139–3157, May 2013.
[26] A. K. Fletcher, S. Rangan, and V. K. Goyal, “Necessary and sufficient conditions for sparsity pattern recovery,” IEEE Transactions on Information Theory, vol. 55, no. 12, pp. 5758–5772, Dec 2009.
[27] R. Gribonval, H. Rauhut, K. Schnass, and P. Vanderheyden, “Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms,” Journal of Fourier Analysis and Applications, vol. 14, no. 5, pp. 655–687, Dec 2008.
[28] Y. C. Eldar and H. Rauhut, “Average case analysis of multichannel sparse recovery using convex relaxation,” IEEE Transactions on Information Theory, vol. 56, no. 1, pp. 505–519, Jan 2010.
[29] G. Reeves and M. Gaetp, “A note on optimal support recovery in compressed sensing,” 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, pp. 1576–1580, Nov 2009.
[30] J. Zhu, D. Baron, and F. Krzakala, “Performance limits for noisy multimeasurement vector problems,” IEEE Transactions on Signal Processing, vol. 65, no. 9, pp. 2444–2454, May 2017.