Many Access for Small Packets Based on Precoding and Sparsity-aware Recovery

Ronggui Xie[1], Huarui Yin[1], Xiaohui Chen[1], and Zhengdao Wang[2]

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China

[2]Department of Electrical and Computer Engineering, Iowa State University

Abstract

Modern mobile terminals produce massive small data packets. For these short-length packets, it is inefficient to follow the current multiple access schemes to allocate transmission resources due to heavy signaling overhead. We propose a non-orthogonal many-access scheme that is well suited for the future communication systems equipped with many receive antennas. The system is modeled as having a block-sparsity pattern with unknown sparsity level (i.e., unknown number of transmitted messages). Block precoding is employed at each single-antenna transmitter to enable the simultaneous transmissions of many users. The number of simultaneously served active users is allowed to be even more than the number of receive antennas. Sparsity-aware recovery is designed at the receiver for joint user detection and symbol demodulation. To reduce the effects of channel fading on signal recovery, normalized block orthogonal matching pursuit (BOMP) algorithm is introduced, and based on its approximate performance analysis, we develop interference cancellation based BOMP (ICBOMP) algorithm. The ICBOMP performs error correction and detection in each iteration of the normalized BOMP. Simulation results demonstrate the effectiveness of the proposed scheme in small packet services, as well as the advantages of ICBOMP in improving signal recovery accuracy and reducing computational cost.

Index Terms

small packet, non-orthogonal many access, precoding, block-sparsity, error correction and detection, order statistics

I. Introduction

Intelligent terminals such as smart phones and tablets are widely used nowadays, and their numbers are expected to grow in the near future. It is likely that many applications in these

intelligent terminals are simultaneously requesting to communicate with a same base station (BS). The resulting data packets to be delivered over a cellular network are usually very short, i.e., small packets [1], [2]. Mobile applications such as instant messaging are the main producers of small packets. In the services of small packet, the number of users simultaneously requesting services may even be comparable with the length of coding block, which makes the channel a many-access channel [3], [4]. On the other hand, the service requests of small packet are frequent and irregular [12]. Although many users may simultaneously have packets to send, they represent only a small percentage of the total users connected to the network.

In current systems, small packet services put great burden on the cellular network. One factor is due to the low efficiency of the transmitted packet. Even very short conversation packet is accompanied by regular overhead such as identity information. More importantly, for frequent and random transmission requests, when following the existing multiple access schemes that allocate orthogonal resources for different transmissions, signaling overhead related to scheduling these packets is very heavy. For example, even the scheduling-based schemes which are widely deployed in systems such as long term evolution (LTE) to ensure the quality of services, will produce massive overhead in dynamic resource programming. Future wireless communication systems, e.g., 5G network [11], are required to offer a lot of such connections, which draws much attention from academia and industry to propose flexible and efficient protocols, e.g., [3]–[10] and references therein.

In this paper, for the services of small packet, we propose a non-orthogonal many access scheme for uplink. The proposed scheme is based on precoding at the transmitters and sparsity-aware recovery at the receiver. The main motivation is to allow for a large number of users to transmit simultaneously, although each user may transmit only a small amount of data. Block precoding is employed to enable the simultaneous transmissions of many users. The sparsity-aware recovery is designed to detect the user activity and recover the transmitted data jointly. We say a user is *active* if the user is transmitting data. The user activity, including the number of active users and the identities of active users, is not known at the BS. Besides frame-level synchronization, no competition for resources or other coordination are required. This saves the signaling overhead related to allocating resources for transmission, and improves the resource utilization efficiency.

The contributions of our work are as follows:

1) *Block precoding and block-sparse system modeling:* We apply block precoding to short-length message at each transmitter, and by considering the user activities, we develop a block-sparse system model with unknown sparsity level. The established model takes full advantages of the block transmission behavior and is suitable for sparse recovery based algorithms.
2) *Block-sparse recovery algorithms:* Signal recovery includes user detection and the symbol demodulation in our scenario. To mitigate the influences of channel state on recovery, we modify the existing BOMP algorithm into a normalized BOMP algorithm. And then based on the normalized BOMP, we develop an interference cancellation (IC) based BOMP (ICBOMP) algorithm. The ICBOMP improves upon the normalized BOMP by taking advantage of the availability of error correction and detection coding. By the ICBOMP algorithm, we achieve better signal recovery accuracy and lower computational cost. The price is a slightly decreased data rate due to coding.
3) *Signal recovery analysis:* We approximately analyze the performance of signal recovery by the normalized BOMP algorithm. Considering the characteristic of the sparse recovery, order statistics [34] is used for the approximate analysis. Among many mutually-exclusive cases of successful user detection, we choose the most likely one to analyze its probability. The obtained probability is used to serve as the approximate lower bound for the probability of successful detection. Based on the successful detection, we perform analysis for symbol error rate (SER).

Thanks to the precoding operation and the designed sparsity-aware recovery, our scheme enables the system to serve many active users simultaneously. The number of active users can be even more than the number of antennas at the BS. When BS is equipped with a large-scale antenna array [13], [14], it can offer small packet services to a very large number of users. At the same time, as demonstrated by the sparse-recovery framework in [19], our scheme can also provide benefits in saving the identity information and reducing the decoding delay. Therefore, the proposed scheme is especially suitable for the future 5G system.

The most related works to our proposed scheme are sparsity-aware recovery based multi-user detection studied in papers such as [3]–[5], [15]–[23] and references therein. In most of the referred works, sparse transmissions have been considered for the detection task. Some of the works address the problem of user detection and the afterwards symbol demodulation. A

new notion of capacity related to many-access channel is introduced in [3], [4]. Full duplex is virtually realized in [5], where on and off slots are respectively assigned for transmitting to and receiving from neighbors, and finally sparse recovery is done at the receiver when one frame transmission completes. In [15], based on convex optimization, the authors introduce sparsity-aware detectors for the multi-user detection in code-division multiple access (CDMA) systems. The fact that sparse recovery algorithms such as OMP can provide better performance than single-user detection is demonstrated in [17]. Works in [19], [21], [23] show that, compared with some well known schemes, the CS based multiple access scheme can save the identity information overhead and reduce the decoding delay. A CS based multi-user detection scheme for machine-to-machine (M2M) communication can be found in [18]. Particularly, authors in [18] establish a block-sparse system model and adopt a novel coding scheme to improve the user detection accuracy.

Our work differs from the above referred works. All works in [3]–[5], [15]–[19], [21], [22] are about scenario of single-antenna receiver. Works in [16], [17], [19]–[23] establish sparsity model where nonzero coefficients to detect are just randomly located among all possible positions in a sparse vector, which neglects the fact that the transmission of a message could last for a few symbol slots, i.e., block transmission. Block transmission is considered in [5], [18]. BOMP algorithm is adopted in [18] for the data recovery and an activity-aware channel coding is used to improve the recovery performance, but coding scheme in it is not fully incorporated into the greedy algorithm to reduce the computational cost.

The rest of the paper is organized as follows. In Section II, the system model of block sparsity is given. In Section III, we introduce the normalized BOMP algorithm to recover the transmitted signals and analyze its performance. Improved algorithm version ICBOMP is proposed in Section IV. Section V presents the numerical simulation results that demonstrate the effectiveness of the proposed scheme. Some issues are discussed in Section VI. Finally, conclusions are presented in Section VII.

Notation: Vectors and matrices are denoted by boldface lowercase and uppercase letters, respectively. The 2-norm of a vector $\mathbf{v}$ is denoted as $\|\mathbf{v}\|_2$, and $\mathbf{v}^T$ denotes its transpose. The identity matrix of size $n \times n$ is denoted as $\mathbf{I}_n$. For a matrix $\mathbf{A}$, its conjugate transpose is denoted as $\mathbf{A}^H$. The (i,j)-th entry of a matrix $\mathbf{A}$ is denoted as $[\mathbf{A}]_{i,j}$. Operation $vec(\mathbf{A})$ denotes vectorization of $\mathbf{A}$ by column stacking. For a subset $I \subset [N] := \{1, 2, \cdots, N\}$ and

matrix $\mathbf{A} := [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_N]$ consisting of N sub-matrices (blocks), where each sub-matrix has an equal size, $\mathbf{A}_I$ stands for a sub-matrix of $\mathbf{A}$ whose block indices are in I; for a vector $\mathbf{v} := [\mathbf{v}_1^T, \mathbf{v}_2^T, \cdots, \mathbf{v}_N^T]^T$, $\mathbf{v}_I$ is similarly defined. Value $|I|$ stands for the cardinality of set I. Given two sets I_1 and I_2, $I_1 \backslash I_2 := I_1 \cap I_2^c$. For a real number r, $\lfloor r \rfloor$ and $\lceil r \rceil$ respectively stand for its floor and ceiling. Operators $\otimes$, Tr and $\mathbb{E}$ respectively stand for the Kronecker product, the trace of a matrix and the expectation of a random variable. Re stands for taking the real part from a complex number, and $\binom{m}{k}$ is the binomial coefficient for $m \geq k$. The real Gaussian PDF with mean μ and variance σ^2 is denoted as

$$f^{(\mathrm{N})}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1}$$

II. System Model

Consider an uplink system with N mobile users, each with a single antenna, and a BS with M antennas. When a terminal is admitted to the network, it becomes an *online* user. We assume that there are N_{a} *active* users, out of the total N online users, that simultaneously have data to transmit. It is not required N_{a} be known *a priori* or $N_{\mathrm{a}} < M$. Actually in practical systems, N_{a} is usually unknown and it is possible that $N_{\mathrm{a}} \geq M$.

We make the following additional assumptions on the system considered.

1) The channels are block-fading. They remain constant for a certain duration and then change independently.
2) The transmissions are in blocks and the users are synchronized at the block level. We assume that each frame of transmission consists of T symbols, which all fall within one channel coherent interval. Value T depends on the communication standard.
3) The antennas at the BS, as well as the antennas among users, are sufficiently apart to yield spatially independent channels.
4) The BS always has perfect channel state information (CSI) of online users.

Since the lengths of small packets are usually shorter than T, we extend their lengths by precoding to fully utilize the available resources. Let $\mathbf{s}_n \in \mathbb{C}^{d\times 1}$ denote the symbol vector to be transmitted by user n, with $d < T$, $n = 1, 2, \cdots, N$. User n applies a precoding to $\mathbf{s}_n$ to yield

$$\mathbf{x}_n = \mathbf{P}_n \mathbf{s}_n \tag{2}$$

where $\mathbf{P}_n$ is a complex precoding matrix of size $T \times d$. We assume that $\mathbf{P}_n$ is normalized such that each column has 2-norm equal to 1. The entries of $\mathbf{x}_n$ are transmitted in T successive time slots.[1] The received signals at the receiver within one frame can be written as

$$\mathbf{Y} = \sqrt{\rho_0} \sum_{n=1}^{N} \mathbf{h}_n \mathbf{x}_n^T + \mathbf{Z} = \sqrt{\rho_0} \sum_{n=1}^{N} \mathbf{h}_n \mathbf{s}_n^T \mathbf{P}_n^T + \mathbf{Z} \tag{3}$$

where ρ_0 is the signal to noise ratio (SNR) of the uplink, $\mathbf{Y}$ is the noisy measurement of size $M \times T$, $\mathbf{Z} \in \mathbb{C}^{M \times T}$ represents the additive Gaussian noise matrix of complex-valued, and $\mathbf{h}_n \in \mathbb{C}^{M \times 1}$ represents the complex channel response from user n to the BS. Without loss of generality, each element of $\mathbf{h}_n$, $\mathbf{Z}$ and nonzero $\mathbf{s}_n$ is assumed to have a zero mean and a unit variance. Using the linear algebra identity $vec(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})vec(\mathbf{B})$, we can rewrite the received signal as

$$vec(\mathbf{Y}) = \sqrt{\rho_0} \sum_{n=1}^{N} (\mathbf{P}_n \otimes \mathbf{h}_n)\mathbf{s}_n + vec(\mathbf{Z}). \tag{4}$$

Define $\mathbf{y} := vec(\mathbf{Y})$, $\mathbf{B}_n := \mathbf{P}_n \otimes \mathbf{h}_n$, $\mathbf{B} := [\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_N]$ and $\mathbf{s} := [\mathbf{s}_1^T, \mathbf{s}_2^T, \cdots, \mathbf{s}_N^T]^T$. Then (4) can be rewritten as

$$\mathbf{y} = \sqrt{\rho_0}\mathbf{B}\mathbf{s} + \mathbf{z}. \tag{5}$$

In the above formulation, we have assumed that all messages have an equal length d. We view d as the maximum length of the messages for all users within a frame. For the users whose message lengths are less than d, we assume their messages have been zero-padded to d before precoding. For those users that are not active, we view their transmitted symbols as all zeros. Throughout this paper, we say $\mathbf{s}_n$ is a block of $\mathbf{s}$ and $\mathbf{B}_n$ is a block of $\mathbf{B}$.

Since typically only a small percentage of the online users are active, nonzero signals are only located in a small fraction of blocks in $\mathbf{s}$ and all other blocks are zero. When $MT > Nd$, the receiver design is easy. Our consideration is limited to case $MT < Nd$, where (5) has the same form as an overloaded CDMA system. The equivalent spreading chip sequences in $\mathbf{B}$ no longer have constant amplitude and are generated by combining precoding codes and channel gains. When precoding matrix $\mathbf{P}_n$ is well designed, matrix $\mathbf{B}$ can meet the restricted isometry

[1] In fact, T can be generally regarded as the number of resource units in time, frequency and code domains. For example, T resource units can be composed of m subcarriers and $\lceil T/m \rceil$ successive time slots for each subcarrier.

property (RIP) requirement for measurement matrix in the CS literature [24], [25]. Therefore, from the viewpoint of signal recovery, (5) can be viewed as a block-sparsity CS model [27].

Although the information of active users is unknown at the BS, it is assumed that the BS knows the sparsity of transmission, so that it can perform sparsity-aware recovery. Let I be the set containing the unknown indices of active users, with $|I| = N_{\mathrm{a}} \leq N_{\mathrm{a\,max}}$. This means that the unknown number of active users is at most $N_{\mathrm{a\,max}}$. We assume $N_{\mathrm{a\,max}}$ is known at the receiver.
Remark 1: The precoding scheme is proposed because in reality, T is usually longer than the lengths of small packets. Also, the precoding scheme contributes to solving the signal recovery problem in the situation where $N_{\mathrm{a}} > M$.
Remark 2: Each user knows its own precoding matrix and the BS knows all users' precoding matrices. Since any two precoding matrices are not allowed to be identical, each precoding matrix represents a unique user. Therefore, no additional identity information of users is required.
Remark 3: The precoding matrix should have full column rank for data recovery. Additionally, we assume each column of the precoding matrices is normalized to have unit energy.

III. Normalized BOMP Algorithm For Signal Recovery

Sparsity-aware recovery has been widely studied in the frame work of CS [24], [25]. Initial work in CS treats sparse weighting coefficients as just randomly located among all possible positions in a data vector, i.e., random sparsity. When block-sparsity of the vector is taken into account, it is possible to obtain better recovery performance and reduce the number of required measurements [26]–[29].

As for signal recovery algorithm to be used at the BS, we first apply the known BOMP algorithm and modify it to our problem. The normalized BOMP algorithm is therefore introduced. Then, we analyze the performance of the normalized BOMP, including user detection and signal demodulation. The order statistics [34] is essential for our analysis.

A. Normalized BOMP algorithm

The main idea of BOMP is that, at each iteration, it chooses one block from $\mathbf{B}$ which has the maximum correlation with the residual signal. After that, it will use all the previously selected blocks to update the signals by solving a least-square (LS) problem [27]. Due to $\mathbf{B}_n = \mathbf{P}_n \otimes \mathbf{h}_n$ and to mitigate the negative influence of $\mathbf{h}_n$ on user detection, it is beneficial

to use the norm of $\mathbf{h}_n$ to normalize the columns of $\mathbf{B}_n$ when calculating the correlation value. Idea of similar modification can also be found in [18] to combat the influences of channels in a CDMA system. Detailed iterations of normalized BOMP is presented in the Algorithm 1, where $c_{j,k} = \|\mathbf{B}_j^H \mathbf{r}_{k-1}\|_2^2 / \|\mathbf{h}_j\|_2^2$ is the squared correlation coefficient of normalized.

Algorithm 1 normalized BOMP

1: **Input:** $\mathbf{B}$, $\mathbf{y}$, ρ_0, all channel vectors, iterations K (necessarily $K \leq \lfloor \frac{MT}{d} \rfloor$ or directly $K = N_{\mathrm{a\,max}}$)

2: **Initialization:** index set $\Lambda_0 = \phi$, residual signal $\mathbf{r}_0 = \mathbf{y}$

3: **for all** $k \in \{1, 2, \cdots, K\}$ **do**

1) calculate the $c_{j,k}$ for $j \in \{[N] \backslash \Lambda_{k-1}\}$: $c_{j,k} = \|\mathbf{B}_j^H \mathbf{r}_{k-1}\|_2^2 / \|\mathbf{h}_j\|_2^2$

2) find the index λ_k: $\lambda_k = \arg\max_{j \in \{[N] \backslash \Lambda_{k-1}\}} c_{j,k}$

3) augment the index set: $\Lambda_k = \Lambda_{k-1} \cup \{\lambda_k\}$

4) update the signals by LS algorithm: $\bar{\mathbf{s}}_{\Lambda_k} = \arg\min_{\mathbf{s}_0} \|\mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_k}\mathbf{s}_0\|_2$

5) update the residual signals: $\mathbf{r}_k = \mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_k}\bar{\mathbf{s}}_{\Lambda_k}$

4: **end for**

5: **Output:** $\bar{\mathbf{s}}_K$: the reconstruction of most likely signal-bearing blocks in the index set Λ_K

With the reconstructed coefficients, we will then demodulate the transmitted messages.

B. Analysis of group user detection probability

Since the identities of the active users are unknown to the BS, the BS need to first perform detection of the active users' identities before symbol demodulation can be performed. At the k-th iteration, in order to guarantee that an inactive user is not falsely added to the list of users to be demodulated, we need $\max_{j \in I} c_{j,k} > \max_{j \notin I} c_{j,k}$. In our analysis below, we will use group user detection success rate (GUDSR) as the performance metric, which is defined as the probability that all active users are selected by the recovery algorithm, i.e., $I \subset \Lambda_K$. When $K > N_{\mathrm{a}}$, as many as $K - N_{\mathrm{a}}$ non-active users will be falsely selected as active, this is unavoidable for our normalized BOMP algorithm because the number N_{a} of active users is unknown and no other side information is available or utilized to decide whether a user is actually active. The situation can be greatly improved when error control coding is used (see Section IV).

Exact active user detection probability analysis is difficult for several reasons, including 1) the distribution of the entry of $\mathbf{B}$ is difficult to characterize because it involves both the linear precoding code and the channel fading, and 2) error propagation in the normalized BOMP iterations cannot be avoided. To make the analysis tractable, we make the following approximations:

1) We assume that the entries of precoding matrices are i.i.d. complex Gaussian variables, with zero mean and variance $\frac{1}{T}$. When d and T are large enough, such approximation is reasonable for the analysis of the statistics of the correlation coefficients.
2) Although the normalized BOMP algorithm works with the channel realizations, we evaluate the performance by averaging over channel fading statistics, which makes the analysis tractable. The result will therefore depend on the channel statistics rather than specific channel realizations.
3) We approximate the distributions of the $c_{j,k}$ for both $j \notin I$ and $j \in I$ as Gaussian. Even though by definition $c_{j,k}$ is non-negative, the approximation is reasonable when dT is large.

In the following, we will analyze the performance of user detection by several steps. The order of selecting active users is firstly studied and then order statistics is applied for the analysis.

1) User selection order: At the start of the k-th iteration, the residual signal from the previous iteration is

$$\mathbf{r}_{k-1} = \mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_{k-1}}\bar{\mathbf{s}}_{\Lambda_{k-1}} \tag{6}$$

$$= \mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H\mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H\mathbf{y} \tag{7}$$

$$= \sqrt{\rho_0}\sum_{i\in I\setminus\Lambda_{k-1}}\mathbf{B}_i\mathbf{s}_i + \tilde{\mathbf{z}}_k \tag{8}$$

where $\tilde{\mathbf{z}}_k = \mathbf{z} - \mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H\mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H\left(\sqrt{\rho_0}\sum_{i\in I\setminus\Lambda_{k-1}}\mathbf{B}_i\mathbf{s}_i + \mathbf{z}\right)$, and the second term of $\tilde{\mathbf{z}}_k$ is the propagating error of the normalized BOMP resulting from previous iterations. We will first evaluate the means of the squared correlation coefficients $c_{j,k}$ between the residual signal $\mathbf{r}_{k-1}$ and the normalized signature sequences $\frac{1}{\|\mathbf{h}_j\|_2}\mathbf{B}_j$ for both active and inactive users. Let $\mu_{0,k,j}$ and $\mu_{1,k,j}$ denote the mean values of $c_{j,k}$ for $j \notin I$ (inactive user) and $j \in I$ (active user), respectively. Let s_k be the number of active users that have been selected by the first k

iterations. Let σ_k^2 denote the variance of each element of $\tilde{\mathbf{z}}_k$. Then we have

$$\mu_{0,k,j} \approx (N_{\mathrm{a}} - s_{k-1})\frac{\rho_0 d^2}{T} + d\sigma_k^2, \tag{9}$$

$$\mu_{1,k,j} \approx \frac{\rho_0 d}{T}(d + T - 1)\mathbb{E}\left\{\mathbf{h}_j^H \mathbf{h}_j\right\} + \rho_0(N_{\mathrm{a}} - s_{k-1} - 1)\frac{d^2}{T} + d\sigma_k^2. \tag{10}$$

Results in (9) and (10) have been obtained by averaging over all channel realizations. See Appendix A for the derivations of (9) and (10) and the approximations involved. Generally speaking, among the remaining active users that have not been selected, the one with the largest $\mathbf{h}_j^H \mathbf{h}_j$ is expected to be selected at each iteration. This can also be observed from (10) — if the channels are random, then the one with the largest energy $\mathbb{E}\left\{\mathbf{h}_j^H \mathbf{h}_j\right\}$ will result in the largest mean in the squared correlation coefficient. Next, we will apply order statistics to derive the distributions of $\mathbf{h}_j^H \mathbf{h}_j$ for different active users.

2) Order statistics of channel coefficients: Without loss of generality, we assume $I = [N_{\mathrm{a}}]$ and $\mathbf{h}_1^H \mathbf{h}_1 \geq \mathbf{h}_2^H \mathbf{h}_2 \geq \cdots \geq \mathbf{h}_{N_{\mathrm{a}}}^H \mathbf{h}_{N_{\mathrm{a}}}$. Indices from $N_{\mathrm{a}} + 1$ to N are for the inactive users. For each user j, the *unordered* random variable $\mathbf{h}_j^H \mathbf{h}_j$ follows a chi-squared distribution with $2M$ degrees of freedom. The PDF $f(x)$ and CDF $F(x)$ of the unordered $\mathbf{h}_j^H \mathbf{h}_j$ are given as follows:

$$f(x) = \exp(-x)\frac{x^{M-1}}{(M-1)!} \tag{11}$$

$$F(x) = 1 - \exp(-x)\sum_{k=0}^{M-1}\frac{x^k}{k!} \tag{12}$$

both of which hold for $x \geq 0$, and are zero when $x < 0$. By order statistics [34], the PDF of the *ordered* $\mathbf{h}_n^H \mathbf{h}_n, n = 1, 2, \cdots, N_{\mathrm{a}}$, is given as

$$f_n(x) = \frac{N_{\mathrm{a}}!}{(N_{\mathrm{a}} - n)!(n-1)!}F^{N_{\mathrm{a}}-n}(x)[1 - F(x)]^{n-1} f(x) \tag{13}$$

for $x \geq 0$. The mean $\mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\}$ can then be computed based on the above PDF.

3) Variance of residual error: With K iterations, there are many cases for a successful group user detection. To simplify the study of GUDSR, we consider only one case for simplicity: the first N_{a} iterations select all the N_{a} active users and the selection order is based on the descending order of their $\mathbf{h}_n^H \mathbf{h}_n$. Since there are other cases where the active users can be correctly identified, the GUDSR we obtain by considering only one case of the successful detection will serve as a lower bound to the true GUDSR. The assumed selection order is in line with a similar idea used to analyze the performance of successive IC in [34]. Next, we will analyze the probability

of the occurrence of the considered case. Our analysis of GUDSR will just involves the first $N_{\rm a}$ iterations. At the k-th iteration, the algorithm selects the k-th active user, i.e., the active user with $\mathbf{h}_k^H \mathbf{h}_k$.

When the above selection order is considered in the analysis, for $1 \leq k \leq N_{\rm a}$ and after the previous $k-1$ iterations, the variance σ_k^2 of $\tilde{\mathbf{z}}_k$ (c.f. (8)) is given by

$$\sigma_k^2 = \frac{1}{MT}\mathbb{E}\left\{\tilde{\mathbf{z}}_k^H \tilde{\mathbf{z}}_k\right\} \tag{14}$$

$$= \frac{1}{MT}\mathbb{E}\left\{\mathbf{z}^H \left[\mathbf{I}_{MT} - \mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H \mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H\right]\mathbf{z}\right\} \tag{15}$$

$$+ \frac{\rho_0}{MT}\sum_{n=k}^{N_{\rm a}}\mathbb{E}\left\{\mathbf{s}_n^H \mathbf{B}_n^H \mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H \mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H \mathbf{B}_n \mathbf{s}_n\right\} \tag{16}$$

$$\approx \frac{1}{MT}\,\mathrm{Tr}\left\{\mathbb{E}\left\{\mathbf{I}_{MT} - \mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H \mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H\right\}\right\} \tag{17}$$

$$+ \frac{\rho_0 d}{M^2T^2}\,\mathrm{Tr}\left\{\mathbb{E}\left\{\mathbf{B}_{\Lambda_{k-1}}\left(\mathbf{B}_{\Lambda_{k-1}}^H \mathbf{B}_{\Lambda_{k-1}}\right)^{-1}\mathbf{B}_{\Lambda_{k-1}}^H\right\}\right\}\sum_{n=k}^{N_{\rm a}}\mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} \tag{18}$$

$$= 1 - \frac{(k-1)d}{MT} + \frac{\rho_0 (k-1)d^2}{M^2T^2}\sum_{n=k}^{N_{\rm a}}\mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} \tag{19}$$

where the identity $\mathbf{B}_n^H \mathbf{B}_n = \left(\mathbf{P}_n^H \otimes \mathbf{h}_n^H\right)\left(\mathbf{P}_n \otimes \mathbf{h}_n\right) = \left(\mathbf{P}_n^H \mathbf{P}_n\right) \otimes \left(\mathbf{h}_n^H \mathbf{h}_n\right) = \left(\mathbf{P}_n^H \mathbf{P}_n\right)\left(\mathbf{h}_n^H \mathbf{h}_n\right)$ has been used in the above derivation.

4) Statistics of the squared correlation: At the k-th iteration, by our previous assumption on the active user selection order, the algorithm will select the k-th active user ranked according to channel strengths. Since the statistics of the squared correlation coefficient $c_{j,k}$ for inactive user $j \notin I$ does not depend on the inactive user index, we use $\mu_{0,k}$ and $\sigma_{0,k}^2$ to denote the mean and variance for $c_{j,k}$ for $j \notin I$. For $c_{k,k}$, which corresponds to the k-th active user, we denote

its mean and variance as $\mu_{1,k}$ and $\sigma_{1,k}^2$, respectively. We have the following results[2]

$$\mu_{0,k} \approx \frac{\rho_0 d^2}{MT} \sum_{n=k}^{N_\text{a}} \mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} + d\sigma_k^2 \tag{20}$$

$$\sigma_{0,k}^2 \approx d(d+1)\sigma_{\mathbf{r}_{k-1}}^4 - \mu_{0,k}^2 \tag{21}$$

$$\mu_{1,k} \approx \frac{\rho_0 d}{T}(d+T-1)\mathbb{E}\left\{\mathbf{h}_k^H \mathbf{h}_k\right\} + \frac{\rho_0 d^2}{MT} \sum_{n=k+1}^{N_\text{a}} \mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} + d\sigma_k^2 \tag{22}$$

$$\sigma_{1,k}^2 \approx \frac{\rho_0^2}{T^2}(d^2+d)(d+T-1)^2 \mathbb{E}\left\{\left(\mathbf{h}_k^H \mathbf{h}_k\right)^2\right\} + d(d+1)\sigma_{\tilde{\mathbf{r}}_k}^4 \tag{23}$$

$$+ \rho_0 d\left[\frac{3}{4} + 2d + \frac{2d^2}{T} + \frac{d}{4T} - \frac{d^2}{2T^2}\right]\sigma_{\tilde{\mathbf{r}}_k}^2 \mathbb{E}\left\{\mathbf{h}_k^H \mathbf{h}_k\right\} - \mu_{1,k}^2 \tag{24}$$

where $\sigma_{\mathbf{r}_{k-1}}^2$ is the variance of each elements of the residual signal $\mathbf{r}_{k-1}$, given by

$$\sigma_{\mathbf{r}_{k-1}}^2 = \frac{1}{MT}\mathbb{E}\left\{\mathbf{r}_{k-1}^H \mathbf{r}_{k-1}\right\} \tag{25}$$

$$= \frac{\rho_0}{MT} \sum_{n \in I \setminus \Lambda_{k-1}} \mathbf{s}_n^H \mathbf{B}_n^H \mathbf{B}_n \mathbf{s}_n + \frac{1}{MT}\tilde{\mathbf{z}}_k^H \tilde{\mathbf{z}}_k \tag{26}$$

$$\approx \frac{\rho_0 d}{MT} \sum_{n=k}^{N_\text{a}} \mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} + \sigma_k^2 \tag{27}$$

and $\sigma_{\tilde{\mathbf{r}}_k}^2$ is the variance of each elements of $\tilde{\mathbf{r}}_k := \mathbf{r}_{k-1} - \sqrt{\rho_0}\mathbf{B}_k \mathbf{s}_k$, given as

$$\sigma_{\tilde{\mathbf{r}}_k}^2 \approx \frac{\rho_0 d}{MT} \sum_{n=k+1}^{N_\text{a}} \mathbb{E}\left\{\mathbf{h}_n^H \mathbf{h}_n\right\} + \sigma_k^2, \tag{28}$$

which is reduced to $\sigma_{\tilde{\mathbf{r}}_k}^2 = \sigma_k^2$ when $k = N_\text{a}$. The derivations of (20)-(24) are presented in the Appendix B. With the Gaussian assumption on $c_{j,k}$ in general and $c_{k,k}$ in particular, the PDF of $c_{k,k}$ for the k-th active user is $f_{1,k}(x) = f^{(\text{N})}(x|\mu_{1,k}, \sigma_{1,k}^2)$. The PDF of $c_{j,k}$ for a generic inactive user $j \notin I$ is given by $f_{0,k}(x) = f^{(\text{N})}(x|\mu_{0,k}, \sigma_{0,k}^2)$, which does not depend on the index of the inactive user.

5) Successful user detection probability: False alarm in the detection is generated in the first N_a iterations of the normalized BOMP algorithm when an inactive user has a larger squared correlation $c_{j,k}$ than an active user. Using order statistics, the PDF of the $\max_{j \notin I} c_{j,k}$ is given by

$$f_{0,k,\text{max}}(x) = (N - N_\text{a})[F_{0,k}(x)]^{N-N_\text{a}-1} f_{0,k}(x) \tag{29}$$

[2]When $k = N_\text{a}$, the $\sigma_{0,k}^2$ can be more precisely given by using the result of (48) as the secondary moment.

where $F_{0,k}(x)$ is the CDF of $c_{j,k}$ for $j \notin I$.

Let E_k denote the event that the k-th active user is selected at the k-th iteration, i.e., $c_{k,k} > \max_{j \notin I} c_{j,k}$, whose probability is given by

$$P(E_k) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{x_1} f_{0,k,\max}(x_0) dx_0 \right] f_{1,k}(x_1) dx_1, \tag{30}$$

The successful group user detection of the considered decoding order has a probability as follows

$$P^{(\mathrm{u})} = \prod_{k=1}^{N_\mathrm{a}} P\left(E_k | E_1, E_2, \cdots, E_{k-1}\right) \approx \prod_{k=1}^{N_\mathrm{a}} P(E_k) \tag{31}$$

where the last approximation is valid when the probability of $E_1, \ldots, E_{k-1}$ is high, namely when the correct detection probability is high.

Remark 4: Due to the approximations made in the derivation of $P^{(\mathrm{u})}$, it can be used as an approximate lower bound of the true GUDSR only. In the simulations section, we will demonstrate the usefulness of the derived lower bound using numerical examples, when the SNR is not too small.

Remark 5: The above analysis about GUDSR shows that, large N_a and N are not good for user detection when other parameters are given. It is because they make the gap between normalized correlation coefficients of the last active user and non-active user small, which is quantified by the order statistics.

Remark 6: Probability of successful detection of *individual* active users is another useful performance metric. However it is more difficult to analyze. We will resort to numerical simulations for their evaluations. The error rate of group detection is naturally typically higher than that of individual detection because any individual detection error causes the group detection failure.

C. Symbol error rate analysis

For the analysis of symbol demodulation error probability, we assume that quadrature phase shift keying (QPSK) modulation is used as an example. Other modulation schemes can be similarly considered.

When all the N_a active users and $K - N_\mathrm{a}$ non-active users have been selected by K iterations of normalized BOMP, signal update is finally as follows

$$\bar{\mathbf{s}}_{\Lambda_K} = \mathbf{s}_{\Lambda_K} + \frac{1}{\sqrt{\rho_0}} \left(\mathbf{B}_{\Lambda_K}^H \mathbf{B}_{\Lambda_K} \right)^{-1} \mathbf{B}_{\Lambda_K}^H \mathbf{z} \tag{32}$$

where $\mathbf{B}_{\Lambda_K}$ is $MT \times Kd$ matrix composed of K selected blocks.

When the entries of $\mathbf{B}_{\Lambda_K}$ are i.i.d. complex Gaussian variables, the demodulation performance of (32) has been analyzed through an upper bound in [35]. Our model is different in several aspects: 1) each column of $\mathbf{B}_{\Lambda_K}$ is a Kronecker product of a precoding vector and a single-input multiple-output channel response vector; 2) every d columns of $\mathbf{B}_{\Lambda_K}$ share the same channel vector; 3) the precoding vectors are normalized. Despite these differences, we can obtain an approximated SER expression by assuming that the entries of $\mathbf{B}_{\Lambda_K}$ are i.i.d. complex Gaussian variables with zero mean and variance $1/T$. Under this assumption, all symbols of all active users have the same average (ergodic) performance, where the average is over channel fading and noise statistics. As a result we can focus on the analysis on any one symbol, say the first symbol of the first (unordered) active user.

Assuming without loss of generality that the first column of $\mathbf{B}_{\Lambda_K}$ belongs to an active user. Let $G = 1/\left[\left(\mathbf{B}_{\Lambda_K}^H \mathbf{B}_{\Lambda_K}\right)^{-1}\right]_{1,1}$, which is a random variable that affects the SNR of the useful symbol. Let $\mathbf{b}_1$ be the first column of $\mathbf{B}_{\Lambda_K}$. Using the result in [35, Appendix A], we can express G as

$$G = \sum_{m=1}^{MT-Kd+1} |b_m|^2 \tag{33}$$

where b_m, $m = 1, \ldots, MT - KT + 1$, are i.i.d. and they have the same distribution as the entries of $\mathbf{b}_1$, namely zero-mean, circularly symmetric complex Gaussian, with variance $1/T$. As a result, G is a scaled chi-squared distributed random variable with degrees of freedom $2(MT - Kd + 1)$, with PDF given as follows:

$$f_G(g) = \frac{T^{MT-Kd+1}}{(MT - Kd)!} \cdot e^{-Tg} g^{MT-Kd}. \tag{34}$$

Conditioned on $G = g$, the SER of the symbol of interest is given by

$$\text{SER}(g) = \text{erfc}\left(\sqrt{0.5\rho_0 g}\right) - \left[\frac{1}{2}\text{erfc}\left(\sqrt{0.5\rho_0 g}\right)\right]^2. \tag{35}$$

The average SER can then be evaluated as

$$\text{SER} = \int_0^\infty \text{SER}(g) \cdot f_G(g) dg. \tag{36}$$

Remark 7: When $MT - Kd + 1 \gg 1$, G is well approximated by a constant

$$G_0 = \frac{MT - Kd + 1}{T} = M - \frac{Kd - 1}{T}, \tag{37}$$

according to the law of large numbers. The average SER is then given by $\text{SER}(g)\big|_{g=G_0}$. That is, compared to a single user QPSK transmission scheme, the performance for a multi-user system as we proposed suffers an multiplicative SNR loss factor of G_0. Large M is preferred for better performance as expected. However, there is an effect of diminishing returns based on (37). Intuitively, a large K is preferred for increasing user detection probability. On the other hand, large K causes more SNR degradation. When Kd approaches MT, symbol error probability is expected to be large.

IV. ICBOMP Algorithm For Signal Recovery

The previous analyses of normalized BOMP indicate that a smaller iteration number K can provide a lower SER. However, since N_{a} is unknown at the BS, K must be large enough to accommodate a possibly large N_{a}. As Kd approaches MT, the performance of symbol demodulation may degrade significantly. During the iterations, many blocks are expected to have only a few demodulation errors, which can be corrected if an error control code is employed. Error control coding is widely used for correcting demodulation errors due to noise and interferences, and error detection codes such as cyclic redundancy check (CRC) codes can be utilized to indicate whether the decoded packets are error-free.

This motivates us to consider error control coding at the transmitters and an improved ICBOMP algorithm for recovery. The ICBOMP algorithm makes use of error correction and detection to carry out perfect interference cancellation in each iteration of the normalized BOMP. Perfect interference cancellation means that if some blocks of signals are correctly recovered before the iterations have been finished, they are regarded as interference to the following iterations and will be canceled to end their update. The idea of interference cancellation can be found in [34].

The ICBOMP is presented in Algorithm 2.

A. ICBOMP algorithm

In Algorithm 2, $\bar{\text{s}}^i_{\Lambda_k}$ is the i-th block of $\bar{\text{s}}_{\Lambda_k}$ obtained by the LS reconstruction. Since some blocks of signals may have been correctly recovered and then canceled by the previous $k-1$ iterations, it has $1 \leq i \leq |\Lambda_{k-1}| + 1 \leq k$. Vector $\tilde{\text{s}}_{\Lambda_k}$ is the output of $\bar{\text{s}}_{\Lambda_k}$ after error correction and detection. Let $\tilde{\Lambda}_k$ denote the index set for error-free blocks at the k-th iteration.

Algorithm 2 ICBOMP

Input: the same as as normalized BOMP

Initialization: index set $\Lambda_0 = \phi$, residual signal $\mathbf{r}_0 = \mathbf{y}$

for all $k \in \{1, 2, \cdots, K\}$ **do**

1) initialize a temporary index set: $\tilde{\Lambda}_k = \phi$
2) calculate the $c_{j,k}$ for $j \in \left\{[N] \backslash \{\Lambda_{k-1}\ \tilde{\Lambda}_{k-1} \cdots \tilde{\Lambda}_1\}\right\}$: $c_{j,k} = \|\mathbf{B}_j^H \mathbf{r}_{k-1}\|_2^2 / \|\mathbf{h}_j\|_2^2$
3) find the index λ_k: $\lambda_k = \arg\max_{j \in \left\{[N] \backslash \{\Lambda_{k-1}\ \tilde{\Lambda}_{k-1} \cdots \tilde{\Lambda}_1\}\right\}} c_{j,k}$
4) augment the index set: $\Lambda_k = \Lambda_{k-1} \cup \{\lambda_k\}$
5) update the signals by LS algorithm: $\bar{\mathbf{s}}_{\Lambda_k} = \arg\min_{\mathbf{s}_0} \|\mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_k}\mathbf{s}_0\|_2$
6) **for all** $i \in \{1, 2, \cdots, |\Lambda_k|\}$ **do** the following error correction and detection:
   6a) If $\bar{\mathbf{s}}^i_{\Lambda_k}$ after error correction and detection is error-free, then $\tilde{\mathbf{s}}^i_{\Lambda_k}$ is the decoded version of $\bar{\mathbf{s}}^i_{\Lambda_k}$; the original index of this i-th block should be added into $\tilde{\Lambda}_k$.
   6b) If $\bar{\mathbf{s}}^i_{\Lambda_k}$ after error correction and detection is not error-free, then $\tilde{\mathbf{s}}^i_{\Lambda_k}$ is the same as $\bar{\mathbf{s}}^i_{\Lambda_k}$ without correction.
7) **end for**
8) perform perfect cancellation: $\mathbf{y} = \mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\tilde{\Lambda}_k}\tilde{\mathbf{s}}_{\tilde{\Lambda}_k}$
9) update the index sets: $\Lambda_k = \Lambda_k \backslash \{\tilde{\Lambda}_k\}$
10) update the residual signals: $\mathbf{r}_k = \mathbf{y} - \sqrt{\rho_0}\mathbf{B}_{\Lambda_k}\tilde{\mathbf{s}}_{\Lambda_k}$

end for

Output: $\tilde{\mathbf{s}}_K$ and the already correctly recovered signal blocks

It can be observed from the comparison of ICBOMP and normalized BOMP algorithms that, their main difference occurs after signals have been updated by the LS method. The ICBOMP will perform error correction and detection for each block in $\bar{\mathbf{s}}_{\Lambda_k}$. And when some blocks of signals are decided to have been correctly recovered, ICBOMP will perform the perfect cancellation. When perfect cancellation is allowed, ICBOMP can offer the following benefits:

1) *Higher user detection successful rate and signal demodulation accuracy:* By error correction and detection, the accuracy of signal demodulation is improved, and this will reduce the error propagation which is the main drawback of the (normalized) BOMP. In turn, this is beneficial for the detection of remaining users and their signals' demodulation. Additionally,

since error correction and detection are performed iteration by iteration for signals that are not previously error-free, the signals have a higher probability to be correctly recovered.

2) *Lower computational cost:* When some blocks of signals have been exactly recovered, The ICBOMP regards them as interference to the following iterations and no longer update these signals. For the LS update, more blocks indicate more computational cost. Therefore, the ICBOMP can reduce the computational cost.

The benefit of ICBOMP is from the perfect interference cancellation after error correction and detection. The price paid is a slightly decreased data rate due to coding and a certain amount of additional calculations in decoding. However, the reduced computational cost outweighs the additional computational cost of decoding. As demonstrated by simulations, when the SNR is not too small and under the same condition, by the BCH coding scheme, the ICBOMP recovery runs much faster than the normalized BOMP. The performance analysis by the ICBOMP recovery will require a lot of future work.

V. Numerical Results

The simulation studies for verifying the proposed scheme are presented in this section. In all simulations, QPSK is applied for data modulation. Channel vectors are i.i.d. complex Gaussian with zero-mean and unit-variance for each element. The N_{a} active users are chosen uniformly at random among all N online users. We will choose the frame length T to be a multiple of the maximum length of short messages, $T = 5d$ in all our presented simulations. All packets are assumed to have 100 nonredundant message-bearing symbols, i.e., 200 information bits. As for precoding design, we simply generate a complex Gaussian matrix and then normalize its columns to produce $\mathbf{P}_n$, $n = 1, 2, \cdots, N$.

In our simulations, the SNR is defined as E_s/N_0, where E_s is the symbol energy and N_0 is the noise spectral density. The SER is computed as follows: when a demodulated symbol of an active user is different from its original symbol, we claim an error; if an active user is missed to detect, then all its d symbols are treated as erroneous. When an active user is selected before algorithm finishes, we claim one successful detection, and the number of all successful detections divided by N_{a} is UDSR.

A. Simulations with normalized BOMP

In the first two experiments, we have the simulations of normalized BOMP (labeled as NBOMP) when BS are equipped with 8 antennas.

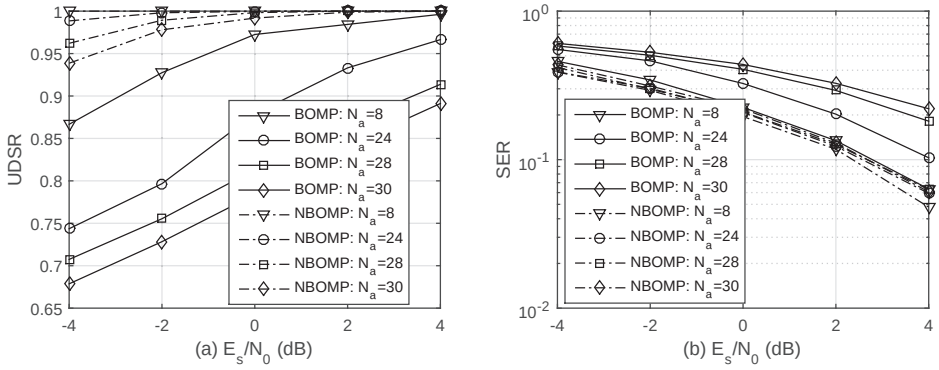

Fig. 1. performance of the normalized BOMP and the BOMP, $(M, N, d, K) = (8, 80, 100, 30)$

Test Case 1: Figure 1 compares the performance of normalized BOMP and BOMP algorithms. As we can see, for both algorithms, when the number of active users increases, lower UDSR occurs. Obviously, the normalization approach contributes to a much higher UDSR, which directly results in a much lower SER, especially when N_{a} is large like $N_{\mathrm{a}} \geq 24$. We also observe that, as long as UDSR exceeds a certain value, for example, UDSR $> 85\%$, SERs of different numbers of active users are very close.

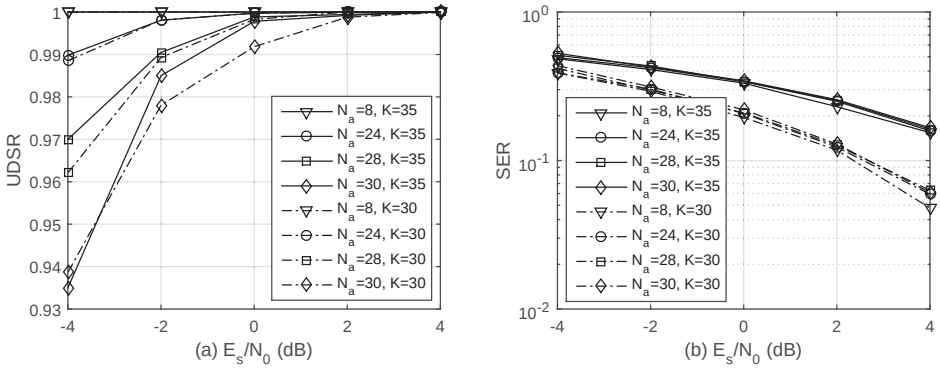

Fig. 2. performance of the normalized BOMP, $(M, N, d) = (8, 80, 100)$

Test Case 2: Influences of iteration number K on the performance of USDR and SER are given in the Figure 2. It can be drawn from the results that, UDSR performance is improved

with a larger K, see Figure 2 (a). While Figure 2 (b) shows that, a larger K will degrade the performance of symbol demodulation, which is also indicated by our previous analysis result of SER. Additionally, it is noted that, when $E_s/N_0 = 4$dB, even 30 iterations are enough to select 30 active users.

B. Simulations with ICBOMP

In this part, we report results on ICBOMP algorithm. Channel coding BCH(255,223) that can correct at most 4 erroneous bits and CRC of 16 bits are acting as error correction and detection codes. With necessary zero padding, we map 200 raw data bits and 16 CRC bits into 248 bits. Therefore, it has $d = 124$ and $T = 5d = 620$. BCH(255,223) still offers a high coding rate. It should also be stated that BCH coding is not the only choice, we can use other more efficient coding schemes. Compared with the previous simulations, 0.6dB reduction in E_s/N_0 is caused due to channel coding at the rate $\frac{223}{255}$.

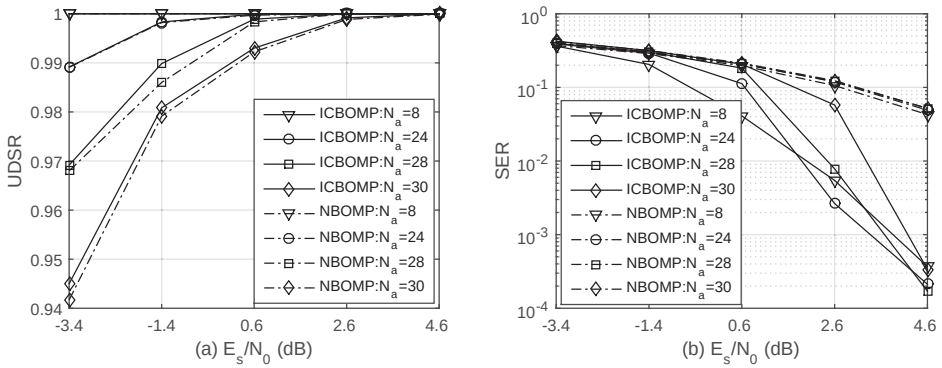

Fig. 3. performance of the ICBOMP and the normalized BOMP, $(M, N, K, d) = (8, 80, 30, 124)$

Test Case 3: Figure 3 depicts the performance when ICBOMP is exploited. Comparison with normalized BOMP is also made, where the same error correction and detection are only performed at final iteration of the normalized BOMP. As we can see, ICBOMP improves the performance, especially in symbol demodulation when E_s/N_0 increases. In fact, whether correction and detection coding is used when iterations finish almost makes no difference for the normalized BOMP. However, for ICBOMP, since error correction and detection are performed at each iteration, each block of signals has a much higher probability to be correctly recovered. Interestingly, we observe that, when almost all active users can be selected, a larger N_a offers a lower SER for a

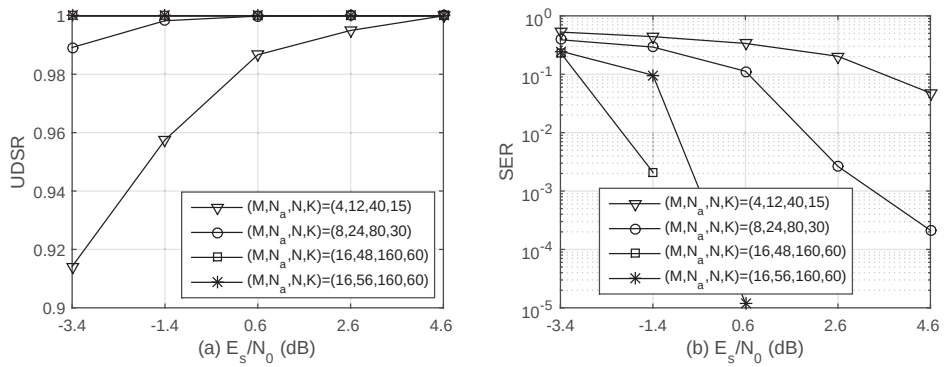

Fig. 4. performance of ICBOMP when BS is equipped with different numbers of antennas, $d = 124$

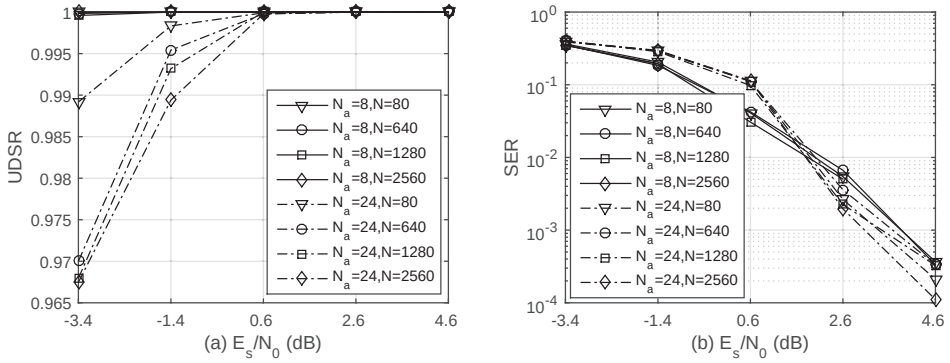

Fig. 5. performance of the ICBOMP when there are different numbers of online users, $(M, K, d) = (8, 30, 124)$

same N. This result cannot be obviously deduced from analysis. It is related to the order statistics of powers of channel vectors. Such phenomenon is also expected to appear in the normalized BOMP when there are not too many active users.

Test Case 4: In Figure 4 we have the simulation results when BS is equipped with different numbers of antennas, respectively with 4, 8, and 16 antennas. To have a fair comparison, we set different parameters to guarantee the ratios of $(M : N : N_{\mathrm{a}} : K)$ at a same level, and to show the benefits of $M = 16$, one additional result of larger N_{a}/M $(i.e., N_{\mathrm{a}} = 56)$ is presented. The results show that, more antennas at BS can offer remarkable benefits, both in user detection and symbol demodulation, and these benefits are much more than linear with the antenna number M.

Test Case 5: It is desirable that a system can accommodate many online users. When BS is equipped with 8 antennas, the UDSRs and SERs of different numbers of online users are depicted

in the (a) and (b) of Figure 5. As we can see, although a larger number of online users N decreases the UDSR, the degradation from $N = 80$ to $N = 2560$ is not very obvious for a same N_{a}. When N_{a} is not large, like $N_{\text{a}} = 8$ in our simulation, such degradation can be neglected. SERs of different numbers of online users are rather close, as shown in the Figure 5 (b). It means that our proposed scheme with ICBOMP receiver for user detection and symbol demodulation is applicable for a system with many online users. As the same in Figure 3, when all active users can be selected in the higher SNR regime, SER of larger N_{a} is lower.

Combination of the results in Figure 4 and Figure 5 gives us the confidence that, when BS is equipped with many antennas, e.g., in massive MIMO systems, more online users and active users are allowed to exist. This makes the proposed scheme especially suitable for the future 5G systems to support massive connections.

Different frame lengths also offer different performance. As a conclusion, since T acts as block length in channel coding, larger T offers better performance in the relatively high SNR regime.

C. Computational cost reduction by ICBOMP

Test Case 6: In this part, we cite one example to demonstrate the benefit of ICBOMP in computational cost reduction. The results are shown in the Figure 6 and they are obtained when $E_s/N_0 = 4.6$dB. The ordinate is the number of blocks for signal update at each iteration, which is directly related to the computational cost. The result of the (normalized) BOMP is also presented. It shows that under the given condition, some blocks of signals can be correctly recovered before the iterations finish. For example, for $N_{\text{a}} = 30$ under the given condition, the maximum number of blocks of signals to update is about 22, and after that, the number quickly decreases. However, for the (normalized) BOMP, since no blocks will be canceled, its number of blocks for update increases linearly with the iteration number. The benefit of ICBOMP in reducing the computational cost is therefore demonstrated.

D. Results of the approximate analysis

Test Case 7: In this part, we show two examples to compare our analyses with the simulations, including the user detection in Figure 7 (a) and the symbol demodulation in Figure 7 (b). User detection performance is quantified using GUDSR.

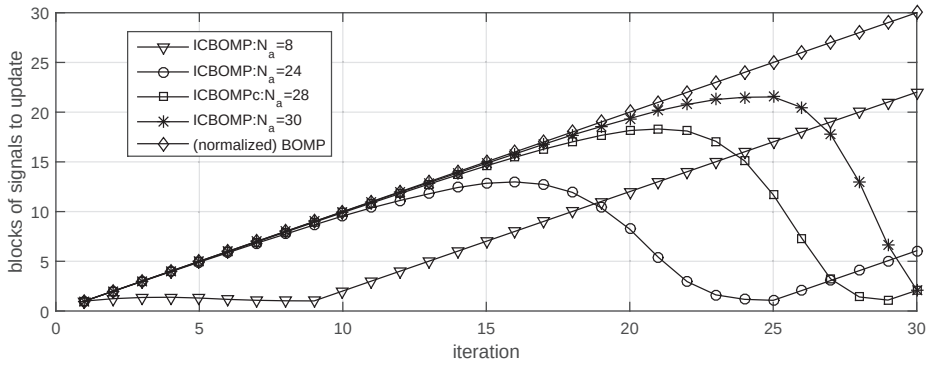

Fig. 6. computational cost reduction by the ICBOMP, $(M, N, K, d) = (8, 80, 30, 124)$

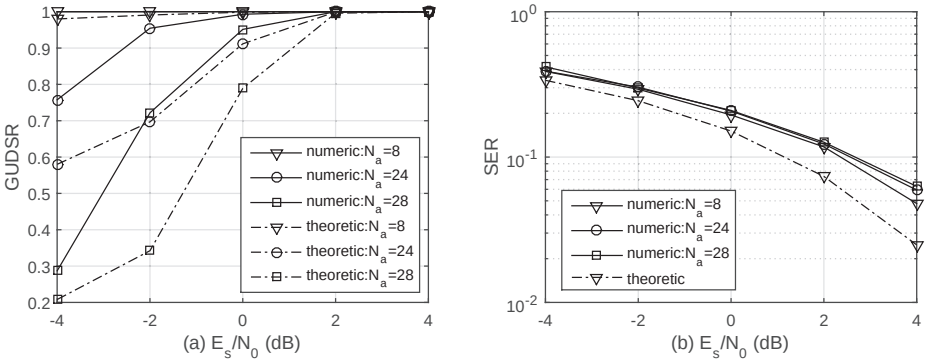

Fig. 7. comparison of simulation and analysis of the normalized BOMP when $(M, N, K, d) = (8, 80, 30, 100)$

It can be observed from the Figure 7 (a) that, our lower bound analysis about GUDSR of the normalized BOMP has certain gap to the simulation results when E_s/N_0 is below a certain level. It is because in this regime, except the most likely case we considered for successful user detection, other possible cases also contribute a high probability for the GUDSR. As E_s/N_0 increases, since the cited case plays a more and more dominant role in the successful user detection, gradually, this gap narrows and the analysis presents a tighter lower bound. On the other hand, our approximate analysis about the SER shows stable performance, about 1.5dB better than the simulation result. Since our SER analysis is based on the case that all active users have been successfully identified, it remains the same for different numbers of the active users. That is why only one analytical curve is presented.

E. Probabilistic transmission

Test Case 8: In order to exactly test the influences of different parameters such as M and N on the performance, we have assumed a constant N_{a} for each set of parameters. However, random transmission with certain active probability p makes more sense for certain applications. In this part, we have simulations for random transmission where each user sends a small packet according to a probability p. Since $N = 1280$, probabilities $p = 0.625\%, 1.875\%$ are expected to respectively produce $N_{\text{a}} = 8, 24$ active users. We set $K = 40$ for this probabilistic transmission case. See Figure 8 for the results.

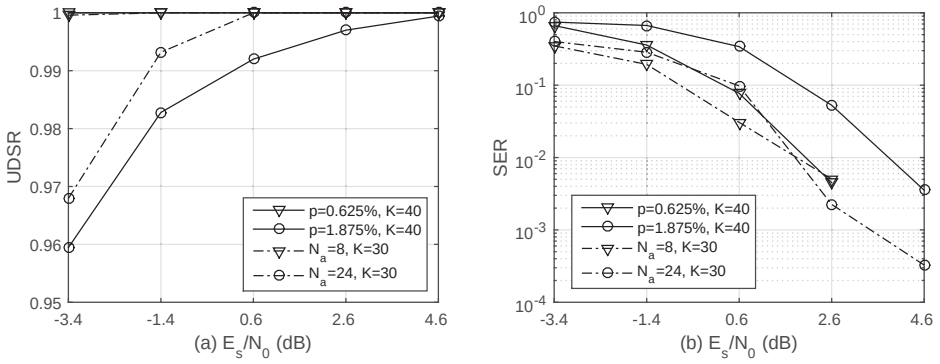

Fig. 8. performance of the ICBOMP when each user transmits with a certain probability, $(M, N) = (8, 1280)$

Under the given condition, almost all results of different active probabilities show performance degradation when compared with the results of given Np active users, especially when p is large. When $p = 0.625\%$, this degradation is not severe and SER degradation mainly comes from a larger iteration number K, which has been previously demonstrated. Such slightly performance degradation is also the case when $p = 1.25\%$, where $N_{\text{a}} = 16$ is expected. When $p = 1.875\%$, if we use Gaussian distribution to estimate the number of active users, then N_{a} can be larger than 47 with a probability 15.85%. Since the number of active users can be even larger than K, it will certainly make some active users missed to detect. This missed alarm greatly harms the performance of symbol demodulation. In fact, the ICBOMP receiver can handle this practical problem to some extent. When signals of one active user have been correctly recovered, set $K = K + 1$. Then, there are more iterations for user detection. Fundamentally, the most effective technique to solve this problem is to equip the BS with more antennas to support more online users and active users.

VI. DISCUSSIONS

In this section, we discuss three issues that are related to our proposed scheme.

A. Comparison with known schemes

Our proposed scheme is different from the existing schemes that are based on the orthogonal access patterns, e.g., time-division multiple access (TDMA), CDMA and single carrier orthogonal frequency division multiplexing (SC-FDMA). TDMA is mostly for the 2G wireless systems, CDMA is deployed in the 3G wireless systems and SC-FDMA is proposed for the uplink in the LTE systems [30]. To meet the requirements of future 5G wireless network, e.g., offering massive connectivity, several non-orthogonal multiple access schemes have been proposed [6]–[9], including sparse code multiple access (SCMA) [6], [7]. Since the number of resources are strictly limited for orthogonal multiple access schemes, the number of users that can be served is limited. Furthermore, resource scheduling is necessary. In the case of burst transmissions of many small packets, dynamic resource scheduling leads to significant signaling overhead. In contrast, our scheme requires no such dynamic scheduling and hence improves the resource utilization efficiency. During one frame interval, since different precoding codes are used for different symbols and even the same symbol can be assigned different precoding codes at different times, the interferences among symbols are averaged. This is favorable for improving the recovery performance. Our system also has robust performance in detecting user activities by incorporating error correction and detection into the recovery algorithm. The comparisons among TDMA, CDMA, SC-FDMA, SCMA and our scheme are summarized in Table I. For the future 5G network, our proposed non-orthogonal many access scheme is promising to address the challenges produced by the transmissions of small packets.

B. Message segmentation for practical transmission

Although the length of small packet is short, block-sparse signal recovery by ICBOMP still requires heavy computations, especially when d is not very small and there are many receiving antennas. To reduce such cost, messages of small packet can be segmented into shorter parts further. These segments can be successively transmitted and signal recovery is successively performed for them. For different segments of one small packet, we can even pre-assign different precoding matrices to them. There are several benefits of using such segmented transmission.

TABLE I

COMPARISONS OF DIFFERENT SCHEMES

	TDMA/SC-FDMA	CDMA	SCMA	Our Scheme
Coding Level	／	symbol	bit	frame
Codebook Design	／	independent of CSI	dependent on CSI	independent of CSI
Resource Scheduling	necessary	necessary	necessary	unnecessary
Massive Connectivity support (N_{a} per resource block with T units)	limited, at most 1	limited, at most $\frac{T}{d}$	unclear when $M > 2$	$\alpha\frac{MT}{d}$ (e.g., $\alpha > 0.7$ when $T = 5d$, $M = 16$ in Figure 4)

Firstly, it can reduce the computational cost for signal recovery. Secondly, it can alleviate the length requirement for constant duration of the block-fading channel. In scenario where channel varies in short period, pilots can be inserted between two segments for the CSI estimation or renewal. Additionally, it can also avoid having highly correlated blocks in $\mathbf{B}$ for large duration. Highly correlated channel vectors can be essentially decorrelated at the BS when different precoding matrices for different segmentation periods are used.

C. Sparsity level estimation

Since the receiver does not know the number of active users N_{a}, a large number of iterations are set for the recovery algorithms. When N_{a} is small and user detection can be successful with less number of iterations. However, unnecessary iterations not only increase the computational cost but also weaken the ability for an exact signal recovery [31]. Such phenomenon has also been observed in our previous analysis and simulations. Our numerical simulation indicate that, when $E_s/N_0 \geq 2.6$dB and with the conditions $M = 8$, $N = 2560$, $T = 5d = 620$, 24 iterations for ICBOMP will correctly identify all the 24 active users with a probability exceeding 99.99%. In this case, technique for sparsity level (i.e., N_{a}) estimation will provide benefits. For example, work to estimate the random sparsity level can be found in [31], [32]. However for block-sparsity scenario, more effective technique will be useful.

VII. CONCLUSIONS

In this paper, we proposed an uplink many access scheme for the services of small packet. The proposed scheme combines the techniques of block precoding at the transmitters and sparsity-aware recovery at the receiver. The proposed non-orthogonal transmission scheme is applicable

to future wireless communication systems with many receive antennas. The reason is that in the future communications, small packets play a more and more important role due to the rapid development of the mobile Internet. The overall throughput of such systems is currently hampered by small packets because of the heavy signaling overhead related to transmitting them. Designed to solve this problem, our scheme can greatly reduce the signaling overhead and therefore guarantees a high throughput. Besides, future communication systems are supposed to have massive antennas at receiver, which enables the proposed scheme to offer massive connections for the frequent and random transmissions of small packets.

APPENDIX A
BRIEF DERIVATIONS OF (9) AND (10)

Due to the space limitation, we only present the important steps of the following derivations. The derivations are based on the previously made approximations.

At the k-th iteration, the residual signal vector is given by (8). For $i \neq j$, symbol vector $\mathbf{s}_i$ is independent of $\mathbf{s}_j$ and channel vector $\mathbf{h}_i$ is independent of $\mathbf{h}_j$. When $j \notin I$, we have that

$$\begin{aligned}
\mu_{0,k,j} &= \mathbb{E}\left\{\frac{\|\mathbf{B}_j^H\mathbf{r}_{k-1}\|_2^2}{\mathbf{h}_j^H\mathbf{h}_j}\right\} = \mathbb{E}\left\{\frac{\mathbf{r}_{k-1}^H\mathbf{B}_j\mathbf{B}_j^H\mathbf{r}_{k-1}}{\mathbf{h}_j^H\mathbf{h}_j}\right\} \\
&= \mathbb{E}\left\{\frac{\rho_0\sum_{i\in I\setminus\Lambda_{k-1}}\mathbf{s}_i^H\mathbf{B}_i^H\mathbf{B}_j\mathbf{B}_j^H\mathbf{B}_i\mathbf{s}_i+\tilde{\mathbf{z}}_k^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{z}}_k}{\mathbf{h}_j^H\mathbf{h}_j}\right\} \\
&\approx \rho_0(N_{\mathrm{a}}-s_{k-1})\mathbb{E}\left\{\mathbf{s}_i^H\mathbf{P}_i^H\mathbf{P}_j\mathbf{P}_j^H\mathbf{P}_i\mathbf{s}_i\right\}+\sigma_k^2\,\mathrm{Tr}\left\{\mathbb{E}\left\{\mathbf{P}_j^H\mathbf{P}_j\right\}\right\} \\
&\approx \rho_0(N_{\mathrm{a}}-s_{k-1})\tfrac{d^2}{T}+d\sigma_k^2
\end{aligned} \tag{38}$$

where the identity $\mathbf{B}_i^H\mathbf{B}_j = \left(\mathbf{P}_i^H\otimes\mathbf{h}_i^H\right)\left(\mathbf{P}_j\otimes\mathbf{h}_j\right) = \left(\mathbf{P}_i^H\mathbf{P}_j\right)\otimes\left(\mathbf{h}_i^H\mathbf{h}_j\right) = \left(\mathbf{P}_i^H\mathbf{P}_j\right)\left(\mathbf{h}_i^H\mathbf{h}_j\right)$, the fact $\mathbb{E}\left\{\tilde{\mathbf{z}}_k^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{z}}_k\right\} = \sigma_k^2\,\mathrm{Tr}\left\{\mathbb{E}\left\{\left(\mathbf{P}_j^H\mathbf{P}_j\right)\left(\mathbf{h}_j^H\mathbf{h}_j\right)\right\}\right\}$ and approximation $\mathbb{E}\left\{\frac{\mathbf{h}_j\mathbf{h}_j^H}{\mathbf{h}_j^H\mathbf{h}_j}\right\}\approx\frac{I_M}{M}$ have been used for the above derivations. $\tilde{\mathbf{z}}_k$ is assumed to be vector with i.i.d. variables. By the assumption that entries of precoding matrices are 0-mean and $\frac{1}{T}$-variance i.i.d. Gaussian variables, $\mathbb{E}\left\{\mathbf{s}_i^H\mathbf{P}_i^H\mathbf{P}_j\mathbf{P}_j^H\mathbf{P}_i\mathbf{s}_i\right\} = \mathbb{E}\left\{\sum_{m=1}^d\sum_{n=1}^d\sum_{r=1}^d\mathbf{s}_{i,m}^H\mathbf{P}_{i:m}^H\mathbf{P}_{j:r}\mathbf{P}_{j:r}^H\mathbf{P}_{i:n}\mathbf{s}_{i,n}\right\} = \frac{d^2}{T}$, where $\mathbf{s}_{i,n}$ and $\mathbf{P}_{i:n}$ respectively denote the n-th element of $\mathbf{s}_i$ and n-th column of $\mathbf{P}_i$.

When $j \in I$, we have

$$\begin{aligned}
\mu_{1,k,j} &= \mathbb{E}\left\{\frac{\|\mathbf{B}_j^H\mathbf{r}_{k-1}\|_2^2}{\mathbf{h}_j^H\mathbf{h}_j}\right\} \\
&= \mathbb{E}\left\{\frac{\rho_0\mathbf{s}_j^H\mathbf{B}_j^H\mathbf{B}_j\mathbf{B}_j^H\mathbf{B}_j\mathbf{s}_j+\rho_0\sum_{i\in I\setminus\{\Lambda_{k-1},j\}}\mathbf{s}_i^H\mathbf{B}_i^H\mathbf{B}_j\mathbf{B}_j^H\mathbf{B}_i\mathbf{s}_i+\tilde{\mathbf{z}}_k^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{z}}_k}{\mathbf{h}_j^H\mathbf{h}_j}\right\} \\
&\approx \rho_0\mathbb{E}\left\{\left[\mathbf{s}_j^H\left(\mathbf{P}_j^H\mathbf{P}_j\right)^2\mathbf{s}_j\right]\left(\mathbf{h}_j^H\mathbf{h}_j\right)\right\}+\rho_0(N_{\mathrm{a}}-s_{k-1}-1)\tfrac{d^2}{T}+d\sigma_k^2 \\
&\approx \rho_0\tfrac{d}{T}(d+T-1)\mathbb{E}\left\{\mathbf{h}_j^H\mathbf{h}_j\right\}+\rho_0(N_{\mathrm{a}}-s_{k-1}-1)\tfrac{d^2}{T}+d\sigma_k^2
\end{aligned} \tag{39}$$

where the result of $\mu_{1,k,j}$ in (38) is used, and $\mathbf{s}_j^H \left(\mathbf{P}_j^H \mathbf{P}_j\right)^2 \mathbf{s}_j$ is also expanded for the above evaluation.

APPENDIX B
BRIEF DERIVATIONS OF (20)-(24)

The derivations are also based on the previously made approximations.

It should be noted that, $\mathbb{E}\left\{\frac{\mathbf{h}_i^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_i}{\mathbf{h}_j^H \mathbf{h}_j}\right\} \approx \frac{I_M}{M}\mathbb{E}\left\{\mathbf{h}_i^H \mathbf{h}_i\right\}$, which is I_M when $\mathbf{h}_i^H \mathbf{h}_i$ is not ordered. In the derivations of Appendix A, by evaluating the $\mathbb{E}\left\{\mathbf{h}_i^H \mathbf{h}_i\right\}$ for different active users with order statistics, we can respectively obtain $\mu_{0,k}$ in (20) and $\mu_{1,k}$ in (22).

For $j \notin I$, by the independence of $\mathbf{B}_j$ and $\mathbf{r}_{k-1}$, it has that $\mathbb{E}\left\{\frac{\mathbf{r}_{k-1}^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{r}_{k-1}}{\mathbf{h}_j^H \mathbf{h}_j}\right\} = d\sigma^2_{\mathbf{r}_{k-1}}$. By this, we can assume that each element of $\frac{\mathbf{B}_j^H \mathbf{r}_{k-1}}{\sqrt{\mathbf{h}_j^H \mathbf{h}_j}}$ is a complex Gaussian variable with a mean 0 and a variance $\sigma^2_{\mathbf{r}_{k-1}}$. The variable $\frac{\left(\mathbf{r}_{k-1}^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{r}_{k-1}\right)}{\mathbf{h}_j^H \mathbf{h}_j}$ then follows a chi-squared distribution with $2d$ degrees of freedom. Therefore, we can have the following evaluation

$$\mathbb{E}\left\{\left(\frac{\|\mathbf{B}_j^H \mathbf{r}_{k-1}\|_2^2}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} = \mathbb{E}\left\{\left(\frac{\mathbf{r}_{k-1}^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{r}_{k-1}}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} \approx d(d+1)\sigma^4_{\mathbf{r}_{k-1}} \tag{40}$$

which gives the result in (21).

Define $\tilde{\mathbf{r}}_k := \mathbf{r}_{k-1} - \sqrt{\rho_0}\mathbf{B}_k \mathbf{s}_k$, each of whose zero-mean elements has a variance given by (25). Then for $j = k \in I$, we have

$$\begin{aligned}
\mathbb{E}\left\{\left(\frac{\|\mathbf{B}_j^H \mathbf{r}_{k-1}\|_2^2}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} &= \mathbb{E}\left\{\frac{\left(\rho_0 \mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j + 2\sqrt{\rho_0} Re\left\{\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k\right\} + \tilde{\mathbf{r}}_k^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k\right)^2}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\} \\
&= \rho_0^2 \mathbb{E}\left\{\left(\frac{\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} + 2\rho_0 \mathbb{E}\left\{\frac{\left|\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k\right|^2}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\} \\
&\quad + \mathbb{E}\left\{\left(\frac{\tilde{\mathbf{r}}_k^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} + 2\rho_0 \mathbb{E}\left\{\frac{\left(\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j\right)\left(\tilde{\mathbf{r}}_k^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k\right)}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\}.
\end{aligned} \tag{41}$$

With the same approach to evaluate $\mathbb{E}\left\{\frac{\mathbf{s}_i^H \mathbf{B}_i^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_i \mathbf{s}_i}{\mathbf{h}_j^H \mathbf{h}_j}\right\}$ in (38), we have

$$\mathbb{E}\left\{\left(\frac{\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j}{\mathbf{h}_j^H \mathbf{h}_j}\right)^2\right\} \approx \frac{\rho_0^2}{T^2}(d^2+d)(d+T-1)^2 \mathbb{E}\left\{\left(\mathbf{h}_j^H \mathbf{h}_j\right)^2\right\}. \tag{42}$$

Furthermore, we have the following result

$$\begin{aligned}
\mathbb{E}\left\{\frac{\left(\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j\right)\left(\tilde{\mathbf{r}}_k^H \mathbf{B}_j \mathbf{B}_j^H \tilde{\mathbf{r}}_k\right)}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\} &= \mathbb{E}\left\{\frac{\left(\mathbf{s}_j^H \mathbf{B}_j^H \mathbf{B}_j \mathbf{B}_j^H \mathbf{B}_j \mathbf{s}_j\right) \mathrm{Tr}\left\{\mathbf{B}_j^H \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k^H \mathbf{B}_j\right\}}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\} \\
&\approx \sigma^2_{\tilde{\mathbf{r}}_k} \mathbb{E}\left\{\frac{\left(\mathbf{s}_j^H \mathbf{P}_j^H \mathbf{P}_j \mathbf{P}_j^H \mathbf{P}_j \mathbf{s}_j\right)\left(\mathbf{h}_j^H \mathbf{h}_j\right)^2 \mathrm{Tr}\left\{\mathbf{P}_j^H \mathbf{P}_j\right\}\left(\mathbf{h}_j^H \mathbf{h}_j\right)}{\mathbf{h}_j^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_j}\right\} \\
&= d\sigma^2_{\tilde{\mathbf{r}}_k} \mathbb{E}\left\{\mathbf{s}_j^H \mathbf{P}_j^H \mathbf{P}_j \mathbf{P}_j^H \mathbf{P}_j \mathbf{s}_j\right\} \mathbb{E}\left\{\mathbf{h}_j^H \mathbf{h}_j\right\} \\
&\approx \sigma^2_{\tilde{\mathbf{r}}_k} \frac{d^2}{T}(d+T-1)\mathbb{E}\left\{\mathbf{h}_j^H \mathbf{h}_j\right\}.
\end{aligned} \tag{43}$$

Similar to have (40), we obtain

$$\mathbb{E}\left\{\left(\frac{\tilde{\mathbf{r}}_k^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{r}}_k}{\mathbf{h}_j^H\mathbf{h}_j}\right)^2\right\} \approx d(d+1)\sigma_{\tilde{\mathbf{r}}_k}^4. \tag{44}$$

Besides, the following derivation holds for $j \in I$

$$\begin{aligned}\mathbb{E}\left\{\frac{\left|\mathbf{s}_j^H\mathbf{B}_j^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{r}}_k\right|^2}{\mathbf{h}_j^H\mathbf{h}_j\mathbf{h}_j^H\mathbf{h}_j}\right\} &= \mathbb{E}\left\{\frac{\mathbf{s}_j^H\mathbf{B}_j^H\mathbf{B}_j\mathbf{B}_j^H\tilde{\mathbf{r}}_k\tilde{\mathbf{r}}_k^H\mathbf{B}_j\mathbf{B}_j^H\mathbf{B}_j\mathbf{s}_j}{\mathbf{h}_j^H\mathbf{h}_j\mathbf{h}_j^H\mathbf{h}_j}\right\} \\ &\approx \sigma_{\tilde{\mathbf{r}}_k}^2\mathbb{E}\left\{\left[\mathbf{s}_j^H\left(\mathbf{P}_j^H\mathbf{P}_j\right)^3\mathbf{s}_j\right]\left(\mathbf{h}_j^H\mathbf{h}_j\right)\right\} \\ &= d\sigma_{\tilde{\mathbf{r}}_k}^2\ \bar{\lambda}^3\mathbb{E}\left\{\mathbf{h}_j^H\mathbf{h}_j\right\}\end{aligned} \tag{45}$$

where $\bar{\lambda}^3$ is the expected eigenvalue of $\left(\mathbf{P}_j^H\mathbf{P}_j\right)^3$.

Define $\beta := \frac{d}{T} \in (0,1)$, by the asymptotic result of Theorem 2.35 of random matrix theory in [33], which is also a good approximation for reasonably small-scale matrix dimensions, the eigenvalues of $\left(\mathbf{P}_j^H\mathbf{P}_j\right)$ have an empirical distribution

$$f_\beta(x) = \frac{1}{2\pi\beta x}\sqrt{(x-a)(b-x)} \tag{46}$$

for $x \in [a,b]$, where $a = \left(1-\sqrt{\beta}\right)^2$ and $b = \left(1+\sqrt{\beta}\right)^2$.

Then $\bar{\lambda}^3$ can be approximately evaluated as

$$\bar{\lambda}^3 \approx \int_{\left(1-\sqrt{\beta}\right)^2}^{\left(1+\sqrt{\beta}\right)^2} \frac{x^3}{2\pi\beta x}\sqrt{\left[x-\left(1-\sqrt{\beta}\right)^2\right]\left[\left(1+\sqrt{\beta}\right)^2-x\right]}dx = \frac{1}{8}\left(3+\frac{9d}{T}-\frac{2d^2}{T^2}\right). \tag{47}$$

Similarly, by the eigenvalue approach, $\mathbb{E}\left\{\left[\mathbf{s}_j^H\left(\mathbf{P}_j^H\mathbf{P}_j\right)^2\mathbf{s}_j\right]\right\} \approx \frac{d}{T}(d+T)$, which is almost the same as the achieved $\frac{d}{T}(d+T-1)$ in (43).

Finally, by substituting the results from (42) to (47) into (41) and by replacing j with k (indicating the k-th active user), we can have the derived result in (24).

It should be noted that, when $k = N_\mathrm{a}$, the $\mathbb{E}\left\{\left(\frac{\|\mathbf{B}_j^H\mathbf{r}_{k-1}\|_2^2}{\mathbf{h}_j^H\mathbf{h}_j}\right)^2\right\}$ for $j \notin I$ can be more precisely evaluated. As the same to have the final result of (41), the more precise evaluation is obtained by the expansion approach. It produces the following result

$$\begin{aligned}\mathbb{E}\left\{\left(\frac{\|\mathbf{B}_j^H\mathbf{r}_{N_\mathrm{a}-1}\|_2^2}{\mathbf{h}_j^H\mathbf{h}_j}\right)^2\right\} \approx\ & \frac{\rho_0^2 d}{M^2T^3}\left[4(dT+1)+(d^2-d)(dT+d+T)\right]\mathbb{E}\left\{\left(\mathbf{h}_{N_\mathrm{a}}^H\mathbf{h}_{N_\mathrm{a}}\right)^2\right\} \\ & +2\rho_0\frac{d^2}{MT^2}(dT+d+T-1)\mathbb{E}\left\{\mathbf{h}_{N_\mathrm{a}}^H\mathbf{h}_{N_\mathrm{a}}\right\}\sigma_{N_\mathrm{a}}^2 + d(d+1)\sigma_{N_\mathrm{a}}^4.\end{aligned} \tag{48}$$

REFERENCES

[1] R. E. Grinter and L. Palen, "Instant messaging in teen life," in *Proceedings of the ACM conference on Computer supported cooperative work*, pp. 21-30, 2002.

[2] Z. Xiao, L. Guo, and J. Tracey, "Understanding instant messaging traffic characteristics," in *Distributed Computing Systems, IEEE International Conference on*, pp. 51-51, 2007.

[3] X. Chen and D. Guo, "Gaussian many-access channels: definition and symmetric capacity," *IEEE Information Theory Workshop*, pp. 1-5, 2013.

[4] X. Chen and D. Guo, "Many-access channels: the gaussian case with random user activities," *IEEE International Symposium on Information Theory*, pp. 3127-3131, 2014.

[5] L. Zhang and D. Guo, "Virtual full duplex wireless broadcasting via compressed sensing," *Networking, IEEE/ACM Transactions on*, vol. 22, no. 5, pp. 1659-1671, 2014.

[6] H. Nikopour and H. Baligh, "Sparse code multiple access," *Personal, Indoor and Mobile Radio Communications (PIMRC), IEEE International Symposium on*, pp. 332 C 336, 2013.

[7] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," *in Proc. IEEE VTC Fall 2014*, pp. 1C5, 2014.

[8] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communication Magazine*, vol. 53, no. 9, pp. 74-81, 2015.

[9] Z. Yuan, G. Yu, and W. Li, "Multi-user shared access for 5G," *Telecommun. Network Technology*, vol. 5, no. 5, pp. 28C30, 2015.

[10] R. Xie, H. Yin, Z. Wang and X. Chen, "A novel uplink data transmission scheme for small packets in massive MIMO system," *Communications in China (ICCC), 2014 IEEE/CIC International Conference on*, pp. 375-379, 2014.

[11] P. Pirinen, "A brief overview of 5G research activities," in *5G for Ubiquitous Connectivity (5GU), 2014 1st IEEE International Conference on*, pp. 17-22, 2014.

[12] X. He, P. P. C. Lee, L. Pan, et. al., "A panoramic view of 3G data/control-plane traffic: mobile device perspective," *NETWORKING 2012, Springer Berlin Heidelberg*, pp. 318-330, 2012.

[13] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, Feb. 2014.

[14] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40-60, 2013.

[15] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *Communications, IEEE Transactions on*, vol. 59, no. 2, pp. 454-465, 2011.

[16] X, Yao, Y. C. Eldar, and A. Goldsmith. "Reduced-dimension multiuser detection," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3858-3874, 2013.

[17] A. K. Fletcher, S. Rangan, and V. K Goyal, "On-off random access channels: a compressed sensing framework," http://arxiv.org/pdf/0903.1022v2.pdf.

[18] 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, 2013.

[19] R. Mao and H. Li, "A novel multiple access scheme via compressed sensing with random data traffic," *IEEE Journal of Communications and Networks*, vol. 12, no. 4, pp. 308-316, 2010.

[20] B. C. Jung and W. Chang, "A message passing algorithm for compressed sensing in wireless random access networks," *Communications (APCC), Asia-Pacific Conference on*, pp. 463-464, 2013.

[21] H. Li, R. Mao, L. Lai, and R. C. Qiu, "Compressed meter reading for delay-sensitive and secure load report in smart grid," in *IEEE Int. Conf. on Smart Grid Commun*, pp. 114-119, 2010.

[22] J. Luo and D. Guo, "Neighbor discovery in wireless networks using compressed sensing with Reed-Muller codes," in *IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 154-160, 2011.

[23] R. H. Y. Louie, W. Hardjawana, Y. Li, and B. Vucetic, "Distributed multiple-access for wireless communications: compressed sensing with multiple antennas," *IEEE Globecom*, pp. 3622-3627, 2012.

[24] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.

[25] E. J. Candés and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21-30, 2008.

[26] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: from theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053-4085, 2011.

[27] Z. B. Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp.1032-1047, 2011.

[28] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp.1982-2001, 2010.

[29] H. Kwon and B. D. Rao, "On the benefits of the block-sparsity structure in sparse signal recovery," *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 3685-3688, 2012.

[30] H. G. Myung, J. Lim and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *Vehicular Technology Magazine, IEEE*, vol. 1, no. 3, pp. 30-38, 2006.

[31] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," *in 42nd Asilomar Conf. on Signals, Systems and Computers*, pp. 581-587, 2008.

[32] K. Qiu and A. Dogandzic, "Sparse signal reconstruction via ECME hard thresholding," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4655-4666, 2012.

[33] A. M. Tulino and S. Verdú, *Random matrix theory and wireless communications*. Vol. 1. Now Publishers Inc, 2004.

[34] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a DS/CDMA system," *IEEE Journal of Selected Topics in Communications*, vol. 12, no. 5, pp.796-807, 1994.

[35] J. H. Winters, J. Salz and R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *Communications, IEEE Transactions on*, vol. 42, no. 234, pp. 1740-1751, 1994.

[36] J. Alan and Z. Daniel, *Table of integrals, series, and products*, 7nd edition, Academic Press, 2007.