Massive Device Connectivity with Massive MIMO

Liang Liu and Wei Yu
Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON, M5S 3G4, Canada
Emails: lianguot.liu@utoronto.ca, weiyu@comm.utoronto.ca

***Abstract*—This paper studies a single-cell uplink massive device communication scenario in which a large number of single-antenna devices are connected to the base station (BS), but user traffic is sporadic so that at a given coherence interval, only a subset of users are active. For such a system, active user detection and channel estimation are key issues. To accommodate many simultaneously active users, this paper studies an asymptotic regime where the BS is equipped with a large number of antennas. A grant-free two-phase access scheme is adopted where user activity detection and channel estimation are performed in the first phase, and data is transmitted in the second phase. Our main contributions are as follows. First, this paper shows that despite the non-orthogonality of pilot sequences (which is necessary for accommodating a large number of potential devices), in the asymptotic massive multiple-input multiple-output (MIMO) regime, both the missed detection and false alarm probabilities can be made to go to zero by utilizing compressed sensing techniques that exploit sparsity in user activities. Further, this paper shows that despite the guaranteed success in user activity detection, the non-orthogonality of pilot sequences nevertheless can cause significantly larger channel estimation error as compared to the conventional massive MIMO system, thus lowering the overall achievable transmission rate. This paper quantifies the cost due to device detection and channel estimation and illustrates its effect on the optimal pilot length for massive device connectivity.**

I. Introduction

Massive connectivity is a central requirement for future wireless cellular networks in which a large number of devices may be connected to the base station (BS). A key characteristics for device traffic is that device activities are sporadic, so that within any given time only a subset of devices are active. Thus, reliable detection of active devices along with channel estimation are important for system design. This paper studies a two-phase grant-free multiple-access scheme for device communication in which the BS simultaneously identifies the active devices and estimates their channels based on their pilot sequences in the first phase, and data transmissions take place in the second phase. Because of the large number of potentially active devices in the system and limited coherence time, the devices cannot be all assigned orthogonal pilot sequences. The main goal of this paper is to characterize the device activity detection, channel estimation, and data transmission performance of such a multiple-access scenario.

In order to accommodate a large number of simultaneously active devices, this paper further assumes that the BS is equipped with a large number of antennas. Our main contributions are analytic results establishing that in certain massive multiple-input multiple-output (MIMO) regime, where the number of BS antennas, the number of potential devices, and the number of active devices all go to infinity, accurate user activity detection can always be guaranteed in terms of both missed detection and false alarm probabilities; yet the non-orthogonality of pilot sequences nevertheless incurs significant channel estimation error, thereby lowering the overall achievable rate. The analytic results further allow the optimization of the pilot sequence length and show that the pilots need to be longer in the massive device connectivity setting as compared to the conventional massive MIMO system.

The main technique used in this paper is approximate message passing (AMP) [1]–[4] for compressed sensing, in recognition of the fact that device transmissions are sporadic, so one can take advantage of sparse optimization techniques for activity detection. Sparse optimization has been used in the past for the mass connectivity problem, for example, in [5]–[7] which study joint user activity detection and channel estimation in various settings, and in [8], [9] where information theoretic analysis is carried out. The approach in this paper is based on [6], which analytically characterizes the probabilities of false alarm and missed detection by exploiting the state evolution of the AMP algorithm in the single-antenna case. This paper generalizes [6] to the massive MIMO setting in providing asymptotic activity detection performance analysis.

This paper further characterizes the achievable device transmission rates. We show that the state evolution of AMP allows the mean square error (MSE) for channel estimation to be analytically derived, thereby providing a characterization of user achievable rates for massive device connectivity in the asymptotic massive MIMO regime.

II. System Model

Consider the uplink of a single-cell cellular network consisting of N users, denoted by the set $\mathcal{N}=\{1,\cdots,N\}$. The BS is equipped with M antennas; each user is equipped with one antenna. The complex uplink channel vector from user n to the BS is denoted by $\boldsymbol{h}_n\in\mathbb{C}^{M\times 1}$, $n=1,\cdots,N$. This paper adopts a block-fading model, in which the channel coefficients follow independent quasi-static flat fading. Within each coherence block, $\boldsymbol{h}_n$'s remain constant, but they vary independently from block to block. The channel vector $\boldsymbol{h}_n$ is modeled as $\boldsymbol{h}_n=\sqrt{\beta_n}\boldsymbol{g}_n$, $\forall n$, where $\boldsymbol{g}_n\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$ denotes the Rayleigh fading component, and β_n denotes the path-loss and shadowing component, so that $\boldsymbol{h}_n\sim\mathcal{CN}(\boldsymbol{0},\beta_n\boldsymbol{I}_n)$, $\forall n$. The path-loss and shadowing components depend on the user locations and are assumed to be known at the BS.

To model the sporadic nature of user traffic, we assume that the users are synchronized and in each coherence block each user accesses the channel with probability ϵ in an i.i.d. manner. We define the activity indicator for user n in each block as:

$$\alpha_n = \begin{cases} 1, & \text{if user } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \qquad \forall n, \tag{1}$$

so that $\Pr(\alpha_n = 1) = \epsilon$, $\Pr(\alpha_n = 0) = 1 - \epsilon$, $\forall n$. Further, we define the set of active users within a coherence block as

$$\mathcal{K} = \{n : \alpha_n = 1, n = 1, \cdots, N\}. \tag{2}$$

The number of active users is denoted as $K = |\mathcal{K}|$.

This paper adopts a grant-free multiple-access scheme, in which each coherence block of length T symbols is divided into two phases. In the first phase, the active users send their pilot sequences of length L symbols to the BS synchronously, and the BS jointly detects the user activities, i.e., α_n's, as well as the active users' channels, i.e., $\boldsymbol{h}_n$'s, $\forall n \in \mathcal{K}$. In the second phase, the active users send their data to the BS using the remaining $T - L$ symbols, and the BS decodes these messages based on the knowledge of user activities and channels obtained in the first phase.

For the massive connectivity scenario with a large number of potential devices, the length of pilot sequence is typically smaller than the total number of devices, i.e., $L < N$. In this case, it is not possible to assign mutually orthogonal sequences to all the users. This paper assumes that each user n is assigned a unique pilot sequence $\boldsymbol{a}_n = [a_{n,1}, \cdots, a_{n,L}]^T \in \mathbb{C}^{L\times 1}$, whose entries are generated randomly according to an i.i.d. complex Gaussian distribution with zero mean and variance $1/L$, i.e. $a_{n,l} \sim \mathcal{CN}(0, 1/L)$, so that each pilot sequence has unit norm, i.e., $\|\boldsymbol{a}_n\|^2 = 1$, as $L \to \infty$. It is further assumed that the pilot sequences of all the users are known at the BS.

The goal of this paper is to analyze the performance of joint user detection and channel estimation using the above non-orthogonal pilots in Phase I, and subsequently to characterize the user achievable rate in Phase II. To facilitate analysis, we consider certain asymptotic regime where $N \to \infty$, so that $K \to \epsilon N$, and the empirical distribution of $\beta_1, \cdots, \beta_N$ converges to a fixed distribution denoted by p_β. Moreover, to support K active users, the pilot length L for channel estimation and the number of BS antennas M both need to be in the order of K. The goal is to utilize analytic results in the asymptotic regime where N, K, M, L all go to infinity in certain ways to derive analytic insight for practical systems with large but finite system parameters.

III. User Activity Detection and Channel Estimation via AMP

Consider the first phase of massive device transmission in which each user sends its pilot sequence synchronously through the channel. Define ρ^{pilot} as the identical transmit power of the active users in the first transmission phase. The transmit signal of user n can be expressed as $\alpha_n\sqrt{\xi}\boldsymbol{a}_n$, where $\xi = L\rho^{\text{pilot}}$ denotes the total transmit energy of each active user in the first phase. The received signal at the BS is then

$$\boldsymbol{Y} = \sqrt{\xi}\sum_{n\in\mathcal{N}} \alpha_n \boldsymbol{a}_n \boldsymbol{h}_n^T + \boldsymbol{Z}, \tag{3}$$

where $\boldsymbol{Y} \in \mathbb{C}^{L\times M}$ is the matrix of received signals across M antennas over L symbols, and $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_M]$ with $\boldsymbol{z}_m \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $\forall m$, is the additive white Gaussian noise (AWGN) at the BS. Now define $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_N]$. Further, let $\boldsymbol{x}_n = \alpha_n\boldsymbol{h}_n$ and define $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^T$. Then, the training phase can be modeled as the following matrix equation

$$\boldsymbol{Y} = \sqrt{\xi}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{Z}, \tag{4}$$

where the rows of the matrix $\boldsymbol{X}$ follow a Bernoulli Gaussian distribution:

$$p_{\boldsymbol{x}_n} = (1-\epsilon)\delta_0 + \epsilon p_{\boldsymbol{h}_n}, \quad \forall n. \tag{5}$$

Here, δ_0 denotes the point mass measure at zero, and $p_{\boldsymbol{h}_n}$ denotes the distribution of user n's channel $\boldsymbol{h}_n \sim \mathcal{CN}(\boldsymbol{0}, \beta_n\boldsymbol{I})$.

The goal of the BS in Phase I is to detect the user activities and to estimate the user channels by recovering $\boldsymbol{X}$ based on the noisy observation $\boldsymbol{Y}$. As $\boldsymbol{X}$ is row sparse, i.e., many $\boldsymbol{x}_n$'s are zero, such a reconstruction problem is a compressed sensing problem. Further, as the sparsity pattern is sensed at multiple antennas, this is known as a multiple measurement value (MMV) compressed sensing problem.

Among many powerful compressed sensing techniques, this paper adopts a low-complexity AMP algorithm to recover the row-sparse matrix $\boldsymbol{X}$. In the rest of this section, we first briefly review the vector version of the AMP algorithm, then evaluate its asymptotic performance for user activity detection and channel estimation, respectively.

A. Vector AMP Algorithm with MMSE Denoiser

The general form of the vector AMP algorithm proceeds at each iteration as follows [3], [4], [10]. Starting with $\boldsymbol{R}^0 = \boldsymbol{Y}$,

$$\boldsymbol{x}_n^{t+1} = \eta_{t,n}((\boldsymbol{R}^t)^H\boldsymbol{a}_n + \boldsymbol{x}_n^t), \tag{6}$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}^{t+1} + \frac{N}{L}\boldsymbol{R}^t\sum_{n=1}^{N}\frac{\eta'_{t,n}((\boldsymbol{R}^t)^H\boldsymbol{a}_n + \boldsymbol{x}_n^t)}{N}, \tag{7}$$

where t is the index of the iteration, $\boldsymbol{X}^t = [\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_N^t]^T$ is the estimate of $\boldsymbol{X}$ at iteration t, and $\boldsymbol{R}^t = [\boldsymbol{r}_1^t, \cdots, \boldsymbol{r}_L^t]^T \in \mathbb{C}^{L\times M}$ denotes the corresponding residual. The algorithm performs in (6) a matching filtering of the residual for each user n using its pilot sequence, followed by a denoising step using an appropriately designed denoiser $\eta_{t,n}(\cdot) : \mathbb{C}^{M\times 1} \to \mathbb{C}^{M\times 1}$. The residual is then updated in (7), but corrected with a so-called Onsager term involving $\eta'_{t,n}(\cdot)$, the first-order derivative of $\eta_{t,n}(\cdot)$.

A remarkable property of the AMP algorithm is that when applied to the compressed sensing problem with the entries of the sensing matrix $\boldsymbol{A}$ generated from i.i.d. Gaussian distribution, its detection performance in certain asymptotic regime can be accurately predicted by the so-called *state evolution*. The asymptotic regime is when $L, K, N \to \infty$, while their ratios converge to some fixed positive values $N/L \to \omega$ and $K/N \to \epsilon$ with $\omega, \epsilon \in (0, \infty)$, while keeping the total transmit power fixed at ξ. Note that we fix the total transmit power rather than allowing it to scale with L here in this hypothetical asymptotic system in order to carry out the state evolution analysis. (This implies that the per-symbol power goes down to zero.) The analysis is then used to predict the system performance at finite (but large) L, K, N and $\xi = L\rho^{\text{pilot}}$. As shown in [10], this approach is found to corroborate very well with simulation results.

Specifically, let $\beta \sim p_\beta$; define a random vector $\boldsymbol{X}_\beta \in \mathbb{C}^{M\times 1}$ with a distribution $(1-\epsilon)\delta_0 + \epsilon p_{\boldsymbol{h}_\beta}$, where $p_{\boldsymbol{h}_\beta}$ denotes the distribution $\boldsymbol{h}_\beta \sim \mathcal{CN}(\boldsymbol{0}, \beta\boldsymbol{I})$. Let $\boldsymbol{V} \in \mathbb{C}^{M\times 1} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ be independent of $\boldsymbol{X}_\beta$. Define random vectors

$$\hat{\boldsymbol{X}}_{t,\beta} = \boldsymbol{X}_\beta + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{V}. \tag{8}$$

The state evolution is the following recursion [2]–[4], [10]:

$$\boldsymbol{\Sigma}_0 = \frac{\sigma^2}{\xi}\boldsymbol{I} + \omega\mathbb{E}[\boldsymbol{X}_\beta \boldsymbol{X}_\beta^H], \tag{9}$$

and

$$\boldsymbol{\Sigma}_{t+1} = \frac{\sigma^2}{\xi}\boldsymbol{I} + \omega\mathbb{E}\Big[(\eta_{t,\beta}(\hat{\boldsymbol{X}}_{t,\beta}) - \boldsymbol{X}_\beta)(\eta_{t,\beta}(\hat{\boldsymbol{X}}_{t,\beta}) - \boldsymbol{X}_\beta)^H\Big], \tag{10}$$

where $\boldsymbol{\Sigma}_t$ is referred to as the state, and the expectation is over β, $\boldsymbol{X}_\beta$ and $\boldsymbol{V}$. Note that $\eta_{t,n}(\cdot)$ is replaced by $\eta_{t,\beta}(\cdot)$ for convenience.

The state evolution analysis [2]–[4] says that in the vector AMP algorithm, applying the denoiser to $(\boldsymbol{a}_n^H \boldsymbol{R}^t)^H + \boldsymbol{x}_n^t$ as in (6) is statistically equivalent to applying the denoiser to

$$\hat{\boldsymbol{x}}_{t,n} = \boldsymbol{x}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{v}_n = \alpha_n \boldsymbol{h}_n + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{v}_n, \quad \forall n. \tag{11}$$

The key advantage of this equivalent signal model is the decoupling of the estimation between different users, which allows us to design the denoiser $\eta_{t,n}(\cdot)$ based on the above decoupled signal model.

This paper adopts an MMSE denoiser for minimizing the MSE for user detection and channel training. Specifically, in the tth iteration of the AMP algorithm, the MMSE denoiser $\eta_{t,n}(\cdot)$ is set to be the conditional expectation $\mathbb{E}[\boldsymbol{X}_n|\hat{\boldsymbol{X}}_{t,n}]$. This denoiser has been derived in [10], and is shown below in a slightly different form in order to highlight its structural dependence in M:

$$\eta_{t,n}(\hat{\boldsymbol{x}}_{t,n}) = \mathbb{E}[\boldsymbol{X}_n|\hat{\boldsymbol{X}}_{t,n}] = \phi_{t,n}\beta_n(\beta_n\boldsymbol{I} + \boldsymbol{\Sigma}_t)^{-1}\hat{\boldsymbol{x}}_{t,n}, \tag{12}$$

where

$$\phi_{t,n} = \frac{1}{1 + \frac{1-\epsilon}{\epsilon}\exp\left(-\frac{M}{2}\left(\pi_{t,n} - \psi_{t,n}\right)\right)}, \tag{13}$$

$$\pi_{t,n} = \frac{\hat{\boldsymbol{x}}_{t,n}^H(\boldsymbol{\Sigma}_t^{-1} - (\boldsymbol{\Sigma}_t + \beta_n\boldsymbol{I})^{-1})\hat{\boldsymbol{x}}_{t,n}}{M}, \tag{14}$$

$$\psi_{t,n} = \frac{\log\det(\boldsymbol{I} + \beta_n\boldsymbol{\Sigma}_t^{-1})}{M}. \tag{15}$$

It is worth noting that if all the users are active, i.e., $\epsilon = 1$, it then follows that $\phi_{t,n} = 1$, $\forall n$, in which case the denoiser reduces to the widely used linear MMSE channel estimator: $\eta_{t,n}(\hat{\boldsymbol{x}}_{t,n}) = \beta_n(\beta_n\boldsymbol{I} + \boldsymbol{\Sigma}_t)^{-1}\hat{\boldsymbol{x}}_{t,n}$. With unknown user activity, however, the above MMSE denoiser is non-linear.

Next, we observe that when the MMSE denoiser as given in (12) is used, $\boldsymbol{\Sigma}_t$ as defined in (10) is always a diagonal matrix with identical diagonal entries, i.e.,

$$\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}, \quad \forall t \geq 0. \tag{16}$$

Intuitively, this is because the channels across the BS antennas are assumed to be uncorrelated. As a result, the MMSE denoiser given in (12) is reduced to

$$\eta_{t,n}(\hat{\boldsymbol{x}}_{t,n}) = \phi_{t,n}\frac{\beta_n}{\beta_n + \tau_t^2}\hat{\boldsymbol{x}}_{t,n} \tag{17}$$

where $\phi_{t,n}$ is as given in (13), and

$$\pi_{t,n} = \left(\frac{1}{\tau_t^2} - \frac{1}{\tau_t^2 + \beta_n}\right)\frac{\hat{\boldsymbol{x}}_{t,n}^H\hat{\boldsymbol{x}}_{t,n}}{M}, \tag{18}$$

$$\psi_{t,n} = \log\left(1 + \frac{\beta_n}{\tau_t^2}\right). \tag{19}$$

B. User Activity Detection in the Massive MIMO Regime

The structure of the MMSE denoiser (17), (13), (18) and (19) suggests the following user activity detector. Observe that when $M \to \infty$, $\phi_{t,n}$ in (13) reduces to a threshold function:

$$\phi_{t,n} = \begin{cases} 1, & \text{if } \pi_{t,n} > \psi_{t,n}, \\ 0, & \text{if } \pi_{t,n} < \psi_{t,n}, \\ \epsilon, & \text{otherwise}. \end{cases} \tag{20}$$

Thus, after the tth iteration of the AMP, the detector can declare a user as active if $\pi_{t,n}$ is larger than the threshold $\psi_{t,n}$, and as inactive if $\pi_{t,n}$ is smaller than the threshold $\psi_{t,n}$. For such a detector, we can define the missed detection and false alarm probabilities for user n after t iterations as

$$P_{t,n}^{\text{MD}}(M) = \Pr(\pi_{t,n} \leq \psi_{t,n}|\alpha_n \neq 0), \tag{21}$$

and

$$P_{t,n}^{\text{FA}}(M) = \Pr(\pi_{t,n} \geq \psi_{t,n}|\alpha_n = 0), \tag{22}$$

respectively, as functions of M, the number of BS antennas. The following establishes the optimality of this vector AMP based user activity detector in the massive MIMO regime.

Theorem 1: Consider the user activity detector (20) based on vector AMP with MMSE denoiser. In the asymptotic regime where the number of users N, the number of active users K, and the length of the pilot sequences L all go to infinity, while their ratios converge to some fixed positive values, i.e., $N/L \to \omega$ and $K/N \to \epsilon$ with $\omega, \epsilon \in (0, \infty)$, under fixed total transmit power ξ, assuming state evolution equation (9)-(10), for any user n, we have that both the probabilities of false alarm and missed detection go to zero as the number of antennas at the BS goes to infinity, i.e., $\lim\limits_{M\to\infty} P_{t,n}^{\text{FA}}(M) \to 0$ and $\lim\limits_{M\to\infty} P_{t,n}^{\text{MD}}(M) \to 0$, $\forall t, n$.

The intuition behind the proof of Theorem 1 is as follows. As $M \to \infty$, applying the strong law of large numbers to (18), we have

$$\pi_{t,n} \to \begin{cases} \beta_n/\tau_t^2, & \text{if } \alpha_n = 1 \\ \beta_n/(\beta_n + \tau_t^2), & \text{if } \alpha_n = 0 \end{cases} \tag{23}$$

almost surely. Based on the fact that $a > \log(1+a) > \frac{a}{1+a}$ for any $a > 0$, and that β_n/τ_t^2 is lower bounded by a constant strictly greater than zero, we can conclude that the detector based on the comparison between $\pi_{t,n}$ as in (23) and $\psi_{t,n}$ as in (19) is asymptotically always correct.

C. Asymptotic Analysis of Channel Estimation Error

The AMP algorithm directly gives a characterization of channel estimation error. It can be shown that the MSE term in (10) reduces to the following in the massive MIMO regime:

$$\begin{aligned}&\mathbb{E}\left[(\eta_{t,\beta}(\hat{\boldsymbol{X}}_{t,\beta}) - \boldsymbol{X}_\beta)(\eta_{t,\beta}(\hat{\boldsymbol{X}}_{t,\beta}) - \boldsymbol{X}_\beta)^H\right] \\ &= \epsilon\mathbb{E}\left[\frac{\beta\tau_t^2}{\beta+\tau_t^2}\boldsymbol{I}\right] + \mathbb{E}\left[\phi_{t,\beta}(1-\phi_{t,\beta})\frac{\beta^2}{(\beta+\tau_t^2)^2}\hat{\boldsymbol{x}}_{t,\beta}\hat{\boldsymbol{x}}_{t,\beta}^H\right] \\ &\to \epsilon\mathbb{E}\left[\frac{\beta\tau_t^2}{\beta+\tau_t^2}\boldsymbol{I}\right], \qquad \text{as} \quad M\to\infty, \end{aligned} \tag{24}$$

where we note that $\phi_{t,\beta}$ is asymptotically either 0 or 1 as $M \to \infty$. Intuitively, the above MSE can be interpreted as the product of the user activity probability ϵ and the MSE of channel estimation if the user activity is known, $\frac{\beta\tau_t^2}{\beta+\tau_t^2}\boldsymbol{I}$.

Putting (24) into (10), the state evolution reduces to the following scalar equation:

$$\tau_{t+1}^2 = \frac{\sigma^2}{\xi} + \omega\epsilon\mathbb{E}_\beta\left[\frac{\beta\tau_t^2}{\beta+\tau_t^2}\right]. \tag{25}$$

We note the following properties of the above recursion. First, there is an interesting phase transition phenomenon in AMP [11]: the MSE achieved in (24) for the case when $\omega\epsilon < 1$ is significantly smaller than that achieved for the case when $\omega\epsilon > 1$. Thus, to control channel estimation error, we need to have $\omega\epsilon < 1$, i.e., $L > K$. Second, the limit of τ_t^2 as $t \to \infty$ is of particular interest. Assuming $\omega\epsilon < 1$, we can show that there exists a unique fixed-point solution to (25). This unique fixed-point, denoted as τ_∞^2, is the limit of τ_t^2 after AMP converges.

The fixed-point solution gives us an analytical expression for channel estimation error under AMP. Let $\tilde{\boldsymbol{h}}_n$ denote the estimated channel for user n after convergence of the AMP algorithm (6)-(7); let $\Delta\boldsymbol{h}_n = \boldsymbol{h}_n - \tilde{\boldsymbol{h}}_n$ denote the corresponding channel estimation error. The state evolution analysis of AMP implies that if user n is active in one coherence block, $\tilde{\boldsymbol{h}}_n$ must have the same second-order statistics as $\eta_{\infty,n}(\hat{\boldsymbol{x}}_{\infty,n})$, where $\eta_{\infty,n}(\cdot)$ is the converged MMSE denoiser applied on the signal model (11) with $\alpha_n = 1$. When M goes to infinity, the covariance matrices for the estimated channel and the corresponding channel estimation error, after the convergence of AMP, become respectively

$$\mathrm{cov}(\tilde{\boldsymbol{h}}_n, \tilde{\boldsymbol{h}}_n) \to \frac{\beta_n^2}{\beta_n+\tau_\infty^2}\boldsymbol{I}, \tag{26}$$

$$\mathrm{cov}(\Delta\boldsymbol{h}_n, \Delta\boldsymbol{h}_n) \to \frac{\beta_n\tau_\infty^2}{\beta_n+\tau_\infty^2}\boldsymbol{I}. \tag{27}$$

Although τ_∞^2 does not have a closed-form expression, we can characterize its asymptotic value in the high signal-to-noise ratio (SNR) regime. Specifically, suppose that the path-loss variable β is bounded by $\beta \in [\beta_{\min}, \beta_{\max}]$, and further that $\frac{\xi\beta_{\min}}{\sigma^2} \to \infty$. Then, in the regime $\omega\epsilon < 1$, it can be shown that the unique fixed-point solution to (25) approaches

$$\tau_\infty^2 \to \frac{\sigma^2}{\xi(1-\omega\epsilon)}. \tag{28}$$

As the path loss exponent β for each user depends on its distance to the BS, the above condition on $\frac{\xi\beta_{\min}}{\sigma^2}$ implies that the SNR of the farthest user in the cell must be high.

IV. Achievable Rate of Maximal Ratio Combining

We can now characterize the achievable rate for information transmission of Phase II, assuming the user activity detection and channel estimation analysis of Phase I in the previous section. The equivalent baseband signal received of Phase II is

$$\boldsymbol{y}^{\mathrm{data}} = \sum_{n\in\mathcal{K}} \boldsymbol{h}_n\sqrt{\rho^{\mathrm{data}}}s_n + \boldsymbol{z}^{\mathrm{data}}, \tag{29}$$

where $s_n \sim \mathcal{CN}(0,1)$ denotes the transmit symbol of user $n \in \mathcal{K}$, ρ^{data} denotes the identical transmit power of all the users in the data transmission phase, and $\boldsymbol{z}^{\mathrm{data}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ denotes the AWGN at the BS.

Assuming that the BS views the estimated channel as the true channel and applies a linear maximal ratio combining (MRC) beamformer $\boldsymbol{w}_k = \tilde{\boldsymbol{h}}_k$ to the received signal $\boldsymbol{y}^{\mathrm{data}}$ for user message decoding, following standard bounding technique based on the worst case uncorrelated noise [12], the uplink achievable rate of active user k can be written as

$$R_k = \frac{T-L}{T}\mathbb{E}[\log_2(1+\gamma_k)], \quad \forall k \in \mathcal{K}, \tag{30}$$

where the signal-to-interference-plus-noise ratio (SINR) of user k given a channel realization is

$$\gamma_k = \frac{\|\tilde{\boldsymbol{h}}_k\|^4}{\sum\limits_{n\in\mathcal{K}, n\neq k}|\tilde{\boldsymbol{h}}_k^H\tilde{\boldsymbol{h}}_n|^2 + \|\tilde{\boldsymbol{h}}_k\|^2\sum\limits_{n\in\mathcal{K}}\frac{\beta_n\tau_\infty^2}{\beta_n+\tau_\infty^2} + \frac{\sigma^2}{\rho^{\mathrm{data}}}\|\tilde{\boldsymbol{h}}_k\|^2}. \tag{31}$$

To derive analytical results, we further consider an asymptotic massive MIMO regime where both M and K go to infinity, while keeping their ratio fixed, and utilize the channel estimation error characterization of Phase I. This is in fact a different asymptotic regime as in Phase I, but we nevertheless combine the analyses as ultimately both analyses are used to provide performance projection at finite (but large) N, L, K, M.

Theorem 2: Consider an uplink massive MIMO system with M BS antennas serving K users. Suppose that the estimated channels $\tilde{\boldsymbol{h}}_k$ and channel estimation errors $\Delta\boldsymbol{h}_k$ are Gaussian distributed with the covariance matrices (26) and (27). In the asymptotic regime where both K and M go to infinity but $K/M \to \mu$ with $\mu \in (0,\infty)$, the achievable rate for each user, assuming MRC at the BS, is given by (30), where

$$\gamma_k^{\mathrm{MRC}} \to \frac{\beta_k^2}{\mu\mathbb{E}[\beta](\beta_k+\tau_\infty^2)}, \quad \forall k. \tag{32}$$

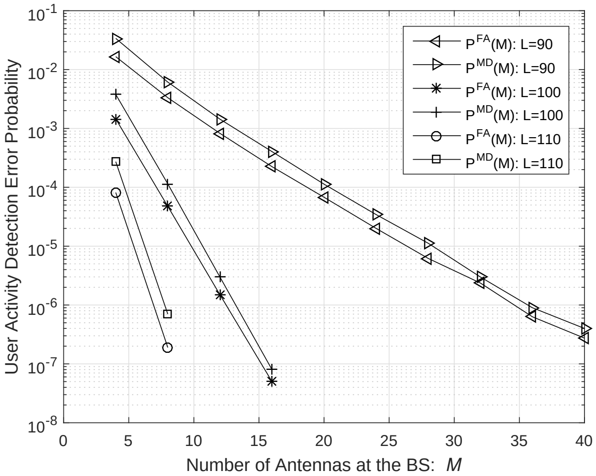

Fig. 1. Probabilities of missed detection and false alarm versus the number of antennas at the BS when $K = 100$ out of $N = 2000$ users are active.

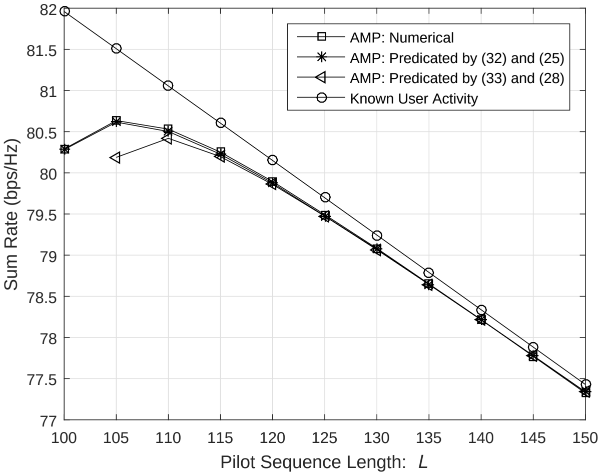

Fig. 2. User sum-rate without and with prior activity information when $K = 100$ out of $N = 2000$ users are active and BS antennas $M = 128$.

If we further assume a high SNR regime with τ_∞ as characterized in (28) and also assume finite but large L, K, M, the user SINR can now be written as

$$\gamma_k \approx \frac{M\beta_k^2}{K\mathbb{E}[\beta]\left(\beta_k + \frac{\sigma^2}{\rho^{\text{pilot}}(L-K)}\right)}, \quad \forall k. \tag{33}$$

By contrasting with the massive MIMO system with known user activity so that orthogonal pilots can be assigned to the K active users for channel training, for which [13]

$$\gamma_k \approx \frac{M\beta_k^2}{K\mathbb{E}[\beta]\left(\beta_k + \frac{\sigma^2}{\rho^{\text{pilot}}L}\right)}, \quad \forall k, \tag{34}$$

it is clear that the cost of user activity detection lies in the increase in the effective channel estimation error, due to the non-orthogonality of the pilot sequences.

V. Numerical Example

Fig. 1 shows the missed detection and false alarm probabilities versus the number of antennas at the BS, for a numerical example with $N = 2000$ users, among which $K = 100$ users are active in each coherence block of $T = 1000$ and the pilot length is $L = 90, 100, 110$, where the users are distributed randomly in distance between 500m and 1km from the BS. The path loss model is given as $\beta_n = -128.1 - 36.7\log_{10}(d_n)$ in dB, with $\rho^{\text{pilot}} = \rho^{\text{data}} = 23$dBm and AWGN at -169dBm/Hz over 1MHz. It is observed that both $P^{\text{MD}}(M)$ and $P^{\text{FA}}(M)$ decrease significantly when L increases from 90 to 110. Moreover, as M increases, both decrease towards zero, which verifies Theorem 1.

Fig. 2 shows the user sum rate versus pilot length with $M = 128$ antennas at the BS. It is observed that the theoretical rate characterization of Theorem 2, where τ_∞^2 is predicted by either the fixed-point solution to (25) or high SINR approximation (28), matches the numerical result well, especially when $L \geq 110$. Moreover, we observe that the optimal pilot length needs to be longer when user activities are not known in advance.

VI. Conclusion

This paper quantifies the cost of user activity detection and channel estimation for massive connectivity. Our main results show that while massive MIMO at the BS essentially guarantees perfect user activity detection, there is nevertheless a marked cost due to the extra interference introduced by imperfect channel estimation when non-orthogonal pilot sequences are used. Such cost can be analytically quantified leveraging results in vector AMP state evolution.

References

[1] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 918, Nov. 2009.

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

[3] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jul. 2011, pp. 2168–2172.

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

[5] X. Xu, X. Rao, and V. K. N. Lau, “Active user detection and channel estimation in uplink CRAN systems,” in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2727–2732.

[6] Z. Chen and W. Yu, “Massive device activity detection by approximate message passing,” in *IEEE Inter. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Mar. 2017.

[7] H. F. Schepker, C. Bockelmann, and A. Dekorsy, “Exploiting sparsity in channel and data estimation for sporadic multi-user communication,” in *Inter. Symp. Wireless Commun. Sys. (ISWCS)*, Aug. 2013, pp. 1–5.

[8] X. Chen, T.-Y. Chen, and D. Guo, “Capacity of Guassian many-access channels,” *IEEE Trans. Inf. Theory*, 2017, to appear.

[9] W. Yu, “On the fundamental limits of massive connectivity,” in *Information Theory and Application (ITA), Workshop*, Feb. 2017.

[10] Z. Chen, F. Sohrabi, and W. Yu, “Sparse activity detection for massive connectivity,” *IEEE Trans. Signal Process.*, 2017, submitted.

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

[12] B. Hassibi and B. M. Hochwald, “How much training is needed in multiple-antenna wireless links?” *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[13] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser MIMO systems,” *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.