Chapter 1

Grant-Free Random Access via Covariance-Based Approach

Ya-Feng Liu,[1*] Wei Yu,[2] Ziyue Wang,[1,3] Zhilin Chen,[2] and Foad Sohrabi[4]

[1] *Institute of Computational Mathematics and Scientific/ Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190, Beijing, China*
[2] *The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, ON M5S 3G4, Toronto, Canada*
[3] *School of Mathematical Sciences, University of Chinese Academy of Sciences, 100049, Beijing, China*
[4] *Nokia Bell Labs, NJ 07974, New Jersey, Murray Hill, USA*

*Corresponding Author: Ya-Feng Liu; yafliu@lsec.cc.ac.cn

Abstract: This chapter presents the theory and algorithms for a covariance-based approach for device activity detection in a grant-free random access protocol. We consider the device activity detection problem in massive multi-input multi-output (MIMO) systems, where active devices transmit their non-orthogonal signature sequences to the base stations (BSs), and the BSs cooperatively detect the active devices based on the received signals. The device activity detection problem can be formulated as a maximum likelihood estimation (MLE) problem. Because the sample covariance matrix of the received signals is a sufficient statistic for the device activity

pattern in the MLE formulation, the approach based on solving the MLE formulation is often called the covariance-based approach. In this chapter, we study the covariance-based approach in both single-cell and multi-cell massive MIMO systems. More specifically, we first present necessary and sufficient conditions on the problem input parameters to ensure a vanishing error probability as the number of antennas at the BS(s) tends to infinity. We then show that the number of active devices that can be detected by the covariance-based approach (in each cell) can scale quadratically as the length of the devices' signature sequence. In addition to the asymptotic performance analysis, we also present efficient coordinate descent (CD) algorithms and their accelerated variants for solving the device activity detection problem. Numerical results verify the accuracy of the asymptotic analysis results and illustrate the efficiency of the CD algorithms.

Keywords: Coordinate descent (CD), covariance-based approach, device activity detection, massive machine-type communication (mMTC), massive multi-input multi-output (MIMO), massive random access, phase transition analysis

1.1. Introduction

Massive machine-type communication (mMTC) is expected to play a crucial role in the fifth-generation (5G) cellular systems and beyond [Bockelmann et al., 2016]. One of the main challenges in mMTC is massive random access, in which a massive number of devices with sporadic data traffic wish to connect to the network in the uplink [Chen et al., 2021]. Conventional cellular systems provide random access for human-type communications by employing a set of orthogonal sequences, from which every active device randomly and independently selects one sequence to transmit as a pilot for requesting access [Dahlman et al., 2013]. However, when the number of active devices is com-

parable to the number of available orthogonal sequences, this uncoordinated random access approach inevitably leads to collisions (with high probability), which will result in a (severe) delay of the data transmission stage because multiple rounds of retransmission signaling are required to resolve the collisions. As such, the above random access scheme is generally not suitable for mMTC.

To reduce the communication latency, grant-free random access schemes are proposed [Liu et al., 2018], where the active devices directly transmit the data signals after transmitting their preassigned non-orthogonal signature sequences without first obtaining permissions from the base-stations (BSs). The BSs first identify the active devices based on the signatures, then decode the data. In this paradigm, no handshake is needed. However, the non-orthogonality of the signature sequences would cause both intra-cell and inter-cell interference which pose unique challenges in the task of device activity detection. This chapter studies the theory and algorithms for device activity detection in the grant-free random access protocol.

There are generally two mathematical optimization formulations of the device activity detection problem. In the first formulation, the device activity detection problem is formulated as a compressed sensing (CS) problem, in which the instantaneous channel state information (CSI) and the device activity are jointly recovered by exploiting the sparsity in the device activity pattern [Senel and Larsson, 2018, Liu and Yu, 2018, Chen et al., 2018]. When the CSI is not needed (e.g., when the data are embedded in the pilot sequence [Senel and Larsson, 2018, Chen et al., 2019]) and the BSs are equipped with a large number of antennas, it is also possible to jointly estimate the device activities (and the channel large-scale fading components) by exploiting the channel sta-

tistical information via maximum likelihood estimation (MLE). This approach is proposed in [Haghighatshoar et al., 2018] and termed as the covariance-based approach because the detection relies on the sample covariance matrix of the received signal. As compared to the CS approach, this covariance-based approach has the advantage of being able to detect many more active devices due to its quadratic scaling law [Fengler et al., 2021, Haghighatshoar et al., 2018, Chen et al., 2022]. It is worthwhile remarking that even in situation where the CSI is needed, the covariance-based approach can still play an important role, e.g., in a three-phase protocol [Kang and Yu, 2022], where in the first phase, the BSs apply the covariance-based approach to detect device activities; in the second phase, the BSs transmit a common feedback message to all the active devices to schedule them in orthogonal transmission slots; and finally, in the third phase, the BSs estimate channels and detect data from the active devices. Since the users are scheduled in orthogonal channels in the three-phase protocol [Kang and Yu, 2022], the channel estimation performance is expected to be better than that of the grant-free protocol [Liu et al., 2018] based on non-orthogonal pilots.

This chapter focuses on the covariance-based approach for the device activity detection problem. We mainly study two questions. The first question is how many active devices can be successfully identified out of a large number of potential devices by the covariance-based approach given a pilot sequence length and assuming a fixed set of non-orthogonal pilot sequences. The answer to the above question leads to a theoretical characterization of the detection performance of the covariance-based approach. The second question is how to efficiently and correctly identify the active devices. To answer the above question, we present several computationally efficient algorithms for solving the

device activity detection problem.

The rest of this chapter is organized as follows. Section 1.2 and Section 1.3 study the theory and algorithms for the covariance-based approach for the device activity detection problem in the single-cell and multi-cell massive multiple-input multiple-output (MIMO) systems, respectively. Section 1.4 discusses some practical issues and presents two interesting extensions. Finally, Section 1.5 concludes this chapter and lists some possible directions for future research.

1.2. Device Activity Detection in Single-Cell Massive MIMO

1.2.1. System Model and Problem Formulation

In this section, we consider an uplink single-cell massive random access scenario with N single-antenna devices communicating with a BS equipped with M antennas. We assume a block fading channel model, i.e., the channel coefficients remain constant for a coherence interval. We also assume that the user traffic is sporadic, i.e., only $K \ll N$ devices are active during each coherence interval. For the purpose of device identification, each device n is preassigned a unique signature sequence $\mathbf{s}_n = [s_{1n}, s_{2n}, \ldots, s_{Ln}]^T \in \mathbb{C}^L$, where L is the sequence length. In the pilot phase, we assume that all the active devices transmit their signature sequences synchronously at the same time. (We consider a more practical asynchronous scenario in Section 1.4.2.) The objective is to detect which subset of devices are active based on the received signal at the BS.

Let $a_n \in \{0, 1\}$ denote the activity of device n in a given coherence interval, i.e., $a_n = 1$ if the device is active and $a_n = 0$ otherwise. The channel vector between the BS and device n is modeled as a random vector $\sqrt{g_n}\mathbf{h}_n$, where $g_n \geq 0$ is the large-scale fading component due to path-loss and shadowing, and $\mathbf{h}_n \in \mathbb{C}^M$ is the Rayleigh fading component following $\mathcal{CN}(\mathbf{0}, \mathbf{I})$. The received signal $\mathbf{Y} \in \mathbb{C}^{L\times M}$ at the BS in the pilot phase can be expressed as

$$\mathbf{Y} = \sum_{n=1}^{N} a_n \mathbf{s}_n \sqrt{g_n}\mathbf{h}_n^T + \mathbf{W} = \mathbf{S}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{H} + \mathbf{W}, \tag{1.1}$$

where $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_N) \in \mathbb{R}^{N\times N}$ with $\gamma_n = a_n g_n$ is a diagonal matrix indicating both the device activity a_n and the large-scale fading component g_n, $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N] \in \mathbb{C}^{L\times N}$ is the signature sequence matrix (which is assumed to be known at the BS), $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]^T \in \mathbb{C}^{N\times M}$ is the channel matrix, and $\mathbf{W} \in \mathbb{C}^{L\times M}$ is the normalized effective independent and identically distributed (i.i.d.) Gaussian noise with variance σ_w^2. We let $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_N]^T \in \mathbb{R}^N$ denote the diagonal entries of $\boldsymbol{\Gamma}$ and use $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ interchangeably throughout this section.

Following the approach proposed in [Fengler et al., 2021, Haghighatshoar et al., 2018], we first use MLE to estimate $\boldsymbol{\gamma}$ from $\mathbf{Y}$, then thereafter obtain the device activity indicator a_n from $\boldsymbol{\gamma}$. The idea is to treat $\boldsymbol{\gamma}$ as a set of deterministic but unknown parameters, and to model $\mathbf{Y}$ as an observation that follows the conditional distribution $p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$ based on the statistics of $\mathbf{h}_n$ and $\mathbf{W}$. To compute the likelihood $p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$, we first observe from (1.1) that given $\boldsymbol{\gamma}$, the columns of $\mathbf{Y}$, denoted by $\mathbf{y}_m \in \mathbb{C}^L$, $1 \leq m \leq M$, are independent due to the i.i.d. channel coefficients over different antennas. In particular, each column $\mathbf{y}_m$ follows a complex Gaussian distribution as $\mathbf{y}_m \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma})$, where

the covariance matrix is given by

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{y}_m \mathbf{y}_m^H\right] = \mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma_w^2 \mathbf{I} = \sum_{n=1}^{N} \gamma_n \mathbf{s}_n \mathbf{s}_n^H + \sigma_w^2 \mathbf{I}. \tag{1.2}$$

Due to the independence of the columns of $\mathbf{Y}$, the likelihood function $p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$ can be computed as

$$\begin{aligned} p(\mathbf{Y} \,|\, \boldsymbol{\gamma}) &= \prod_{m=1}^{M} \frac{1}{|\pi \boldsymbol{\Sigma}|} \exp\left(-\mathbf{y}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{y}_m\right) \\ &= \frac{1}{|\pi \boldsymbol{\Sigma}|^M} \exp\left(-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \mathbf{Y}\mathbf{Y}^H\right)\right), \end{aligned} \tag{1.3}$$

where $|\cdot|$ denotes the determinant of a matrix. The maximization of $\log p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$ is equivalent to the minimization of $-\frac{1}{M} \log p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$, so the MLE problem can be formulated as

$$\min_{\boldsymbol{\gamma}} \quad \log |\boldsymbol{\Sigma}| + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\Sigma}}\right) \tag{1.4a}$$

$$\text{s.t.} \quad \boldsymbol{\gamma} \geq 0, \tag{1.4b}$$

where

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{M} \mathbf{Y}\mathbf{Y}^H = \frac{1}{M} \sum_{m=1}^{M} \mathbf{y}_m \mathbf{y}_m^H \tag{1.5}$$

is the sample covariance matrix of the received signal averaged over different antennas, and the constraint $\boldsymbol{\gamma} \geq 0$ is due to the fact that $\gamma_n = a_n g_n$. Throughout this chapter, we focus on the massive MIMO regime where M is large, which ensures that the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ in (1.5) is a good approximation of the true covariance matrix in (1.2).

We observe from (1.4) that the MLE problem depends on $\mathbf{Y}$ through the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. For this reason, the approach based on solving

the formulation in (1.4) is termed as the covariance-based approach in the literature. As M increases, $\widehat{\mathbf{\Sigma}}$ tends to the true covariance matrix of $\mathbf{Y}$, but the size of the optimization problem does not change. As such, the complexity of solving (1.4) does not scale with M. This is a desirable property especially for the massive MIMO systems.

It is noteworthy to mention an alternative way to model the device activity detection problem is through the following non-negative least squares (NNLS) formulation [Fengler et al., 2021, Haghighatshoar et al., 2018]:

$$\min_{\boldsymbol{\gamma}} \quad \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}} \right\|_F^2 \tag{1.6a}$$

$$\text{s.t.} \quad \boldsymbol{\gamma} \geq 0. \tag{1.6b}$$

The above NNLS formulation tries to match the true covariance matrix $\mathbf{\Sigma}$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ as much as possible under the Frobenius norm metric. The optimization problem (1.6) is convex. However, it has been shown in [Fengler et al., 2021, Chen et al., 2022] that the detection performance of the NNLS formulation (1.6) is much worse than that of the MLE formulation (1.4). Therefore, we focus on the MLE formulation (1.4) in this chapter.

1.2.2. Phase Transition Analysis

In this section, we assume that the MLE problem (1.4) is solved to global optimality and analyze the asymptotic properties of the true MLE solution $\hat{\boldsymbol{\gamma}}^{(M)}$ in the massive MIMO regime where $M \to \infty$. Although the global minimizer of (1.4) may not be easily found in practice due to its nonconvex nature, simulation results show that the analysis still provides useful insights into the per-

formance of practical algorithms for solving the problem (1.4). In Section 1.2.3, we present efficient algorithms for solving the MLE problem (1.4).

For notational clarity, let $\boldsymbol{\gamma}^0$ denote the true parameter to be estimated. We aim to answer the following theoretical question in this section: what are the conditions on the system parameters N, K, and L such that the MLE solution $\hat{\boldsymbol{\gamma}}^{(M)}$ can approach the true parameter $\boldsymbol{\gamma}^0$ as $M \to \infty$? The answer to this question helps identify the desired operating regime in the space of N, K, and L for getting an accurate estimate $\hat{\boldsymbol{\gamma}}^{(M)}$ via MLE with massive MIMO. The phase transition analysis result in this section is mainly from [Chen et al., 2022].

Since the Fisher information matrix $\mathbf{J}(\boldsymbol{\gamma})$ plays a key role in the analysis of MLE, we first provide an explicit expression for $\mathbf{J}(\boldsymbol{\gamma})$.

Theorem 1.1: *Consider the likelihood function in* (1.3)*, where* $\boldsymbol{\gamma}$ *is the parameter to be estimated, and define* $\mathbf{P} = \mathbf{S}^H \left(\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma_w^2\mathbf{I}\right)^{-1}\mathbf{S}$*. The associated* $N \times N$ *Fisher information matrix of* $\boldsymbol{\gamma}$ *is given by*

$$\mathbf{J}(\boldsymbol{\gamma}) = M\left(\mathbf{P} \odot \mathbf{P}^*\right), \tag{1.7}$$

where $\odot$ *is the element-wise product, and* $(\cdot)^*$ *is the conjugate operation.*

Next, we present a necessary and sufficient condition such that $\hat{\boldsymbol{\gamma}}^{(M)}$ can approach $\boldsymbol{\gamma}^0$ in the large M limit.

Theorem 1.2: *Consider the MLE problem* (1.4) *for device activity detection with given signature sequence matrix* $\mathbf{S} \in \mathbb{C}^{L\times N}$ *and noise variance* σ_w^2*, and let* $\hat{\boldsymbol{\gamma}}^{(M)}$ *be a sequence of solutions of* (1.4) *as* M *increases. Let* $\boldsymbol{\gamma}^0$ *be the true*

parameter whose $N-K$ zero entries are indexed by $\mathcal{I}$, i.e.,

$$\mathcal{I} = \{i \mid \gamma_i^0 = 0\}. \tag{1.8}$$

Define

$$\mathcal{N} = \left\{\mathbf{x} \in \mathbb{R}^N \mid \mathbf{x}^T \mathbf{J}(\boldsymbol{\gamma}^0)\mathbf{x} = 0\right\}, \tag{1.9}$$

$$\mathcal{C} = \left\{\mathbf{x} \in \mathbb{R}^N \mid x_i \geq 0, i \in \mathcal{I}\right\}, \tag{1.10}$$

where x_i is the i-th entry of $\mathbf{x}$. Then a necessary and sufficient condition for the consistency of $\hat{\boldsymbol{\gamma}}^{(M)}$, i.e., $\hat{\boldsymbol{\gamma}}^{(M)} \to \boldsymbol{\gamma}^0$ as $M \to \infty$, is that the intersection of $\mathcal{N}$ and $\mathcal{C}$ is the zero vector, i.e., $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$.

The sets $\mathcal{N}$ and $\mathcal{C}$ in Theorem 1.2 can be interpreted as follows: $\mathcal{N}$ is the null space of $\mathbf{J}(\boldsymbol{\gamma}^0)$, which contains all directions $\mathbf{x}$ from $\boldsymbol{\gamma}^0$ along which the likelihood function stays unchanged, i.e., $p(\mathbf{Y} \,|\, \boldsymbol{\gamma}^0) = p(\mathbf{Y} \,|\, \boldsymbol{\gamma}^0 + t\mathbf{x})$ holds for any sufficiently small positive t and any $\mathbf{x} \in \mathcal{N}$; $\mathcal{C}$ is a cone, which contains vectors whose coordinates indexed by $\mathcal{I}$ are always nonnegative—in other words, directions $\mathbf{x}$ from $\boldsymbol{\gamma}^0$ along which $\boldsymbol{\gamma}^0 + t\mathbf{x} \in [0, +\infty)^N$ holds for any sufficiently small positive t. The condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ guarantees that the likelihood function $p(\mathbf{Y} \,|\, \boldsymbol{\gamma})$ in the feasible neighborhood of $\boldsymbol{\gamma}^0$ is not identical to $p(\mathbf{Y} \,|\, \boldsymbol{\gamma}^0)$, so that the true parameter $\boldsymbol{\gamma}^0$ is uniquely identifiable in the feasible region via the likelihood function maximization. See [Chen et al., 2022, Fig. 1] (and the related discussion) for an illustration of the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$.

Since there is generally no closed-form characterization of $\mathcal{N} \cap \mathcal{C}$, we cannot analytically verify the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ for a given $\boldsymbol{\gamma}^0$ and $\mathbf{J}(\boldsymbol{\gamma}^0)$. However, by noting that $\mathcal{N}$ and $\mathcal{C}$ are both convex sets, we can numerically

test whether the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ holds. By further exploiting the positive semidefiniteness of $\mathbf{J}(\boldsymbol{\gamma}^0)$, the following theorem turns the verification of $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ into a linear program (LP).

Theorem 1.3: *Given* $\mathbf{S}$, σ_w^2, *and* $\boldsymbol{\gamma}^0$, *let* $\mathbf{J}(\boldsymbol{\gamma}^0)$ *be the Fisher information matrix in* (1.7)*; let* $\mathbf{A} \in \mathbb{R}^{(N-K)\times(N-K)}$ *be a submatrix of* $\mathbf{J}(\boldsymbol{\gamma}^0)$ *indexed by* $\mathcal{I}$; *let* $\mathbf{C} \in \mathbb{R}^{K\times K}$ *be a submatrix of* $\mathbf{J}(\boldsymbol{\gamma}^0)$ *indexed by* $\mathcal{I}^c$, *where* $\mathcal{I}^c$ *is the complement of* $\mathcal{I}$ *with respect to* $\{1, 2, \ldots, N\}$; *and let* $\mathbf{B} \in \mathbb{R}^{(N-K)\times K}$ *be a submatrix of* $\mathbf{J}(\boldsymbol{\gamma}^0)$ *with rows and columns indexed by* $\mathcal{I}$ *and* $\mathcal{I}^c$, *respectively. Then the condition* $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ *(in Theorem 1.2) is equivalent to: (i)* $\mathbf{C}$ *is invertible; and (ii) the following problem is feasible*

$$\text{find} \quad \mathbf{x} \tag{1.11a}$$

$$\text{s.t.} \quad (\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)\mathbf{x} > \mathbf{0}, \tag{1.11b}$$

where vector $\mathbf{x} \in \mathbb{R}^{N-K}$.

Theorem 1.3 offers a way of identifying the phase transition of the MLE problem numerically. More specifically, suppose that $\mathbf{S}$ and $\mathcal{I}$ are generated randomly according to some distribution for any fixed N, L, and K (e.g., $\mathbf{S}$ is Gaussian and the elements in $\mathcal{I}$ are uniformly selected from $\{1, 2, \ldots, N\}$). We can then use (1.11) to test the consistency of the MLE solution for each realization of $\mathbf{S}$ and $\mathcal{I}$. This enables us to numerically characterize the region in the space of N, L, and K such that $\hat{\boldsymbol{\gamma}}^{(M)}$ can approach $\boldsymbol{\gamma}^0$ in the large M limit. We present some simulation results on the phase transition of the MLE problem in Section 1.2.4.

The last theorem in this section shows the quadratic scaling law of the

covariance-based approach, i.e., the maximum number of active devices K that can be correctly detected by the covariance-based approach increases quadratically with the length of the signature sequence L. Intuitively, the covariance-based approach tries to match the sample covariance matrix and the true covariance matrix (by taking the derivative of the objective in (1.4) with respect to $\mathbf{\Sigma}$), hence the number of effective observations in the covariance-based approach is in the order of L^2. The technical reason behind this quadratic scaling law is that the set $\mathcal{N}$ defined in (1.9) is equivalent to the null space of the matrix $\widehat{\mathbf{S}} = [\mathbf{s}_1^* \otimes \mathbf{s}_1, \ldots, \mathbf{s}_N^* \otimes \mathbf{s}_N] \in \mathbb{C}^{L^2 \times N}$ (where $\otimes$ is the Kronecker product) [Chen et al., 2022], and the matrix $\widehat{\mathbf{S}}$ enjoys the null space property under such a scaling law.

Theorem 1.4: *Let $\mathbf{S} \in \mathbb{C}^{L \times N}$ be the signature sequence matrix whose columns are uniformly drawn from the sphere of radius $\sqrt{L}$ in an i.i.d. fashion. There exist some constants c_1 and c_2 whose values do not depend on K, L, and N such that if*

$$K \leq c_1 L^2 / \log^2(eN/L^2), \tag{1.12}$$

then $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ (in Theorem 1.2) holds true with probability at least $1 - \exp(-c_2 L)$.

1.2.3. Coordinate Descent Algorithms

The optimization problem (1.4) is not convex due to the fact that $\mathrm{tr}(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}})$ is convex whereas $\log|\mathbf{\Sigma}|$ is concave in $\boldsymbol{\gamma}$. However, various practical algorithms are designed and shown to have excellent performance in terms of computational efficiency and detection error probability for solving problem (1.4). Ex-

amples of these algorithms include coordinate descent (CD) [Haghighatshoar et al., 2018, Chen et al., 2019] and accelerated variants [Wang et al., 2021b, Dong et al., 2022], expectation maximization/minimization (EM) (i.e., sparse Bayesian learning) [Wipf and Rao, 2007], gradient descent [Wang et al., 2021a], sparse iterative covariance-based estimation (SPICE) [Stoica et al., 2011], etc. Among the above algorithms, the CD algorithm that iteratively updates the variable associated with each device is popular for solving the MLE problem in the covariance-based approach. The reason for its popularity is that each of its subproblems (i.e., the optimization of the original objective with respect to only one of the variables) admits a closed-form solution [Haghighatshoar et al., 2018], which makes it easily implementable. In this section, we introduce the CD algorithm and its accelerated variant for solving problem (1.4).

1.2.3.1. Coordinate Descent Algorithm

The basic idea of the CD algorithm for solving problem (1.4) is to update each coordinate of the unknown variable $\boldsymbol{\gamma}$ iteratively (while keeping the others fixed) until convergence. In particular, fixing all the other variables except γ_{i_n}, the problem reduces to a univariate optimization problem, which admits a closed-form solution due to its special structure [Haghighatshoar et al., 2018], shown in lines 5 and 6 of Algorithm 1.

In addition to the coordinate update strategy (i.e., how to update the selected variable), the coordinate selection strategy (i.e., selecting which coordinate to update) also plays a vital role in the CD algorithm. Two commonly used strategies are random permutation (which randomly permutes all coordinates and then updates the coordinate one by one according to the order in the permutation) and random selection (which randomly picks one coordinate

Algorithm 1 Coordinate descent algorithm for solving problem (1.4)

1: Initialize $\boldsymbol{\gamma} = \mathbf{0}$, $\boldsymbol{\Sigma}^{-1} = \sigma_w^{-2}\mathbf{I}$;
2: **repeat** [*one iteration*]
3: Randomly select a permutation $i_1, i_2, \ldots, i_N$ of the coordinate indices $\{1, 2, \ldots, N\}$;
4: **for** $n = 1$ to N **do**
5: $d = \max\left\{\frac{\mathbf{s}_{i_n}^H \boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\mathbf{s}_{i_n} - \mathbf{s}_{i_n}^H \boldsymbol{\Sigma}^{-1}\mathbf{s}_{i_n}}{(\mathbf{s}_{i_n}^H \boldsymbol{\Sigma}^{-1}\mathbf{s}_{i_n})^2}, -\gamma_{i_n}\right\}$;
6: $\gamma_{i_n} \leftarrow \gamma_{i_n} + d$;
7: $\boldsymbol{\Sigma}^{-1} \leftarrow \boldsymbol{\Sigma}^{-1} - \frac{d\,\boldsymbol{\Sigma}^{-1}\mathbf{s}_{i_n}\mathbf{s}_{i_n}^H \boldsymbol{\Sigma}^{-1}}{1 + d\,\mathbf{s}_{i_n}^H \boldsymbol{\Sigma}^{-1}\mathbf{s}_{i_n}}$;
8: **end for**
9: **until** $\|\operatorname{Proj}(\boldsymbol{\gamma} - \nabla f(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2 < \varepsilon$;
10: Output $\boldsymbol{\gamma}$.

from all coordinates at a time). Note that in Algorithm 1, we adopt the random permutation strategy; see line 3 of Algorithm 1. Based on our experience of solving problem (1.4), CD equipped with the random permutation strategy is more efficient than that equipped with the random selection strategy. Due to the randomness in the coordinate selection strategy, the CD algorithm sometimes is called random CD in the literature.

The dominant complexity of (random) CD Algorithm 1 is the matrix-vector multiplications in lines 5–7, whose complexity is $\mathcal{O}(L^2)$. Note that in line 7, a rank-one update of $\boldsymbol{\Sigma}^{-1}$ is used to reduce the computational cost and improve the computational efficiency. The overall complexity of Algorithm 1 is $\mathcal{O}(INL^2)$, where I is the total number of iterations. Because the complexity of the CD algorithm is linear in N and quadratic in L, it is particularly suitable for low-latency mMTC scenarios where N is often large and L is often small.

Once Algorithm 1 returns $\boldsymbol{\gamma}$, in order to do the detection, we still need to employ the element-wise thresholding to determine a_n from γ_n, the n-th entry of $\boldsymbol{\gamma}$, using a threshold l_{th}, i.e., $a_n = 1$ if $\gamma_n \geq l_{th}$ and $a_n = 0$ otherwise. The

probabilities of missed detection and false alarm can be traded off by setting different values for l_{th}.

1.2.3.2. Active Set Coordinate Descent Algorithm

In this section, we present a computationally more efficient active set CD algorithm for solving the device activity detection problem in (1.4). Note that most of the devices are inactive (i.e., $K \ll N$). The basic idea of the active set CD algorithm is to exploit the sparsity in the device activity pattern to avoid unnecessary computations on the inactive devices and improve the computational efficiency of the CD algorithm (i.e., Algorithm 1). In particular, at each iteration, the active set CD algorithm first selects a small subset of all devices, termed as the active set, which contains a number of devices that contribute the most to the deviation from the first-order optimality condition of the optimization problem (1.4), then applies the CD algorithm to update the selected variables in the active set.

We first present the first-order optimality condition of the optimization problem (1.4). Let $f(\boldsymbol{\gamma})$ denote the objective function of problem (1.4). Then, for any $n = 1, 2, \ldots, N$, the gradient of $f(\boldsymbol{\gamma})$ with respect to γ_n is

$$[\nabla f(\boldsymbol{\gamma})]_n = \mathbf{s}_n^H \boldsymbol{\Sigma}^{-1} \mathbf{s}_n - \mathbf{s}_n^H \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1} \mathbf{s}_n.$$

The first-order (necessary) optimality condition of problem (1.4) is

$$[\nabla f(\boldsymbol{\gamma})]_n \begin{cases} = 0, & \text{if } \gamma_n > 0; \\ \geq 0, & \text{if } \gamma_n = 0, \end{cases} \quad \forall\, n, \tag{1.13}$$

which is equivalent to $\text{Proj}(\boldsymbol{\gamma} - \nabla f(\boldsymbol{\gamma})) - \boldsymbol{\gamma} = \mathbf{0}$, where $\text{Proj}(\cdot)$ denotes the

Algorithm 2 Active set CD algorithm for solving problem (1.4)

1: Initialize $\boldsymbol{\gamma}^0 = \mathbf{0}$, $k = 0$, $\boldsymbol{\delta}^0 > \mathbf{0}$, and $\varepsilon > 0$;
2: **repeat** [*one iteration*]
3: Update $\boldsymbol{\delta}^k$;
4: Select the active set $\mathcal{A}^k$ according to (1.14);
5: Apply lines 5–7 of Algorithm 1 to update all coordinates in $\mathcal{A}^k$ *only once* in the order of a random permutation;
6: Set $k \leftarrow k + 1$;
7: **until** $\| \operatorname{Proj}(\boldsymbol{\gamma}^k - \nabla f^k) - \boldsymbol{\gamma}^k \|_2 < \varepsilon$;
8: Output $\boldsymbol{\gamma}^k$.

projection operator (onto the feasible region of the corresponding problem). It can be shown that the complexity of computing $\nabla f(\boldsymbol{\gamma})$ is $\mathcal{O}(NL^2)$.

The selection strategy of the active set $\mathcal{A}^k$ at a given feasible point $\boldsymbol{\gamma}^k$ in [Wang et al., 2021a] is mainly based on the degree of the violation of the first-order optimality condition (1.13), which is given by

$$\mathcal{A}^k = \left\{ n \mid \gamma_n^k > 0 \text{ and } \left| \nabla f_n^k \right| > \delta_n^k \right\} \cup \left\{ n \mid \gamma_n^k = 0 \text{ and } \nabla f_n^k < -\delta_n^k \right\}, \quad (1.14)$$

where ∇f_n^k denotes $\left[\nabla f(\boldsymbol{\gamma}^k)\right]_n$ and $\boldsymbol{\delta}^k \in \mathbb{R}_+^N$ is a threshold vector that changes with iteration. The choice of the threshold vector $\boldsymbol{\delta}^k$ in (1.14) is important in balancing the competing goals of improving the objective function and reducing the computational cost at the k-th iteration. A method that works well in practice is to choose a relatively large $\boldsymbol{\delta}^0$ at first and update $\boldsymbol{\delta}^k$ by a multiplicative factor less than one at each iteration.

The active set CD algorithm for solving problem (1.4) is summarized as Algorithm 2. Note that in line 5 of Algorithm 2, lines 5–7 of Algorithm 1 is used to update all coordinates in the selected active set $\mathcal{A}^k$. It is worth remarking here that it is also possible to choose other algorithms. In other words, the active set strategy can be used to accelerate any algorithm (that

does not properly exploit the sparsity in the device activity pattern) for solving problem (1.4), such as those mentioned at the beginning of Section 1.2.3.

1.2.4. Performance Evaluation

In this section, we present some simulation results to validate the accuracy of the phase transition analysis, and compare the existing algorithms for solving the activity detection problem in (1.4). We use the same system parameters as in [Chen et al., 2022]. More specifically, we consider a single cell of radius 1000 m and the channel path-loss is modeled as $128.1 + 37.6 \log_{10}(d)$. We consider the worst-case scenario that all devices are located in the cell edge such that the large-scale fading components g_n's are the same for all devices. The power spectrum density of the background noise is -169 dBm/Hz over 10 MHz and the transmit power of each device is set to 25 dBm. All signature sequences of length L are uniformly drawn from the sphere of radius $\sqrt{L}$ in an i.i.d. fashion as required in Theorem 1.4.

We solve the LP in (1.11) to numerically test the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ in Theorem 1.2 under a variety of choices of L and K, given $N = 900$ or $N = 3600$. Fig. 1.1 plots the region of $(L^2/N, K/N)$ in which the condition is satisfied or not. The result is obtained based on 100 random realizations of $\mathbf{S}$ and $\boldsymbol{\gamma}^0$ for each given K and L. The error bars indicate the range beyond which either all realizations or no realization satisfy the condition. To validate the prediction by Theorem 1.2, we also run the CD algorithm to solve the MLE problem (1.4) by replacing the sample covariance matrix with the true covariance matrix (implying that $M \to \infty$). We then identify the region of $(L^2/N, K/N)$ in which the active devices can be perfectly detected, thus

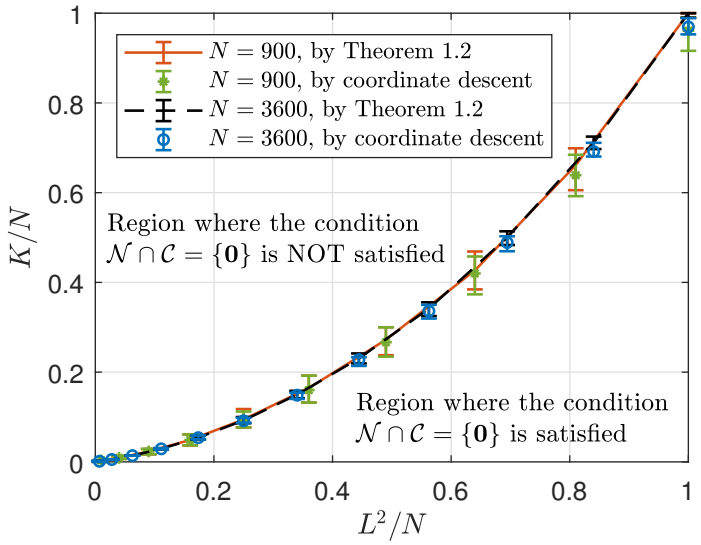

Figure 1.1: Phase transition of the covariance-based approach for device activity detection in the single-cell scenario.

obtaining the phase transition curve empirically. We observe that the curves obtained by Theorem 1.2 and by the CD algorithm match pretty well. We also observe from Fig. 1.1 that K is approximately proportional to L^2, which verifies the scaling law in Theorem 1.4.

Next, we compare the efficiency of existing algorithms (including CD, active set CD, EM [Wipf and Rao, 2007], and SPICE [Stoica et al., 2011]) for solving problem (1.4). Fig. 1.2 plots the decrease of the probability of error as the algorithms run with $M = 128$, $L = 60$, $N = 3000$, and $K = 100$. The result is obtained by averaging over 1000 random realizations of $\mathbf{S}$, $\boldsymbol{\gamma}^0$, and the channel matrix $\mathbf{H}$. For each realization, we record the variable $\boldsymbol{\gamma}$ and calculate the corresponding probability of error when the algorithms run to a fixed moment. We observe from Fig. 1.2 that CD can take less time to achieve a better probability of error than EM and SPICE. It can also be observed from Fig. 1.2 that active set CD is more efficient than CD (due to the active set se-

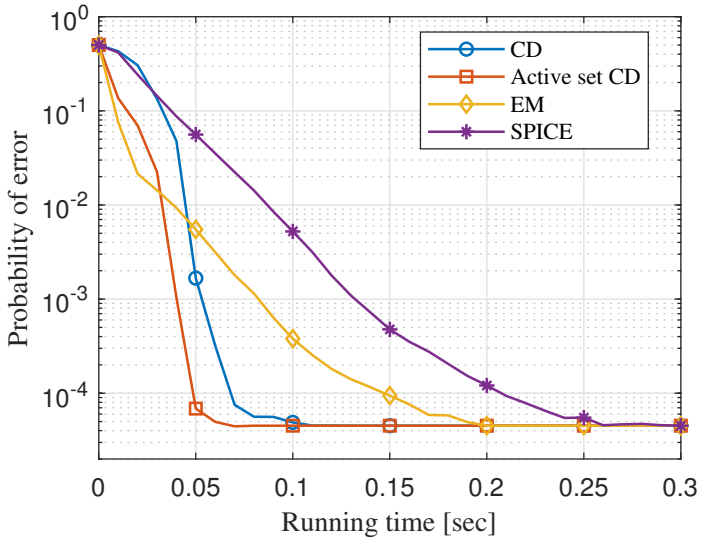

Figure 1.2: Comparison of the probability of error of existing algorithms versus running time in the single-cell scenario.

lection strategy). We also observe from Fig. 1.2 that all of these four algorithms can achieve the same low probability of error if their running time is allowed to be sufficiently long, implying that they converge to the same solution to problem (1.4) (albeit the problem is generally nonconvex).

1.3. Device Activity Detection in Multi-Cell Massive MIMO

1.3.1. System Model and Problem Formulation

In this section, we consider the multi-cell case, i.e., an uplink massive MIMO system consisting of B cells, where each cell contains one BS equipped with M antennas and N single-antenna devices. We assume that a cloud radio access

network (C-RAN) architecture is used for inter-cell interference mitigation, in which all B BSs are connected to a central unit (CU) via fronthaul links such that the signals received at the BSs can be jointly processed at the CU. Similar to the single-cell scenario, we also assume that only $K \ll N$ devices are active in each cell during any coherence interval. For device identification, each device n in cell j is preassigned a unique signature sequence $\mathbf{s}_{jn} \in \mathbb{C}^L$ with L being the sequence length. Let a_{jn} be a binary variable with $a_{jn} = 1$ for active devices and $a_{jn} = 0$ for inactive devices. The channel between device n in cell j and BS b is denoted as $\sqrt{g_{bjn}}\mathbf{h}_{bjn}$, where $g_{bjn} \geq 0$ is the large-scale fading coefficient, and $\mathbf{h}_{bjn} \in \mathbb{C}^M$ is the i.i.d. Rayleigh fading component that follows $\mathcal{CN}(\mathbf{0}, \mathbf{I})$.

Assume that all active devices synchronously transmit their preassigned signature sequences to the BSs in the uplink pilot stage. Then, the received signal at BS b can be expressed as

$$\begin{aligned}\mathbf{Y}_b &= \sum_{n=1}^{N} a_{bn}\mathbf{s}_{bn}g_{bbn}^{\frac{1}{2}}\mathbf{h}_{bbn}^T + \sum_{j\neq b}\sum_{n=1}^{N} a_{jn}\mathbf{s}_{jn}g_{bjn}^{\frac{1}{2}}\mathbf{h}_{bjn}^T + \mathbf{W}_b \\ &= \mathbf{S}_b\mathbf{A}_b\mathbf{G}_{bb}^{\frac{1}{2}}\mathbf{H}_{bb} + \sum_{j\neq b}\mathbf{S}_j\mathbf{A}_j\mathbf{G}_{bj}^{\frac{1}{2}}\mathbf{H}_{bj} + \mathbf{W}_b, \end{aligned} \tag{1.15}$$

where $\mathbf{S}_j = [\mathbf{s}_{j1}, \mathbf{s}_{j2}, \ldots, \mathbf{s}_{jN}] \in \mathbb{C}^{L\times N}$ is the signature sequence matrix of the devices in cell j, $\mathbf{A}_j = \mathrm{diag}(a_{j1}, a_{j2}, \ldots, a_{jN})$ is a diagonal matrix that indicates the activity of all devices in cell j, $\mathbf{G}_{bj} = \mathrm{diag}(g_{bj1}, g_{bj2}, \ldots, g_{bjN})$ contains the large-scale fading components between the devices in cell j and BS b, $\mathbf{H}_{bj} = [\mathbf{h}_{bj1}, \mathbf{h}_{bj2}, \ldots, \mathbf{h}_{bjN}]^T \in \mathbb{C}^{N\times M}$ is the Rayleigh fading channel between the devices in cell j and BS b, and $\mathbf{W}_b$ is the additive Gaussian noise that follows $\mathcal{CN}(\mathbf{0}, \sigma_w^2\mathbf{I})$ with σ_w^2 being the variance of the background noise normalized by the transmit power.

For notational simplicity, let $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_B] \in \mathbb{C}^{L \times BN}$ denote the signature matrix of all devices, and let $\mathbf{G}_b = \mathrm{diag}(\mathbf{G}_{b1}, \mathbf{G}_{b2}, \ldots, \mathbf{G}_{bB}) \in \mathbb{R}^{BN \times BN}$ denote the matrix containing large-scale fading components between all devices and BS b. Let $\mathbf{A} = \mathrm{diag}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_B) \in \mathbb{R}^{BN \times BN}$ be a diagonal matrix that indicates the activity of all devices, and let $\mathbf{a} \in \mathbb{R}^{BN}$ denote its diagonal entries. We use $\mathbf{A}$ and $\mathbf{a}$ interchangeably throughout this section.

The device activity detection problem in the multi-cell massive MIMO scenario is to detect the active devices from the received signals $\mathbf{Y}_b$, $b = 1, 2, \ldots, B$. In this section, we assume that the large-scale fading coefficients are known, i.e., the matrices $\mathbf{G}_b$ for all b are known at the BSs. This assumption holds true if all devices are stationary so that their large-scale fadings are fixed and can be obtained before the detection. In this case, the device activity detection problem is equivalent to estimating the activity indicator vector $\mathbf{a}$.

Note that for each BS b, the Rayleigh fading components and noises are both i.i.d. Gaussian over the antennas. Therefore, for a given $\mathbf{a}$, the columns of the received signal $\mathbf{Y}_b$ in (1.15) denoted by $\mathbf{y}_{bm}$, $m = 1, 2, \ldots, M$, are i.i.d. Gaussian vectors, that is $\mathbf{y}_{bm} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_b)$, where the covariance matrix $\mathbf{\Sigma}_b$ is given by

$$\mathbf{\Sigma}_b = \frac{1}{M}\mathbb{E}\left[\mathbf{Y}_b \mathbf{Y}_b^H\right] = \mathbf{S}\mathbf{G}_b\mathbf{A}\mathbf{S}^H + \sigma_w^2\mathbf{I}. \tag{1.16}$$

Since the received signals $\mathbf{Y}_b$, $b = 1, 2, \ldots, B$, are independent due to the i.i.d. Rayleigh fading channels, the likelihood function can be computed as

$$p(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_B \,|\, \mathbf{a}) = \prod_{b=1}^{B} p(\mathbf{Y}_b \,|\, \mathbf{a}).$$

Hence the MLE problem is equivalent to the minimization of $-\frac{1}{M}\sum_{b=1}^{B} \log p(\mathbf{Y}_b \,|\, \mathbf{a})$.

Thus, the overall problem formulation [Chen et al., 2021] is

$$\min_{\mathbf{a}} \quad \sum_{b=1}^{B} \left(\log |\mathbf{\Sigma}_b| + \mathrm{tr} \left(\mathbf{\Sigma}_b^{-1} \widehat{\mathbf{\Sigma}}_b \right) \right) \tag{1.17a}$$

$$\text{s. t.} \quad a_{bn} \in [0, 1], \, \forall \, b, n, \tag{1.17b}$$

where $\widehat{\mathbf{\Sigma}}_b = \mathbf{Y}_b \mathbf{Y}_b^H / M$ is the sample covariance matrix of the received signals at BS b.

1.3.2. Phase Transition Analysis

In this section, we wish to characterize the asymptotic detection performance of the solution to problem (1.17) as the number of antennas M tends to infinity and in particular, reveal how the number of cells B (and the inter-cell interference) affects the detection performance. To be specific, we want to answer the following question: given the system parameters L, B, and N, how many active devices can be correctly detected by solving the MLE problem (1.17) as $M \to \infty$?

We first present a necessary and sufficient condition for the consistency of the MLE estimator via solving problem (1.17) [Chen et al., 2021], which can be seen as an extension of Theorems 1.1 and 1.2 in the single-cell scenario to the multi-cell scenario.

Theorem 1.5: *Consider the MLE problem* (1.17) *with a given signature sequence matrix* $\mathbf{S}$, *large-scale fading component matrices* $\{\mathbf{G}_b\}$, *and noise variance* σ_w^2. *Let* $\hat{\mathbf{a}}^{(M)}$ *be the solution to* (1.17) *when the number of antennas* M *is given, and let* $\mathbf{a}^0$ *be the true activity indicator vector whose* $B(N-K)$ *zero entries are indexed by* $\mathcal{I}$, *i.e.,* $\mathcal{I} = \{i \mid a_i^0 = 0\}$.

(i) The Fisher information matrix of $\mathbf{a}$ *is given by*

$$\mathbf{J}(\mathbf{a}) = M \sum_{b=1}^{B} \left(\mathbf{P}_b \odot \mathbf{P}_b^* \right), \tag{1.18}$$

where $\mathbf{P}_b = \mathbf{G}_b^{\frac{1}{2}} \mathbf{S}^H \left(\mathbf{S}\mathbf{G}_b\mathbf{A}\mathbf{S}^H + \sigma_w^2 \mathbf{I} \right)^{-1} \mathbf{S}\mathbf{G}_b^{\frac{1}{2}}$.

(ii) Define

$$\mathcal{N} = \left\{ \mathbf{x} \in \mathbb{R}^{BN} \mid \mathbf{x}^T \mathbf{J}(\mathbf{a}^0) \mathbf{x} = 0 \right\}, \tag{1.19}$$

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^{BN} \mid x_i \geq 0 \text{ if } i \in \mathcal{I},\, x_i \leq 0 \text{ if } i \notin \mathcal{I} \right\}. \tag{1.20}$$

Then a necessary and sufficient condition for $\hat{\mathbf{a}}^{(M)} \to \mathbf{a}^0$ *as* $M \to \infty$ *is that the intersection of* $\mathcal{N}$ *and* $\mathcal{C}$ *is the zero vector, i.e.,* $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$.

Notice that in the multi-cell scenario, the large-scale fading coefficients are involved in the definition of the Fisher information matrix in (1.18), which is different from the single-cell scenario. Therefore, the large-scale fading coefficients play a central role in the phase transition analysis in the multi-cell scenario, i.e., the feasible set of the system parameters under which the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ (in Theorem 1.5) holds true. To establish the scaling law result in the multi-cell scenario, we first specify the assumption on the large-scale fading coefficients.

Assumption 1.1: The multi-cell system consists of B hexagonal cells with radius R. The BSs are in the center of the corresponding cells. In this system, the large-scale fading components decrease exponentially with the distance [Rappaport, 2002], i.e.,

$$g_{bjn} = P_0 \left(\frac{d_0}{d_{bjn}} \right)^{\alpha}, \tag{1.21}$$

where P_0 is the received power at the point with distance d_0 from the transmitting antenna, d_{bjn} is the BS-device distance between device n in cell j and BS b, and α is the path-loss exponent.

Now we present an analytic scaling law result by establishing a sufficient condition for $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ (in Theorem 1.5) [Wang et al., 2023].

Theorem 1.6: *Let* $\mathbf{S} \in \mathbb{C}^{L \times BN}$ *be the signature sequence matrix whose columns are uniformly drawn from the sphere of radius* $\sqrt{L}$ *in an i.i.d. fashion. Under Assumption 1.1 with* $\alpha > 2$*, there exist positive constants* c_1 *and* c_2 *independent of system parameters* $K, L, N,$ *and* B*, such that if*

$$K \leq c_1 L^2 / \log^2(eBN/L^2), \tag{1.22}$$

then the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ *(in Theorem 1.5) holds with probability at least* $1 - \exp(-c_2 L)$.

Theorem 1.6 shows that, with a sufficiently large M, the maximum number of active devices that can be correctly detected in each cell by solving the MLE problem (1.17) scales as $\mathcal{O}(L^2)$ as shown in (1.22). The scaling law in (1.22) in the multi-cell scenario is approximately the same as Theorem 1.4 in the single-cell scenario [Fengler et al., 2021, Chen et al., 2022], which provides important insights that solving the MLE problem (1.17) can detect almost as many active devices in each cell in the multi-cell scenario as solving problem (1.4) in the single-cell scenario. Notice that $\alpha > 2$ in Assumption 1.1 holds true for most channel models and application scenarios, see [Rappaport, 2002, Chap. 4]. Therefore, the inter-cell interference is not a limiting factor of the phase transition because B affects K only through $\log(B)$ in (1.22).

1.3.3. Coordinate Descent Algorithms

We now apply the CD algorithm to the multi-cell case. CD is one of the most efficient algorithms for solving problem (1.17) [Chen et al., 2021]. At each iteration, the algorithm randomly permutes the indices of all variables and then updates the variables one by one according to their order in the permutation. For any particularly given coordinate (b, n), the CD algorithm needs to solve the following one-dimensional subproblem

$$\min_{d} \quad \sum_{j=1}^{B} \left(\log\left(1 + d\, g_{jbn} \mathbf{s}_{bn}^H \boldsymbol{\Sigma}_j^{-1} \mathbf{s}_{bn}\right) - \frac{d\, g_{jbn} \mathbf{s}_{bn}^H \boldsymbol{\Sigma}_j^{-1} \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{\Sigma}_j^{-1} \mathbf{s}_{bn}}{1 + d\, g_{jbn} \mathbf{s}_{bn}^H \boldsymbol{\Sigma}_j^{-1} \mathbf{s}_{bn}} \right) \tag{1.23a}$$

$$\text{s.t.} \quad d \in [-a_{bn}, 1 - a_{bn}] \tag{1.23b}$$

in order to possibly update the variable a_{bn}. The closed-form solution for the above problem generally does not exist, which is different from the single-cell case. Fortunately, problem (1.23) can be transformed into a polynomial root-finding problem of degree $2B - 1$, which can further be solved by computing the eigenvalues of the companion matrix formed using the coefficients of the corresponding polynomial function [McNamee, 2007, Chapter 6]. The computational complexity of this approach to solve problem (1.23) is $\mathcal{O}(B^3)$. The CD algorithm is summarized in Algorithm 3. The overall complexity of Algorithm 3 is $\mathcal{O}\left(IBN\left(BL^2 + B^3\right)\right)$, where I is the total number of iterations.

Below we briefly mention two ways of further accelerating Algorithm 3. Notice that, at each iteration, Algorithm 3 treats all coordinates equally and tries to update all of them. However, due to the sparsity of the solution to problem (1.17), there are a lot of coordinates (b, n) for which problem (1.23) has to be solved but a_{bn} does not change, i.e., the corresponding solution to problem (1.23) will be zero. Such computations are unnecessary and slow down

Algorithm 3 Coordinate descent algorithm for solving problem (1.17)

Initialize $\mathbf{a} = \mathbf{0}$, $\mathbf{\Sigma}_b^{-1} = \sigma_w^{-2}\mathbf{I}$, $b = 1, 2, \ldots B$, and $\varepsilon > 0$;
repeat [*one iteration*]
    Randomly select a permutation $\{i_1, i_2, \ldots, i_{BN}\}$ of the coordinate indices $\{1, 2, \ldots, BN\}$;
    for $n = 1$ to BN **do**
        Solve problem (1.23) to obtain d;
        $a_{i_n} \leftarrow a_{i_n} + d$;
        $\mathbf{\Sigma}_b^{-1} \leftarrow \mathbf{\Sigma}_b^{-1} - \frac{d\, g_{bi_n} \mathbf{\Sigma}_b^{-1} \mathbf{s}_{i_n} \mathbf{s}_{i_n}^H \mathbf{\Sigma}_b^{-1}}{1 + d\, g_{bi_n} \mathbf{s}_{i_n}^H \mathbf{\Sigma}_b^{-1} \mathbf{s}_{i_n}}$, $b = 1, \ldots, B$;
    end for
until $\| \operatorname{Proj}(\mathbf{a} - \nabla f(\mathbf{a})) - \mathbf{a} \|_2 < \varepsilon$;
Output $\mathbf{a}$.

Algorithm 3. The active set selection strategy can be used to reduce this kind of unnecessary computations and further improve the computational efficiency of Algorithm 3. This accelerated version of Algorithm 3 is called active set CD in Section 1.3.4. More details along this direction can be found in [Wang et al., 2021b].

Another way of accelerating Algorithm 3 is to inexactly or approximately solve the subproblem in (1.23). More specifically, for a given cell b, instead of considering all B cells as in (1.23), it is reasonable (and desirable) to only consider a cluster of cells that are close to cell b (and neglect those that are far away from cell b). In this case, the degree of the polynomial function associated with the derivative of the objective function of (1.23) will be much smaller, which improves the efficiency of solving the corresponding subproblem. This accelerated version of Algorithm 3 is called clustering-based CD in Section 1.3.4. More details along this direction can be found in [Ganesan et al., 2021].

1.3.4. Performance Evaluation

In this section, we present some simulation results to validate the accuracy of the phase transition analysis, and compare the existing CD types of algorithms for solving the multi-cell device activity detection problem in (1.17). We use the same system parameters as in [Chen et al., 2021]. More specifically, we consider a multi-cell system consisting of hexagonal cells and all potential devices within each cell are uniformly distributed. In the simulations, the radius of each cell is 500 m; the channel path-loss is modeled as $128.1 + 37.6 \log_{10}(d)$ as in Assumption 1.1, where d is the corresponding BS-device distance in km; the transmit power of each device is set to 23 dBm, and the background noise power is -169 dBm/Hz over 10 MHz. All signature sequences of length L are uniformly drawn from the sphere of radius $\sqrt{L}$ in an i.i.d. fashion.

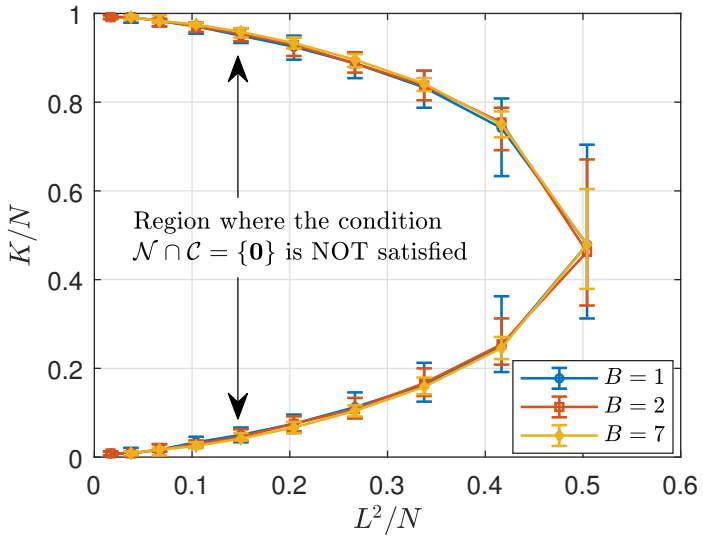

Figure 1.3: Phase transition of the covariance-based approach for device activity detection in the multi-cell scenario with B cells.

We numerically test the condition $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ in Theorem 1.5 under

a variety of choices for L and K, given $N = 240$ and $B = 1, 2, 7$. Fig. 1.3 plots the region of $(L^2/N, K/N)$ in which the condition is satisfied or not. The result is obtained based on 100 random realizations of $\mathbf{S}$ and $\mathbf{a}^0$ for each given K and L. We observe from Fig. 1.3 that: (i) the curves with different B's overlap with each other, implying that the phase transition for $\mathcal{N} \cap \mathcal{C} = \{\mathbf{0}\}$ is almost independent of B (under Assumption 1.1 with the path-loss exponent $\alpha > 2$); the maximum number of identifiable active devices K is approximately proportional to L^2. These observations are consistent with the phase transition analysis in Theorem 1.6.

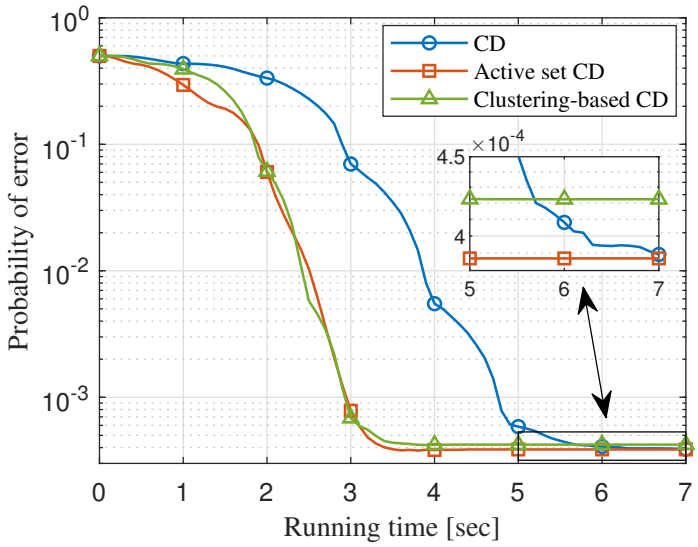

Figure 1.4: Comparison of the probability of error of existing algorithms versus running time in the multi-cell scenario.

Next, we compare the efficiency of CD and its accelerated versions (i.e., active set CD [Wang et al., 2021b] and clustering-based CD [Ganesan et al., 2021]). Fig. 1.4 plots the decrease in the probability of error as the algorithms run with $B = 7$, $M = 128$, $L = 50$, $N = 1000$, and $K = 50$. The result is

obtained by averaging over 1000 Monte-Carlo runs, and the number of clusters of clustering-based CD is chosen to be 2. It can be observed from Fig. 1.4 that active set CD and clustering-based CD are significantly more efficient than CD. It can also be observed from Fig. 1.4 that CD and active set CD can achieve the same low probability of error if their running time is allowed to be sufficiently long, but the probability of error of clustering-based CD (due to the coarse approximation) is slightly worse than those of CD and active set CD. The simulation results show that active set CD has the best efficiency and detection performance among the compared algorithms.

1.4. Practical Issues and Extensions

In this section, we discuss two practical issues in the previous system models and problem formulations and present two interesting extensions.

1.4.1. Joint Device Data and Activity Detection

This section considers a grant-free massive random access scenario for mMTC with very small data payloads as first investigated in [Senel and Larsson, 2018], in which each device maintains a unique set of preassigned 2^J signature sequences. When a device is active, it sends J bits of data by transmitting one sequence from the set. By detecting which sequences are received, the BS acquires both the identity of the active devices as well as the J-bit messages from each of the active devices.

The joint device data and activity detection problem can be formulated as

the MLE problem in both the single-cell and multi-cell scenarios. We illustrate the problem formulation in the single-cell scenario. The considered scenario is the same as that in Section 1.2, except that each device n now has a unique signature sequence set $\mathcal{S}_n = \{\mathbf{s}_{n,1}, \mathbf{s}_{n,2}, \ldots, \mathbf{s}_{n,Q}\}$, where $\mathbf{s}_{n,q} \in \mathbb{C}^L$, $1 \leq q \leq Q \triangleq 2^J$, and L is the signature sequence length. When device n is active and needs to send J bits of data, it selects one sequence from $\mathcal{S}_n$ to transmit.

Using the same technique as in Section 1.2.1, the joint device data and activity detection problem can be formulated as [Chen et al., 2019]

$$\min_{\boldsymbol{\gamma}} \quad \log\left|\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma_w^2\mathbf{I}\right| + \operatorname{tr}\left(\left(\mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma_w^2\mathbf{I}\right)^{-1}\widehat{\boldsymbol{\Sigma}}\right) \tag{1.24a}$$

$$\text{s.t.} \quad \boldsymbol{\gamma} \geq \mathbf{0}, \tag{1.24b}$$

where $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N] \in \mathbb{C}^{L\times NQ}$ with $\mathbf{S}_n = [\mathbf{s}_{n,1}, \mathbf{s}_{n,2}, \ldots, \mathbf{s}_{n,Q}] \in \mathbb{C}^{L\times Q}$ and $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \ldots, \boldsymbol{\gamma}_N^T]^T \in \mathbb{C}^{NQ}$ with $\boldsymbol{\gamma}_n = [\gamma_{n,1}, \gamma_{n,2}, \ldots, \gamma_{n,Q}]^T \in \mathbb{C}^Q$. Problem (1.24) takes the same form as problem (1.4) and hence it can be efficiently solved by the CD algorithm and its accelerated active set variant, i.e., Algorithms 1 and 2.

1.4.2. Device Activity Detection in Asynchronous Systems

This section considers a more practical grant-free massive random access scenario where all active devices asynchronously transmit their preassigned signature sequences to the BS [Liu and Liu, 2021]. We adopt the same notations as in Section 1.2. We introduce a new notation $\tau_n \in \{0, 1, \ldots, \tau_{\max}\}$ to denote the delay of the transmitted packet of each active device $n \in \mathcal{K}$, which means that

each active device $n \in \mathcal{K}$ starts to transmit its signature sequence $\mathbf{s}_n$ at the beginning of the (τ_n+1)-th time slot. In the above, $\tau_{\max}$ is the maximum allowed delay for all the devices and is assumed to be known at the BS. However, the delay of each active device is unknown and needs to be estimated.

Given the delay τ_n, define the effective signature sequence of device n as

$$\bar{\mathbf{s}}_{n,\tau_n} = [\underbrace{0,\ldots,0}_{\tau_n}, \mathbf{s}_n, \underbrace{0,\ldots,0}_{\tau_{\max}-\tau_n}]^T,\ n = 1,\ldots,N. \tag{1.25}$$

In this case, the received signal $\mathbf{Y} \in \mathbb{C}^{(L+\tau_{\max})\times M}$ from time slot 1 to time slot $L + \tau_{\max}$ is expressed as

$$\mathbf{Y} = \sum_{n=1}^{N} a_n \bar{\mathbf{s}}_{n,\tau_n} \sqrt{g_n} \mathbf{h}_n^T + \mathbf{W}. \tag{1.26}$$

Then, the joint device and delay detection problem of estimating $\mathcal{K}$ and associated $\{\tau_k\}_{k\in\mathcal{K}}$ boils down to the sequence detection problem where the sequences are given in (1.25).

Using the same technique as in Section 1.2.1, the joint activity and delay detection problem can be formulated as [Wang et al., 2022]

$$\min_{\boldsymbol{\gamma}} \quad \log|\boldsymbol{\Sigma}| + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}\right) \tag{1.27a}$$

$$\text{s. t.} \quad \boldsymbol{\gamma} \geq 0, \tag{1.27b}$$

$$\|\boldsymbol{\gamma}_n\|_0 \leq 1,\ n = 1,2,\ldots,N, \tag{1.27c}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{M}\mathbf{Y}\mathbf{Y}^H$ is the sample covariance of the received signal in (1.26), $\boldsymbol{\Sigma} = \mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^H + \sigma_w^2\mathbf{I}$, and $\boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \ldots, \boldsymbol{\gamma}_N^T]^T$, and

$$\boldsymbol{\gamma}_n = [\gamma_{n,0}, \gamma_{n,1}, \ldots, \gamma_{n,\tau_{\max}}]^T.$$

In (1.27c), $\|\boldsymbol{\gamma}_n\|_0$ denotes the number of nonzero elements in the vector $\boldsymbol{\gamma}_n$ and the constraint is because there is at most one possible delay for each device. Although $\{\bar{\mathbf{s}}_{n,\tau_n}\}$ are no longer i.i.d., so a phase transition analysis would be challenging, problem (1.27) can still be efficiently solved by the CD algorithm with constraint (1.27c) explicitly enforced during the CD iterations [Wang et al., 2022] and by the penalty based algorithm [Li et al., 2022] which penalizes constraint (1.27c) and solves an equivalent penalty formulation.

1.5. Conclusions

This chapter studies the device activity detection problem for grant-free massive random access with massive MIMO. The covariance-based approach is employed to formulate the device activity detection problem as an MLE problem in both single-cell and multi-cell scenarios. In this chapter, we analyze the asymptotic detection performance of the covariance-based approach as the number of antennas at the BS(s) goes to infinity, including a phase transition analysis. We also present efficient CD types of algorithms for solving the nonconvex detection problem. Finally, we discuss some practical issues in the device activity detection problem and present two extensions of practical interest.

We conclude this chapter with a brief discussion of two future research directions. First, most of the existing phase transition analysis (e.g., Theorems 1.4 and 1.6) crucially relies on the assumption that the signature/pilot sequences of devices are uniformly and randomly drawn from a sphere in an i.i.d. fashion. It would be interesting to extend the current phase transition analysis to more practical ways of generating the signature sequences, for in-

stance, each entry of the device's signature sequence is uniformly drawn from the discrete $\{\pm 1 \pm j\}$, where j is the imaginary unit. Second, the chapter focuses on the massive MIMO system. However, low-resolution ADCs are often employed in the massive MIMO system to reduce hardware cost and power consumption. In this case, the BSs can only observe a coarsely quantized version of the received signals in (1.1) or (1.15). The extension of the covariance-based approach to the massive MIMO system with low-resolution ADCs and the study on how the quantization errors affect the detection performance would be of great interest.

Bibliography

C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, Č. Stefanović, P. Popovski, and A. Dekorsy. Massive machine-type communications in 5G: Physical and MAC-layer solutions. *IEEE Commun. Mag.*, 54(9):59–65, Sept. 2016. doi: 10.1109/MCOM.2016.7565189.

X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober. Massive access for 5G and beyond. *IEEE J. Sel. Areas Commun.*, 39(3): 615–637, Mar. 2021. doi: 10.1109/JSAC.2020.3019724.

Z. Chen, F. Sohrabi, and W. Yu. Sparse activity detection for massive connectivity. *IEEE Trans. Signal Process.*, 66(7):1890–1904, Apr. 2018. doi: 10.1109/TSP.2018.2795540.

Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu. Covariance based joint activity and data detection for massive random access with massive MIMO. In *Proc. IEEE Int. Conf. Commun. (ICC)*, pages 1–6, Shanghai, China, May 2019. doi: 10.1109/ICC.2019.8761672.

Z. Chen, F. Sohrabi, and W. Yu. Sparse activity detection in multi-cell massive MIMO exploiting channel large-scale fading. *IEEE Trans. Signal Process.*, 69:3768–3781, Jun. 2021. doi: 10.1109/TSP.2021.3090679.

Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu. Phase transition analysis for covariance based massive random access with massive MIMO. *IEEE Trans. Inf. Theory*, 68(3):1696–1715, Mar. 2022. doi: 10.1109/TIT.2021.3132397.

E. Dahlman, S. Parkvall, and J. Skold. *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2nd edition, 2013.

J. Dong, J. Zhang, Y. Shi, and J. H. Wang. Faster activity and data detection in massive random access: A multiarmed bandit approach. *IEEE Internet of Things J.*, 9(15):13664–13678, Aug. 2022. doi: 10.1109/JIOT.2022.3142185.

A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire. Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver. *IEEE Trans. Inf. Theory*, 67(5): 2925–2951, May 2021. doi: 10.1109/TIT.2021.3065291.

U. K. Ganesan, E. Björnson, and E. G. Larsson. Clustering-based activity detection algorithms for grant-free random access in cell-free massive MIMO. *IEEE Trans. Commun.*, 69(11):7520–7530, Nov. 2021. doi: 10.1109/TCOMM.2021.3102635.

S. Haghighatshoar, P. Jung, and G. Caire. Improved scaling law for activity detection in massive MIMO systems. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 381–385, Vail, CO, USA, Jun. 2018. doi: 10.1109/ISIT.2018.8437359.

J. Kang and W. Yu. Scheduling versus contention for massive random access in massive MIMO systems. *IEEE Trans. Commun.*, 70(9):5811–5824, Sept. 2022. doi: 10.1109/TCOMM.2022.3190904.

Y. Li, Q. Lin, Y.-F. Liu, B. Ai, and Y.-C. Wu. Asynchronous activity detection for cell-free massive MIMO: From centralized to distributed algorithms. *IEEE Trans. Wireless Commun.*, 2022. doi: 10.1109/TWC.2022.3211967.

L. Liu and Y.-F. Liu. An efficient algorithm for device detection and channel estimation in asynchronous IoT systems. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4815–4819, Toronto, ON, Canada, Jun. 2021. doi: 10.1109/ICASSP39728.2021.9413870.

L. Liu and W. Yu. Massive connectivity with massive MIMO —Part I: Device activity detection and channel estimation. *IEEE Trans. Signal Process.*, 66 (11):2933–2946, Jun. 2018. doi: 10.1109/TSP.2018.2818082.

L. Liu, E. G. Larsson, W. Yu, P. Popovski, Č. Stefanović, and E. De Carvalho. Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things. *IEEE Signal Process. Mag.*, 35(5):88–99, Sept. 2018. doi: 10.1109/MSP.2018.2844952.

J. M. McNamee. *Numerical Methods for Roots of Polynomials-Part I.* The Netherlands: Elsevier, Amsterdam, 2007.

T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 2002.

K. Senel and E. G. Larsson. Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach. *IEEE Trans. Commun.*, 66(12):6164–6175, Dec. 2018. doi: 10.1109/TCOMM.2018.2866559.

P. Stoica, P. Babu, and J. Li. SPICE: A sparse covariance-based estimation method for array processing. *IEEE Trans. Signal Process.*, 59(2):629–638, 2011. doi: 10.1109/TSP.2010.2090525.

Z. Wang, Y.-F. Liu, and L. Liu. Covariance-based joint device activity and delay detection in asynchronous mMTC. *IEEE Signal Process. Lett.*, 29: 538–542, Jan. 2022. doi: 10.1109/LSP.2022.3144853.

Z. Wang, Y.-F. Liu, Z. Wang, and W. Yu. Scaling law analysis for covariance based activity detection in cooperative multi-cell massive MIMO. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes, Greece, Jun. 2023.

Z. Wang, Z. Chen, Y.-F. Liu, F. Sohrabi, and W. Yu. An efficient active set algorithm for covariance based joint data and activity detection for massive random access with massive MIMO. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4840–4844, Toronto, ON, Canada, Jun. 2021. doi: 10.1109/ICASSP39728.2021.9413525.

Z. Wang, Y.-F. Liu, Z. Chen, and W. Yu. Accelerating coordinate descent via active set selection for device activity detection for multi-cell massive random access. In *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pages 366–370, Lucca, Italy, Sept. 2021. doi: 10.1109/SPAWC51858.2021.9593150.

D. P. Wipf and B. D. Rao. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Trans. Signal Process.*, 55(7):3704–3716, Jul. 2007. doi: 10.1109/TSP.2007.894265.