Exact Joint Sparse Frequency Recovery via Optimization Methods

Zai Yang and Lihua Xie, *Fellow, IEEE*[*]

May 2014

Abstract

Frequency recovery/estimation from samples of superimposed sinusoidal signals is a classical problem in statistical signal processing. Its research has been recently advanced by atomic norm techniques which deal with continuous-valued frequencies and completely eliminate basis mismatches of existing compressed sensing methods. This work investigates the frequency recovery problem in the presence of multiple measurement vectors (MMVs) which share the same frequency components, termed as joint sparse frequency recovery and arising naturally from array processing applications. ℓ_0- and ℓ_1-norm-like formulations, referred to as atomic ℓ_0 norm and the atomic norm, are proposed to recover the frequencies and cast as (nonconvex) rank minimization and (convex) semidefinite programming, respectively. Their guarantees for exact recovery are theoretically analyzed which extend existing results with a single measurement vector (SMV) to the MMV case and meanwhile generalize the existing joint sparse compressed sensing framework to the continuous dictionary setting. In particular, given a set of N regularly spaced samples per measurement vector it is shown that the frequencies can be exactly recovered via solving a convex optimization problem once they are separate by at least (approximately) $\frac{4}{N}$. Under the same frequency separation condition, a random subset of N regularly spaced samples of size $O\left(K \log K \log N\right)$ per measurement vector is sufficient to guarantee exact recovery of the K frequencies and missing samples with high probability via similar convex optimization. Extensive numerical simulations are provided to validate our analysis and demonstrate the effectiveness of the proposed method.

Keywords: Array processing, atomic norm, compressed sensing, direction of arrival (DOA) estimation, frequency recovery/estimation, joint sparsity, low rank matrix completion, multiple measurement vector (MMV).

1 Introduction

Suppose that we observe equispaced samples (at the Nyquist sampling rate) of a number of L sinusoidal signals:

$$y_{jt}^{o} = \sum_{k=1}^{K} s_{kt} e^{i2\pi j f_k}, \quad (j,t) \in \boldsymbol{J} \times [L], \tag{1}$$

denoted by matrix $\boldsymbol{Y}^o = \left[y_{jt}^o\right] \in \mathbb{C}^{N\times L}$, on the index set $\boldsymbol{\Omega} \times [L]$, where N is the number of equispaced samples per sinusoidal signal, $\boldsymbol{\Omega} \subset \boldsymbol{J} = \{0, 1, \dots, N-1\}$, and $[L] = \{1, 2, \dots, L\}$. That is, we have L measurement vectors corresponding to the L columns of $\boldsymbol{Y}^o$. Here (j,t) indexes the entries of $\boldsymbol{Y}^o$, $i = \sqrt{-1}$, $f_k \in \mathbb{T} \triangleq [0,1]$ denotes the kth normalized frequency (the starting point 0 and the ending point 1 are identical), $s_{kt} \in \mathbb{C}$ is the (complex) amplitude of the kth frequency component composing the tth sinusoidal signal, and K is the number of the components which is unknown but typically small. Moreover, we let $M = |\boldsymbol{\Omega}| \leq N$ be the sample size

[*] Part of the results of this paper is to be presented at the 2014 IEEE Workshop on Statistical Signal Processing (SSP) [1]. The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: {yangzai, elhxie}@ntu.edu.sg).

of each measurement vector. Following from the literature of spectral analysis [2, 3], the observed data $\boldsymbol{Y}_{\boldsymbol{\Omega}}^{o} \triangleq \left\{y_{jt}^{o}\right\}_{(j,t)\in\boldsymbol{\Omega}\times[L]}$ are called complete if $M = N$ (i.e., $\boldsymbol{\Omega} = \boldsymbol{J}$ or $\boldsymbol{Y}_{\boldsymbol{\Omega}}^{o} = \boldsymbol{Y}^{o}$) and incomplete otherwise. In the latter case, the unobserved data on the complementary index set $\overline{\boldsymbol{\Omega}} \times [L]$, $\overline{\boldsymbol{\Omega}} = \boldsymbol{J}\backslash\boldsymbol{\Omega}$, are called missing data. Let $\mathcal{T} = \{f_1, \dots, f_K\}$ denote the set of the frequencies. The problem concerned in this paper is to recover $\mathcal{T}$ given the observed data, which is referred to as the problem of joint sparse frequency recovery in the sense that the multiple measurement vectors (MMVs) (i.e., the L columns of $\boldsymbol{Y}_{\boldsymbol{\Omega}}^{o}$) share the same K frequencies. After $\mathcal{T}$ is obtained, the amplitudes $\{s_{kt}\}$ and the missing data can be easily obtained by a simple least-squares method.

In the single measurement vector (SMV) case where $L = 1$, this frequency recovery problem is known as line spectral estimation and has wide applications in communications, radar, sonar, seismology, astronomy and so on [2, 3]. With complete data, Prony's method can recover the frequencies once $K \leq \frac{N}{2}$ by root-finding regardless of the values of the frequencies (see, e.g., [4]). However, it does not work in the presence of missing data and is sensitive to measurement noise. With the development of sparse signal representation and later the compressed sensing (CS) concept [5, 6], which studies the recovery of a sparse signal from a number of linear measurements much less than its ambient dimension, sparse methods for frequency recovery have been popular in the past decade. But unfortunately, the frequencies of interest are usually assumed to lie on a fixed grid of the frequency domain $\mathbb{T}$ because the development of CS so far has been focused on signals that can be sparsely represented under a finite discrete dictionary. Under the aforementioned assumption the observation model (1) can be written into an underdetermined system of linear equations and then sparse methods are applied to solve the sparse signal involved whose support corresponds to the frequency set $\mathcal{T}$. Typical sparse methods include combinatorial optimization or ℓ_0 (pseudo-)norm minimization, its convex relaxation or ℓ_1 norm minimization, and greed methods such as orthogonal matching pursuit (OMP) [7, 8]. The ℓ_0 minimization has the best theoretical guarantee, typically ensuring exact recovery once $K \leq \frac{M}{2}$, however, it is NP-hard and cannot be practically solved. In conventional wisdom, the maximal K allowed in the ℓ_1 minimization and OMP for guaranteed exact recovery is inversely proportional to a metric called coherence which, however, increases dramatically as the grid gets fine. On the other hand, due to the grid selection, basis mismatches become a major problem of CS-based methods and many modifications have been proposed to alleviate this drawback (see, e.g., [9–12]). A breakthrough came up recently. Candès and Fernandez-Granda [13] deal directly with the continuous frequency recovery problem and therefore completely eliminate the basis mismatches. In particular, they consider the complete data case and show that the frequencies can be exactly recovered via convex optimization once any two frequencies are separate by at least $\frac{4}{N}$. That means, up to $K = \frac{N}{4}$ frequencies can be recovered within a polynomial time under the frequency separation condition. The convex optimization is based on the so-called total variation norm or atomic norm which extends the ℓ_1 norm to the continuous frequency setting and is formulated as semidefinite programming (SDP) [14, 15]. Inspired by [13], Tang *et al.* [16] study the problem of continuous frequency recovery from partial observations (i.e., incomplete data) based on the atomic norm minimization. Under the same frequency separation condition, they show that a number of $M \geq O(K \log K \log N)$ randomly located measurements is sufficient to guarantee exact recovery with high probability. Several subsequent works on this topic have been done and an incomplete list includes [17–22].

In the MMV case, an example at hand is direction of arrival (DOA) estimation in array processing [2, 23]. In particular, suppose that K farfield, narrowband sources impinge on an array of sensors and one wants to know their directions. The output of the sensor array can be modeled by (1), where each frequency corresponds to one source direction. The sampling index set $\boldsymbol{\Omega}$ therein represents the geometry of the sensor array. To be specific, $\boldsymbol{\Omega} = \boldsymbol{J}$ in the complete data case corresponds to an N-element uniform linear array (ULA) while $\boldsymbol{\Omega} \subsetneq \boldsymbol{J}$ denotes a sparse linear array (SLA). Each measurement vector is composed of the output of the sensor array at one snapshot, and the MMVs are obtained by taking measurements at multiple snapshots, where the source directions are assumed constant during the time window. MUSIC [24] is prominent for this joint sparse frequency recovery problem, however, it is sensitive to correlations of the sources and requires a sufficient number of snapshots such that the sample variance can capture the whole signal subspace. In lieu of sparse methods in the SMV case, joint sparse recovery techniques have been widely studied in the MMV case which, besides the sparse property, exploit the prior knowledge that the MMVs share the same sparsity profile, known as joint sparsity [25–39]. It has been vastly reported that the performance of joint sparse recovery can be generally improved by increasing the number of measurement vectors. Theoretical results include [27] on the ℓ_0 norm under a mild condition, and [30] on greed methods and [34] on the ℓ_1 norm under the assumption that the joint sparse signals are randomly drawn, in particular, the rows of the source signals $[s_{kt}] \in \mathbb{C}^{K\times L}$ in (1) are at general positions. However, it is worth noting that the theoretical guarantee of any joint sparse recovery technique cannot be improved without additional assumptions of the sparse signals of interest. To see this, suppose that all the columns of $[s_{kt}]$ are identical up to some scale factors (any two sources/rows of $[s_{kt}]$ are identical up to a scale factor as well and usually said to be coherent in array processing) and therefore, the MMVs are simply scaled replica of a SMV and do not provide additional information for the recovery. In this respect, the results of [30] and [34] are referred to as the *average case* analysis while those accounting for the aforementioned extreme case are called *worst case* analysis. Similarly to the SMV case, the joint sparse recovery methods rely on discretization/gridding of the frequency domain and suffer from basis mismatches. Unlike the SMV case in which the continuous frequency recovery methods have been recently studied, results are rare on the joint sparse frequency recovery concerned in this paper. To the best of our knowledge, the only known discretization-free/gridless method is introduced in our previous work [40] based on a statistical perspective and utilizing a weighted covariance fitting (WCF) criterion in [41]. In the main context of this paper we will show that this WCF technique is related to the MMV atomic norm method to be proposed in this paper.

In this paper, we present optimization methods for joint sparse frequency recovery and theoretically analyze their performances in the noiseless setting. We firstly consider a continuous ℓ_0 norm formulation, referred to as the atomic ℓ_0 norm, and present its theoretical guarantees for exact frequency recovery, which extends the conventional discrete problem formulation, concept and result to the continuous setting. We next consider its convex relaxation, referred to as the (MMV) atomic norm, and investigate its theoretical guarantees with complete and incomplete data separately. In particular, given the complete data we prove that the frequencies can be exactly recovered by solving a convex optimization problem once they are separate by at least $\frac{4}{N}$. Under the same frequency separation condition, a number of $O\left(K\log K\log N\right)$ randomly located samples per measurement vector is sufficient to guarantee that the frequencies and the missing data can be exactly recovered with high probability by solving an atomic norm minimization problem. As a result, our analysis with the atomic norm extends the results of [13, 16] to the MMV case. Since no or very mild assumptions are made for the source signals in our analysis, the *worst case* theoretical guarantees above do not improve as the number of measurement vectors increases. Moreover, we formulate the atomic ℓ_0 norm and the atomic norm problems as rank minimization and SDP, respectively. Extensive numerical simulations are carried out which validate our theoretical results and further demonstrate that the frequency separation condition required for exact recovery can be relaxed in general as the number of measurement vectors increases but cannot in the worst case. Our results provide theoretical guidance for the practical array processing applications and will inspire further studies such as the average case analysis.

Notations used in this paper are as follows. $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers respectively. $\mathbb{T}$ denotes the unit circle $[0,1]$ by identifying the beginning and ending points. Boldface letters are reserved for vectors and matrices. For an integer N, $[N]\triangleq\{1,\cdots,N\}$. $|\cdot|$ denotes the amplitude of a scalar or cardinality of a set. $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_{\text{F}}$ denote the ℓ_1, ℓ_2 and Frobenius norms respectively. $\boldsymbol{A}^T$ and $\boldsymbol{A}^H$ are the matrix transpose and conjugate transpose of $\boldsymbol{A}$ respectively. x_j is the jth entry of a vector $\boldsymbol{x}$, and $\boldsymbol{A}_j$ denotes the jth row of a matrix $\boldsymbol{A}$. Unless otherwise stated, $\boldsymbol{x}_{\boldsymbol{\Omega}}$ and $\boldsymbol{A}_{\boldsymbol{\Omega}}$ respectively reserve the entries of $\boldsymbol{x}$ and the rows of $\boldsymbol{A}$ in the index set $\boldsymbol{\Omega}$. For a vector $\boldsymbol{x}$, diag $(\boldsymbol{x})$ is a diagonal matrix with $\boldsymbol{x}$ being its diagonal. $\boldsymbol{x}\succeq\boldsymbol{0}$ means $x_j\geq 0$ for all j. rank $(\boldsymbol{A})$ denotes the rank of a matrix $\boldsymbol{A}$ and tr $(\boldsymbol{A})$ the trace. For positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A}\geq\boldsymbol{B}$ means that $\boldsymbol{A}-\boldsymbol{B}$ is positive semidefinite. $\mathbb{E}\left[\cdot\right]$ denotes the expectation and $\mathbb{P}\left(\cdot\right)$ the probability of an event. $\widehat{f}$ is an estimator of f. For notational simplicity, a random variable and its numerical value will not be distinguished.

The rest of the paper is organized as follows. Section 2 presents the atomic ℓ_0 norm method and studies its theoretical guarantees. Section 3 turns to the convex relaxation method and its theoretical results. Section 4 discusses connections of our methods to prior arts. Sections 5 and 6 present proofs of two main theorems in Section 3. Section 7 provides numerical simulations and finally Section 8 concludes this paper.

2 Frequency Recovery via Nonconvex Optimization

2.1 Atomic ℓ_0 Norm and Its Use for Frequency Recovery

We exploit the sparsity to solve the problem of joint sparse frequency recovery. In particular, we seek a set of frequencies of the minimum length among the infinitely many candidates which can express the observed data $\boldsymbol{Y}^o_{\boldsymbol{\Omega}}$. To state it formally, we denote $\boldsymbol{a}\left(f\right) = \left[1, e^{i2\pi f}, \cdots, e^{i2\pi(N-1)f}\right]^T \in \mathbb{C}^N$ and $\boldsymbol{s}_k = \left[s_{k1}, \cdots, s_{kL}\right] \in \mathbb{C}^{1\times L}$. Then (1) can be written as

$$\boldsymbol{Y}^o = \sum_{k=1}^{K} \boldsymbol{a}\left(f_k\right)\boldsymbol{s}_k = \sum_{k=1}^{K} c_k \boldsymbol{a}\left(f_k\right)\boldsymbol{\phi}_k = \sum_{k=1}^{K} c_k \boldsymbol{a}\left(f_k, \boldsymbol{\phi}_k\right), \tag{2}$$

where $c_k = \left\|\boldsymbol{s}_k\right\|_2 > 0$ and $\boldsymbol{\phi}_k = c_k^{-1}\boldsymbol{s}_k$ with $\left\|\boldsymbol{\phi}_k\right\|_2 = 1$. Let $\mathbb{S}^{2L-1} = \left\{\boldsymbol{\phi} : \boldsymbol{\phi} \in \mathbb{C}^{1\times L}, \left\|\boldsymbol{\phi}\right\|_2 = 1\right\}$ denote the unit complex $L-1$-sphere or real $2L-1$-sphere. We define the continuous dictionary or the set of atoms

$$\mathcal{A} \triangleq \left\{\boldsymbol{a}\left(f, \boldsymbol{\phi}\right) = \boldsymbol{a}\left(f\right)\boldsymbol{\phi} : f \in \mathbb{T}, \boldsymbol{\phi} \in \mathbb{S}^{2L-1}\right\}. \tag{3}$$

We see by (2) that $\boldsymbol{Y}^o$ is a linear combination of a number of K atoms in $\mathcal{A}$. In particular, we say that a decomposition of $\boldsymbol{Y}^o$ as in (2) is an atomic decomposition of order K if $c_k > 0$ and the frequencies f_k are distinct. For $\boldsymbol{Y} \in \mathbb{C}^{M\times L}$, we define the smallest number of atoms that can express it as its atomic ℓ_0 (pseudo-)norm:

$$\left\|\boldsymbol{Y}\right\|_{\mathcal{A},0} = \inf\left\{\widehat{K} : \boldsymbol{Y} = \sum_{k=1}^{\widehat{K}} c_k \boldsymbol{a}_k, \boldsymbol{a}_k \in \mathcal{A}, c_k > 0\right\}. \tag{4}$$

So, we propose to recover the frequencies by minimizing the atomic ℓ_0 norm of some $\boldsymbol{Y}$ which is consistent with the observed data on $\boldsymbol{\Omega} \times [L]$, i.e., to solve the following optimization problem:

$$\min_{\boldsymbol{Y}} \left\|\boldsymbol{Y}\right\|_{\mathcal{A},0}, \text{ subject to } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}^o_{\boldsymbol{\Omega}}. \tag{5}$$

For $\boldsymbol{u} \in \mathbb{C}^N$, denote by $T\left(\boldsymbol{u}\right) \in \mathbb{C}^{N\times N}$ a (Hermitian) Toeplitz matrix with

$$T\left(\boldsymbol{u}\right) = \begin{bmatrix} u_1 & u_2 & \cdots & u_M \\ u_2^H & u_1 & \cdots & u_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_M^H & u_{M-1}^H & \cdots & u_1 \end{bmatrix},$$

where u_j is the jth entry of $\boldsymbol{u}$. We provide a finite dimensional formulation of $\left\|\boldsymbol{Y}\right\|_{\mathcal{A},0}$ in the following proposition.

Proposition 2.1. *$\left\|\boldsymbol{Y}\right\|_{\mathcal{A},0}$ defined in (4) equals the optimal value of the following rank minimization problem:*

$$\min_{\boldsymbol{W},\boldsymbol{u},\boldsymbol{U}} \textit{rank}\left(\boldsymbol{U}\right), \textit{ subject to } \boldsymbol{U} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix} \textit{ and } \boldsymbol{U} \geq \boldsymbol{0}. \tag{6}$$

The proof of Proposition 2.1 is based on the classical Vandermonde decomposition lemma stated as follows.

Lemma 2.1 ([2,42]). *Any positive semidefinite Toeplitz matrix $T\left(\boldsymbol{u}\right) \in \mathbb{C}^{N\times N}$ of rank r has an order-r Vandermonde decomposition:*

$$T\left(\boldsymbol{u}\right) = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H,$$

where $\boldsymbol{A} = \left[\boldsymbol{a}\left(f_1\right), \ldots, \boldsymbol{a}\left(f_r\right)\right]$ and $\boldsymbol{P} = \text{diag}\left(p_1, \ldots, p_r\right)$ with $p_j > 0$. Moreover, the decomposition is unique if $r \leq N-1$. ■

Remark 2.1. *The Vandermonde decomposition is not unique if $T(\boldsymbol{u})$ has full rank. In fact, we can arbitrarily choose $f_1 \in \mathbb{T}$ and let $p_1 = \left[\boldsymbol{a}(f_1)^H T(\boldsymbol{u})^{-1} \boldsymbol{a}(f_1)\right]^{-1}$. It follows that the residue $T(\boldsymbol{u}) - p_1 \boldsymbol{a}(f_1)\boldsymbol{a}(f_1)^H$ is of rank $N-1$ and remains a positive semidefinite Toeplitz matrix, which admits a unique Vandermonde decomposition of order $N-1$ and further results in a decomposition of $T(\boldsymbol{u})$ of order N.*

Proof of Proposition 2.1: Let $K = \|\boldsymbol{Y}\|_{\mathcal{A},0}$ and $K^* = \text{rank}(\boldsymbol{U}^*)$, where $(\boldsymbol{W}^*, \boldsymbol{u}^*, \boldsymbol{U}^*)$ denotes an optimal solution to the rank minimization problem (6). We need to show that $K = K^*$. On one hand, since $\boldsymbol{U}^* \geq \boldsymbol{0}$ we have $T(\boldsymbol{u}^*) \geq \boldsymbol{0}$ and $r = \text{rank}(T(\boldsymbol{u}^*)) \leq K^*$. It follows from Lemma 2.1 that $T(\boldsymbol{u}^*) = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H$, where $\boldsymbol{A} = [\boldsymbol{a}(f_1), \cdots, \boldsymbol{a}(f_r)]$, $\boldsymbol{P} = \text{diag}(p_1, \ldots, p_r)$ and $p_j > 0$. Moreover, $\boldsymbol{Y}$ lies in the range space of $T(\boldsymbol{u}^*)$ and therefore, there exists $\boldsymbol{S} \in \mathbb{C}^{r \times L}$ such that $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} = \sum_{j=1}^{r} \boldsymbol{a}(f_j)\boldsymbol{S}_j$, where $\boldsymbol{S}_j$ denotes the jth row of $\boldsymbol{S}$. By the definition of the atomic ℓ_0 norm in (4) we have $K \leq r \leq K^*$.

On the other hand, let $\boldsymbol{Y} = \sum_{j=1}^{K} c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right) = \boldsymbol{A}\boldsymbol{S}$ be an atomic decomposition of $\boldsymbol{Y}$, where $\boldsymbol{A}$ is similarly defined as before and $\boldsymbol{S} = \left[c_1\boldsymbol{\phi}_1^T, \ldots, c_K\boldsymbol{\phi}_K^T\right]^T$. Let $T(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H$ and $\boldsymbol{W} = \boldsymbol{S}^H\boldsymbol{P}^{-1}\boldsymbol{S}$ for arbitrary $p_j > 0$, $j \in [K]$. Then

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T(\boldsymbol{u}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}^H\boldsymbol{P}^{-1} \\ \boldsymbol{A} \end{bmatrix} \boldsymbol{P} \left[\boldsymbol{P}^{-1}\boldsymbol{S} \quad \boldsymbol{A}^H\right] \geq \boldsymbol{0}.$$

As a result, $(\boldsymbol{W}, \boldsymbol{u}, \boldsymbol{U})$ defines a feasible solution of the rank minimization problem (6), and we have that $K^* \leq \text{rank}(\boldsymbol{U}) \leq \text{rank}(\boldsymbol{P}) = K$. ■

Proposition 2.1 presents a rank minimization problem to characterize the atomic ℓ_0 norm. It follows that (5) can be formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{Y},\boldsymbol{W},\boldsymbol{u},\boldsymbol{U}} \text{rank}(\boldsymbol{U}), \\ &\text{subject to } \boldsymbol{U} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T(\boldsymbol{u}) \end{bmatrix}, \boldsymbol{U} \geq \boldsymbol{0}, \text{ and } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}^o_{\boldsymbol{\Omega}}. \end{aligned} \tag{7}$$

In (7) we need to complete a (structured positive semidefinite) low rank matrix $\boldsymbol{U}$ with partial access to its entries. Therefore, we establish a link between the frequency recovery problem and low rank matrix completion [43, 44]. Note that a similar rank minimization problem is presented in [16] in the SMV case, where the rank is put on the matrix $T(\boldsymbol{u})$ rather than the full matrix $\boldsymbol{U}$. Later we will see that (7) has a clearer connection to the convex relaxation method presented in Section 3.

Suppose we could solve the rank minimization problem in (7). According to the proof of Proposition 2.1, we can obtain the solution of the frequencies by the Vandermonde decomposition of $T(\boldsymbol{u}^*)$, where $\boldsymbol{u}^*$ is an optimal solution. While Lemma 2.1 states the existence of the decomposition, readers are referred to [22, Appendix A] for a detailed approach to its realization. We next investigate theoretical guarantees of the proposed atomic ℓ_0 norm method.

2.2 Spark of Continuous Dictionary

To analyze the atomic ℓ_0 norm minimization problem (5) or equivalently (7), we generalize the concept of spark in [45] to the case of continuous dictionary. We define the following continuous dictionary with respect to the index set $\boldsymbol{\Omega}$ similarly to (3):

$$\mathcal{A}^1_{\boldsymbol{\Omega}} \triangleq \left\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f) : f \in \mathbb{T}\right\}. \tag{8}$$

Definition 2.1 (Spark of continuous dictionary)**.** *Given the continuous dictionary $\mathcal{A}^1_{\boldsymbol{\Omega}}$, the quantity spark of $\mathcal{A}^1_{\boldsymbol{\Omega}}$, denoted by spark $\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right)$, is the smallest number of atoms of $\mathcal{A}^1_{\boldsymbol{\Omega}}$ which are linearly dependent.*

In general, it is NP-hard to compute spark $\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right)$ given some sampling index set $\boldsymbol{\Omega}$. Some preliminary results are presented in the following proposition.

Proposition 2.2. *The following results hold about spark* $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right)$ *with* $|\mathbf{\Omega}| = M$*:*

1. $2 \leq$ *spark* $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) \leq M+1$,
2. *spark* $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) = 2$ *if and only if the elements of*

$$\boldsymbol{\mathcal{D}} \triangleq \left\{m_1 - m_2 : m_1, m_2 \in \mathbf{\Omega}, m_1 \geq m_2\right\} \tag{9}$$

   is not coprime, and

3. *spark* $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) = M+1$ *if* $\mathbf{\Omega}$ *consists of* M *consecutive integers.*

Proof. 1) spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) \geq 2$ since all atoms of $\mathcal{A}_{\mathbf{\Omega}}^{1}$ are nonzero. spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) \leq M+1$ because $|\mathbf{\Omega}| = M$ and any $M+1$ atoms of $\mathcal{A}_{\mathbf{\Omega}}^{1}$ are linearly dependent.

2) spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) = 2$ if and only if there exist $f_1 \neq f_2$ such that $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_1\right)$ and $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_2\right)$ are linearly dependent. We next prove the equivalence between the linear dependence and the non-coprimality.

Sufficiency: If $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_1\right)$ and $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_2\right)$ are linearly dependent, i.e., $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_1\right) = ce^{i\theta}\boldsymbol{a}_{\mathbf{\Omega}}\left(f_2\right)$ with $c > 0$ and $\theta \in \mathbb{R}$, then it holds for any $m \in \mathbf{\Omega}$ that

$$2\pi m\left(f_1 - f_2\right) \equiv \theta \mod 2\pi$$

and thus $\left(m_1 - m_2\right)\left(f_1 - f_2\right)$ is an integer for any $m_1, m_2 \in \mathbf{\Omega}$, i.e., $d\left(f_1 - f_2\right)$ is an integer for any $d \in \boldsymbol{\mathcal{D}}$. It follows that $f_1 - f_2$ is a rational number. Let $\left|f_1 - f_2\right| = \frac{b_1}{b_2} < 1$, where b_1, b_2 are coprime positive integers with $b_2 \geq 2$. Consequently, b_2 is a common divisor of the elements of $\boldsymbol{\mathcal{D}}$.

Necessity: Suppose integer $b \geq 2$ is a common divisor of the elements of $\boldsymbol{\mathcal{D}}$. Consider any two $f_1, f_2 \in \mathbb{T}$ satisfying that $f_1 = f_2 + \frac{1}{b}$. Then, for any $m \in \mathbf{\Omega}$,

$$e^{i2\pi m f_1} = e^{i2\pi\frac{m}{b}}e^{i2\pi m f_2} = e^{i2\pi\frac{\Omega_1}{b}}e^{i2\pi m f_2}.$$

The last equality holds since b evenly divides $m - \Omega_1 \in \boldsymbol{\mathcal{D}}$. It follows that $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_1\right)$ and $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_2\right)$ are linearly dependent.

3) If $\mathbf{\Omega}$ consists of M consecutive integers, say, $m, m+1, \cdots, m+M-1$ with $m \geq 0$, then $\boldsymbol{a}_{\mathbf{\Omega}}\left(f_j\right)$, $j \in [M]$, are linearly independent for any M distinct $f_j \in \mathbb{T}$ since the determinant

$$\begin{aligned}
&\left|\left[\boldsymbol{a}_{\mathbf{\Omega}}\left(f_1\right), \ldots, \boldsymbol{a}_{\mathbf{\Omega}}\left(f_M\right)\right]\right| \\
&= e^{i2\pi m\sum_{j=1}^{M} f_j} \prod_{1\leq j\leq l\leq M}\left(e^{i2\pi f_j} - e^{i2\pi f_l}\right) \\
&\neq 0.
\end{aligned} \tag{10}$$

As a result, any M atoms of $\mathcal{A}_{\mathbf{\Omega}}^{1}$ are linearly independent, which together with spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) \leq M+1$ concludes that spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right) = M+1$. ■

Proposition 2.2 presents the range of spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right)$ with respect to the sampling index set $\mathbf{\Omega}$. A necessary and sufficient condition is provided under which spark $\left(\mathcal{A}_{\mathbf{\Omega}}^{1}\right)$ achieves the lower bound 2, where there exist two atoms that are linearly dependent. In this case, in fact, for any atom in $\mathcal{A}_{\mathbf{\Omega}}^{1}$ we can always find one such that they are linearly dependent following from the proof above. However, such $\mathbf{\Omega}$ is rare. To see this, suppose that $\mathbf{\Omega}$ is chosen uniformly at random. Then it satisfies the condition with probability no greater than

$$\frac{\sum_{p\leq\lfloor\frac{N}{M}\rfloor \text{ is prime}} p\binom{\lceil\frac{N}{p}\rceil}{M}}{\binom{N}{M}}, \tag{11}$$

where $\binom{N}{M}$ denotes the number of M-combinations given an N-element set. It is clear that the probability equals 0 when $M > \frac{N}{2}$. Numerically, the probability is less than 1.2×10^{-3}, 1.8×10^{-7}, 3.2×10^{-12} when $N = 100$ and $M = 10,\ 20,\ 30$ respectively. A sufficient (but unnecessary) condition is also provided in the third part under which $\mathcal{A}_{\boldsymbol{\Omega}}^1$ achieves the upper bound $M+1$. We say that $\mathcal{A}_{\boldsymbol{\Omega}}^1$ is a full spark dictionary if spark $\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) = M+1$ following from the discrete setting in [46]. We will see the benefits of a full spark dictionary for frequency recovery in the ensuing subsection.

2.3 Theoretical Guarantees of the Atomic ℓ_0 Norm

We provide theoretical guarantees of the atomic ℓ_0 norm minimization in (5) or the rank minimization in (7) for frequency recovery in this subsection. In particular, we have the following result, which can be considered as a continuous version of [27, Theorem 2.4].

Theorem 2.1. *$\boldsymbol{Y}^o = \sum_{j=1}^K c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$ is the unique optimizer to (5) or (7) if*

$$K < \frac{\textit{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1 + \textit{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right)}{2}. \tag{12}$$

Moreover, the atomic decomposition above is the unique one satisfying that $K = \left\|\boldsymbol{Y}^o\right\|_{\mathcal{A},0}$.

Proof. Suppose that there exists $\widetilde{\boldsymbol{Y}} \neq \boldsymbol{Y}^o$ satisfying that $\widetilde{\boldsymbol{Y}}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ and $\left\|\widetilde{\boldsymbol{Y}}\right\|_{\mathcal{A},0} \leq \left\|\boldsymbol{Y}^o\right\|_{\mathcal{A},0} = K$, and $\widetilde{\boldsymbol{Y}} = \sum_{k=1}^{\widetilde{K}} \widetilde{c}_j \boldsymbol{a}\left(\widetilde{f}_j, \widetilde{\boldsymbol{\phi}}_j\right)$ is an atomic decomposition of order $\widetilde{K} \leq K$. Let $\boldsymbol{A}_1 = \left[\boldsymbol{a}\left(f\right)\right]_{f \in \mathcal{T} \backslash \left\{\widetilde{f}_j\right\}}$ (the matrix consisting of those $\boldsymbol{a}\left(f\right)$, $f \in \mathcal{T} \backslash \left\{\widetilde{f}_j\right\}$), $\boldsymbol{A}_{12} = \left[\boldsymbol{a}\left(f\right)\right]_{f \in \mathcal{T} \cap \left\{\widetilde{f}_j\right\}}$ and $\boldsymbol{A}_2 = \left[\boldsymbol{a}\left(f\right)\right]_{f \in \left\{\widetilde{f}_j\right\} \backslash \mathcal{T}}$. In addition, let $K_{12} = \left|\mathcal{T} \cap \left\{\widetilde{f}_j\right\}\right|$ and $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{A}_{12} & \boldsymbol{A}_2\end{bmatrix}$. Then we have

$$\begin{aligned} \boldsymbol{Y}^o &= \begin{bmatrix}\boldsymbol{A}_1 \ \boldsymbol{A}_{12}\end{bmatrix} \begin{bmatrix}\boldsymbol{S}_1 \\ \boldsymbol{S}_{12}\end{bmatrix}, \\ \widetilde{\boldsymbol{Y}} &= \begin{bmatrix}\boldsymbol{A}_{12} \ \boldsymbol{A}_2\end{bmatrix} \begin{bmatrix}\boldsymbol{S}_{21} \\ \boldsymbol{S}_2\end{bmatrix}, \end{aligned}$$

where $\boldsymbol{S}_1$, $\boldsymbol{S}_{12}$, $\boldsymbol{S}_{21}$ and $\boldsymbol{S}_2$ are matrices of proper dimensions. It follows that

$$\boldsymbol{Y}^o - \widetilde{\boldsymbol{Y}} = \boldsymbol{A} \begin{bmatrix}\boldsymbol{S}_1 \\ \boldsymbol{S}_{12} - \boldsymbol{S}_{21} \\ -\boldsymbol{S}_2\end{bmatrix} \neq \boldsymbol{0},$$

and thus

$$\begin{bmatrix}\boldsymbol{S}_1 \\ \boldsymbol{S}_{12} - \boldsymbol{S}_{21} \\ -\boldsymbol{S}_2\end{bmatrix} \neq \boldsymbol{0}. \tag{13}$$

On the other hand, it follows from $\widetilde{\boldsymbol{Y}}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ that

$$\boldsymbol{A}_{\boldsymbol{\Omega}} \begin{bmatrix}\boldsymbol{S}_1 \\ \boldsymbol{S}_{12} - \boldsymbol{S}_{21} \\ -\boldsymbol{S}_2\end{bmatrix} = \boldsymbol{0}.$$

Note that $\boldsymbol{A}_{\boldsymbol{\Omega}}$ is composed of atoms in $\mathcal{A}_{\boldsymbol{\Omega}}^1$, and has a nontrivial null space by (13). Then by the definition of spark we have

$$\text{rank}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) \geq \text{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1. \tag{14}$$

Moreover, for the nullity (dimension of the null space) of $\boldsymbol{A}_{\boldsymbol{\Omega}}$ it holds that

$$\begin{aligned} \text{nullity}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) &\geq \text{rank}\left(\begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12}-\boldsymbol{S}_{21} \\ -\boldsymbol{S}_2 \end{bmatrix}\right) \\ &\geq \text{rank}\left(\begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12}-\boldsymbol{S}_{21} \end{bmatrix}\right) \\ &\geq \text{rank}\left(\begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12} \end{bmatrix}\right) - \text{rank}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{S}_{21} \end{bmatrix}\right) \\ &\geq \text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right) - K_{12}. \end{aligned} \tag{15}$$

Consequently, the equality

$$\#\text{columns of } \boldsymbol{A}_{\boldsymbol{\Omega}} = \text{rank}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) + \text{nullity}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right)$$

together with (14) and (15) yields that $K + \widetilde{K} - K_{12} \geq \text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right) - 1 + \text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right) - K_{12}$, and then

$$2K \geq K + \widetilde{K} \geq \text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right) - 1 + \text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right),$$

which contradicts the condition in (12).

To show the uniqueness of the atomic decomposition, we observe that the condition in (12) implies that $K < \text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right) - 1$ since $\text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right) \leq K$. According to the definition of spark, any K atoms in $\mathcal{A}^1_{\boldsymbol{\Omega}}$ are linearly independent. Therefore, the atomic decomposition is unique given the set of frequencies $\mathcal{T} = \{f_j\}_{j=1}^K$. Now suppose there exists another decomposition $\boldsymbol{Y}^o = \sum_{j=1}^{\widetilde{K}} \widetilde{c}_j \boldsymbol{a}\left(\widetilde{f}_j, \widetilde{\boldsymbol{\phi}}_j\right)$ with $\widetilde{K} \leq K$ and a different set of frequencies $\left\{\widetilde{f}_j\right\}$ (i.e., there exists some $\widetilde{f}_j \notin \mathcal{T}$). Note that we have repetitively used the same notations for notational simplicity and we similarly define the other notations. Once again we have (13) since $\boldsymbol{A}_2$ is nonempty followed by $\boldsymbol{S}_2 \neq \boldsymbol{0}$. The rest of the proof follows from the same arguments. ■

It is generally difficult to compute $\text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right)$ given $\boldsymbol{\Omega}$, which makes the condition (12) hard to check in a particular scenario. However, it is checkable in the special complete data case, which is shown in the following corollary. Note that the rank minimization problem (7) can still help to recover the frequencies though (5) admits a trivial solution.

Corollary 2.1 (Complete data). *$\boldsymbol{Y}^o = \sum_{j=1}^K c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$ is the unique atomic decomposition satisfying that $K = \left\|\boldsymbol{Y}^o\right\|_{\mathcal{A},0}$ if*

$$K < \frac{N + \mathit{rank}\left(\boldsymbol{Y}^o\right)}{2}. \tag{16}$$

Proof. The result follows from $\text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right) = N + 1$ given $\boldsymbol{\Omega} = \boldsymbol{J}$, which holds according to the third part of Proposition 2.2. ■

Theorem 2.1 shows that the frequencies can be exactly recovered by minimizing the atomic ℓ_0 norm or equivalently solving the rank minimization problem (7) provided that the sparsity level K is sufficiently small. In particular, the upper bound of K depends on a particular sampling index set $\boldsymbol{\Omega}$ and the observed data $\boldsymbol{Y}^o_{\boldsymbol{\Omega}}$. In the SMV case where $\text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right) = 1$ and therefore, the condition becomes $K < \frac{1}{2}\text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right)$ (or $K \leq \frac{N}{2}$ in the complete data case). As we take more measurement vectors, we have a chance to recover more complex signals by potentially increasing $\text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right)$ (or $\text{rank}\left(\boldsymbol{Y}^o\right)$), which is practically relevant in array processing applications. When $\text{rank}\left(\boldsymbol{Y}^o_{\boldsymbol{\Omega}}\right)$ achieves the maximum value K, the sparsity level K can be as large as $\text{spark}\left(\mathcal{A}^1_{\boldsymbol{\Omega}}\right) - 2$ (or $N-1$ in the complete data case), which is consistent with the result in array processing (see e.g., [40]).

Remark 2.2. *An unpleasant scenario is in the presence of coherent sources. In an extreme case where all the sources are coherent, i.e., all the rows of $[s_{kt}]$ are identical up to some scale factors, it holds that rank $\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^{o}\right)=1$ as in the SMV case and taking more measurement vectors does not improve the number of frequencies recoverable. In fact, we can easily verify that all the measurement vectors are identical up to scale factors as well and thus provide the same amount of information for frequency recovery as a SMV. In this case, it is easy to verify that the optimal solution of $\boldsymbol{u}$ to (7) remains constant when we add or delete columns of $\boldsymbol{Y}$.*

Remark 2.3. *Since the upper bound increases with spark $\left(\mathcal{A}_{\boldsymbol{\Omega}}^{1}\right)$, an interesting topic in the future is to study selection of the sampling index set $\boldsymbol{\Omega}$, which in array processing corresponds to design of the geometry of the SLA, such that the spark is maximized. While preliminary results have been obtained in the discrete frequency setting (the frequencies are fixed on a uniform grid) in [46], it is worth noting that the continuous setting concerned here is more practical and challenging.*

3 Frequency Recovery via Convex Relaxation

3.1 Convex Relaxation and Semidefinite Formulation

The atomic ℓ_0 norm exploits sparsity directly, however, it is nonconvex and the formulated rank minimization problem cannot be globally solved with a practical algorithm. To avoid the nonconvexity, we utilize convex relaxation to relax the atomic ℓ_0 norm. In particular, it can be relaxed in two ways from two different perspectives. One is to relax the atomic ℓ_0 norm to the atomic ℓ_1 norm, or simply the atomic norm, which is defined as the gauge function of conv $(\mathcal{A})$, the convex hull of $\mathcal{A}$ [15]:

$$\begin{aligned}\left\|\boldsymbol{Y}\right\|_{\mathcal{A}} &\triangleq \inf\left\{t>0: \boldsymbol{Y}\in t\text{conv}\left(\mathcal{A}\right)\right\}\\ &= \inf\left\{\sum_k c_k : \boldsymbol{Y}=\sum_k c_k\boldsymbol{a}_k, c_k>0, \boldsymbol{a}_k\in\mathcal{A}\right\}.\end{aligned} \tag{17}$$

The atomic norm $\left\|\cdot\right\|_{\mathcal{A}}$ induced by $\mathcal{A}$ defined in (3) is a norm and thus convex by the property of the gauge function. The concept of atomic norm is firstly introduced in [15] and the authors argue that the atomic norm minimization problem is the best convex heuristic for recovering simple models with respect to a given atomic set. The other way of convex relaxation is based on a perspective of rank minimization illustrated by (6) and to relax the pseudo rank norm to the nuclear norm or equivalently the trace norm for a positive semidefinite matrix, i.e., to replace rank $(\boldsymbol{U})$ by tr $(\boldsymbol{U})$ in (6). It is worthy noting that the nuclear norm has been extensively studied for low rank matrix completion and recovery [43,44,47]. Interestingly enough, the two ways of convex relaxation are equivalent, which is shown in the following result.

Theorem 3.1. $\left\|\boldsymbol{Y}\right\|_{\mathcal{A}}$ *defined in (17) equals the optimal value of the following SDP:*

$$\min_{\boldsymbol{W},\boldsymbol{u},\boldsymbol{U}} \frac{1}{2\sqrt{N}} tr\left(\boldsymbol{U}\right), \text{ subject to } \boldsymbol{U}=\begin{bmatrix}\boldsymbol{W} & \boldsymbol{Y}^H\\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right)\end{bmatrix} \text{ and } \boldsymbol{U}\geq\boldsymbol{0}. \tag{18}$$

Before presenting the proof of Theorem 3.1, we clarify some notations first. We use the following identity whenever $\boldsymbol{R}\geq\boldsymbol{0}$:

$$\boldsymbol{y}^H\boldsymbol{R}^{-1}\boldsymbol{y}=\min t, \text{ subject to } \begin{bmatrix}t & \boldsymbol{y}^H\\ \boldsymbol{y} & \boldsymbol{R}\end{bmatrix}\geq 0. \tag{19}$$

It follows that $\boldsymbol{y}^H\boldsymbol{R}^{-1}\boldsymbol{y}$ is finite if and only if $\boldsymbol{y}$ is in the range space of $\boldsymbol{R}$. In fact, (19) is equivalent to defining $\boldsymbol{y}^H\boldsymbol{R}^{-1}\boldsymbol{y}\triangleq\lim_{\sigma\to 0_+}\boldsymbol{y}^H\left(\boldsymbol{R}+\sigma\boldsymbol{I}\right)^{-1}\boldsymbol{y}$ when $\boldsymbol{R}$ loses rank. The following result will be used to prove Theorem 3.1.

Lemma 3.1. *Given $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}^H \geq \boldsymbol{0}$, it holds that $\boldsymbol{y}^H \boldsymbol{R}^{-1} \boldsymbol{y} = \min \left\| \boldsymbol{s} \right\|_2^2$, subject to $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}$.*

Proof. We need only to show that for any $\boldsymbol{s}$ satisfying $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}$ it holds that $\boldsymbol{y}^H \boldsymbol{R}^{-1} \boldsymbol{y} \leq \left\| \boldsymbol{s} \right\|_2^2$, or equivalently, $\begin{bmatrix} \left\| \boldsymbol{s} \right\|_2^2 & \boldsymbol{y}^H \\ \boldsymbol{y} & \boldsymbol{R} \end{bmatrix} \geq 0$. The conclusion follows from that

$$\begin{bmatrix} \left\| \boldsymbol{s} \right\|_2^2 & \boldsymbol{y}^H \\ \boldsymbol{y} & \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \left\| \boldsymbol{s} \right\|_2^2 & \boldsymbol{s}^H \boldsymbol{A}^H \\ \boldsymbol{A}\boldsymbol{s} & \boldsymbol{A}\boldsymbol{A}^H \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}^H \\ \boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}^H \\ \boldsymbol{A} \end{bmatrix}^H \geq 0.$$

∎

Proof of Theorem 3.1: Provided $\boldsymbol{U} \geq \boldsymbol{0}$, we have equivalently $T(\boldsymbol{u}) \geq \boldsymbol{0}$ and $\boldsymbol{W} \geq \boldsymbol{Y}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{Y}$. So, we need to show that

$$\left\| \boldsymbol{Y} \right\|_{\mathcal{A}} = \min_{\boldsymbol{u}} \frac{\sqrt{N}}{2} u_1 + \frac{1}{2\sqrt{N}} \text{tr} \left(\boldsymbol{Y}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{Y} \right), \text{ subject to } T(\boldsymbol{u}) \geq \boldsymbol{0}, \tag{20}$$

where u_1 is the first entry of $\boldsymbol{u}$ as before. By the Vandermonde decomposition we have $T(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H = \left[\boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}} \right] \left[\boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}} \right]^H$, where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{f})$ is composed of columns $\boldsymbol{a}(f_j)$, and $\boldsymbol{P} = \text{diag}(\dots, p_j, \dots)$ with $p_j > 0$. Moreover, we can verify that $u_1 = \sum p_j$. For the tth column of $\boldsymbol{Y}$, say $\boldsymbol{y}_{:t}$, it holds by Lemma 3.1 that

$$\begin{aligned} \boldsymbol{y}_{:t}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{y}_{:t} &= \min_{\boldsymbol{v}} \left\| \boldsymbol{v} \right\|_2^2, \text{ subject to } \boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}} \boldsymbol{v} = \boldsymbol{y}_{:t} \\ &= \min_{\boldsymbol{s}} \left\| \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{s} \right\|_2^2, \text{ subject to } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}_{:t} \\ &= \min_{\boldsymbol{s}} \boldsymbol{s}^H \boldsymbol{P}^{-1} \boldsymbol{s}, \text{ subject to } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}_{:t}. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{tr} \left(\boldsymbol{Y}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{Y} \right) &= \sum_{t=1}^{N} \boldsymbol{y}_{:t}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{y}_{:t} \\ &= \min_{\boldsymbol{S}, \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S} = \boldsymbol{Y}} \text{tr} \left(\boldsymbol{S}^H \boldsymbol{P}^{-1} \boldsymbol{S} \right). \end{aligned}$$

We complete the proof via the following equalities:

$$\begin{aligned} & \min_{\boldsymbol{u}} \frac{\sqrt{N}}{2} u_1 + \frac{1}{2\sqrt{N}} \text{tr} \left(\boldsymbol{Y}^H \left[T(\boldsymbol{u}) \right]^{-1} \boldsymbol{Y} \right) \\ = & \min_{\substack{\boldsymbol{f}, \boldsymbol{p} \succeq \boldsymbol{0}, \boldsymbol{S} \\ \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S} = \boldsymbol{Y}}} \frac{\sqrt{N}}{2} \sum_j p_j + \frac{1}{2\sqrt{N}} \text{tr} \left(\boldsymbol{S}^H \boldsymbol{P}^{-1} \boldsymbol{S} \right) \\ = & \min_{\substack{\boldsymbol{f}, \boldsymbol{p} \succeq \boldsymbol{0}, \boldsymbol{S} \\ \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S} = \boldsymbol{Y}}} \frac{\sqrt{N}}{2} \sum_j p_j + \frac{1}{2\sqrt{N}} \sum_j \left\| \boldsymbol{S}_j \right\|_2^2 p_j^{-1} \\ = & \min_{\boldsymbol{f}, \boldsymbol{S}} \sum_j \left\| \boldsymbol{S}_j \right\|_2, \text{ subject to } \boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{f}) \boldsymbol{S} \\ = & \min_{\boldsymbol{f}, \boldsymbol{c} \succeq \boldsymbol{0}} \sum_j c_j, \text{ subject to } \boldsymbol{Y} = \sum_j c_j \boldsymbol{a}(f_j, \boldsymbol{\phi}_j) \\ = & \left\| \boldsymbol{Y} \right\|_{\mathcal{A}}. \end{aligned} \tag{21}$$

The second last equality holds by the substitutions $c_j = \left\|\boldsymbol{S}_j\right\|_2$ and $\boldsymbol{\phi}_j = c_j^{-1}\boldsymbol{S}_j$, followed by $\boldsymbol{Y} = \boldsymbol{A}\left(\boldsymbol{f}\right)\boldsymbol{S} = \sum_j \boldsymbol{a}\left(f_j\right)\boldsymbol{S}_j = \sum_j c_j\boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$, where $\boldsymbol{S}_j$ denotes the jth row of $\boldsymbol{S}$. The last equality follows from the definition of the atomic norm. ■

Remark 3.1. *The semidefinite formulation (18) of the atomic norm was firstly reported in our conference paper [1]. When preparing this paper, we found that the same result was also independently obtained in [48].*

Remark 3.2. *Other optimization methods can be studied in the future for joint sparse frequency recovery by borrowing ideas in low rank matrix recovery (see, e.g., [49, 50]). As an example, the (nonconvex) log-determinant criterion* $\log\left|\boldsymbol{U} + \epsilon\boldsymbol{I}\right|$*, where* $\epsilon > 0$ *is a small number, has been studied in [1] and implemented by a series of SDPs via a majorization-maximization (MM) process.*

Theorem 3.1 shows that the atomic norm can be formulated as an SDP and thus can be globally solved using standard SDP solvers such as SDPT3 [51]. Note that the semidefinite formulation (18) generalizes the results in the SMV case in [13, 16, 17]. As a result, by the convex relaxation we propose the following optimization problem for frequency recovery:

$$\min_{\boldsymbol{Y}} \left\|\boldsymbol{Y}\right\|_{\mathcal{A}}, \text{ subject to } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o, \tag{22}$$

or equivalently,

$$\begin{aligned} &\min_{\boldsymbol{Y},\boldsymbol{W},\boldsymbol{u},\boldsymbol{U}} \text{tr}\left(\boldsymbol{U}\right), \\ &\text{subject to } \boldsymbol{U} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix}, \boldsymbol{U} \geq \boldsymbol{0}, \text{ and } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o. \end{aligned} \tag{23}$$

For a matrix $\boldsymbol{V} \in \mathbb{C}^{N\times L}$, the dual norm of the atomic norm is defined as

$$\begin{aligned} \left\|\boldsymbol{V}\right\|_{\mathcal{A}}^* &= \sup_{\left\|\boldsymbol{a}\right\|_{\mathcal{A}} \leq 1} \left\langle \boldsymbol{V}, \boldsymbol{a}\right\rangle_{\mathbb{R}} \\ &= \sup_{\boldsymbol{a}\left(f,\boldsymbol{\phi}\right)\in\mathcal{A}} \left\langle \boldsymbol{V}, \boldsymbol{a}\left(f,\boldsymbol{\phi}\right)\right\rangle_{\mathbb{R}} \\ &= \sup_{f\in\mathbb{T},\boldsymbol{\phi}\in\mathbb{S}^{2L-1}} \left\langle \boldsymbol{a}(f)^H\boldsymbol{V}, \boldsymbol{\phi}\right\rangle_{\mathbb{R}} \\ &= \sup_{f\in\mathbb{T}} \left\|\boldsymbol{a}(f)^H\boldsymbol{V}\right\|_2, \end{aligned} \tag{24}$$

where $\left\langle \boldsymbol{A}, \boldsymbol{B}\right\rangle_{\mathbb{R}} = \Re\text{tr}\left(\boldsymbol{B}^H\boldsymbol{A}\right)$ for two matrices (or column/row vectors) $\boldsymbol{A}$ and $\boldsymbol{B}$ of proper dimensions and $\Re$ takes the real part of a complex argument. The dual problem of (22) is thus

$$\begin{aligned} &\max_{\boldsymbol{V}\in\mathbb{C}^{N\times L}} \left\langle \boldsymbol{V}_{\boldsymbol{\Omega}}, \boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right\rangle_{\mathbb{R}}, \\ &\text{subject to } \left\|\boldsymbol{V}\right\|_{\mathcal{A}}^* \leq 1, \boldsymbol{V}_{\overline{\boldsymbol{\Omega}}} = \boldsymbol{0} \end{aligned} \tag{25}$$

following from a standard Lagrangian analysis [52]. On the other hand, the dual problem of (22) has the following semidefinite formulation

$$\begin{aligned} &\max_{\boldsymbol{V}\in\mathbb{C}^{N\times L},\boldsymbol{H}\in\mathbb{C}^{N\times N}} \left\langle \boldsymbol{V}_{\boldsymbol{\Omega}}, \boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right\rangle_{\mathbb{R}}, \\ &\text{subject to } \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{V}^H \\ -\boldsymbol{V} & \boldsymbol{H} \end{bmatrix} \geq \boldsymbol{0}, \\ &\sum_{k=1}^{N-j} H_{k,k+j} = \left\{ \begin{array}{ll} 1, & j = 0, \\ 0, & j \in \left[N-1\right], \end{array} \right. \\ &\boldsymbol{V}_{\overline{\boldsymbol{\Omega}}} = \boldsymbol{0} \end{aligned} \tag{26}$$

based on the semidefinite formulation (23) and a standard Lagrangian analysis. That means, the first two constraints in (26) can equivalently characterize $\|\boldsymbol{V}\|_{\mathcal{A}}^{*} \leq 1$. We empirically find that the dual problem (26) can be solved more efficiently than the primal problem (23) using, for example, SDPT3, while most solvers can return an optimal solution of (23) (as the dual of (26)) for free when solving (26).

3.2 Frequency Retrieval

Given an optimal solution $(\boldsymbol{Y}^*, \boldsymbol{u}^*)$ of (23), we can retrieve the frequencies of interest and obtain corresponding atomic decomposition(s) of $\boldsymbol{Y}^*$ as follows. Consider first the case where $T(\boldsymbol{u}^*)$ is rank deficient, i.e., $K^* \triangleq \text{rank}(T(\boldsymbol{u}^*)) \leq N-1$. Then we can obtain a unique Vandermonde decomposition $T(\boldsymbol{u}^*) = \boldsymbol{A}(\boldsymbol{f}^*)\,\text{diag}(\boldsymbol{p}^*)\,\boldsymbol{A}(\boldsymbol{f}^*)^H$ of order K^* by Lemma 2.1. Since $\boldsymbol{Y}^*$ is in the range space of $T(\boldsymbol{u}^*)$, there exist unique $\boldsymbol{S}^*$, c_j^* and $\boldsymbol{\phi}_j^*$ corresponding to the variables in the proof of Theorem 3.1 such that

$$\boldsymbol{Y}^* = \boldsymbol{A}(\boldsymbol{f}^*)\,\boldsymbol{S}^* = \sum_{j=1}^{K^*} c_j^* \boldsymbol{a}\left(f_j^*, \boldsymbol{\phi}_j^*\right)$$

is an atomic decomposition of $\boldsymbol{Y}^*$ with

$$\|\boldsymbol{Y}^*\|_{\mathcal{A}} = \sum c_j^* = \sum \left\|\boldsymbol{S}_j^*\right\|_2 = \sqrt{N} \sum p_j^* = \sqrt{N} u_1^*.$$

That means, we have obtained a unique atomic decomposition that achieves the atomic norm. On the other hand, when $T(\boldsymbol{u}^*)$ has full rank, there exist infinitely many Vandermonde decompositions of $T(\boldsymbol{u}^*)$ of order N according to Remark 2.1. Consequently, we can obtain infinitely many atomic decompositions of $\boldsymbol{Y}^*$ of order N with each achieving the atomic norm. Therefore, it is impossible to exactly recover the frequencies.

3.3 Theoretical Guarantees of Convex Relaxation: Complete Data

To guarantee exact frequency recovery with the convex relaxation, we require that the frequencies of interest be appropriately separate. To quantify the separation, we define *minimum separation* of a set of frequencies as follows.

Definition 3.1 (Minimum separation)**.** *For a finite subset $\mathcal{T} \subset \mathbb{T}$, the minimum separation of $\mathcal{T}$ is defined as the closest wrap-around distance between any two elements,*

$$\Delta_{\mathcal{T}} = \inf_{a,b \in \mathcal{T}: a \neq b} \min\left\{|a-b|, 1-|a-b|\right\}.$$

For example, the distance between 0 and $\frac{3}{4}$ equals $\frac{1}{4}$.

Inspired by [13] and [16], we study the cases of complete data and incomplete data separately. Our analysis in the former is deterministic and applicable once $N \geq 257$ while that in the latter is based on non-asymptotic analysis of random matrices [53]. Since the rows of the source signals $[s_{kt}]$ are allowed to be correlated or even coherent in our analysis, it is *worst case* analysis as opposed to the *average case* analysis in [30, 34].

Given the complete data, the proposed atomic norm minimization problem (22) has a trivial solution, however, the semidefinite formulation (23) can be used to recover the frequencies. Its theoretical guarantee is presented in the following result, which generalizes that in the SMV case in [13].

Theorem 3.2 (Complete data)**.** *$\boldsymbol{Y}^o = \sum_{j=1}^{K} c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$ is the unique atomic decomposition satisfying that $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{j=1}^{K} c_j$ if $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$ and $N \geq 257$.*[1]

[1] The condition $N \geq 257$ is more like a technical requirement but not an obstacle in practice (see numerical simulations in Section 7).

Proof. See Section 5. ■

Remark 3.3 (Necessary separation for uniform recovery)**.** *Theorem 3.2 presents a result of uniform recovery in the sense that it holds for all source signals $[s_{kt}]$ once the frequency separation condition $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor} \approx 4N^{-1}$ is satisfied. Consider the observation model (1) which contains $2KL + K$ real variables ($2KL$ for $[s_{kt}]$ and K for $\{f_k\}$) and $2NL$ real measurements. As a result, for guaranteed uniform recovery, a necessary condition is that $2KL + K \leq 2NL$ or equivalently $K \leq \frac{2NL}{2L+1}$. On the other hand, given the sparsity level K, we can always select a set of equispaced frequencies which has a minimum separation of $\frac{1}{K} \geq \left(1 + \frac{1}{2L}\right) N^{-1}$. Therefore, a necessary condition of the minimum separation for uniform recovery is $\Delta_{\mathcal{T}} \geq \left(1 + \frac{1}{2L}\right) N^{-1}$.*

We do not impose any assumption on the source signals $[s_{kt}]$ in Theorem 3.2. Therefore, it can be applied uniformly to all kinds of source signals including correlated or even coherent sources. When all the sources are coherent as discussed in Remark 2.2, taking more measurement vectors does not increase the information for frequency recovery. In this case, in fact, it is easy to show that the proposed atomic norm minimization problem (23) produces the same solution of $\boldsymbol{u}$ up to a positive scale factor and thus the same frequency solution when increasing/decreasing the number of measurement vectors. As a result, we cannot expect a better theoretical guarantee than that in the SMV case. Our contribution by Theorem 3.2 is showing that in the presence of MMVs we can confidently recover the frequencies via a single convex optimization problem by exploiting the joint sparsity therein. Though our *worst case* analysis does not improve over that in [13] on the SMV case, it will be shown in Section 7 via numerical simulations that the proposed joint sparse frequency recovery method improves the recovery performance significantly when the source signals are at general positions. We pose this *average case* analysis as a future work.

3.4 Theoretical Guarantees of Convex Relaxation: Incomplete Data

It is difficult to analyze the atomic norm minimization problem (22) or (23) with a specific sampling index set $\boldsymbol{\Omega}$. Instead, we assume that it is selected uniformly at random. Our result is presented in the following theorem, which generalizes that in the SMV case in [16] with modifications.

Theorem 3.3 (Incomplete data)**.** *Suppose we observe $\boldsymbol{Y}^o = \sum_{j=1}^K c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$ on the index set $\boldsymbol{\Omega} \times [L]$, where $\boldsymbol{\Omega} \subset \boldsymbol{J}$ is of size M and selected uniformly at random. Assume that $\left\{\boldsymbol{\phi}_j\right\}_{j=1}^K \subset \mathbb{S}^{2L-1}$ are independent random variables with $\mathbb{E}\boldsymbol{\phi}_j = \boldsymbol{0}$. If $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$, then there exists a numerical constant C such that*

$$M \geq C \max\left\{\log^2 \frac{\sqrt{L}N}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{L}N}{\delta}\right\} \tag{27}$$

is sufficient to guarantee that, with probability at least $1 - \delta$, $\boldsymbol{Y}^o$ is the unique optimizer to (22) or (23) and $\boldsymbol{Y}^o = \sum_{j=1}^K c_j \boldsymbol{a}\left(f_j, \boldsymbol{\phi}_j\right)$ is the unique atomic decomposition satisfying that $\left\|\boldsymbol{Y}^o\right\|_{\mathcal{A}} = \sum_{j=1}^K c_j$.

Proof. See Section 6. ■

Remark 3.4. *We emphasize that the dependence of the sample size M per snapshot on L is for controlling the probability of successful recovery. To make it clear, we consider the case where we seek to recover the columns of $\boldsymbol{Y}^o$ independently via the SMV atomic norm minimization. Then with M satisfying (27) at $L = 1$ as in [16], each column of $\boldsymbol{Y}^o$ can be recovered with probability $1 - \delta$. It follows that $\boldsymbol{Y}^o$ can be exactly recovered with probability at least $1 - L\delta$. However, if we attempt to recover $\boldsymbol{Y}^o$ via a single convex optimization problem that we propose, then with the same number of measurements the success probability is improved to $1 - \sqrt{L}\delta$ (replacing δ in (27) by $\sqrt{L}\delta$).*

Remark 3.5.

1. *Compared to [16] on the SMV case, Theorem 3.3 relaxes the assumption of* $\left\{\boldsymbol{\phi}_j\right\}_{j=1}^K$*. In particular, the phases are assumed in the former drawn i.i.d. from the uniform distribution on the unit sphere. In contrast, they are only required to be independent and have zero means.*

2. *The relaxation of the assumption has significant impact on the practical DOA estimation problem in two aspects. On one hand, each* $\boldsymbol{\phi}_j$ *corresponds to one source and therefore, they do not necessarily obey an identical distribution. On the other hand, under the assumption of Theorem 3.3 the sources are allowed to be (spatially) coherent or temporally correlated, which can be encountered in practical scenarios. For example, the source signals* $[s_{kt}]$ *are temporarily (column-wise) correlated if the rows of* $[s_{kt}]$ *are i.i.d. Gaussian with mean zero and non-diagonal covariance. Two sources (or two rows of* $[s_{kt}]$*) are certain to be coherent (identical up to a scale factor) if they are i.i.d. Gaussian with mean zero and covariance of rank one, where the resulting* $\boldsymbol{\phi}_j$*'s satisfy the assumptions in Theorem 3.3. In contrast, any two sources with* $\boldsymbol{\phi}_j$ *drawn i.i.d. from the uniform distribution on* $\mathbb{S}^{2L-1}$ *are uncorrelated.*

Theorem 3.3 shows that, if we fix the number of measurement vectors L, then a number of $O\left(K\log K\log N\right)$ samples per measurement vector are sufficient to recover the frequencies and the missing data by solving the convex optimization problem (23) provided that the frequencies are sufficiently separate. When applied to array processing, it means that a number of $O\left(K\log K\log N\right)$ sensors are sufficient to exactly determine the directions of K sources.

Our proofs of Theorems 3.2 and 3.3, which are deferred to Sections 5 and 6 respectively, are inspired by those in [13] and [16] and follow similar procedures. The main challenge of our proofs is dealing with vector-valued dual polynomials induced by the MMV problem.

4 Connection to Prior Arts

4.1 Grid-based Joint Sparse Recovery

The problem of joint sparse frequency recovery concerned in this paper has been widely studied within the framework of CS, typically under the topic of DOA estimation. Since CS has been focused on signals that can be sparsely represented under a finite discrete dictionary, discretization/gridding of the frequency domain has become standard, or equivalently, the frequencies are assumed to lie on a fixed grid. Now recall the atomic ℓ_p norm, $p\in\{0,1\}$, in (4) and (17). It is easy to show that, for $p\in\{0,1\}$,

$$\left\|\boldsymbol{Y}\right\|_{\mathcal{A},p}=\inf\left\{\sum_k\left\|\boldsymbol{s}_k\right\|_2^p:\boldsymbol{Y}=\sum_k\boldsymbol{a}\left(f_k\right)\boldsymbol{s}_k,f_k\in\mathbb{T}\right\}, \tag{28}$$

where $\boldsymbol{s}_k\in\mathbb{C}^{1\times L}$. As a result, the atomic ℓ_0 norm (or the atomic norm) generalizes the $\ell_{2,0}$ norm (or the $\ell_{2,1}$ norm) in grid-based joint sparse recovery methods (see, e.g., [25,35]) to the continuous frequency setting. It is worth noting that, since the grid-based methods rely on the ideal assumption that the true frequencies lie exactly on a fixed grid, we cannot expect exact frequency recovery when it fails unlike the continuous recovery technique studied in this paper. Moreover, even if the assumption holds, existing coherence or RIP-based analysis for the $\ell_{2,1}$-based joint sparse recovery is very conservative, compared to the results in this paper, due to the high coherence in the presence of a dense grid. Readers are referred to [13] for detailed discussions in the SMV case.

4.2 Gridless Joint Sparse Recovery

To the best of our knowledge, the only discretization-free/gridless technique for joint sparse frequency recovery is introduced in [40] prior to this work, termed as the sparse and parametric approach (SPA).[2] Differently from the atomic norm technique proposed in this paper, SPA is from a statistical perspective and based on a weighted covariance fitting (WCF) criterion. But we show next that the two methods are connected. Given the complete data case as an example. In the limiting noiseless case, SPA is equivalent to the following optimization problem:

$$\min_{\boldsymbol{u}\in\mathbb{C}^N, T(\boldsymbol{u})\geq\boldsymbol{0}} \text{tr}\left(\widehat{\boldsymbol{R}}T\left(\boldsymbol{u}\right)^{-1}\widehat{\boldsymbol{R}}\right)+\text{tr}\left(T\left(\boldsymbol{u}\right)\right), \tag{29}$$

where $\widehat{\boldsymbol{R}}=\frac{1}{L}\boldsymbol{Y}^o\boldsymbol{Y}^{oH}$ denotes the sample covariance. Let $\boldsymbol{V}=\left(\frac{1}{N}\boldsymbol{Y}^{oH}\boldsymbol{Y}^o\right)^{\frac{1}{2}}\in\mathbb{C}^{L\times L}$. Then we have the following equalities/equivalences:

$$\begin{aligned}
(29) &= \min_{\boldsymbol{u},T(\boldsymbol{u})\geq\boldsymbol{0}} \frac{N}{L^2}\text{tr}\left(\left(\boldsymbol{Y}^o\boldsymbol{V}\right)^H T\left(\boldsymbol{u}\right)^{-1}\boldsymbol{Y}^o\boldsymbol{V}\right)+\text{tr}\left(T\left(\boldsymbol{u}\right)\right)\\
&= \min_{\boldsymbol{W},\boldsymbol{u}} \text{tr}\left(\boldsymbol{W}\right)+\text{tr}\left(T\left(\boldsymbol{u}\right)\right),\text{ subject to } \begin{bmatrix}\boldsymbol{W} & \frac{\sqrt{N}}{L}\left(\boldsymbol{Y}^o\boldsymbol{V}\right)^H\\ \frac{\sqrt{N}}{L}\boldsymbol{Y}^o\boldsymbol{V} & T\left(\boldsymbol{u}\right)\end{bmatrix}\geq\boldsymbol{0}\\
&= 2\sqrt{N}\left\|\frac{\sqrt{N}}{L}\boldsymbol{Y}^o\boldsymbol{V}\right\|_{\mathcal{A}}\\
&= \frac{2N}{L}\left\|\boldsymbol{Y}^o\boldsymbol{V}\right\|_{\mathcal{A}},
\end{aligned}$$

where the third equality follows from the semidefinite formulation of the atomic norm in Theorem 3.1. As a result, SPA is equivalent to computing the atomic norm of $\boldsymbol{Y}^o\boldsymbol{V}$. Note by (2) that

$$\boldsymbol{Y}^o\boldsymbol{V}=\sum_{k=1}^K\boldsymbol{a}\left(f_k\right)\left(\boldsymbol{s}_k\boldsymbol{V}\right). \tag{30}$$

So, SPA is the atomic norm method proposed in this paper with modifications of the source signals. In the special SMV case where $L=1$ and $\boldsymbol{V}$ is a positive scalar, the two techniques are exactly equivalent. We note that a similar result holds with incomplete data and will be omitted.

5 Proof of Theorem 3.2

5.1 Dual Certificate

We have considered the general case where $\boldsymbol{J}=\{0,\cdots,N-1\}$ in this paper. For technical reasons, our proof is mainly focused on the symmetric case where $\boldsymbol{J}=\{-2n,\ldots,2n\}$ with $n=\lfloor\frac{N-1}{4}\rfloor$, in which all our previous results hold as well with the substitution $N=|\boldsymbol{J}|$. We complete the proof by showing that the same result holds in the general case.

The following proposition presents the so-called *dual certificate* which guarantees the optimality of a certain atomic decomposition.

[2] Another related paper published online is [54], however, in this paper the authors reformulate the MMV problem into a SMV one, instead of exploiting the joint sparsity, and then solve the problem within the framework in [13]. Due to the loss of data structure, the capability of frequency recovery might be degraded.

Proposition 5.1. *Suppose that the atomic set $\mathcal{A}$ is composed of atoms defined by $\boldsymbol{a}(f, \boldsymbol{\phi})$ whose rows are indexed by the set $\boldsymbol{J}$ being either $\{-2n, \dots, 2n\}$ or $\{0, \cdots, N-1\}$. Then $\boldsymbol{Y}^o = \sum_{k=1}^K c_k \boldsymbol{a}(f_k, \boldsymbol{\phi}_k)$ is the unique atomic decomposition satisfying that $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{k=1}^K c_k$ if $K \leq |\boldsymbol{J}|$ and there exists a vector-valued dual polynomial $Q : \mathbb{T} \to \mathbb{C}^{1\times L}$*

$$Q(f) = \langle \boldsymbol{V}, \boldsymbol{a}(f) \rangle \triangleq \boldsymbol{a}(f)^H \boldsymbol{V} \tag{31}$$

satisfying that

$$Q(f_k) = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \tag{32}$$

$$\|Q(f)\|_2 < 1, \quad f \in \mathbb{T} \backslash \mathcal{T}, \tag{33}$$

where the coefficient matrix $\boldsymbol{V} \in \mathbb{C}^{|\boldsymbol{J}|\times L}$.[3]

Proof. For a vector-valued polynomial $Q(f) = \boldsymbol{a}(f)^H \boldsymbol{V}$ satisfying (32) and (33), we have by (24) that

$$\|\boldsymbol{V}\|_{\mathcal{A}}^* = \sup_{f\in\mathbb{T}} \left\|\boldsymbol{a}(f)^H \boldsymbol{V}\right\|_2 = 1.$$

Based on (32) and (33) it is easy to show that

$$\begin{aligned}\langle \boldsymbol{V}, \boldsymbol{Y}^o \rangle_{\mathbb{R}} &= \left\langle \boldsymbol{V}, \sum_{k=1}^K c_k \boldsymbol{a}(f_k, \boldsymbol{\phi}_k) \right\rangle_{\mathbb{R}} \\ &= \sum_{k=1}^K c_k \left\langle \boldsymbol{a}(f_k)^H \boldsymbol{V}, \boldsymbol{\phi}_k \right\rangle_{\mathbb{R}} \\ &= \sum_{k=1}^K c_k \geq \|\boldsymbol{Y}^o\|_{\mathcal{A}}.\end{aligned}$$

The last inequality follows from the definition of the atomic norm. On the other hand, it holds that $\langle \boldsymbol{V}, \boldsymbol{Y}^o \rangle_{\mathbb{R}} \leq \|\boldsymbol{V}\|_{\mathcal{A}}^* \|\boldsymbol{Y}^o\|_{\mathcal{A}} \leq \|\boldsymbol{Y}^o\|_{\mathcal{A}}$, implying that $\langle \boldsymbol{V}, \boldsymbol{Y}^o \rangle_{\mathbb{R}} = \|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{k=1}^K c_k$. We next show the uniqueness. Suppose that there exists another decomposition $\boldsymbol{Y}^o = \sum_k \widetilde{c}_k \boldsymbol{a}\left(\widetilde{f}_k, \widetilde{\boldsymbol{\phi}}_k\right)$ satisfying also that $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_k \widetilde{c}_k$. There must exist some $\widetilde{f}_k \notin \mathcal{T}$ due to linear independence between $\{\boldsymbol{a}(f_k)\}_{k=1}^K$, $K \leq |\boldsymbol{J}|$, for both the settings of $\boldsymbol{J}$, otherwise the two decompositions will be identical. We complete the proof by the following contradiction:

$$\begin{aligned}&\sum_k \widetilde{c}_k = \|\boldsymbol{Y}^o\|_{\mathcal{A}} = \langle \boldsymbol{V}, \boldsymbol{Y}^o \rangle_{\mathbb{R}} \\ &= \left\langle \boldsymbol{V}, \sum_k \widetilde{c}_k \boldsymbol{a}\left(\widetilde{f}_k, \widetilde{\boldsymbol{\phi}}_k\right) \right\rangle_{\mathbb{R}} \\ &= \sum_{\widetilde{f}_k \in \mathcal{T}} \widetilde{c}_k \left\langle Q\left(\widetilde{f}_k\right), \widetilde{\boldsymbol{\phi}}_k \right\rangle_{\mathbb{R}} + \sum_{\widetilde{f}_k \notin \mathcal{T}} \widetilde{c}_k \left\langle Q\left(\widetilde{f}_k\right), \widetilde{\boldsymbol{\phi}}_k \right\rangle_{\mathbb{R}} \\ &< \sum_{\widetilde{f}_k \in \mathcal{T}} \widetilde{c}_k + \sum_{\widetilde{f}_k \notin \mathcal{T}} \widetilde{c}_k = \sum_k \widetilde{c}_k.\end{aligned}$$

■

The rest of the proof is focused on construction of a dual polynomial as in Proposition 5.1, for which the condition $K \leq |\boldsymbol{J}|$ naturally holds. We revisit the proof in [13] for the SMV case in the ensuing subsection and present our proof for the general MMV case after that.

[3] Here we have abused the notation of inner-product since $\boldsymbol{V} \in \mathbb{C}^{|\boldsymbol{J}|\times L}$ and $\langle \boldsymbol{V}, \boldsymbol{a}(f) \rangle = \boldsymbol{a}(f)^H \boldsymbol{V} \in \mathbb{C}^{1\times L}$.

5.2 Revisiting the SMV Case

We consider only the case where $\boldsymbol{J} = \{-2n, \ldots, 2n\}$. The reason will be clear in the next subsection. Candès and Fernandez-Granda [13] consider a polynomial $q : \mathbb{T} \rightarrow \mathbb{C}$ of the following form:

$$q(f) = \sum_{f_k \in \mathcal{T}} \alpha_k \mathcal{K}(f - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \mathcal{K}'(f - f_k),$$

where $\mathcal{K}(f)$ is the squared Fejér kernel

$$\mathcal{K}(f) = \left[\frac{\sin(\pi(n+1)f)}{(n+1)\sin(\pi f)}\right]^4 = \sum_{j=-2n}^{2n} g_n(j)\, e^{-i2\pi j f}$$

with coefficients

$$g_n(j) = \frac{1}{n+1} \sum_{k=\max(j-n-1,-n-1)}^{\min(j+n+1,n+1)} \left(1 - \frac{|k|}{n+1}\right)\left(1 - \frac{|j-k|}{n+1}\right) \tag{34}$$

obeying that $0 < g_n(j) \leq 1$, $j = -2n, \ldots, 2n$. Denote by $\mathcal{K}'$, $\mathcal{K}''$ and $\mathcal{K}'''$ the first three derivatives of $\mathcal{K}$. To obtain the coefficients $\{\alpha_k\}$ and $\{\beta_k\}$ such that q satisfies the constraints of the dual polynomial, they require for $f_j \in \mathcal{T}$ that

$$q(f_j) = \sum_{f_k \in \mathcal{T}} \alpha_k \mathcal{K}(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \mathcal{K}'(f_j - f_k) = \phi_j, \tag{35}$$

$$q'(f_j) = \sum_{f_k \in \mathcal{T}} \alpha_k \mathcal{K}'(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \mathcal{K}''(f_j - f_k) = 0. \tag{36}$$

The equality (35) ensures that $q(f)$ satisfies the interpolation condition (32), while (36) is used to guarantee the inequality (33). (35) and (36) can be written more compactly as

$$\begin{bmatrix} \boldsymbol{D}_0 & c_0^{-1}\boldsymbol{D}_1 \\ -c_0^{-1}\boldsymbol{D}_1 & -c_0^{-2}\boldsymbol{D}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix}, \tag{37}$$

where $[D_0]_{jk} = \mathcal{K}(f_j - f_k)$, $[D_1]_{jk} = \mathcal{K}'(f_j - f_k)$, $[D_2]_{jk} = \mathcal{K}''(f_j - f_k)$, $c_0 = \sqrt{|\mathcal{K}''(0)|} = \sqrt{\frac{4\pi^2 n(n+2)}{3}}$, $\boldsymbol{\Phi} \in \mathbb{C}^K$ is the vector by stacking $\{\phi_j\}$ together and similarly for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{C}^K$. The system of linear equations above is rescaled following from [16] such that the coefficient matrix of (37) is symmetric, positive definite, and very close to identity. Then we have

$$\begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \\ -c_0\boldsymbol{D}_2^{-1}\boldsymbol{D}_1 \end{bmatrix} \boldsymbol{D}_3^{-1}\boldsymbol{\Phi}, \quad \boldsymbol{D}_3 \triangleq \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{D}_2^{-1}\boldsymbol{D}_1, \tag{38}$$

where $\boldsymbol{D}_3$ is the Schur complement, and the solution $\begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix}$ is shown to be close to $\begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix}$ with $\|\boldsymbol{\alpha}\|_\infty$ and $\|\boldsymbol{\beta}\|_\infty$ upper bounded.

To prove that the dual polynomial $q(f)$ with the specified coefficients satisfies the condition (33), the authors divide the the frequency domain $\mathbb{T}$ into two regions near to and far from the frequencies $\mathcal{T}$:

$$\mathcal{T}_{\text{near}} = \cup_{k=1}^{K} [f_k - \nu, f_k + \nu], \tag{39}$$

$$\mathcal{T}_{\text{far}} = \mathbb{T} \backslash \mathcal{T}_{\text{near}} \tag{40}$$

with $\nu = 8.245 \times 10^{-2}\frac{1}{n}$. $|q(f)| < 1$ is shown directly on $\mathcal{T}_{\text{far}}$, and on each continuous interval of $\mathcal{T}_{\text{near}}$, $|q(f)|$ is shown to be strictly concave and obtains the maximum 1 at each element in $\mathcal{T}$. We summarize some results in [13] which will be used later.

Lemma 5.1. *Assume $\Delta_{\mathcal{T}} \geq \Delta_{\mathit{min}} \triangleq \frac{1}{n}$ and $n \geq 64$.*

1.

$$\begin{aligned} \left\| \boldsymbol{I} - \boldsymbol{D}_0 \right\|_\infty &\leq 6.253 \times 10^{-3}, \\ \left\| \boldsymbol{D}_1 \right\|_\infty &\leq 0.1528n, \\ \left\| \left| \mathcal{K}''(0) \right| \boldsymbol{I} - \boldsymbol{D}_2 \right\|_\infty &\leq 4.212n^2, \\ \left\| \boldsymbol{I} - \boldsymbol{D}_3^{-1} \right\|_\infty &\leq 8.824 \times 10^{-3}. \end{aligned}$$

2. For $f \in [-\nu, \nu]$ (or $f \in [0, \nu] \cup [1-\nu, 1]$),

$$\begin{aligned} 1 \geq K(f) &\geq 0.9539, \\ -11.69n^2 \geq K''(f) &\geq -13.57n^2. \end{aligned}$$

3. For $f \in \mathcal{T}_{\mathit{near}}$,

$$\begin{aligned} \sum_{f_k \in \mathcal{T}} \left| \mathcal{K}'(f - f_k) \right| &\leq 1.272n, \\ \sum_{f_k \in \mathcal{T}} \left| \mathcal{K}'''(f - f_k) \right| &\leq 193.9n^3; \end{aligned}$$

for $f \in \mathcal{T}_{\mathit{near}} \backslash [f_j - \nu, f_j + \nu]$, $f_j \in \mathcal{T}$,

$$\begin{aligned} \sum_{f_k \notin \mathcal{T} \backslash \{f_j\}} \left| \mathcal{K}(f - f_k) \right| &\leq 6.279 \times 10^{-3}, \\ \sum_{f_k \notin \mathcal{T} \backslash \{f_j\}} \left| \mathcal{K}''(f - f_k) \right| &\leq 4.212n^2; \end{aligned}$$

and for $f \in \mathcal{T}_{\mathit{far}}$,

$$\begin{aligned} &1.008824 \sum_{f_k \in \mathcal{T}} \left| \mathcal{K}(f - f_k) \right| \\ &\quad + \frac{0.01647}{n} \sum_{f_k \in \mathcal{T}} \left| \mathcal{K}'(f - f_k) \right| \leq 0.99992. \end{aligned}$$

5.3 Proof of Theorem 3.2

Theorem 3.2 is a direct result of the proposition below.

Proposition 5.2. *Under the assumptions of Theorem 3.2, there exists a low-frequency vector-valued polynomial as in (31) satisfying (32) and (33) for $\boldsymbol{J}$ being either $\{-2n, \ldots, 2n\}$ or $\{0, \cdots, N-1\}$, where $K \leq |\boldsymbol{J}|$ naturally holds.*

To prove Proposition 5.2, we first show that once we can construct such a dual polynomial in the symmetric case, then a similar polynomial can be obtained in the general case.

Assume $N = 4n + n_0$ with $n = \lfloor \frac{N-1}{4} \rfloor$ and $n_0 = 1, 2, 3, 4$. Suppose that $\boldsymbol{Y}^o$ has a decomposition in the general case

$$\begin{aligned}\boldsymbol{Y}^o &= \sum_{k=1}^{K} c_k \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(N-1)f_k} \end{bmatrix} \boldsymbol{\phi}_k \\ &= \sum_{k=1}^{K} c_k \begin{bmatrix} e^{i2\pi(-2n)f_k} \\ \vdots \\ e^{i2\pi(2n)f_k} \\ \vdots \\ e^{i2\pi(2n+n_0-1)f_k} \end{bmatrix} \underbrace{e^{i2\pi(2n)f_k}\boldsymbol{\phi}_k}_{\widetilde{\boldsymbol{\phi}}_k}.\end{aligned} \tag{41}$$

Moreover, suppose that we can construct a dual polynomial $\widetilde{Q}(f) = \sum_{j=-2n}^{2n} \widetilde{\boldsymbol{V}}_j e^{-i2\pi jf}$ in the symmetric case satisfying that

$$\begin{aligned}\widetilde{Q}(f_k) &= \widetilde{\boldsymbol{\phi}}_k, \quad f_k \in \mathcal{T}, \\ \left\|\widetilde{Q}(f)\right\|_2 &< 1, \quad f \in \mathbb{T} \backslash \mathcal{T},\end{aligned}$$

and $K = |\mathcal{T}| \leq 4n + 1 \leq N$. Now define $\boldsymbol{V} \in \mathbb{C}^{N \times L}$ with

$$\boldsymbol{V}_j = \begin{cases} \widetilde{\boldsymbol{V}}_{j-2n}, & j = 0, \ldots, 4n, \\ \boldsymbol{0}, & \text{otherwise}, \end{cases}$$

and

$$\begin{aligned}Q(f) &= \sum_{j=0}^{N-1} \boldsymbol{V}_j e^{-i2\pi jf} \\ &= \sum_{j=0}^{4n} \widetilde{\boldsymbol{V}}_{j-2n} e^{-i2\pi jf} \\ &= e^{-i2\pi(2n)f} \widetilde{Q}(f).\end{aligned}$$

It can be readily verified that

$$\begin{aligned}Q(f_k) &= e^{-i2\pi(2n)f_k} \widetilde{\boldsymbol{\phi}}_k = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \\ \|Q(f)\|_2 &= \left\|\widetilde{Q}(f)\right\|_2 < 1, \quad f \in \mathbb{T} \backslash \mathcal{T}.\end{aligned}$$

That means, we have constructed a dual polynomial $Q(f)$ for the general case as in (31) satisfying (32) and (33) with $K \leq N$.

We prove Proposition 5.2 in the symmetric case in the rest of this subsection and thus complete the proof of Theorem 3.2. Inspired by [13], we construct the following vector-valued dual polynomial:

$$Q(f) = \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \mathcal{K}(f - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \mathcal{K}'(f - f_k), \tag{42}$$

where for each k, $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \in \mathbb{C}^{1\times L}$ are vector-valued coefficients. Again, for $f_j \in \mathcal{T}$, we impose the following equalities:

$$Q\left(f_j\right) = \sum_{f_k\in\mathcal{T}} \boldsymbol{\alpha}_k \mathcal{K}\left(f_j - f_k\right) + \sum_{f_k\in\mathcal{T}} \boldsymbol{\beta}_k \mathcal{K}'\left(f_j - f_k\right) = \boldsymbol{\phi}_j, \tag{43}$$

$$Q'\left(f_j\right) = \sum_{f_k\in\mathcal{T}} \boldsymbol{\alpha}_k \mathcal{K}'\left(f_j - f_k\right) + \sum_{f_k\in\mathcal{T}} \boldsymbol{\beta}_k \mathcal{K}''\left(f_j - f_k\right) = \mathbf{0}, \tag{44}$$

or equivalently,

$$\begin{bmatrix} \boldsymbol{D}_0 & c_0^{-1}\boldsymbol{D}_1 \\ -c_0^{-1}\boldsymbol{D}_1 & -c_0^{-2}\boldsymbol{D}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} \\ \mathbf{0} \end{bmatrix}, \tag{45}$$

where we denote $\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1^T, \dots, \boldsymbol{\phi}_K^T\right]^T \in \mathbb{C}^{K\times L}$, $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_K^T\right]^T \in \mathbb{C}^{K\times L}$ and similarly for $\boldsymbol{\beta} \in \mathbb{C}^{K\times L}$. Now we have $2KL$ equations with $2KL$ variables in total. Therefore, we can uniquely determine $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as in (38) following from (45). Note that the coefficient matrix in (45) is close to identity and hence, $\begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix}$ should be close to $\begin{bmatrix} \boldsymbol{\Phi} \\ \mathbf{0} \end{bmatrix}$. Differently from that in the SMV case, however, $\{\boldsymbol{\alpha}_k\}$ and $\{\boldsymbol{\beta}_k\}$ are vector-valued. The difficulty of the remaining proof is to provide tight bounds for them. To do this, we introduce the concept of $\ell_{2,\infty}$ matrix norm and its induced operator norm as follows.

Definition 5.1. *We define the $\ell_{2,\infty}$ norm of a matrix $\boldsymbol{X} \in \mathbb{C}^{d_1\times d_2}$ as*

$$\|\boldsymbol{X}\|_{2,\infty} = \max_j \|\boldsymbol{X}_j\|_2$$

and its induced norm of a linear operator $\boldsymbol{\mathcal{P}} : \mathbb{C}^{d_1\times d_2} \to \mathbb{C}^{d_3\times d_2}$ as

$$\|\boldsymbol{\mathcal{P}}\|_{2,\infty} = \sup_{\boldsymbol{X}\neq\mathbf{0}} \frac{\|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty}}{\|\boldsymbol{X}\|_{2,\infty}} = \sup_{\|\boldsymbol{X}\|_{2,\infty}\leq 1} \|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty},$$

where d_1, d_2 and d_3 are positive integers.

By the definition, we have $\|\boldsymbol{\Phi}\|_{2,\infty} = 1$ and expect to bound $\|\boldsymbol{\alpha}\|_{2,\infty}$ and $\|\boldsymbol{\beta}\|_{2,\infty}$ using the induced norm of the operators $\boldsymbol{D}_j$, $j = 0, \dots, 3$. So we need to calculate the induced norm first. Interestingly, the induced $\ell_{2,\infty}$ norm is identical to the ℓ_∞ norm, which is given in the following result.

Lemma 5.2. $\|\boldsymbol{\mathcal{P}}\|_{2,\infty} = \|\boldsymbol{\mathcal{P}}\|_\infty$ *for any linear operator $\boldsymbol{\mathcal{P}}$ defined by a matrix $\boldsymbol{P}$ such that $\boldsymbol{\mathcal{P}}\boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$ for any $\boldsymbol{X}$ of proper dimension.*

Proof. According to the definition of the induced ℓ_∞ norm, we have $\|\boldsymbol{\mathcal{P}}\|_\infty = \|\boldsymbol{P}\|_\infty = \max_j \|\boldsymbol{P}_j\|_1$, where $\boldsymbol{P}_j$ is the jth row of $\boldsymbol{P}$.

We first show that $\|\boldsymbol{\mathcal{P}}\|_{2,\infty} \geq \|\boldsymbol{\mathcal{P}}\|_\infty$. Denote by $\boldsymbol{X}_{:1}$ and $\boldsymbol{X}_{:-1}$ respectively the first column and the rest of $\boldsymbol{X}$. Note that if $\boldsymbol{X}_{:-1} = \mathbf{0}$, then $\|\boldsymbol{X}\|_{2,\infty} = \|\boldsymbol{X}_{:1}\|_\infty$ and $\|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty} = \|\boldsymbol{P}\boldsymbol{X}_{:1}\|_\infty$. Hence,

$$\begin{aligned} \|\boldsymbol{\mathcal{P}}\|_{2,\infty} &\geq \sup_{\|\boldsymbol{X}\|_{2,\infty}\leq 1, \boldsymbol{X}_{:-1}=\mathbf{0}} \|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty} \\ &= \sup_{\|\boldsymbol{X}_{:1}\|_\infty\leq 1} \|\boldsymbol{P}\boldsymbol{X}_{:1}\|_\infty \\ &= \|\boldsymbol{P}\|_\infty = \|\boldsymbol{\mathcal{P}}\|_\infty . \end{aligned}$$

We next show that $\|\boldsymbol{\mathcal{P}}\|_{2,\infty} \leq \|\boldsymbol{\mathcal{P}}\|_{\infty}$ to complete the proof. Note that

$$\begin{aligned}\|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty}^2 &= \max_j \|\boldsymbol{P}_j\boldsymbol{X}\|_2^2 \\ &= \max_j \boldsymbol{P}_j\boldsymbol{X}\boldsymbol{X}^H\boldsymbol{P}_j^H \\ &= \max_j \operatorname{tr}\left(\boldsymbol{R}\boldsymbol{P}_j^H\boldsymbol{P}_j\right),\end{aligned}$$

where $\boldsymbol{R} \triangleq \boldsymbol{X}\boldsymbol{X}^H$. It is easy to show that if $\|\boldsymbol{X}\|_{2,\infty} \leq 1$, then $\boldsymbol{R} \geq \boldsymbol{0}$ and $\boldsymbol{R} \preccurlyeq \boldsymbol{1}$, and vise versa, where $\boldsymbol{R} \preccurlyeq \boldsymbol{1}$ means $R_{jl} \leq 1$ for all the entries R_{jl} of $\boldsymbol{R}$. That means, the two constraints $\boldsymbol{R} \geq \boldsymbol{0}$ and $\boldsymbol{R} \preccurlyeq \boldsymbol{1}$ are equivalent to $\|\boldsymbol{X}\|_{2,\infty} \leq 1$. As a result,

$$\begin{aligned}\|\boldsymbol{\mathcal{P}}\|_{2,\infty}^2 &= \sup_{\|\boldsymbol{X}\|_{2,\infty}\leq 1} \|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty}^2 \\ &= \sup_{\|\boldsymbol{X}\|_{2,\infty}\leq 1} \max_j \operatorname{tr}\left(\boldsymbol{R}\boldsymbol{P}_j^H\boldsymbol{P}_j\right) \\ &= \max_j \sup_{\|\boldsymbol{X}\|_{2,\infty}\leq 1} \left\langle \operatorname{vec}(\boldsymbol{R}), \operatorname{vec}\left(\boldsymbol{P}_j^H\boldsymbol{P}_j\right)\right\rangle \\ &= \max_j \sup_{\boldsymbol{R}\geq\boldsymbol{0},\boldsymbol{R}\preccurlyeq\boldsymbol{1}} \left\langle \operatorname{vec}(\boldsymbol{R}), \operatorname{vec}\left(\boldsymbol{P}_j^H\boldsymbol{P}_j\right)\right\rangle \\ &\leq \max_j \sup_{\boldsymbol{R}\geq\boldsymbol{0},\boldsymbol{R}\preccurlyeq\boldsymbol{1}} \|\operatorname{vec}(\boldsymbol{R})\|_\infty \left\|\operatorname{vec}\left(\boldsymbol{P}_j^H\boldsymbol{P}_j\right)\right\|_1 \\ &\leq \max_j \|\boldsymbol{P}_j\|_1^2 \\ &= \|\boldsymbol{\mathcal{P}}\|_\infty^2,\end{aligned}$$

where vec $(\cdot)$ denotes the vectorized form of a matrix argument, i.e., by stacking all its columns as a column vector. The last inequality follows from that $\|\operatorname{vec}(\boldsymbol{R})\|_\infty \leq 1$ provided $\boldsymbol{R} \preccurlyeq \boldsymbol{1}$, and $\left\|\operatorname{vec}\left(\boldsymbol{P}_j^H\boldsymbol{P}_j\right)\right\|_1 = \|\boldsymbol{P}_j\|_1^2$. ■

Following from Lemma 5.2, we use only $\|\boldsymbol{\mathcal{P}}\|_\infty$ rather than $\|\boldsymbol{\mathcal{P}}\|_{2,\infty}$ to denote the induced $\ell_{2,\infty}$ norm hereafter for notational simplicity, which also avoids ambiguities between the matrix norm and the induced operator norm of a matrix (a matrix can also denote a linear operator). The following lemma provides upper bounds for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Lemma 5.3. *Under the assumptions of Proposition 5.2, the matrices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{C}^{K\times L}$ determined by (45) satisfy that*

$$\begin{aligned}\|\boldsymbol{\alpha} - \boldsymbol{\Phi}\|_{2,\infty} &\leq 8.824 \times 10^{-3}, \\ \|\boldsymbol{\beta}\|_{2,\infty} &\leq \beta^{max} \triangleq \frac{1.647}{n} \times 10^{-2}.\end{aligned}$$

It follows that $1 - 8.824 \times 10^{-3} \triangleq \alpha^{min} \leq \|\boldsymbol{\alpha}_j\|_2 \leq \alpha^{max} \triangleq 1 + 8.824 \times 10^{-3}$, $j = 1, \ldots, K$.

Proof. As in (38) we have

$$\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} \\ -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\end{bmatrix}\boldsymbol{D}_3^{-1}\boldsymbol{\Phi}. \tag{46}$$

It follows that $\boldsymbol{\alpha} - \boldsymbol{\Phi} = \left(\boldsymbol{D}_3^{-1} - \boldsymbol{I}\right)\boldsymbol{\Phi}$, and

$$\|\boldsymbol{\alpha} - \boldsymbol{\Phi}\|_{2,\infty} \leq \left\|\boldsymbol{D}_3^{-1} - \boldsymbol{I}\right\|_\infty \|\boldsymbol{\Phi}\|_{2,\infty} \leq 8.824 \times 10^{-3}.$$

The interval of $\|\boldsymbol{\alpha}_j\|_2$ is a direct result of the inequality

$$\left\|\boldsymbol{\phi}_j\right\|_2 - \left\|\boldsymbol{\alpha}_j - \boldsymbol{\phi}_j\right\|_2 \leq \|\boldsymbol{\alpha}_j\|_2 \leq \left\|\boldsymbol{\phi}_j\right\|_2 + \left\|\boldsymbol{\alpha}_j - \boldsymbol{\phi}_j\right\|_2$$

provided $\left\|\boldsymbol{\phi}_j\right\|_2 = 1$, $j \in [K]$. Utilizing $\|\boldsymbol{AB}\|_\infty \leq \|\boldsymbol{A}\|_\infty \|\boldsymbol{B}\|_\infty$ and $\left\|\boldsymbol{A}^{-1}\right\|_\infty \leq \frac{1}{1-\|\boldsymbol{I}-\boldsymbol{A}\|_\infty}$, we have

$$\begin{aligned}
\|\boldsymbol{\beta}\|_{2,\infty} &\leq \left\|\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\boldsymbol{D}_3^{-1}\right\|_\infty \\
&\leq \frac{\|\boldsymbol{D}_1\|_\infty \left\|\boldsymbol{D}_3^{-1}\right\|_\infty}{|\mathcal{K}''(0)| - \||\mathcal{K}''(0)|\,\boldsymbol{I} - \boldsymbol{D}_2\|_\infty} \\
&\leq \frac{1.647}{n} \times 10^{-2}.
\end{aligned}$$

■

We now complete the proof of Proposition 5.2 by showing that the polynomial $Q(f)$ constructed above satisfies (33). On $\mathcal{T}_{\text{far}}$,

$$\begin{aligned}
\|Q(f)\|_2 &= \left\| \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \mathcal{K}(f - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \mathcal{K}'(f - f_k) \right\|_2 \\
&\leq \sum_{f_k \in \mathcal{T}} \|\boldsymbol{\alpha}_k\|_2 |\mathcal{K}(f - f_k)| + \sum_{f_k \in \mathcal{T}} \|\boldsymbol{\beta}_k\|_2 \left|\mathcal{K}'(f - f_k)\right| \\
&\leq \alpha^{\max} \sum_{f_k \in \mathcal{T}} |\mathcal{K}(f - f_k)| + \beta^{\max} \sum_{f_k \in \mathcal{T}} \left|\mathcal{K}'(f - f_k)\right| \\
&\leq 0.99992.
\end{aligned} \tag{47}$$

On each interval $[f_j - \nu, f_j + \nu]$, $f_j \in \mathcal{T}$, since $\|Q(f)\|_2 = 1$ and $Q'(f) = \mathbf{0}$, we need only to show that $\|Q(f)\|_2^2$ is a concave function, i.e.,

$$\frac{1}{2}\frac{\mathrm{d}^2 \|Q(f)\|_2^2}{\mathrm{d}f^2} = \left\|Q'(f)\right\|_2^2 + \Re\left\{Q''(f)\, Q(f)^H\right\} < 0.$$

We observe that

$$\begin{aligned}
\left\|Q'(f)\right\|_2 &= \left\| \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \mathcal{K}'(f - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \mathcal{K}''(f - f_k) \right\|_2 \\
&\leq \alpha^{\max} \sum_{f_k \in \mathcal{T}} \left|\mathcal{K}'(f - f_k)\right| + \beta^{\max} \left|\mathcal{K}''(f - f_j)\right| \\
&\quad + \beta^{\max} \sum_{f_k \notin \mathcal{T} \setminus \{f_j\}} \left|\mathcal{K}''(f - f_k)\right| \\
&\leq 1.576n.
\end{aligned}$$

Moreover,

$$\begin{aligned} Q''\left(f\right)Q\left(f\right)^{H} &= \left[\sum_{f_k\in\mathcal{T}}\boldsymbol{\alpha}_k\mathcal{K}''\left(f-f_k\right)+\sum_{f_k\in\mathcal{T}}\boldsymbol{\beta}_k\mathcal{K}'''\left(f-f_k\right)\right] \\ &\quad\times\left[\sum_{f_k\in\mathcal{T}}\boldsymbol{\alpha}_k\mathcal{K}\left(f-f_k\right)+\sum_{f_k\in\mathcal{T}}\boldsymbol{\beta}_k\mathcal{K}'\left(f-f_k\right)\right]^{H} \\ &= \left\|\boldsymbol{\alpha}_j\right\|_2^2\mathcal{K}''\left(f-f_j\right)\mathcal{K}\left(f-f_j\right) \\ &\quad+\mathcal{K}''\left(f-f_j\right)\sum_{f_k\notin\mathcal{T}\backslash\{f_j\}}\boldsymbol{\alpha}_j\boldsymbol{\alpha}_k^H\mathcal{K}\left(f-f_k\right) \\ &\quad+\mathcal{K}''\left(f-f_j\right)\sum_{f_k\in\mathcal{T}}\boldsymbol{\alpha}_j\boldsymbol{\beta}_k^H\mathcal{K}'\left(f-f_k\right) \\ &\quad+\left[\sum_{f_k\notin\mathcal{T}\backslash\{f_j\}}\boldsymbol{\alpha}_k\mathcal{K}''\left(f-f_k\right)\right]Q\left(f\right)^{H} \\ &\quad+\left[\sum_{f_k\in\mathcal{T}}\boldsymbol{\beta}_k\mathcal{K}'''\left(f-f_k\right)\right]Q\left(f\right)^{H}, \end{aligned}$$

$$\begin{aligned} \left\|\boldsymbol{\alpha}_j\right\|_2^2\mathcal{K}''\left(f-f_j\right)\mathcal{K}\left(f-f_j\right) &\le \left(\alpha^{\text{min}}\right)^2\times\left(-11.69n^2\right)\times 0.9539 \\ &= -10.96n^2, \end{aligned}$$

and

$$\begin{aligned} \left\|Q\left(f\right)\right\|_2 &\le \alpha^{\text{max}}\left|\mathcal{K}\left(f-f_j\right)\right|+\alpha^{\text{max}}\sum_{f_k\notin\mathcal{T}\backslash\{f_j\}}\left|\mathcal{K}\left(f-f_k\right)\right| \\ &\quad+\beta^{\text{max}}\sum_{f_k\in\mathcal{T}}\left|\mathcal{K}'\left(f-f_k\right)\right| \\ &\le 1.036. \end{aligned}$$

It follows that

$$\begin{aligned} \Re\left\{Q''\left(f\right)Q\left(f\right)^{H}\right\} &\le -10.96n^2 \\ &\quad+\left(\alpha^{\text{max}}\right)^2\left|\mathcal{K}''\left(f-f_j\right)\right|\sum_{f_k\notin\mathcal{T}\backslash\{f_j\}}\left|\mathcal{K}\left(f-f_k\right)\right| \\ &\quad+\alpha^{\text{max}}\beta^{\text{max}}\left|\mathcal{K}''\left(f-f_j\right)\right|\sum_{f_k\in\mathcal{T}}\left|\mathcal{K}'\left(f-f_k\right)\right| \\ &\quad+\alpha^{\text{max}}\sum_{f_k\notin\mathcal{T}\backslash\{f_j\}}\left|\mathcal{K}''\left(f-f_k\right)\right|\left\|Q\left(f\right)\right\|_2 \\ &\quad+\beta^{\text{max}}\sum_{f_k\in\mathcal{T}}\left|\mathcal{K}'''\left(f-f_k\right)\right|\left\|Q\left(f\right)\right\|_2 \\ &\le -2.875n^2. \end{aligned}$$

Finally, we conclude that

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}^2 \|Q(f)\|_2^2}{\mathrm{d}f^2} &= \|Q'(f)\|_2^2 + \Re\left\{Q''(f)\,Q(f)^H\right\} \\ &\leq -0.3910n^2.\end{aligned}$$

6 Proof of Theorem 3.3

6.1 Dual Certificate

For incomplete data, we also begin with the dual certificate stated in the proposition below, which generalizes Proposition 5.1 for complete data.

Proposition 6.1. *Suppose that the atomic set $\mathcal{A}$ is composed of atoms defined by $\boldsymbol{a}(f, \boldsymbol{\phi})$ whose rows are indexed by the set $\boldsymbol{J}$ being either $\{-2n, \ldots, 2n\}$ or $\{0, \cdots, N-1\}$. Then $\boldsymbol{Y}^o = \sum_{k=1}^K c_k \boldsymbol{a}(f_k, \boldsymbol{\phi}_k)$ is the unique optimizer to (22) or (23) if $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}} \subset \mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent and there exists a vector-valued dual polynomial $Q : \mathbb{T} \to \mathbb{C}^{1\times L}$*

$$Q(f) = \boldsymbol{a}(f)^H \boldsymbol{V} \tag{48}$$

satisfying that

$$Q(f_k) = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \tag{49}$$

$$\|Q(f)\|_2 < 1, \quad f \in \mathbb{T}\backslash\mathcal{T}, \tag{50}$$

$$\boldsymbol{V}_j = \boldsymbol{0}_{L\times 1}, \quad j \notin \boldsymbol{\Omega}, \tag{51}$$

where the coefficient matrix $\boldsymbol{V} \in \mathbb{C}^{|\boldsymbol{J}|\times L}$. Moreover, $\boldsymbol{Y}^o = \sum_{k=1}^K c_k \boldsymbol{a}(f_k, \boldsymbol{\phi}_k)$ is the unique atomic decomposition satisfying that $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{k=1}^K c_k$.

Proof. The proof is similar to that of Proposition 5.1. Note that the dual problem of (22) or (23) is (25), and strong duality holds between them. We can similarly show that $\|\boldsymbol{V}\|_{\mathcal{A}}^* \leq 1$ and $\langle \boldsymbol{V}_{\boldsymbol{\Omega}}, \boldsymbol{Y}_{\boldsymbol{\Omega}}^o\rangle_{\mathbb{R}} = \|\boldsymbol{Y}^o\|_{\mathcal{A}}$. Since $(\boldsymbol{Y}^o, \boldsymbol{V})$ is primal-dual feasible, we conclude that $\boldsymbol{Y}^o$ is a primal optimal solution due to the strong duality. Suppose that there exists another optimal solution $\widetilde{\boldsymbol{Y}} \neq \boldsymbol{Y}^o$ with an atomic decomposition $\widetilde{\boldsymbol{Y}} = \sum_k \widetilde{c}_k \boldsymbol{a}\left(\widetilde{f}_k, \widetilde{\boldsymbol{\phi}}_k\right)$ satisfying that $\left\|\widetilde{\boldsymbol{Y}}\right\|_{\mathcal{A}} = \sum_k \widetilde{c}_k = \|\boldsymbol{Y}^o\|_{\mathcal{A}}$. Since $\widetilde{\boldsymbol{Y}}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ and that $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}} \subset \mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent, there must exist some $\widetilde{f}_k \notin \mathcal{T}$, otherwise the two decompositions will be identical and $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}^o$. Consequently, we have the following contradiction:

$$\begin{aligned}\sum_k \widetilde{c}_k = \left\|\widetilde{\boldsymbol{Y}}\right\|_{\mathcal{A}} &= \left\langle \boldsymbol{V}, \widetilde{\boldsymbol{Y}}\right\rangle_{\mathbb{R}} \\ &= \left\langle \boldsymbol{V}, \sum_k \widetilde{c}_k \boldsymbol{a}\left(\widetilde{f}_k, \widetilde{\boldsymbol{\phi}}_k\right)\right\rangle_{\mathbb{R}} \\ &= \sum_{\widetilde{f}_k \in \mathcal{T}} \widetilde{c}_k \left\langle Q\left(\widetilde{f}_k\right), \widetilde{\boldsymbol{\phi}}_k\right\rangle_{\mathbb{R}} + \sum_{\widetilde{f}_k \notin \mathcal{T}} \widetilde{c}_k \left\langle Q\left(\widetilde{f}_k\right), \widetilde{\boldsymbol{\phi}}_k\right\rangle_{\mathbb{R}} \\ &< \sum_{\widetilde{f}_k \in \mathcal{T}} \widetilde{c}_k + \sum_{\widetilde{f}_k \notin \mathcal{T}} \widetilde{c}_k = \sum_k \widetilde{c}_k,\end{aligned}$$

showing that $\boldsymbol{Y}^o$ is the unique optimizer. With similar arguments, we can conclude the uniqueness of the atomic decomposition of $\boldsymbol{Y}^o$. ■

Remark 6.1. *The condition that* $\left\{\boldsymbol{a}_{\boldsymbol{\Omega}}\left(f_k\right)\right\}_{f_k\in\mathcal{T}}\subset\mathcal{A}_{\boldsymbol{\Omega}}^1$ *are linearly independent is used to prove the uniqueness of the optimizer to (22) or (23). From this point of view, Proposition II.4 of [16] or at least its proof for the special SMV case is flawed, where this condition is omitted. However, we will show later that it can be satisfied for free when we construct the dual polynomial* $Q\left(f\right)$.

6.2 Revisiting the SMV Case

Similarly to the proof of Theorem 3.2, we first revisit the SMV case in [16], where a scalar-valued dual polynomial is constructed inspired by [13] and proven to satisfy the three constraints listed in Proposition 6.1. While the uniform sampling model is difficult to analyze directly, an equivalent Bernoulli observation model is studied instead following from [5], where we say "equivalent" in the sense that the probability that (22) or (23) fails to recover the original signal $\boldsymbol{Y}^o$ under the uniform model is at most twice of that under the Bernoulli model. Assume in the Bernoulli model that the samples indexed by $\boldsymbol{J}=\left\{-2n,\dots,2n\right\}$ are observed independently with probability $p=\frac{M}{4n}$, i.e., we observe about M entries on average. In mathematics, we let $\left\{\delta_j\right\}_{j\in\boldsymbol{J}}$ be i.i.d. Bernoulli random variables such that

$$\mathbb{P}\left(\delta_j=1\right)=p, \tag{52}$$

where $\delta_j=1$ or 0 indicates whether we observe the jth entry in $\boldsymbol{J}$. It follows that the sampling index set $\boldsymbol{\Omega}=\left\{j:\delta_j=1\right\}$.

To deal with the randomness of the observation model, [16] considers a random dual polynomial

$$\overline{q}\left(f\right)=\sum_{f_k\in\mathcal{T}}\alpha_k\overline{\mathcal{K}}\left(f-f_k\right)+\sum_{f_k\in\mathcal{T}}\beta_k\overline{\mathcal{K}}'\left(f-f_k\right), \tag{53}$$

where $\overline{\mathcal{K}}\left(f\right)$, a random analog of $\mathcal{K}\left(f\right)$, denotes a random kernel as follows:

$$\begin{aligned}\overline{\mathcal{K}}\left(f\right)&=\frac{1}{n+1}\sum_{j\in\boldsymbol{\Omega}}g_n\left(j\right)e^{-i2\pi jf}\\&=\frac{1}{n+1}\sum_{j=-2n}^{2n}\delta_jg_n\left(j\right)e^{-i2\pi jf}\end{aligned} \tag{54}$$

with $g_n\left(j\right)$ defined in (34). It is clear that $\mathbb{E}\overline{\mathcal{K}}\left(f\right)=p\mathcal{K}\left(f\right)$ and similarly for its derivatives. The proof of [16] is mainly based on that the random kernel $\overline{\mathcal{K}}\left(f\right)$ is concentrated tightly around its expectation $p\mathcal{K}\left(f\right)$ as the sample size is large enough. As in the complete data case, let us denote $\left[\overline{D}_0\right]_{jk}=\overline{\mathcal{K}}\left(f_j-f_k\right)$, $\left[\overline{D}_1\right]_{jk}=\overline{\mathcal{K}}'\left(f_j-f_k\right)$ and $\left[\overline{D}_2\right]_{jk}=\overline{\mathcal{K}}''\left(f_j-f_k\right)$, and then obtain the system of linear equations

$$\begin{bmatrix}\overline{\boldsymbol{D}}_0 & c_0^{-1}\overline{\boldsymbol{D}}_1\\ -c_0^{-1}\overline{\boldsymbol{D}}_1 & -c_0^{-2}\overline{\boldsymbol{D}}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ c_0\boldsymbol{\beta}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Phi}\\ \boldsymbol{0}\end{bmatrix} \tag{55}$$

as in (37) by imposing conditions as in (35) and (36), where c_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$ are defined in the same manner. The authors of [16] show that the coefficient matrix $\overline{\boldsymbol{D}}=\begin{bmatrix}\overline{\boldsymbol{D}}_0 & c_0^{-1}\overline{\boldsymbol{D}}_1\\ -c_0^{-1}\overline{\boldsymbol{D}}_1 & -c_0^{-2}\overline{\boldsymbol{D}}_2\end{bmatrix}$ in (55) is concentrated tightly around its expectation $p\boldsymbol{D}$, where $\boldsymbol{D}$ is the deterministic analog of $\overline{\boldsymbol{D}}$. That means, $\overline{\boldsymbol{D}}$ is close to $p\boldsymbol{I}$ with high probability, and hence, $\begin{bmatrix}\boldsymbol{\alpha}\\ c_0\boldsymbol{\beta}\end{bmatrix}$ is well defined by (55) and close to $p^{-1}\begin{bmatrix}\boldsymbol{\Phi}\\ \boldsymbol{0}\end{bmatrix}$ with high probability.

According to (53) and (55), $\overline{q}\left(f\right)$ satisfies the interpolation condition (49) and the support condition (51). To further prove that it obeys (50), the authors show that the random perturbation between $\overline{q}\left(f\right)$ (and its derivatives) and its deterministic analog $q\left(f\right)$ (and its derivatives) in Section 5.2 can be arbitrarily small provided the sample size $|\boldsymbol{\Omega}|$ is sufficiently large, firstly on a set of grid points and then extended to the continuous domain $\mathbb{T}$. The proof is completed by showing that $|\overline{q}(f)|$ (or its second derivative) can be arbitrarily close to $|q(f)|$ (or its second derivative) on $\mathcal{T}_{\text{far}}$ (or $\mathcal{T}_{\text{near}}$).

We list some useful results in [16] which are also necessary in our later proof. We start with some notations. For $\tau \in \left(0, \frac{1}{4}\right]$, define the event

$$\mathcal{E}_{1,\tau} = \left\{ \left\| p^{-1}\overline{\boldsymbol{D}} - \boldsymbol{D} \right\|_2 \leq \tau \right\}.$$

On $\mathcal{E}_{1,\tau}$, $\overline{\boldsymbol{D}}$ is guaranteed to be invertible. Then we introduce the partitions $\overline{\boldsymbol{D}}^{-1} = \left[\overline{\boldsymbol{L}}\ \overline{\boldsymbol{R}}\right]$ and $\boldsymbol{D}^{-1} = \left[\boldsymbol{L}\ \boldsymbol{R}\right]$, where $\overline{\boldsymbol{L}}$, $\overline{\boldsymbol{R}}$, $\boldsymbol{L}$ and $\boldsymbol{R}$ are of the same dimension. Let

$$\overline{\boldsymbol{v}}_l(f) = c_0^{-l} \begin{bmatrix} \overline{\mathcal{K}}^{(l)}(f - f_1)^H \\ \vdots \\ \overline{\mathcal{K}}^{(l)}(f - f_K)^H \\ c_0^{-1}\overline{\mathcal{K}}^{(l+1)}(f - f_1)^H \\ \vdots \\ c_0^{-1}\overline{\mathcal{K}}^{(l+1)}(f - f_K)^H \end{bmatrix} \tag{56}$$

and similarly define its deterministic analog $\boldsymbol{v}_l(f)$, where $\overline{\mathcal{K}}^{(l)}$ denotes the lth derivative of $\overline{\mathcal{K}}$ and $f_j \in \mathcal{T}$. Let δ be a small positive number and C a constant which may vary from instance to instance.

Lemma 6.1. *Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$. We have $\mathbb{P}(\mathcal{E}_{1,\tau}) \geq 1 - \delta$ if*

$$M \geq \frac{50}{\tau^2} K \log \frac{2K}{\delta}.$$

Lemma 6.2. *Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$. Let $\tau \in \left(0, \frac{1}{4}\right]$ and consider a finite set $\mathbb{T}_{grid} = \{f_d\} \subset \mathbb{T}$. Then, we have*

$$\begin{aligned} &\mathbb{P}\left[\sup_{f_d \in \mathbb{T}_{grid}} \left\| \overline{\boldsymbol{L}}^H \left(\overline{\boldsymbol{v}}_l(f_d) - p\boldsymbol{v}_l(f_d)\right) \right\|_2 \geq 4\left(2^{2l+3}\sqrt{\frac{K}{M}} + \frac{n}{M} a \overline{\sigma}_l\right), l = 0, 1, 2, 3\right] \\ &\leq 64\left|\mathbb{T}_{grid}\right| e^{-\gamma a^2} + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right) \end{aligned}$$

for some constant $\gamma > 0$, where $\overline{\sigma}_l^2 = 2^{4l+1}\frac{M}{n^2}\max\left\{1, \frac{2^4 K}{\sqrt{M}}\right\}$ and $0 < a \leq \begin{cases} \sqrt{2}M^{\frac{1}{4}}, & \text{if } \frac{2^4 K}{\sqrt{M}} \geq 1, \\ \frac{\sqrt{2}}{4}\sqrt{\frac{M}{K}}, & \text{otherwise.} \end{cases}$

Lemma 6.3. *Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$. On the event $\mathcal{E}_{1,\tau}$, we have*

$$\left\| \left(\overline{\boldsymbol{L}} - p^{-1}\boldsymbol{L}\right)^H p\boldsymbol{v}_l(f) \right\|_2 \leq C\tau$$

for some constant $C > 0$.

6.3 Proof of Theorem 3.3

Theorem 3.3 is a direct result of the following proposition.

Proposition 6.2. *Under the assumptions of Theorem 3.3, then there exists a numerical constant C such that*

$$M \geq C \max\left\{\log^2 \frac{\sqrt{L}N}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{L}N}{\delta}\right\}$$

is sufficient to guarantee that with probability at least $1-\delta$ there exists a vector-valued polynomial as in (48) satisfying (49)–(51) for $\boldsymbol{J}$ being either $\{-2n, \ldots, 2n\}$ or $\{0, \cdots, N-1\}$, where $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}} \subset \mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent.

Similarly to the case of complete data, we first argue that we can consider only the symmetric case where $\boldsymbol{J} = \{-2n, \ldots, 2n\}$. In particular, we attempt to show that, under the assumptions of Proposition 6.2, if we can construct a vector-valued dual polynomial in the symmetric case, then a similar polynomial can be obtained in the general case.

Our arguments are similar to those in Subsection 5.3 for complete data. We omit the details but emphasize some important features. For a fixed set of frequencies $\mathcal{T}$, $\left\{\widetilde{\boldsymbol{\phi}}_k = e^{i2\pi(2n)f_k}\boldsymbol{\phi}_k\right\}_{k=1}^K$ defined in (41) are mutually independent, with $\left\|\widetilde{\boldsymbol{\phi}}_k\right\|_2 = 1$ and $\mathbb{E}\widetilde{\boldsymbol{\phi}}_k = \boldsymbol{0}$ under the same assumptions of $\{\boldsymbol{\phi}_k\}_{k=1}^K$. Moreover, the linear independence of $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}}$ remains unchanged.

The rest of the proof is to show the existence of a dual polynomial as in Proposition 6.2 in the symmetric case. Based on Lemma 6.1 we can easily obtain the following result, which shows the linear independence of $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}}$.

Lemma 6.4. *Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$. Then there exists a numerical constant C such that $M \geq CK \log \frac{K}{\delta}$ is sufficient to guarantee that, with probability at least $1-\delta$, $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}}$ are linearly independent.*

Proof. According to (54), we have

$$\begin{aligned}\left[\overline{D}_0\right]_{jk} &= \overline{\mathcal{K}}(f_j - f_k) \\ &= \frac{1}{n+1}\sum_{l \in \boldsymbol{\Omega}} g_n(l)\, e^{-i2\pi l(f_j - f_k)} \\ &= \frac{1}{n+1}\boldsymbol{a}_{\boldsymbol{\Omega}}(f_j)^H \boldsymbol{G}\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k),\end{aligned}$$

where $\boldsymbol{G} = \operatorname{diag}(g_n(l))_{l \in \boldsymbol{\Omega}}$ is a diagonal matrix composed of the entries $\{g_n(l)\}$, $l \in \boldsymbol{\Omega}$, and thus has full rank. Denote $\boldsymbol{A}_{\boldsymbol{\Omega}} = [\boldsymbol{a}_{\boldsymbol{\Omega}}(f_1), \ldots, \boldsymbol{a}_{\boldsymbol{\Omega}}(f_K)] \in \mathbb{C}^{M \times K}$, i.e., the matrix formed by $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}}$. It follows that

$$\boldsymbol{D}_0 = \frac{1}{n+1}\boldsymbol{A}_{\boldsymbol{\Omega}}^H \boldsymbol{G}\boldsymbol{A}_{\boldsymbol{\Omega}}.$$

On the event $\mathcal{E}_{1,\tau}$, since $\boldsymbol{D}$, as well as $\boldsymbol{D}_0$, is invertible, $\boldsymbol{A}_{\boldsymbol{\Omega}}$ must have full column rank provided $M \geq K$, i,e., $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}}$ are linearly independent. Let us fix $\tau = \frac{1}{4}$. Then the bound of M follows directly from Lemma 6.1. ■

Remark 6.2. *For a specific sampling index set $\boldsymbol{\Omega}$, it is generally infeasible to compute the quantity spark $\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right)$ as mentioned in Section 2.2. Here we show that if $\boldsymbol{\Omega}$ is selected uniformly at random, then $M \geq O(K \log K)$ is sufficient to guarantee that with high probability any K atoms in $\mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent under a frequency separation condition. Define the* conditional spark *of $\mathcal{A}_{\boldsymbol{\Omega}}^1$ as*

$$\textit{spark}_c\left(\mathcal{A}_{\boldsymbol{\Omega}}^1, \Delta\right) = \inf\left\{\widehat{K} : \{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{k=1}^{\widehat{K}} \subset \mathcal{A}_{\boldsymbol{\Omega}}^1 \ \textit{ are linearly dependent with } \Delta_{\{f_k\}} \geq \Delta\right\}.$$

Then we have $\textit{spark}_c\left(\mathcal{A}_{\boldsymbol{\Omega}}^1, \frac{1}{n}\right) \geq K+1$. This will be of independent interest.

Inspired by [16], we construct the following vector-valued polynomial

$$\overline{Q}(f) = \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \overline{\mathcal{K}}(f - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \overline{\mathcal{K}}'(f - f_k), \tag{57}$$

where $\overline{\mathcal{K}}$ is the random kernel, and $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are vector-valued coefficients as in the complete data case. It is clear by (54) that $\overline{Q}(f)$ above satisfies the support condition (51). To make it satisfy (49) and (50), we impose again that

$$\begin{aligned} \overline{Q}(f_j) &= \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \overline{\mathcal{K}}(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \overline{\mathcal{K}}'(f_j - f_k) = \boldsymbol{\phi}_j, \\ \overline{Q}'(f_j) &= \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k \overline{\mathcal{K}}'(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \boldsymbol{\beta}_k \overline{\mathcal{K}}''(f_j - f_k) = \mathbf{0}, \end{aligned}$$

i.e.,

$$\overline{\boldsymbol{D}} \begin{bmatrix} \boldsymbol{\alpha} \\ c_0 \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} \\ \mathbf{0} \end{bmatrix},$$

where $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T\right]^T \in \mathbb{C}^{K \times L}$ and similarly for $\boldsymbol{\beta} \in \mathbb{C}^{K \times L}$. It follows that, on the event $\mathcal{E}_{1,\tau}$, $\begin{bmatrix} \boldsymbol{\alpha} \\ c_0 \boldsymbol{\beta} \end{bmatrix} = \overline{\boldsymbol{D}}^{-1} \begin{bmatrix} \boldsymbol{\Phi} \\ \mathbf{0} \end{bmatrix} = \overline{\boldsymbol{L}} \boldsymbol{\Phi}$, and

$$\begin{aligned} c_0^{-l} \overline{Q}^{(l)}(f) &= \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k c_0^{-l} \overline{\mathcal{K}}^{(l)}(f - f_k) + \sum_{f_k \in \mathcal{T}} c_0 \boldsymbol{\beta}_k \cdot c_0^{-(l+1)} \overline{\mathcal{K}}^{(l+1)}(f - f_k) \\ &= \overline{\boldsymbol{v}}_l(f)^H \overline{\boldsymbol{L}} \boldsymbol{\Phi} \triangleq \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) \right\rangle, \end{aligned} \tag{58}$$

where $\overline{\boldsymbol{v}}_l(f)$ is defined in (56). As in [16], we decompose $\overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f)$ into three parts:

$$\overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) = \boldsymbol{L}^H \boldsymbol{v}_l(f) + \overline{\boldsymbol{L}}^H (\overline{\boldsymbol{v}}_l(f) - p\boldsymbol{v}_l(f)) + \left(\overline{\boldsymbol{L}} - p^{-1}\boldsymbol{L}\right)^H p\boldsymbol{v}_l(f), \tag{59}$$

which results in a decomposition on $c_0^{-l} \overline{Q}^{(l)}(f)$:

$$\begin{aligned} c_0^{-l} \overline{Q}^{(l)}(f) &= \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) \right\rangle \\ &= \left\langle \boldsymbol{\Phi}, \boldsymbol{L}^H \boldsymbol{v}_l(f) \right\rangle + \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H (\overline{\boldsymbol{v}}_l(f) - p\boldsymbol{v}_l(f)) \right\rangle \\ &\quad + \left\langle \boldsymbol{\Phi}, \left(\overline{\boldsymbol{L}} - p^{-1}\boldsymbol{L}\right)^H p\boldsymbol{v}_l(f) \right\rangle \\ &= c_0^{-l} Q^{(l)}(f) + I_1^l(f) + I_2^l(f), \end{aligned} \tag{60}$$

where $Q^{(l)}(f)$ is the dual polynomial (42) in the complete data case with $c_0^{-l} Q^{(l)}(f) = \left\langle \boldsymbol{\Phi}, \boldsymbol{L}^H \boldsymbol{v}_l(f) \right\rangle$. Here we have defined

$$\begin{aligned} I_1^l(f) &= \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H (\overline{\boldsymbol{v}}_l(f) - p\boldsymbol{v}_l(f)) \right\rangle, \\ I_2^l(f) &= \left\langle \boldsymbol{\Phi}, \left(\overline{\boldsymbol{L}} - p^{-1}\boldsymbol{L}\right)^H p\boldsymbol{v}_l(f) \right\rangle. \end{aligned}$$

Denote the event

$$\mathcal{E}_2 = \left\{ \sup_{f_d \in \mathbb{T}_{\text{grid}}} c_0^{-l} \left\| \overline{Q}^{(l)} - Q^{(l)} \right\|_2 \leq \frac{\epsilon}{3}, l = 0, 1, 2, 3 \right\}.$$

We attempt to prove that $\mathcal{E}_2$ happens, i.e., $c_0^{-l} \overline{Q}^{(l)}(f)$ and $c_0^{-l} Q^{(l)}(f)$ are close to each other on a fixed grid, with high probability provided M is sufficiently large, which is summarized in the following proposition.

Proposition 6.3. *Suppose $\mathbb{T}_{grid} \subset \mathbb{T}$ is a finite set of grid points. Under the assumptions of Proposition 6.2, there exists a numerical constant C such that if*

$$M \geq C \frac{1}{\epsilon^2} \max \left\{ \log \frac{\left|\mathbb{T}_{grid}\right|}{\delta} \log \frac{L\left|\mathbb{T}_{grid}\right|}{\delta}, K \log \frac{K}{\delta} \log \frac{L\left|\mathbb{T}_{grid}\right|}{\delta} \right\}, \tag{61}$$

then

$$\mathbb{P}\left(\mathcal{E}_2\right) \geq 1 - \delta.$$

To prove Proposition 6.3, we need to show that $I_1^l(f)$ and $I_2^l(f)$ are small on $\mathbb{T}_{\text{grid}}$. Differently from the SMV case in [16], however, both $I_1^l(f)$ and $I_2^l(f)$ are vector-valued and the difficulty is to provide tight bounds on their norms. To do this, we present the following lemma which corresponds to the vector-form Heoffding's inequality.

Lemma 6.5 (Vector-form Heoffding's inequality)**.** *Let the rows of $\mathbf{\Phi} \in \mathbb{C}^{K \times L}$ be sampled independently on the complex hyper-sphere $\mathbb{S}^{2L-1}$ with zeros means. Then, for all $\boldsymbol{w} \in \mathbb{C}^K$, $\boldsymbol{w} \neq \mathbf{0}$, and $t \geq 0$,*

$$\mathbb{P}\left(\left\|\langle \mathbf{\Phi}, \boldsymbol{w} \rangle\right\|_2 \geq t\right) \leq (L+1)\, e^{-\frac{t^2}{8\|\boldsymbol{w}\|_2^2}}.$$

Proof. See Appendix A. ■

The vector-form Heoffding's inequality generalizes the common scenario with $L = 1$ which is used in [16] to bound the scalars I_1^l and I_2^l by utilizing Lemma 6.2 and Lemma 6.3, respectively. Following the same procedures as in [16] with minor modifications of the coefficients, we can bound I_1^l and I_2^l on $\mathbb{T}_{\text{grid}}$ with high probability, which is summarized in the ensuing Lemma 6.6 and Lemma 6.7 respectively. We omit their proofs and interested readers are referred to the proofs of Lemmas IV.8 and IV.9 in [16].

Lemma 6.6. *Under the assumptions of Proposition 6.2, there exists a numerical constant C such that if*

$$M \geq C \max \left\{ \frac{1}{\epsilon^2} \max \left(K \log \frac{L\left|\mathbb{T}_{grid}\right|}{\delta},\ \ \log \frac{\left|\mathbb{T}_{grid}\right|}{\delta} \log \frac{L\left|\mathbb{T}_{grid}\right|}{\delta} \right), K \log \frac{K}{\delta} \right\},$$

then we have

$$\mathbb{P}\left\{ \sup_{f_d \in \mathbb{T}_{grid}} \left\| I_1^l\left(f_d\right) \right\|_2 \leq \epsilon, l = 0, 1, 2, 3 \right\} \geq 1 - 12\delta.$$

Lemma 6.7. *Under the assumptions of Proposition 6.2, there exists a numerical constant C such that if*

$$M \geq C \frac{1}{\epsilon^2} K \log \frac{K}{\delta} \log \frac{L\left|\mathbb{T}_{grid}\right|}{\delta},$$

then we have

$$\mathbb{P}\left\{ \sup_{f_d \in \mathbb{T}_{grid}} \left\| I_2^l\left(f_d\right) \right\|_2 < \epsilon, l = 0, 1, 2, 3 \right\} \geq 1 - 8\delta.$$

Now we can conclude Proposition 6.3 by combining (60), Lemma 6.6 and Lemma 6.7 with suitable redefinition of C, ϵ and δ.

We have shown in Proposition 6.3 that $c_0^{-l} \overline{Q}^{(l)}(f)$ and $c_0^{-l} Q^{(l)}(f)$ are close to each other on a fixed grid $\mathbb{T}_{\text{grid}} \subset \mathbb{T}$ with high probability. We extend the result to the whole circle $\mathbb{T}$ in the following proposition.

Proposition 6.4. *Under the assumptions of Proposition 6.2, there exists a numerical constant C such that if*

$$M \geq C\frac{1}{\epsilon^2}\max\left\{\log\frac{\sqrt{L}n^3}{\epsilon\delta}\log\frac{L^{\frac{3}{2}}n^3}{\epsilon\delta}, K\log\frac{K}{\delta}\log\frac{L^{\frac{3}{2}}n^3}{\epsilon\delta}\right\}, \tag{62}$$

then with probability $1-\delta$*, we have*

$$\sup_{f\in\mathbb{T}} c_0^{-l}\left\|\overline{Q}^{(l)}(f) - Q^{(l)}(f)\right\|_2 \leq \epsilon, \quad l = 0,1,2,3.$$

Proof. See Appendix B. ■

Now we are ready to show that the vector-valued polynomial $\overline{Q}(f)$ constructed above satisfies (50), and so complete the proof together with Lemma 6.4.

Lemma 6.8. *Under the assumptions of Proposition 6.2, there exists constant C such that if*

$$M \geq C\max\left\{\log^2\frac{\sqrt{L}n}{\delta}, K\log\frac{K}{\delta}\log\frac{\sqrt{L}n}{\delta}\right\},$$

then with probability $1-\delta$ *we have*

$$\left\|\overline{Q}(f)\right\|_2 < 1, \quad f \in \mathbb{T}\backslash\mathcal{T}.$$

Proof. The proof is based on Proposition 6.4 and the results in the complete data case in Section 6.3. On $\mathcal{T}_{\text{far}}$, we have

$$\left\|\overline{Q}(f)\right\|_2 \leq \|Q(f)\|_2 + \left\|\overline{Q}(f) - Q(f)\right\|_2 \leq 0.99992 + \epsilon.$$

So, it suffices to choose $\epsilon = 10^{-5}$. On each interval $[f_j - \nu, f_j + \nu]$, $f_j \in \mathcal{T}$, we also accomplish the proof by showing that

$$\frac{1}{2}\frac{\mathrm{d}^2\left\|\overline{Q}(f)\right\|_2^2}{\mathrm{d}f^2} = \left\|\overline{Q}'(f)\right\|_2^2 + \Re\left\{\overline{Q}''(f)\overline{Q}(f)^H\right\} < 0.$$

According to the results in Section 6.3, it is easy to verify that

$$\left\|c_0^{-l}Q^{(l)}\right\|_2 \leq C_1, \quad l = 0,1,2,$$

for a numerical constant C_1. Let $\Delta_Q = \overline{Q} - Q$. By Proposition 6.4 we have

$$\left\|c_0^{-l}\Delta_Q^{(l)}\right\|_2 \leq \epsilon, \quad l = 0,1,2.$$

Moreover, with simple derivations we have

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}^2\left\|\overline{Q}(f)\right\|_2^2}{\mathrm{d}f^2} - \frac{1}{2}\frac{\mathrm{d}^2\|Q(f)\|_2^2}{\mathrm{d}f^2}\\&= \Re\left\{2Q'\Delta_Q'^H + Q''\Delta_Q^H + \Delta_Q''Q^H + \Delta_Q''\Delta_Q^H\right\} + \left\|\Delta_Q'\right\|_2^2.\end{aligned}$$

It follows that

$$\begin{aligned}&c_0^{-2}\left|\frac{1}{2}\frac{\mathrm{d}^2\left\|\overline{Q}(f)\right\|_2^2}{\mathrm{d}f^2} - \frac{1}{2}\frac{\mathrm{d}^2\|Q(f)\|_2^2}{\mathrm{d}f^2}\right|\\&\leq 2\left\|c_0^{-1}Q'\right\|_2\left\|c_0^{-1}\Delta_Q'\right\|_2 + \left\|c_0^{-2}Q''\right\|_2\|\Delta_Q\|_2\\&\quad + \left\|c_0^{-2}\Delta_Q''\right\|_2\|Q\|_2 + \left\|c_0^{-2}\Delta_Q''\right\|_2\|\Delta_Q\|_2 + \left\|c_0^{-1}\Delta_Q'\right\|_2^2\\&\leq 4C_1\epsilon + 2\epsilon^2.\end{aligned}$$

Recall that $c_0^2 \leq \frac{8\pi^2}{3}n^2$ as $n \geq 2$. So we have

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}^2\left\|\overline{Q}\left(f\right)\right\|_2^2}{\mathrm{d}f^2} &\leq \frac{1}{2}\frac{\mathrm{d}^2\left\|Q\left(f\right)\right\|_2^2}{\mathrm{d}f^2} + \left|\frac{1}{2}\frac{\mathrm{d}^2\left\|\overline{Q}\left(f\right)\right\|_2^2}{\mathrm{d}f^2} - \frac{1}{2}\frac{\mathrm{d}^2\left\|Q\left(f\right)\right\|_2^2}{\mathrm{d}f^2}\right| \\ &\leq -0.3910n^2 + c_0^2\left(4C_1\epsilon + 2\epsilon^2\right) \\ &\leq \left[-0.3910 + \frac{8}{3}\pi^2\left(4C_1\epsilon + 2\epsilon^2\right)\right]n^2 \\ &< 0\end{aligned}$$

as ϵ is a sufficiently small numerical value. With this choice of ϵ, we obtain from (62) the following (slightly stronger) condition

$$M \geq C\max\left\{\log^2\frac{\sqrt{L}n}{\delta}, K\log\frac{K}{\delta}\log\frac{\sqrt{L}n}{\delta}\right\}.$$

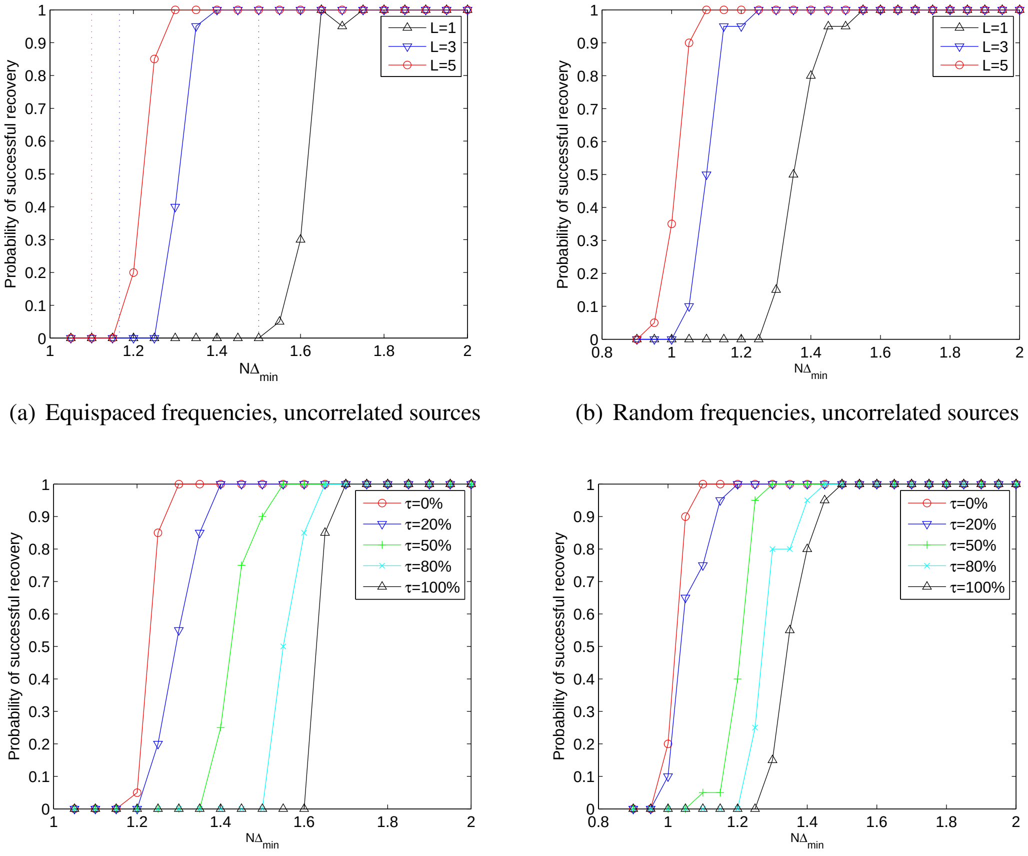
■

7 Numerical Simulations

7.1 Complete Data

We consider the complete data case and test the frequency recovery performance of the proposed atomic norm method with respect to the frequency separation condition. In particular, we consider two types of frequencies, equispaced and random, and two types of source signals, uncorrelated and coherent. We fix $N = 128$ and vary Δ_{min} (a lower bound of the minimum separation of frequencies) from $1.05N^{-1}$ (or $0.9N^{-1}$ for random frequencies) to $2N^{-1}$ at a step of $0.05N^{-1}$. In the case of equispaced frequencies, for each Δ_{min} we select a set of frequencies $\mathcal{T}$ of the maximal cardinality $\lfloor\Delta_{\text{min}}^{-1}\rfloor$ with the frequency separation $\Delta_{\mathcal{T}} = \frac{1}{\lfloor\Delta_{\text{min}}^{-1}\rfloor} \geq \Delta_{\text{min}}$. In the case of random frequencies, we generate the frequency set $\mathcal{T}$, $\Delta_{\mathcal{T}} \geq \Delta_{\text{min}}$, by repetitively adding new frequencies (generated uniformly at random) till no more can be added. Therefore, any two adjacent frequencies in $\mathcal{T}$ are separate by a value in the interval $[\Delta_{\text{min}}, 2\Delta_{\text{min}})$. It follows that $|\mathcal{T}| \in \left(\frac{1}{2}\Delta_{\text{min}}^{-1}, \Delta_{\text{min}}^{-1}\right]$. We empirically find that $\mathbb{E}\,|\mathcal{T}| \approx \frac{3}{4}\Delta_{\text{min}}^{-1}$ which is the mid-point of the interval above.

We first consider uncorrelated sources, where the source signals $\boldsymbol{S} = [s_{kt}] \in \mathbb{C}^{K\times L}$ in (1) are drawn i.i.d. from a standard complex Gaussian distribution. Moreover, we consider the number of measurement vectors $L = 1, 3,$ and 5. For each value of Δ_{min} and each type of frequencies, we carry out 20 Monte Carlo runs and calculate the success rate of frequency recovery. In each run, we generate a set of frequencies $\mathcal{T}$ and $\boldsymbol{S} \in \mathbb{C}^{K\times 5}$ and obtain the complete data $\boldsymbol{Y}^o$. For each value of L, we attempt to recover the frequencies by the proposed atomic norm method, implemented by SDPT3 [51] in Matlab, based on the first L columns of $\boldsymbol{Y}^o$. Given the frequency solution $\mathcal{T}^* = \left\{f_j^*\right\}_{j=1}^{K^*}$ and *mean amplitude* $\left\{\check{c}_j^* \triangleq \frac{c_j^*}{\sqrt{L}}\right\}_{j=1}^{K^*}$ (see Section 3.2), we denote by $\check{\boldsymbol{c}}_K^*$ the vector of the largest K amplitudes (the rest by $\check{\boldsymbol{c}}_{-K}^*$) and by $\mathcal{T}_K^*$ the corresponding set of frequencies. We may consider the rest of the frequencies in $\mathcal{T}^*$ are spurious peaks with amplitudes $\check{\boldsymbol{c}}_{-K}^*$. The recovery is considered successful if $K^* \geq K$, root mean squared error (RMSE) of frequency $\frac{\left\|\mathcal{T}_K^* - \mathcal{T}\right\|_2}{\sqrt{K}} < 1 \times 10^{-8}$ and maximum absolute error of amplitude $\left\|\begin{bmatrix}\check{\boldsymbol{c}}_K^* - \check{\boldsymbol{c}} \\ \check{\boldsymbol{c}}_{-K}^*\end{bmatrix}\right\|_\infty < 1 \times 10^{-4}$, where $\check{\boldsymbol{c}}$ denotes the true vector of amplitudes. Our simulation results are presented in Figs. 1(a) and 1(b), which verify the conclusion of Theorem 3.2 that the frequencies can be exactly recovered by the proposed atomic norm method under a frequency separation condition. Moreover, when we take more measurement vectors, the performance of recovery improves and it seems that a weaker frequency separation condition is sufficient to guarantee exact frequency recovery in this case of uncorrelated sources. By comparing Fig. 1(a) and Fig. 1(b), we



(a) Equispaced frequencies, uncorrelated sources

(b) Random frequencies, uncorrelated sources

(c) $L = 5$, equispaced frequencies, τ coherent sources

(d) $L = 5$, random frequencies, τ coherent sources

Figure 1: Results of probability of successful frequency recovery with respect to the minimum frequency separation Δ_{min}. The vertical dotted lines in (a) indicate the necessary frequency separation condition that $\Delta_{\mathcal{T}} \geq \left(1 + \frac{1}{2L}\right) N^{-1}$.

also observe that a stronger frequency separation condition is required in the case of equispaced frequencies where more frequencies are present and located more closely. It is worthy noting that with the equispaced frequencies the minimum separation at which the atomic norm method starts to succeed is close to the necessary separation condition in Remark 3.3 that $\Delta_{\mathcal{T}} \geq \left(1 + \frac{1}{2L}\right) N^{-1}$.

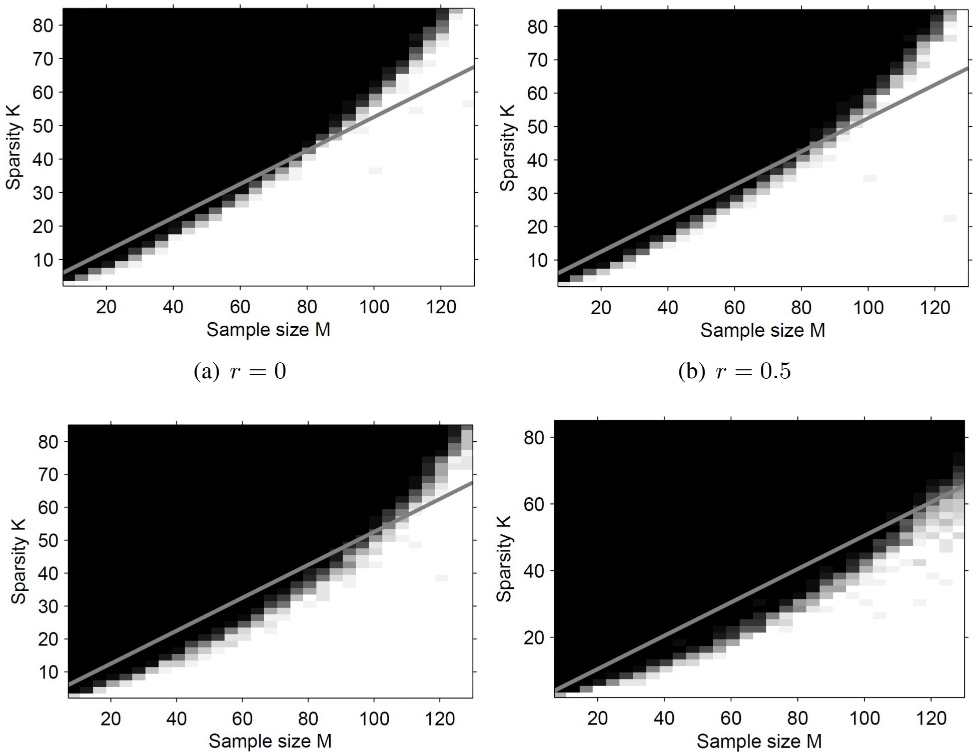
We next consider coherent sources. In this simulation, we fix $L = 5$ and consider different percentages, denoted by τ, of the K source signals which are coherent (identical up to a complex scale factor). That is, $\tau = 0\%$ refers exactly to the case of uncorrelated sources considered previously. $\tau = 100\%$ means that all the sources signals are coherent and the MMV case is equivalent to the SMV case. For each type of frequencies, we consider five values of τ ranging from 0% to 100% and calculate each success rate over 20 Monte Carlo runs. Our simulation results are presented in Figs. 1(c) and 1(d). It is shown that, as the percentage of coherent sources increases, the success rate decreases and a stronger frequency separation condition is required for exact frequency recovery. As τ equals the extreme value 100%, the curves of success rate approximately match those at $L = 1$ in Figs. 1(a) and 1(b), verifying that taking more measurement vectors does not necessarily improve the performance of frequency recovery. Finally, we report the computational speed of the proposed atomic norm method. It takes about 11s to solve one SDP on average and the CPU times differ slightly at the three values of L. About 22 hours are used in total to produce the data used in Fig. 1.

7.2 Incomplete Data

For incomplete data, we study the so-called phase transition phenomenon in the (M, K) plane. In particular, we fix $N = 128$, $L = 5$ and $\Delta_{\text{min}} = 1.2N^{-1}$, and study the performance of our proposed atomic norm minimization method in signal and frequency recovery with different settings of the source signal. The frequency set $\mathcal{T}$ is randomly generated with $\Delta_{\mathcal{T}} \geq \Delta_{\text{min}}$ and $|\mathcal{T}| = K$ (differently from that in the last subsection, the process of adding frequencies is terminated as $|\mathcal{T}| = K$). In our simulation, we vary $M = 8, 12, \dots, 128$ and at each M, $K = 2, 4, \dots, \min(M, 84)$ since it is difficult to generate a set of frequencies with $K > 84$ under the aforementioned frequency separation condition. In this simulation, we consider temporarily correlated sources. In particular, suppose that each row of source signal $\boldsymbol{S}$ has a Toeplitz covariance matrix $\boldsymbol{R}(r) = \begin{bmatrix} 1 & r & \dots & r^4 \\ r & 1 & \dots & r^3 \\ \vdots & \vdots & \ddots & \vdots \\ r^4 & r^3 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{5\times 5}$ (up to a positive scale factor). Therefore, $r = 0$ means that the source signals at different snapshots are uncorrelated while $r = \pm 1$ means completely correlated. In our simulation, we first generate $\boldsymbol{S}_0$ from an i.i.d. standard complex Gaussian distribution and then let $\boldsymbol{S}(r) = \boldsymbol{S}_0 \boldsymbol{R}(r)^{\frac{1}{2}}$, where we consider $r = 0, 0.5, 0.9, 1$. For each combination (M, K), we carry out 20 Monte Carlo runs and calculate the rate of successful recovery with respect to each r. The recovery is considered successful if $K^* \geq K$, the relative RMSE of data recovery $\|\boldsymbol{Y}^* - \boldsymbol{Y}^o\|_{\text{F}} / \|\boldsymbol{Y}^o\|_{\text{F}} < 1 \times 10^{-8}$, the RMSE of frequency recovery $< 1 \times 10^{-6}$ and the maximum absolute error of amplitude recovery $< 1 \times 10^{-4}$, where K^* and the last two metrics are defined as in the previous simulation, and $\boldsymbol{Y}^*$ denotes the solution of $\boldsymbol{Y}$.

Our simulation results are presented in Fig. 2, where a transition from perfect recovery to complete failure can be observed in each subfigure. More frequencies can be recovered when more samples are observed. Moreover, when the correlations between the MMVs, indicated by r, increase, the phase of successful recovery decreases, and at the same time the phase transition becomes less clear. In the extreme case where $r = 1$, all the measurement vectors are completed correlated, which in fact is equivalent to the SMV case. Therefore, by comparing Fig. 2(d) and the first three subfigures, we can conclude that the performance of frequency recovery improves in general when more measurement vectors are observed. It is also worth noting that $M = 128$ corresponds to the complete data case considered in the previous simulation.

We also plot the line $K = \frac{1}{2}(M + L)$ in Figs. 2(a)-2(c) and $K = \frac{1}{2}(M + 1)$ in Fig. 2(d) (straight gray lines) which are upper bounds of the sufficient condition in Theorem 2.1 for the atomic ℓ_0 norm minimization. We see



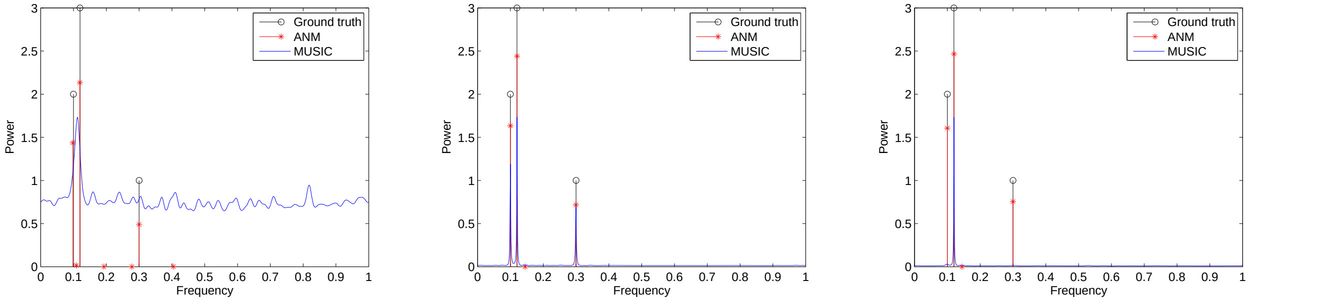
(a) $r = 0$

(b) $r = 0.5$

(c) $r = 0.9$

(d) $r = 1$

Figure 2: Results of phase transition with $N = 128$ and $\Delta_{\text{min}} = 1.2N^{-1}$.



(a) $L = 1$ (b) $L = 5$, with uncorrelated sources (c) $L = 5$, with sources 1 and 3 coherent

Figure 3: Frequency recovery/estimation of atomic norm minimization (ANM) and MUSIC in the presence of noise, with (a) $L = 1$, (b) $L = 5$ and uncorrelated sources, and (c) $L = 5$ and coherent sources.

that successful recoveries can be obtained even above the lines, indicating good performance of the proposed atomic norm minimization method. Again, we report the computational speed. It takes about 13s on average to solve each problem and almost 200 hours in total to generate the whole data set used in Fig. 2.

7.3 The Noisy Case

While this paper has been focused on the noiseless case, one naturally wonders the performance of the proposed method in the practical noisy case. We provide a simple simulation as follows to show this. We set $N = 50$, $M = 20$ with $\boldsymbol{\Omega}$ randomly generated, $K = 3$ sources with frequencies of 0.1, 0.12 and 0.3 and powers of 2, 3 and 1 respectively, and $L = 5$. The signals of each source are generated with constant amplitude and random phases. Complex white Gaussian noise is added to the observed samples with noise variance $\sigma^2 = 0.1$. We attempt to denoise the observed noisy signal $\boldsymbol{Y}^o_{\boldsymbol{\Omega}}$ and recover the frequency components by solving the following optimization problem:

$$\min_{\boldsymbol{Y}} \left\|\boldsymbol{Y}\right\|_{\mathcal{A}}, \text{ subject to } \left\|\boldsymbol{Y}^o_{\boldsymbol{\Omega}} - \boldsymbol{Y}_{\boldsymbol{\Omega}}\right\|_{\text{F}}^2 \leq \epsilon, \tag{63}$$

where ϵ, set to $\left(ML + 2\sqrt{ML}\right)\sigma^2$ (mean + twice standard deviation), upper bounds the Frobenius norm of the noise with large probability. Our simulation results of one Monte Carlo run are presented in Fig. 3. The SMV case is studied in Fig. 3(a), where only the first measurement vector is used for frequency recovery. It is shown that the three frequency components are correctly identified by the atomic norm minimization method while MUSIC fails. The MMV case is studied in Fig. 3(b) with uncorrelated sources, where both atomic norm minimization and MUSIC succeed to identify the three frequency components. The coherent source case is presented in Fig. 3(c), where we modify source 3 in Fig. 3(b) such that it is coherent with source 1. MUSIC fails to detect the coherent sources as expected while the proposed method still performs well. All of the three subfigures show that spurious frequency components can be present using the atomic norm minimization method, however, their powers are insignificant. To be specific, the spurious components have about 0.4% of the total powers in Fig. 3(a), and this number is on the magnitude of 10^{-6} in the latter two subfigures. Since the numerical results imply that the proposed method is robust to noise, a theoretical analysis should be investigated in future studies. The proposed method takes about 1.5s in each scenario. Finally, it is worthy noting that the proposed method requires to know the noise level while MUSIC needs the source number.

8 Conclusion

We studied the joint sparse frequency recovery problem in this paper which arises in practical array processing applications. We presented nonconvex and convex optimization methods for its solution and analyzed their theoretical guarantees under no or very weak assumptions of the source signals. Our results extend the SMV atomic norm methods and their theoretical guarantees in [13, 16] to the MMV case, extend the existing discrete joint sparse recovery framework to the continuous dictionary setting, and provide theoretical guidance for the array processing applications. While this paper is focused on the worst case analysis, it will be interesting to investigate in the future the average case under stronger assumptions of the source signals, in which numerical simulations of this paper suggest that a weaker frequency separation condition is sufficient for exact recovery as the number of measurement vectors increases.

A Proof of Lemma 6.5

The vector-form Heoffding's inequality in Lemma 6.5 is a corollary of the following matrix-form Heoffding's inequality in [55].

Lemma A.1 (Matrix-form Heoffding's inequality). *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices with dimension d, and let $\{\boldsymbol{A}_k\}$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies*

$$\mathbb{E}\boldsymbol{X}_k = \boldsymbol{0} \text{ and } \boldsymbol{X}_k^2 \leq \boldsymbol{A}_k^2 \text{ almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{max}\left(\sum_k \boldsymbol{X}_k\right) \geq t\right\} \leq d e^{-\frac{t^2}{8\sigma^2}},$$

where $\sigma^2 \triangleq \left\|\sum_k \boldsymbol{A}_k^2\right\|_2$ and $\lambda_{max}(\cdot)$ denotes the largest eigenvalue.

We now prove Lemma 6.5 based on the *dilation* technique [56]. In particular, the dilation of a vector $\boldsymbol{v} \in \mathbb{C}^{1\times L}$ is a self-adjoint matrix $\varphi(\boldsymbol{v}) = \begin{bmatrix} 0 & \boldsymbol{v} \\ \boldsymbol{v}^H & \boldsymbol{0} \end{bmatrix} \in \mathbb{C}^{(L+1)\times(L+1)}$. It is not difficult to show that $\varphi(\boldsymbol{v})^2 = \begin{bmatrix} \|\boldsymbol{v}\|_2^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{v}^H\boldsymbol{v} \end{bmatrix} \leq \|\boldsymbol{v}\|_2^2 \boldsymbol{I}_{L+1}$ and $\lambda_{\max}(\varphi(\boldsymbol{v})) = \|\varphi(\boldsymbol{v})\|_2 = \|\boldsymbol{v}\|_2$.

We let $\boldsymbol{X}_k = \varphi\left(w_k^H \boldsymbol{\phi}_k\right)$ and $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k = \varphi\left(\sum_k w_k^H \boldsymbol{\phi}_k\right) = \varphi(\langle \boldsymbol{\Phi}, \boldsymbol{w}\rangle)$, where $\boldsymbol{\phi}_k$ is the kth row of $\boldsymbol{\Phi}$ with $\|\boldsymbol{\phi}_k\|_2 = 1$. It follows that $\{\boldsymbol{X}_k\}$ are independent matrices of dimension $d = L+1$ with $\mathbb{E}\boldsymbol{X}_k = \boldsymbol{0}$ according to the assumptions of $\{\boldsymbol{\phi}_k\}$. By the aforementioned properties of the dilation, we have that $\boldsymbol{X}_k^2 \leq \left\|w_k^H \boldsymbol{\phi}_k\right\|_2^2 \boldsymbol{I}_{L+1} = |w_k|^2 \boldsymbol{I}_{L+1}$ and $\lambda_{\max}(\boldsymbol{Y}) = \|\langle \boldsymbol{\Phi}, \boldsymbol{w}\rangle\|_2$. So, we can choose $\boldsymbol{A}_k^2 = |w_k|^2 \boldsymbol{I}_{L+1}$ and it follows that $\sigma^2 = \left\|\sum_k \boldsymbol{A}_k^2\right\|_2 = \|\boldsymbol{w}\|_2^2$. We complete the proof by applying the matrix-form Heoffding's inequality to $\{\boldsymbol{X}_k\}$:

$$\mathbb{P}\left(\|\langle \boldsymbol{\Phi}, \boldsymbol{w}\rangle\|_2 \geq t\right) \leq (L+1)\, e^{-\frac{t^2}{8\|\boldsymbol{w}\|_2^2}}.$$

B Proof of Proposition 6.4

The proof is based on the following Bernstein's polynomial inequality.

Lemma B.1 (Bernstein's polynomial inequality, [57]). *Let $q(z)$ be any polynomial of degree n on complex numbers with derivative $q'(z)$. Then,*

$$\sup_{|z|\leq 1} \left|q'(z)\right| \leq n \sup_{|z|\leq 1} |q(z)|.$$

On the event $\mathcal{E}_{1,\tau} \cap \mathcal{E}_2$, we make the following decomposition for some $f \in \mathbb{T}$ and $f_d \in \mathbb{T}_{\text{grid}}$:

$$\begin{aligned} & c_0^{-l} \left\| \overline{Q}^{(l)}(f) - Q^{(l)}(f) \right\|_2 \\ & \le c_0^{-l} \left\| \overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right\|_2 + c_0^{-l} \left\| \overline{Q}^{(l)}(f_d) - Q^{(l)}(f_d) \right\|_2 \\ & \quad + c_0^{-l} \left\| Q^{(l)}(f_d) - Q^{(l)}(f) \right\|_2 . \end{aligned} \tag{64}$$

The second term has been bounded in Proposition 6.3. We next provide a upper bound for $c_0^{-l} \left\| \overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right\|_2$, while the same bound is applicable to $c_0^{-l} \left\| Q^{(l)}(f_d) - Q^{(l)}(f) \right\|_2$ under similar arguments. Since $\overline{Q}^{(l)}(f) \in \mathbb{C}^{1\times L}$ is vector-valued and $\|\boldsymbol{v}\|_2 \le \sqrt{L}\|\boldsymbol{v}\|_\infty$ for any $\boldsymbol{v} \in \mathbb{C}^{1\times L}$. We attempt to bound $\overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d)$ elementwise. Denote $\overline{Q}_j^{(l)}(f)$ the jth entry of $\overline{Q}^{(l)}(f)$. Then we have

$$c_0^{-l} \left| \overline{Q}_j^{(l)}(f) \right| \le \left| \left\langle \boldsymbol{\phi}_{:j}, \overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) \right\rangle \right| \le Cn^2$$

for some constant C following from [16] by noticing that $\left\| \boldsymbol{\phi}_{:j} \right\|_2 \le \|\boldsymbol{\Phi}\|_{\text{F}} = \sqrt{K}$, where $\boldsymbol{\phi}_{:j}$ denotes the jth column of $\boldsymbol{\Phi}$. Viewing $c_0^{-l}\overline{Q}_j^{(l)}(f)$ as a polynomial of $z = e^{-i2\pi f}$ of degree $2n$, we get by applying the Bernstein's polynomial inequality that

$$\begin{aligned} & \left| c_0^{-l}\overline{Q}_j^{(l)}(f_a) - c_0^{-l}\overline{Q}_j^{(l)}(f) \right| \\ & \le \left| e^{-i2\pi f_a} - e^{-i2\pi f_b} \right| \sup_z \left| \frac{\mathrm{d} c_0^{-l}\overline{Q}_j^{(l)}(z)}{\mathrm{d}z} \right| \\ & \le \left| e^{-i\pi(f_a+f_b)} \cdot 2\sin\left(\pi\left(-f_a+f_b\right)\right) \right| \cdot 2n \sup_f \left| c_0^{-l}\overline{Q}_j^{(l)}(f) \right| \\ & \le Cn^3 \left| f_a - f_b \right| . \end{aligned}$$

It follows that

$$\begin{aligned} c_0^{-l} \left\| \overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right\|_2 & \le \sqrt{L} c_0^{-l} \left\| \overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right\|_\infty \\ & \le C\sqrt{L}n^3 \left| f_a - f_b \right| . \end{aligned}$$

Then we can select $\mathbb{T}_{\text{grid}}$ satisfying $\left|\mathbb{T}_{\text{grid}}\right| < \frac{3C\sqrt{L}n^3}{\epsilon}$ such that for any $f \in \mathbb{T}$ there exists a point $f_d \in \mathbb{T}_{\text{grid}}$ satisfying that $|f - f_d| \le \frac{\epsilon}{3C\sqrt{L}n^3}$. Consequently, we have $c_0^{-l} \left\| \overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right\|_2 \le \frac{\epsilon}{3}$, which together with (64) gives that on $\mathcal{E}_{1,\tau} \cap \mathcal{E}_2$

$$c_0^{-l} \left\| \overline{Q}^{(l)}(f) - Q^{(l)}(f) \right\|_2 \le \epsilon, \text{ for any } f \in \mathbb{T}.$$

Finally, (62) is a direct consequence of (61) by inserting that $\left|\mathbb{T}_{\text{grid}}\right| < \frac{3C\sqrt{L}n^3}{\epsilon}$. When (62) is satisfied, we have $\mathbb{P}\left(\mathcal{E}_{1,\tau} \cap \mathcal{E}_2\right) \ge 1 - 2\delta$ by Proposition 6.3 and Lemma 6.1.

Acknowledgement

Z. Yang would like to thank Gongguo Tang for fruitful discussions.

References

[1] Z. Yang and L. Xie, "Continuous compressed sensing with a single or multiple measurement vectors," in *2014 IEEE Workshop on Statistical Signal Processing (SSP), Gold Coast, Australia, available online at https://dl.dropboxusercontent.com/u/34897711/SSP14.pdf*, June 2014.

[2] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson/Prentice Hall Upper Saddle River, NJ, 2005.

[3] Y. Wang, J. Li, and P. Stoica, "Spectral analysis of signals: the missing data case," *Synthesis Lectures on Signal Processing Series*, vol. 1, no. 1, pp. 1–102, 2006.

[4] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 31–40, 2008.

[5] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[6] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[8] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[9] L. Hu, Z. Shi, J. Zhou, and Q. Fu, "Compressed sensing of complex sinusoids: An approach based on dictionary refinement," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3809–3822, 2012.

[10] Z. Yang, C. Zhang, and L. Xie, "Robustly stable signal recovery in compressed sensing with structured matrix perturbation," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4658–4671, 2012.

[11] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, 2013.

[12] C. Austin, J. Ash, and R. Moses, "Dynamic dictionary algorithms for model order and parameter estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5117–5130, 2013.

[13] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics, DOI: 10.1002/cpa.21455*, 2013.

[14] S. Aleksanyan, A. Apozyan, V. Z. Dumanyan, K. A. Khachatryan, E. Nazari, A. Pahlevanyan, and H. Rostami, "Real and complex analysis," *Mathematics in Armenia*, vol. 54, p. 21, 1944.

[15] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[16] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[17] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing, DOI: 10.1109/TSP.2013.2273443*, 2013.

[18] E. J. Candès and C. Fernandez-Granda, “Super-resolution from noisy data,” *Journal of Fourier Analysis and Applications, DOI: 10.1007/s00041-013-9292-3*, 2013.

[19] Y. Chen and Y. Chi, “Robust spectral compressed sensing via structured matrix completion,” *arXiv preprint arXiv:1304.8126*, 2013.

[20] J.-M. Azais, Y. De Castro, and F. Gamboa, “Spike detection from inaccurate samplings,” *arXiv preprint arXiv:1301.5873*, 2013.

[21] G. Tang, B. N. Bhaskar, and B. Recht, “Near minimax line spectral estimation,” *Available online at http://arxiv.org/abs/1303.4348*, 2013.

[22] Z. Yang and L. Xie, “On gridless sparse methods for line spectral estimation from complete and incomplete data,” *Available at https://dl.dropboxusercontent.com/u/34897711/GLS.pdf*, 2014.

[23] H. Krim and M. Viberg, “Two decades of array signal processing research: The parametric approach,” *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[24] R. Schmidt, “A signal subspace approach to multiple emitter location spectral estimation,” Ph.D. dissertation, Stanford University, 1981.

[25] D. Malioutov, M. Cetin, and A. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[26] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[27] J. Chen and X. Huo, “Theoretical results on sparse representations of multiple-measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[28] M. Fornasier and H. Rauhut, “Recovery algorithms for vector-valued data with joint sparsity constraints,” *SIAM Journal on Numerical Analysis*, vol. 46, no. 2, pp. 577–613, 2008.

[29] M. Mishali and Y. C. Eldar, “Reduce and boost: Recovering arbitrary sets of jointly sparse vectors,” *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4692–4702, 2008.

[30] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, “Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 655–687, 2008.

[31] M. Kowalski, “Sparse regression using mixed norms,” *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303–324, 2009.

[32] S. Ji, D. Dunson, and L. Carin, “Multitask compressive sensing,” *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[33] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[34] Y. Eldar and H. Rauhut, “Average case analysis of multichannel sparse recovery using convex relaxation,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[35] M. Hyder and K. Mahata, “Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation,” *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4646–4655, 2010.

[36] E. Van Den Berg and M. Friedlander, “Theoretical and empirical results for recovery from multiple measurements,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2516–2527, 2010.

[37] J. M. Kim, O. K. Lee, and J. C. Ye, “Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing,” *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 278–301, 2012.

[38] K. Lee, Y. Bresler, and M. Junge, “Subspace methods for joint sparse recovery,” *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3613–3641, 2012.

[39] M. E. Davies and Y. C. Eldar, “Rank awareness in joint sparse recovery,” *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1135–1146, 2012.

[40] Z. Yang, L. Xie, and C. Zhang, “A discretization-free sparse and parametric approach for linear array signal processing,” *IEEE Transactions on Signal Processing, accepted with madatory minor revisions (AQ), available at http://arxiv.org/abs/1312.7695*, 2014.

[41] P. Stoica, P. Babu, and J. Li, “SPICE: A sparse covariance-based estimation method for array processing,” *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, 2011.

[42] U. Grenander and G. Szegö, *Toeplitz forms and their applications.* Univ of California Press, 1958.

[43] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[44] B. Recht, “A simpler approach to matrix completion,” *The Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[45] J. B. Kruskal, “Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics,” *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[46] B. Alexeev, J. Cahill, and D. G. Mixon, “Full spark frames,” *Journal of Fourier Analysis and Applications*, vol. 18, no. 6, pp. 1167–1194, 2012.

[47] B. Recht, M. Fazel, and P. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2007.

[48] Y. Chi, “Joint sparsity recovery for spectral compressed sensing,” *Available online at http://arxiv.org/abs/1311.2229*, 2013.

[49] K. Mohan and M. Fazel, “Iterative reweighted algorithms for matrix rank minimization,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3441–3473, 2012.

[50] M. Malek-Mohammadi, M. Babaie-Zadeh, A. Amini, and C. Jutten, “Recovery of low-rank matrices under affine constraints via a smoothed rank function,” *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 981–992, 2014.

[51] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, “SDPT3–a MATLAB software package for semidefinite programming, version 1.3,” *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 545–581, 1999.

[52] S. P. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[53] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *Available at http://arxiv.org/abs/1011.3027*, 2010.

[54] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *Available online at http://arxiv.org/abs/1312.7793*, 2013.

[55] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[56] V. Paulsen, *Completely bounded maps and operator algebras*. Cambridge University Press, 2002, vol. 78.

[57] A. Schaeffer, "Inequalities of A. Markoff and S. Bernstein for polynomials and related functions," *Bulletin of the American Mathematical Society*, vol. 47, no. 8, pp. 565–579, 1941.