On the Sample Complexity of Multichannel Frequency Estimation via Convex Optimization

Zai Yang, *Member, IEEE*, Jinhui Tang, *Senior Member, IEEE*,
Yonina C. Eldar, *Fellow, IEEE*, and Lihua Xie, *Fellow, IEEE*

Abstract—The use of multichannel data in line spectral estimation (or frequency estimation) is common for improving the estimation accuracy in array processing, structural health monitoring, wireless communications, and more. Recently proposed atomic norm methods have attracted considerable attention due to their provable superiority in accuracy, flexibility, and robustness compared with conventional approaches. In this paper, we analyze atomic norm minimization for multichannel frequency estimation from noiseless compressive data, showing that the sample size per channel that ensures exact estimation decreases with the increase of the number of channels under mild conditions. In particular, given L channels, order $K (\log K) \left(1 + \frac{1}{L} \log N\right)$ samples per channel, selected randomly from N equispaced samples, suffice to ensure with high probability exact estimation of K frequencies that are normalized and mutually separated by at least $\frac{4}{N}$. Numerical results are provided corroborating our analysis.

Index Terms—Multichannel frequency estimation, atomic norm, convex optimization, sample complexity, average case analysis, compressed sensing.

I. INTRODUCTION

LINE spectral estimation or frequency estimation is a fundamental problem in statistical signal processing [2]. It refers to the process of estimating the frequency and amplitude parameters of several complex sinusoidal waves from samples of their superposition. In this paper, we consider a compressive multichannel setting in which each channel

observes a subset of equispaced samples and the sinusoids among the multiple channels share the same frequency profile.

The full $N \times L$ data matrix Y^o is composed of equispaced samples $\left\{y_{jl}^o\right\}$ that are given by

$$y_{jl}^o = \sum_{k=1}^{K} e^{i2\pi(j-1)f_k} s_{kl}, \quad j = 1, \ldots, N, \ l = 1, \ldots, L, \quad (1)$$

where N is the full sample size per channel, L is the number of channels, $i = \sqrt{-1}$, and s_{kl} denotes the unknown complex amplitude of the kth frequency in the lth channel. We assume that the frequencies $\{f_k\}$ are normalized by the sampling rate and belong to the unit circle $\mathbb{T} = [0, 1]$, where 0 and 1 are identified. Letting $a(f) = \left[1, e^{i2\pi f}, \ldots, e^{i2\pi(N-1)f}\right]^T$ represent a complex sinusoid of frequency f, and $s_k = [s_{k1}, \ldots, s_{kL}]$ the coefficient vector of the kth component, we can express the full data matrix Y^o as

$$Y^o = \sum_{k=1}^{K} a(f_k) s_k. \quad (2)$$

In the absence of noise, the goal of compressive multichannel frequency estimation is to estimate the frequencies $\{f_k\}$ and the amplitudes $\{s_{kl}\}$, given a subset of the rows of Y^o. In the case of $L = 1$ this problem is referred to as continuous/off-the-grid compressed sensing [3], or spectral super-resolution by swapping the roles of frequency and time (or space) [4].

Multichannel frequency estimation appears in many applications such as array processing [2], [5], structural health monitoring [6], wireless communications [7], radar [8]–[10], and fluorescence microscopy [11]. In array processing, for example, one needs to estimate the directions of several electromagnetic sources from outputs of several antennas that form an antenna array. In the common uniform-linear-array case, the rows of Y^o correspond to (spatial) locations of the antennas, each column of Y^o corresponds to one (temporal) snapshot, and each frequency has a one-to-one mapping to the direction of one source. Multichannel data are available if multiple data snapshots are acquired. The compressive data case arises when the antennas form a sparse linear array [12]. In structural health monitoring, a number of sensors are deployed to collect vibration data of a physical structure (e.g., a bridge or building) that are then used to estimate the structure's modal parameters such as natural frequencies

and mode shapes. In this case, the rows of Y^o correspond to the (temporal) sampling instants and each column to one (spatially) individual sensor. Multichannel data can be acquired by deploying multiple sensors. The use of compressive data has potential to relax the requirements on data acquisition, storage and transmission.

Due to its connection to array processing, multichannel frequency estimation has a long history dating back to at least World War II when the Bartlett beamformer was proposed to estimate the directions of enemy planes. The use of multichannel data can improve estimation performance by taking more samples and exploiting redundancy among the data channels. As the number of channels increases, the Fisher information about the frequency parameters increases if the data are independent among channels [13], and the number of identifiable frequencies increases in general as well [14]. However, in the worst case where all vector coefficients $\{s_k\}$ are identical up to scaling factors, referred to as coherent sources in array processing, the multichannel data will be identical up to scaling factors. Therefore, the frequency estimation performance cannot be improved by increasing the number of channels. The possible presence of coherent sources is known to form a difficult scenario for multichannel frequency estimation.

In multichannel frequency estimation, the sample data is a highly nonlinear function of the frequencies. Consequently, (stochastic and deterministic) maximum likelihood estimators [5] are difficult to compute. Subspace-based methods were proposed to circumvent nonconvex optimization and dominated research on this topic for several decades. However, these techniques have drawbacks such as sensitivity to coherent sources and limited applicability in the case of missing data or when the number of channels is small [2], [5]. Convex optimization approaches have recently been considered and are very attractive since they can be globally solved in polynomial time. In addition, they are applicable to various kinds of noise and data structures, are robust to coherent sources and have provable performance guarantees [3], [4], [15]–[21]. However, the performance of such techniques under multichannel data has not been thoroughly derived. Extensive numerical results presented in [18], [19], and [22] show that the frequency estimation performance improves (in terms of accuracy and resolution) with the number of channels in the case of noncoherent sources; nevertheless, few theoretical results have been developed to support this empirical behavior.

In this paper, we provide a rigorous analysis of one of the empirical findings: The sample size per channel that ensures exact frequency estimation from noiseless data can be reduced as the number of channels increases. Our main result is described in the following theorem.

Theorem 1: Suppose we observe the $N \times L$ data matrix $Y^o = \sum_{k=1}^{K} a(f_k) s_k$ on rows indexed by $\Omega \subset \{1, \ldots, N\}$, where Ω is of size M, selected uniformly at random, and given. Assume that the vector phases $\phi_k = \frac{s_k}{\|s_k\|_2}$ are independent random variables that are drawn uniformly from the unit L-dimensional complex sphere, and that the frequency support

$\mathcal{T} = \{f_k\}$ *satisfies the minimum separation condition*

$$\Delta_{\mathcal{T}} := \min_{p \neq q} |f_p - f_q| > \frac{1}{\lfloor (N-1)/4 \rfloor}, \quad (3)$$

where the distance is wrapped around on the unit circle. Then, there exists a numerical constant C, which is fixed and independent of the parameters K, N, L and δ, such that

$$M \geq C \max \left\{ \log^2 \frac{N}{\delta}, K \left(\log \frac{K}{\delta} \right) \left(1 + \frac{1}{L} \log \frac{N}{\delta} \right) \right\} \quad (4)$$

is sufficient to guarantee that, with probability at least $1 - \delta$, the full data Y^o and its frequency support $\mathcal{T}$ can be uniquely produced by solving a convex optimization problem (defined in (10) or (12) below).

Theorem 1 provides a non-asymptotic analysis for multichannel frequency estimation that holds for arbitrary number of channels L. According to (4), the sample size per channel that ensures exact frequency estimation decreases as L increases. Consider the case of interest in which the lower bound is dominated by the second term in it, which is true once the number of frequencies is modestly large such that $K \log \frac{K}{\delta} \geq \min \left\{ \log^2 \frac{N}{\delta}, L \log \frac{N}{\delta} \right\}$. The sample size per channel then becomes

$$M \geq C K \left(\log \frac{K}{\delta} \right) \left(1 + \frac{1}{L} \log \frac{N}{\delta} \right), \quad (5)$$

where the lower bound is a monotonically decreasing function of L. In fact, to solve for $\{f_k\}$ and $\{s_{kl}\}$ which consist of $(2L + 1)K$ real unknowns, the minimum number of complex-valued samples is $\frac{1}{2}(2L+1)K$. Consequently, the sample size per channel for any method must satisfy

$$M \geq \frac{1}{2L}(2L + 1)K = K \left(1 + \frac{1}{2L} \right). \quad (6)$$

Our sample order in (5) nearly reaches this information-theoretic rate up to logarithmic factors of K and N. In addition, they share an identical decreasing rate with respect to L.

The assumption on the vector phases $\phi_k = \frac{s_k}{\|s_k\|_2}$ is made in Theorem 1 to avoid coherent sources. It is inspired by [23] and [24] which show performance gains of multiple channels in the context of sparse recovery. This assumption is satisfied by the rotational invariance of Gaussian random vectors if $\{s_{kl}\}$ are independent complex Gaussian random variables with

$$s_{kl} \sim \mathcal{CN}(0, p_k),$$

where the variance p_k refers to the power of the kth source. This latter assumption is usually adopted in the literature to derive the Cramer-Rao bound or to analyze the theoretical performance of subspace-based methods [5].

Note that the $4/N$ minimum separation condition (3) comes from the single-channel analysis in [3] and [4]. It is reasonably tight in the sense that, if the separation is under $2/N$, then the data corresponding to two different sets of frequencies can be practically the same and it becomes nearly impossible to distinguish them [22]. This condition is conservative in practice, and a numerical approach was implemented in [25] to further improve resolution. Such a separation condition is not required for subspace-based methods, at least in theory.

A corollary of Theorem 1 can be obtained when the frequencies lie on a known grid. This scenario may be unrealistic but is often assumed in practice to simplify the algorithm design.

Corollary 1: Consider the problem setup of Theorem 1 when in addition the frequencies $\{f_k\}$ lie on a given fixed grid (that can be nonuniform or arbitrarily fine). Under the assumptions of Theorem 1 and given the same number of samples, the frequencies can be exactly estimated with at least the same probability by solving an $\ell_{2,1}$ norm minimization problem (defined in (15) below).

It is worth noting that Corollary 1 holds regardless of the density of the grid, while the grid is used to define the $\ell_{2,1}$ norm minimization problem. In fact, Theorem 1 corresponds to the extreme case of Corollary 1 in which the grid becomes infinitely fine and contains infinitely many points. In this case, the $\ell_{2,1}$ norm minimization problem in Corollary 1 becomes computationally intractable. In contrast, the convex optimization problem in Theorem 1 is still applicable.

The rest of the paper is organized as follows. Section II revisits prior art in multichannel frequency estimation and related topics and discusses their connections to this work. Section III presents the convex optimization methods mentioned in Theorem 1 and Corollary 1. The former method is referred to as atomic norm minimization originally introduced in [18] and [19]. The latter is known as $\ell_{2,1}$ norm minimization [24], [26]. Section IV presents simulation results to verify our main theorem. Section V provides the detailed proof of Theorem 1. Finally, Section VI concludes this paper.

Throughout the paper, $\log a\, b$ means $(\log a)\, b$ unless we write $\log(ab)$. The notations $\mathbb{R}$ and $\mathbb{C}$ denote the set of real and complex numbers, respectively. The normalized frequency interval is identified with the unit circle $\mathbb{T} = [0, 1]$. Boldface letters are reserved for vectors and matrices. The ℓ_1, ℓ_2 and Frobenius norms are written as $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_F$ respectively, and $|\cdot|$ is the magnitude of a scalar or the cardinality of a set. For matrix A, A^T is the transpose, A^H is the conjugate transpose, $\text{rank}(A)$ denotes the rank, and $\text{tr}(A)$ is the trace. The fact that a square matrix A is positive semidefinite is expressed as $A \geq 0$. Unless otherwise stated, x_j is the jth entry of a vector x, A_{jl} is the (j, l)th entry of a matrix A, and A_j denotes the jth row. Given an index set Ω, x_Ω and A_Ω are the subvector and submatrix of x and A, respectively, that are formed by the rows of x and A indexed by Ω. For a complex argument, $\Re$ and $\Im$ return its real and imaginary parts. The big O notation is written as $O(\cdot)$. The expectation of a random variable is denoted by $\mathbb{E}[\cdot]$, and $\mathbb{P}(\cdot)$ is the probability of an event.

II. PRIOR ART

The advantages of using multichannel data have been well understood for parametric approaches for frequency estimation, e.g., maximum likelihood and subspace-based methods. Under proper assumptions on $\{s_{kl}\}$, which are similar to those in Theorem 1, these techniques are asymptotically efficient in the number of channels L and therefore their accuracy improves with the increase of L [27]. However, few theoretical results on their accuracy are available with fixed L, in contrast to Theorem 1 which holds for any L.

Note that theoretical guarantees of two common subspace-based methods, MUSIC [28] and ESPRIT [29], have recently been developed in [30] and [31] in the single-channel full data case. For multichannel sparse recovery in discrete models, MUSIC has been applied or incorporated into other methods, with theoretical guarantees, see [32]–[35].

Theorem 1 is related to the so-called average-case analysis for multichannel sparse recovery in compressed sensing problems [23], [24], [36]. In such settings, the goal is to recover multiple sparse signals sharing the same support from their linear measurements. Average case analysis (as opposed to the worst case) attempts to analyze an algorithm's average performance for generic signals. Under an assumption on the sparse signals similar to that on $\{s_{kl}\}$ in Theorem 1, the papers [23] and [24] showed that, if the columns of the sensing matrix are weakly correlated (in terms of mutual coherence [37] or the restricted isometry property (RIP) [38]), then the probability of successful sparse recovery improves as the number of channels increases. While [23] used the greedy method of orthogonal matching pursuit (OMP) [39], [40], the work in [24] was focused on $\ell_{2,1}$ norm minimization. The proof of Theorem 1 is partially inspired by these results. However, the frequency parameters are continuously valued so that the weak correlation assumption is not satisfied. In addition to this, Theorem 1 provides an explicit expression of the sample size per channel as a decreasing function of the number of channels.

Multichannel sparse recovery methods have been applied to multichannel frequency estimation using, e.g., $\ell_{2,1}$ norm minimization as in Corollary 1 [26]. These approaches utilize the fact that the number of frequencies is small, and attempt to find the smallest set of frequencies on a grid describing the observed data. It was empirically shown in a large number of publications (see, e.g., [26], [41], [42]) that, compared with conventional nonparametric and parametric methods, such sparse techniques have improved resolution and relaxed requirements on the number of channels and the type of sources and noise. However, these approaches can be applied only if the continuous frequency domain is discretized so that the frequencies are restricted to a finite set of grid points. Due to discretization errors and high correlations among the candidate frequency components, few theoretical guarantees were developed to support the good empirical results. Corollary 1 provides theoretical support under the assumption of no discretization errors.

Assume in Corollary 1 a uniform grid of size equal to N (without taking into account discretization errors). The case of $L = 1$ then degenerates to the compressed sensing problem studied in [43], where the weak correlation assumption can be satisfied with sample complexity $O\left(K \log^2 K \log N\right)$ [44]. This means that given approximately the same number of samples, the resolution can be increased from $4/N$ to $1/N$ if a stricter assumption is made on the frequencies. In this case, the benefits of using multichannel data have been shown in [23] and [24], as discussed previously.

The gap between the discrete and continuous frequency setups has been closed in the past few years. In the pioneering work of Candès and Fernandez-Granda [4], the authors

studied the single-channel full data setting and proposed a convex optimization method (to be specific, semidefinite program (SDP)) in the continuous domain, referred to as total variation minimization, that is a continuous analog of ℓ_1 minimization in the discrete setup. They proved that the true frequencies can be recovered if they are mutually separated by at least $4/N$—the minimum separation condition (3). The idea and method were then extended by Tang *et al.* [3] to compressive data using atomic norm minimization [45] which is equivalent to the total variation norm. They showed that exact frequency localization occurs with high probability under the $4/N$ separation condition if $M \geq O(K \log K \log N)$ samples are given. This result is a special case of Theorem 1 in the single-channel case.

Multichannel frequency estimation using atomic norm was studied by Yang and Xie [18], Li and Chi [19], and Li *et al.* [21]. The paper [18] showed that exact frequency estimation occurs under the same $4/N$ separation condition with full data. In the compressive setting, a sample complexity of $O(K \log K \log(\sqrt{L}N))$ per channel was derived under the assumption of centered independent random phases $\{\boldsymbol{\phi}_k\}$. Compared with Theorem 1, the assumption on $\{\boldsymbol{\phi}_k\}$ is weaker but the sample complexity is larger. Unlike Theorem 1, the sources are allowed to be coherent in [18] and therefore the above results do not shed any light on the advantage of multichannel data. These results may be considered as a worst-case analysis, if we refer to Theorem 1 as an average-case analysis.

A slightly different compressive setting was considered in [19], in which the samples are taken uniformly at random from the entries (as opposed to the rows) of $\boldsymbol{Y}^o$, and an average per-channel sample complexity of $O(K \log(LK) \log(LN))$ was stated under an assumption on $\{s_{kl}\}$ similar to that in Theorem 1. This result is weaker than that of Theorem 1 and again fails to show the advantage of multichannel data.[1]

A different means for data compression was considered in [21], in which each column of $\boldsymbol{Y}^o$ is compressed by using a distinct sensing matrix generated from a Gaussian distribution. In this case $M \geq \max \left\{ \frac{8}{L} \log \frac{1}{\delta} + 2, \ CK \log N \right\}$ samples per channel suffice to guarantee exact frequency estimation with probability at least $1 - \delta$ by using atomic norm minimization, where C is a constant. The sample size M decreases with the number of channels if the lower bound is dominated by the first term or equivalently if $L(CK \log N - 2)$ is less than a constant depending on δ. Evidently, such a benefit from multichannel data diminishes if any one of L, K or N becomes large. This benefit is a result of the different sampling matrices applied on each channel, which we do not require in our analysis.

Another convex optimization method operating on the continuum for multichannel frequency estimation is gridless SPICE (GLS) [42], [46], [47]. It was shown to be equivalent to an atomic norm method for small L and a weighted atomic norm method for large L [25], [48]. Under an assumption on $\{s_{kl}\}$ similar to that in Theorem 1, GLS is asymptotically

efficient in L. However, such benefits from multichannel data have not been analyzed for small L.

Finally, sparse estimation methods using nonconvex optimization have been proposed in [49] and [50], but few theoretical guarantees on their estimation accuracy have been derived. Readers are referred to [48] for a review on sparse methods for multichannel frequency estimation.

III. MULTICHANNEL ATOMIC NORM MINIMIZATION

Atomic norm [45] provides a generic approach to finding a sparse representation of a signal by exploiting particular structures in it. It generalizes the ℓ_1 norm for sparse signal recovery and the nuclear norm for low rank matrix recovery. For multichannel frequency estimation, the set of atoms is defined as [18], [19]:

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

where

$$\mathbb{S}^{2L-1} := \left\{ \boldsymbol{\phi} \in \mathbb{C}^{1 \times L} : \ \|\boldsymbol{\phi}\|_2 = 1 \right\}$$

denotes the unit L-dimensional complex sphere, or equivalently, the unit $2L$-dimensional real sphere. The atomic norm of a multichannel signal $\boldsymbol{Y} \in \mathbb{C}^{N \times L}$ is defined as the gauge function of the convex hull of $\mathcal{A}$:

$$\|\boldsymbol{Y}\|_{\mathcal{A}} = \inf \left\{ t > 0 : \ \boldsymbol{Y} \in t \,\mathrm{conv}\,(\mathcal{A}) \right\}$$
$$= \inf \left\{ \sum_k c_k : \ \boldsymbol{Y} = \sum_k c_k \boldsymbol{a}(f_k)\boldsymbol{\phi}_k, \ c_k > 0 \right\}$$
$$= \inf \left\{ \sum_k \|\boldsymbol{s}_k\|_2 : \ \boldsymbol{Y} = \sum_k \boldsymbol{a}(f_k)\boldsymbol{s}_k \right\}. \tag{7}$$

When only the rows of $\boldsymbol{Y}^o$ indexed by $\Omega \subset \{1, \ldots, N\}$ are observed, which form the submatrix $\boldsymbol{Y}^o_\Omega$, the following atomic norm minimization problem was introduced to recover the full data matrix $\boldsymbol{Y}^o$ and its frequencies [18], [19]:

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_{\mathcal{A}},$$
$$\text{subject to } \boldsymbol{Y}_\Omega = \boldsymbol{Y}^o_\Omega. \tag{8}$$

In particular, by (8) we attempt to find the signal that is consistent with the observed data and has the smallest atomic norm. The problem in (8) can then be cast as an SDP in order to solve it, in two ways. The first was proposed in [18] and [19] (and later reproduced in [51]) by writing $\|\boldsymbol{Y}\|_{\mathcal{A}}$ as:

$$\|\boldsymbol{Y}\|_{\mathcal{A}} = \min_{\boldsymbol{X}, t} \frac{1}{2}\mathrm{tr}\,(\boldsymbol{X}) + \frac{1}{2}t_0,$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & \boldsymbol{T} \end{bmatrix} \geq \boldsymbol{0}. \tag{9}$$

It follows that (8) can be equivalently written as:

$$\min_{\boldsymbol{X}, t, \boldsymbol{Y}} \frac{1}{2}\mathrm{tr}\,(\boldsymbol{X}) + \frac{1}{2}t_0,$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & \boldsymbol{T} \end{bmatrix} \geq \boldsymbol{0} \ \text{ and } \ \boldsymbol{Y}_\Omega = \boldsymbol{Y}^o_\Omega. \tag{10}$$

In (9) and (10), $\boldsymbol{T}$ is an $N \times N$ Hermitian Toeplitz matrix and is defined as $T_{ml} = t_{l-m}$, $1 \leq m \leq l \leq N$, and $\boldsymbol{t} = [t_j]_{j=0}^{N-1}$.

---

[1]Note that an explicit proof of this result was not included in [19], which was also pointed out in [21, Footnote 8].

Alternatively, an SDP can also be provided for the following dual of (8):

$$\max_{V} \langle V_\Omega, Y^o_\Omega \rangle_{\mathbb{R}},$$
$$\text{subject to } \|V\|^*_{\mathcal{A}} \leq 1 \text{ and } V_{\Omega^c} = 0, \quad (11)$$

where $\langle A, B \rangle_{\mathbb{R}} = \Re\text{tr}\left(B^H A\right)$ is the inner product of matrices A and B. The dual atomic norm $\|V\|^*_{\mathcal{A}}$ is defined as:

$$\|V\|^*_{\mathcal{A}} = \sup_{a \in \text{conv}(\mathcal{A})} \langle V, a \rangle_{\mathbb{R}}$$
$$= \sup_{a \in \mathcal{A}} \langle V, a \rangle_{\mathbb{R}}$$
$$= \sup_{f} \left\| a^H(f) V \right\|_2.$$

It follows that the constraint $\|V\|^*_{\mathcal{A}} \leq 1$ is equivalent to

$$\left\| a^H(f) V \right\|_2 \leq 1 \quad \text{for all} f \in \mathbb{T}.$$

Using theory of positive trigonometric polynomials [52], the above constraint can be cast as a linear matrix inequality (LMI) so that the dual SDP is given by:

$$\max_{V} \langle V_\Omega, Y^o_\Omega \rangle_{\mathbb{R}},$$
$$\text{subject to} \begin{cases} \begin{bmatrix} I & V^H \\ V & H \end{bmatrix} \geq 0, \\ \text{tr}(H) = 1, \\ \sum_{n=1}^{N-j} H_{n,n+j} = 0, \ j = 1, \ldots, N-1, \\ V_{\Omega^c} = 0. \end{cases} \quad (12)$$

Both (10) and (12) can be solved using off-the-shelf SDP solvers. In fact, when one is solved using a primal-dual algorithm, the solution to the other is given simultaneously. While the solved signal Y is given by the primal solution, the frequencies in Y can be retrieved from either the primal or the dual solution. In particular, given the primal solution T, the frequencies can be obtained from its so-called Vandermonde decomposition [2] given as:

$$T = \sum_{j=1}^{r} c_j a(f_j) a^H(f_j), \quad (13)$$

where $r = \text{rank}(T)$ and $c_j > 0$. This decomposition is unique in the typical case of rank $(T) < N$ and can be computed using subspace-based methods such as ESPRIT [29]. The atomic decomposition of Y can then be obtained by solving for the coefficients $\{s_j\}_{j=1}^{r}$ from (2) using a least squares method. Interestingly, it holds that $c_j = \|s_j\|_2$. This means that, besides the frequencies, their magnitudes are also given by the Vandermonde decomposition of T.

Given the solution V in the dual, $Q(f) = a^H(f)V$ is referred to as the (vector) dual polynomial. The frequencies in Y can be identified by those values f_j satisfying

$$\left\| Q(f_j) \right\|_2 = \left\| a^H(f_j)V \right\|_2 = 1.$$

Subsequently, the coefficients in the atomic decomposition can be solved for in the same manner.

The dimensionality of both the primal SDP in (10) and the dual SDP in (12) increases as the number of channels L increases. To reduce the computational workload when L is large, a dimensionality reduction technique was introduced

in [25] that, in the case of $L > \text{rank}\left(Y^o_\Omega\right)$, reduces the number of channels from L to rank $\left(Y^o_\Omega\right)$ and at the same time produces the same T, from which both the frequencies and their magnitudes are obtained. In particular, for any $\widetilde{Y^o_\Omega}$ satisfying $\widetilde{Y^o_\Omega}\widetilde{Y^o_\Omega}^H = Y^o_\Omega Y^{oH}_\Omega$, whose number of columns may be as small as rank $\left(Y^o_\Omega\right)$, the solution to T remains unchanged if, in (10), we replace Y^o_Ω with $\widehat{Y^o_\Omega}$ and properly change the dimensions of X and Y. This means that the same T can be obtained by solving the following SDP:

$$\min_{X,t,Y} \frac{1}{2}\text{tr}(X) + \frac{1}{2}t_0,$$
$$\text{subject to} \begin{bmatrix} X & Y^H \\ Y & T \end{bmatrix} \geq 0 \text{ and } Y_\Omega = \widetilde{Y^o_\Omega}. \quad (14)$$

Similar techniques were also reported in [51] and [53].

Suppose that the source powers are approximately constant across different channels. Then the magnitude of Y^o_Ω increases and may even become unbounded as L increases. To render (14) solvable for large L, we consider a minor modification by shrinking $\widetilde{Y^o_\Omega}$ by $\sqrt{L}$ so that $\widetilde{Y^o_\Omega}\widetilde{Y^o_\Omega}^H = \frac{1}{L}Y^o_\Omega Y^{oH}_\Omega$ (note that $\frac{1}{L}Y^o_\Omega Y^{oH}_\Omega$ is the sample covariance matrix). Consequently, the solution to T shrinks by $\sqrt{L}$, resulting in the same frequencies and scaled magnitudes. As L approaches infinity, the sample covariance matrix approaches the data covariance matrix by the law of large numbers and therefore, given the data covariance matrix R, this modification enables us to deal with the case of $L \to \infty$ by choosing $\widetilde{Y^o_\Omega}$ satisfying $\widetilde{Y^o_\Omega}\widetilde{Y^o_\Omega}^H = R$. This result will be used in Section IV to study the numerical performance of atomic norm minimization as $L \to \infty$.

Finally, if we assume that the frequencies lie on a given fixed grid, denoted by $\left\{\tilde{f}_g\right\}_{g=1}^{G} \subset \mathbb{T}$, where G is the grid size, then the atomic norm minimization in (8) and (10) can be simplified to the following $\ell_{2,1}$ norm minimization problem [26]:

$$\min_{\{\tilde{s}_g\}} \sum_{g=1}^{G} \|\tilde{s}_g\|_2,$$
$$\text{subject to } \sum_{g=1}^{G} a_\Omega\left(\tilde{f}_g\right)\tilde{s}_g = Y^o_\Omega. \quad (15)$$

In (15), $a_\Omega(\cdot)$ is a subvector of $a(\cdot)$, and the frequency support $\mathcal{T}$ can be identified from the solutions to $\{\tilde{s}_g\}$ that are nonzero. This $\ell_{2,1}$ norm minimization problem is used in Corollary 1 for frequency estimation.

IV. NUMERICAL SIMULATIONS

We now present numerical simulations demonstrating the decrease and its rate of the required sample size per channel with the increase of the number of channels. Numerical results provided in previous publications [18], [19], [22], [25] show the advantages of taking more channel signals in either improving the probability of successful frequency estimation or enhancing the resolution.

In our simulations, we let $N = 128$, $K = 10$, $L \in \{1, 2, 4, 8, 16, \infty\}$, and $M \in \{10, 12, \ldots, 50\}$. Note that the case of $L = \infty$ can be dealt with by using the dimensionality reduction technique presented in Section III. For each
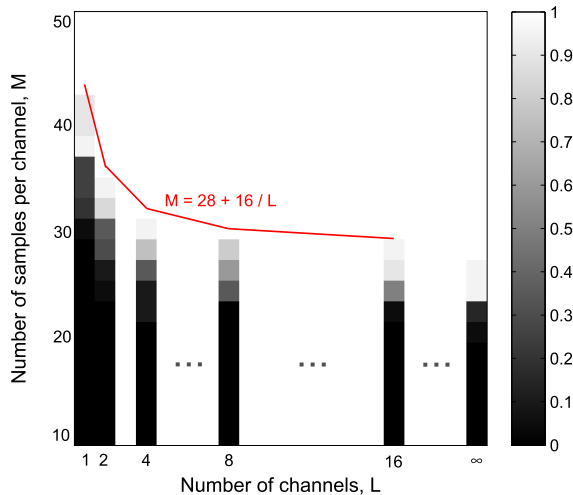
Fig. 1. Success rates of atomic norm minimization for multichannel frequency estimation, with $N = 128$, $K = 10$, $L = 1, 2, 4, 8, 16, \infty$ and $M = 10, 12, \ldots, 50$, under the minimum separation condition in (3). White means complete success and black means complete failure. The red curve plots $M = 28 + 16/L$. The decreasing behavior and rate of the sample size with respect to the number of channels match those predicted in Theorem 1.

pair (L, M), 20 Monte Carlo runs are carried out. In each run, the frequencies $\{f_k\}$ are randomly generated and satisfy the minimum separation condition (3), and the amplitudes $\{s_{kl}\}$ are independently generated from a standard complex Gaussian distribution. After the full data matrix $\boldsymbol{Y}^o$ is computed according to (1), M rows of $\boldsymbol{Y}^o$ are randomly selected and fed into atomic norm minimization for frequency estimation implemented using CVX [54]. The frequencies are said to be successfully estimated if the root mean squared error, computed as

$$\sqrt{\frac{\sum_{k=1}^{K} \left| f_k - \hat{f}_k \right|^2}{K}},$$

is less than 10^{-4}, where $\hat{f}_k$ denotes the estimate of f_k. We also calculate the rate of successful estimation over the 20 runs. Our simulation results are presented in Fig. 1. It can be seen that the required sample size per channel for exact frequency estimation decreases with the number of channels. Evidently, the decrease rate matches well with that predicted in Theorem 1 even in such a low-dimensional problem setup.

V. PROOF OF THEOREM 1

We now prove Theorem 1. The proof of Corollary 1 is similar and is given in Appendix F. Our proof is inspired by the analysis for multichannel frequency estimation in [18], the average case analysis for multichannel sparse recovery in [24], and several earlier publications on compressed sensing and super-resolution [3], [4], [23], [43]. We mainly follow the steps of [18], with non-trivial modifications. To make the proof concise and self-contained, we first revisit the result in [18] and then highlight the key differences in our problem.

A. Revisiting Previous Analysis

The following theorem is proved in [18].

Theorem 2: Suppose we observe the $N \times L$ data matrix $\boldsymbol{Y}^o = \sum_{k=1}^{K} c_k \boldsymbol{a}(f_k) \boldsymbol{\phi}_k$ on rows indexed by $\Omega \subset \{1, \ldots, N\}$,

where Ω is of size M, selected uniformly at random, and given. Assume that the vector phases $\{\boldsymbol{\phi}_k\}$ are independent and centered random variables and that the frequencies $\{f_k\}$ satisfy the same separation condition as in Theorem 1. Then, there exists a numerical constant C such that

$$M \geq C \max \left\{ \log^2 \frac{\sqrt{L}N}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{L}N}{\delta} \right\} \quad (16)$$

is sufficient to guarantee that, with probability at least $1 - \delta$, the full data $\boldsymbol{Y}^o$ and its frequency support $\mathcal{T}$ can be uniquely produced by solving (8) [or equivalently (10)].

According to [18], the optimality of a solution to (8) can be validated using a dual certificate that is provided in the following proposition.

Proposition 1: The matrix $\boldsymbol{Y}^o = \sum_{k=1}^{K} c_k \boldsymbol{a}(f_k) \boldsymbol{\phi}_k$ is the unique optimizer of (8) if $\{\boldsymbol{a}_\Omega(f_k)\}_{f_k \in \mathcal{T}}$ are linearly independent and if there exists a vector-valued dual polynomial $\overline{Q} : \mathbb{T} \to \mathbb{C}^{1 \times L}$

$$\overline{Q}(f) = \boldsymbol{a}^H(f) \boldsymbol{V} \quad (17)$$

satisfying

$$\overline{Q}(f_k) = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \quad (18)$$
$$\left\| \overline{Q}(f) \right\|_2 < 1, \quad f \in \mathbb{T} \backslash \mathcal{T}, \quad (19)$$
$$\boldsymbol{V}_j = \boldsymbol{0}, \quad j \in \Omega^c, \quad (20)$$

where $\boldsymbol{V}$ is an $N \times L$ matrix and $\boldsymbol{V}_j$ denotes its jth row. Moreover, $\boldsymbol{Y}^o = \sum_{k=1}^{K} c_k \boldsymbol{a}(f_k) \boldsymbol{\phi}_k$ is the unique atomic decomposition achieving the atomic norm with $\|\boldsymbol{Y}^o\|_\mathcal{A} = \sum_{k=1}^{K} c_k$.

Applying Proposition 1, Theorem 2 is proved in [18] by showing the existence of an appropriate dual polynomial $\overline{Q}(f)$ under the assumptions of Theorem 2 (note that in this process the condition of linear independence of $\{\boldsymbol{a}_\Omega(f_k)\}_{f_k \in \mathcal{T}}$ is also satisfied). To explicitly construct the dual polynomial $\overline{Q}(f)$, it is shown in [18] that we may consider the symmetric case where the rows of $\boldsymbol{Y}^o$ are indexed by $\mathcal{J} := \{-2n, \ldots, 2n\}$ instead of $\{1, \ldots, N\}$, where $N = 4n + n_0$ with $n_0 = 1, 2, 3, 4$. Moreover, we may equivalently consider the Bernoulli observation model following from [3] and [43], in which each row of $\boldsymbol{Y}^o$ is observed independently with probability $p = \frac{M}{4n}$, rather than the uniform observation model as in Theorems 1 and 2. The remainder of the proof consists of two steps.

1) Consider the full data case where Ω^c is empty and construct $\overline{Q}(f)$, referred to as $Q(f)$, satisfying (18) and

$$\|Q(f)\|_2^2 \leq 1 - c_1 n^2 (f - f_k)^2,$$
$$f \in \mathcal{N}_k := \left(f_k - \frac{0.16}{n}, \ f_k + \frac{0.16}{n} \right), \quad (21)$$

$$\|Q(f)\|_2 \leq 1 - c_2, \quad f \in \mathcal{F} := \mathbb{T} \backslash \bigcup_{k=1}^{K} \mathcal{N}_k, \quad (22)$$

$$\frac{d^2 \|Q(f)\|_2^2}{df^2} \leq -2c_1 n^2, \quad f \in \mathcal{N}_k, \quad (23)$$

where c_1 and c_2 are positive constants. Note that (21) and (22) form a stronger condition than (19).

2) In the compressive data case of interest, construct a random polynomial $\overline{Q}(f)$ satisfying (18) and (20) with respect to the Bernoulli sampling scheme. Show that, if M satisfies (16), then $\overline{Q}(f)$ (and its first and second derivatives) is close to $Q(f)$ (and its first and second derivatives) on the whole unit circle $\mathbb{T}$ with high probability under the assumptions of Theorem 2, and $\overline{Q}(f)$ also satisfies (21) and (22). As a result, $\overline{Q}(f)$ is a polynomial as required in Proposition 1.

In the ensuing subsections, we prove Theorem 1 following the aforementioned steps of the proof of Theorem 2. The difference is in the second step above, namely, showing that $\overline{Q}(f) - Q(f)$ and its derivatives are arbitrarily small on the unit circle $\mathbb{T}$ with high probability under the assumptions of Theorem 1, with M in (4) being a decreasing function of the number of channels L. In particular, we show in Subsection V-B how the dual certificate $\overline{Q}(f)$ is constructed, summarize in Subsection V-C some useful lemmas shown in [3], [4], and [18], analyze in Subsection V-D the residue $\overline{Q}(f) - Q(f)$, and complete the proof in Subsection V-E.

B. Construction of $\overline{Q}(f)$

To construct $\overline{Q}(f)$, we start with the squared Fejér kernel

$$\mathcal{K}(f) = \left[\frac{\sin(\pi(n+1)f)}{(n+1)\sin(\pi f)} \right]^4 = \sum_{j=-2n}^{2n} g_n(j) e^{-i2\pi jf} \quad (24)$$

with coefficients

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

$$\times \left(1 - \frac{|j-k|}{n+1}\right)$$

obeying $0 < g_n(j) \le 1$, $j = -2n, \ldots, 2n$. This kernel equals unity at the origin and decays rapidly away from it. Let $\{\delta_j\}_{j \in \mathcal{J}}$ be i.i.d. Bernoulli random variables with

$$\mathbb{P}(\delta_j = 1) = p = \frac{M}{4n}.$$

It follows that

$$\Omega = \{j \in \mathcal{J} : \delta_j = 1\}$$

with $\mathbb{E}|\Omega| \approx M$. This allows us to write a compressive-data analog of the squared Fejér kernel as

$$\overline{\mathcal{K}}(f) = \sum_{j \in \Omega} g_n(j) e^{-i2\pi jf} = \sum_{j=-2n}^{2n} \delta_j g_n(j) e^{-i2\pi jf} \quad (25)$$

and define the vector-valued polynomial $\overline{Q}(f)$ as

$$\overline{Q}(f) = \sum_{f_k \in \mathcal{T}} \alpha_k \overline{\mathcal{K}}(f - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \overline{\mathcal{K}}^{(1)}(f - f_k), \quad (26)$$

where α_k and β_k are $1 \times L$ vector coefficients to specify and the superscript l denotes the lth derivative.

It is evident that $\overline{Q}(f)$ in (26) satisfies the support condition in (20). According to [3], $\overline{\mathcal{K}}$ and its derivatives are concentrated around their expectations, $p\mathcal{K}$ and its derivatives, if M is large enough. As a result, $\overline{Q}(f)$ is expected to peak

near $f_k \in \mathcal{T}$ if $\overline{Q}(f)$ is dominated by the first term and the coefficients α_k and β_k are appropriately chosen. To make $\overline{Q}(f)$ satisfy (18), we impose for any $f_j \in \mathcal{T}$,

$$\sum_{f_k \in \mathcal{T}} \alpha_k \overline{\mathcal{K}}(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \overline{\mathcal{K}}^{(1)}(f_j - f_k) = \phi_j. \quad (27)$$

To satisfy (19), a necessary condition is that the derivative of $\left\| \overline{Q}(f) \right\|_2^2$ vanishes at $f_j \in \mathcal{T}$, leading to

$$\left. \frac{d \left\| \overline{Q}(f) \right\|_2^2}{df} \right|_{f=f_j} = 2\Re \left\{ \overline{Q}^{(1)}(f_j) \overline{Q}^H(f_j) \right\}$$

$$= 2\Re \left\{ \overline{Q}^{(1)}(f_j) \phi_j^H \right\}$$

$$= 0. \quad (28)$$

Consequently, one feasible choice is to let, for any $f_j \in \mathcal{T}$,

$$\overline{Q}^{(1)}(f_j) = \sum_{f_k \in \mathcal{T}} \alpha_k \overline{\mathcal{K}}^{(1)}(f_j - f_k) + \sum_{f_k \in \mathcal{T}} \beta_k \overline{\mathcal{K}}^{(2)}(f_j - f_k)$$

$$= \mathbf{0}. \quad (29)$$

The conditions in (27) and (29) consist of $2KL$ equations and therefore may determine the coefficients $\{\alpha_k\}$ and $\{\beta_k\}$ (and hence $\overline{Q}(f)$). Define the $K \times K$ matrices $\overline{D}_l$, $l = 0, 1, 2$ such that $[\overline{D}_l]_{jk} = \overline{\mathcal{K}}^{(l)}(f_j - f_k)$ (note that the zeroth derivative is itself). Then, (27) and (29) lead to the following system of linear equations:

$$\overline{D} \begin{bmatrix} \alpha \\ c_0 \beta \end{bmatrix} = \begin{bmatrix} \overline{D}_0 & c_0^{-1} \overline{D}_1 \\ -c_0^{-1} \overline{D}_1 & -c_0^{-2} \overline{D}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ c_0 \beta \end{bmatrix} = \begin{bmatrix} \Phi \\ \mathbf{0} \end{bmatrix}, \quad (30)$$

where Φ is the $K \times L$ matrix formed by stacking $\{\phi_j\}$ together, with α and β similarly defined. The constant $c_0 = \sqrt{|\mathcal{K}^{(2)}(0)|} = \sqrt{\frac{4\pi^2 n(n+2)}{3}}$ is introduced so that the coefficient matrix $\overline{D}$ is symmetric and well-conditioned [3]. Note that all compressive-data (random) quantities defined above such as $\overline{Q}$ and $\overline{D}$ have full-data (deterministic) analogs, denoted by Q and D by removing the overlines and obtained by replacing $\overline{\mathcal{K}}$ in their expressions with $\mathcal{K}$ [refer Q to (21) and (22)]. The remaining task is to show that, if M satisfies (4), then $\overline{D}$ is invertible and $\overline{Q}(f)$ can be uniquely determined under the assumptions of Theorem 1. In addition, show that $\overline{Q}(f)$ satisfies (19), which together with Proposition 1 completes the proof.

C. Useful Lemmas

To complete the proof we rely on some useful results shown in [3], [4], and [18] that are summarized below.

First consider the invertibility of $\overline{D}$. For $\tau \in (0, \frac{1}{4}]$, define the event

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

Let δ, ϵ be small positive numbers and C a constant, which are independent of the parameters K, M, N and L, and may vary from instance to instance. Then we have the following lemma which, in addition to the invertibility of $\overline{D}$, also guarantees linear independence of $\{a_\Omega(f_k)\}_{f_k \in \mathcal{T}}$ as required in Proposition 1.

Lemma 1 [3], [4], [18]: Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$ *and* $n \geq 64$ *and let* $\tau \in \left(0, \frac{1}{4}\right]$. *Then,* $\boldsymbol{D}$ *is invertible,* $\overline{\boldsymbol{D}}$ *is invertible on* $\mathcal{E}_{1,\tau}$, $\{a_{\Omega}(f_k)\}_{f_k \in \mathcal{T}}$ *are linearly independent on* $\mathcal{E}_{1,\tau}$, *and* $\mathbb{P}(\mathcal{E}_{1,\tau}) \geq 1 - \delta$ *if*

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

On $\mathcal{E}_{1,\tau}$, we introduce the partitions $\overline{\boldsymbol{D}}^{-1} = \begin{bmatrix} \overline{\boldsymbol{L}} & \overline{\boldsymbol{R}} \end{bmatrix}$ and $\boldsymbol{D}^{-1} = \begin{bmatrix} \boldsymbol{L} & \boldsymbol{R} \end{bmatrix}$, where $\overline{\boldsymbol{L}}, \overline{\boldsymbol{R}}, \boldsymbol{L}$ and $\boldsymbol{R}$ are all $2K \times K$ matrices. For $l = 0, 1, 2$, let

$$\overline{\boldsymbol{v}}_l(f) = c_0^{-l} \begin{bmatrix} \overline{\mathcal{K}}^{(l)H}(f - f_1) \\ \vdots \\ \overline{\mathcal{K}}^{(l)H}(f - f_K) \\ c_0^{-1} \overline{\mathcal{K}}^{(l+1)H}(f - f_1) \\ \vdots \\ c_0^{-1} \overline{\mathcal{K}}^{(l+1)H}(f - f_K) \end{bmatrix} \quad (31)$$

and similarly define its deterministic analog $\boldsymbol{v}_l(f)$, where $f_j \in \mathcal{T}$ and the superscript H denotes the complex conjugate for a scalar. It follows that on the event $\mathcal{E}_{1,\tau}$,

$$\begin{bmatrix} \boldsymbol{\alpha} \\ c_0 \boldsymbol{\beta} \end{bmatrix} = \overline{\boldsymbol{D}}^{-1} \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix} = \overline{\boldsymbol{L}} \boldsymbol{\Phi}, \quad (32)$$

$$\begin{aligned} c_0^{-l} \overline{Q}^{(l)}(f) &= \sum_{f_k \in \mathcal{T}} \boldsymbol{\alpha}_k c_0^{-l} \overline{\mathcal{K}}^{(l)}(f - f_k) \\ &\quad + \sum_{f_k \in \mathcal{T}} c_0 \boldsymbol{\beta}_k \cdot c_0^{-(l+1)} \overline{\mathcal{K}}^{(l+1)}(f - f_k) \\ &= \overline{\boldsymbol{v}}_l^H(f) \overline{\boldsymbol{L}} \boldsymbol{\Phi} \\ &:= \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) \right\rangle, \quad (33) \end{aligned}$$

where the notation of inner-product is abused since the result of the product is a vector rather than a scalar.

We write $\overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f)$ in (33) in three parts:

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

which results in the following decomposition of $c_0^{-l} \overline{Q}^{(l)}(f)$:

$$\begin{aligned} c_0^{-l} \overline{Q}^{(l)}(f) &= \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H \overline{\boldsymbol{v}}_l(f) \right\rangle \\ &= \left\langle \boldsymbol{\Phi}, \boldsymbol{L}^H \boldsymbol{v}_l(f) \right\rangle \\ &\quad + \left\langle \boldsymbol{\Phi}, \overline{\boldsymbol{L}}^H (\overline{\boldsymbol{v}}_l(f) - p\boldsymbol{v}_l(f)) \right\rangle \\ &\quad + \left\langle \boldsymbol{\Phi}, \left(\overline{\boldsymbol{L}} - p^{-1}\boldsymbol{L}\right)^H p\boldsymbol{v}_l(f) \right\rangle \\ &= c_0^{-l} Q^{(l)}(f) + I_1^l(f) + I_2^l(f), \quad (34) \end{aligned}$$

where we have defined

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

As a result of (34), a connection between $\overline{Q}(f)$ and $Q(f)$ is established. Our goal is to show that the random

perturbations $I_1^l(f) + I_2^l(f)$, $l = 0, 1, 2$ can be arbitrarily small when M is sufficiently large, meaning that $c_0^{-l} \overline{Q}^{(l)}(f)$ is concentrated around $c_0^{-l} Q^{(l)}(f)$. To this end, we need the following results shown in [3].

Lemma 2 [3]: 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*

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

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 3 [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 $C > 0$.

D. Analysis of $\overline{Q}(f) - Q(f)$ and Its Derivatives

In this subsection we show that the quantities $c_0^{-l} \overline{Q}^{(l)}(f) - c_0^{-l} Q^{(l)}(f)$, $l = 0, 1, 2$ can be arbitrarily small on the unit circle. We first consider a set of finite grid points $\mathbb{T}_{grid} \subset \mathbb{T}$ and define the event

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

The following result states that $\mathcal{E}_2$ occurs with high probability if M is sufficiently large.

Proposition 2: Suppose $\mathbb{T}_{grid} \subset \mathbb{T}$ *is a finite set of grid points. Under the assumptions of Theorem 1, there exists a numerical constant* C *such that if*

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

$$\left. K \log \frac{K}{\delta} \left(1 + \frac{1}{L} \log \frac{|\mathbb{T}_{grid}|}{\delta} \right) \right\}, \quad (35)$$

then

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

To prove Proposition 2, we need to show that both $I_1^l(f)$ and $I_2^l(f)$ are small on $\mathbb{T}_{grid}$. To make the lower bound on M decrease as L increases, inspired by [24], we use the following lemma that generalizes the Bernstein inequality for Steinhaus sequences in [55, Proposition 16] to higher dimensions.

Lemma 4 [24]: Let $\boldsymbol{0} \neq \boldsymbol{w} \in \mathbb{C}^K$ *and* $\{\boldsymbol{\phi}_k\}_{k=1}^K$ *be a series of independent random vectors that are uniformly distributed on the complex sphere* $\mathbb{S}^{2L-1}$. *Then, for all* $t > \|\boldsymbol{w}\|_2$,

$$\mathbb{P}\left(\left\| \sum_{k=1}^K w_k \boldsymbol{\phi}_k \right\|_2 \geq t \right) \leq e^{-L\left(\frac{t^2}{\|\boldsymbol{w}\|_2^2} - \log \frac{t^2}{\|\boldsymbol{w}\|_2^2} - 1 \right)}. \quad (36)$$

The following result will be used to move the dependence on L from the upper bound on the probability in (36) to the sample size M.

Lemma 5: Let $y(x) \geq 1$ be the solution to the equation $y - \log y - 1 = x$ for $x \geq 0$. Then, $y(x)$ is monotonically increasing in x and $1 + x \leq y(x) \leq 2(1 + x)$.

Proof: See Appendix A. ∎

Applying Lemmas 4 and 5, we show in the following two lemmas that both $I_1^l(f)$ and $I_2^l(f)$ can be arbitrarily small on $\mathbb{T}_{grid}$ with M being a decreasing function of L.

Lemma 6: Under the assumptions of Theorem 1, there exists a numerical constant C such that if

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

$$\left. \frac{1}{\epsilon^2} \log^2 \frac{|\mathbb{T}_{grid}|}{\delta}, \; K \log \frac{K}{\delta} \right\},$$

then we have

$$\mathbb{P} \left\{ \sup_{f_d \in \mathbb{T}_{grid}} \left\| I_1^l(f_d) \right\|_2 \leq \epsilon, \; l = 0, 1, 2 \right\} \geq 1 - 9\delta.$$

Proof: See Appendix B. ∎

Lemma 7: Under the assumptions of Theorem 1, there exists a numerical constant C such that if

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

then we have

$$\mathbb{P} \left\{ \sup_{f_d \in \mathbb{T}_{grid}} \left\| I_2^l(f_d) \right\|_2 < \epsilon, \; l = 0, 1, 2 \right\} \geq 1 - 6\delta.$$

Proof: See Appendix C. ∎

Proposition 2 is a direct consequence of combining Lemmas 6 and 7. We next extend Proposition 2 from the set of finite grid points $\mathbb{T}_{grid}$ to the whole unit circle $\mathbb{T}$. To this end, for any $f \in \mathbb{T}$, we make the following decomposition:

$$\overline{Q}^{(l)}(f) - Q^{(l)}(f)$$
$$= \left(\overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right) + \left(\overline{Q}^{(l)}(f_d) - Q^{(l)}(f_d) \right)$$
$$+ \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right). \tag{37}$$

Because (37) holds for any $f_d \in \mathbb{T}_{grid}$, the inequalities in (38) then follow (shown at the bottom of this page).

The second term in (38) can be arbitrarily small according to Proposition 2. We next show that the first term can also

be arbitrarily small. Recall that $c_0^{-l} \overline{Q}^{(l)}(f) = \left\langle \Phi, \overline{L}^H \overline{v}_l(f) \right\rangle$, $c_0^{-l} Q^{(l)}(f) = \left\langle \Phi, L^H v_l(f) \right\rangle$, and thus

$$c_0^{-l} \left(\overline{Q}^{(l)}(f_d) - \overline{Q}^{(l)}(f) \right) + c_0^{-l} \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right)$$
$$= \left\langle \Phi, \overline{L}^H (\overline{v}_l(f_d) - \overline{v}_l(f)) + L^H (v_l(f_d) - v_l(f)) \right\rangle. \tag{39}$$

The magnitudes of $L^H (v_l(f) - v_l(f_d))$ and $\overline{L}^H (\overline{v}_l(f)) - \overline{v}_l(f_d)$ in (39) are controlled in the following lemma.

Lemma 8: Assume $\Delta_{\mathcal{T}} \geq \frac{1}{n}$. On the event $\mathcal{E}_{1,\tau}$, we have

$$\left\| L^H (v_l(f) - v_l(f_d)) \right\|_2 \leq C n^2 |f - f_d|, \tag{40}$$

$$\left\| \overline{L}^H (\overline{v}_l(f) - \overline{v}_l(f_d)) \right\|_2 \leq C n^3 |f - f_d| \tag{41}$$

for $f, f_d \in \mathbb{T}$ and some constant $C > 0$.

Proof: See Appendix D. ∎

Applying Lemmas 8, 4 and 5, we show in the following lemma that the first term on the right hand side of (38) can be arbitrarily small if $\mathbb{T}_{grid}$ is properly chosen, where the grid size $|\mathbb{T}_{grid}|$ decreases with L.

Lemma 9: Suppose $\mathbb{T}_{grid}$ is a finite set of uniform grid points. Under the assumptions of Theorem 1, there exists a constant C such that if $M \geq CK \log \frac{K}{\delta}$ and

$$|\mathbb{T}_{grid}| = \left\lceil C \frac{n^3}{\epsilon} \sqrt{1 + \frac{1}{L} \log \frac{1}{\delta}} \right\rceil, \tag{42}$$

then

$$\mathbb{P} \left(\sup_{f \in \mathbb{T}} \inf_{f_d \in \mathbb{T}_{grid}} c_0^{-l} \left\| \left(\overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right) \right. \right.$$
$$\left. \left. + \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right) \right\|_2 < \epsilon, \; l = 0, 1, 2 \right)$$
$$\geq 1 - 6\delta.$$

Proof: See Appendix E. ∎

Combining Proposition 2 and Lemma 9, we have the following proposition, where the bound on M in (43) is obtained by modifying (35) using (42) with the relaxation

$$|\mathbb{T}_{grid}| = \left\lceil C \frac{n^3}{\epsilon} \sqrt{1 + \frac{1}{L} \log \frac{1}{\delta}} \right\rceil$$
$$\leq C \frac{n^3}{\epsilon} \sqrt{2 \log \frac{1}{\delta}}$$
$$\leq \sqrt{2} C \frac{n^3}{\epsilon \sqrt{\delta}}.$$

$$\sup_{f \in \mathbb{T}} c_0^{-l} \left\| \overline{Q}^{(l)}(f) - Q^{(l)}(f) \right\|_2$$

$$\leq \sup_{f \in \mathbb{T}} \inf_{f_d \in \mathbb{T}_{grid}} \left\{ c_0^{-l} \left\| \left(\overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right) + \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right) \right\|_2 + c_0^{-l} \left\| \overline{Q}^{(l)}(f_d) - Q^{(l)}(f_d) \right\|_2 \right\}$$

$$\leq \sup_{f \in \mathbb{T}} \inf_{f_d \in \mathbb{T}_{grid}} c_0^{-l} \left\| \left(\overline{Q}^{(l)}(f) - \overline{Q}^{(l)}(f_d) \right) + \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right) \right\|_2 + \sup_{f_d \in \mathbb{T}_{grid}} c_0^{-l} \left\| \overline{Q}^{(l)}(f_d) - Q^{(l)}(f_d) \right\|_2. \tag{38}$$

Proposition 3: Under the assumptions of Theorem 1, there exists a numerical constant C such that if

$$M \geq C \frac{1}{\epsilon^2} \max \left\{ \log^2 \frac{n}{\epsilon\delta}, K \log \frac{K}{\delta} \left(1 + \frac{1}{L} \log \frac{n}{\epsilon\delta} \right) \right\}, \quad (43)$$

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

$$\sup_{f \in \mathbb{T}} c_0^{-l} \left\| \overline{Q}^{(l)}(f) - Q^{(l)}(f) \right\|_2 \leq \epsilon, \quad l = 0, 1, 2. \quad (44)$$

E. Completion of the Proof

We have shown by Proposition 3 that $\overline{Q}(f)$ (and its derivatives) can be arbitrarily close to $Q(f)$ (and its derivatives) with high probability provided that M satisfies (43). Following the same steps as in [56], we can show that $\overline{Q}(f)$ also satisfies (21), (22) and (23), if ϵ is taken to be a small value. In particular, it follows from (22) and (44) that for $f \in \mathcal{F}$,

$$\left\| \overline{Q}(f) \right\|_2 \leq \| Q(f) \|_2 + \left\| \overline{Q}(f) - Q(f) \right\|_2$$
$$\leq 1 - c_2 + \epsilon. \quad (45)$$

For $f \in \mathcal{N}_k$, it follows from (44) and the proof of [56, Lemma 6.8] that

$$\left| \frac{d^2 \left\| \overline{Q}(f) \right\|_2^2}{df^2} - \frac{d^2 \| Q(f) \|_2^2}{df^2} \right|$$
$$\leq \left(8C_1\epsilon + 4\epsilon^2 \right) c_0^2$$
$$\leq \frac{8}{3} \pi^2 \left(8C_1\epsilon + 4\epsilon^2 \right) n^2,$$

where C_1 is a constant. Then,

$$\frac{d^2 \left\| \overline{Q}(f) \right\|_2^2}{df^2}$$
$$\leq \frac{d^2 \| Q(f) \|_2^2}{df^2} + \left| \frac{d^2 \left\| \overline{Q}(f) \right\|_2^2}{df^2} - \frac{d^2 \| Q(f) \|_2^2}{df^2} \right|$$
$$\leq \left(-2c_1 + \frac{8}{3} \pi^2 \left(8C_1\epsilon + 4\epsilon^2 \right) \right) n^2. \quad (46)$$

Letting

$$\epsilon = \frac{1}{2} \min \left\{ c_2, \frac{3c_1}{4\pi^2 (8C_1 + 4)} \right\} \quad (47)$$

and substituting (47) into (45) and (46), we have

$$\left\| \overline{Q}(f) \right\|_2 \leq 1 - c_2', \quad f \in \mathcal{F}, \quad (48)$$

$$\frac{d^2 \left\| \overline{Q}(f) \right\|_2^2}{df^2} \leq -2c_1' n^2, \quad f \in \mathcal{N}_k, \quad (49)$$

where c_1' and c_2' are positive constants as well. By consecutively applying $\left\| \overline{Q}(f_k) \right\|_2^2 = 1$ (according to (27)), $\left. \frac{d \left\| \overline{Q}(f) \right\|_2^2}{df} \right|_{f=f_k} = 0$ (according to (28)) and (49), we have for $f \in \mathcal{N}_k$,

$$\left\| \overline{Q}(f) \right\|_2^2 = 1 + \int_{f_k}^{f} \frac{d \left\| \overline{Q}(s) \right\|_2^2}{ds} ds$$
$$= 1 + \int_{f_k}^{f} ds \int_{f_k}^{s} \frac{d^2 \left\| \overline{Q}(t) \right\|_2^2}{dt^2} dt$$
$$\leq 1 + \int_{f_k}^{f} ds \int_{f_k}^{s} -2c_1' n^2 dt$$
$$= 1 - c_1' n^2 (f - f_k)^2. \quad (50)$$

It follows from (48) and (50) that $\overline{Q}(f)$ satisfies (19). Therefore, $\overline{Q}(f)$ is a polynomial satisfying the constraints in (18)–(20), as required in Proposition 1. Finally, substituting (47) into (43) gives the bound on M in (4), completing the proof.

VI. CONCLUSION

A rigorous analysis was performed in this paper to confirm the observation that the frequency estimation performance of atomic norm minimization improves as the number of channels increases. The sample size per channel that ensures exact frequency estimation from noiseless data was derived as a decreasing function of the number of channels. Numerical results were provided that agree with our analysis.

While we have shown the performance gain of the use of multichannel data in reducing the sample size per channel, future work is to answer the question as to whether the resolution can be improved by increasing the number of channels. A positive answer to the question was suggested by empirical evidence presented in [18], [19], [22], and [25]. Another interesting and important future work is to analyze the noise robustness of atomic norm minimization with compressive data.

APPENDIX

A. Proof of Lemma 5

Consider x as a function of $y \geq 1$:

$$x(y) = y - \log y - 1.$$

Since the derivative $x'(y) = 1 - 1/y > 0$, as $y > 1$, we have that $x(y) \geq 0$ is monotonically increasing on $[1, +\infty)$. As a result, $y(x)$ is the inverse function of $x(y)$ and is monotonically increasing.

To show the second part of the lemma, we first note that

$$y = x + \log y + 1 \geq x + 1.$$

Then, using the fact that $\log y \leq \frac{1}{2} y$ for any $y \geq 1$, we have

$$x = y - \log y - 1 \geq y - \frac{1}{2} y - 1 = \frac{1}{2} y - 1.$$

It follows that $y \leq 2(1 + x)$, completing the proof.

B. Proof of Lemma 6

The proof of this lemma follows similar steps as those of the proofs of [3, Lemma IV.8] and [56, Lemma 6.6]. A main difference is the use of Lemma 4, rather than Hoeffding's inequality.

Recall that $I_1^l(f) = \left\langle \mathbf{\Phi}, \overline{\mathbf{L}}^H \left(\overline{\mathbf{v}}_l(f) - p\mathbf{v}_l(f) \right) \right\rangle$, where the rows of $\mathbf{\Phi}$ are independent random vectors that are uniformly distributed on the unit sphere. Conditioned on a particular realization $\omega \in \mathcal{E}$ where

$$\mathcal{E} = \left\{ \omega : \sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\| \overline{\mathbf{L}}^H \left(\overline{\mathbf{v}}_l(f_d) - p\mathbf{v}_l(f_d) \right) \right\|_2 < \lambda_l, \right.$$

$$\left. l = 0, 1, 2, 3 \right\},$$

Lemma 4 and the union bound then imply, for $\epsilon > \lambda_l$,

$$\mathbb{P}\left(\sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\| \left\langle \mathbf{\Phi}, \, \overline{\mathbf{L}}^H \left(\overline{\boldsymbol{v}}_l \left(f_d\right) - p\boldsymbol{v}_l\left(f_d\right)\right)\right\rangle\right\|_2 > \epsilon \middle| \omega\right)$$

$$\leq \left|\mathbb{T}_{\text{grid}}\right| e^{-L\left(\frac{\epsilon^2}{\lambda_l^2} - \log \frac{\epsilon^2}{\lambda_l^2} - 1\right)}. \quad (51)$$

It immediately follows that

$$\mathbb{P}\left(\sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\| \left\langle \mathbf{\Phi}, \, \overline{\mathbf{L}}^H \left(\overline{\boldsymbol{v}}_l \left(f_d\right) - p\boldsymbol{v}_l\left(f_d\right)\right)\right\rangle\right\|_2 > \epsilon\right)$$

$$\leq \left|\mathbb{T}_{\text{grid}}\right| e^{-L\left(\frac{\epsilon^2}{\lambda_l^2} - \log \frac{\epsilon^2}{\lambda_l^2} - 1\right)} + \mathbb{P}\left(\mathcal{E}^c\right).$$

Setting

$$\lambda_l = 4\left(2^{2l+3}\sqrt{\frac{K}{M}} + \frac{n}{M}a\bar{\sigma}_l\right) \quad (52)$$

in $\mathcal{E}$ and applying Lemma 2 yields

$$\mathbb{P}\left(\sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\| \left\langle \mathbf{\Phi}, \, \overline{\mathbf{L}}^H \left(\overline{\boldsymbol{v}}_l \left(f_d\right) - p\boldsymbol{v}_l\left(f_d\right)\right)\right\rangle\right\|_2 > \epsilon\right)$$

$$\leq \left|\mathbb{T}_{\text{grid}}\right| e^{-L\left(\frac{\epsilon^2}{\lambda_l^2} - \log \frac{\epsilon^2}{\lambda_l^2} - 1\right)}$$

$$+ 64 \left|\mathbb{T}_{\text{grid}}\right| e^{-\gamma a^2} + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right). \quad (53)$$

For the second term to be no greater than δ, a is chosen such that

$$a^2 = \gamma^{-1} \log \frac{64 \left|\mathbb{T}_{\text{grid}}\right|}{\delta}$$

and it will be fixed from now on. The first term is no greater than δ if

$$\frac{\epsilon^2}{\lambda_l^2} - \log \frac{\epsilon^2}{\lambda_l^2} - 1 \geq \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}.$$

Applying Lemma 5, this holds if

$$\frac{\epsilon^2}{\lambda_l^2} \geq 2\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right). \quad (54)$$

Note that (54) implies $\epsilon > \lambda_l$, which is required to verify (51).

We next derive a bound for M given (54). First consider the case of $2^4 K/\sqrt{M} \geq 1$. The condition in Lemma 2 is $a \leq \sqrt{2}M^{1/4}$ or equivalently

$$M \geq \frac{1}{4}a^4 = \frac{1}{4}\gamma^{-2}\log^2\frac{64\left|\mathbb{T}_{\text{grid}}\right|}{\delta}. \quad (55)$$

In this case, we have $a\bar{\sigma}_l \leq 2^{2l+3}\frac{\sqrt{MK}}{n}$, which inserting into (52) results in

$$\frac{1}{\lambda_l^2} = \frac{1}{16\left(2^{2l+3}\sqrt{\frac{K}{M}} + \frac{n}{M}a\bar{\sigma}_l\right)^2} \geq \frac{1}{4^{2l+5}}\frac{M}{K}.$$

Now consider the other case of $2^4 K/\sqrt{M} < 1$, which will be discussed in two scenarios. If $32s \geq a^2$, then $a\bar{\sigma}_l \leq 2^{2l+3}\frac{\sqrt{MK}}{n}$ which again gives the above lower bound on $\frac{1}{\lambda_l^2}$. Otherwise if $32s \leq a^2$, then $\lambda_l \leq 2^{2l+3}\sqrt{2}\frac{a}{\sqrt{M}}$ and

$$\frac{1}{\lambda_l^2} \geq \frac{1}{2^{4l+7}}\frac{M}{a^2}.$$

Therefore, to make (54) hold true, it suffices to take M satisfying (55) and

$$M \min\left(\frac{1}{4^{2l+5}}\frac{1}{K}, \frac{1}{2^{4l+7}}\frac{1}{a^2}\right) \geq \frac{2}{\epsilon^2}\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right).$$

By the arguments above, the first term on the right hand side of (53) is no greater than δ if

$$M \geq \max\left\{\frac{2}{\epsilon^2}4^{2l+5}K\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right),\right.$$

$$\frac{2}{\epsilon^2}2^{4l+7}\gamma^{-1}\log\frac{64\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right),$$

$$\left.\frac{1}{4}\gamma^{-2}\log^2\frac{64\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right\}. \quad (56)$$

According to Lemma 1, the last term on the right hand side of (53) is no greater than δ if

$$M \geq \frac{50}{\tau^2}K\log\frac{2K}{\delta}. \quad (57)$$

Setting $\tau = \frac{1}{4}$, combining (56) and (57) together, absorbing all constants into one, and using the inequality $\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right) \leq 2\log^2\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}$, we have

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

$$\left.\frac{1}{\epsilon^2}\log^2\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}, \; K\log\frac{K}{\delta}\right\}$$

is sufficient to guarantee

$$\sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\|I_1^l\left(f_d\right)\right\|_2 \leq \epsilon$$

with probability at least $1 - 3\delta$. Applying the union bound then completes the proof.

C. Proof of Lemma 7

Recall that $I_2^l(f) = \left\langle \mathbf{\Phi}, \left(\overline{\mathbf{L}} - p^{-1}\mathbf{L}\right)^H p\boldsymbol{v}_l(f)\right\rangle$. According to Lemma 3, we have on the set $\mathcal{E}_{1,\tau}$

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

for some constant $C > 0$. Applying Lemma 4 and the union bound gives for $\epsilon > C\tau$,

$$\mathbb{P}\left(\sup_{f_d \in \mathbb{T}_{\text{grid}}} \left\|I_2^l\left(f_d\right)\right\|_2 > \epsilon\right)$$

$$\leq \left|\mathbb{T}_{\text{grid}}\right| e^{-L\left(\frac{\epsilon^2}{C\tau^2} - \log\frac{\epsilon^2}{C\tau^2} - 1\right)} + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right).$$

To make the first term no greater than δ, we take τ such that

$$\frac{\epsilon^2}{C\tau^2} - \log\frac{\epsilon^2}{C\tau^2} - 1 \geq \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}.$$

According to Lemma 5, it then suffices to fix τ such that

$$\frac{1}{\tau^2} = C\frac{2}{\epsilon^2}\left(1 + \frac{1}{L}\log\frac{\left|\mathbb{T}_{\text{grid}}\right|}{\delta}\right).$$

To make the second term no greater than δ, it suffices by Lemma 1 to take

$$M \geq \frac{C}{\tau^2} K \log \frac{2K}{\delta}$$
$$= C \frac{1}{\epsilon^2} K \log \frac{2K}{\delta} \left(1 + \frac{1}{L} \log \frac{|\mathbb{T}_{\text{grid}}|}{\delta} \right).$$

Application of the union bound proves the lemma.

D. Proof of Lemma 8

Recall by inserting (24) into $\boldsymbol{v}_l(f)$ as given in (31) that

$$\boldsymbol{v}_l(f) = \frac{1}{n} \sum_{j=-2n}^{j=2n} \left(\frac{i 2\pi j}{c_0} \right)^l g_n(j) e^{i 2\pi f j} \boldsymbol{e}(j)$$

where

$$\boldsymbol{e}(j) = \begin{bmatrix} e^{-i 2\pi f_1 \, j} \\ \vdots \\ e^{-i 2\pi f_K j} \\ \frac{i 2\pi j}{c_0^2} e^{-i 2\pi f_1 \, j} \\ \vdots \\ \frac{i 2\pi j}{c_0^2} e^{-i 2\pi f_K j} \end{bmatrix}.$$

It follows that

$$\boldsymbol{v}_l(f) - \boldsymbol{v}_l(f_d)$$
$$= \frac{1}{n} \sum_{j=-2n}^{j=2n} \left(\frac{i 2\pi j}{c_0} \right)^l g_n(j) \left(e^{i 2\pi f j} - e^{i 2\pi f_d j} \right) \boldsymbol{e}(j).$$

Using the following bounds shown in [3], [4]:

$$\|g_n\|_\infty \leq 1,$$
$$\left| \frac{i 2\pi j}{c_0} \right| \leq 4 \quad \text{when } n \geq 2,$$
$$\|\boldsymbol{e}(j)\|_2^2 \leq 14K \quad \text{when } n \geq 4$$

and the bound

$$\left| e^{i 2\pi f j} - e^{i 2\pi f_d j} \right| = \left| e^{i \pi (f + f_d) j} \cdot 2i \sin \left(\pi (f - f_d) j \right) \right|$$
$$= 2 \left| \sin \left(\pi (f - f_d) j \right) \right|$$
$$\leq 2\pi \left| f - f_d \right| \cdot |j|$$
$$\leq 4 n \pi \left| f - f_d \right|,$$

we have

$$\|\boldsymbol{v}_l(f) - \boldsymbol{v}_l(f_d)\|_2$$
$$\leq \frac{1}{n} \sum_{j=-2n}^{j=2n} \left| \frac{i 2\pi j}{c_0} \right|^l |g_n(j)| \left| e^{i 2\pi f j} - e^{i 2\pi f_d j} \right| \|\boldsymbol{e}(j)\|_2$$
$$\leq \frac{1}{n} \cdot (4n+1) \cdot 4^l \cdot 4 n \pi \left| f - f_d \right| \cdot 14K$$
$$\leq C n K \left| f - f_d \right|$$
$$\leq C n^2 \left| f - f_d \right|.$$

We then obtain

$$\left\| \boldsymbol{L}^H \left(\boldsymbol{v}_l(f) - \boldsymbol{v}_l(f_d) \right) \right\|_2 \leq \|\boldsymbol{L}\|_2 \|\boldsymbol{v}_l(f) - \boldsymbol{v}_l(f_d)\|_2$$
$$\leq C n^2 \left| f - f_d \right|,$$

since $\|\boldsymbol{L}\|_2 \leq \left\| \boldsymbol{D}^{-1} \right\|_2 \leq 1.568$ according to [4].

The bound in (41) can be shown using similar arguments, while the exponent for n increases from 2 to 3 since $\|\overline{\boldsymbol{L}}\|_2 \leq 2 \left\| \boldsymbol{D}^{-1} \right\|_2 p^{-1} \leq C \frac{n}{M} \leq C n$ according to [3].

E. Proof of Lemma 9

It follows from Lemma 8 that

$$\left\| \overline{\boldsymbol{L}}^H \left(\overline{\boldsymbol{v}}_l(f_d) - \overline{\boldsymbol{v}}_l(f) \right) + \boldsymbol{L}^H \left(\boldsymbol{v}_l(f_d) - \boldsymbol{v}_l(f) \right) \right\|_2$$
$$\leq \left\| \overline{\boldsymbol{L}}^H \left(\overline{\boldsymbol{v}}_l(f_d) - \overline{\boldsymbol{v}}_l(f) \right) \right\|_2 + \left\| \boldsymbol{L}^H \left(\boldsymbol{v}_l(f_d) - \boldsymbol{v}_l(f) \right) \right\|_2$$
$$\leq C n^3 \left| f - f_d \right|.$$

We then recall (33) and apply Lemma 4, having for $\epsilon > C n^3 \left| f - f_d \right|$,

$$\mathbb{P}\left(c_0^{-l} \left\| \left(\overline{Q}^{(l)}(f_d) - \overline{Q}^{(l)}(f) \right) \right.\right.$$
$$\left.\left. + \left(Q^{(l)}(f_d) - Q^{(l)}(f) \right) \right\|_2 \geq \epsilon \right)$$
$$\leq e^{-L \left(\frac{\epsilon^2}{C n^6 |f - f_d|^2} - \log \frac{\epsilon^2}{C n^6 |f - f_d|^2} - 1 \right)} + \mathbb{P} \left(\mathcal{E}_{1,\tau}^c \right).$$

For the second term to be no greater than δ, it suffices to take $M \geq C K \log \frac{K}{\delta}$ by Lemma 1. For the first term to be no greater than δ, according to Lemma 5, it suffices to let

$$\frac{\epsilon^2}{C n^6 |f - f_d|^2} \geq 2 \left(1 + \frac{1}{L} \log \frac{1}{\delta} \right),$$

or equivalently

$$\frac{1}{|f - f_d|} \geq C \frac{n^3}{\epsilon} \sqrt{1 + \frac{1}{L} \log \frac{1}{\delta}}. \tag{58}$$

Therefore, to guarantee, for any $f \in \mathbb{T}$, that some $f_d \in \mathbb{T}_{\text{grid}}$ can always be found such that (58) holds, it suffices to let $\mathbb{T}_{\text{grid}}$ be a uniform grid with

$$|\mathbb{T}_{\text{grid}}| = \left\lceil C \frac{n^3}{\epsilon} \sqrt{1 + \frac{1}{L} \log \frac{1}{\delta}} \right\rceil.$$

Application of the union bound completes the proof.

F. Proof of Corollary 1

Similar to atomic norm minimization in Theorem 1, to certify optimality for $\ell_{2,1}$ norm minimization in (15) it suffices to construct a dual certificate $\overline{Q}(f) = \boldsymbol{a}^H(f) \boldsymbol{V}$ satisfying (see also [24, Th. 3.1])

$$\overline{Q}(f_k) = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \tag{59}$$
$$\|\overline{Q}(f)\|_2 < 1, \quad f \in \{\tilde{f}_g\}_{g=1}^G \setminus \mathcal{T}, \tag{60}$$
$$\boldsymbol{V}_j = \boldsymbol{0}, \quad j \in \Omega^c. \tag{61}$$

The only difference from the dual certificate for atomic norm minimization lies in that only finitely many constraints are involved in (60). Therefore, the dual certificate constructed in Section V for atomic norm minimization is naturally a certificate for $\ell_{2,1}$ norm minimization.

REFERENCES

[1] Z. Yang, Y. C. Eldar, and L. Xie, "Average case analysis of compressive multichannel frequency estimation using atomic norm minimization," in *Proc. 23rd Int. Conf. Digit. Signal Process. (DSP)*, 2018.

[2] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.

[3] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Nov. 2013.

[4] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Commun. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, Jun. 2014.

[5] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[6] W. Heylen and P. Sas, *Modal Analysis Theory and Testing*. Leuven, Belgium: Katholieke Univ. Leuven, 2006.

[7] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli, "Estimation of sparse MIMO channels with common support," *IEEE Trans. Commun.*, vol. 60, no. 12, pp. 3705–3716, Dec. 2012.

[8] J. Li and P. Stoica, "MIMO radar with colocated antennas," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 106–114, Sep. 2007.

[9] E. Baransky, G. Itzhak, N. Wagner, I. Shmuel, E. Shoshan, and Y. Eldar, "Sub-Nyquist radar prototype: Hardware and algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 2, pp. 809–822, Apr. 2014.

[10] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[11] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature Methods*, vol. 3, no. 10, pp. 793–796, 2006.

[12] M. Rossi, A. M. Haimovich, and Y. C. Eldar, "Spatial compressive sensing for MIMO radar," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 419–430, Jan. 2014.

[13] M. J. Schervish, *Theory of Statistics*. New York, NY, USA: Springer, 2012.

[14] M. Wax and I. Ziskind, "On unique localization of multiple sources by passive sensor arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 996–1000, Jul. 1989.

[15] E. J. Candès and C. Fernandez-Granda, "Super-resolution from noisy data," *J. Fourier Anal. Appl.*, vol. 19, no. 6, pp. 1229–1254, 2013.

[16] J.-M. Azaïs, Y. De Castro, and F. Gamboa, "Spike detection from inaccurate samplings," *Appl. Comput. Harmon. Anal.*, vol. 38, no. 2, pp. 177–195, 2015.

[17] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 499–512, Jan. 2015.

[18] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5145–5157, Oct. 2016.

[19] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1257–1269, Mar. 2016.

[20] C. Fernandez-Granda, G. Tang, X. Wang, and L. Zheng, "Demixing sines and spikes: Robust spectral super-resolution in the presence of outliers," *Inf. Inference, A J. IMA*, vol. 7, no. 15, pp. 105–168, 2018.

[21] S. Li, D. Yang, G. Tang, and M. B. Wakin, "Atomic norm minimization for modal analysis from random and compressed samples," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1817–1831, Apr. 2018.

[22] C. Fernandez-Granda, "Super-resolution of point sources via convex programming," *Inf. Inference, A J. IMA*, vol. 5, no. 3, pp. 251–303, Sep. 2016.

[23] 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," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 655–687, Dec. 2008.

[24] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[25] Z. Yang and L. Xie, "Enhancing sparsity and resolution via reweighted atomic norm minimization," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 995–1006, Feb. 2016.

[26] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[27] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, May 1989.

[28] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[29] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[30] W. Liao and A. Fannjiang, "MUSIC for single-snapshot spectral estimation: Stability and super-resolution," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 1, pp. 33–67, 2016.

[31] A. Fannjiang. (2016). "Compressive spectral estimation with single-snapshot ESPRIT: Stability and resolution." [Online]. Available: https://arxiv.org/abs/1607.01827

[32] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf. Proc. (ICASSP)*, vol. 3, May 1996, pp. 1688–1691.

[33] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[34] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.

[35] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[36] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[37] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[38] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. Math.*, vol. 3, 2006, pp. 1433–1452.

[39] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, Nov. 1993, pp. 40–44.

[40] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[41] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[42] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, Feb. 2011.

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

[44] I. Haviv and O. Regev, "The restricted isometry property of subsampled Fourier matrices," in *Geometric Aspects of Functional Analysis*. Springer, 2017, pp. 163–179.

[45] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, Dec. 2012.

[46] Z. Yang, L. Xie, and C. Zhang, "A discretization-free sparse and parametric approach for linear array signal processing," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4959–4973, Oct. 2014.

[47] Z. Yang and L. Xie, "On gridless sparse methods for line spectral estimation from complete and incomplete data," *IEEE Trans. Signal Process.*, vol. 63, no. 12, pp. 3139–3153, Jun. 2015.

[48] Z. Yang, J. Li, P. Stoica, and L. Xie, "Sparse methods for direction-of-arrival estimation," in *Academic Press Library in Signal Processing*, vol. 7, R. Chellappa and S. Theodoridis, Eds. New York, NY, USA: Academic, 2018, pp. 509–581.

[49] J. Fang, F. Wang, Y. Shen, H. Li, and R. S. Blum, "Super-resolution compressed sensing for line spectral estimation: An iterative reweighted approach," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4649–4662, Sep. 2016.

[50] X. Wu, W.-P. Zhu, J. Yan, and Z. Zhang, "Two sparse-based methods for off-grid direction-of-arrival estimation," *Signal Process.*, vol. 142, pp. 87–95, Jan. 2018.

[51] C. Steffens, M. Pesavento, and M. E. Pfetsch, "A compact formulation for the $\ell_{2,1}$ mixed-norm minimization problem," *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1483–1497, Mar. 2018.

[52] B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*. Dordrecht, The Netherlands: Springer, 2007.

[53] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2017.

[54] M. Grant and S. Boyd. (2008). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://cvxr.com/cvx/

[55] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 1, pp. 1–24, Jul. 2008.

[56] Z. Yang and L. Xie. (May 2014). "Exact joint sparse frequency recovery via optimization methods." [Online]. Available: https://arxiv.org/abs/1405.6585v1

Zai Yang (S'11–M'14) received the B.S. degree in mathematics in 2007 and the M.S. degree in applied mathematics in 2009 both from Sun Yat-sen (Zhongshan) University, China, and the Ph.D degree in electrical and electronic engineering in 2014 from Nanyang Technological University (NTU), Singapore. He is currently a Professor of the School of Mathematics and Statistics, Xi'an Jiaotong University, China. He was a Research Fellow of NTU from 2013 to 2015 and a Professor of the School of Automation, Nanjing University of Science and Technology from 2016 to 2018. His research interests are in compressed sensing, statistical signal processing, optimization methods, and their applications in array and radar processing and machine learning. He is serving on the editorial board of *Signal Processing*.

Jinhui Tang (M'08–SM'14) received the B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He has authored over 150 papers in top-tier journals and conferences. His current research interests include multimedia analysis and search, computer vision, and machine learning. Dr. Tang was a recipient of the best paper awards in ACM MM 2007, PCM 2011, and ICIMCS 2011, the Best Paper Runner-up in ACM MM 2015, and the best student paper awards in MMM 2016 and ICIMCS 2017. He has served as an Associate Editor for the IEEE TNNLS, IEEE TKDE, and IEEE TCSVT.

Yonina C. Eldar (S'98–M'02–SM'07–F'12) received the B.Sc. degree in Physics in 1995 and the B.Sc. degree in Electrical Engineering in 1996 both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in Electrical Engineering and Computer Science in 2002 from the Massachusetts Institute of Technology (MIT), Cambridge. She is currently a Professor in the Department of Electrical Engineering at the Technion - Israel Institute of Technology, Haifa, Israel, where she holds the Edwards Chair in Engineering. She is also a Visiting Professor with the Research Laboratory of Electronics at MIT, an Adjunct Professor at Duke University, and was a Visiting Professor at Stanford University, Stanford, CA. She is a member of the Israel Academy of Sciences and Humanities (elected 2017), an IEEE Fellow and a EURASIP Fellow. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

Dr. Eldar has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016). She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technions Award for Excellence in Teaching (two times). She received several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize, the UFFC Outstanding Paper Award, the Signal Processing Society Best Paper Award and the IET Circuits, Devices and Systems Premium Award, and was selected as one of the 50 most influential women in Israel. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is the Editor in Chief of Foundations and Trends in Signal Processing, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*. She was Co-Chair and Technical Co-Chair of several international conferences and workshops. She is author of the book "*Sampling Theory: Beyond Bandlimited Systems*" and co-author of the books "*Compressed Sensing*" and "*Convex Optimization Methods in Signal Processing and Communications*", all published by Cambridge University Press.

Lihua Xie (S'91–M'92–SM'97–F'07) received the B.E. and M.E. degrees in electrical engineering from Nanjing University of Science and Technology in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Newcastle, Australia, in 1992. Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a professor and Director, Delta-NTU Corporate Laboratory for Cyber-Physical Systems. He served as the Head of Division of Control and Instrumentation from July 2011 to June 2014. He held teaching appointments in the Department of Automatic Control, Nanjing University of Science and Technology from 1986 to 1989.

Dr Xie's research interests include robust control and estimation, networked control systems, multi-agent networks, localization and unmanned systems. He is an Editor-in-Chief for Unmanned Systems and an Associate Editor for IEEE TRANSACTIONS ON NETWORK CONTROL SYSTEMS. He has served as an editor of IET Book Series in Control and an Associate Editor of a number of journals including IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II. He is an elected member of Board of Governors, IEEE Control System Society (Jan 2016-Dec 2018). Dr Xie is a Fellow of IEEE and a Fellow of IFAC.