# Structured LISTA for Multidimensional Harmonic Retrieval

Rong Fu, Yimin Liu, *Member, IEEE*, Tianyao Huang, and Yonina C. Eldar, *Fellow, IEEE*

***Abstract*—Learned iterative shrinkage thresholding algorithm (LISTA), which adopts deep learning techniques to optimize algorithm parameters from labeled training data, can be successfully applied to small-scale multidimensional harmonic retrieval (MHR) problems. However, LISTA becomes computationally demanding for large-scale MHR because the matrix size of the learned mutual inhibition matrix exhibits quadratic growth with the signal length. These large matrices consume costly memory/computation resources and require a huge amount of labeled data for training. For MHR problems, the mutual inhibition matrix naturally has a Toeplitz structure, implying the degrees of freedom of the matrix can be reduced from quadratic order to linear order. We thereby propose a structured LISTA-Toeplitz network, which imposes Toeplitz structure on the mutual inhibition matrices and applies linear convolution instead of matrix-vector multiplications in traditional LISTA. Both simulation and field tests for air target detection with radar are carried out to validate the performance of the proposed network. For small-scale MHR problems, LISTA-Toeplitz exhibits close or even better recovery accuracy than traditional LISTA, while the former significantly reduces the network complexity and requires much less training data. For large-scale MHR problems, where LISTA is difficult to implement due to the huge size of the matrices, our proposed LISTA-Toeplitz still enjoys good recovery performance.**

***Index Terms*—Compressed sensing, multidimensional harmonic retrieval, iterative shrinkage thresholding algorithm, learned ISTA, Toeplitz structure.**

## I. Introduction

We consider the harmonic retrieval problem, of recovering the underlying frequencies and amplitudes of harmonic signals from observed time-domain samples. This problem appears in a wide range of applications such as wireless communication channel estimation [2], [3], beampattern synthesis [4], direction-of-arrival (DOA) estimation [5]–[7] and range-Doppler estimation [8]. Extended to multi-dimensional frequency models, MHR has been extensively studied in the signal processing literature. Standard methods for MHR include fast Fourier transform (FFT) spectral estimation [9], [10] (periodogram), Welch's method [11] (also called the modified periodogram) and subspace techniques [12], [13]. According to the Nyquist criterion, harmonic signals are generally uniformly sampled at or above the Nyquist rate to avoid aliasing of the spectrum.

In various applications, it is desirable to minimize the required number of Nyquist samples needed for MHR [14]. The demand for sample reduction becomes more significant for MHR problems with large dimensions. Consider DOA estimation with antenna arrays as an example, where each active antenna transmits and receives signals reflected from one or more moving targets to estimate the direction angles. Fewer samples imply fewer active antennas, lowering the hardware cost [15]–[18].

Sparse recovery or compressed sensing (CS) [18], [19] has been suggested to reduce the measurement data size, when the sinusoids are spectrally sparse, namely when there is a small number of harmonics. Particularly, consider a harmonic retrieval problem, where a time-domain signal, consisting of $K$ distinct complex sinusoids, has $M$ Nyquist samples. When the sinusoid frequencies lie precisely on a set of discrete grids, the signal of interest can be sparsely represented by a discrete basis. CS suggests that the sparse signal can be recovered from a random subset with reduced size of $N = O(K \log M)$ out of the $M$ Nyquist samples [14]. Recovery can be performed by finding a sparse representation of the time-domain signal over a discrete harmonic dictionary $\mathbf{\Phi}$. Thus, CS formulates MHR as a linear decoding problem

$$\boldsymbol{y} = \mathbf{\Phi}\boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{C}^N$ is the obtained sub-Nyquist samples, $\mathbf{\Phi} \in \mathbb{C}^{N\times M}$ is the observation matrix, $\boldsymbol{x} \in \mathbb{C}^M$ denotes the sparse spectral representation of the unknown sinusoids, and $\boldsymbol{w} \in \mathbb{C}^N$ is additive noise.

A myriad of methods including greedy algorithms and optimization-based approaches [20]–[24] have been developed to solve CS problems. Among these techniques, a key concern for applications is real-time realization [22]. One of the well-known $\ell_1$-norm regularization techniques is the iterative shrinkage thresholding algorithm (ISTA) [23], which has desirable global rates of convergence to the optimum solution [24]. An accelerated method, fast ISTA (FISTA), can speed up the rate of convergence by adding a momentum term [23]. However, it takes many iterations for ISTA or FISTA to reach a sparse

Manuscript received November 27, 2020; revised March 31, 2021 and May 13, 2021; accepted May 13, 2021. Date of publication June 7, 2021; date of current version June 23, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was E. Chouzenoux. This work is supported by the National Natural Science Foundation of China under Grant 61801258. Part of this paper was presented at the IEEE Radar Conference (RadarConf), Boston, MA, USA, April 2019 [1]. *(Corresponding author: Tianyao Huang.)*

Rong Fu, Yimin Liu, and Tianyao Huang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: fu-r16@mails.tsinghua.edu.cn; yiminliu@tsinghua.edu.cn; huangtianyao2009@gmail.com).

Yonina C. Eldar is with the Faculty of Math and CS, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Digital Object Identifier 10.1109/TSP.2021.3086593

1053-587X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

representation, which inevitably gives rise to high computational cost.

The approximation ability of deep learning motivates considering the possibilities of recovering sparse signals at a small computational cost through a neural network. Unfolding the iterations of ISTA, LISTA was proposed by Gregor and LeCun [25], and has demonstrated superior convergence speed over ISTA in both theoretical analysis and empirical results [1], [26], [27]. Following this idea, many researchers [26]–[29] have unfolded other iterative techniques and incorporated them into deep networks, which have shown their efficiency in many recovery problems. For example, approximated message passing (AMP) proposed in [30], can also be unfolded to construct a feed-forward neural network called learned AMP [26], [27]. It is common for these reconstruction networks to learn some intended network variables for the model information in the dictionary matrix $\mathbf{\Phi}$. However, in some cases like large-scale CS problems, the dictionary $\mathbf{\Phi}$ has many columns so that there will be a huge amount of network variables to learn. In particular, the size of the mutual inhibition matrix, which depends on the Gram matrix $\mathbf{\Phi}^H\mathbf{\Phi}$, scales as $M^2$. Training a network with high-dimensional variables takes much more training time and memory and requires large training datasets to avoid overfitting.

To reduce the number of trainable parameters, many convolutional extensions based on LISTA were proposed for the applications of optical image super-resolution, denoising or inpainting [31]–[33]. These networks exploit the fact that dictionary matrices in these applications are Toeplitz matrices or concatenations of Toeplitz matrices, which allows to impose the Toeplitz structure constraint on the learned matrices in LISTA and replace the matrix-multiplication operations in LISTA with convolutions. Since the degrees of freedom (DoF) of a $M \times M$ Toeplitz matrix are $O(M)$, such convolutional networks reduce the number of variables to be learned significantly.

The dictionaries of MHR problems have a Fourier structure. Thus, these convolutional extensions are not directly applicable. However, the Gram matrix of the dictionary in MHR is a Toeplitz matrix, which motivates us to explore the redundancy of parameters in LISTA by Toeplitz projections. Many structured networks [34]–[38] have been proposed by adopting similar parameter-sharing strategies, i.e., taking advantage of the specific structure in the model to achieve a high compression ratio and nearly retain the performance of state of the art models. Based on LISTA, a structured network [1] has been proposed to deal with DOA estimation problems, which can be viewed as a 1D harmonic retrieval problem. In this paper, we extend this structured network to MHR. To build such a network, we formulate the signal model for compressive MHR and explore the Toeplitz-related structure of the Gram matrix both in 1D and 2D settings. In our network, referred to as LISTA-Toeplitz, we impose a Toeplitz structure constraint on the mutual inhibition matrix and replace the corresponding matrix-multiplication operations with convolution filters, while the rest of the components in LISTA are not modified.

For comparison, we also consider an alternative convolutional extension of LISTA named convolutional-LISTA (ConvLISTA) which imposes a convolutional prior on the dictionary matrix and forces all the linear blocks in LISTA to be convolutional. To demonstrate the effectiveness of 1D and 2D LISTA-Toeplitz networks, we assess the performance of oracle LISTA, LISTA-Toeplitz and ConvLISTA with synthetic data and also apply LISTA-Toeplitz to real radar data for air target recovery. The proposed network, designed for signal models where the Gram matrix of the dictionary has a Toeplitz structure, yields fast and accurate reconstruction on both synthetic and real data, and outperforms ConvLISTA in harmonic retrieval scenarios. The main contributions of this paper are as follows:

1) We introduce LISTA-Toeplitz which exploits the Toeplitz structure in one-dimensional (1D) harmonic retrieval problems and the doubly block-Toeplitz structure in two-dimensional (2D) harmonic retrieval, and imposes a Toeplitz structure restriction on the learned mutual inhibition matrices in LISTA. LISTA-Toeplitz achieves model order reduction in the number of network variables, significantly reducing the network complexity and training size.
2) We use linear convolution to perform the multiplication by the Toeplitz matrices, which significantly relieves the storage burden. It can further reduce computation complexity by using efficient algorithms to speed up the calculation of linear convolution.
3) Both simulated and real data validate the effectiveness of our proposed network and demonstrate its improved recovery performance over traditional LISTA and its previous convolutional extension ConvLISTA.

The rest of this paper is organized as follows. The signal model in MHR, the conventional CS solutions, and our motivations to use LISTA and LISTA-Toeplitz, are introduced in Section II. In Section III we explore the Toeplitz structure in harmonic retrieval problems and develop the corresponding LISTA-Toeplitz network. To show the effectiveness of the proposed networks, in Section IV we apply our network to 1D and 2D harmonic retrieval problems with both synthetic and real data. Section V concludes the paper.

*Notation:* The symbol $\mathbb{C}$ represents the set of complex numbers. Correspondingly, $\mathbb{C}^M$ and $\mathbb{C}^{M\times N}$ are the sets of the $M$-dimensional ($M$-D) vectors and $M \times N$ matrices of complex numbers, respectively. The subscripts $[\cdot]_i$ and $[\cdot]_{i,k}$ are used for the $i$-th entry of a vector and the $i$-th row, $k$-th column entry of a matrix. We let $[\cdot]$ and $\{\cdot\}$ denote a vector/matrix and a set, respectively. We use a set in subscript to construct a vector/matrix or set, e.g., for a set $\mathcal{N} := \{0, 1, \ldots, N-1\}$ and vectors $\boldsymbol{x}_n \in \mathbb{C}^M$, $n \in \mathcal{N}$, $[\boldsymbol{x}_n]_{n\in\mathcal{N}}$ and $\{\boldsymbol{x}_n\}_{n\in\mathcal{N}}$ representing the matrix $[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}] \in \mathbb{C}^{M\times N}$ and the set $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}\}$, respectively. To distinguish from the set $\mathcal{N}$, we use the asterisk notation in the superscript to denote $\mathcal{N}^* := \{-N+1, -N+2, \ldots, N-1\}$. The transpose and Hermitian transpose are written as the superscripts $(\cdot)^T$ and $(\cdot)^H$, respectively. For a vector, $\|\cdot\|_0$ and $\|\cdot\|_q$ denote the $\ell_0$ "norm" and $\ell_q$ norm, $q \geq 1$, respectively. Operators $*$, $\circ$ and $\otimes$ represent linear convolution, element-wise multiplication and Kronecker product [39], respectively.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

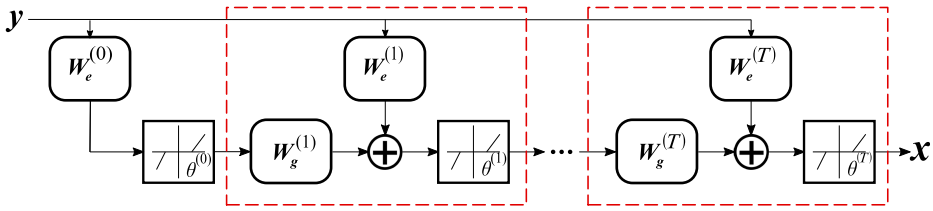


Fig. 1. Block diagram of LISTA [25], in which the red dashed-line boxes indicate the process of one ISTA iteration.

## II. Motivation and Problem Formulation

Here, we introduce the motivation for proposing the LISTA-Toeplitz algorithm, i.e., accelerating ISTA-based methods for MHR problems. To this aim, we first review preliminaries on the typical ISTA approach and its variations, including FISTA and LISTA, in Subsection II-A. Then in Subsection II-B, we formulate the MHR signal model and show that the inherent Toeplitz structure in MHR can be exploited to improve existing iterative algorithms, yielding the LISTA-Toeplitz algorithm. The flow of LISTA-Toeplitz is detailed in Section III.

### A. Preliminaries

In the framework of CS, the dictionary matrix $\boldsymbol{\Phi}$ in (1) is usually under-determined. ISTA methods harness a sparse prior to estimate $\boldsymbol{x}$ by using regularized regression [40], [41]

$$\min_{\boldsymbol{x}} f\left(\boldsymbol{\Phi x}, \boldsymbol{y}\right) + \lambda \|\boldsymbol{x}\|_1, \tag{2}$$

where $f(\boldsymbol{\Phi x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2$ measures the error, and $\lambda$ is a regularization parameter controlling the sparsity penalty characterized by the $\ell_1$ norm.

Standard ISTA iteratively performs proximal gradient descent with respect to the cost function in (2) [23], [42]. Specifically, the sparse solution in the $(t+1)$-th iteration, denoted $\boldsymbol{x}^{(t+1)}$, is formed by the following recursion:

$$\boldsymbol{x}^{(t+1)} = \mathcal{S}_{\frac{\lambda}{L}}\left(\frac{1}{L}\boldsymbol{\Phi}^H\boldsymbol{y} + \left(\boldsymbol{I} - \frac{1}{L}\boldsymbol{\Phi}^H\boldsymbol{\Phi}\right)\boldsymbol{x}^{(t)}\right), \tag{3}$$

where $\lambda$ is the regularization parameter in (2), $L$ is the Lipschitz constant $L = \lambda_{\max}(\boldsymbol{\Phi}^H\boldsymbol{\Phi})$, and $\lambda_{\max}(\cdot)$ represents the maximum eigenvalue of a Hermitian matrix. The element-wise soft-threshold operator $\mathcal{S}_\theta(\cdot)$ is defined as

$$\left[\mathcal{S}_\theta\left(\boldsymbol{u}\right)\right]_i = \text{sign}\left(\left[\boldsymbol{u}\right]_i\right)\left(\left|\left[\boldsymbol{u}\right]_i\right| - \theta\right)_+, \tag{4}$$

where $\text{sign}(\cdot)$ returns the sign of a scalar, $(\cdot)_+$ means $\max(\cdot, 0)$, and $\theta$ is the threshold.

ISTA can be very accurate in recovering sparse signals. However, it may take thousands of iterations for convergence [22], which motivates many variants of ISTA with a higher convergence rate, including the well-known FISTA and LISTA techniques. FISTA can be viewed as a Nesterov's accelerated version of ISTA. It is know that ISTA and FISTA requires $\mathcal{O}(1/\varepsilon)$ and $\mathcal{O}(1/\varepsilon^2)$ iterations for achieving a desired accuracy of $\varepsilon$ [43], [44].

To further accelerate ISTA, Gregor and LeCun proposed a neural network containing only several layers, named LISTA [25], [28]. Each layer is an unfolded version of the ISTA iteration (3) that can be rewritten as

$$\boldsymbol{x}^{(t+1)} = \mathcal{S}_{\theta^{(t)}}\left(\boldsymbol{W}_e^{(t)}\boldsymbol{y} + \boldsymbol{W}_g^{(t)}\boldsymbol{x}^{(t)}\right), \tag{5}$$

where the terms $\lambda/L$, $\frac{1}{L}\boldsymbol{\Phi}^H\boldsymbol{y}$ and $(\boldsymbol{I} - \frac{1}{L}\boldsymbol{\Phi}^H\boldsymbol{\Phi})$ in (3) are replaced by $\theta^{(t)}$, $\boldsymbol{W}_e^{(t)} \in \mathbb{C}^{M\times N}$ and $\boldsymbol{W}_g^{(t)} \in \mathbb{C}^{M\times M}$, respectively. The matrices $\boldsymbol{W}_e^{(t)}$ and $\boldsymbol{W}_g^{(t)}$ are named the *filter matrix* and the *mutual inhibition matrix*, respectively. Fig. 1 illustrates a LISTA network structure with $T$ layers.

In contrast to ISTA and FISTA, where parameters in each iteration are identical and are calculated analytically or set manually, LISTA treats the tuple $\{\boldsymbol{W}_e^{(t)}, \boldsymbol{W}_g^{(t)}, \theta^{(t)}\}$ in each layer, $t = 0, 1, \ldots, T-1$, as variables to learn from some predefined training data using back-propagation [28]. The network performs better when the values of parameters differ in each layer. We omit the superscript $(t)$ in the following notation for simplicity. Numerical results show that LISTA can achieve virtually the same accuracy as the original ISTA using nearly two orders of magnitude fewer iterations [26], [45].

Nevertheless, a challenge in LISTA is that there are many variables to learn. In general $N \ll M$. For example, we have $N = 64$ and $M = 1024$ for air target recovery shown in Subsection IV-C. With a large-scale sparse signal $\boldsymbol{x} \in \mathbb{C}^M$, the mutual inhibition matrix $\boldsymbol{W}_g$ of size $M \times M$, which is much larger than $\boldsymbol{W}_e$, plays a dominant role among all the variables, especially when allowing $\boldsymbol{W}_g$ to vary across layers. A large neural network also leads to a high computational burden. The time complexity for LISTA is mainly due to matrix multiplication $\boldsymbol{W}_g^{(t)}\boldsymbol{x}^{(t)}$, which requires $\mathcal{O}(M^2)$ floating point operations (flops) at each layer. In addition, large neural networks require careful tuning of hyper-parameters such as learning rates and initialization values to avoid training problems such as overfitting [46] and gradient vanishing [47].

This motivates us to tailor the network to fit our specific problem so that we can impose restrictions on $\boldsymbol{W}_g$ in order to reduce the computational burden. For certain CS problems like MHR, the mutual inhibition matrix naturally has a Toeplitz structure, which can be exploited to reduce dimension.

### B. Toeplitz Structure in MHR Problems

We next introduce the signal model of MHR problems, associated with many practical applications including 1D, 2D DOA estimation and range-Doppler recovery in radar systems. The Toeplitz structures of 1D and 2D harmonic retrieval problems are discussed in Subsection II-B1 and II-B2, respectively.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

*1) Toeplitz Structure of 1D Harmonic Retrieval:* First, we consider a 1D harmonic retrieval problem, following the signal model presented in [14]. Assume that there are $K$ distinct complex-valued sinusoids, with their normalized frequencies denoted by $f_k$, $k = 1, 2, \ldots, K$. The observation taken at the $i$-th time instant can be written as a superposition of these $K$ complex-valued sinusoids.

$$y_i = \sum_{k=1}^{K} a_k e^{j2\pi f_k i}, \tag{6}$$

where $a_k$ denotes the complex amplitude of the $k$-th sinusoid.

We uniformly discretize frequencies into $M = M_1$ grids and assume that the $K$ sinusoids are on the grids, i.e., $f_k = \frac{m_k}{M_1}$ where $m_k \in \mathcal{M}_1 := \{0, 1, \ldots, M_1 - 1\}$ . Thus we recast (6) in matrix form as

$$\boldsymbol{y}^\star = \boldsymbol{\Psi}\boldsymbol{x}, \tag{7}$$

where $\boldsymbol{x} \in \mathbb{C}^M$ contains only $K$ nonzero elements, corresponding to the complex amplitudes of the $K$ sinusoids, and $\boldsymbol{\Psi}$ is the $M_1 \times M_1$ discrete Fourier matrix $\boldsymbol{F}_{M_1}$ whose $(i, m)$-th entry is

$$[\boldsymbol{\Psi}]_{i,m} = [\boldsymbol{F}_{M_1}]_{i,m} = e^{\mathrm{j}2\pi \frac{m}{M_1} i}, \;\; i, m \in \mathcal{M}_1. \tag{8}$$

We consider compressive measurements, where only $N$ entries of $\boldsymbol{y}^\star$ are observed with $N \ll M_1$. To store the indices of the selected entries from $\boldsymbol{y}^\star$, we define a subset $\Omega$ of cardinality $N$ randomly chosen from the set $\mathcal{M}_1$. Then we use a row-subsampled matrix $\boldsymbol{R}$ to select $N$ rows of $\boldsymbol{\Psi}$ corresponding to the elements in $\Omega$, i.e., $[\boldsymbol{R}]_{n,m} = 1$, where $m$ is the $n$-th element of $\Omega$ while other entries in the $n$-th row are zeros. The sub-sampled observations are denoted by $\boldsymbol{y} \in \mathbb{C}^N$, with

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{y}^\star = \boldsymbol{R}\boldsymbol{\Psi}\boldsymbol{x}. \tag{9}$$

Here, we use $\boldsymbol{\Phi} = \boldsymbol{R}\boldsymbol{\Psi}$ to represent the new dictionary matrix, consisting of $N$ sub-sampled rows of the full dictionary $\boldsymbol{\Psi}$.

The Gram matrix of this dictionary matrix is $\boldsymbol{\Phi}^H\boldsymbol{\Phi} = \boldsymbol{F}_{M_1}^H \boldsymbol{R}^H \boldsymbol{R} \boldsymbol{F}_{M_1}$, which can be formulated as

$$\boldsymbol{\Phi}^H\boldsymbol{\Phi} = \sum_{m=0}^{M-1} \mathbb{1}_\Omega(m) \boldsymbol{\phi}_m \boldsymbol{\phi}_m^H, \tag{10}$$

where $\mathbb{1}_\Omega(m)$ is an indicator function which equals 1 when $m \in \Omega$ and zero otherwise, and $\boldsymbol{\phi}_m$ is the $m$-th column of the Fourier matrix $\boldsymbol{F}_{M_1}$. It can be verified that $\boldsymbol{\phi}_m \boldsymbol{\phi}_m^H$ is a Hermitian Toeplitz matrix. Consequently, the Gram matrix can be viewed as a sum of $N$ Hermitian Toeplitz matrices, which is also a Hermitian Toeplitz matrix. Note that for a Hermitian Toeplitz matrix, the DoF is $M$. If we disregard the Hermitian symmetry, the DoF is slightly larger, $2M - 1$. In both cases, the DoF are much less than $M^2$, that of a general matrix without structure.

*2) Doubly-Block Toeplitz Structure of 2D Harmonic Retrieval:* For 2D harmonic retrieval, we also assume that there are $K$ distinct complex-valued sinusoids, with their frequencies denoted by $\boldsymbol{f}_k = [f_{1,k}, f_{2,k}]^T \in [0, 1)^2$, $k = 1, 2, \ldots, K$. Following the 2D problem presented in [14], the observations taken at time index $\boldsymbol{i} = [i_1, i_2]^T$ can be represented by a 2D tensor, $\boldsymbol{Y}^\star \in \mathbb{C}^{M_1 \times M_2}$, where $M_p$ denotes the number of indices in the $p$-th dimension. Each element is computed as

$$y_{\boldsymbol{i}}^\star = [\boldsymbol{Y}^\star]_{i_1,i_2} = \sum_{k=1}^{K} a_k e^{j2\pi \boldsymbol{f}^T \boldsymbol{i}}, \tag{11}$$

where $a_k$ denotes the complex amplitude of the $k$-th sinusoid. We then uniformly sample $M_p$ points in the $p$-th frequency dimension, yielding $M_p$ grid points, encapsulated in the set of grid points $\{m/M_p\}_{m \in \mathcal{M}_p}$, and assume that the frequencies $f_{p,k}$ belong to such a set.

Let $M = M_1 M_2$ be the number of total frequency grid points in all dimensions. Considering compressive measurements, we define the dictionary matrix as a Kronecker product of Fourier matrices with respect to all dimensions, given by $\boldsymbol{\Phi} = \boldsymbol{R}(\boldsymbol{F}_{M_1} \otimes \boldsymbol{F}_{M_2})$. The Gram matrix can then be formulated as

$$\begin{aligned}
\boldsymbol{\Phi}^H\boldsymbol{\Phi} &= (\boldsymbol{F}_{M_1} \otimes \boldsymbol{F}_{M_2})^H \boldsymbol{R}^H \boldsymbol{R} (\boldsymbol{F}_{M_1} \otimes \boldsymbol{F}_{M_2}) \\
&\overset{(a)}{=} \sum_{m_2=0}^{M_2-1} \sum_{m_1=0}^{M_1-1} \mathbb{1}_\Omega\{(m_1, m_2)\} \left(\boldsymbol{\phi}_{m_1} \otimes \boldsymbol{\psi}_{m_2}\right) \left(\boldsymbol{\phi}_{m_1}^H \otimes \boldsymbol{\psi}_{m_2}^H\right) \\
&\overset{(b)}{=} \sum_{m_2=0}^{M_2-1} \sum_{m_1=0}^{M_1-1} \mathbb{1}_\Omega\{(m_1, m_2)\} \left(\boldsymbol{\phi}_{m_1} \boldsymbol{\phi}_{m_1}^H\right) \otimes \left(\boldsymbol{\psi}_{m_2} \boldsymbol{\psi}_{m_2}^H\right),
\end{aligned}$$

where $\boldsymbol{\phi}_{m_1}$ and $\boldsymbol{\psi}_{m_2}$ denote the $m_1$-th, $m_2$-th column of the Fourier matrix $\boldsymbol{F}_{M_1}$ and $\boldsymbol{F}_{M_2}$, respectively. The equality (a) is a consequence of the fact that taking the complex conjugate or transpose before carrying out the Kronecker product yields the same result as doing so afterward, and (b) follows the mixed product property of the Kronecker product [39].

Similarly, both $\boldsymbol{A} := \boldsymbol{\phi}_{m_1} \boldsymbol{\phi}_{m_1}^H \in \mathbb{C}^{M_1 \times M_1}$ and $\boldsymbol{B} := \boldsymbol{\psi}_{m_2} \boldsymbol{\psi}_{m_2}^H \in \mathbb{C}^{M_2 \times M_2}$ are Hermitian Toeplitz matrices. Thus, the Toeplitz matrix $\boldsymbol{A}$ can be represented by a $(2M_1 - 1)$ dimensional vector, denoted by $[a_l]_{l \in \mathcal{M}_1^*}$, where $\mathcal{M}_1^* := \{-M_1 + 1, -M_1 + 2, \ldots, M_1 - 1\}$ and we use the asterisk notation in the superscript to distinguish it from the set $\mathcal{M}_p$. Particularly, the $(i, j)$-th element of $\boldsymbol{A}$ can be denoted as $[\boldsymbol{A}]_{i,j} = a_{i-j}$. According to the definition of Kronecker product, we have

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} \boldsymbol{C}_0 & \boldsymbol{C}_{-1} & \cdots & \boldsymbol{C}_{-M_1+1} \\ \boldsymbol{C}_1 & \boldsymbol{C}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{C}_{-1} \\ \boldsymbol{C}_{M_1-1} & \cdots & \boldsymbol{C}_1 & \boldsymbol{C}_0 \end{bmatrix},$$

where $\boldsymbol{C}_i = a_i \boldsymbol{B}$, $i \in \mathcal{M}_1^*$.

The matrix $\boldsymbol{A} \otimes \boldsymbol{B}$ has a so-called doubly-block Toeplitz structure [48], [49], i.e., each block matrix is itself a Toeplitz matrix and is repeated down the diagonals of the entire matrix. In the literature, such a matrix is also called two-fold block Toeplitz [14] or Toeplitz-block-Toeplitz matrix [50]. After the sum operation over $m_1$ and $m_2$, the doubly-block Toeplitz structure is preserved, implying that the Gram matrix in the 2D MHR problem is also doubly-block Toeplitz. Observing the structure of $\boldsymbol{\Phi}^H\boldsymbol{\Phi}$, we find that it is constructed by two Hermitian Toeplitz matrices, with DoFs given by $M_1$ and $M_2$, respectively. To compromise on the representational probability of LISTA-Toeplitz, we ignore the Hermitian structure. Consequently, the

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

DoF of the Gram matrix $\boldsymbol{\Phi}^H\boldsymbol{\Phi}$ is $(2M_1-1)(2M_2-1)$, much less than $M^2=M_1^2M_2^2$, the number of elements in $\boldsymbol{\Phi}^H\boldsymbol{\Phi}$.

Evidently, the Gram matrix $\boldsymbol{\Phi}^H\boldsymbol{\Phi}$ possesses Toeplitz or Toeplitz related structure. Hence, the mutual inhibition matrix $\boldsymbol{W}_g$ to learn, which corresponds to $(\boldsymbol{I}-\frac{1}{L}\boldsymbol{\Phi}^H\boldsymbol{\Phi})$, is also Toeplitz structured and thus compressible. Inspired by this phenomenon in MHR problems, we propose a heuristically structured network called LISTA-Toeplitz by imposing a Toeplitz structure on $\boldsymbol{W}_g$ in LISTA. Applying such a structure-imposing approach benefits performance by the reduction in space complexity. We next explain how to build the proposed LISTA-Toeplitz network.

## III. LISTA-Toeplitz Network Design

As shown in Section II-B, 1D/2D harmonic retrieval problems naturally possess a Toeplitz/doubly-block Toeplitz structure on the Gram matrix $\boldsymbol{\Phi}^H\boldsymbol{\Phi}$. In this section, we impose the Toeplitz structure restriction on the corresponding variables of the LISTA network, yielding the LISTA-Toeplitz network. The motivation for using the Toeplitz structure in building the network architecture is multi-fold: 1) Neural networks with structured weight matrices will reduce the model order in the number of network variables enabling the application to large-scale sparse recovery problems; 2) It is much more efficient to train such a model-based network.

### A. 1D LISTA-Toeplitz Network

We first consider 1D harmonic retrieval problems and design the LISTA-Toeplitz network by exploiting the Toeplitz structure in the mutual inhibition matrix $\boldsymbol{W}_g\in\mathbb{C}^{M\times M}$, where $M$ is the number of 1D grid points.

Since a $M$ dimensional Toeplitz matrix can be represented by a $2M-1$ dimensional vector, we denote such a vector by $\boldsymbol{h}:=[h_m]_{m\in\mathcal{M}^*}^T\in\mathbb{C}^{2M-1}$. Consequently, the matrix $\boldsymbol{W}_g$ is expressed by $[\boldsymbol{W}_g]_{i,k}=h_{i-k}, i,k\in\mathcal{M}$, i.e.

$$\boldsymbol{W}_g=\begin{bmatrix} h_0 & h_{-1} & h_{-2} & \cdots & h_{-M+1}\\ h_1 & h_0 & h_{-1} & \ddots & \vdots\\ h_2 & \ddots & \ddots & \ddots & h_{-2}\\ \vdots & \cdots & \ddots & \ddots & h_{-1}\\ h_{M-1} & \cdots & h_2 & h_1 & h_0 \end{bmatrix}. \tag{12}$$

Here, we disregard the Hermitian structure of $\boldsymbol{W}_g$, to achieve a compromise between the expressive power and compactness of the network architecture.

Based on the Toeplitz structure of $\boldsymbol{W}_g$, the multiplication operation $\boldsymbol{W}_g\boldsymbol{x}$ in (5) can be expressed by a linear convolution between two vectors and realized by off-the-shelf toolboxes [51], [52]. To see this, note that the linear convolution between $\boldsymbol{h}$ and $\boldsymbol{x}\in\mathbb{C}^M$, denoted $\boldsymbol{h}*\boldsymbol{x}$, yields a $M$ dimensional vector with entries given by

$$[\boldsymbol{h}*\boldsymbol{x}]_i=\sum_{k=0}^{M-1}h_{i-k}[\boldsymbol{x}]_k,\ i\in\mathcal{M}. \tag{13}$$

The right hand side in (13) also equals $[\boldsymbol{W}_g\boldsymbol{x}]_i$, according to the definition of $\boldsymbol{W}_g$ in (12) and matrix multiplication. Hence, we have

$$\boldsymbol{W}_g\boldsymbol{x}=\boldsymbol{h}*\boldsymbol{x}. \tag{14}$$

Substituting (14) into (5) implies a proximal mapping using linear convolution, given by

$$\boldsymbol{x}^{(t+1)}=\mathcal{S}_{\theta^{(t)}}\left(\boldsymbol{W}_e\boldsymbol{y}+\boldsymbol{h}*\boldsymbol{x}^{(t)}\right). \tag{15}$$

Based on (15), we replace each layer of LISTA with a structured network, which yields the 1D version of LISTA-Toeplitz, given in Fig. 2. Furthermore, considering the Hermitian symmetry of the Gram matrix, we can furthermore force $\boldsymbol{h}\in\mathbb{C}^{2M-1}$ to be centrohermitian, i.e., $h_m=h_{-m}^*,\ m\in\mathcal{M}$. Thus we also construct another convolutional network named LISTA-ToeplitzSymm to compare with LISTA and LISTA-Toeplitz networks.

Comparing Fig. 1 and 2, we note that the difference between conventional LISTA and LISTA-Toeplitz lies in the realization of the multiplication $\boldsymbol{W}_g\boldsymbol{x}$. The latter uses a dimension-reduced vector $\boldsymbol{h}$ to replace the large matrix $\boldsymbol{W}_g$, and applies linear convolution (14). The reduction in space complexity and computational complexity by using structured matrices is significant. Taking advantage of the Toeplitz structure in the MHR model, LISTA-Toeplitz achieves superior accuracy-compactness-speed trade-offs, which are detailed as follows.

a) With a compressed network structure, LISTA-Toeplitz enjoys comparable recovery performance as LISTA while LISTA-ToeplitzSymm suffers from slight performance degradation. We evaluate the performance of LISTA, LISTA-Toeplitz and LISTA-ToeplitzSymm with different numbers of layers varying from 1 to 11 and different noise power $\sigma^2$ increasing from $-40$ dB up to 11 dB. The recovered NMSE of each network is shown in Fig. 3. Here, we set the number of grid points $M=512$ and the number of observation samples $N=64$. According to the signal model (9), we perform 100 Monte Carlo trials for each noise level to determine the recovered NMSE. More details on network hyper-parameters and initialization values are shown in Appendix B. In Fig. 3, the darker color represents better recovery performance. As increasing the number of layers $T$ does not show obvious improvement when $T>7$, both LISTA and LISTA-Toeplitz are able to reconstruct signals using around $T=10$ layers with noise power less than 0 dB. On the other hand, the recovered NMSE of LISTA-ToeplitzSymm is worse than the former two networks approximately 3 dB. Thus, in later sections only the Toeplitz structure is fully exploited to reduce the complexity of networks in LISTA while the Hermitian symmetry structure is disregarded.

b) For space complexity, the memory demand of the network is reduced, which also contributes to lowering the cost and power consumption of the networks [53]. While 1D LISTA needs to update the weight matrix $\boldsymbol{W}_g$ of size $M^2$, our proposed 1D LISTA-Toeplitz network only needs to learn $2M-1$ elements for $\boldsymbol{W}_g$. Thus, the space complexity of

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

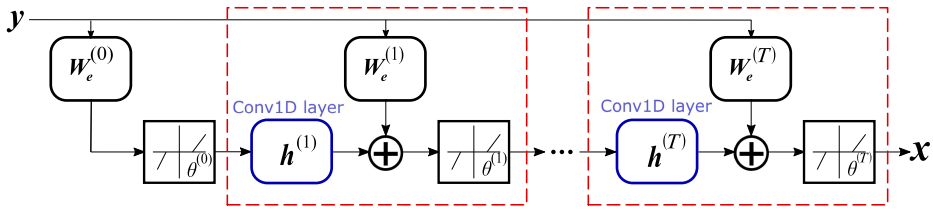


Fig. 2. Block diagram of 1D LISTA-Toeplitz network. The red dashed-line boxes indicate each layer of LISTA-Toeplitz, corresponding to the whole process of an ISTA iteration. Blue boxes represent convolutional layers (14), which highlight the modification of LISTA-Toeplitz over standard LISTA.

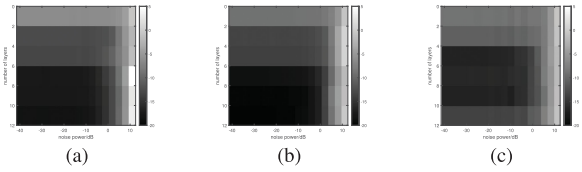


Fig. 3. Recovered NMSE (dB) of (a) LISTA and (b) LISTA-Toeplitz (c) LISTA-ToeplitzSymm containing $T = 1$ to 11 layers under different noise level.

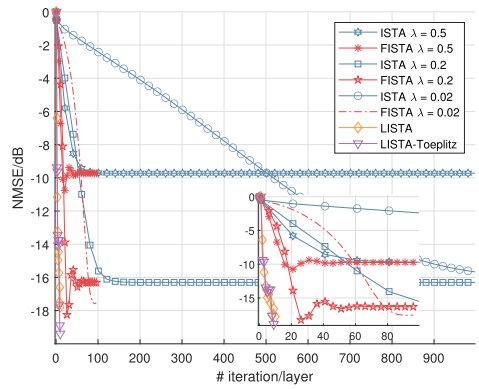


Fig. 4. The NMSE (dB) of different algorithms and networks in each iteration/layer with noise power $\sigma^2 = -4$ dB.

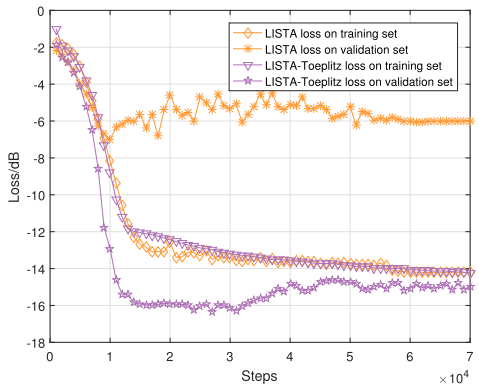


Fig. 5. The learning curve using limited training samples of LISTA and LISTA-Toeplitz.

LISTA-Toeplitz is $O(M)$. This is a significant advantage compared to LISTA which requires $O(M^2)$ parameters. For large-scale harmonic retrieval problems, our proposed LISTA-Toeplitz decreases memory requirements by a factor of $M$, which greatly relieves the storage burden.

c) Inherited from LISTA, LISTA-Toeplitz has superior convergence speed over ISTA and FISTA. The recovered NMSE results for each iteration (ISTA and FISTA) or layer (LISTA and LISTA-Toeplitz) are shown in Fig. 4, where we use a fixed number of layers $T = 10$ and set the noise power $\sigma^2 = -4$ dB. For ISTA and FISTA algorithms, the regularization parameter $\lambda$ in (2) is set as 0.02, 0.2 and 0.5. It is shown that ISTA and FISTA suffer from an inherent trade-off between recovery performance and convergence speed based on the choice of $\lambda$; a larger $\lambda$ leads to faster convergence speed but a poorer reconstruction result. On the other hand, LISTA and LISTA-Toeplitz learn the parameters such as $\lambda$ by end-to-end training to achieve the best estimation performance with limited computational time. As the total computation time is proportional to the number of iterations, LISTA and LISTA-Toeplitz converge faster than the original ISTA and FISTA thereby reducing the computational cost.

Resorting to linear convolution has some additional advantages. There are many methods to improve the computational efficiency of linear convolution. For example, after training, we can use FFT to speed up the computation process of linear convolution by substituting $\boldsymbol{h} * \boldsymbol{x}^{(t)}$ with $\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{h}) \circ \mathcal{F}(\boldsymbol{x}^{(t)}))$, where $\circ$ denotes element-wise multiplication and $\mathcal{F}$ denotes the FFT operator while $\mathcal{F}^{-1}$ is the inverse FFT operator. Thus in LISTA-Toeplitz, the linear convolution in (15) can be computed in the Fourier domain with time complexity $O(M \log M)$ instead of $O(M^2)$ in traditional LISTA.

Because of the huge reduction in network variables, the proposed structured network performs better on limited training data. To validate the sample efficiency of LISTA-Toeplitz, we provide another experiment under the same signal model $\boldsymbol{\Phi} \in \mathbb{C}^{64\times 512}$ and the number of network layers is set as 10. In this setting, LISTA has more than 2 million parameters to learn. However, we only prepare 1000000 samples for training and another 1000 samples as the validation set. According to the practical criterion in [54]–[56], LISTA with millions of parameters will easily overfit this limited training dataset. The learning curves indicated in Fig. 5 also help to identify whether LISTA and LISTA-Toeplitz overfit when using the same training and validation set. We observe that after updating 10000 steps LISTA starts to over-learn the training data as its loss on validation set no longer decreases. On the other hand, LISTA-Toeplitz learns well on this data set, and retains its recovery quality even with less training samples, demonstrating that taking advantage of the Toeplitz structure enhances sample efficiency in training.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

TABLE I
THE ONE-LAYER COMPLEXITY OF THE LISTA AND LISTA-TOEPLITZ NETWORKS

| Network | Time | Space |
|---|---|---|
| LISTA | $O(M^2)$ | $O(M^2)$ |
| LISTA-Toeplitz | $O(M \log M)$ | $O(M)$ |

To summarize, different from other network compression approaches such as Principal Filter Analysis (PFA) which provides a heuristic compression factor [57], the Toeplitz constraint used in the proposed methods enables model order reduction and provides huge computational complexity reduction: The storage requirement is reduced from $O(M^2)$ to $O(M)$ and the computational complexity can be reduced from $O(M^2)$ to $O(M \log M)$. Table I compares the time and space complexity of LISTA and 1D LISTA-Toeplitz.These benefits are also applicable to higher dimensional harmonic retrieval problems, though the network architecture is slightly different, as discussed in the next subsection.

### *B. 2D LISTA-Toeplitz Network*

We next extend the LISTA-Toeplitz network to 2D harmonic retrieval problems. As shown in Subsection II-B2, the mutual inhibition matrix $\boldsymbol{W}_g$ naturally follows a doubly-block Toeplitz structure, which requires a slight change in the network settings of LISTA-Toeplitz. For clarity, we denote such a network by 2D LISTA-Toeplitz. Recall that in the 2D harmonic retrieval problem, we use $M_1$ and $M_2$ to denote the cardinality of the grid sets in the first and second dimension, respectively, and $M = M_1 M_2$. The sparse vector $\boldsymbol{x}$ has a block structure, expressed as $\boldsymbol{x} = [\boldsymbol{x}_{m_2}^T]_{m_2 \in \mathcal{M}_2}^T$, where the sub-vector $\boldsymbol{x}_{m_2}$ is denoted by $\boldsymbol{x}_{m_2} = [x_{m_1,m_2}]_{m_1 \in \mathcal{M}_1}^T \in \mathbb{C}^{M_1}$.

Following a similar procedure as the 1D case, we first use a dimension reduced matrix $\boldsymbol{H} \in \mathbb{C}^{(2M_1-1)\times(2M_2-1)}$ to represent $\boldsymbol{W}_g \in \mathbb{C}^{M_1 M_2 \times M_1 M_2}$. This is possible, because the DoF of $\boldsymbol{W}_g$ is $(2M_1 - 1) \times (2M_2 - 1)$ due to its Toeplitz structure. For convenience, $\boldsymbol{H}$ is expressed as $\boldsymbol{H} := [h_{m_1,m_2}]_{m_1 \in \mathcal{M}_1^*, m_2 \in \mathcal{M}_2^*}$. Here we also ignore the Hermitian structure.

To link between $\boldsymbol{H}$ and $\boldsymbol{W}_g$, we first define $2M_2 - 1$ Toeplitz sub-matrices $\boldsymbol{H}_m \in \mathbb{C}^{M_1 \times M_1}, m \in \mathcal{M}_2^*$, which are constructed from the $m$-th column of $\boldsymbol{H}$ as

$$\boldsymbol{H}_m = \begin{bmatrix} h_{0,m} & h_{-1,m} & \cdots & h_{-M_1+1,m} \\ h_{1,m} & h_{0,m} & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{-1} \\ h_{M_1-1,.m} & \cdots & h_{1,m} & h_{0,m} \end{bmatrix}. \tag{16}$$

With these Toeplitz matrices, we construct a doubly-block Toeplitz matrix $\mathcal{T}(\boldsymbol{H}) \in \mathbb{C}^{M_1 M_2 \times M_1 M_2}$ as

$$\mathcal{T}(\boldsymbol{H}) := \begin{bmatrix} \boldsymbol{H}_0 & \boldsymbol{H}_{-1} & \cdots & \boldsymbol{H}_{-M_2+1} \\ \boldsymbol{H}_1 & \boldsymbol{H}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{H}_{-1} \\ \boldsymbol{H}_{M_2-1} & \cdots & \boldsymbol{H}_1 & \boldsymbol{H}_0 \end{bmatrix}. \tag{17}$$

According to (16) and (17), we can verify that $\boldsymbol{W}_g = \mathcal{T}(\boldsymbol{H})$ if the elements of $\boldsymbol{H}$ satisfy

$$h_{i-j,s-t} = [\boldsymbol{H}_{s-t}]_{i,j} = [\boldsymbol{W}_g]_{i+(s-1)\times M_1, j+(t-1)\times M_1}, \tag{18}$$

where $m_2 = s - t \in \mathcal{M}_2^*$ denotes the index of sub-matrix $\boldsymbol{H}_{m_2}$ in the $\boldsymbol{W}_g$, and $m_1 = i - j \in \mathcal{M}_2^*$ denotes the position of the element inside the sub-matrix $\boldsymbol{H}_{m_2}$.

We then prove that the result of $\boldsymbol{W}_g \boldsymbol{x}$ can be calculated by linear convolution operations as follows. Since the vector $\boldsymbol{x}$ has a nested structure, we define $\boldsymbol{X} := [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_{M_2}] \in \mathbb{C}^{M_1 \times M_2}$, which satisfies that $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X})$. The linear convolution operation between two matrices is defined as

$$[\boldsymbol{H} * \boldsymbol{X}]_{s,t} = \sum_{j=0}^{M_2-1} \sum_{i=0}^{M_1-1} h_{s-i,t-j} x_{i,j}, \tag{19}$$

where $s \in \mathcal{M}_1, t \in \mathcal{M}_2$. Then,

$$\begin{aligned} [\boldsymbol{W}_g \boldsymbol{x}]_l &= \sum_{k=1}^{M_1 M_2} [\boldsymbol{W}_g]_{l,k} [\boldsymbol{x}]_k \\ &= \sum_{j=1}^{M_2} \sum_{i=1}^{M_1} [\boldsymbol{W}_g]_{l,i+jM_1} x_{i,j}. \end{aligned} \tag{20}$$

Comparing (19) and (20), we conclude that $\boldsymbol{W}_g \boldsymbol{x} = \mathrm{vec}(\boldsymbol{H} * \boldsymbol{X})$ by setting $l = s + tM_1$ in the subscript in (20). Thus, the unfolded ISTA iteration (5) can be rewritten as

$$\boldsymbol{x}^{(t+1)} = \mathcal{S}_{\theta^{(t)}} \left( \boldsymbol{W}_e \boldsymbol{y} + \mathrm{vec} \left( \boldsymbol{H} * \boldsymbol{X}^{(t)} \right) \right), \tag{21}$$

where $\boldsymbol{X}^{(t)} \in \mathbb{C}^{M_1 \times M_2}$ is obtained by reshaping $\boldsymbol{x}^{(t)} \in \mathbb{C}^{M_1 M_2}$, i.e., $\boldsymbol{X}^{(t)} = [\boldsymbol{x}_1^{(t)}\ \boldsymbol{x}_2^{(t)}\ \cdots\ \boldsymbol{x}_{M_2}^{(t)}]$. We illustrate the 2D LISTA-Toeplitz network in Fig. 6.

As with 1D LISTA-Toeplitz, the 2D version significantly reduces the number of variables in the weight matrix compared with its counterpart LISTA, from $O(M_1^2 M_2^2)$ to $O(M_1 M_2)$. For a typical case where $M_* = M_1 = M_2$, the reduction in spacial complexity, from $O(M_*^4)$ to $O(M_*^2)$, is more notable than the 1D case. In $p$-dimension harmonic retrieval problems, with $M_* = M_1 = M_2 = \cdots = M_P$, the LISTA-Toeplitz network can be extended, and the benefits of exploiting the Toeplitz structure becomes more dominant: The memory demand for storing the variables to learn will be decreased from $O(M_*^{2P})$ to $O(M_*^P)$. As discussed in Subsection III-A, the compressed network also contributes to making the network more trainable, lowering the cost and power consumption.

Note that this convolutional prior is imposed on the Gram matrix of the dictionary matrix rather than the dictionary itself. That is why we replace matrix multiplication with linear convolution just for the mutual inhibition part $\boldsymbol{W}_g \boldsymbol{x}^{(t)}$. For comparison, we also try making the linear multiplication by $\boldsymbol{W}_e$ to be a convolution, although there is no equivalence between $\boldsymbol{W}_e \boldsymbol{y}$ and $\boldsymbol{h}_e * \boldsymbol{y}$ where $\boldsymbol{W}_e \in C^{M \times N}$ and $\boldsymbol{h}_e \in C^{M+N-1}$. We achieve the architecture of this convolutional network, named ConvLISTA, as follows

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \mathcal{S}_{\theta^{(t)}} \left( \boldsymbol{h}_e * \boldsymbol{y} + \boldsymbol{h} * \boldsymbol{x}^{(t)} \right), \text{for 1D MHR}, \\ \mathcal{S}_{\theta^{(t)}} \left( \boldsymbol{h}_e * \boldsymbol{y} + \mathrm{vec} \left( \boldsymbol{H} * \boldsymbol{X}^{(t)} \right) \right), \text{for 2D MHR}. \end{cases}$$

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

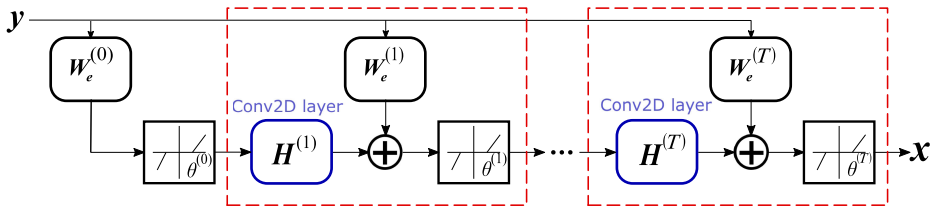


Fig. 6. Block diagram of 2D LISTA-Toeplitz network. The red dashed-line boxes indicate each layer of LISTA-Toeplitz network, corresponding to the whole process of an ISTA iteration. Blue boxes represent linear convolutional layers (14), which highlight the modification of LISTA-Toeplitz over standard LISTA.

In MHR problems, the recovery performance of ConvLISTA, where both linear blocks are implemented by convolutions, is worse than LISTA and our LISTA-Toeplitz networks. The numerical results of these networks are shown in the next section.

## IV. Practical Applications and Numerical Results

In this section, we perform numerical experiments involving 1D and 2D harmonic retrieval problems to demonstrate the effectiveness of the proposed LISTA-Toeplitz in comparison with the original LISTA network, ConvLISTA and conventional iterative algorithms including ISTA, FISTA and AMP. As presented in the sequel, we showcase our performance by using both synthetic data and real data. In Subsection IV-A and IV-B, we generate synthetic data for both 1D and 2D harmonic retrieval problems, respectively. The reconstruction quality of these methods is measured in terms of normalized mean squared error (NMSE) and hit rate. Here, NMSE means the mean squared error of the recovered signal $\hat{\boldsymbol{x}}$ normalized by the power of ground truth $\boldsymbol{x}$, given by

$$\text{NMSE} = \text{E}\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 / \|\boldsymbol{x}\|_2. \tag{22}$$

Hit rate is defined as the percentage of the number of correctly recovered components to the total number of nonzero entries in $\boldsymbol{x}$, i.e., the sparsity $K$ of signal $\boldsymbol{x}$. For every nonzero entry $[\boldsymbol{x}]_i$, it is defined to be correctly recovered if $i \in S^K(\widehat{\boldsymbol{x}})$ where $S^K(\boldsymbol{v}) = \{i_1, i_2, \ldots, i_K | |\boldsymbol{v}_{i_1}| \geq |\boldsymbol{v}_{i_2}| \cdots \geq |\boldsymbol{v}_{i_K}| \cdots \geq |\boldsymbol{v}_{i_M}|\}$ which includes the indexes of elements with the largest $K$ magnitudes in the vector $\boldsymbol{v} \in \mathbb{C}^M$.

In Subsection IV-C, we apply our proposed 2D LISTA-Toeplitz to real radar measurements, verifying its performance on real data.

In the following experiments, we set the maximum number of iterations for ISTA, FISTA and AMP at 1000, 100 and 50, respectively, while LISTA, ConvLISTA and LISTA-Toeplitz have only 10 layers. For LISTA, ConvLISTA and our LISTA-Toeplitz methods, the performance relies on not only the network structure but also the amount of training data, which will be specified in the following experiments individually. Following the practical criterion in [54]–[56], the size of training data is set at roughly ten times the number of unknown parameters in the network.

For each training pair, we generate sparse signals $\boldsymbol{x}$ as i.i.d. Bernoulli-Gaussian with the fixed sparsity $K$, where the sparse support follows a Bernoulli distribution and each non-zero entry follows a Gaussian distribution. Using a fixed dictionary based on harmonic retrieval model, the observations $\boldsymbol{y}$ are computed according to (9) with noise power $\sigma^2$. After training, we evaluate the well-trained networks on different sets of test data with varying noise power without retraining the networks.

More details on network settings and training procedures of LISTA-Toeplitz are in the Appendix, where we introduce some modifications to handle complex-valued data in Appendix A, and discuss training details in Appendix B.

### A. 1D Harmonic Retrieval With Synthetic Data

To show the effectiveness of the proposed 1D LISTA-Toeplitz network, we formulate a 1D harmonic retrieval problem in two different cases: noiseless and noisy synthetic data. We evaluate the performance of our proposed LISTA-Toeplitz and compare it with the baseline LISTA and ConvLISTA, as well as three traditional iterative algorithms including ISTA, FISTA and AMP. Here, the measurement vector $\boldsymbol{y}$ is generated using the signal model (9), where $N = 64$ samples are randomly selected from $M = 512$ measurements. The normalized frequency [0,1) is equally divided into $M = 512$ grids, making $\Delta\omega = 2\pi/M$. The number of signal components is $K = 5$ unless stated otherwise, and their amplitudes are Gaussian distributed. In training, the noise power was fixed as $\sigma^2 = 0.4$.

For ConvLISTA and LISTA-Toeplitz, we generate 1000000 trials for training and 1000 trials for verification and testing. As a benchmark, we also train a ten-layer LISTA network, which has around $10\,M^2 \approx 2 \times 10^6$ parameters to learn. To avoid overfitting, we generate $10^7$ trails for the traditional LISTA network.

First, we consider noiseless cases and examine the recovery performance using a single trial. The frequencies of the true sinusoids are set on and off the predefined frequency points, respectively, with recovery results shown in Fig. 7 and 8. In our simulation, we set 5 targets with random amplitudes, marked by circles. In the on-the-grid case, ConvLISTA can only recover 2 targets of the highest amplitude, while all the other methods (ISTA, FISTA, AMP, LISTA and LISTA-Toeplitz) successfully recover all targets. This demonstrates that forcing the linear block $\boldsymbol{W}_e\boldsymbol{y}$ to be convolutions leads to bad recovery performance in 1D harmonic retrieval problems, while imposing the Toeplitz structure constraint on $\boldsymbol{W}_g$ has negligible performance loss. We also find that LISTA-Toeplitz has close performance to other methods, successfully recovering the on-the-grid and off-the-grid sinusoids, while it has a significant reduction in network parameters compared with traditional LISTA. We also present simulation results showing that LISTA-Toeplitz is able to recover the off-the-grid frequencies. Comparing with Fig. 7,

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

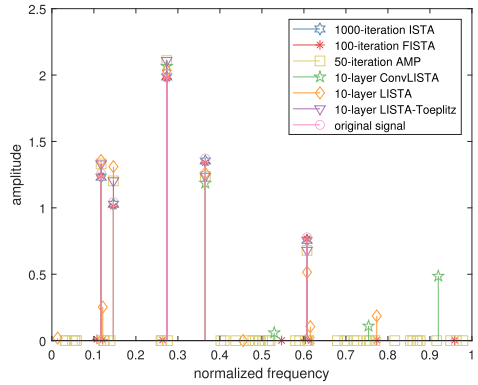


Fig. 7. Recovered results of multiple harmonic components in an on-the-grid case.

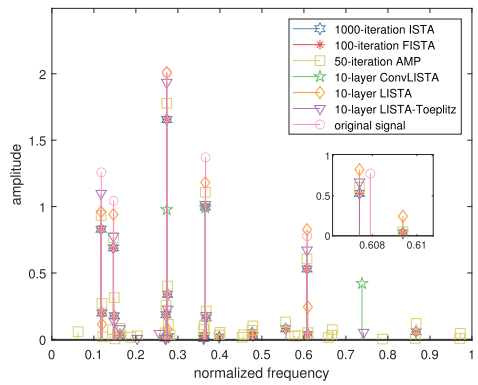


Fig. 8. Recovered results of multiple harmonic components in an off-the-grid case.

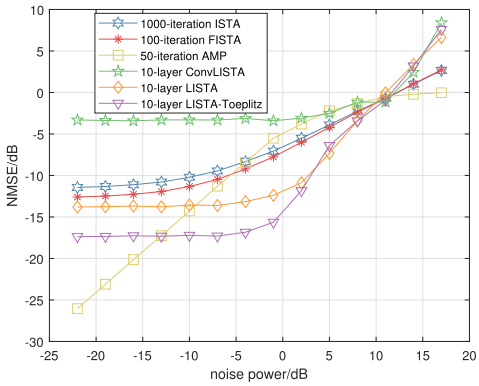


Fig. 9. The NMSEs of 1D harmonic retrieval between six methods (ISTA, FISTA, AMP, ConvLISTA, LISTA, LISTA-Toeplitz).

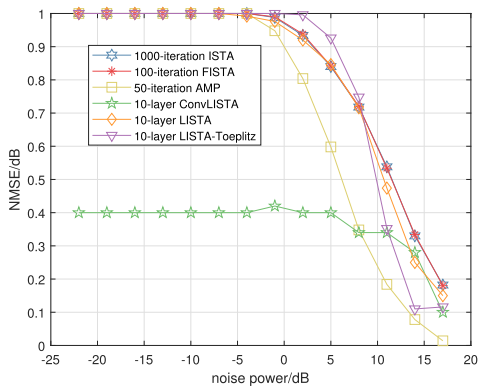


Fig. 10. The hit rates of 1D harmonic retrieval between six methods (ISTA, FISTA, AMP, ConvLISTA, LISTA, LISTA-Toeplitz).

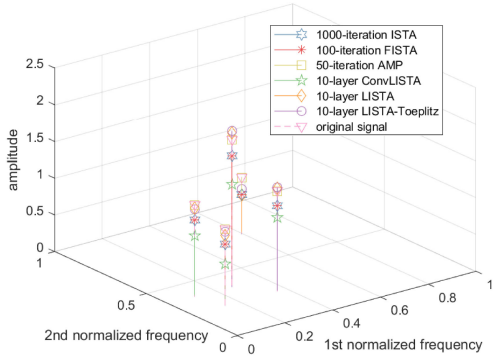


Fig. 11. Recovered 2D plane of multiple harmonic components in an on-the-grid case.

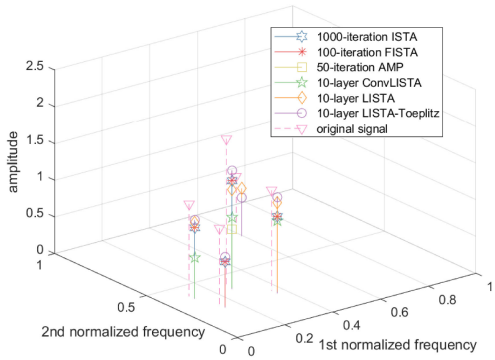


Fig. 12. Recovered 2D plane of multiple harmonic components in an off-the-grid case.

where frequencies are on-the-grid, we shift the true frequencies one quarter of the grid size away from the on-the-grid values in Fig. 8. Particularly, we set the actual frequencies to $f_k = (m_k + 1/4)/M$, where $m_k$ is the grid index nearest to the $k$-th sinusoid, $k = 1, 2, \ldots, K$. As discussed in [58], the off-the-grid case causes a mismatch between the assumed and the actual frequencies, which increases the sidelobe pedestal problem. However, the proposed LISTA-Toeplitz method successfully recovers the neighboring grid points.

Next, we consider noisy cases, where we use NMSE and hit rate as performance metrics to evaluate the tested recovery methods versus noise power. Based on the same sparse signal as the previous experiment, we change the noise power, and calculate the NMSE and hit rates by averaging 1000 Monte Carlo trials, yielding Fig. 9 and 10, respectively. As shown in Fig. 9, the NMSE of each method decreases monotonically with noise power. When the noise power is no larger than 10 dB, the

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

TABLE II
THE PARAMETERS AND INITIALIZATION OF THE LISTA, LISTA-TOEPLITZ AND CONVLISTA NETWORKS

| Network | Learned parameters | Initialization |
|---|---|---|
| LISTA | $\{\boldsymbol{W}_e^{(t)}, \boldsymbol{W}_g^{(t)}, \theta^{(t)}\}_{t=0}^{T-1}$ | $\boldsymbol{W}_e^{(t)} = \frac{1}{L}\widehat{\boldsymbol{\Phi}}^T$, $\boldsymbol{W}_g^{(t)} = \boldsymbol{0}$, $\theta^{(t)} = \frac{\lambda}{L}$ |
| 1D LISTA-Toeplitz | $\{\boldsymbol{W}_e^{(t)}, \boldsymbol{h}^{(t)}, \theta^{(t)}\}_{t=0}^{T-1}$ | $\boldsymbol{W}_e^{(t)} = \frac{1}{L}\widehat{\boldsymbol{\Phi}}^T$, $\boldsymbol{h}^{(t)} = \boldsymbol{0}$, $\theta^{(t)} = \frac{\lambda}{L}$ |
| 2D LISTA-Toeplitz | $\{\boldsymbol{W}_e^{(t)}, \boldsymbol{H}^{(t)}, \theta^{(t)}\}_{t=0}^{T-1}$ | $\boldsymbol{W}_e^{(t)} = \frac{1}{L}\widehat{\boldsymbol{\Phi}}^T$, $\boldsymbol{H}^{(t)} = \boldsymbol{0}$, $\theta^{(t)} = \frac{\lambda}{L}$ |
| 1D ConvLISTA | $\{\boldsymbol{h}_e^{(t)}, \boldsymbol{h}^{(t)}, \theta^{(t)}\}_{t=0}^{T-1}$ | $\boldsymbol{h}_e^{(t)} = \boldsymbol{0}$, $\boldsymbol{h}^{(t)} = \boldsymbol{0}$, $\theta^{(t)} = \frac{\lambda}{L}$ |
| 1D ConvLISTA | $\{\boldsymbol{h}_e^{(t)}, \boldsymbol{H}^{(t)}, \theta^{(t)}\}_{t=0}^{T-1}$ | $\boldsymbol{h}_e^{(t)} = \boldsymbol{0}$, $\boldsymbol{H}^{(t)} = \boldsymbol{0}$, $\theta^{(t)} = \frac{\lambda}{L}$ |

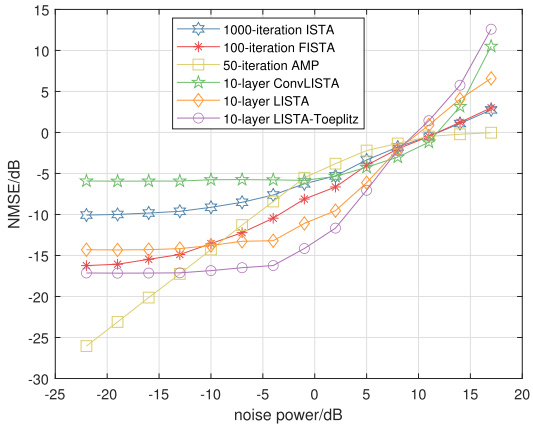


Fig. 13. Simulation results of six methods (ISTA, FISTA, AMP, ConvLISTA, LISTA, LISTA-Toeplitz) on the 2D harmonic retrieval problem in terms of NMSE.

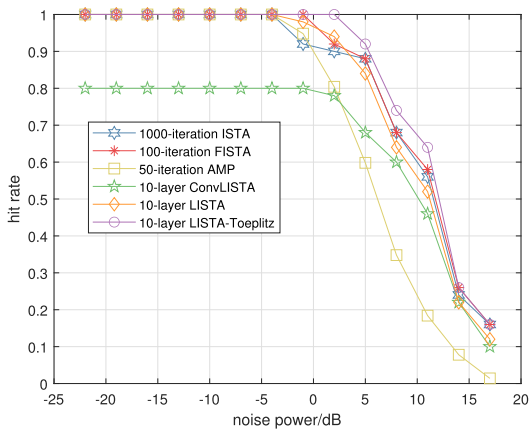


Fig. 14. Simulation results of six methods (ISTA, FISTA, AMP, ConvLISTA, LISTA, LISTA-Toeplitz) on the 2D harmonic retrieval problem in terms of hit rate.

ten-layer LISTA and LISTA-Toeplitz achieve higher accuracy than that of FISTA and ISTA, whose maximum numbers of iterations are set to 100 and 1000, respectively, much larger than the counterparts of learning-based methods. AMP can recover signals with very low NMSE especially without noise or when the noise power is less than -15 dB, but it gets worse quickly when the noise power increases while LISTA and LISTA-Toeplitz are more robust. We also find that the proposed LISTA-Toeplitz has the lowest NMSE when the noise power is between -10 dB and 0 dB, and has close performance to LISTA when the noise

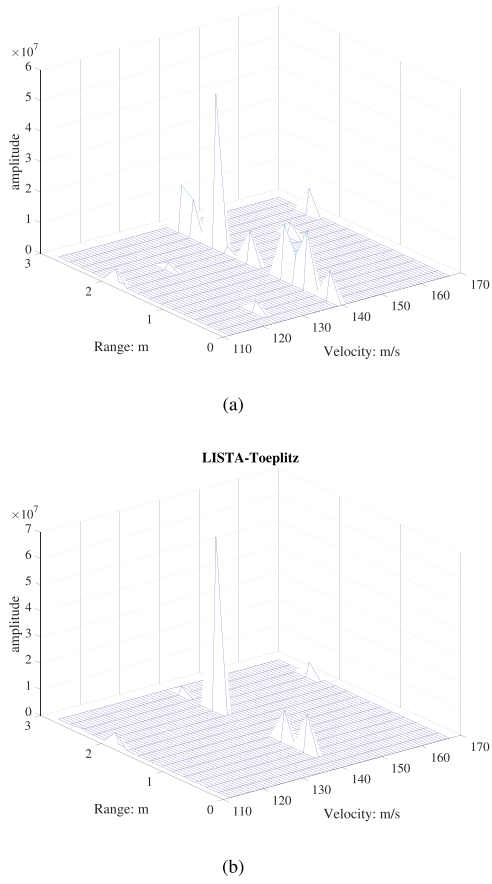


Fig. 15. The reconstructed air target using (a) ISTA and (b) LISTA-Toeplitz.

power is larger than 0 dB. As shown in Fig. 10, under reasonably small noise levels (less than 5 dB), the LISTA-Toeplitz leads to the highest hit rates, which shows its advantage in finding the frequencies. On the other hand, under noise power above 10 dB, all the methods have very poor performance as the NMSE is larger than 0 dB and hit rate is less than 0.6.

### B. 2D Harmonic Retrieval With Synthetic Data

Simulations were also conducted to show the performance of LISTA-Toeplitz for 2D harmonic retrieval problems under both noiseless and noisy cases.

In the 2D case, simulations are configured as follows. The number of full observations in two dimensions are $M_1 = 8$ and $M_2 = 64$, respectively. From these overall $M_1 M_2$ observations, we randomly select $N = 64$ samples. The number of distinct sinusoids is $K = 5$. In these simulations, we generate 1000000

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

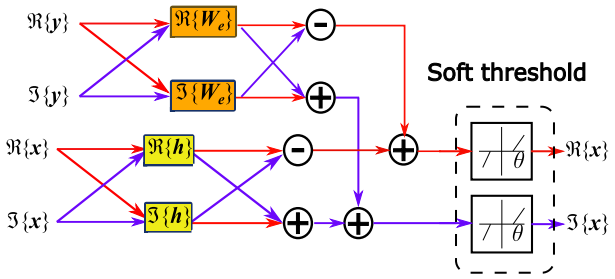


Fig. 16. Block diagrams of one layer of LISTA-Toeplitz applied to complex-valued data.

samples to train the ConvLISTA and our LISTA-Toeplitz network, while we use $10^7$ samples to train the ten-layer LISTA network. For both learned networks, we prepare 1000 samples for the validation and test sets.

Similarly to the 1D case presented in Subsection IV-A, we first provide a single noiseless trial to intuitively demonstrate the recovery output of the tested algorithms. When the frequencies are all on the grid points, the results are shown in Fig. 11. For clarity of presentation, only components with amplitudes larger than 0.5 are shown in the plot. All the algorithms except ConvLISTA work well in reconstructing sparse signals. In our simulation, we set 5 targets with random amplitudes, marked by triangles. In Fig. 11, ConvLISTA can only recover 4 targets of the highest amplitude, while all the other methods succeed in recovering all 5 targets. Thus, to guarantee recovery performance in harmonic retrieval problems, it is suggested to only make the block for the mutual inhibition part to be convolutional and leave the others as standard linear.

We also consider the case when the frequencies are off the grid. Here, the frequency values are one quarter of the grid size away from its nearest grid point along the second frequency dimension, i.e., $f_{p,k} = (m_k + 1/4)/M_p$, where $m_k$ is the grid index of the $k$-th sinusoid, $p = 2$, $k = 1, 2, \ldots, K$. Fig. 12 presents the results of an off-the-grid example, and shows that the four methods successfully find the nearest grid points corresponding to those off-the-grid ground truth. Particularly, LISTA-Toeplitz achieves comparable performance with other methods by learning a smaller number of network parameters.

We then consider noisy scenarios, and evaluate the performance of tested methods versus noise power in terms of both NMSE and hit rates, leading to results shown in Fig. 13 and 14, respectively. When the noise power is $-5$ dB, the NMSE of LISTA-Toeplitz is $-17.2$ dB, lower than the other methods. From Fig. 14, the hit rate curves of the four methods are close: The hit rates of ISTA and FISTA are almost the same, slightly lower than LISTA-Toeplitz. Similar to the 1D case, we also observe the bad NMSE and hit rate of all the methods, especially for LISTA and its variants. In comparison to original ISTA and FISTA, LISTA and LISTA-Toeplitz improve upon ISTA by achieving a similar result using one to two orders of magnitude fewer iterations, while LISTA-Toeplitz improves upon LISTA with faster training and higher reconstruction quality.

### C. Experimental Data Set for Air Target Recovery

To further investigate the performance of the 2D LISTA-Toeplitz network, experimental scenarios were carried out, where we used randomized stepped frequency radar (RSFR) to detect civil aircraft in air and estimate the range-Doppler parameters of its dominant scatterers. The configuration of the radar is introduced in [59]. Here, we formulate the range-Doppler reconstruction problem as a 2D harmonic retrieval problem, and resort to the proposed 2D LISTA-Toeplitz network. The range and Doppler domain correspond to two frequency domains, with $M_1 = 64$ and $M_2 = 16$. The observations, the radar echoes of transmitted $N = 64$ pulses, are regarded as samples of the virtual full $M_1 M_2$ observations [16].

Under such settings, we compare the 2D LISTA-Toeplitz with the conventional ISTA method, disregarding the standard LISTA network, because the LISTA network has too many variables (more than $10^7$ parameters) to learn and it is hard to train such a huge LISTA network. Based on the observation matrix of RSFR, we generate 100000 samples to train LISTA-Toeplitz. The recovered range-Doppler parameters of scatterers are indicated in Fig. 15. Both ISTA and LISTA-Toeplitz recover some scatterers laid on the grids of velocity 140 m/s, which accords with the fact that the aircraft is flying away from the radar with a radial relative velocity approximately of 140 m/s. This experiment validates the correctness of the proposed method in real applications.

## V. Conclusion

This paper considered the compressive MHR problem, which estimates frequency components in multidimensional spectra based on compressive measurements. We proposed a structured network to solve MHR problems by revealing and exploiting the Toeplitz structure in the Gram matrix of the measurement matrix. Compared with traditional LISTA, the proposed network, LISTA-Toeplitz, greatly reduces the dimension of network variables, which consequently reduces the time and space complexity and makes the network easier to train. Other convolutional extensions such as ConvLISTA which is not designed based on the specific signal model suffer major degradation in recovery performance. We presented both simulated and real data of 1D/2D harmonic retrieval problems to verify the performance of the proposed algorithms. In the numerical examples, our network shows excellent recovery performance in terms of both NMSE and hit rate, better than ConvLISTA and comparable to ISTA, FISTA and LISTA.

## Appendix

In this appendix, we illustrate some additional specifications when constructing and training a LISTA-Toeplitz network for complex-valued applications like harmonic retrieval. In Appendix VI-A, we give a detailed description of the necessary extension for our proposed network when applied to complex-valued cases. Some details on the training process are provided in Appendix VI-B, including dataset generation, parameter initialization and training strategies.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

### A. Complex-Valued Network Implementation

In Section III, we redesign the original LISTA network by converting multiplications of Toeplitz matrices to linear convolutions. Here we further extend it to a complex-valued framework, because most off-the-shelf deep learning toolboxes, e.g., PyTorch and TensorFlow, are only applicable to real-valued networks. Particularly, we transform the complex-valued linear convolution as well as matrix multiplication and non-linear operations to their real-valued counterparts.

Different from some existing complex-valued network extensions [60], [61], which separately feed a double-size real-valued network with the real and imaginary parts of the inputs, discarding the links between real and imaginary parts, we maintain the structure of the complex-valued data and operations, as discussed in the sequel.

The main difficulty is on the non-linear operator of the complex numbers in each layer, i.e., the soft-threshold operator defined in (4), which can be rewritten as

$$\mathcal{S}_\theta\left([\boldsymbol{x}]_i\right) = [\boldsymbol{x}]_i\left(1 - \frac{\theta}{\max\left(\left|[\boldsymbol{x}]_i\right|, \theta\right)}\right), \tag{23}$$

which equals 0 when $|[\boldsymbol{x}]_i| < \theta$, and yields $\mathcal{S}_\theta([\boldsymbol{x}]_i) = [\boldsymbol{x}]_i(1 - \frac{\theta}{|[\boldsymbol{x}]_i|}) = \mathrm{sign}([\boldsymbol{x}]_i)(|[\boldsymbol{x}]_i| - \theta)$ when $|[\boldsymbol{x}]_i| \geq \theta$. To be compatible with the real-valued deep learning toolboxes, we realize (23) with real-valued components. To this end, we first compute the absolute value of $\boldsymbol{x}$. Then, the real and imaginary parts of (23) are computed by

$$\begin{aligned}\Re\{\mathcal{S}_\theta\left([\boldsymbol{x}]_i\right)\} &= \Re\{[\boldsymbol{x}]_i\}\left(1 - \frac{\theta}{\max\left(\left|[\boldsymbol{x}]_i\right|, \theta\right)}\right),\\ \Im\{\mathcal{S}_\theta\left([\boldsymbol{x}]_i\right)\} &= \Im\{[\boldsymbol{x}]_i\}\left(1 - \frac{\theta}{\max\left(\left|[\boldsymbol{x}]_i\right|, \theta\right)}\right),\end{aligned} \tag{24}$$

respectively, where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and imaginary parts of a complex argument, respectively.

For the linear matrix multiplication and convolution operator, we implement their complex-valued counterpart operators through two cross coupling real-valued channels, e.g., for 2D linear convolution operator, $\boldsymbol{H} * \boldsymbol{X} = (\Re\{\boldsymbol{H}\} * \Re\{\boldsymbol{X}\} - \Im\{\boldsymbol{H}\} * \Im\{\boldsymbol{X}\}) + \mathrm{j}(\Re\{\boldsymbol{H}\} * \Im\{\boldsymbol{X}\} + \Im\{\boldsymbol{H}\} * \Re\{\boldsymbol{X}\})$.

With the above preparations, we construct each layer of the complex-valued networks with its real-valued components thus facilitating the use of off-the-shelf toolboxes. These adaptations for complex data are introduced for all the networks, including LISTA, LISTA-Toeplitz and ConvLISTA. The structure of complex-valued LISTA-Toeplitz is illustrated in Fig. 16, where each colored block is implemented by real-valued linear convolution or matrix multiplication operators.

### B. Training Details

In this appendix, we describe some details regarding training our proposed network, including three main parts: 1) generating data set, 2) initializing network parameters, and 3) training strategies.

*1) Data Preparation:* To evaluate the performance of the proposed method, we generate three data sets: training, validation and test sets. We first prepare $N_{tr}$ training samples with known labels, i.e., the ground truth of sparse signal $\boldsymbol{x}$. In these training samples, we denote by $\{\boldsymbol{y}_i\}_{i=1}^{N_{tr}}$ and $\{\boldsymbol{x}_i\}_{i=1}^{N_{tr}}$ the observation and label sets, respectively, which obey $\boldsymbol{y}_i = \boldsymbol{\Phi}\boldsymbol{x}_i$, $i = 1, 2, \ldots, N_{tr}$. Denote two matrices $\boldsymbol{Y} \in \mathbb{C}^{N\times N_{tr}}$ and $\boldsymbol{X} \in \mathbb{C}^{M\times N_{tr}}$ by

$$\begin{aligned}\boldsymbol{Y} &= \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_{N_{tr}}\end{bmatrix},\\ \boldsymbol{X} &= \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{N_{tr}}\end{bmatrix}.\end{aligned} \tag{25}$$

The choice of $N_{tr}$ is related to the initialization of network parameters, and will be discussed in the subsequent subsection.

After training, we generate another $N_{vl}$ and $N_{ts}$ samples for validation and testing, respectively. The validation data set is used for determining some hyper-parameters of the network.

*2) Network Parameter Initialization:* In LISTA-Toeplitz, each layer has a set of parameters to learn, e.g., $\{\boldsymbol{W}_e, \boldsymbol{h}, \theta^{(k)}\}$ for a 1D LISTA-Toeplitz or $\{\boldsymbol{W}_e, \boldsymbol{H}, \theta^{(t)}\}$ for a 2D counterpart. As both LISTA and LISTA-Toeplitz have $\boldsymbol{W}_e$ to learn, we keep the same initialization strategy among different networks, i.e. identically computing the initial value of the common parameter $\boldsymbol{W}_e$ while others are initialized as zeros. Following the steps in [27], we initialize the network parameter $\boldsymbol{W}_e \in \mathbb{C}^{M\times N}$ in LISTA and LISTA-Toeplitz networks by generating a coarse estimate of the dictionary matrix from training pairs, computed as

$$\widehat{\boldsymbol{\Phi}} = \boldsymbol{Y}(\boldsymbol{X}^H\boldsymbol{X})^{-1}\boldsymbol{X}^H. \tag{26}$$

Then, the learned matrix $\boldsymbol{W}_e$ is initialized as $\boldsymbol{W}_e = \frac{1}{L}\widehat{\boldsymbol{\Phi}}^T$, where $L = \lambda_{\max}(\widehat{\boldsymbol{\Phi}}^H\widehat{\boldsymbol{\Phi}})$. The initial values of $\boldsymbol{W}_g$ (for LISTA), $\boldsymbol{h}$ or $\boldsymbol{H}$ (for 1D or 2D LISTA-Toeplitz, respectively) are chosen as zeros in the experiments. They can also be initialized by random values. We summarize the initialization for each learned parameter in all the unfolded networks in Table II.

The initialization also affects the required number of training data. It has been widely accepted as a practical criterion in many published prediction modeling studies that the number of samples needs to be more than roughly ten times the degrees of freedom in the model [54]–[56]. For a $T$-layer 1D LISTA-Toeplitz network, the numbers of parameters in $\boldsymbol{W}_g$ and $\boldsymbol{W}_e$ are around $O(2MT)$ and $O(MNT)$, respectively. Here, since $\boldsymbol{W}_e$ has a good choice of initial value, it can be learnt efficiently with a small number of training samples. However, $\boldsymbol{W}_g$ is initialized as zeros or random numbers, which are updated from training data using back-propagation. Therefore, in our simulations, we generate around $N_{tr} = O(20MT)$ pairs of sparse signals and corresponding measurement signals to obtain a well-trained network.

*3) Training Protocol:* With these $N_{tr}$ labeled training data and initialized parameters, we train the network implemented by TensorFlow with the strategy described below.

To provide a trade-off between the allowed approximation error and the improved convergence speed, all LISTA-based networks tune their parameters by implementing end-to-end task-driven training, i.e., minimize the approximation error between the true signal $\boldsymbol{x}$ and the network output $\widehat{\boldsymbol{x}}$ under low computational cost. As a quantitative metric of the

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

approximation error, we use

$$\text{NMSE} = \sum_{i=1}^{N_{tr}} \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|_2 / \sum_{i=1}^{N_{tr}} \|\boldsymbol{x}_i\|_2, \tag{27}$$

where $\widehat{\boldsymbol{x}}_i$ is the output signal reconstructed by the proposed neural network with respect to the observation $\boldsymbol{y}_i$. Smaller NMSE means better reconstruction quality. In the training process, LISTA and its variants have plenty of annotated training data pairs, which means we know the true signal $\boldsymbol{x}$ corresponding to each observation $\boldsymbol{y}$. Thus, we can learn the optimal solution as close as possible to the true signal $\boldsymbol{x}$, while ISTA and FISTA algorithms are designed to solve a Lasso problem without knowledge of $\boldsymbol{x}$.

## REFERENCES


[1] R. Fu, T. Huang, Y. Liu, and Y. C. Eldar, "Compressed LISTA exploiting toeplitz structure," in *Proc. IEEE Radar Conf.*, 2019, pp. 1–6.

[2] X. Liu, N. D. Sidiropoulos, and T. Jiang, "Multidimensional harmonic retrieval with applications in MIMO wireless channel sounding," in*Space-Time Process. for MIMO Communi.*, A. Gershman and N. Sidiropoulos, Eds. New York, NY, USA: Wiley, 2005, ch. 2, pp. 41–75.

[3] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[4] J. Xiong and W. Wang, "Sparse reconstruction-based beampattern synthesis for multi-carrier frequency diverse array antenna," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 3395–3398.

[5] R. D. Balakrishnan and H. M. Kwon, "A new inverse problem based approach for azimuthal DOA estimation," in *Proc. Global Telecommun. Conf.*, vol. 4, 2004, pp. 2187–2191.

[6] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoustical Soc. Amer.*, vol. 136, pp. 260–271, Jul. 2014.

[7] N. D. Nion, D. Sidiropoulos, "Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5693–5705, Nov. 2010.

[8] T. Huang and Y. Liu, "Compressed sensing for a frequency agile radar with performance guarantees," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 1057–1061.

[9] C. Bingham, M. Godfrey, and J. Tukey, "Modern techniques of power spectrum estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 56–66, Jun. 1967.

[10] Q. Zhang, J. Zhu, N. Zhang, and Z. Xu, "Multidimensional variational line spectra estimation," *IEEE Signal Process. Lett.*, vol. 27, no. 15, pp. 945–949, May 2020.

[11] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.

[12] M. Pesavento, "Exploiting multiple shift invariances in harmonic retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 2101–2104.

[13] L. Ma, Y. Jin, and L. Wang, "A cross four-order cumulants based harmonic retrieval tls-esprit method in the estimation of harmonic parameters," in *Proc. 2nd Int. Conf. Mechanic Automat. Control Eng.*, 2011, pp. 12–15.

[14] Y. Chi and Y. Chen, "Compressive two-dimensional harmonic retrieval via atomic norm minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1030–1042, Feb. 2015.

[15] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, Nov. 2014.

[16] T. Huang, Y. Liu, X. Xu, Y. C. Eldar, and X. Wang, "Analysis of frequency agile radar via compressed sensing," *IEEE Trans. Signal Process.*, vol. 66, no. 23, pp. 6228–6240, Dec. 2018.

[17] D. Cohen and Y. C. Eldar, "Sub-nyquist radar systems: Temporal, spectral, and spatial compression," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 35–58, Nov. 2018.

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

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

[20] H. Fang and H. R. Yang, "Greedy algorithms and compressed sensing," *Acta Automatica Sinica*, vol. 37, no. 12, pp. 1413–1421, 2011.

[21] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, "Proximal gradient algorithms: Applications in signal processing," 2018, *arXiv:1803.01621*.

[22] A. Draganic, I. Orovic, and S. Stankovic, "On some common compressive sensing recovery algorithms and applications - review paper," *Facta Universitatis - Ser.: Electron. Energetics*, vol. 30, pp. 477–510, Apr. 2017.

[23] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *Siam J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[24] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model Simul.*, vol. 4, no. 4, pp. 1168–1200, 2006.

[25] K. Gregor and Y. Lecun, "Learning fast approximations of sparse coding," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 399–406.

[26] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 2017.

[27] M. Borgerding and P. Schniter, "Onsager-corrected deep learning for sparse linear inverse problems," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2016, pp. 227–231.

[28] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[29] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-Net for compressive sensing MRI," in *Advances in Neural Information Processing Systems* 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 10–18.

[30] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," in *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, 2009.

[31] H. Sreter and R. Giryes, "Learned convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2191–2195.

[32] R. J. G. van Sloun, R. Cohen, and Y. C. Eldar, "Deep learning in ultrasound imaging," in *Proc. IEEE*, vol. 108, no. 1, pp. 11–29, Jan. 2020.

[33] G. Dardikman-Yoffe and Y. C. Eldar, "Learned SPARCOM: unfolded deep super-resolution microscopy," *Opt. Exp.*, vol. 28, no. 19, pp. 27 736–27 763, Sep. 2020.

[34] S. Liao, A. Samiee, C. Deng, Y. Bai, and B. Yuan, "Compressing deep neural networks using toeplitz matrix: Algorithm design and fpga implementation," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 1443–1447.

[35] A. Araujo, B. Negrevergne, Y. Chevaleyre, and J. Atif, "Understanding and training deep diagonal circulant neural networks," in *Proc. Eur. Conf. Artificial Intell.*, 2020.

[36] V. Sindhwani, T. N. Sainath, and S. Kumar, "Structured transforms for small-footprint deep learning," in *Proc. 28th Int. Conf. Neural Inform. Process. Syst.*, vol 2, Cambridge, MA, USA: MIT Press, pp. 3088–3096.

[37] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S. F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2857–2865.

[38] L. Zhao, S. Liao, Y. Wang, J. Tang, and B. Yuan, "Theoretical properties for neural networks with weight matrices of low displacement rank," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 4082–4090.

[39] H. Zhang and F. Ding, "On the Kronecker products and their applications," *J. Appl. Math.*, Jun. 2013, doi: 10.1155/2013/296185.

[40] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Stat. Society: Ser. B (Stat. Methodol.)*, vol. 73, no. 3, pp. 267–288, 2011.

[41] T. Hastie, R. Tibshirani, J. H. Friedman, and J. Franklin, "The elements of statistical learning: Data mining, inference and prediction. by," *Math. Intelligencer*, vol. 27, no. 2, pp. 83–85, 2009.

[42] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629–654, 2008.

[43] F. Chen, L. Shen, B. W. Suter, and Y. Xu, "Nesterov's algorithm solving dual formulation for compressed sensing," *J. Comput. Appl. Math.*, vol. 260, pp. 1–17, 2014.

[44] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proc. Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[45] R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro, "Tradeoffs between convergence speed and reconstruction accuracy in inverse problems," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1676–1690, Apr. 2018.


Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.
[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[48] P. A. Roebuck and S. Barnett, "A survey of toeplitz and related matrices," *Int. J. Syst. Sci.*, vol. 9, no. 8, pp. 921–934, 1978.
[49] M. Wax and T. Kailath, "Efficient inversion of doubly block toeplitz matrix," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 8, 1983, pp. 170–173.
[50] M. Oudin and J. P. Delmas, "Asymptotic generalized eigenvalue distribution of toeplitz-block-toeplitz matrices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 3309–3312.
[51] G. Nguyen *et al.*, "Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey," *Artif. Intell. Rev.*, vol. 52, pp. 77–124, Jan. 2019.
[52] T. Hope, Y. S. Resheff, and I. Lieder, *Learning TensorFlow: A Guide to Building Deep Learning Systems*, 1st ed. O'Reilly Media, 2017.
[53] T. E. Potok, C. Schuman, S. R. Young, R. M. Patton, and G. Chakma, "A study of complex deep learning networks on high performance, neuromorphic, and quantum computers," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 14, no. 2, pp. 1–21, 2018.
[54] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 76–92, Aug. 2016.
[55] M. van Smeden *et al.*, "Sample size for binary logistic prediction models: Beyond events per variable criteria," *Stat. Methods Med. Res.*, vol. 28, no. 8, pp. 2455–2474, 2019, pMID: 29966490.
[56] S. Lei, H. Zhang, K. Wang, and Z. Su, "How training data affect the accuracy and robustness of neural networks for image classification," in *Proc. Int. Conf. Learn. Representations*, 2019.
[57] X. Suau, u. Zappella, and N. Apostoloff, "Filter distillation for network compression," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 3129–3138.
[58] Y. Chi, L. Scharf, A. Pezeshki, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, pp. 2182–2195, Jun. 2011.
[59] L. Wang, T. Huang, and Y. Liu, "Randomized stepped frequency radars exploiting block sparsity of extended targets: A theoretical analysis," *IEEE Trans. Signal Process.*, vol. 69, no. 12, pp. 1378–1393, Feb. 2021.
[60] R. Hänsch, "Complex-valued multi-layer perceptrons-an application to polarimetric SAR data," *Photogrammetric Eng. Remote Sens.*, vol. 76, Sep. 2010, Art no. 1081.
[61] R. Haensch and O. Hellwich, "Complex-valued convolutional neural networks for object detection in polsar data," in *Proc. 8th Eur. Conf. Synthetic Aperture Radar*, Jun. 2010, pp. 1–4.

**Rong Fu** received the B.S. degree from Beihang University, Beijing, China in 2012. She is currently working toward the Ph.D. degree with the Intelligence Sensing Laboratory, Department of Electronic Engineering, Tsinghua University, Beijing, China. Her current research interests include statistical signal processing, compressive sensing, optimization methods, model-based deep learning, and their applications in signal detection and parameter estimation.

**Yimin Liu** (Member, IEEE) received the B.S. and Ph.D. degrees (both with Hons.) in electronics engineering from Tsinghua University, Beijing, China, in 2004 and 2009, respectively. In 2004, he was with the Intelligence Sensing Lab, Department of Electronic Engineering, Tsinghua University. He is currently an Associate Professor with Tsinghua, where he is studying on new concept radar and other microwave sensing technologies. His current research interests include radar theory, statistic signal processing, compressive sensing and their applications in radar, spectrum sensing, and intelligent transportation systems.

**Tianyao Huang** received the B.S. degree in telecommunication engineering from the Harbin Institute of Technology, Harbin, China, in 2009 and the Ph.D degree in electronics engineering from the Tsinghua University, Beijing, China, in 2014. From 2014 to 2017, he was a radar Researcher with the Aviation Industry Corporation of China. In July 2017, he joined the Intelligent Sensing Lab, Department of Electronic Engineering, Tsinghua University, as an Assistant Professor. His current research interests include signal processing, compressed sensing, and joint radar communications system design.

**Yonina C. Eldar** (Fellow, IEEE) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2002.

She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She was previously a Professor with the Department of Electrical Engineering, Technion, where she held the Edwards Chair in engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, and an Adjunct Professor with Duke University, Durham, NC, USA, and was a Visiting Professor with Stanford. She is a Member of the Israel Academy of Sciences and Humanities (elected 2017) and a EURASIP Fellow. She is also the author of the book *Sampling Theory: Beyond Bandlimited Systems* and coauthor of four other books published by Cambridge University Press. Her research interests include statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging, and optics.

She was the recipient of 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 with the Technion-Israel Institute of Technology, Haifa, Israel and an Alon Fellow. She was the recipient of 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 with the Technion, the Muriel and David Jacknow Award for Excellence in Teaching, and the Technions Award for Excellence in Teaching (two times). She was also the recipient of 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 and in Asia. She is also a highly cited Researcher. 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 the *Foundations and Trends in Signal Processing*, a Member of the IEEE Sensor Array and Multichannel Technical Committee and was on several other IEEE committees. She was a Signal Processing Society Distinguished Lecturer, a Member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and was 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 the Co-Chair and the Technical Co-Chair of several international conferences and workshops. She is a Member of the Israel Academy of Sciences and Humanities (elected 2017) and a EURASIP Fellow.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 04,2021 at 06:39:02 UTC from IEEE Xplore. Restrictions apply.